



The USPTO Genetic Sequence Database, USGENE[®], on STN[®]

Martin Goffman – SequenceBase Corporation
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 sequence searchable databases

- **REGISTRY**
 - Chemical Abstracts Service (CAS) Registry File
- **DGENE**
 - Thomson Scientific GENESEQ™
- **PCTGEN**
 - WIPO/PCT Patent Application Biosequences
- **USGENE®**
 - The USPTO Genetic Sequence Database

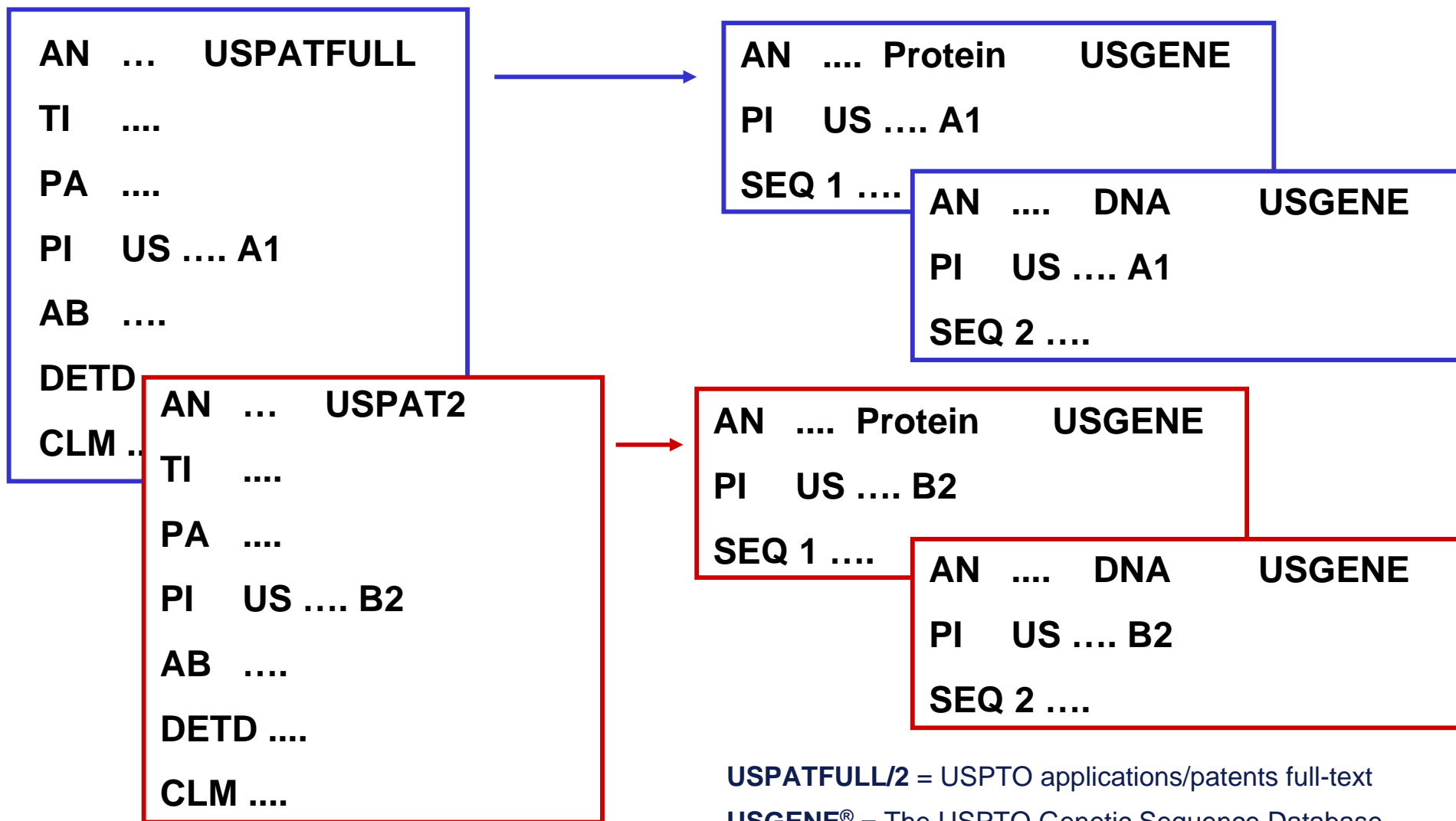
See *Effective patent sequence searching on STN*:

