

# STN<sup>®</sup>

Synergies and Surprises –  
USGENE<sup>®</sup> and DGENE Multifile Patent  
Sequence Searching on STN<sup>®</sup>

Robert Austin – FIZ Karlsruhe

# Agenda

2

- DGENE and USGENE database content
- The importance of DWPI patent families
- Multifile “best-practice” technique using BLAST
- Step-by-step walk through a multifile search
- Overview of case-study search results
- Examples of unique USGENE retrieval
- Conclusions

# Thomson Scientific GENESEQ (DGENE)

3

- Largest value-added patent sequence database
- Used routinely by all major patent offices\*
- Sequences from the basic patents of the 40 authorities of the *Derwent World Patents Index*®
- Bibliography, enhanced title, abstract, indexing and patent location provided for each sequence
- Patent Family and Legal Status display
- Updated every two weeks
- 1981 - present

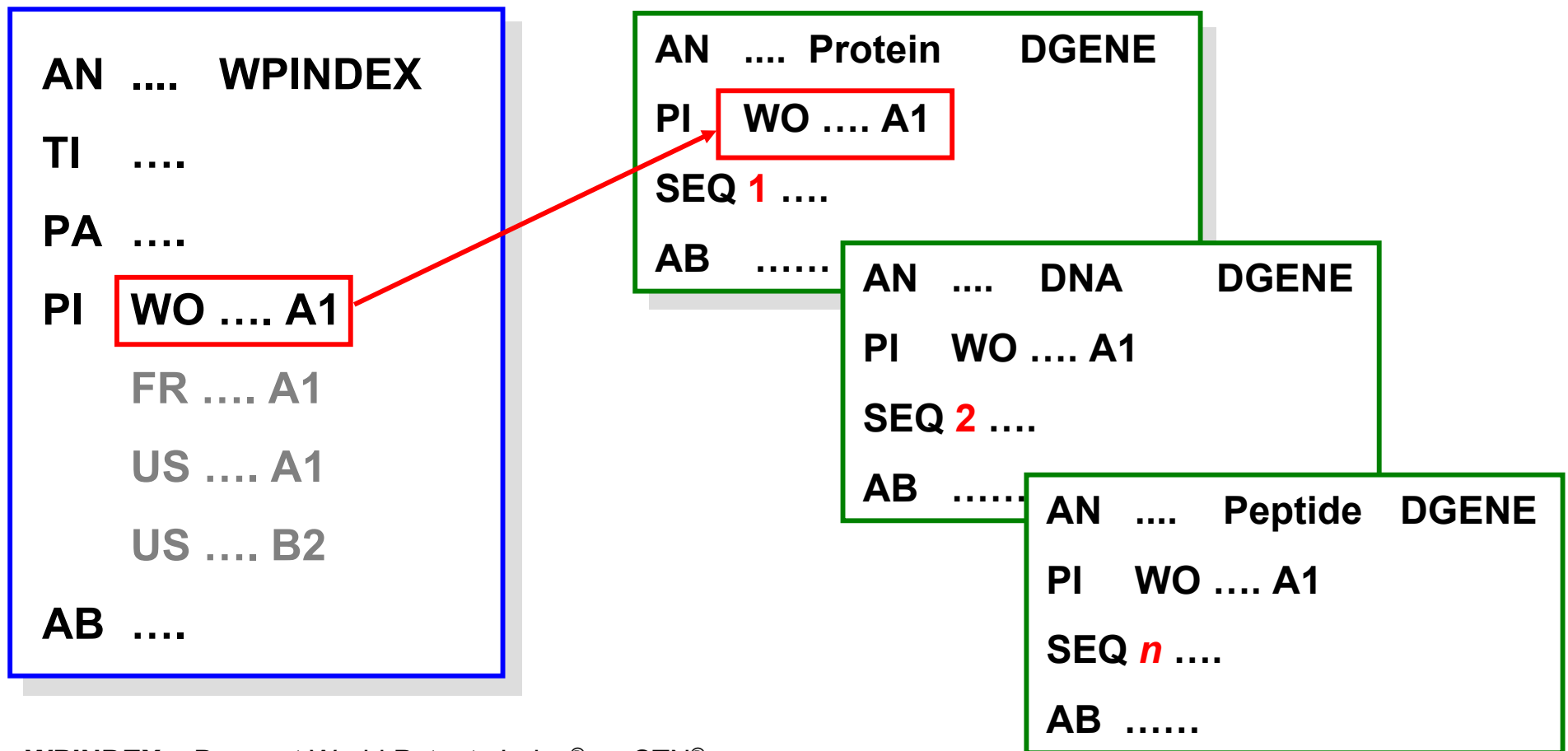
\* See page 11: [http://www.trilateral.net/projects/biotechnology/search\\_guidebook\\_vers\\_1.pdf](http://www.trilateral.net/projects/biotechnology/search_guidebook_vers_1.pdf)

# The Derwent World Patents Index (DWPI<sup>SM</sup>)

4

- Largest value-added global patent database
- Enhanced patent titles and abstracts
  - Improve search recall and relevance
  - Reduce time required to review results
- Intellectually compiled patent families
  - Precise access to equivalent documents
- Comprehensive classification and indexing
  - Improves search recall and relevance

# Relationship between DWPI patent family & DGENE sequence database



WPINDEX = Derwent World Patents Index® on STN®

DGENE = GENESEQ™ on STN

# What exactly is the “value-add” in DGENE?

6

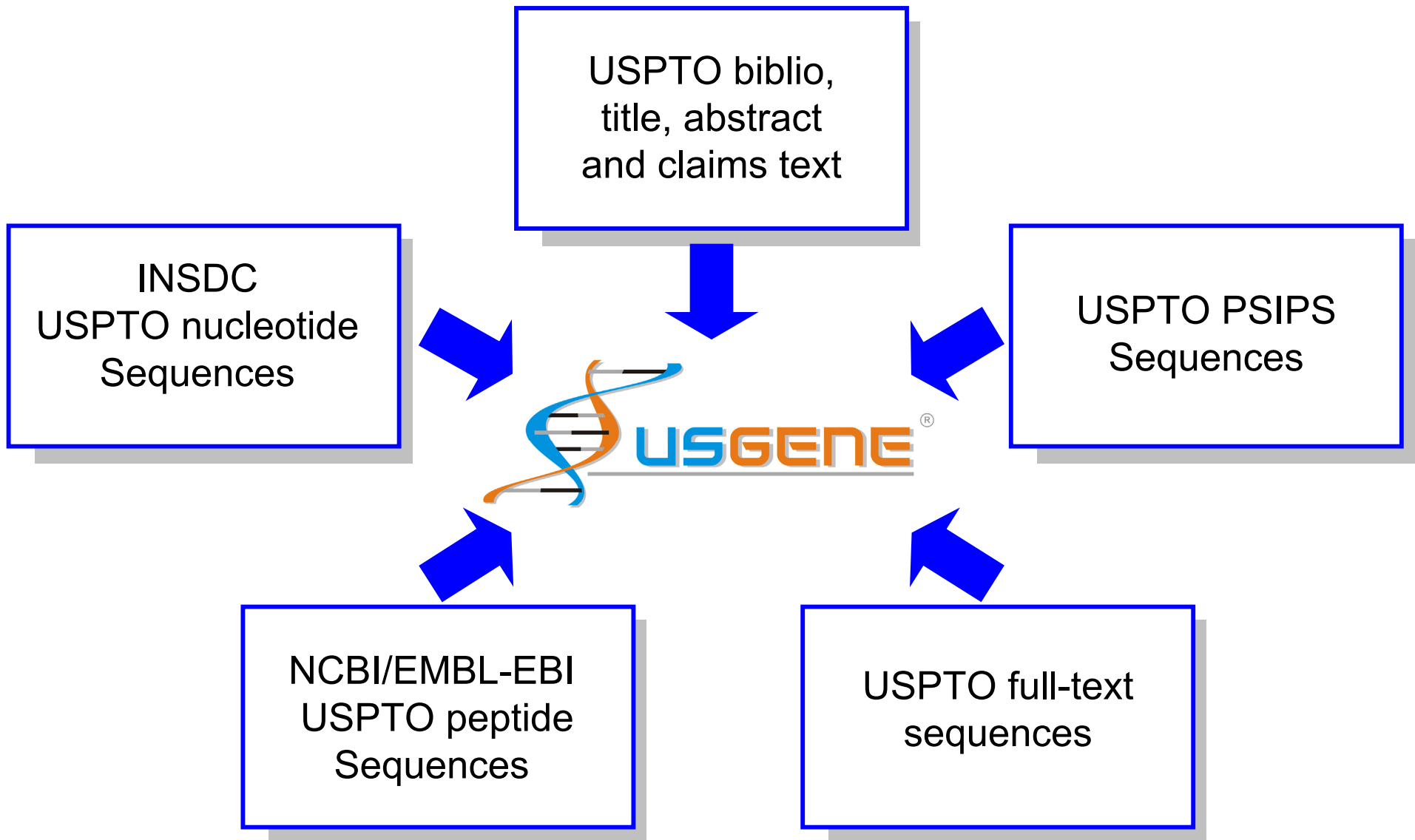
- DWPI patent title, concise sequence description, abstract and keyword indexing *per sequence*
  - Illuminate the context of *each sequence* within the invention
  - Superior text based refinement of sequence searches
  - Efficient scanning and review of search results for relevance
- Feature tables for sequence modifications/annotations
  - Extensive detailed annotations are provided by Indexers
- Patent sequence location (claim, example, etc)
  - Assigned manually by Thomson Indexers
  - Ability to filter searches to those described in the claims
- Sequences intellectually derived by Indexers
  - Provides unique sequence hits not disclosed in formal listings

