

STN[®]

Multifile Patent Sequence
Searching on STN[®]

Robert Austin – FIZ Karlsruhe

Agenda

- STN sequence searchable databases
- 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
- Comparisons and conclusions

STN sequence searchable databases

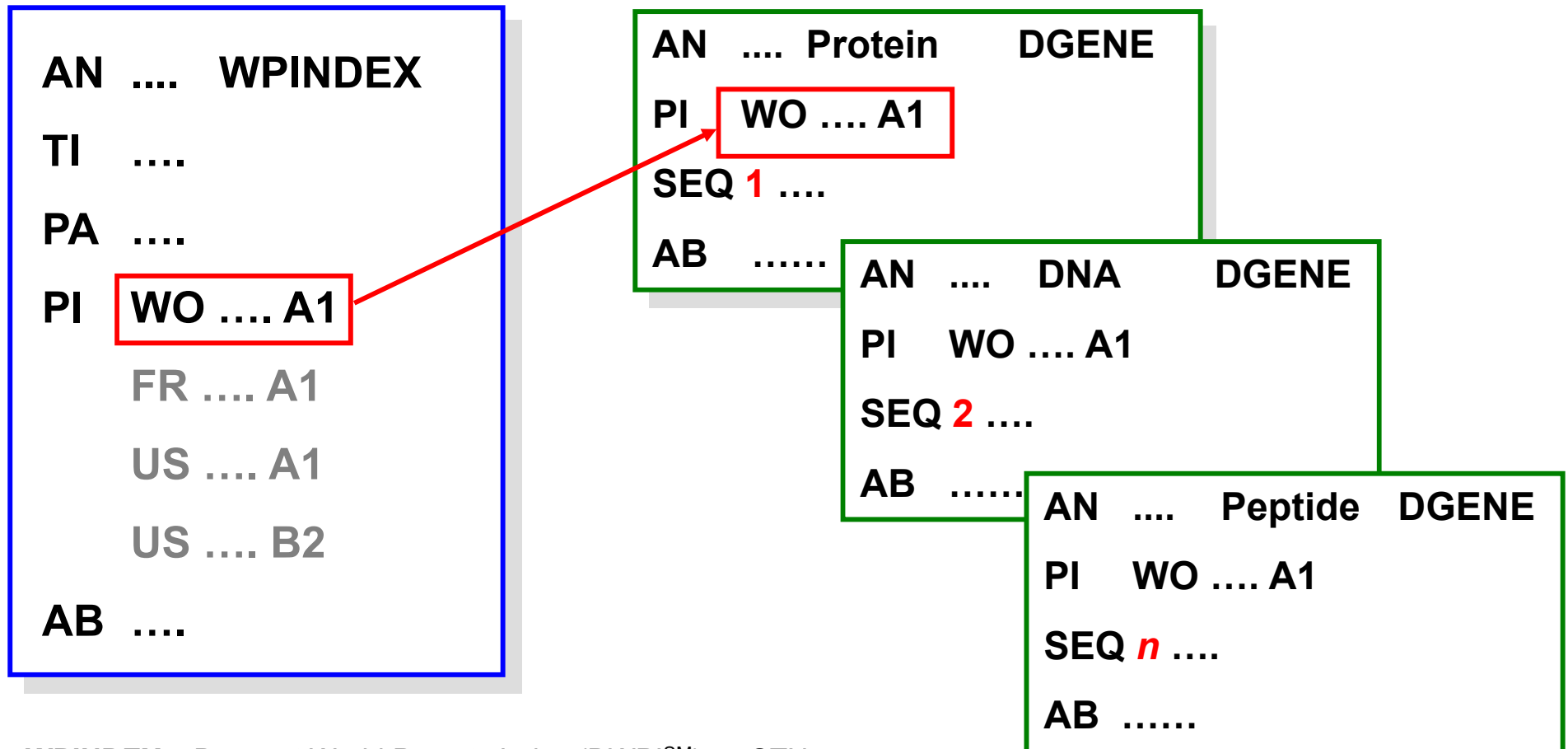
- **DGENE**
 - Thomson Reuters GENESEQ™
 - Value-added patent sequence data from around the globe
- **USGENE®**
 - SequenceBase USPTO Genetic Sequence Database
 - A new and unique access point to USPTO sequence data
- **PCTGEN**
 - WIPO/PCT Patent Application Biosequences
 - The complete collection of e-published sequences from WIPO
- **CAS REGISTRYSM**
 - Chemical Abstracts Service (CAS) REGISTRY
 - Worldwide value-added patent and non-patent sequence data

Thomson Reuters GENESEQ (DGENE)

- 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: www.trilateral.net/projects/biotechnology/search_guidebook_vers_1.pdf

Relationship between DWPI patent family and DGENE sequence database



WPINDEX = Derwent World Patents Index (DWPISM) on STN

DGENE = GENESEQ on STN

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

- DWPI patent title, concise sequence description, abstract, and keyword indexing *per sequence*
 - Context of *each sequence* illuminated 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 provided by indexers
- Patent sequence location (claim, example, etc.)
 - Assigned manually by Thomson Reuters indexers
 - Flexible filtering of searches to those described in the claims
- Sequences intellectually derived by indexers
 - Unique sequence hits not disclosed in formal listings

Some editorial insights regarding WIPO/PCT sequences indexed in DGENE

- On average 120 WIPO/PCT basic patents have sequences indexed into DGENE each week
- Of these, about 15-20 may have electronic listings available – the rest are keyed manually
 - Sequences are independently double-keyed with a guaranteed accuracy of 99.995% (1 in 20,000)
- About 15% of PCTs with electronic listings have extra sequences indexed from the specification
- Typically 1 or 2 documents per week will also have intellectually derived sequences indexed, based upon the wording of the patent claims

Source: Colin Williams, GENESEQ Editorial & Content Manager, Thomson Reuters (12/2006)

Derived sequences are intellectually created by indexers from wording in the patent text

AN AEJ92622 protein DGENE
TI Hydrolyzing/synthesizing carboxylic acid ester/amide from chiral/prochiral reactants for preparing e.g. pharmaceuticals, comprises contacting reactants with a polypeptide having hydrolytic activity.
IN Svendsen A; Vind J; Brask J; De Maria L
PA (NOVO) NOVOZYMES AS.
PI WO 2006084470 A2 20060817 17
AI WO 2006-DK76 20060210
PRAI EP 2005-388012 20050210
PSL Claim 16
DED 19 OCT 2006 (first entry)
LA English
OS 2006-560037 [57]
DESC Variant fungal lipolytic hydrolase #2.
KW hydrolysis; lipase; pharmaceutical; pesticide; enzyme; mutein.
ORGN Thermomyces lanuginosus. Synthetic.
AB The new invention relates to a enzymatic method of hydrolyzing or synthesizing carboxylic acid ester or amide from chiral or prochiral reactants, by providing reactants for hydrolysis or synthesis, and contacting the reactants with a polypeptide which has hydrolytic activity on ester or amide, and a sequence 50% homologous to Thermomyces lanuginosus lipase. Also described is a polypeptide, which has hydrolase activity on an ester or amide substrate, and has an amino acid sequence that has at least 80% identity to SEQ ID No: 5 and compared to SEQ ID No: 5 comprises a substitution corresponding to I90Q, N92TD, F95Y, F113Y, I202M, V203GM, L269T and 270F. . . .

In this example, the indexer has intellectually derived this sequence from the wild-type lipolytic hydrolase.

Indexers explain exactly how they derived the sequence at the end of the abstract

The polypeptide is at least 80% homologous to any of SEQ ID No: 1-6 being amino acid sequences of lipolytic enzymes of fungus such as *Rhizomucor miehei* (SWISSPROT P19515), *Rhizopus delemar*, *Fusarium oxysporum*, *Penicillium camemberti* (SWISSPROT P25234), *Thermomyces lanuginosus* (SWISSPROT 059952) and *Thermomyces ibadanensis* The method is useful in the preparation of pharmaceuticals or pesticides, where the synthesis includes synthesis of 2-butyl propionate. This sequence is a variant fungal lipolytic hydrolase (lipase), V203M T231R N233R. This sequence is not shown in the specification, but was created by the indexer using the information given in claim 16.


```
SQL      269
SEQ      1  evsqdlfnqf nlfagysaaa ycgknndapa gtnitctgna cpevekadat
        51  flysfedsgv gdvtgflald ntnklivlsf rgsrsienwi gnlnfdlkei
        101 ndicsgcrgh dgftsswrsv adtlrpkved avrehpdyrv vftghslgga
        151 latvagadlr gng
        201 dimprlppre fgy
        251 nipdipahlw yfg
```

The indexer has added explanatory sentences to the abstract and annotations to the feature table.