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

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

Relationship between USPATFULL/2 and USGENE databases



USPATFULL/2 = USPTO applications/patents full-text

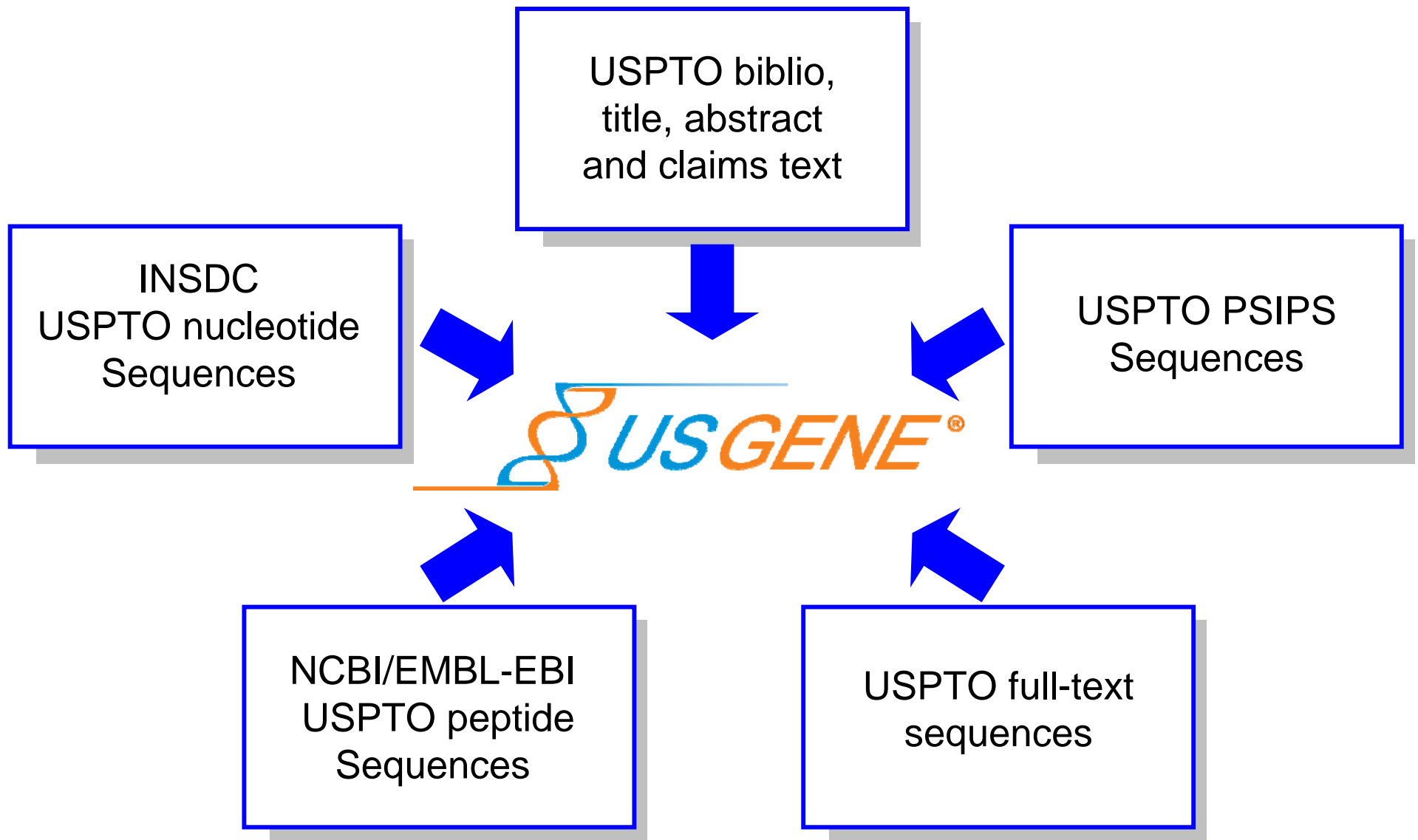
USGENE® = The USPTO Genetic Sequence Database

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.

USGENE combines these sequences with bibliographic data and claims text



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)
IN Simmonds Peter (Edinburgh, GB)
Yap Peng Lee (Edinburgh, GB) (4)
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 described, together with those of a newly discovered type 6. Unique
(7) 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.

ALL display format.

See (1) - (7) on slide 11.

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) QPAVIPDREVLVYQQFDEM (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-QPAVIPDREVLVYQQFDEN].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

All USGENE sequences are provided in STN standardized format

```
SSO  PROTEIN; USPTO; GRANTED (9)
ORGN  Hepititis C Virus (10)
SQL   19 (11)
SEQ
      1 ecaskaalie egqrmaeml (12)
```

ALL display format (cont.)

```
FEATURE TABLE: (13)
Key      |Location |
=====+=====+=====
VARIANT  |(0)...(0)|HCV TYPE 2 NS4 REGION
```

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

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

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

In contrast, EMBL patent records have minimal bibliographic and text data

General Information			
Accession #	AAA00521		
SRS Entry ID	USPO_PRT:AAA00521		
Molecule Type	PRT		
Sequence Length	40		
Entry Data Class	STANDARD		
Sequence Version	AAA00521.1		
Creation Date	21-MAY-1993		
UniParc	UPI0000035113		
Description			
Description	Sequence 1 from Patent US 4563352.		
Organism	Unknown		
References			
1.	Rivier; J.E.F.; Spiess; J. and Vale; W.W. Jr.; Human pancreatic GRF Patent number US4563352 -A/1 07-JAN-1986; The Salk Institute For Biological Studies; San Diego, CA Position 1-40		
Features			
Key	Location	Qualifier	Value
source	1..40		
Sequence			
Characteristics	Length: 40 AA		
Sequence	<pre>>uspo_prt AAA00521 AAA00521 Sequence 1 from Patent US 4563352. YADAIFTNSYRKVLGQLSARKLLQDIMSRQQGESNQERGA</pre>		

EMBL USPTO peptide sequence records, like this one, are also available at NCBI, but not at DDBJ.

PSIPS sequence records also have minimal bibliographic and text data

United States Patent and Trademark Office

[Home](#) | [Site Index](#) | [Search](#) | [FAQ](#) | [Glossary](#) | [Guides](#) | [Contacts](#) | [eBusiness](#) | [eBiz alerts](#) | [News](#) | [Help](#)

Publication Site for Issued and Published Sequences (PSIPS)

PSIPS View Sequence(s): 2 for 6825322

Here is the list of the requested sequences.

Sequence listings provided at PSIPS are those which exceed 300 pages in length.

Sequence ID No:

```

(2) INFORMATION FOR SEQ ID NO: 2:
  (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 938 amino acids
      (B) TYPE: amino acid
      (D) TOPOLOGY: linear
  (ii) MOLECULE TYPE: protein
  (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:
Met Ser Thr Met Arg Leu Leu Thr Leu Ala Leu Leu Phe Ser Cys Ser
  1          5          10          15
Val Ala Arg Ala Ala Cys Asp Pro Lys Ile Val Asn Ile Gly Ala Val
          20          25          30
Leu Ser Thr Arg Lys His Glu Gln Met Phe Arg Glu Ala Val Asn Gln
          35          40
Ala Asn Lys Arg His Gly Ser Trp
          50          55
Val Thr His Lys Pro Asn Ala Ile
          65          70
Asp Leu Ile Ser Ser Gln Val Tyr
          85
Thr Pro Asn Asp His Phe Thr Pro
          100
Phe Tyr Arg Ile Pro Val Leu Gly
          115          120
Ser Asp Lys Ser Ile His Leu Ser Phe Leu Arg Thr Val Pro Pro Tyr
          130          135          140
Ser His Gln Ser Ser Val Trp Phe Glu Met Met Arg Val Tyr Ser Trp
          145          150          155          160
Asn His Ile Ile Leu Leu Val Ser Asp Asp His Glu Gly Arg Ala Ala
          
```

[First Sequence](#)
[Next Sequence](#)
[Previous Sequence](#)
[Last Sequence](#)

[Full Text Patent](#)
[PSIPS Home Page](#)
[NCBI Home](#)
[PIW and AIW Search Home Page](#)
[Document Services Division](#)
[USPTO Home](#)

[Help Page](#)
[FAQ](#)

Sequences are either available in WIPO ST.25 format, or the previous USPTO standard format.

