

STN[®]

To PSIPS and Beyond

Exploring the content and utility of
USGENESM

Robert Austin – FIZ Karlsruhe

Agenda

- Public sources of USPTO sequence data
- USGENE database content
- The 7 basic steps of USGENE BLAST®
- USGENE BLAST search example
- Comparisons and conclusions

See also *Effective patent sequence searching - Part III:*

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

Public sources of USPTO sequence data

3






- International Nucleotide Sequence Database Collaboration (INSDC)
 - www.insdc.org
- USPTO Protein database
 - NCBI: www.ncbi.nlm.nih.gov
 - EMBL-EBI: www.ebi.ac.uk
- Publication Site for Issued and Published Sequences (PSIPS)
 - <http://seqdata.uspto.gov/>

International Nucleotide Sequence Database Collaboration (INSDC)

4

- NCBI/EMBL/DDBJ collaboration (Genbank)
- Direct submissions (>77% with no references)
- Information as given by the submission author
- Patent division data from USPTO, EPO and JPO
- 66.8 million nucleotide sequence records
 - Including 3.66 million in the patent division, of which 1.07 million come from USPTO Issued Patents
- Updated daily
- 1982- present

NCBI and EMBL-EBI also provide a collection of USPTO peptide data

| | USPTO Nucleotide | USPTO Peptide | Home Page |
|------|---|---|--|
| EMBL |  |  | www.ebi.ac.uk |
| NCBI |  |  | www.ncbi.nlm.nih.gov |
| DDBJ |  | | www.ddbj.nig.ac.jp |

USPTO Nucleotide and Peptide sequences are included in the EMBL Nucleotide Database, and the Patent Proteins collection respectively.

Search Options

- Select the databanks you want to search
- Enter your search terms in the Quick Search box, or choose a query form from below

[Standard Query Form](#)
[Extended Query Form](#)

You can browse through all the entries in any databanks. First, select the databanks you want to browse, then click:

[Browse Entries](#)

Available Databanks

- Expand all Collapse all
- Literature, Bibliography and Reference
- Gene Dictionaries and Ontologies
- Nucleotide sequence databases
 - EMBL
 - IPD-KIR
 - EMBL (Coding Sequences)
 - EMBL ID/Accession Mapping
 - EMBL MGA
 - Nucleotide sequence databases - subsections**
 - EMBL (Updates)
 - EMBL (Whole Genome Shotgun release)
 - EMBL (Contig updates)
 - EMBL (Annotated Cons release)
 - RefSeq Genome (Updates)
- Nucleotide related databases
- UniProt Universal Protein Resource
- Other protein sequence databases
 - Active protein sequence databases**
 - Patent Proteins
 - IPI
 - Refseq Proteome (Release)
 - RefSeq Proteome (Updates)
 - Deprecated Protein Databases**
 - Swall(SPTR)
 - PIR
 - RemTrEMBL
- Protein function, structure and interaction databases
- Enzymes, reactions and metabolic pathway databases
- Mutation and SNP databases

- Patent DNA
- EMBL (Contig)
- Genome Reviews
- RefSeq Genome
- IMGT/LIGM-DB
- EMBL (Contigs expanded)
- RefSeq Genome
- IMGT/HLA
- EMBL (Annotated Cons)
- LiveLists

The EPO nucleotide sequence database, and the EPO, JPO and USPTO protein databases may also be searched separately.

- EPO Proteins
- JPO Proteins
- USPTO Proteins
- IPI History
- MHCIN
- RefSeq Proteome
- SWISSCHANGE
- BCIPEP

EMBL-EBI SRS: <http://srs.ebi.ac.uk/>

7

[Text Entry](#) | [Patent Protein Entry](#) | [Related Data](#)

[Reset](#) [Previous Entry](#) Entry 2 of 319 from [Query 1](#) [Next Entry](#)

Entry Information
Entry from: [USPTO Proteins](#)

Entry Options
Launch analysis tool:
 [Launch](#)
Link to related information: [Link](#)
Save entry: [Save](#)
View: [Printer Friendly](#)

Go to: [General](#) [Description](#) [References](#) [Sequence](#)

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](#)

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

Description

Description Sequence 1 from Patent US 4563352.
Organism Unknown

EMBL patent sequence records have minimal searchable patent bibliographic and text information.