FEATURE TABLE:

Key	Location	Qualifier	
Modified-site	203	note	"Wild type Val replaced by Met"
Modified-site	231	note	"Wild type Thr replaced by Arg"
Modified-site	233	note	"Wild type Asn replaced by Arg"

SequenceBase USPTO Genetic Sequence Database (USGENE)

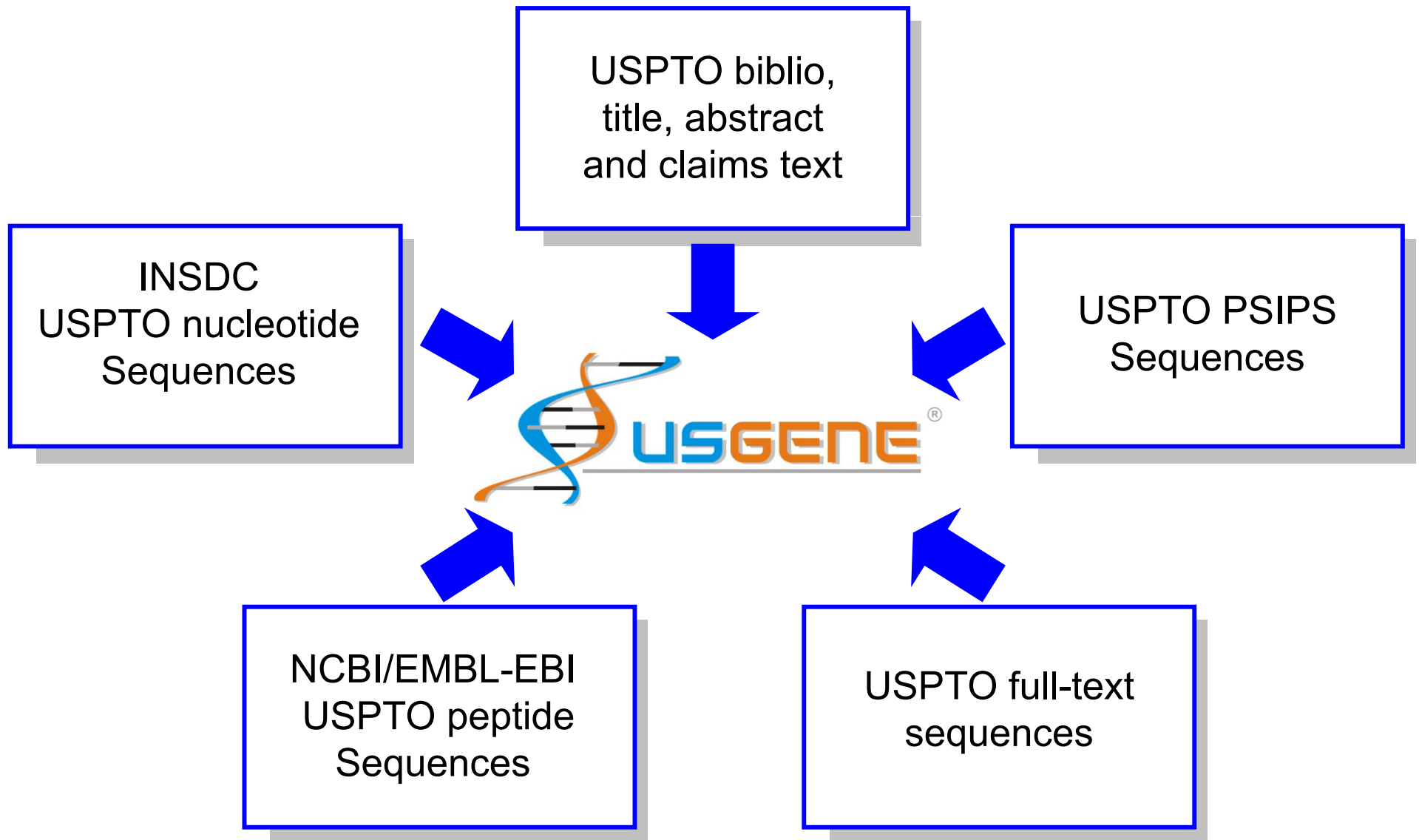
- Sequences from all relevant USPTO published patent applications and granted (issued) patents
- Assignee and full inventor names; publication, application and parent case PCT numbers and dates; original publication **title**, **abstract**, and **claims**
- Organism name, Sequence Length, Molecule Type, SEQ ID NO, Feature Tables and **Patent Sequence Location**
- **Patent Family and Legal Status display** 
- Updated weekly – within **3 days** of publication
- 1982 – present

USGENE consolidates unique USPTO sequence data from different sources

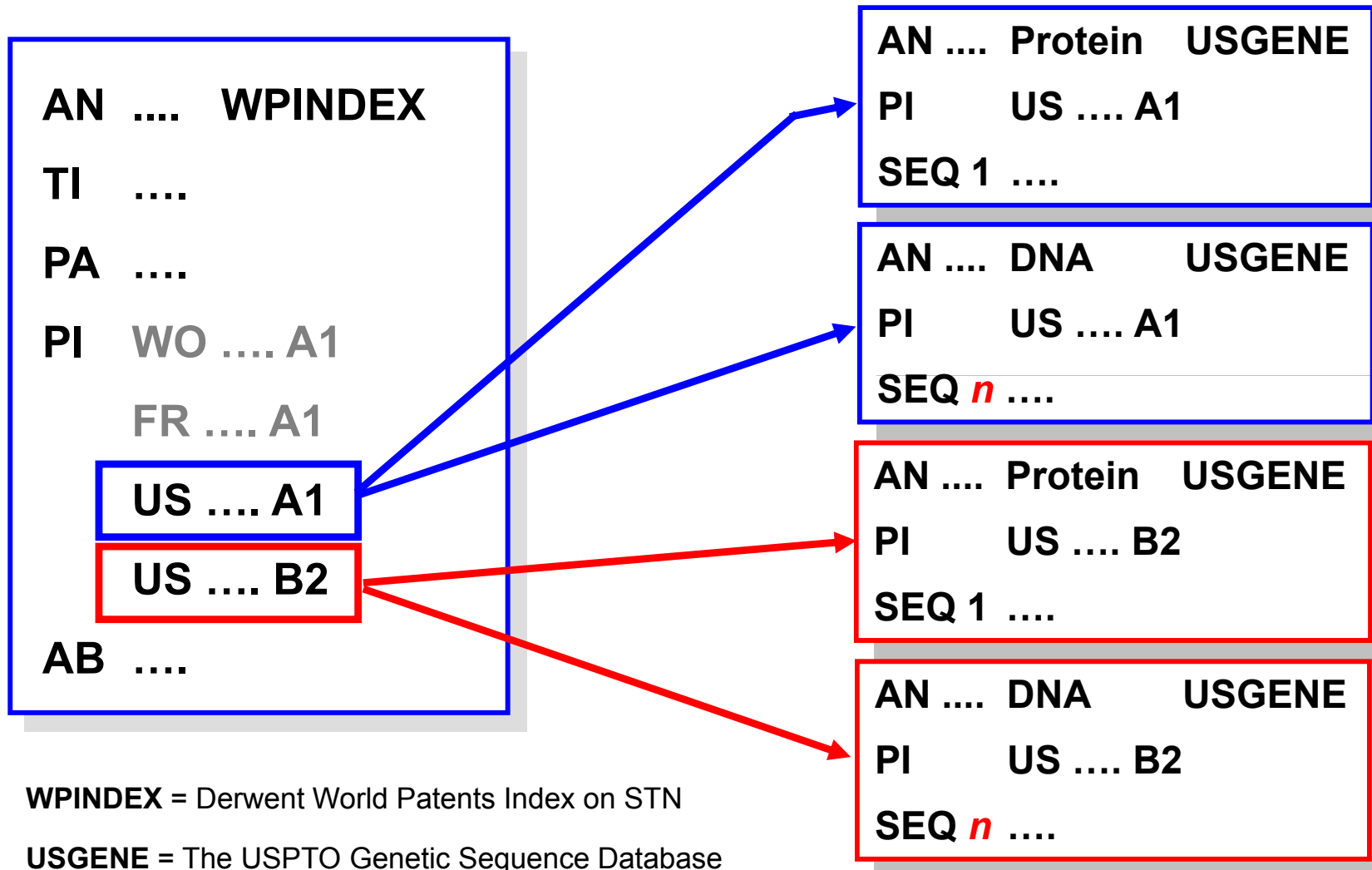
1. USPTO Publication Site for Issued and Published Sequences (PSIPS)
 - The official mega-publication download site, 2001-date
2. International Nucleotide Sequence Database Collaboration (INSDC) (NCBI/EMBL/DDBJ, Genbank)
 - U.S. granted patent nucleotide sequences, 1982-date
3. USPTO Protein Database (NCBI/EMBL)
 - U.S. granted patent protein/peptide sequences, 1982-date
4. USPTO Published Applications and Patents Full-Text
 - Filling in omissions, coverage gaps and to enhance timeliness

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

USGENE combines these sequences with bibliographic data and claims text



Relationship between DWPI patent family and USGENE sequence databases



USGENE sequence records are available within 3 days of publication by the USPTO

```
L3 ANSWER 1 OF 1 USGENE COPYRIGHT 2009 SEQUENCEBASE CORP on STN
AN 20080256649.126 Protein USGENE
TI Novel Acetylcholinesterase Gene Responsible for resistance and Applications Thereof (Published
IN Weill Mylene (Montpellier, FR); Fort Philippe Raymond Michel (Montpellier, FR); Pasteur Nico
PA CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE (PARIS CEDEX FR)
PI US 20080256649 A1 20081016
AI US 2003-518072 20030619
RLI WO 2003-FR1876 20030619
PSL Claim 4; SEQ ID NO 126
ED 20081017
AB The invention relates to a gene responsible for resistance in mosquitoes, which encodes an acetylcholinesterase protein (AChE) and a protein AChE1) and
ECLM US20080256649 A1: 1. An insect acetylcholinesterase, characterized in that it comprises a central catalytic region which has an amino acid sequence selected from the group consisting of the sequence SEQ ID NO 1 and the sequences exhibiting at least 60% identity or 70% similarity with the sequence SEQ ID NO 1, with the exclusion . . . .
SSO PROTEIN; USPTO; APPLICATION
ORGN Anopheles gambiae
SQL 737
SEQ
1 meirgllmgr lrlgrrmvpl gllgvt
51 igshqlsaaa gvglssqsaq sgslas
```

AN 20080256649.126 is SEQ ID NO: 126 from US20080069867.