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 CAplusSM 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**

Example: sequence listing variation between patent family members

```
L1 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 19951220 (199508) EN
EP 698101 A1 19960228 (199508) EN
JP 09500009 W 19970107 (199508) EN
AU 695259 B 19980811 (199508) EN
EP 698101 B1 20041101 (199508) EN
DE 69434116 E 20041209 (199508) EN
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

Agenda

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

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

The 7 basic 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

The 7 basic steps of USGENE BLAST

1) SAVE, UPLOAD and VERIFY the sequence query text file (L1)

➤ Upload options

- STN Express: Use UPLOAD command or Upload Query Wizard (STN Express 8.01+)
- STN on the Web: Use Upload feature or Sequence Assistant (link below)

➤ Verify the sequence with D LQUE

STN on the Web Sequence Search Assistant:

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

Requirements for sequences for the STN Express Upload Query Wizard

- Sequence queries must be saved individually in text (.txt) format
- Files may
 - be 3 letter codes (amino acids) or single letter
 - have header information as seen in, e.g. WIPO ST.25, USPTO PSIPS or EMBL formats
 - include sequence count numbers
- Query (.txt) files must
 - be 10,000 characters or less
 - not have any lines longer than 300 characters
- After upload to STN verify with D LQUE

Examples of formats that work

DETD SEQUENCE CHARACTERISTICS:

SEQ ID NO: 4

LENGTH: 724

USPATFULL/USPAT2 format

TYPE: PRT

ORGANISM: Artificial Sequence

FEATURE:

OTHER INFORMATION: Description of Artificial Sequence; Note = synthetic construct

SEQUENCE: 4

Met Ser Phe Val Asp His Pro Pro Asp Trp Leu Glu Glu Val Gly Glu

1

5

Gly Leu Arg Glu Phe Leu

20

<210> SEQ ID NO 137

<211> LENGTH: 951

<212> TYPE: DNA

USPTO PSIPS ST.25 format

<213> ORGANISM: Zea mays

<400> SEQUENCE: 137

accgagccg acttccggtt cactggccac gacgggacgt gcatctcaa actgaaaaat 60

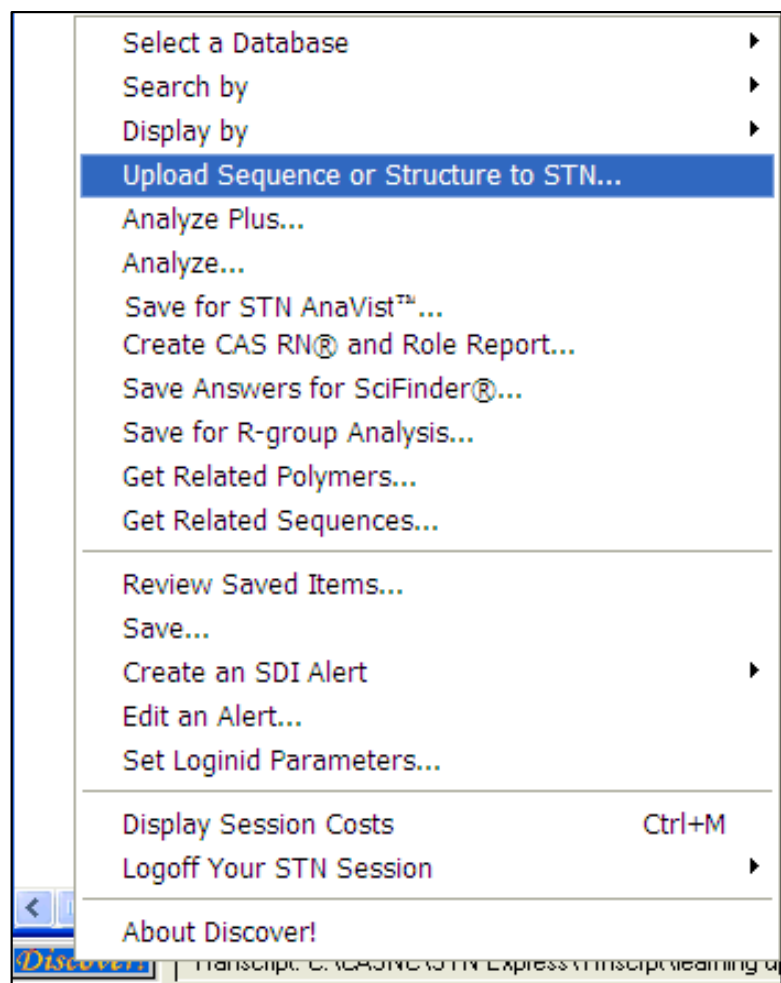
acaagggttg tatecataga ttcttctgag cgtgtgcca tcaactacga gagagcgtg 120

cagaagccg tggcgacca gctgttagt gccagcattg aagcatctcg gcgcgcgttc 180

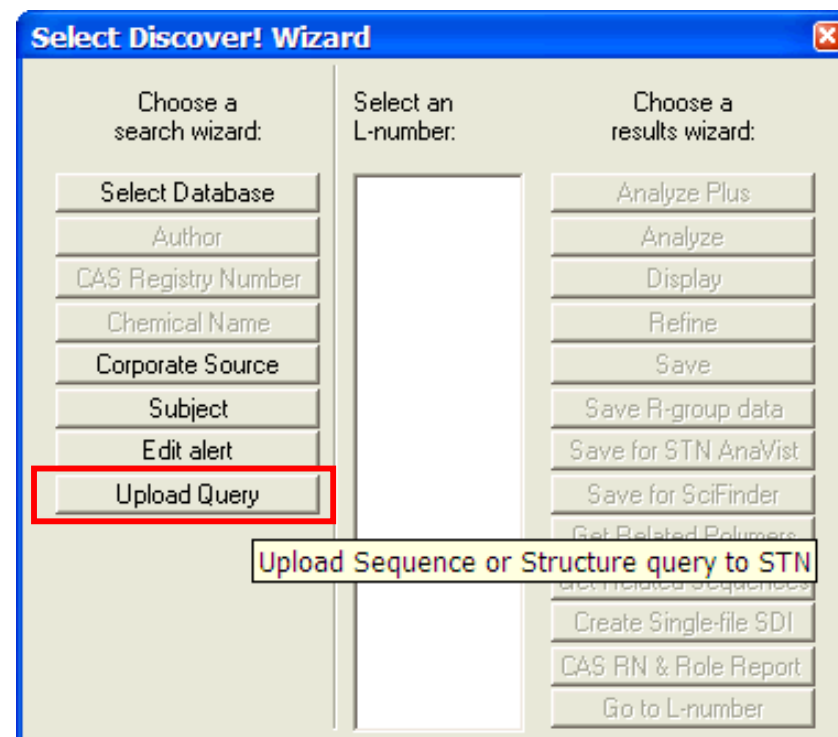
cagctctaca gttctggcat ctctgacggg agatgcggga cgtacctgga ccacggtgtg 240

a) Choose the Upload Query Wizard

From the Discover! button menu

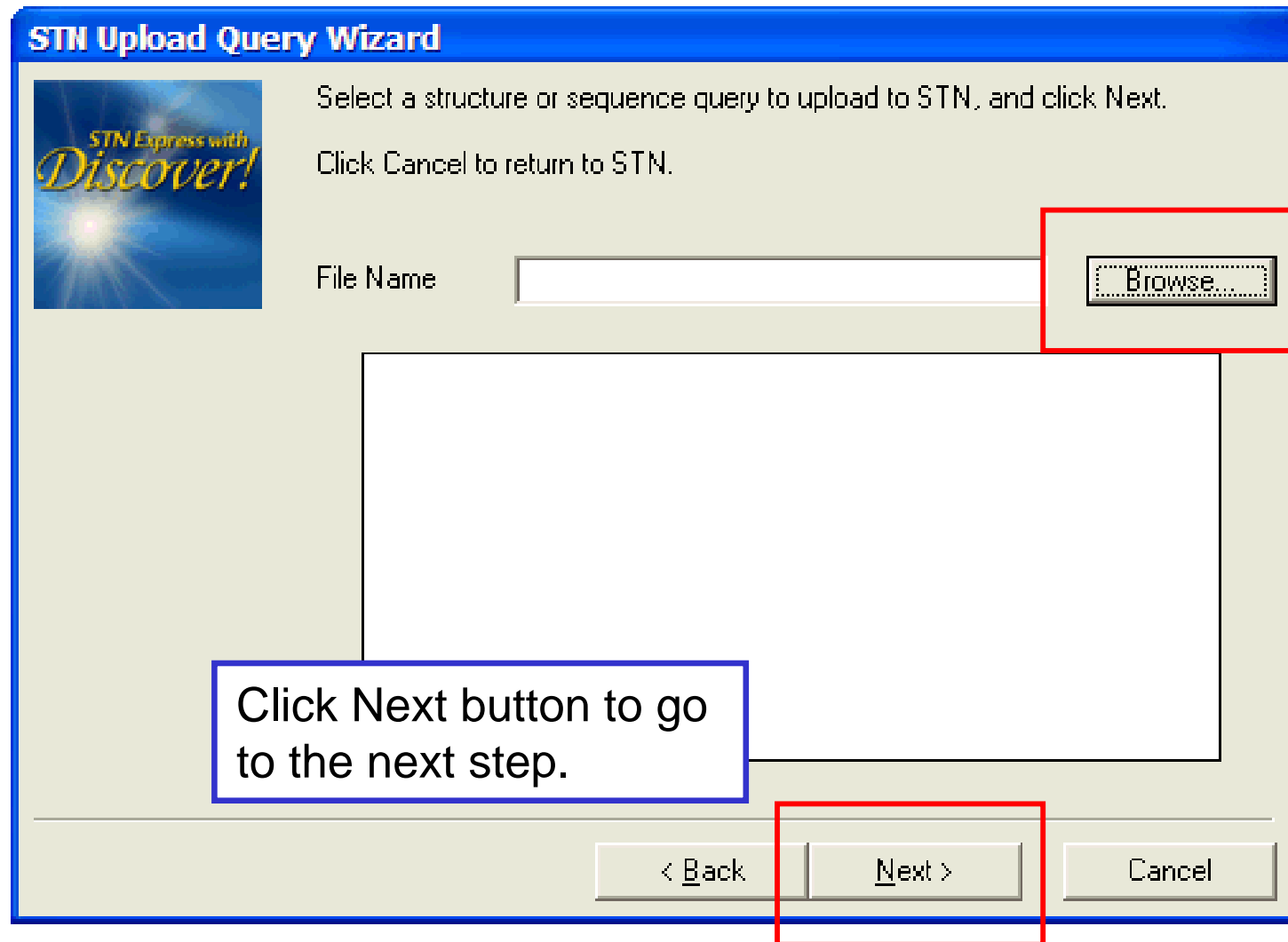


OR

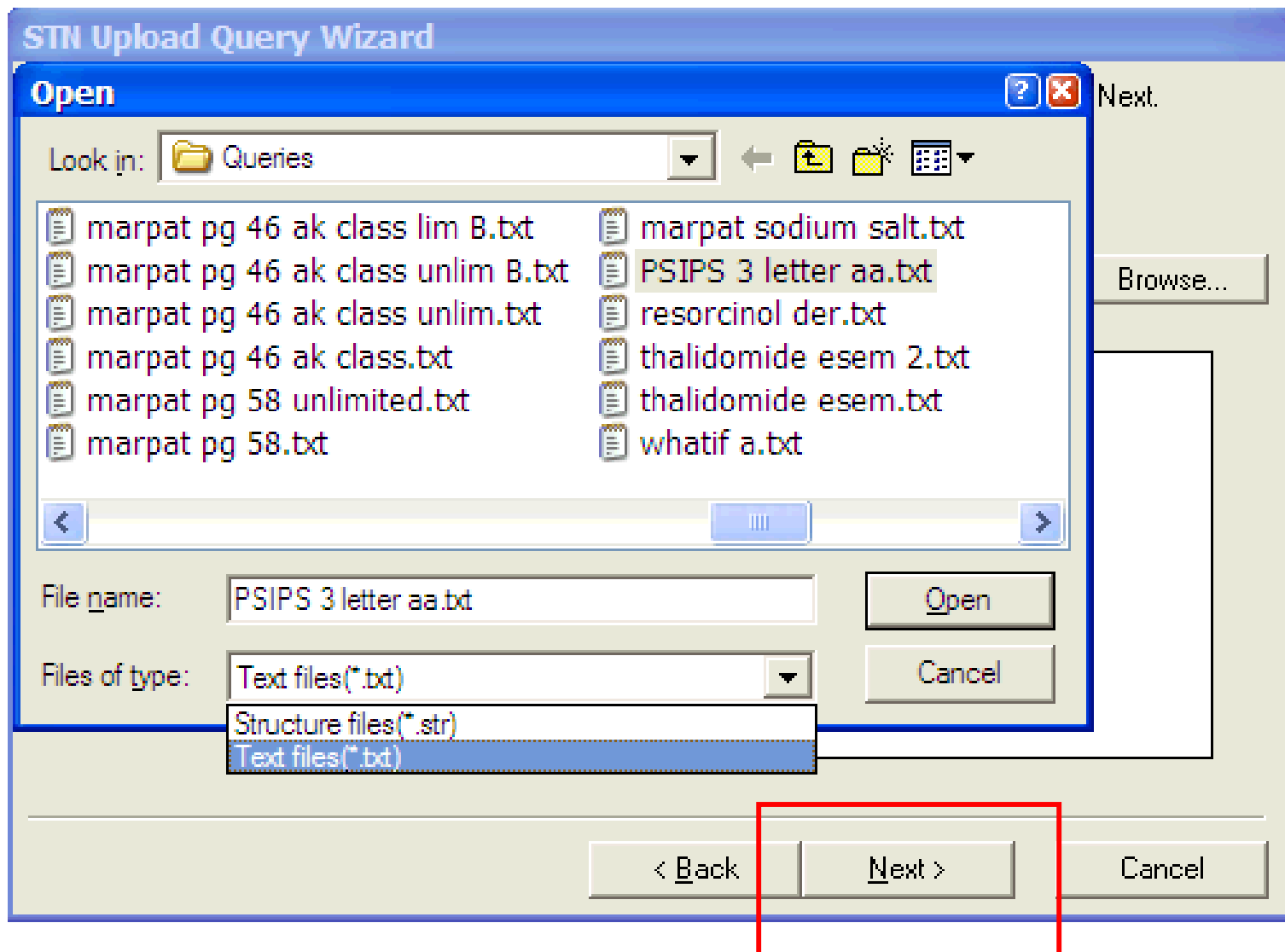


From the Select Discover!
Wizard window

b) Browse to locate sequence file



c) Change File type to .txt



d) Verify it's the right query!

STN Upload Query Wizard

Select a structure or sequence query to upload to STN, and click Next.
Click Cancel to return to STN.

File Name