# USGENE is the USPTO Genetic Sequence Database

7

- Sequences captured from all relevant USPTO published patent applications and granted (issued) 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 **3 days** of publication
- 1982 – present

# USGENE combines sequences with bibliographic data and claims text



# An individual publication is represented by one or more USGENE sequence records

(12) <b>United States Patent</b> <b>Higo et al.</b>	(10) <b>Patent No.:</b> <b>US 7,255,990 B2</b>
	(45) <b>Date of Patent:</b> <b>Aug. 14, 2007</b>
(54) <b>METHOD FOR SCREENING GENES EXPRESSING AT DESIRED SITES</b>	(56) <b>References Cited</b>
(75) Inventors: <b>Kenichi Higo</b> , Tsukuba (JP); <b>Masao Iwamoto</b> , Tsukuba (JP)	U.S. PATENT DOCUMENTS
(73) Assignee: <b>National Institute of Agrobiological Sciences</b> , Ibaraki (JP)	5,569,831 A * 10/1996 DellaPenna ..... 800/286
(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 399 days.	6,576,815 B1* 6/2003 Higo et al. .... 800/287
(21) Appl. No.: <b>10/221,596</b>	OTHER PUBLICATIONS
(22) PCT Filed: <b>Nov. 21, 2001</b>	Iwamoto et al. Atourist element in the 5'-flanking region of the catalase gene CatA reveals evolutionary relationships among Oryza species with various genome types. 1999. Mol. Gen. Genetics 262:493-500.*
(86) PCT No.: <b>PCT/JP01/10195</b>	Iwamoto et al. Mol. Gen. Genet. vol. 262:493-500. 1999.*
§ 371 (c)(1), (2), (4) Date: <b>Sep. 11, 2002</b>	McSteen et al. Development. vol. 125:2359-2369. Sep. 1998.*
(87) PCT Pub. No.: <b>WO03/044227</b>	Higo et al. Plant Molecular Biology vol. 30:505-521. 1996.*
PCT Pub. Date: <b>May 30, 2003</b>	* cited by examiner
(65) <b>Prior Publication Data</b>	<i>Primary Examiner</i> —Gary Benzion
US 2004/0086855 A1 May 6, 2004	<i>Assistant Examiner</i> —Heather Calamita
(51) <b>Int. Cl.</b>	(74) <i>Attorney, Agent, or Firm</i> —Perkins Coie LLP; Jacqueline F. Mahoney
<b>C12Q 1/68</b> (2006.01)	
(52) <b>U.S. Cl.</b> ..... <b>435/6</b>	(57) <b>ABSTRACT</b>
(58) <b>Field of Classification Search</b> ..... <b>435/6,</b>	The present invention relates to a method for inferring a plant organ, in which a certain gene is to be expressed, using a part of a base sequence, a method for searching for a gene which is to be expressed at a desired site, and a composition, kit, system and program for carrying out these methods. The present invention also relates to a method for inferring a plant organ, in which a plant gene is to be expressed, based on information about the presence or absence of a base sequence which is highly similar to a transposable element in the vicinity of a protein coding region of a plant gene.
435/91.1, 91.2; 536/23.1, 23.5	
See application file for complete search history.	<b>9 Claims, 10 Drawing Sheets</b>

**AN .... Protein USGENE**  
**PI US .... B2**  
**SEQ 1 ....**

**AN .... DNA USGENE**  
**PI US .... B2**  
**SEQ 2 ....**

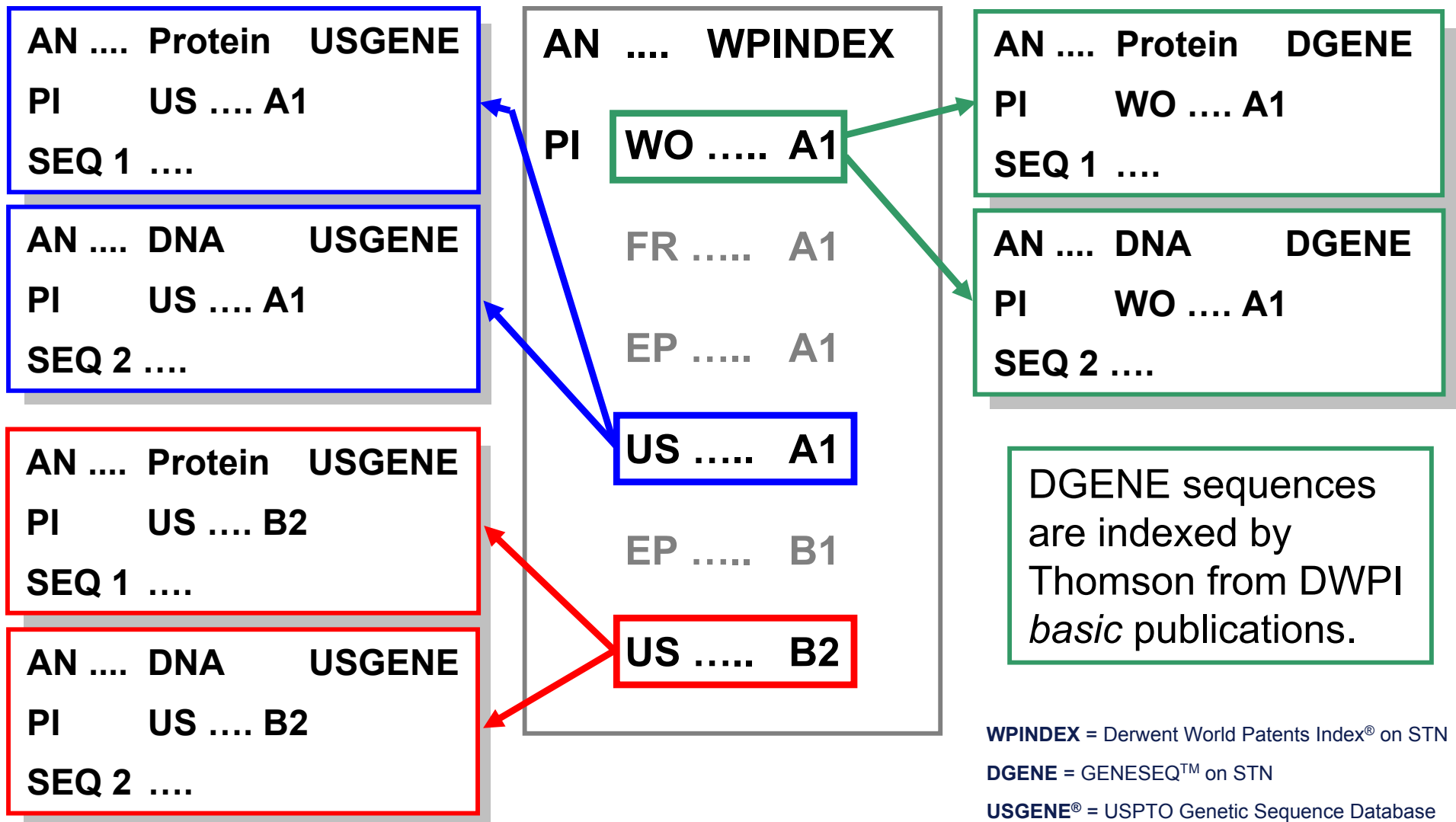


**AN .... cDNA USGENE**  
**PI US .... B2**  
**SEQ n ....**

# USGENE is an essential additional tool for tackling business critical searches

- DGENE provides curated and indexed patent sequence data from the DWPI *basic* publication
  - 61% of *basics* are WIPO/PCT published applications
  - Updated biweekly, typically 65 days from publication
- USGENE provides all available sequence data from the USPTO as a single merged resource
  - Both **U.S. patents** and **U.S. published applications**
  - Updated weekly, within **3 days** of USPTO publication
- Sequence listing variation often occurs between PCT and U.S. granted patent publication stages
  - Especially important, e.g. for freedom-to-operate