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 `>uspo_prt|AAA00521|AAA00521 Sequence 1 from Patent US 4563352.
YADAIFTNSYRKVLGQLSARKLLQDIMSRQQGESNQERGA`

Go to: [General](#) [Description](#) [References](#) [Sequence](#)

Publication Site for Issued and Published Sequences (PSIPS)

- Since 2001 this service provides electronic access to sequence listings which would be over 300 pages in length if printed out in full
- There are 2,400+ USPTO listings available, 200 from Issued Patents and 2,200 from published applications, representing 25+ million sequences
- The PSIPS sequences cannot be searched at the site – they can only be downloaded
- Most listings are in ST.25 format – a minority are available in a previous USPTO standard format

PSIPS sequence listing download:
<http://seqdata.uspto.gov/>

Publication Site for Issued and Published Sequences (PSIPS)

PSIPS Home Page

Welcome to the PSIPS 2.6 Home Page! At present, this system stores and retrieves large Sequence Listings and tables that have been included in a granted and/or published US patent application. Shorter Sequence Listings (i.e., less than 300 pages) and smaller table sections (i.e., less than 200 contiguous pages) are accessible via the Patents- and Applications-On-The-Web home pages. You can view individual sequences or tables, or download Sequence Listings, tables, or other mega items for any PSIPS document. [More Information](#)

Two Easy Ways To Search For A PSIPS Document!

By Document/Patent ID

This method is the one to choose if you know the document or patent ID in advance. It is the most direct method of finding the patent data you are looking for.

Enter Document ID/Patent ID:

By Publication Date Range

You may not know the document or patent ID in advance, in which case this method will allow you to search for a document or patent ID based upon a date range.

Enter Date Range:

Document ID Examples

Sample document number:
US06979557B2

Sample :
6183957

Sample Re-issue:
RE000123

Sample Design:
D0000126

All documents by Doc ID:
Click on Submit Button without entering Doc ID

Date Range Examples

Published on or after March 15, 2001:
03/15/2001- or May 15, 2001- or 20010515-

Published
-05/12/2001

Published
11/06/2000

Published
05/01/2000

Published
May 1, 2000

All documents
Click on Submit Button without entering Date

PSIPS provides sequence listings for Published Applications and Issued Patents. Most listings are in WIPO ST.25 format.

[Data Entry Help](#) | [PSIPS Help](#) | [PSIPS FAQ](#)

Most PSIPS listings are provided in the USPTO variant of WIPO ST.25 format

10

```
<210> SEQ ID NO 1
<211> LENGTH: 314
<212> TYPE: DNA
<213> ORGANISM: Corynebacterium glutamicum
<220> FEATURE:
<221> NAME/KEY: CDS
<222> LOCATION: (1)..(291)
<223> OTHER INFORMATION: RXA02548
<400> SEQUENCE: 1
cca ccg atc tac ttc tcc cac gac cgc gaa gtt ttc
Pro Pro Ile Tyr Phe Ser His Asp Arg Glu Val Phe
      1              5              10
atg tgg ctg acc gca ggc gag tgg ggt gga cca aag aag ggc gag gag   96
Met Trp Leu Thr Ala Gly Glu Trp Gly Gly Pro Lys Lys Gly Glu Glu
      20              25              30
atc gtc acc aag act gtc cgc tac cgc acc gtc ggc gat atg tcc tgc   144
Ile Val Thr Lys Thr Val Arg Tyr Arg Thr Val Gly Asp Met Ser Cys
      35              40              45
```

| <u>Information</u> | <u>ST.25</u> |
|--------------------|--------------|
| SEQ ID NO | <210> |
| Length | <211> |
| Type | <212> |
| Organism | <213> |
| Feature | <220> |
| Name/key | <221> |
| Location | <222> |
| Sequence | <400> |

For further details about the USPTO and WIPO ST.25 format:

http://www.uspto.gov/web/offices/pac/mpep/documents/appxr_1_823.htm

PSIPS sequence listing download:
<http://seqdata.uspto.gov/>

Publication Site for Issued and Published Sequences (PSIPS)

PSIPS View Sequence(s): 2 for 6825322

Here is the list of the requested sequences.