```
1 vqtvplsrif dhamleahra helaidtyqe feetyipkdq  
kysflhdsqt  
51 sfcfsdsipt psnmeetqgk snlellrisl llieswlepv  
rflrsmfann  
101 lvydtsdsdd yhlkdleeg iqtlmgrled gsrrtgqilk  
qtyskfdtns  
151 hnhdallkny gllycfrkdm dkvetflrmv qcrsvegscg f
```

e) Select STN file to upload to

STN Upload Query Wizard

Select a database from the list below, or more than one database by holding the Ctrl key while making your selection, and click Finish

To exit the Wizard click Cancel.

Databases:

DGENE	Derwent Geneseq Database 1981 - present
PCTGEN	World Patent Application Biosequences

View this Database Summary Sheet on the Web

< Back Finish Cancel

Use PCTGEN to upload queries and verify them (lower connect hour). The resulting L-numbers may be searched in DGENE, PCTGEN or USGENE.

Click Finish for the file to be “scrubbed” and uploaded to STN.

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 vqtvplsrlfdhamleahrahelaidtyqefeetyipkdqkysflhdsqtsfcfsdsi  
ptpsnmeetqqksnlellrislllieswlepvrflrsmfannlvdydtsdsddyhllkd  
leegiqtlmgrledgsrrtgqilkqtyskfdtnshnhdallknygllycfrkdmdkve  
tflrmvqcrsvegscgf
```

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

```
=>
```

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

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.

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)

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)

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.

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)