# USGENE and DGENE capture sequence data from different patent family members



# Agenda

12

- DGENE and USGENE database content
- The importance of DWPI patent families
- **Multifile “best-practice” technique using BLAST**
- Step-by-step walk through a multifile search
- Overview of case-study search results
- Examples of unique USGENE retrieval
- Conclusions

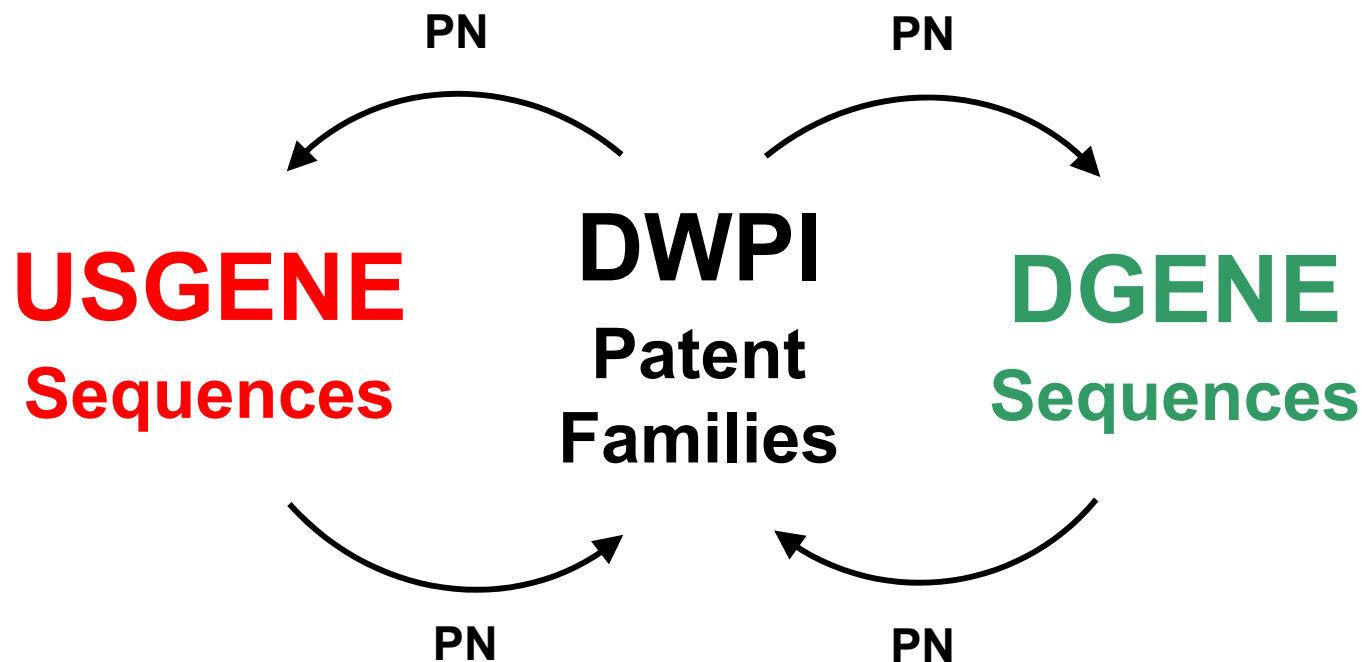
# The development of a “best-practice” recipe for multfile patent sequence searching

13

- A commonly used multfile technique is this
  - RUN BLAST searches in DGENE and USGENE, merge (DUP IDE), FSORT, display the best scoring sequence answer per patent family in full, and then optionally display the rest in a free-of-charge format
- The problem is that two sets of displays have to be recombined, and that family members are often not correctly grouped into patent families
- A best-practice methodology was developed which incorporates DWPI to solve these issues
- An STN script was written which automates this “best-practice” methodology for users

# The “best-practice” recipe for multfile searching incorporates DWPI patent families

14



The connection between DWPI and patent sequence databases DGENE and USGENE is via publication numbers (PN).

# The basic mechanics of the “best-practice” multifile patent sequence search

15

- 1) Ensure preferred file default display formats are set
- 2) UPLOAD the sequence query via *STN Express* (**L1**)
- 3) *USGENE*: BLAST (**L2**); SORT SCORE D (**L3**); review and isolate chosen hits with SORT AN 1-x (**L4**)
- 4) *DGENE*: BLAST (**L5**); SORT SCORE D (**L6**); review and isolate selected hits with SORT AN 1-x (**L7**)
- 5) *WPINDEX*: TRA PN L4 (**L9**); TRA PN L7 (**L11**); combine answer sets L9 OR L11 (**L12**)
- 6) Merge: DUP IDE L4 L7 L12 (**L13**); FSORT (**L14**)
- 7) Display results: D PFAM=1- TOTAL
- 8) Remember to capture your transcript and logoff

# The basic mechanics of a the “best-practice” multifile patent sequence search

16

## Search Question:

Find relevant patent references for *Eukaryotic translation elongation factor 1 gamma* (NP\_001395)

```
MAAGTLYTYPENWRAFKALIAAQYSGAQVRVLSAPPHFHFHGQTNRTPEFLRKFPAGKVPAFEG  
DDGFCVFESENAIAYYVSNEELRGSTPEAAAQVVQWVSFADSDIVPPASTWVFPTLGIMHHNKQ  
ATENAKEEVRRILGLLDAYLKTRTFLVGERVTLADITVVCTLLWLYKQVLEPSFRQAFPNTNR  
WFLTCINQPQFRAVLGEVKLCEKMAQFDAKKFAETQPKKDTPRKEKGSREEKQKPQAERKEEK  
KAAAPAPEEEMDECEQALAAEPKAKDPFAHLPKSTFVLDEFKRKYSNEDTLSVALPYFWEHFD  
KDGWSLWYSEYRFPEELTQTFMSCNLITGMFQRLDKLRKNAFASVILFGTNNSSSISGVWVFR  
GQELAFPLSPDWQVDYESYTWKLDPGSEETQTLVREYFSWEGAFQHVGKAFNQGKIFK
```

( Search conducted on February 6<sup>th</sup>, 2008.)

# 1) Ensure preferred file default display formats are set

```
=> FILE STNGUIDE
=> SET FORMAT .MYUSGENEALIGN TRI ORGN SEQN SEQC ALIGN
=> SET FORMAT .MYDGENEALIGN TRI OS ALIGN
=> SET FORMAT .MYWPINDEX BIB
=> FILE USGENE; SET DFORMAT .MYUSGENEALIGN
=> FILE DGENE; SET DFORMAT .MYDGENEALIGN
=> FILE WPINDEX; SET DFORMAT .MYWPINDEX

=> D FORMAT
```

A simple STN script can be used to issue these commands automatically.