Sequence ID No:

- [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](#)

```
(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          45
Ala Asn Lys Arg His Gly Ser Trp Lys Ile Gln Leu Asn Ala Thr Ser
 50          55          60
Val Thr His Lys Pro Asn Ala Ile Gln Met Ala Leu Ser Val Cys Glu
 65          70          75          80
Asp Leu Ile Ser Ser Gln Val Tyr Ala Ile Leu Val Ser His Pro Pro
 85          90          95
Thr Pro Asn Asp His Phe Thr Pro Thr Pro Val Ser Tyr Thr Ala Gly
100          105          110
Phe Tyr Arg Ile Pro Val Leu Gly Leu Thr Thr Arg Met Ser Ile Tyr
115          120          125
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
165          170          175
Gln Lys Arg Leu Glu Thr Leu Leu Glu Glu Arg Glu Ser Lys Ala Glu
180          185          190
Lys Val Leu Gln Phe Asp Pro Gly Thr Lys Asn Val Thr Ala Leu Leu
195          200          205
Met Glu Ala Lys Glu Leu Glu Ala Arg Val Ile Ile Leu Ser Ala Ser
210          215          220
```

A minority of listings are presented in the pre-ST.25 USPTO standard format.

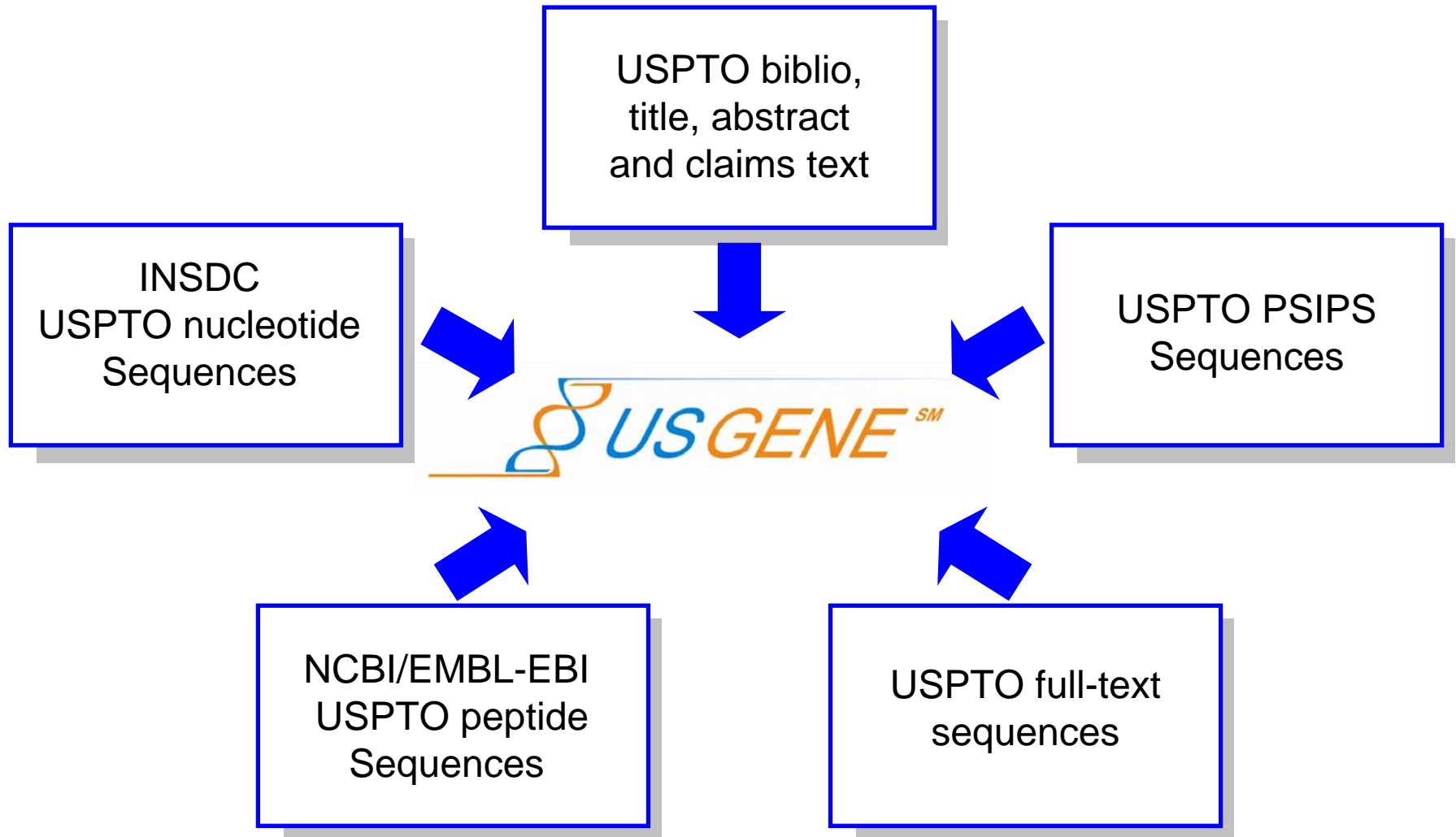
The USPTO Genetic Sequence Database - USGENE