Published Application sequences published each Thursday, are typically available within 1 day of publication on Friday of each week.

AN 20080256649.126 is displayed here in BRIEF format, which includes the Exemplary Claim (ECLM).

USGENE also has an extensive backfile

United States Patent: 5210028 - Mozilla Firefox

http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=

Escherichia coli LC137 transformed with pCL.sub.857 and pPLMu/IGFII

SEQUENCE LISTING (1) GENERAL INFORMATION:

(iii) NUMBER OF SEQUENCES: 4 (2) INFORMATION FOR SEQ ID NO:1: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 67 amino acids (B) TYPE: amino acid (D) TOPOLOGY: linear (ii) MOLECULE TYPE: protein (iii) HYPOTHETICAL: N (iv) ANTI-SENSE: N (vi) ORIGINAL SOURCE: (A) ORGANISM: Human IGF-II (xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:
AlaTyrArgProSerGluThrLeuCysGlyGlyGluLeuValAspThr 151015 LeuGlnPheValCysGlyAsp ArgGlyPheTyrPheSerArgProAla 202530
SerArgValSerArgArgSerArgGlyIleValGluGluCysCysPhe 35 4045 ArgSerCysAspLeuAlaLeuLeuGluThrTyrCysAlaThrProAla 505560 LysSerGlu 65 (2)

INFORMATION FOR SEQ ID NO:2: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 216 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: double (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (iii) HYPOTHETICAL: N (iv) ANTI-SENSE: N (vi) ORIGINAL SOURCE: (A) ORGANISM: Synthetic Human IGF-II DNA Sequence; E.coli pref (B) STRAIN: pBB8/IGF-II (xi) SEQUENCE DESCRIPTION: SEQ ID NO:2: CATATGGCATAACCG CCCGAGCGAGACCCTGTGCGGTGGCGAGCTCGTAGACACTCTGCAG60
TTCGTTTGTGGTGACCCTGGCTTCTACTTCTCTCGTCCTGCTAGCCGTGTATCTCGCCGT120
TCTAGAGGCATCGTTGAAGAGTGCTGTTTCCGCAGCTGTGACTACTGCGCAACTCCAGCAAATCCGAATAAGGATCC216 (2) INFORMATION FOR SEQ ID NO:3: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 233 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: double (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (iii) HYPOTHETICAL: N (iv) ANTI-SENSE: N (vi) ORIGINAL SOURCE: (A) ORGANISM: Human IGF-II DNA Sequence; E.coli pref (B) STRAIN: pIGF-II/3 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:3: GAATTCGACGCTTATGGCTTACAGACCATCCGAAACCTTGTCTCACCTTGCAATTCGTTTGTGGTGACAGAGGTTT CTACTTCTTTCTAGAAGATCCAGAGGTATCGTTGAAGAATGTTGTTTCA GTTGGAAACCTACTGTGCTACCCAGCTAAGTCTGAATGAATGCGTCGAATTC233 (2) INFORMATION FOR SEQ ID NO:4: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 150 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: double (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (iii) HYPOTHETICAL: N (iv) ANTI-SENSE: N (vi) ORIGINAL SOURCE: (A) ORGANISM: Human IGF-II DNA Sequence; E.coli pref (B) STRAIN: pPLMu (xi) SEQUENCE DESCRIPTION: SEQ ID NO:4: CTTACACTTAGTTAAATTGCTAACTTTATAGATTACAAAACCTTACACCATGGTTACGAATTCCCGGGGATCCGTGCACCTGCAGCCAAGGTTTCGGTGATGACGGTGAAAACCTCTGAC 150

Done

This USPTO example is US5210028, which was issued on May 11th, 1993.

Published sequence data like this are identified, extracted, standardized, and loaded into USGENE on STN (compare this to the STN record on the next slide).

Note: US5210028 is an example of a U.S. patent which is not available at NCBI Genbank or EMBL-EBI.

To facilitate precise searching all USGENE sequences are in STN standardized format

```
L1 ANSWER 1 OF 1 USGENE COPYRIGHT 2009 SEQUENCEBASE CORP on STN
AN 5210028.1 protein USGENE
TI Process for the production of unfused IGF-II (Patent)
IN Schmitz Albert (Basel, CH); Marki Walter (M
PA Ciba Geigy Corporation (Ardsley NY)
PI US 5210028 A 19930511
AI US 1990-616470 19901121
AB A process for the preparation of a recombinant IGF-II (rIGF-II)
without a covalently attached foreign protein moiety and without N-
terminal attached methionine or a derivative of methionine or of a
salt of said IGF-II, rIGF-II produced by said method, . . . .
ECLM US5210028 A: We claim:1. A process for the production of a
recombinant IGF-II without a covalently attached foreign protein
moiety and without N-terminal attached methionine or a derivative of
methionine, said process comprising:a) transforming a suitable strain
of E. coli, said strain being a lon.sup.- and htpR.sup.- double
mutant, with a hybrid vector comprising an expression cassette
consisting of the following elements in the 5' to 3' direction, said
elements which are operably linked: an inducible promoter, a
ribosomal binding site, and the code
linked in proper reading frame to a
IGF-II having the amino acid sequenc
SSO PROTEIN; USPTO; GRANTED
ORGN Human IGF II
SQL 67
SEQ
1 ayrpsetlcg gelvdtlqfv cgdrgyfysr pasrvsrrsr giveeccfrs
51 cdlalletyc atpakse
```

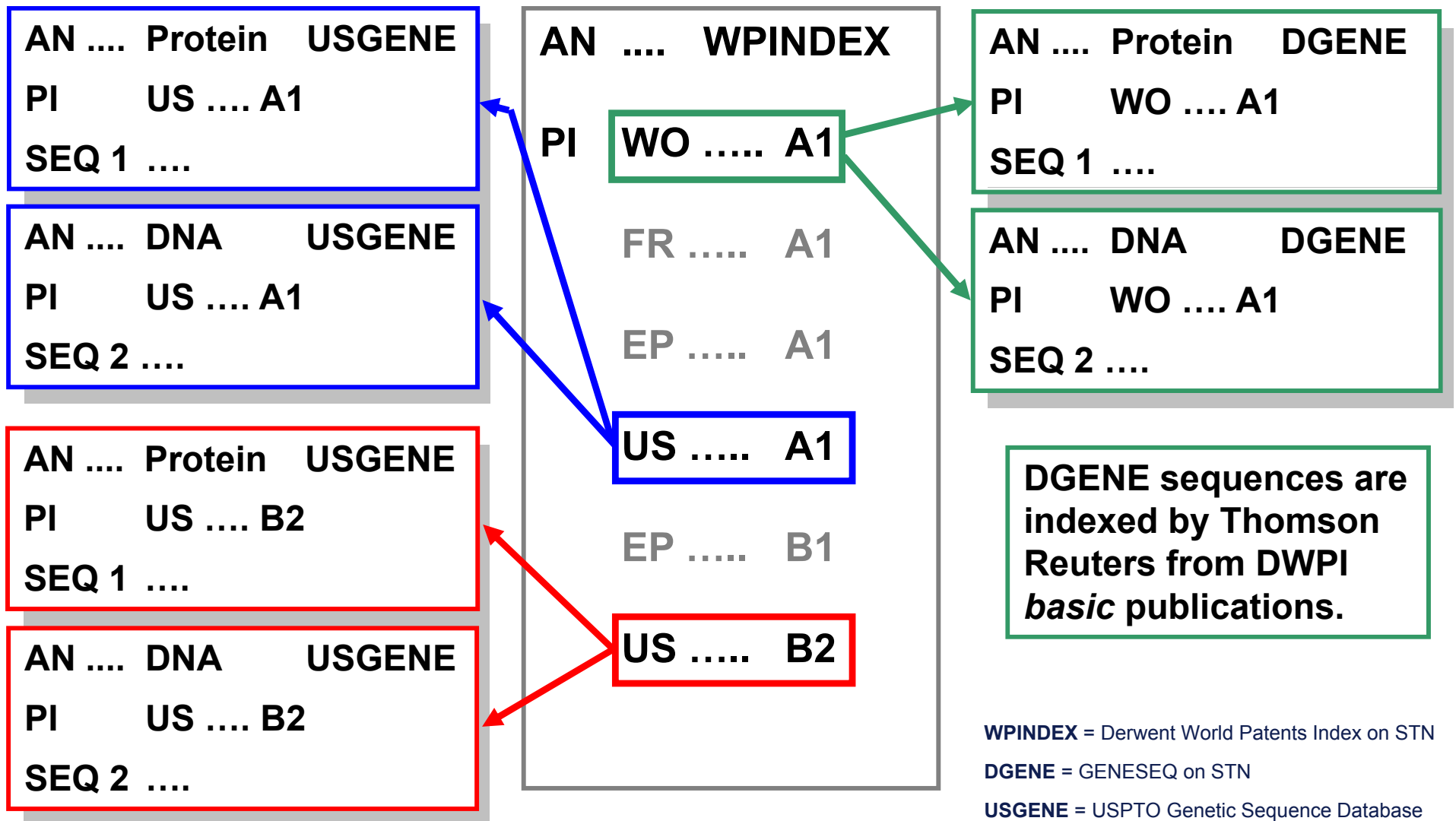
AN 5210028.1 is SEQ ID
NO: 1 from US5210028.