USER-DEFINED FORMAT DEFINITION

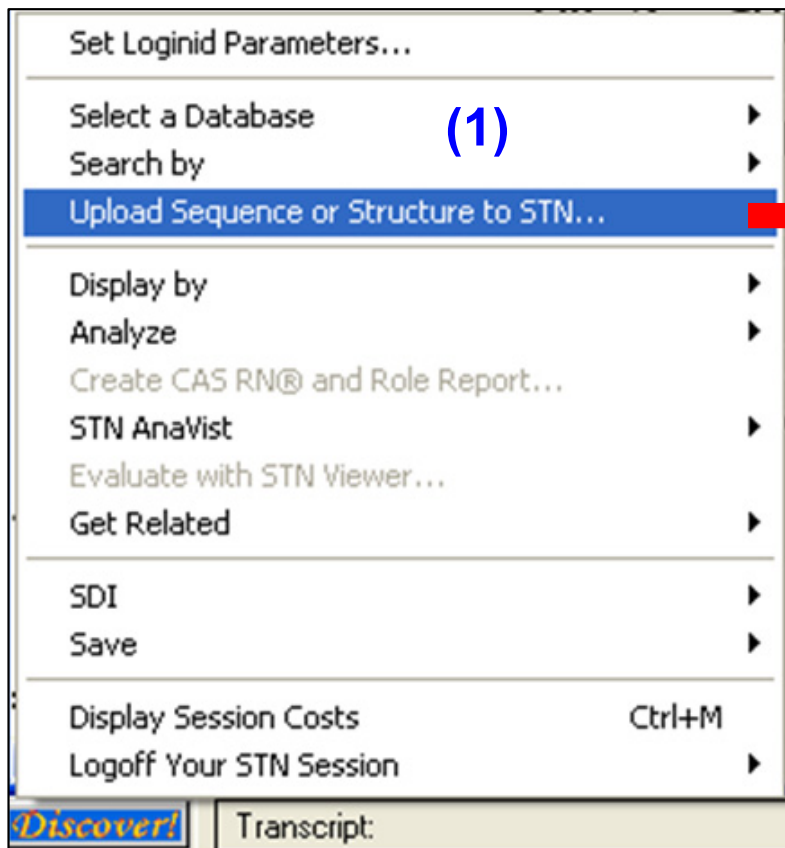
DEFAULT FORMAT  
FOR FILE

-----	-----	-----
.MYDGENEALIGN	TRI OS ALIGN	DGENE
.MYUSGENEALIGN	TRI ORGN SEQN SEQC ALIGN	USGENE
.MYWPINDEX	BIB	WPINDEX

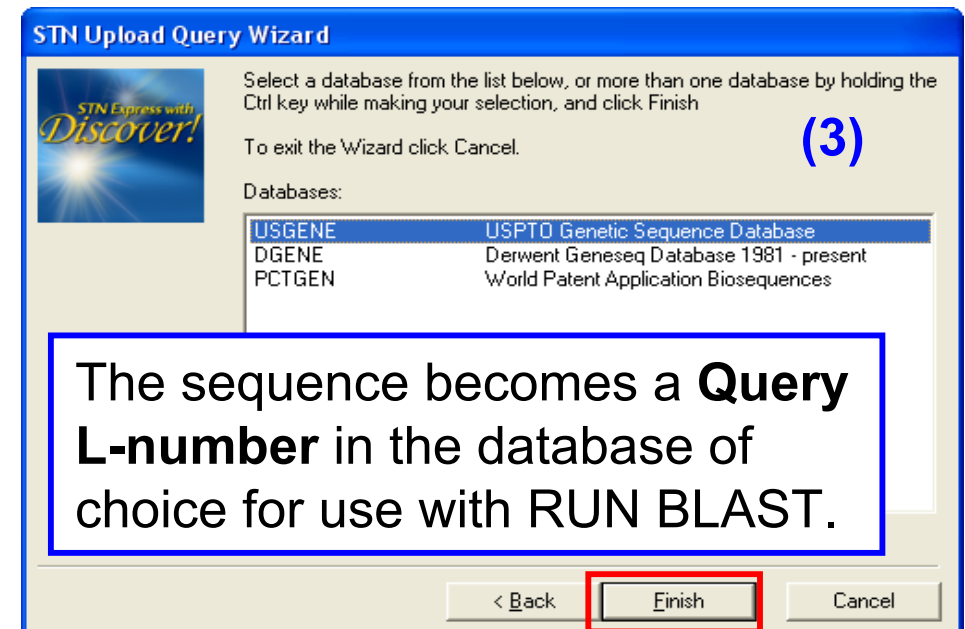
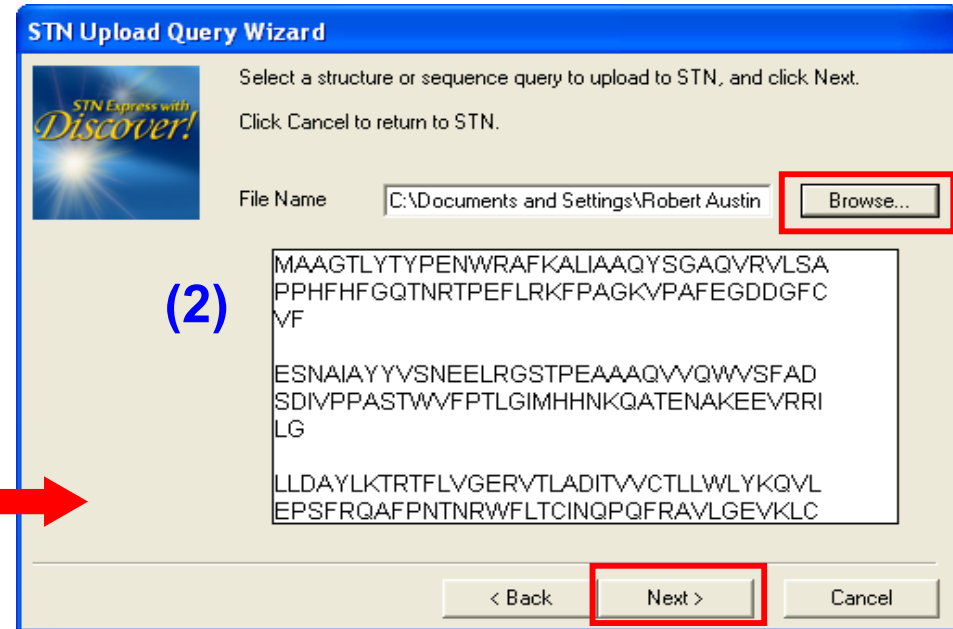
## 2) UPLOAD the sequence query

18

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



From the *Discover!* button menu.



## 2) UPLOAD the sequence query (cont.)

=> FILE USGENE

=> UPL R BLAST

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

UPLOAD SUCCESSFULLY COMPLETED

L1 GENERATED

=> D L1 LQUE

Verify that the UPLOAD was successful with D LQUE.

L1 ANSWER 1 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN  
LQUE MAAGTLYTYPENWRAFKALIAAQYSGAQVRVLSAPPHFHFGQTNRTPEFLRKFPAGKVP  
AFEGDDGFCVFESNAIAYYVSNEELRGSTPEAAAQVVQWVSFADSDIVPPASTWVFPTL  
GIMHHNKQATENAKEEVRRILGLLDAYLKTRTFLVGERVTLADITVVCTLLWLYKQVLE  
PSFRQAFPNTNRWFLTCINQPQFRAVLGEVKLCEKMAQFDAKKFAETQPKKDTPRKEKG  
SREEKQKPQAERKEEKKAAPAPEEEMDECEQALAAEPKAKDPFAHLPKSTFVLDEFKR  
KYSNEDTLSVALPYFWEHFDKDGWSLWYSEYRFPEELTQTFMSCNLI TGMFQRLDKLRK  
NAFASV GSEETQTLV  
REYFSW

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

=>

# 3) RUN the USGENE BLAST search

=> FILE USGENE

USGENE is updated within **3 days** of publication by the USPTO.