12

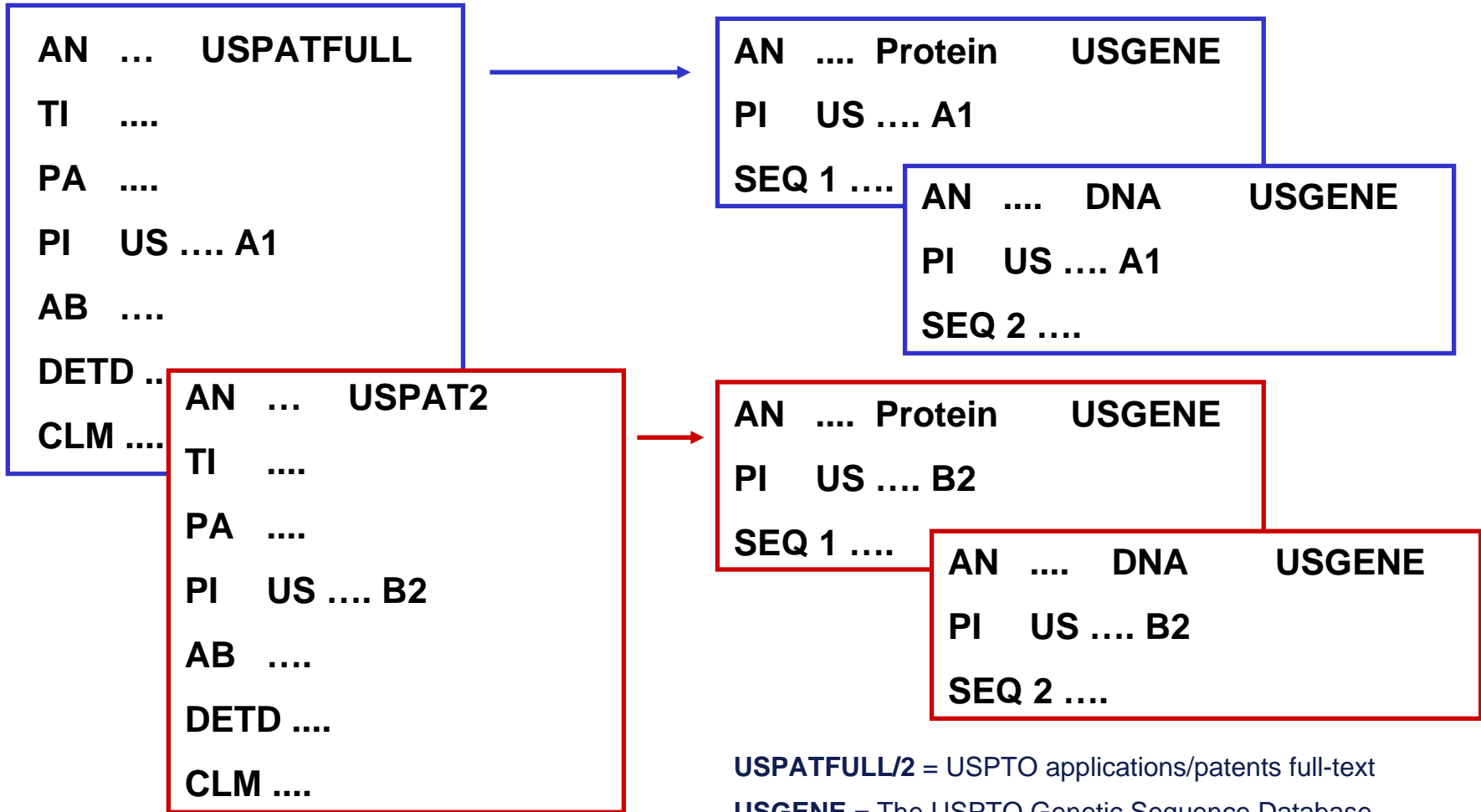
- Produced by the SequenceBase Corporation
- Sequences from all relevant USPTO published patent applications and issued (granted) patents
- Assignee and full inventor names; publication, application and parent case PCT numbers and dates; original publication **title**, **abstract** and **claims**
- Organism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