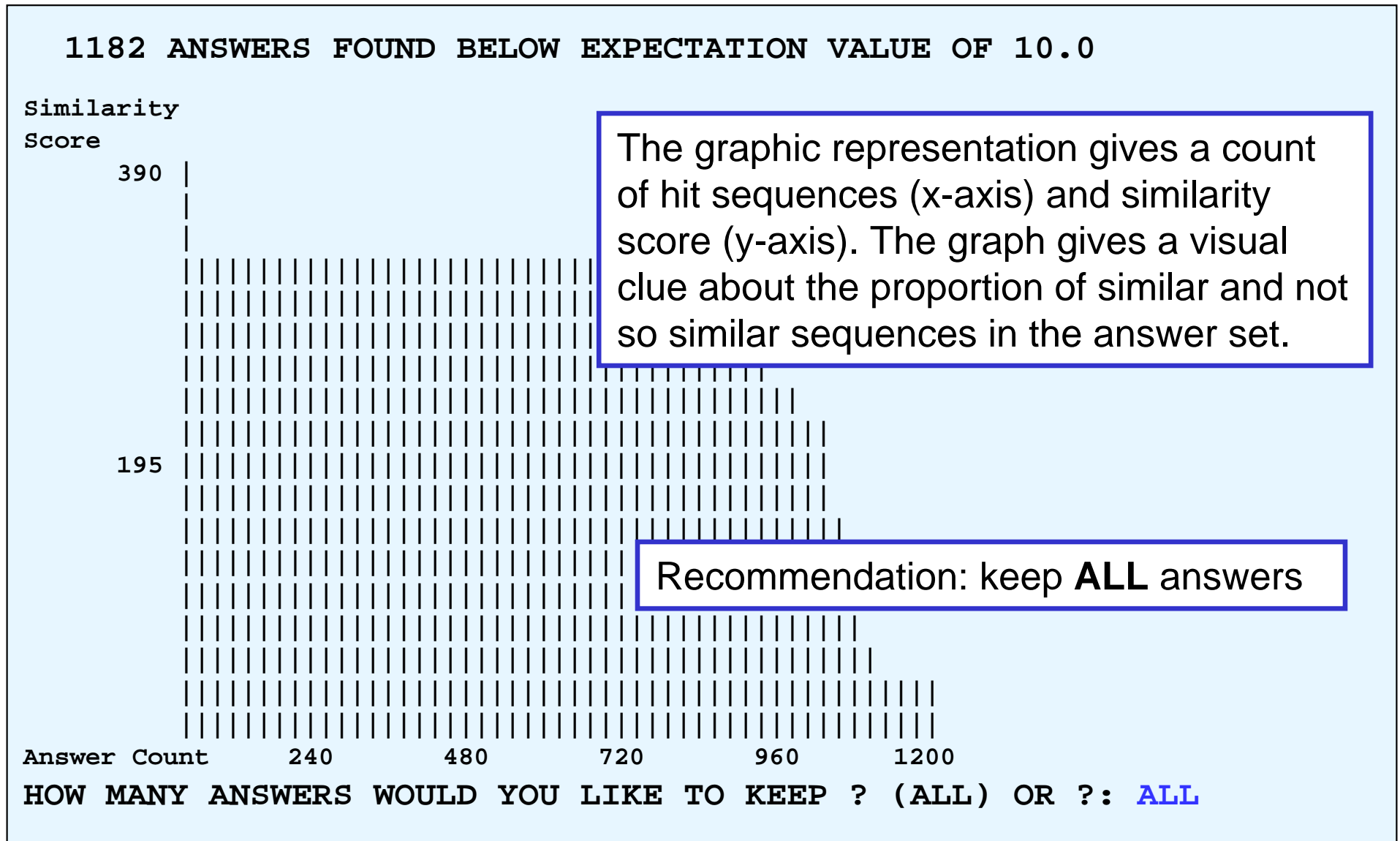
Guidelines from NCBI on the use of Advanced Settings for peptide sequence searching are as follows:

<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)

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

3) Decide how many answers to keep



The 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)

4) SORT by SCORE descending

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

L2 RUN STATEMENT CREATED

```
L2      1182 VQTVPLSRLFDHAMLEAHRAHEL AIDTYQEF EETYIPKDQKYSFLHDSQT
          SFCFSDSIPTPSNMEETQQKSNLELLRISLLLLIESWLEPVRFLRSMFANN
          LVYDTSDDYHLLKDL EEGIQTLMGRLEDGSRRTGQILKQTYSKFDTNS
          HNHDALLKNYGLLYCFR KDMKVETFLRMVQCRSVEGSCGF/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.

The 7 basic steps of USGENE BLAST

- 5) Review answers using a *free-of-charge* format including alignment (ALIGN), while “parked” in the STNGUIDE file
 - D L5 TRI ORGN ALIGN 1-
 - FILE STNGUIDE

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
```

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

```
BLASTALIGN
```

```
Query   = 191 letters
```

```
Length  = 191
```

```
Score   = 387 bits (995), Expect = e-113
```

```
Identities = 189/191 (98%), Positives = 191/191 (99%)
```

```
Query: 1  VQTVPLSRLFDHAMLEAHRAHELAIIDTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
          VQTVPLSRLFDHAML+AHRAH+LAIDTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
Sbjct: 1  VQTVPLSRLFDHAMLQAHRHQLAIDTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
Query: 61 PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDDSDDYHLLKDLEEG
          PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDDSDDYHLLKDLEEG
Sbjct: 61 PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDDSDDYHLLKDLEEG
. . . .
```

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  VQTVPLSRLFDHAMLEAHRAHELAIPTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
        VQTVPLSRLFDHAML+AHRAH+LAIDTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
Sbjct:  1  VQTVPLSRLFDHAMLQAHRAHQLAIDTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
Query:  61  PSNMEETQQKSNLELLRISLLLIESW
        PSNMEETQQKSNLELLRISLLLIESW
Sbjct:  61  PSNMEETQQKSNLELLRISLLLIESW
Query:  121 IQTLMGRLEDGSRRTGQILKQTYSKFDTNSHNHDALLKNYGLLYCFRKDMDKVETFLRMV
        IQTLMGRLEDGSRRTGQILKQTYSKFDTNSHNHDALLKNYGLLYCFRKDMDKVETFLRMV
Sbjct:  147 IQTLMGRLEDGSRRTGQILKQTYSKFDTNSHNHDALLKNYGLLYCFRKDMDKVETFLRMV
        . . . .
```

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

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

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.)

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 VQTVPLSRLFDHAMLEAHRAHELAIPTYQEFEEITYIPKDQKYSFLHDSQT
          SFCFSDSIPTPSNMEETQOKSNLELLLRISLLLIESWLEPVRFLRSMFANN
          LVYDTSDDYHLLKDLLEGIQTLMGRLLEDGSRRTGQILKQTYSKFDTNS
          HNHDALLKNYGLLYCFRKDMDKVETFLRMVQCRSVEGSCGF/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!

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

6) 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.

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

The importance of using the correct BLAST advanced options

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

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

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

```
=> 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".
```

Changing BLAST options is especially important for short sequence queries!

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  GSSFLSPEHQR 11
        GSSFLSPEHQR
Sbjct: 24 GSSFLSPEHQR 34
```

Exploring USGENE search fields

- USGENE is similar in design to DGENE, but has a number of unique additional search fields

/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 Exemplary Claim (/ECLM), e.g. using
 - => S VIRUS/BI,ECLM
 - => SET SFIELDS BI ECLM

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
BRIEF	BIB + AB + ECLM, sequence, sequence source (SSO), feature table (FEAT)
ALL	BRIEF with CLM instead of ECLM

(* Free of charge display formats in USGENE.)