FILE 'USGENE' ENTERED AT 04:59:03 ON 06  
COPYRIGHT (C) 2008 SEQUENCEBASE CORP

FILE LAST UPDATED: 1 FEB 2008 <20080206/UP>  
MOST RECENT PUBLICATION DATE: 31 JAN 2008 <20080131/PD>

FILE COVERS 1982 TO DATE

>>> SIMULTANEOUS LEFT AND RIGHT TRUNCATION (SLART) IS AVAILABLE  
IN THE BASIC INDEX (/BI) AND FEATURE TABLE (/FEAT) FIELDS <<<

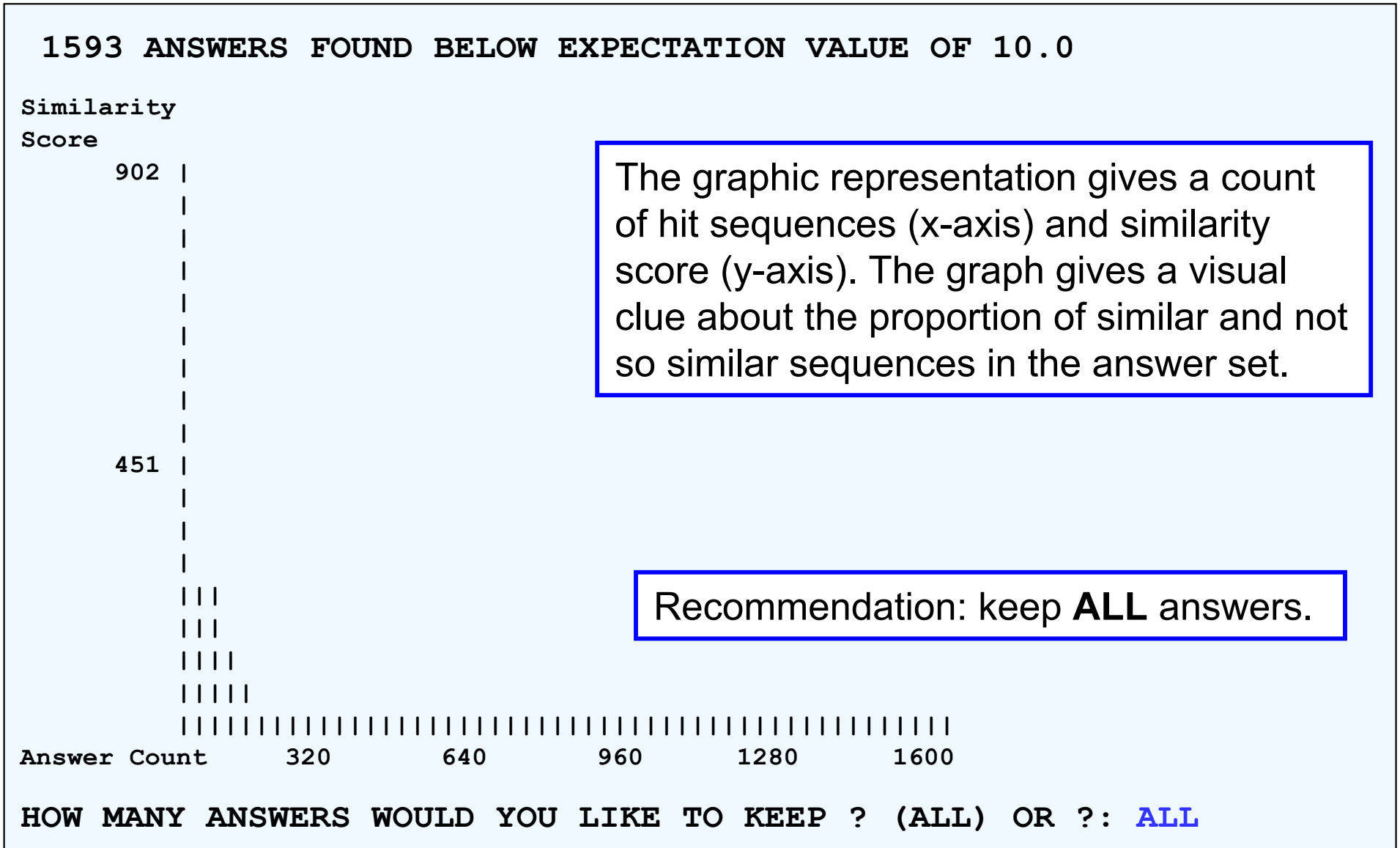
=> RUN BLAST L1 /SQP -F F

Turn the Low Complexity Filter off for the protein (SQP) search using... /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).

# 3) RUN the USGENE BLAST search (cont.)



### 3) RUN the USGENE BLAST search (cont.)

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

L2 RUN STATEMENT CREATED

```
L2      1593 MAAGTLYTYPENWRAFKALIAAQYSGAQVRVLSAPPHFHFGQTNRTPEFL
          RKFPAGKVPAFEGDDGFCVFESNAIAYYVSNEELRGSTPEAAAQVVQWVS
          FADSDIVPPASTWVFPTLGIMHHNKQATENAKEEVRRILGLLDAYLKTRT
          FLVGERVTLADITVVCTLLWLKYQVLEPSFRQAFPNTNRWFLTCINQPQF
          RAVLGEVKLCEKMAQFDAKKFAETQPKKDTPRKEKGSREEKQKPQAEERKE
          EKKAAPAPEEEMDECEQALAAEPKAKDPFAHLPKSTFVLDEFKRKYSNE
          DTLSVALPYFWEHFDDKDGWLSLWYSEYRFPEELTQTFMSCNLITGMFQRLD
          KLRKNAFASVILFGTNNSSSISGVWVFRGQELAFPLSPDWQVDYESYTW
          KLDPGSEETQTLVREYFSWEGAFQHVGKAFNQGKIFK/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 1593 SOR L2 SCORE D

Use SORT SCORE D to sort  
by descending BLAST score.

# 3) RUN the USGENE BLAST search (cont.)

=> D 1-30

Review answers in the free-of-charge default format, including alignment.

L3 ANSWER 1 OF 1593 USGENE COPYE

TI Tissue-and serum-derived glycoproteins and methods of their use  
(PublishedApplication)

MTY Protein

SQL 437

ORGN Homo Sapiens

SEQN 10979

SEQC 14918

BLASTALIGN

Query = 437 letters

Length = 437

Score = 902 bits (2331), Expect = 0.0

Identities = 437/437 (100%), Positives = 437/437 (100%)

The top BLAST score is 902.

Query: 1 MAAGTLYTYPENWRAFKALIAAQYSGAQVRVLSAPPHFHFGQTNRTPEFLRKFPAGKVPA

MAAGTLYTYPENWRAFKALIAAQYSGAQVRVLSAPPHFHFGQTNRTPEFLRKFPAGKVPA

Sbjct: 1 MAAGTLYTYPENWRAFKALIAAQYSGAQVRVLSAPPHFHFGQTNRTPEFLRKFPAGKVPA

Query: 61 FEGDDGFCVFESNAIAYYSNEELRGSTPEAAAQVVQWVSFADSDIVPPASTWVFPTLGI

FEGDDGFCVFESNAIAYYSNEELRGSTPEAAAQVVQWVSFADSDIVPPASTWVFPTLGI

Sbjct: 61 FEGDDGFCVFESNAIAYYSNEELRGSTPEAAAQVVQWVSFADSDIVPPASTWVFPTLGI

. . . .

# 3) RUN the USGENE BLAST search (cont.)

=> D SCORE 1-20

Another way to review quickly is by BLAST SCORE.

L3 ANSWER 1 OF 1593 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN  
SCORE 902

. . . .

L3 ANSWER 10 OF 1593 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN  
SCORE 788

L3 ANSWER 11 OF 1593 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN  
SCORE 717

L3 ANSWER 12 OF 1593 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN  
SCORE 656

. . . .

L3 ANSWER 20 OF 1593 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN  
SCORE 290

=> SOR AN 1-10

PROCESSING COMPLETED FOR L3

L4 10 SOR L3 1-10 AN

Gather selected USGENE hits into a new L-number with SORT AN (L4).

# 4) RUN the DGENE BLAST search

=> FILE DGENE

=> RUN BLAST L1 /SQP -F F

. . . .

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

L5 RUN STATEMENT CREATED

```
L5      1904 MAAGTLYTYPENWRAFKALIAAQYSGAQVRVLSAPPHFHFGQTNRTPEFL
        RKEFPAGKVPAFEGDDGFCVFESNAIAYYVSNEELRGSTPEAAAQVVQVWS
        FADSDIVPPASTWVFPTLGIMHHNKQATENAKEEVRRILGLLDAYLKTRT
        FLVGERVTLADITVVCTLLWLYKQVLEPSFRQAFPNTNRWFLTCINQPQF
        RAVLGEVKLCEKMAQFDACKFAETQPKKDTPRKEKGSREEKQKPQAERKE
        EKKAAPAPEEEMDECEQALAAEPKAKDPFAHLPKSTFVLDEFKRKYSNE
        DTLSVALPYFWEHFDKDGWSLWYSEYRFPEELTQTFMSCNLIITGMFORLD
        KLRKNAFASVILFGTNNSSSISGVWVFRGQELAFPLSPDWQVDYESYTW
        KLDPGSEETQTLVREYFSWEGAFQHVVGKAFNQGKIFK/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 L5

```
L6      1904 SOR L5 SCORE D
```

Turn the Low Complexity Filter off for the protein (SQP) search using... /SQP -F F

Use SORT SCORE D to sort by descending BLAST score.