Compare the STN standardized
USGENE record to the original
data source on the previous slide.

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



Sequence listing variation often occurs between PCT and U.S. granted patent stage

```

L1 ANSWER 1 OF 1 WPINDEX COPYRIGHT 2009 THOMSON REUTERS 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 19
FI 9505224 A 19
EP 698101 A1 19
JP 09500009 W 19
AU 695259 B 19
EP 698101 B1 20
DE 69434116 E 20
US 20050032047 A1 20
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 [WO9425602](#) in DGENE
- 50 sequences from [US20050032047](#) in USGENE
- 58 sequences from [US6881821](#) in USGENE

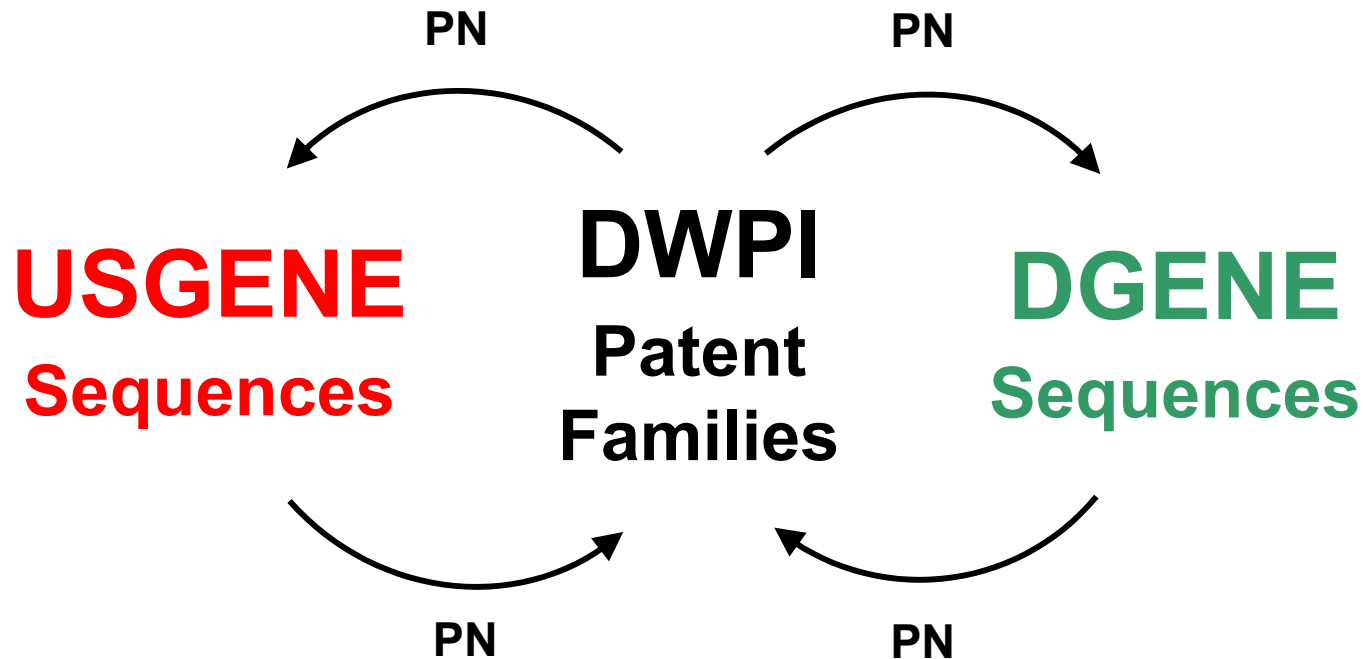
USGENE covers a comprehensive variety of USPTO patent publication types

<u>PK</u>	<u>Patent Kind covered in USGENE (field /PK)</u>
USA1	Published patent application
USA2	Republished patent application
USA9	Corrected published patent application
USA	Granted patent (until 2000)
USB1	Granted patent without pre-grant publication (2001 onwards)
USB2	Granted patent with pre-grant publication (2001 onwards)
USE	Reissued patent
USH1	Statutory Invention Registration
USP1	Published plant patent application
USP2	Granted plant patent without pre-grant publication
USP3	Granted plant patent with pre-grant publication
WOA	WIPO/PCT published patent application (parent case data)

Agenda

- STN sequence searchable databases
- 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
- Comparisons and conclusions

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



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

- 1) Ensure that preferred file default display formats are set
- 2) UPLOAD the sequence query via STN Express[®] (**L1**)
- 3) *USGENE*: BLAST (**L2**); SORT SCORE D (**L3**)
Option: review and isolate chosen hits with SORT AN 1-x (**L4**)
- 4) *DGENE*: BLAST (**L5**); SORT SCORE D (**L6**);
Option: review and isolate chosen 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 final results in file default format: D L14 TOTAL

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

Search Question:

Find relevant patent references for *Eukaryotic translation elongation factor 1 gamma* (NP_001395):

MAAGTLYTYPENWRAFKALIAAQYSGAQVRVLSAPPHFHFGQTNRTPEFLRKFPAGKVPAFEG