USGENE is compiled technologically from several disparate sources

13



Relationship between USPATFULL/2 and USGENE databases



USPATFULL/2 = USPTO applications/patents full-text

USGENE = The USPTO Genetic Sequence Database

A typical USGENE record

L1 ANSWER 1 OF 1 USGENE COPYRIGHT 2007 SEQUENCEBASE CORP on STN
AN 6639063.3883 (1) DNA (2) USGENE
TI EST's and encoded human proteins (Patent) (3)
IN Edwards Jean-Baptiste Dumas Milne (Paris, FR)
Jobert Severin (Paris, FR)
Giordano Jean-Yves (Paris, FR)
PA Genset S A (FR)
PI US 6639063 B1 20031028 (4)
AI US 2000-621976 20000721
ORGN Homo Sapiens (5)
AB The sequences of 5' ESTs and consensus contigated 5' ESTs derived from
mRNAs encoding secreted proteins are disclosed. The 5' ESTs and consensus
contigated 5' ESTs may be used to obtain cDNAs and genomic DNAs
(6) corresponding to the 5' ESTs and consensus contigated 5' ESTs. The 5'
ESTs and consensus contigated 5' ESTs may also be used in diagnostic,
forensic, gene therapy, and chromosome mapping procedures. Upstream
regulatory sequences may also be obtained using the 5' ESTs and consensus
contigated ESTs. The 5' ESTs and consensus contigated 5' ESTs may also be
used to design expression vectors and secretion vectors.

USGENE records include
patent bibliography, author
title and abstract.

A typical USGENE record (cont.)

CLM US6639063 B1: What is claimed is:

USGENE records include searchable patent claims text.

- (7) 1. An isolated, purified, or recombinant polynucleotide encoding a signal peptide, wherein:a) the polynucleotide encodes a signal peptide consisting of residues -102 to -1 of SEQ ID NO: 3986; and b) wherein said signal peptide directs the secretion of a polypeptide when located at the amino terminus of a polypeptide.
2. The isolated, purified, or recombinant polynucleotide of claim 1, wherein: said polynucleotide consists of nucleotides 144 to 449 of SEQ ID NO: 126.
3. An isolated, purified, or recombinant polynucleotide encoding a signal peptide, wherein:a) said polynucleotide, consists of a nucleic acid sequence having at least 90% homology to the polynucleotide of claim 2; andb) said polynucleotide, encodes a signal peptide that

SSO NUCLEIC; PSIPS; GRANTED (8)

USGENE records are compiled from several disparate sources.

A typical USGENE record (cont.)

```
SQL      231  (9)
SEQ
      1 tactactagt ctccttgaag tatatgttgt cgccacatt ttgctgcagt
     51 tcacttttaa ttcctaagaa ggttgttttc acttggtggt tttttaatct
    101 cttaagaatg aatagtagga atattagtag caacacctta aactcatgtc
    151 acattttaat attcacagaa catctacaca cacattatgt tattaggtaa
    201 acaggtggtg acagcctgca ttagttttaa g
```

(10)

FEATURE TABLE:

```
Key      |Location|
=====+=====+=====
CDS      |24..197 |
```

(11)

A typical USGENE record (cont.)