# 4) RUN the DGENE BLAST search (cont.)

=> D 1-30

Review answers in the free-of-charge default format, including alignment.

L6 ANSWER 1 OF 1904 DGENE COPYE  
AN AEL43555 protein DGENE  
TI New human cancer suppressor proteins and DNA, useful for diagnosing, preventing, and treating human cancers, e.g. cancer of the breast, brain, heart, muscles, large intestine, thymus, spleen, kidney, liver, or small intestine.  
DESC Human cancer suppressor protein GIG35.  
KW diagnosis; therapeutic; prophylaxis; gene therapy; cancer; tumor; neoplasm; cytostatic; GIG35; EEF1G.  
SQL 437  
OS 2006-747536 [76]

BLASTALIGN

Query = 437 letters

Length = 437

Score = 902 bits (2331), Expect = 0.0

Identities = 437/437 (100%), Positives = 437/437 (100%)

Query: 1 MAAGTLYTYPENWRAFKALIAAQYSGAQVRVLSAPPHFHFGQTNRTPEFLRKFPAGKVPA  
MAAGTLYTYPENWRAFKALIAAQYSGAQVRVLSAPPHFHFGQTNRTPEFLRKFPAGKVPA  
Sbjct: 1 MAAGTLYTYPENWRAFKALIAAQYSGAQVRVLSAPPHFHFGQTNRTPEFLRKFPAGKVPA

. . . .

The top BLAST score is 902.

# 4) RUN the DGENE BLAST search (cont.)

=> D SCORE 1-20

Another way to review quickly is by BLAST SCORE.

L6 ANSWER 1 OF 1904 DGENE COPYRIGHT 2008 THE THOMSON CORP on STN  
SCORE 902

. . . .

L6 ANSWER 14 OF 1904 DGENE COPYRIGHT 2008 THE THOMSON CORP on STN  
SCORE 880

L6 ANSWER 15 OF 1904 DGENE COPYRIGHT 2008 THE THOMSON CORP on STN  
SCORE 880

L6 ANSWER 16 OF 1904 DGENE COPYRIGHT 2008 THE THOMSON CORP on STN  
SCORE 717

. . . .

L6 ANSWER 20 OF 1904 DGENE COPYRIGHT 2008 THE THOMSON CORP on STN  
SCORE 495

=> SOR AN 1-15

PROCESSING COMPLETED FOR L6  
L7 15 SOR L6 1-15 AN

Gather selected DGENE hits into a new L-number with SORT AN (L7).

# 5) Transfer PNs from USGENE and DGENE and combine answer sets in DWPI

=> FILE WPINDEX

=> TRA L4 PN; TRA L7 PN

L4 = USGENE selected BLAST hits.  
L7 = DGENE selected BLAST hits.

L8            TRANSFER L4 1- PN :  
L9            8 L8

10 USGENE sequence hits (L4)  
found 8 DWPI records (L9)

L10           TRANSFER L7 1- PN :  
L11           12 L10

15 DGENE sequence hits (L7)  
found 12 DWPI records (L11).

=> S L9 OR L11

L12           16 L9 OR L11

Total DWPI records is 16 (L12) – both USGENE and DGENE have found unique DWPI patent families!

# 6) Merge results with duplicate identify (DUP IDE) and sort by patent family (FSORT)

=> DUP IDE L4 L7 L12

DUPLICATE IS NOT AVAILABLE IN 'USGENE',  
ANSWERS FROM THESE FILES WILL BE CONSID

L4 = USGENE selected BLAST hits.  
L7 = DGENE selected BLAST hits.  
L12 = corresponding DWPI records.

FILE 'USGENE' ENTERED AT 05:16:23 ON 06 FEB 2008  
COPYRIGHT (C) 2008 SEQUENCEBASE CORP

FILE 'DGENE' ENTERED AT 05:16:23 ON 06 FEB 2008  
COPYRIGHT (C) 2008 THE THOMSON CORPORATION

FILE 'WPINDEX' ENTERED AT 05:16:23 ON 06 FEB 2008  
COPYRIGHT (C) 2008 THE THOMSON CORPORATION

PROCESSING COMPLETED FOR L4  
PROCESSING COMPLETED FOR L7  
PROCESSING COMPLETED FOR L12

L13            41 DUP IDE L4 L7 L12 (INCLUDES 0 SETS OF DUPLICATES)  
                  ANSWERS '1-10' FROM FILE USGENE  
                  ANSWERS '11-25' FROM FILE DGENE  
                  ANSWERS '26-41' FROM FILE WPINDEX

# 6) Merge results with DUP IDE and sort by patent family (FSORT) (cont.)

=> FSORT L13

L14 41 FSO L13

15 Multi-record Families	Answers 1-41
Family 1	Answers 1-8
Family 2	Answers 9-11
Family 3	Answers 12-13
Family 4	Answers 14-15
Family 5	Answers 16-17
Family 6	Answers 18-19
Family 7	Answers 20-22
Family 8	Answers 23-25
Family 9	Answers 26-28
Family 10	Answers 29-30
Family 11	Answers 31-32
Family 12	Answers 33-34
Family 13	Answers 35-37
Family 14	Answers 38-39
Family 15	Answers 40-41
0 Individual Records	
0 Non-patent Records	

The 16 DWPI records (L12), 10 USGENE sequence hits and 15 DGENE sequence hits belong to 15 FSORT families (L14).

Use the patent family display (PFAM) feature to display selective records from a FSORT L-number

General format of PFAM:

=> D L# PFAM=# RECORD# FORMAT

Examples using PFAM:

=> D PFAM=1-10

1st member of patent family number 1-10 in default display format

=> D PFAM=2 TRI ORGN ALIGN TOTAL

All members of family number 2 in a free sequence review format

# 7) Display results using the customized file default display formats (see slide 17)

=> D PFAM=1- TOTAL

This displays all records (TOTAL), from all families (PFAM=1-) in file default format.

L14 ANSWER 9 OF 41 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN FAMILY2

TI Compositions and methods for the diagnosis and treatment of tumor  
(PublishedApplication)

MTY Protein

SQL 437

ORGN Homo Sapiens

SEQN 2421

SEQC 6355

BLASTALIGN

USGENE hit sequence display(s).

Query = 437 letters

Length = 437

Score = 889 bits (2296), Expect = 0.0

Identities = 430/437 (98%), Positives = 433/437 (98%)

Query: 1 MAAGTLYTYPENWRAFKALIAAQYSGAQVRVLSAPPHFHFGQTNRTPEFLRKFPAGKVPA

MAAGTLYTYPENWRAFKALIAAQYSGAQ+RVLSAPPHFHFGQTNRT EFLRKFPAGKVPA

Sbjct: 1 MAAGTLYTYPENWRAFKALIAAQYSGAQIRVLSAPPHFHFGQTNRTSEFLRKFPAGKVPA

Query: 61 FEGDDGFCVFESNAIAYYVSNEELRGSTPEAAAQVVQWVSFADSDIVPPASTWVFPTLGI

FEGDDGFCVFESNAIAYYVSNEELRGSTPEAAAQVVQWVSFADSDIVPPASTWVFPTLGI

Sbjct: 61 FEGDDGFCVFESNAIAYYVSNEELRGSTPEAAAQVVQWVSFADSDIVPPASTWVFPTLGI

. . . .