DDGFCVFESENAIAYYVSNEELRGSTPEAAAQVVQWVSFADSDIVPPASTWVFPTLGIMHHNKQ
ATENAKEEVRRILGLLDAYLKTRTFLVGERVTLADITVVCTLLWLYKQVLEPSFRQAFPNTR
WFLTCINQPQFRAVLGEVKLCEKMAQFDAKKFAETQPKKDTPRKEKGSREEKQKPQAERKEEK
KAAAPAPEEEMDECEQALAAEPKAKDPFAHLPKSTFVLDEFKRKYSNEDTLSVALPYFWEHFD
KDGWLSLWYSEYRFPEELTQTFMSCNLI TGMFQRLDKLRKNAFASVILFGTNNSSSISGVWVFR
GQELAFPLSPDWQVDYESYTWKLDPGSEETQTLVREYFSWEGAFQHVGKAFNQGKIFK

(Search conducted on October 20th, 2008.)

1) Ensure that preferred user-defined file default display formats are set

=> FILE STNGUIDE

=> SET FORMAT .MYUSGENEALIGN TRI **ORGN SEQN SEQC** SCORE ALIGN

=> SET FORMAT .MYDGENEALIGN TRI OS SCORE ALIGN

=> FILE USGENE; SET DFORMAT .MYUSGENEALIGN

=> FILE DGENE; SET DFORMAT .MYDGENEALIGN

=> FILE WPINDEX; SET DFORMAT BIB

=> D FORMAT

Review all user-defined formats with D FORMAT.

ORGN = Organism Name
SEQN = SEQ ID Number
SEQC = Sequence Count

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

USER-DEFINED FORMAT DEFINITION

DEFAULT FORMAT
FOR FILE

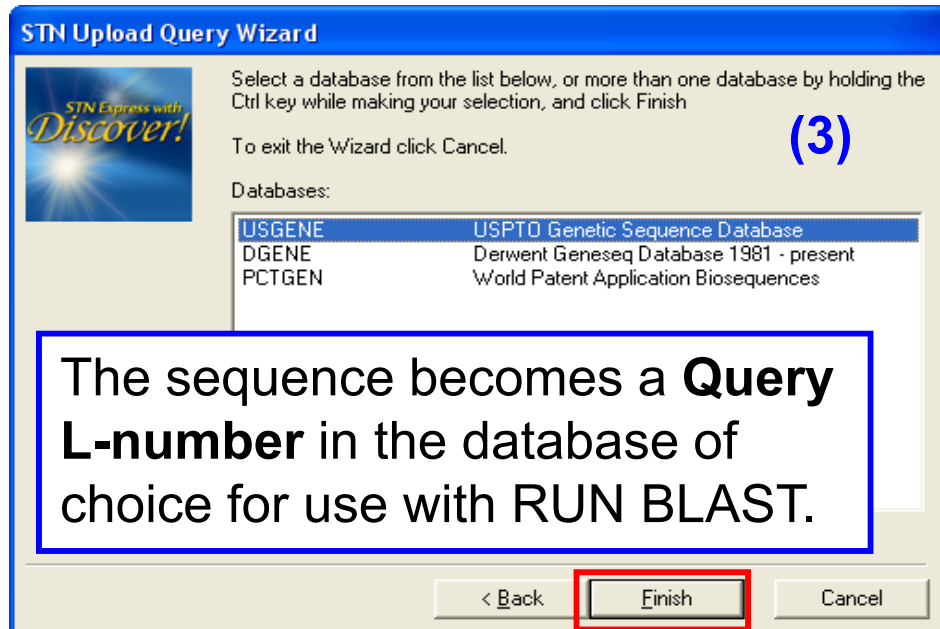
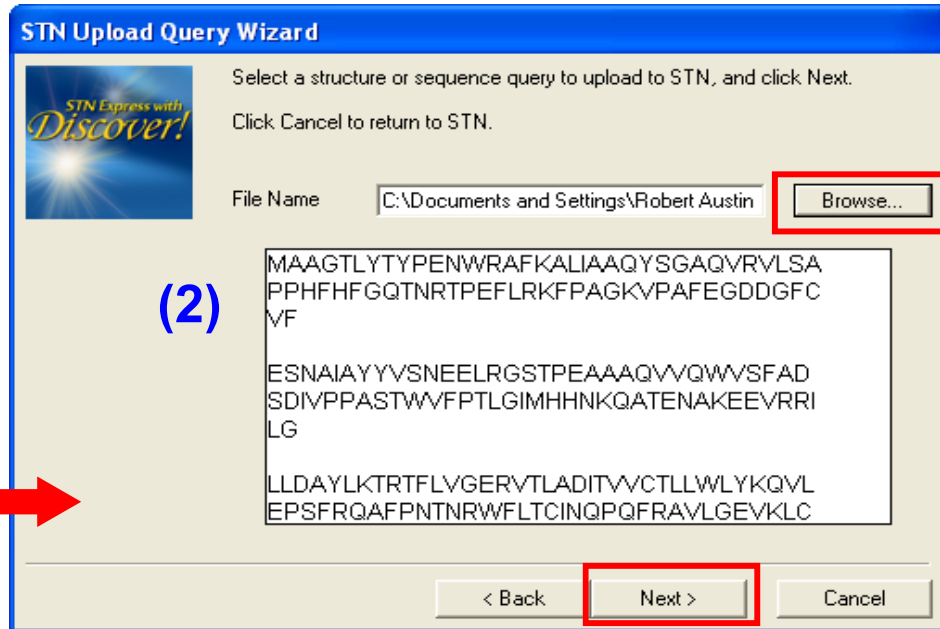
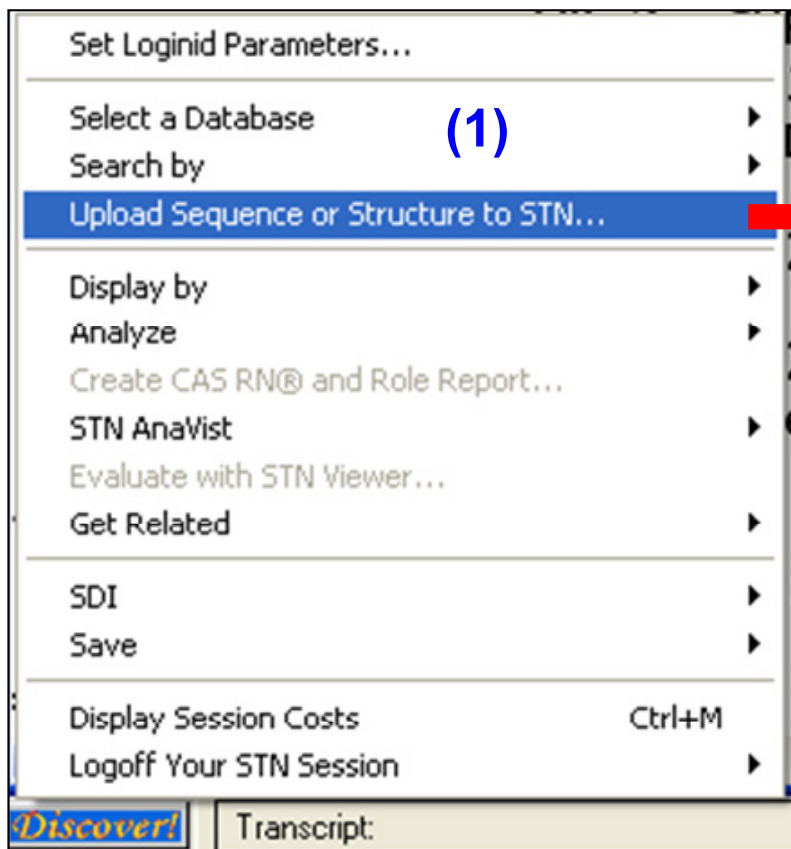
.MYDGENEALIGN TRI OS SCORE ALIGN

DGENE

.MYUSGENEALIGN TRI ORGN SEQN SEQC SCORE ALIGN USGENE

2) UPLOAD the sequence query via STN Express

- (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 MAAGTLYTYPENWRAFKALIAAQYSGAQVRVLSAPPHFHFHQTNRTPEFLRKFPAGKVP
AFEGDDGFCVFESNAIAYYVSNEELRGSTPEAAAQVVQWVSFADSDIVPPASTWVFPTL
GIMHHNKQATENAKEEVRRI LGLLDAYLKTRTFLVGERVTLADITVVCTLLWLYKQVLE
PSFRQAFPNTRWFLTCINQPQFRAVLGEVVKLCEKMAQFDAKKFAETQPKKDTPRKEKG
SREEKQKPQAERKEEKKAAAPAPEEEMDECEQALAAEPKAKDPFAHLPKSTFVLDEFKR
KYSNEDTLSVALPYFWEHFDKDGWSLWYSEYRFPEELTQTFMSCNLIITGMFORLDKLRK
NAFASV
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 11:53:09 ON 22
COPYRIGHT (C) 2008 SEQUENCEBASE CORP

FILE LAST UPDATED: 17 OCT 2008 <20081017/UP>
MOST RECENT PUBLICATION DATE: 16 OCT 2008 <20081016/PD>

FILE COVERS 1982 TO DATE

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

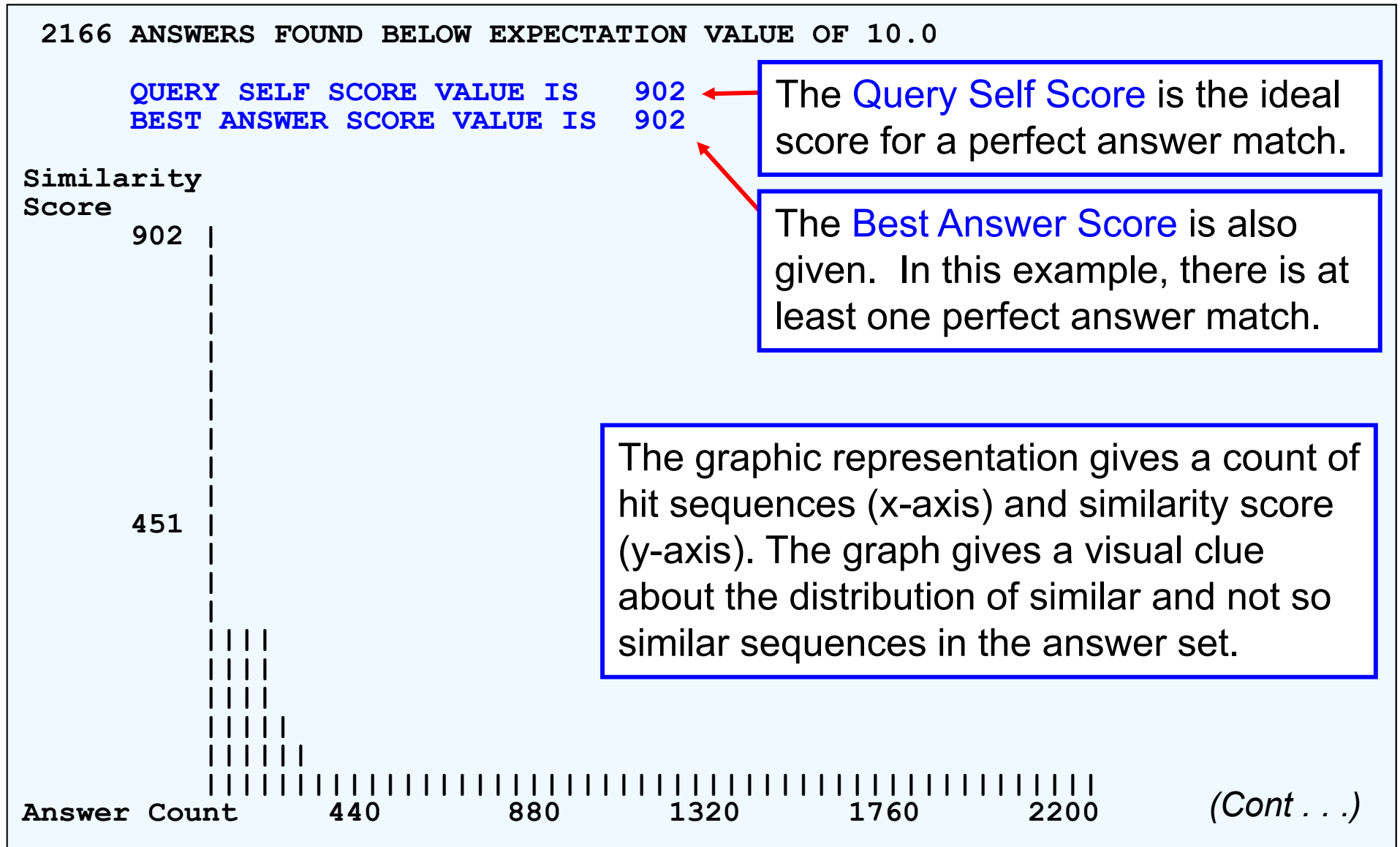
=> **RUN BLAST L1 /SQP -F F**

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.)

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

```
ENTER (ALL) OR ? : 50%
```

In this example, 50% of the Query Self Score is used to select out the best results (L2).