USGENE Original Sequence (SEQO)

=> D SEQO

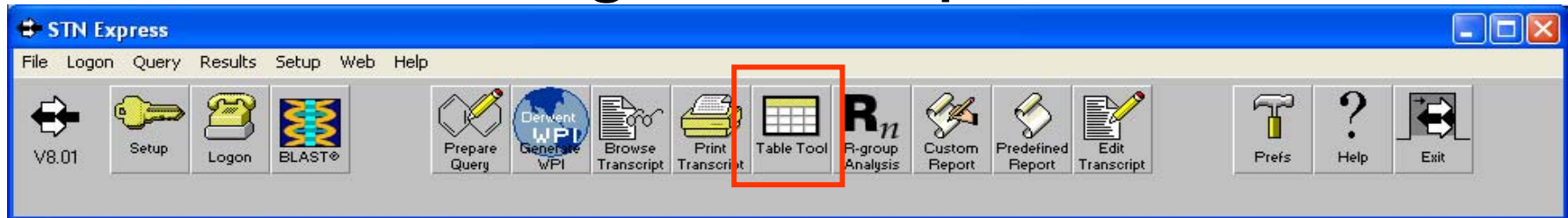
L1 ANSWER 1 OF 1 USGENE COPYRIGHT
SEQO

```
cgctcgcagt ctgtgggccc tccgggaggc ggcggaggtc accgcgggga gaggggcggg      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
ctg ggc aac atg ggg aat cca atg gc
Leu Gly Asn Met Gly Asn Pro Met Ala
                50                50
.....
```

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.

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



STN Online and Results - [Table Output - USGENE POST PROCESS.1b1]











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 methods of making the albumin fusion proteins of the invention and using 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 diseases, 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:1038, 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 letters Length = 28 Score = 26.9 bits (58), Expect = 5e-06 Identities = 11/11 (100%), Positives = 11/11 (100%) Query: 1 GSSFLSPEHQR 11 GSSFLSPEHQR Sbjct: 1 GSSFLSPEHQR 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: 129 or the complement thereof; (b) the full-length coding sequence of the nucleic acid of SEQ ID NO: 129 or the complement thereof; (c) the full-length coding sequence of the cDNA deposited at ATCC accession number 203867 or the complement thereof.	Query = 11 letters Length = 117 Score = 26.9 bits (58), Expect = 2e-05 Identities = 11/11 (100%), Positives = 11/11 (100%) Query: 1 GSSFLSPEHQR 11 GSSFLSPEHQR Sbjct: 24 GSSFLSPEHQR 34	38
7160993.442 Protein USGENE	Nucleic acids encoding PRO4400 polypeptides	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	US7160993 B2: What is claimed is:1. A nucleic acid molecule encoding the nucleic acid sequence of	Query = 11 letters Length = 117 Score = 26.9 bits (58), Expect = 2e-05 Identities = 11/11 (100%), Positives = 11/11 (100%) Query: 1 GSSFLSPEHQR 11 GSSFLSPEHQR Sbjct: 24 GSSFLSPEHQR 34	38

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



Agenda

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

How does USGENE compare to other USPTO sequence data sources?

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

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	

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 (Part I)*:

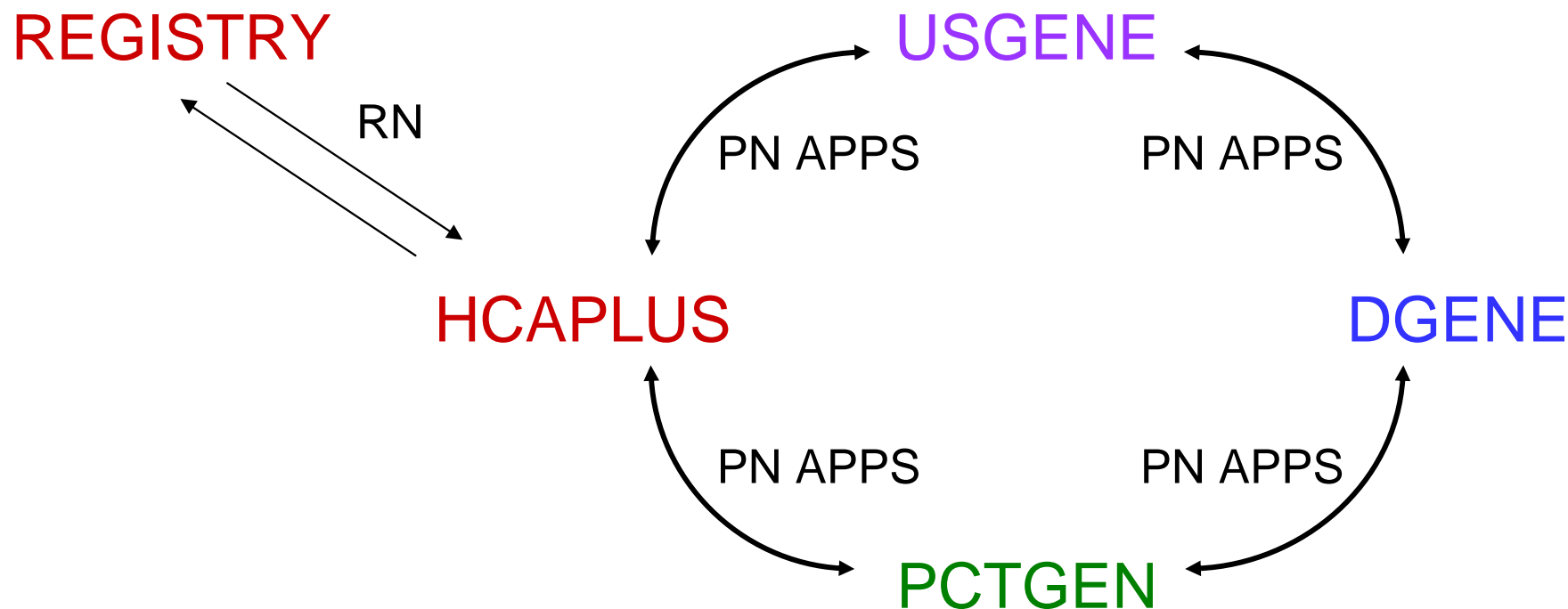
http://www.stn-international.com/training_center/bioseq/epss.pdf

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

<i>Functionality / Options</i>	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 Web</i>	SCM and CAS REGISTRY BLAST (slightly different implementation)	All 3 options (GETSEQ, GETSIM, BLAST) with RUN or with Sequence Search Assistant