18

- 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) Bibliographic information – publication, application, assignee and inventor data
- 5) Organism (where given) – providing the name of the organism from which the sequence is derived
- 6) Original patent abstract.

A typical USGENE record (cont.)

19

- 7) Full published application or patent claims text
- 8) The Sequence Source (SSO) – nucleic or protein; PSIPS/USPTO, NCBI, etc; granted or application
- 9) Sequence Length (SQL)
- 10) Patent sequence - each USGENE record is based upon a sequence
- 11) Feature table including sequence modifications and other features, as provided by the patent applicant

USGENE represents a new tool for tackling business critical searches

- DGENE and REGISTRY patent sequences are indexed by Thomson from the DWPI basic and by CAS from the CAplus 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 CAPLUS COPYRIGHT 2006 ACS on STN
AN 1999:549367 CAPLUS
DN 131:166204
TI A system for screening for
complex between transcripti
control of immune function
IN Mach, Bernard
PA Novimmune S.A., Switz.
FAN.CNT 1











In this example the patent family has:

- 3 sequences from [WO 9942571](#) in REGISTRY
- 4 sequences from [FR 2775003](#) in DGENE
- 6 sequences from [US 6370894](#) in USGENE

| | PATENT NO. | KIND | DATE | APPLICATION NO. | DATE |
|------|--|------|----------|-----------------|----------|
| PI | WO 9942571 | A1 | 19990826 | WO 1999-FR376 | 19990219 |
| | W: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, , . . . | | | | |
| | FR 2775003 | A1 | 19990820 | FR 1998-2025 | 19980219 |
| | AU 9924312 | A1 | 19990906 | AU 1999-24312 | 19990219 |
| | EP 1068310 | A1 | 20010117 | EP 1 | |
| | R: DE, FR, GB | | | | |
| | JP 2002504325 | T2 | 20020212 | JP 2 | |
| | US 6379894 | B1 | 20020430 | US 2 | |
| PRAI | FR 1998-2025 | A | 19980219 | | |
| | WO 1999-FR376 | W | 19990219 | | |

It is critical to search all available resources to ensure a thorough patent literature sequence search

USGENE provides more USPTO sequence data than EMBL-EBI

| | USPTO PGP's | USPTO Patents | USPTO claims text | Value added |
|-----------|---|--|---|---|
| EMBL-EBI | |  | | |
| USGENE |  |  |  | |
| DGENE* |  |  | |  |
| REGISTRY* |  |  | |  |

USGENE also provides the most timely source of USPTO sequence data

| | Update Frequency | Typical Timeliness | Value added |
|-----------|------------------|--------------------|--|
| USGENE | Weekly | 7 days | |
| FASTAlert | Biweekly | 14 days | |
| REGISTRY | Daily | 27 days |  |
| DGENE | Biweekly | 65 days |  |
| EMBL-EBI | Daily | 1-3 months | |

USGENE offers the same sequence searching methods 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 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 ALIGN 1,3,10
- 7) Ensure session transcript was captured and Logoff

1) SAVE, UPLOAD and VERIFY

26

The screenshot displays three overlapping windows from the STN software interface, illustrating the steps to upload a query:

- Window (1): Select Discover! Wizard**
 - Section: Choose a search wizard:
 - Buttons: Select Database, Author, CAS Registry Number, Chemical Name, Corporate Source, Subject, Edit alert, **Upload Query** (highlighted with a red box).
- Window (2): STN Upload Query Wizard**
 - Text: Select a structure or sequence query to upload to STN, and click Next. Click Cancel to return to STN.
 - File Name: C:\CASNC\STN Express\Queries\PSIPS 3
 - Text: SEQ ID NO 136, LENGTH: 247
 - Text: Ala Leu Leu Leu Val, Gly Asn Gly Asp Asn
 - Buttons: Next >, Cancel
- Window (3): STN Upload Query Wizard**
 - Text: 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.
 - Section: Databases:

| | |
|--------|---|
| DGENE | Derwent Geneseq Database 1981 - present |
| PCTGEN | World Patent Application Biosequences |
 - Buttons: Next >, Cancel

At the bottom, a window titled "View this Database Summary" shows buttons for < Back, Finish, and Cancel.