```
L2      RUN STATEMENT CREATED
```

```
L2      19 MAAGTLYTYPENWRAFKALIAAQ
          RKFPAGKVPAPAFEGDDGFCVFESNAIAYIVSNEELRGSTPEAAAQVVQWVS
          FADSDIVPPASTWVFPPTLGMHNNKQATENAKEEVRRILGLLDAYLKTRT
          FLVGERVTLADITVVCTLLWLYKQVLEPSFRQAFPNTNRWFLTCINQPQF
          RAVLGEVKLCEKMAQFDAKKFAETQPKKDTPRKEKGSREEKQKPQAERKE
          EKKAAPAPEEEMDECEQALAAEPKAKDPFAHLPKSTFVLDEFKRKYSNE
          DTLSVALPYFWEHFDDKDGWSLWYSEYRFPEELTQTFMSCNLITGMFQRLD
          KLRKNAFASVILFGTNNSSSISGVWVFRGQELAFPLSPDWQVDYESYTW
          KLDPGSEETQTLVREYFSWEGAFQHVKGAFNQGKIFK/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      19 SOR L2 SCORE D
```

Use SORT SCORE D to sort by descending BLAST score (L3).

3) RUN the USGENE BLAST search (cont.)

=> D 1-19

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

L3 ANSWER 1 OF 19 USGENE COPYRI
TI Biomarkers of cyclin-dependent kinase modulation
(PublishedApplication)

MTY Protein

SQL 437

ORGN Homo Sapiens

SEQN 132

SEQC 2786

Note: The customized USGENE default display format includes the SEQ ID Number (SEQN) and the Sequence Count (SEQC).

SCORE 902 100% of query self score 902

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

Query: 61 FEGDDGFCVFESNAIAYYSNEELRGSTPEAAAQVVQVVSFADSDIVPPASTWVFPTLGI

. . . .

3) RUN the USGENE BLAST search (cont.)

=> D SCORE 1-19

Another way to review quickly is by BLAST SCORE.

L3 ANSWER 1 OF 19 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
SCORE 902 100% of query self score 902

. . . .

L3 ANSWER 13 OF 19 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
SCORE 788 87% of query self score 902

. . . .

The SCORE display field includes the percentage of the Query Self Score.

L3 ANSWER 16 OF 19 USGENE COPYR
SCORE 656 72% of query self score 902

L3 ANSWER 17 OF 19 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
SCORE 656 72% of query self score 902

L3 ANSWER 18 OF 19 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
SCORE 495 54% of query self score 902

. . . .

=> SOR AN 1-17

PROCESSING COMPLETED FOR L3

L4 17 SOR L3 1-17 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

. . . .

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

ENTER (ALL) OR ? : 50%

L5 RUN STATEMENT CREATED

L5 23 MAAGTLYTYPENWRAFKALIAAC
RKFPAGKVPAFEGDDGFCVFESN

. . . .

KLRKNAFASVILFGTNNSSISGVWVFRGQELAFPLSPDWQVDYESYTW
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 23 SOR L5 SCORE D

The same query (L1) can be used again
to repeat the BLAST search in DGENE.

In this example, 50% of the
Query Self Score is used to
select out the best results (L5).

Use SORT SCORE D to sort by
descending BLAST score (L6).

4) RUN the DGENE BLAST search (cont.)

=> D 1-23

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

```
L6 ANSWER 1 OF 23 DGENE COPYRI
TI Human cancer suppressor gene mig 20 and protein encoded therein which
are useful for diagnosis, prevention and treatment of cancers,
particularly ovarian cancer, leukemia, liver cancer and lung cancer.
DESC Human growth-inhibiting gene 35 (GIG35) protein, SEQ ID NO:50.
KW diagnostic test; gene therapy; tumor suppressor; breast tumor;
endocrine-gen.; gynecological; brain tumor; cytostatic;
neuroprotective; muscle tumor; muscular-gen.; large intestine tumor;
gastrointestinal-gen.; t
SQL 437
OS 2007-250439 [25]
SCORE 902 100% of query self score 902
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
. . . .
```

Other Source (OS) = the Accession Number from the corresponding DWPI family record.

4) RUN the DGENE BLAST search (cont.)

=> D SCORE 1-23 ← Another way to review quickly is by BLAST SCORE.

L6 ANSWER 1 OF 23 DGENE COPYRIGHT 2008 THE THOMSON CORP on STN
SCORE 902 100% of query self score 902

. . . .

L6 ANSWER 17 OF 23 DGENE COPYRIGHT 2008 THOMSON REUTERS on STN
SCORE 880 97% of query self score 902

. . . .

L6 ANSWER 20 OF 23 DGENE COPYR
SCORE 656 72% of query self score 902

The SCORE display field includes the percentage of the Query Self Score.

L6 ANSWER 21 OF 23 DGENE COPYRIGHT 2008 THOMSON REUTERS on STN
SCORE 656 72% of query self score 902

L6 ANSWER 22 OF 23 DGENE COPYRIGHT 2008 THOMSON REUTERS on STN
SCORE 495 54% of query self score 902

. . . .

=> SOR AN 1-21

PROCESSING COMPLETED FOR L6

L7 21 SOR L6 1-21 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

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

=> TRA L4 PN; TRA L7 PN

L8 TRANSFER L4 1- PN : 17 USGENE sequence hits (L4)
L9 14 L8 ← found 14 DWPI records (L9).

L10 TRANSFER L7 1- PN : 21 DGENE sequence hits (L7)
L11 18 L10 ← found 18 DWPI records (L11).

=> S L9 OR L11

L12 23 L9 OR L11

Total DWPI records is 23 (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

L4 = USGENE selected BLAST hits.

L7 = DGENE selected BLAST hits.

L12 = corresponding DWPI records.

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

FILE 'USGENE' ENTERED AT 11:57:53 ON 22 OCT 2008
COPYRIGHT (C) 2008 SEQUENCEBASE CORP

FILE 'DGENE' ENTERED AT 11:57:53 ON 22 OCT 2008
COPYRIGHT (C) 2008 THOMSON REUTERS

FILE 'WPINDEX' ENTERED AT 11:57:53 ON 22 OCT 2008
COPYRIGHT (C) 2008 THOMSON REUTERS

PROCESSING COMPLETED FOR L4

PROCESSING COMPLETED FOR L7

PROCESSING COMPLETED FOR L12

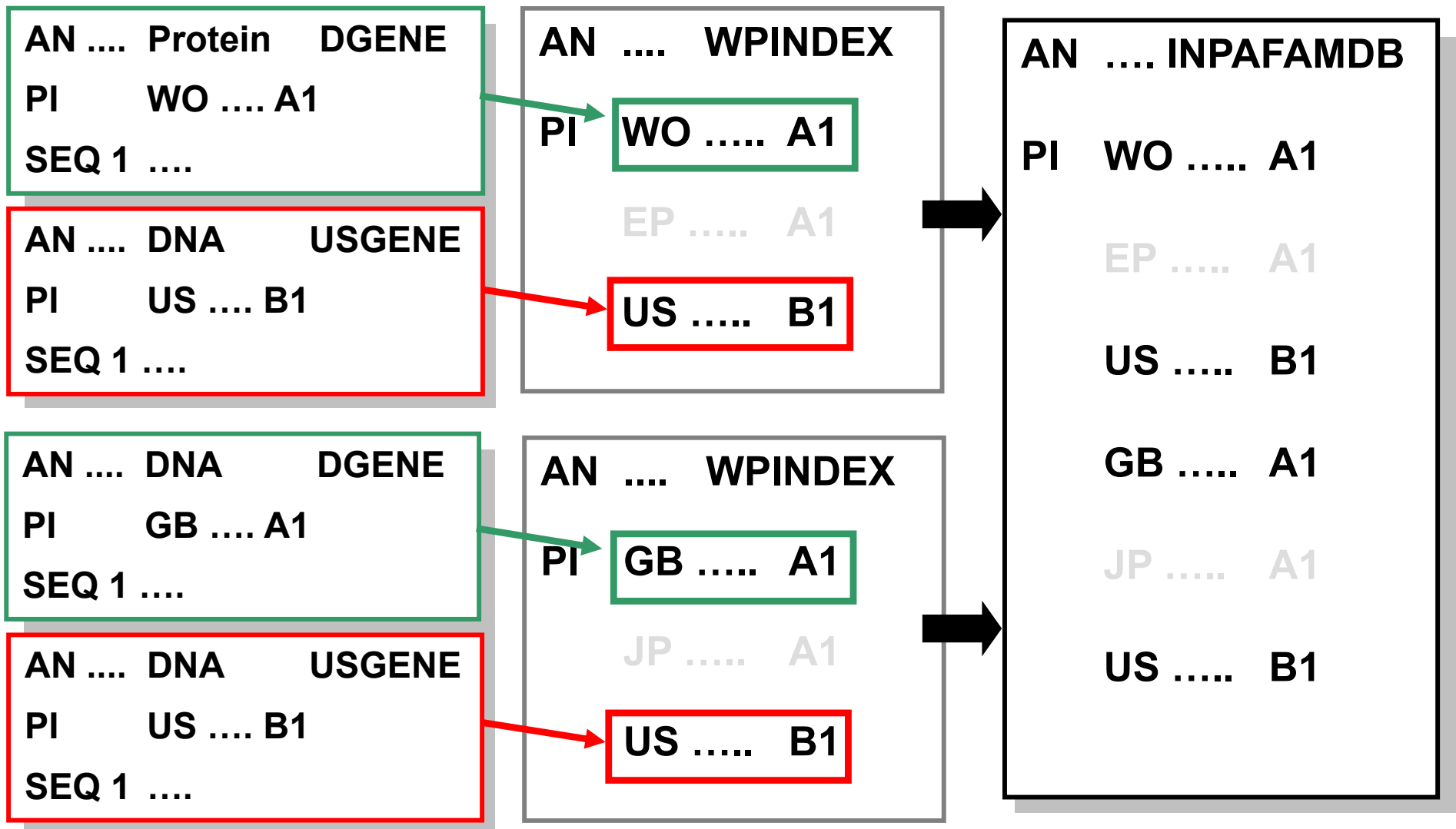
L13 61 DUP IDE L4 L7 L12 (INCLUDES 0 SETS OF DUPLICATES)

ANSWERS '1-17' FROM FILE USGENE

ANSWERS '18-38' FROM FILE DGENE

ANSWERS '39-61' FROM FILE WPINDEX

Note that a FSORT patent family may be represented by one or more DWPI records



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

=> FSORT L13

. . . .

L14 61 FSO L13

21 Multi-record Families	Answers 1-61
Family 1	Answers 1-5
Family 2	Answers 6-8
Family 3	Answers 9-11
Family 4	Answers 12-14
Family 5	Answers 15-22
Family 6	Answers 23-25
Family 7	Answers 26-27
Family 8	Answers 28-29
Family 9	Answers 30-31
Family 10	Answers 32-34
Family 11	Answers 35-37
Family 12	Answers 38-39
Family 13	Answers 40-42
Family 14	Answers 43-45
Family 15	Answers 46-48
Family 16	Answers 49-50
Family 17	Answers 51-52
Family 18	Answers 53-54
Family 19	Answers 55-57
Family 20	Answers 58-59
Family 21	Answers 60-61
0 Individual Records	
0 Non-patent Records	

The 23 DWPI records (L12), 17 USGENE sequence hits and 21 DGENE sequence hits belong to 21 FSORT families (L14).

7) Display results using the customized file default display formats (see slide 24)

=> D L14 TOTAL

This displays all the FSORT grouped records (TOTAL) for the final results (L14) using file default formats.

L14 ANSWER 6 OF 61 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN FAMILY2

TI Biomarkers of cyclin-dependent kinase modulation
(PublishedApplication)

MTY Protein

SQL 437

ORGN Homo Sapiens

SEQN 132

SEQC 2786

SCORE 902 100% of query self score 902

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

.

USGENE hit sequence display(s).

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