# 7) Display results using the customized file default display formats (cont.)

```
L14 ANSWER 10 OF 41 DGENE COPYRIGHT 2008 THE THOMSON CORP on STN FAMILY2
AN ABM80939 protein DGENE
TI New tumor-associated antigenic target polypeptides and nucleic acids,
useful in preparing a medicament for treating or detecting a
proliferative disorder, e.g. breast, lung, colorectal, ovarian or
prostate cancer or tumor.
DESC Tumour-associated antigenic target (TAT) polypeptide PRO81615,
SEQ:2421.
KW Tumour-associated antigenic target; TAT; human; overexpression;
cancer; tumour; diagnosis; cell proliferative disorder; breast
cancer; colorectal cancer; lung cancer; ovarian cancer; . . . .
SQL 437
OS 2004-347921 [32]
BLASTALIGN
Query = 437 letters
Length = 437
Score = 889 bits (2296), Expect = 0.0
Identities = 430/437 (98%), Positives = 433/437 (98%)
Query: 1 MAAGTLYTYPENWRAFKALIAAQYSGAQVRVLSAPPHFHFGQTNRTPEFLRKFPAGKVPA
MAAGTLYTYPENWRAFKALIAAQYSGAQ+RVLSAPPHFHFGQTNRT EFLRKFPAGKVPA
Sbjct: 1 MAAGTLYTYPENWRAFKALIAAQYSGAQIRVLSAPPHFHFGQTNRTSEFLRKFPAGKVPA
. . . .
```

DGENE hit sequence display(s).

# 7) Display results using the customized file default display formats (cont.)

```
L14 ANSWER 11 OF 41 WPINDEX COPYRIGHT 2008 THE THOMSON CORP on STN FAMILY2
AN 2004-347921 [32] WPINDEX
TI New tumor-associated antigenic target polypeptides and nucleic acids,
    useful in preparing a medicament for treating or detecting a
    proliferative disorder, e.g. breast, lung, colorectal, ovarian or
    prostate cancer or tumor
DC B04; D16; S03
IN WU T D; ZHANG Z; ZHOU Y
PA (GETH-C) GENENTECH INC
CYC 105
PIA WO 2004030615 A2 20040415 (200432)* EN 7273[635] <--
    AU 2003295328 A1 20040423 (200465) EN
    EP 1594447 A2 20051116 (200575) EN
    JP 2006516089 W 20060622 (200641) JA 1466
ADT WO 2004030615 A2 WO 2003-US28547 20030929; AU 2003295328 A1
    AU 2003-295328 20030929; EP 1594447 A2 EP 2003-786510 20030929;
    EP 1594447 A2 WO 2003-US28547 20030929; JP 2006516089 W
    WO 2003-US28547 20030929; JP 2006516089 W JP 2004-541530 20030929
FDT AU 2003295328 A1 Based on WO 2004030615 A; EP 1594447 A2
    Based on WO 2004030615 A; JP 2006516089 W Based on WO 2004030615 A
PRAI US 2002-414971P 20021002
```

WPINDEX patent family display.

# Agenda

35

- DGENE and USGENE database content
- The importance of DWPI patent families
- Multifile “best-practice” technique using BLAST
- Step-by-step walk through a multifile search
- **Overview of case-study search results**
- Examples of unique USGENE retrieval
- Conclusions

# Summary of results for *Eukaryotic translation elongation factor 1 gamma* (NP\_001395)

	<b>SEQs</b>	<b>SEQs &gt; 80%</b>	<b>PNs</b>	<b>DWPI Records</b>	<b>FSORT Families</b>
<b>DGENE</b>	1904	15	12	12	11
<b>USGENE</b>	1593	10	9	8*	8
<b>Overlap</b>	-	-	0	4	4
<b>Total</b>	-	-	21	16	15

(\* USGENE US20070224201 was **not** present in DWPI, as of February 6<sup>th</sup>, 2008.)

# Example: USGENE unique retrieval

37

```
L14 ANSWER 18 OF 41 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN FAMILY6
TI Genetic polymorphisms associated with coronary heart disease, methods
of detection and uses thereof (PublishedApplication)
MTY Protein
SQL 437
ORGN Homo Sapiens
SEQN 138
SEQC 17377
BLASTALIGN
```

This USGENE hit sequence uniquely retrieved the DWPI record on the following slide (as of Feb 6<sup>th</sup> 2008).

Query = 437 letters

Length = 437

Score = 889 bits (2296), Expect = 0.0

Identities = 430/437 (98%), Positives = 433/437 (98%)

```
Query: 1 MAAGTLYTYPENWRAFKALIAAQYSGAQVRVLSAPPHFHFGQTNRTPEFLRKFPAGKVPA
MAAGTLYTYPENWRAFKALIAAQYSGAQ+RVLSAPPHFHFGQTNRT EFLRKFPAGKVPA
```

```
Sbjct: 1 MAAGTLYTYPENWRAFKALIAAQYSGAQIRVLSAPPHFHFGQTNRTSEFLRKFPAGKVPA
```

```
Query: 61 FEGDDGFCVFESNAIAYYSNEELRGSTPEAAAQVVQWVSFADSDIVPPASTWVFPTLGI
FEGDDGFCVFESNAIAYYSNEELRGSTPEAAAQVVQWVSFADSDIVPPASTWVFPTLGI
```

```
Sbjct: 61 FEGDDGFCVFESNAIAYYSNEELRGSTPEAAAQVVQWVSFADSDIVPPASTWVFPTLGI
```

```
Query: 121 MHHNKQATENAKEEVRRI LGLLDAYLKTRTFLVGERVTLADITVVCTLLWLYKQVLEPSF
MHHNKQATENAKEEVRRI LGLLDAYLKTRTFLVGERVTLADITVVCTLLWLYKQVLEPSF
```

```
Sbjct: 121 MHHNKQATENAKEEVRRI LGLLDAYLKTRTFLVGERVTLADITVVCTLLWLYKQVLEPSF
```

. . . .

# Example: USGENE unique retrieval (cont.)

38

```
L14 ANSWER 19 OF 41 WPINDEX COPYRIGHT 2008 THE THOMSON CORP on STN FAMILY6
AN 2005-630949 [64] WPINDEX
TI New isolated nucleic acid molecule comprising a single nucleotide
polymorphism, useful for identifying an individual at an increased
risk of developing coronary heart disease, or for treating or
preventing myocardial infarction
DC B04; D16
IN CARGILL M; DEVLIN J; DEVLIN J J;
PA (APPL-N) APPLERA CORP
CYC 108
PIA WO 2005087953 A2 20050922 (200509) EN 199111
US 20060228715 A1 20061012 (200668) EN <--
EP 1745147 A2 20070124 (200708) EN
ADT WO 2005087953 A2 WO 2005-US7453 20050307; US 20060228715 A1
Provisional US 2004-550051P 20040305; US 20060228715 A1 Provisional US
2004-567831P 20040505; US 20060228715 A1 Provisional US 2004-617163P
20041012; US 20060228715 A1 US 2005-73360 20050307; EP 1745147 A2
EP 2005-724897 20050307; EP 1745147 A2 WO 2005-US7453 20050307
FDT EP 1745147 A2 Based on WO 2005087953 A
PRAI US 2004-617163P 20041012
US 2004-550051P 20040305
US 2004-567831P 20040505
US 2005-73360 20050307
```

This relevant DWPI record was uniquely retrieved via a USGENE BLAST search (on Feb 6<sup>th</sup>, 2008).

# Results of applying the same multifile techniques to a second search example

39

## Search Question:

Find relevant patent references for *Human Tumor Necrosis Factor (TNF) alpha* (AAC03542)

VRSSSRTPSDKPVAVVAVANPQAEGQLQWLNRRANALLANGVELRDNQLVVPSEGLYLIYSQVLF  
KGQGCPSHVLTLTHTISRIAVSYQTKVNLLSAIKSPCQRETPRGAEAKPWYEPIYLGGVFQLEK  
GDRLSAEINRPDYLDFAESGQVYFGI IAL

( Search conducted on February 6<sup>th</sup>, 2008.)

# Summary of results for *Human Tumor Necrosis Factor (TNF) alpha* (AAC03542)

	<b>SEQs</b>	<b>SEQs &gt; 80%</b>	<b>PNs</b>	<b>DWPI Records</b>	<b>FSORT Families</b>
<b>DGENE</b>	2871	702	314	314	216
<b>USGENE</b>	2035	332	190	136*	85
<b>Overlap</b>	-	-	14	101	58
<b>Total</b>	-	-	490	348	242

(\* USGENE US20070185017 and US20080025976 were **not** present in DWPI, as of February 6<sup>th</sup>, 2008.)