Multifile sequence searching workflow uses PN, AP, PRN and RLN numbers



See *Effective patent sequence searching on STN (Part V)*:

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

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

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

Databases in Science and Technology - STN International - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.stn-international.com/

STN International | STN/CAS Home | FIZ Karlsruhe Inc | STN on the Web | CAS Summary Sheets | CAS Workshop Schedule | DWPI Reference | INPADOC Home | USPTO search | Esp@cenet

STN

FIZ Karlsruhe Home | Terms and Conditions | Contact
About us | Guided Tour | STN Easy | STN on the Web | Site Map | Search Site

STN Self Services

- Account Setup/Administration
- Free STNewline
- Downloads
- Site Administration

STN Interfaces

- STN AnaVist
- STN Express
- STN Easy
- STN on the Web
- STN Easy for Intranets
- Full-Text Solution

STN Databases

- List A-Z
- by Categories
- Summary Sheets
- Keep & Share

Service/Support

- Help Desk
- Representatives
- Prices/Order Forms
- Academic Page
- FAQ

USGENE Workshop - 10 May 2007

Your Connection to Science and Technology

Get Connected! [Info](#)

STN International connects scientists, engineers and anyone who needs technical information to the world's most complete and authoritative databases.

From Your Desktop [Info](#)

Select your preferred STN interface and:

- ask questions simply or by using sophisticated search commands
- identify published research and patents in all scientific fields
- retrieve original full-text articles and patents on the Web
- search chemical substance information by structure, name, or CAS Registry Numbers (CAS Number)

Be Confident [Info](#)

You can use STN with confidence because the system and the more than 220 databases it brings you are operated by some of the most respected scientific organizations in the world.

INPADOC reloaded - new INPADOCDB available April 29, 2007

STN Service Centers [Info](#)

- FIZ Karlsruhe in Europe
- CAS in North America
- Japan Association for International Chemical Information (JAICI)

What's new

- Database News
- INPADOC Reload**
- DWPI Reload
- IPC Reform
- STN Fixed Fee Plan
- Free STNewlines
- Interface News
- Meetings/Forums
- Exhibitions

Training Center

- Workshops
- e-Seminars
- Getting Started with STN
- Materials for Searching STN
- STN Express Interactive Training
- STN Free Search Preview
- STN Easy Demo

STN Archive

- STNews
- STN Brochures

© FIZ Karlsruhe 2007 Last Update: 04/26/2007 00:10:44 - Imprint

Visit www.sequencebase.com for the latest USGENE reference materials

SequenceBase USGENE USPTO Genetic Sequence Database - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.sequencebase.com/

STN International CRS STN/CAS Home FIZ Karlsruhe Inc STN on the Web CRS Summary Sheets CRS Workshop Schedule DWPI Reference INPADOC Home USPTO search Esp@cenet

SequenceBase

About the SequenceBase Corporation and the USPTO Genetic Sequence Database - USGENE®

The SequenceBase Corporation is a leader in providing biosequence information to the legal, biotech, pharmaceutical, scientific, technical, and academic community. USGENE® is our unparalleled resource for freedom-to-operate, prior-art, validity & infringement patent sequence searches, competitive analysis of organizations with sequence patent publications, and current awareness alerts (SDIs) from the very latest USPTO sequence data. In addition to being the producer of USGENE® on STN®, we provide customized subscription download services of sequence information, with a particular emphasis on patent sequences, to a wide variety of clients. SequenceBase Corporation pools sequence information from many world-wide resources, standardizes its collection and then supplies it in multiple standard formats to suit individual in-house patent bioinformatics information needs. To learn more about SequenceBase products and services, email us at: info@sequencebase.com.

USGENE® - coming soon to STN®!

"Our new USGENE® database provides a unique searchable combination of biological data - sequences, organism names, molecule types, sequence ID numbers and feature tables - with the original USPTO publication text, especially the full-patent claims," commented Martin Goffman, President and CEO, SequenceBase Corporation. "We are delighted to partner with FIZ Karlsruhe to offer STN® patent sequence searchers first access to this new database - an entirely new resource for U.S. freedom-to-operate, prior-art, validity and patent infringement searches."

"The unique combination of USGENE® with prestigious STN® patent sequence databases DGENE (GENESEQ™), PCTGEN and REGISTRY gives life-science intellectual property professionals the ability to make business critical decisions with greater certainty than was possible before," said Sabine Brünger-Weilandt, President and CEO of FIZ Karlsruhe. "FIZ Karlsruhe recognizes that the creation of the USGENE® database by the SequenceBase Corporation is a truly innovative concept, and we join with our customers to welcome its arrival on STN® with great enthusiasm."

[Download the complete press release! \(PDF\)](#)

For further information about USGENE®, please contact:

Martin Goffman President and CEO SequenceBase Corporation Tel: +1 732 549-5433 E-Mail: mgoffman@sequencebase.com www.sequencebase.com	Robert Austin Regional Sales Manager FIZ Karlsruhe Inc Tel: +1 609 333 1466 E-Mail: usgene@fiz-k.com www.fiz-k.com/usgene
---	--

USGENE®

USGENE® is The USPTO Genetic Sequence Database - a ground-breaking new resource for life-science intellectual property professionals.

Downloads

- [Presentation \(PDF\)](#)
- [Workshop Program \(PDF\)](#)
- [USGENE Flyer \(PDF\)](#)
- [Press Release \(PDF\)](#)

Sign-up to STN®

- [North America](#)
- [Europe/Rest-of-World](#)

STN INTERNATIONAL

Home | Disclaimer | [Presentation \(PDF\)](#) | [Workshop Program \(PDF\)](#) | [USGENE Flyer \(PDF\)](#) | [Press Release \(PDF\)](#)

Done

Questions/Comments...?

- Martin Goffman
 - Phone: +1 732 549-5433
 - Email: mgoffman@sequencebase.com
 - www.sequencebase.com
- Robert Austin
 - Phone: +1 609 333 1466
 - Email: usgene@fiz-k.com
 - www.fiz-k.com/usgene



The USPTO Genetic Sequence Database, USGENE[®], on STN[®]

www.fiz-k.com/usgene

www.sequencebase.com