```
L14 ANSWER 7 OF 61 DGENE COPYRIGHT 2008 THOMSON REUTERS on STN FAMILY2
AN ADX05567 protein DGENE
TI Biomarkers useful for predicting or determining the response of a
mammal to a cancer treatment comprising administration of a modulator
of cyclin-dependent kinase activity.
DESC Cyclin-dependent kinase modulation biomarker SEQ ID NO 132.
KW cytostatic; cyclin-dependent kinase; cdk; biomarker.
SQL 437
OS 2005-163068 [17]
SCORE 902 100% of query self score 902
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
        . . . .
```

DGENE hit sequence display(s).

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

```
L14 ANSWER 8 OF 61 WPINDEX COPYRIGHT 2008 THOMSON REUTERS on STN FAMILY2
AN 2005-163068 [17] WPINDEX
TI Biomarkers useful for predicting or determining the response of a
   mammal to a cancer treatment comprising administration of a modulator
   of cyclin-dependent kinase activity
DC B03; B04; D16; S03
IN JACKSON D G; LI M; RUPNOW B A; WEBSTER K; WONG T
   RUPNOW B; WEBSTER K; WONG T
PA (BRIM-C) BRISTOL-MYERS SQUIBB CO; (JACK-I) JACKSON D G; (LIMM-I) LI M;
   (RUPN-I) RUPNOW B A; (WEBS-I) WEBSTER K R; (WONG-I) WONG T W
PIA WO 2005012875 A2 20050210 (200517)* EN 141[9]
   EP 1656542 A2 20060517 (200634) EN
   AU 2004262369 A1 20050210 (200660) EN
   JP 2007507204 W 20070329 (200725) JA 138
   US 20070105114 A1 20070510 (200732) EN
ADT WO 2005012875 A2 WO 2004-US24424 20040729; AU 2004262369 A1 AU
   2004-262369 20040729; EP 1656542 A2 EP 2004-779471 20040729; EP
   1656542 A2 WO 2004-US24424 20040729; JP 2007507204 W WO 2004-US24424
   20040729; JP 2007507204 W JP 20070329 (200725) JA 138
   Provisional US 2003-490890P 20030729
   US24424 20040729; US 20070105114 A1 20070510 (200732) EN
FDT EP 1656542 A2 Based on WO 2005012875 A; JP 2007507204 W
PRAI US 2003-490890P 20030729
     US 2006-567867 20060818
```

DWPI patent family display.

Tip: These arrows indicate the family member(s) retrieved in the USGENE and/or DGENE BLAST searches.

Agenda

- STN sequence searchable databases
- 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**
- Comparisons and conclusions

Summary of results for *Eukaryotic translation elongation factor 1 gamma* (NP_001395)

	SEQs > 70%	PNs	DWPI Records	FSORT Families
USGENE	17	16	14	13
DGENE	21	18	18	17
Overlap	-	4	9	9
Total Unique	-	30	23	21

Example: USGENE unique retrieval

```
L14 ANSWER 26 OF 61 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN FAMILY7
TI Tissue-and serum-derived glycoproteins and methods of their use
(PublishedApplication)
MTY Protein
SQL 437
ORGN Homo Sapiens
SEQN 10979
SEQC 14918
SCORE 902 100% of query self score 902
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
Query: 61 FEGDDGFCVFESNAIAYYSNEELRGSTPEAAAQVVQWVSFADSDIVPPASTWVFPTLGI
FEGDDGFCVFESNAIAYYSNEELRGSTPEAAAQVVQWVSFADSDIVPPASTWVFPTLGI
Sbjct: 61 FEGDDGFCVFESNAIAYYSNEELRGSTPEAAAQVVQWVSFADSDIVPPASTWVFPTLGI
. . . .
```

This USGENE hit sequence uniquely retrieved the DWPI record on the following slide (as of Oct 20th, 2008).

Example: USGENE unique retrieval (cont.)

L14 ANSWER 27 OF 61 WPINDEX COPYRIGHT 2008 THOMSON REUTERS on STN **FAMILY7**
AN 2007-560359 [54] WPINDEX
TI New diagnostic panel comprising detection reagents that are specific
for tissue-derived serum glycoprotein, useful in defining a disease-
associated tissue-derived blood fingerprint or monitoring response to
therapy in a subject
DC A89; B04; D16; K08; S03
IN AEBERSOLD R H; ZHANG H; AEBERSOLD R H
PA (SYST-N) INST SYSTEMS BIOLOGY
CYC 117
PIA WO 2007047796 A2 20070426 (200754) EN
US 20070099251 A1 20070503 (200754) EN <--
EP 1938104 A2 20080702 (200845) EN
AU 2006304605 A1 20070426 (200858) EN
ADT WO 2007047796 A2 WO 2006-US40784 20061017; **US 20070099251 A1**
Provisional US 2005-728044P 20051017; EP 1938104 A2 EP 2006-836381
20061017; **US 20070099251 A1** US 2006-582861 20061017; AU 2006304605 A1
AU 2006-304605 20061017; EP 1938104 A2 PCT Application WO 2006-US40784
20061017
FDT EP 1938104 A2 Based on WO 2007047796 A; AU 2006304605 A1 Based on
WO 2007047796 A
PRAI US 2005-728044P 20051017
US 2006-582861 20061017

This relevant DWPI record was uniquely retrieved via a USGENE BLAST search (as of Oct 20th, 2008).

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

Search Question:

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

VRSSSRTPSDKPVAVHVVANPQAEGQLQWLNRRANALLANGVELRDNQLVVPSEGLYLIYSQVLF
KGQGCPSTHVLLTHTISRIAVSYQTKVNLLSAIKSPCQRETPRGAEAKPWYEPIYLGGVFQLEK
GDRLSAEINRPDYLDFAESGQVYFGI IAL

(Search conducted on October 22nd, 2008.)

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

	SEQs > 80%	PNs	DWPI Records	FSORT Families
USGENE	753	337	216	120
DGENE	735	326	326	225
Overlap	-	15	171	88
Total Unique	-	648	371	257

Example: USGENE unique retrieval

L14 ANSWER 541 OF 1859 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
FAMILY 20

TI Fusion proteins comprising gp39 and CD8 (Patent)

MTY protein

SQL 157

ORGN Not provided

SEQN 5

SEQC 9

SCORE 313 99% of query self score 316

BLASTALIGN

Query = 157 letters

Length = 157

Score = 313 bits (803), Expect = 6e-91

Identities = 155/157 (98%), Positives = 156/157 (98%)

Query: 1 VRSSSRTPSDKPVAVHVVANPQAEGLQWLNRRANALLANGVELRDNQLVVPSEGLYLIYS
VRSSSRTPSDKPVAVHVVANPQAEGLQWLNRRANALLANGVELRDNQLVVPSEGLYLIYS

Sbjct: 1 VRSSSRTPSDKPVAVHVVANPQAEGLQWLNRRANALLANGVELRDNQLVVPSEGLYLIYS

Query: 61 QVLFKGQGCPSTHVLLTHTISRIAVSYQTKVNLLSAIKSPCQRET PRGAEAKPWYEPIYL
QVLFKGQGCPSTHVLLTHTISRIAVSYQTKVNLLSAIKSPCQRET P GAEAKPWYEPIY+

Sbjct: 61 QVLFKGQGCPSTHVLLTHTISRIAVSYQTKVNLLSAIKSPCQRETPEGAEAKPWYEPIYI

Query: 121 GGVFQLEKGDRLSAEINRPDYLDFAESGQVYFGIIAL 157
GGVFQLEKGDRLSAEINRPDYLDFAESGQVYFGIIAL

Sbjct: 121 GGVFQLEKGDRLSAEINRPDYLDFAESGQVYFGIIAL 157

This USGENE hit sequence uniquely retrieved the DWPI record on the following slide (as of Oct 22nd, 2008).

Example: USGENE unique retrieval (cont.)