# Example: USGENE unique retrieval

41

L16 ANSWER 1380 OF 1382 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN  
FAMILY 242

TI Method for purifying a physiologically active substance produced by  
recombinant DNA technique (Patent)

MTY Protein

SQL 155

ORGN Unknown

SEQN 1

SEQC 3

BLASTALIGN

Query = 157 letters

Length = 155

Score = 311 bits (796), Expect = 4e-90

Identities = 154/155 (99%), Positives = 154/155 (99%)

Query: 3 SSSRTPSDKPVAVHVVANPQAEGQLQWLNRRANALLANGVELRDNQLVVPSEGLYLIYSQV  
SSSRTPSDKPVAVHVVANPQAEGQLQWLNRRANALLANGVELRDNQLVVPSEGLYLIYSQV

Sbjct: 1 SSSRTPSDKPVAVHVVANPQAEGQLQWLNRRANALLANGVELRDNQLVVPSEGLYLIYSQV

Query: 63 LFKGQGCPSTHVLLTHTISRIAVSYQTKVNLLSAIKSPCQRETPRGAEAKPWYEPIYLG  
LFKGQGCPSTHVLLTHTISRIAVSYQTKVNLLSAIKSPCQRETP GAEAKPWYEPIYLG

Sbjct: 61 LFKGQGCPSTHVLLTHTISRIAVSYQTKVNLLSAIKSPCQRETPGAEAKPWYEPIYLG

Query: 123 VFQLEKGDRLSAEINRPDYLDFAESGQVYFGIIAL 157  
VFQLEKGDRLSAEINRPDYLDFAESGQVYFGIIAL

Sbjct: 121 VFQLEKGDRLSAEINRPDYLDFAESGQVYFGIIAL 155

This USGENE hit sequence uniquely  
retrieved the DWPI record on the  
following slide (as of Feb 6<sup>th</sup> 2008).

# Example: USGENE unique retrieval (cont.)

```
L16 ANSWER 1381 OF 1382 WPINDEX COPYRIGHT 2008 THE THOMSON CORP on STN
  FAMILY 242
AN 1986-145432 [23] WPINDEX
TI Purificn. of human tissue necrosis factor prod. - comprises
  chromatography on dye bonded crosslinked agarose gel column when prod.
  obtd. by recombinant dna techniques
DC B04; D16
IN HAYASHI H; KAJIHARA J; KAJIWARA J; KIYOTA T
PA (ASAH-C) ASAHI CHEM IND CO LTD; (ASAH-C) ASAHI KASEI KOGYO KK
PIA EP 183198 A 19860604 (1986)
  AU 8550248 A 19860529 (1986)
  JP 61124392 A 19860612 (1986)
  DD 240216 A 19861022 (1987)
  CN 86104077 A 19871202 (1988)
  US 4880915 A 19891114 (1990) EN <--
  CA 1265650 A 19900206 (1990) EN
  EP 183198 B 19900307 (1990) EN
  DE 3576362 G 19900412 (1990) DE
  JP 02029317 B 19900628 (1990) JA
  IL 77101 A 19910131 (1991) EN
  SU 1630602 A 19910223 (1991) RU
  KR 9400542 B1 19940124 (1995) KO
ADT EP 183198 A EP 1985-114806 19851121; JP 61124392 A . . . .
PRAI JP 1984-246184 19841122
```

This relevant DWPI record was uniquely retrieved via a USGENE BLAST search (on Feb 6<sup>th</sup>, 2008).

# Results of applying the same multfile techniques to a third search example

43

## Search Question:

Find relevant patent references for the *Hepatitis C virus 5' region* (M58406)

```
GCCAGCCCCCTGATGGGGGCGACACTCCACCATGAATCACTCCCCTGTGAGGAACTACTGTCTT  
CACGCAGAAAGCGTCTAGCCATGGCGTTAGTATGAGTGTCGTGCAGCCTCCAGGACCCCCCTC  
CCGGGAGAGCCATAGTGGTCTGCGGAACCGGTGAGTACACCGGAATTGCCAGGACGACCGGGTC  
CTTTCTTGGATCAACCCGCTCAATGCCTGGAGATTTGGGCGTGCCCCCGCAAGACTGCTAGCCG  
AGTAGTGTTGGGTCGCGAAAGGCCTTGTGGTACTGCCTGATAGGGTGCTTGCGAGTGCCCCGGG  
AGGTCTCGTAGACCGTGCACC
```

( Search conducted on February 6<sup>th</sup>, 2008.)

# Summary of multifile search results for *Hepatitis C virus 5' region (M58406)*

	<b>SEQs</b>	<b>SEQs &gt; 80%</b>	<b>PNs</b>	<b>DWPI Records</b>	<b>FSORT Families</b>
<b>DGENE</b>	8961	404	174	174	138
<b>USGENE</b>	4117	263	117	65*	52
<b>Overlap</b>	-	-	7	59	47
<b>Total</b>	-	-	284	180	143

(\* USGENE US20080026952 and US7244585 were **not** present in DWPI, as of February 6<sup>th</sup>, 2008.)

# Example: USGENE unique retrieval

```
L18  ANSWER 838 OF 847  USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
      FAMILY 140
TI   Method for inducing hepatitis c virus (hcv) replication in vitro,
      cells and cell lines enabling robust hcv replication and kit therefor
      (PublishedApplication)
MTY  Nucleic acid
SQL  383
ORGN  Epstein Barr virus
SEQN  17
SEQC  37
```

This USGENE hit sequence uniquely retrieved the DWPI record on the following slide (as of Feb 6<sup>th</sup>, 2008).

## BLASTALIGN

```
Query  = 341 letters
Length = 383
Score  = 626 bits (316), Expect = 0.0
Identities = 333/341 (97%)
Strand = Plus / Plus
```

```
Query: 1  gccagccccctgatgggggcgacactccaccatgaatcactcccctgtgaggaactactg
          |||
Sbjct: 1  gccagccccctgatgggggcgacactccaccatgaatcactcccctgtgaggaactactg

Query: 61  tcttcacgcagaaagcgtctagccatggcgtagtatgagtgctcgtgcagcctccaggan
          |||
Sbjct: 61  tcttcacgcagaaagcgtctagccatggcgtagtatgagtgctcgtgcagcctccaggac
```

# Example: USGENE unique retrieval (cont.)

```
L18 ANSWER 839 OF 847 WPINDEX COPYRIGHT 2008 THE THOMSON CORP on STN
  FAMILY 140
AN 2005-122423 [13] WPINDEX
TI Generating an established cell line that produces hepatitis C virus
  (HCV) for identifying a compound with anti-HCV activity, comprises
  transforming peripheral blood mononuclear cells that produce HCV with
  Epstein Barr virus
DC B04; D16; S03
IN LOPEZ LASTRA M; LOPEZ-LASTRA M; SONENBERG N
PA (UYMC-N) UNIV MCGILL; (LOPE-I) LOPEZ LASTRA M
CYC 106
PIA WO 2005005625 A2 20050120 (200513) EN
  CA 2436104 A1 20050114 (200513) EN
  CA 2454540 A1 20050114 (200513) EN
  US 20070099179 A1 20070503 (200731) EN <--
ADT WO 2005005625 A2 WO 2004-CA1009 20040714; CA 2436104 A1
  CA 2003-2436104 20030714; CA 2454540 A1 CA 2004-2454540 20040206;
  US 20070099179 A1 WO 2004-CA1009 20040714; US 20070099179 A1
  US 2006-564886 20060915
PRAI CA 2004-2454540 20040206
```

This relevant DWPI record was uniquely retrieved via a USGENE BLAST search (on Feb 6<sup>th</sup> 2008).

# Conclusions

- GENESEQ (DGENE) is the “industry-standard” prior-art patent sequence database and must be used for every type of patent sequence search
- USGENE is a vital additional resource with an extensive and timely archive of both U.S. Issued Patent and Published Application sequence data
- A “best-practice” approach to multiframe searching uses DWPI to more accurately group together sequence hits from USGENE and DGENE
- USGENE and DGENE often find unique relevant hits and should always be used in combination

# STN<sup>®</sup>

Synergies and Surprises –  
USGENE<sup>®</sup> and DGENE Multifile Patent  
Sequence Searching on STN<sup>®</sup>

Robert Austin – FIZ Karlsruhe