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

The sequence becomes a **Query L-number** in the file of choice for use with RUN BLAST

1) SAVE, UPLOAD and VERIFY (cont.)

27

=> FILE PCTGEN

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

=> UPL R BLAST

UPLOAD SUCCESSFULLY COMPLETED

L1 GENERATED

=> D LQUE L1

L1 ANSWER 1 PCTGEN COPYRIGHT 2006 WIPO on STN
VQTVPLSRLFDHAMLEAHRAHEL AIDTYQEF EETYIPKDQKYSFLHDSQT
SFCFSDSIPTPSNMEETQQKSNLELLRISLLLI ESWLEPVRFLRSMFANN
LVYDTSDDYHLLKDL EEGIQTLMGRLEDGSRRTGQILKQTYSKFDTNS
HNHDALLKNYGLLYCFRKDMDKVETFLRMVQCRSVEGSCGF

=>

The sequence is now ready for searching directly in USGENE, DGENE or PCTGEN using the L-number.

2) RUN the USGENE BLAST search

```
=> FILE USGENE
```

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

Turn the Low Complexity Filter off with the syntax... /SQP -F F (next slide)

```
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.

3) Decide how many answers to keep

1297 ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0

Similarity
Score

387

194

Answer Count

260

520

780

1040

1300

The graphic representation gives a count of hit sequences (x-axis) and similarity score (y-axis). The graph gives a visual clue about the proportion of similar and not so similar sequences in the answer set.

Recommendation: keep **ALL** answers

HOW MANY ANSWERS WOULD YOU LIKE TO KEEP ? (ALL) OR ? : [ALL](#)

4) SORT by SCORE descending

```

L2      RUN STATEMENT CREATED
L2      1297 VQTVPLSRLFDHAMLEAHRAHELAIPTYQEFETYIPKDQKYSFLHDSQT
          SFCFSDSIPTPSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANN
          LVYDTSDDSDDYHLLKDLLEGIQTLMGRLEDGSRRTGQILKQTYSKFDTNS
          HNHDALLKNYGLLYCFRKDMDKVETFLRMVQCRSVEGSCGF/SQP.-F F

```

Answer set arranged by ascending similarity score, enter a

Here the search is further limited to sequences from issued (granted) patents, with application year prior to 1996, and to a specific term within the claims.

```

=> S L2 AND SOMATOMAMMOTROPIN/CLM AND AY<1996 AND GRANTED/SSO
L3      39 L2 AND SOMATOMAMMOTROPIN/CLM AND AY<1996 AND GRANTED/SSO

```

```

=> SOR L3 SCORE D
PROCESSING COMPLETED FOR L3
L4      39 SOR L3 SCORE D

```

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

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

```
=> D L4 TRI ALIGN 1-39; FILE STNGUIDE
```

```
L4      ANSWER 1 OF 39  USGENE COPYRIGHT 2006 SEQUENCEBASE CORP on STN
TI      Recombinant DNA transfer vectors (Patent)
MTY     Protein
SQL     191
ORGN    Unknown
BLASTALIGN
```

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

```
Query  = 191 letters
```

```
Length = 191
```

```
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     PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDDSDDYHLLKDLEEG
              PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDDSDDYHLLKDLEEG
Sbjct: 61     PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDDSDDYHLLKDLEEG
              . . . .
```

6) Display selected answers in a preferred bibliographic format

=> D BIB SSO AB CLM ALIGN 1 3 10

L4 ANSWER 1 OF 39 USGENE COPYRIGHT 2007 SEQU
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
ORGN Unknown
SSO PROTEIN; EMBL; GRANTED
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 c
chorionic somatomammotropin comprising

BLASTALIGN

USGENE has a great deal of utility as a *freedom-to-operate* patent searching tool.

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

33

- 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 ALIGN 1,3,10
- 7) Ensure session transcript was captured and Logoff

See also *Effective patent sequence searching on STN*:

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 the “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[®]

To PSIPS and Beyond

**Exploring the content and utility of
USGENESM**

Robert Austin – FIZ Karlsruhe