```
L14 ANSWER 543 OF 1859 WPINDEX COPYRIGHT 2008 THOMSON REUTERS on STN
  FAMILY 20
AN 1994-076264 [10] WPIX
TI New nucleic acid encoding human gp39 T cell antigen - which is a
  ligand for the CD40 receptor, causing proliferation and
  differentiation of B cells and some cancer cells
DC B04; D16
IN ARUFFO A; ARUFFO A A; ARUFFO A E; HOLLEBAUGH D; HOLLENBAUGH D;
  LEDBETTER J A
PA (BRIM-C) BRISTOL-MYERS SQUIBB CO
PIA EP 585943 A2 19940309 (199410)* EN 39[9]
  AU 9346120 A 19940310 (199415) EN
  NO 9303126 A 19940307 (199416) NO
  CA 2105552 A 19940305 (199420) EN
  FI 9303862 A 19940305 (199420) FI
  ZA 9306491 A 19940525 (1994
  JP 06315383 A 19941115 (1995
  EP 585943 A3 19940706 (1995
  HU 69977 T 19950928 (1995
  NZ 248569 A 19951026 (1996
  US 5540926 A 19960730 (199636) EN 30[9] <--
  AU 677788 B 19970508 (199727) EN
  EP 585943 B1 19980211 (199811) EN 40[8]
  DE 69316948 E 19980319 (199817) DE
  ES 2113980 T3 19980516 (199826) ES
```

This relevant DWPI record was uniquely retrieved via a USGENE BLAST search (on Oct 22nd, 2008).

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

Search Question:

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

```
GCCAGCCCCCTGATGGGGGCGACACTCCACCATGAATCACTCCCCTGTGAGGAACTACTGTCTT  
CACGCAGAAAGCGTCTAGCCATGGCGTTAGTATGAGTGTCGTGCAGCCTCCAGGACCCCCCTC  
CCGGGAGAGCCATAGTGGTCTGCGGAACCGGTGAGTACACCGGAATTGCCAGGACGACCGGGTC  
CTTTCTTGGATCAACCCGCTCAATGCCTGGAGATTTGGGCGTGCCCCCGCAAGACTGCTAGCCG  
AGTAGTGTTGGGTCGCGAAAGGCCTTGTGGTACTGCCTGATAGGGTGCTTGCAGAGTGCCCCGGG  
AGGTCTCGTAGACCGTGCACC
```

(Search conducted on October 22nd, 2008.)

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

	SEQs > 80%	PNs	DWPI Records	FSORT Families
USGENE	590	204	107	82
DGENE	444	191	191	153
Overlap	-	14	95	75
Total Unique	-	381	203	160

Example: USGENE unique retrieval

```
L14  ANSWER 38 OF 1237  USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
      FAMILY3
TI   Fusion polypeptide having the C protein and E1 protein of hepatitis C
      virus (Patent)
MTY  DNA
SQL  9379
ORGN Unknown
SEQN 1
SEQC 2
SCORE 632          93% of query self score 676
BLASTALIGN
      Query  = 341 letters
      Length = 9379
      Score  = 632 bits (319), Expect = 0.0
      Identities = 319/319 (100%)
      Strand = Plus / Plus

Query: 23  cactccaccatgaatcactcccctgtgaggaactactgtcttcacgcagaaagcgtctag
          |||
Sbjct: 1   cactccaccatgaatcactcccctgtgaggaactactgtcttcacgcagaaagcgtctag

Query: 83  ccatggcgtagtatgagtgtcgtgcagcctccaggacccccctcccgggagagccata
          |||
Sbjct: 61  ccatggcgtagtatgagtgtcgtgcagcctccaggacccccctcccgggagagccata
. . . . .
```

This is one of three USGENE hit sequences that uniquely retrieved the DWPI record on the following slide (as of Oct 22nd, 2008).

Example: USGENE unique retrieval (cont.)

L14 ANSWER 39 OF 1237 WPINDEX COPYRIGHT 2008 THOMSON REUTERS on STN
FAMILY3
AN 1998-495550 [42] WPIX
TI Composition for immunotherapy of hepatitis C virus infection -
comprises viral capsid and envelope polypeptides, optionally as
fusion, unable to regulate genes or corresponding DNA
DC B04; D16
IN BARBAN V
PA (AVET-C) AVENTIS PASTEUR; (INMR-C) PASTEUR MERIEUX SERUMS & VACCINS
SA; (SNFI-C) SANOFI PASTEUR
PIA WO 9839030 A1 19980911 (19980911) EN
FR 2760367 A1 19980911 (19980911) EN
AU 9868398 A 19980922 (19990922) EN
EP 1017418 A1 20000712 (20000712) EN
NZ 337138 A 20010427 (200128) EN
US 6284249 B1 20010904 (200154) EN <--
JP 2001513807 W 20010904 (200165) JA 29
US 20020034734 A1 20020321 (200224) EN <--
AU 745442 B 20020321 (200233) EN
US 6538123 B2 20030325 (200325) EN <--
EP 1017418 B1 20080514 (200833) FR
DE 69839489 E 20080626 (200844) DE

This relevant DWPI record was uniquely retrieved via a USGENE BLAST search (on Oct 22nd, 2008).











Agenda

- STN sequence searchable databases
- 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
- **Comparisons and conclusions**

Comparison of USPTO coverage in sequence databases

	Update Frequency	Typical Timeliness	Backfile coverage
USGENE	Weekly	3 days	1982 -
DGENE (DWPI basics)	Biweekly	65 days	1981 -
REGISTRY (CAplus basics)	Daily	27 days	1957 -
NCBI/EMBL	Daily	1-3 months	1982 -

Comparison of USPTO coverage in sequence databases (cont.)

	USPTO Pub Apps	USPTO Patents	USPTO claims text	Editorial value-add
USGENE				
DGENE (DWPI basics)				
REGISTRY (CAplus basics)				
NCBI/EMBL				

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
- USGENE and DGENE often find unique relevant hits and should always be used in combination
- A “best-practice” approach to multiframe searching uses DWPI to more accurately group together sequence hits from USGENE and DGENE
- The “best-practice” approach can be extended to incorporate PCTGEN and CAplus/REGISTRY

Resources for sequence searching on STN

- *Sequence Searching on STN* modular workshop
www.stn-international.com/sequence_searching.html
 - Sequence Code Match (SCM) searching
 - DGENE, USGENE, PCTGEN content and searching
 - CAS REGISTRY and REGISTRY BLAST
 - Multifile searching using USGENE and DGENE
- USGENE resources, reference materials and FAQ
www.sequencebase.com
- More on the new percent option for BLAST & GETSIM
www.stn-international.com/New_sequence_search.html

STN[®]

Multifile Patent Sequence
Searching on STN[®]

Robert Austin – FIZ Karlsruhe