

STN[®]

Effective patent sequence searching
on STN[®]

Jim Brown – FIZ Karlsruhe

Agenda

- DGENE, PCTGEN, USGENE[®] database content
- The 7 basic steps of BLAST[®]
- BLAST and Patent Family SORT (FSORT)
- Post-processing BLAST search results
- Similarity searching GETSIM (FASTA)
- Offline BATCH search mode
- Sequence Code Match searching (GETSEQ)
- Multifile patent sequence search example

STN[®] sequence searchable databases

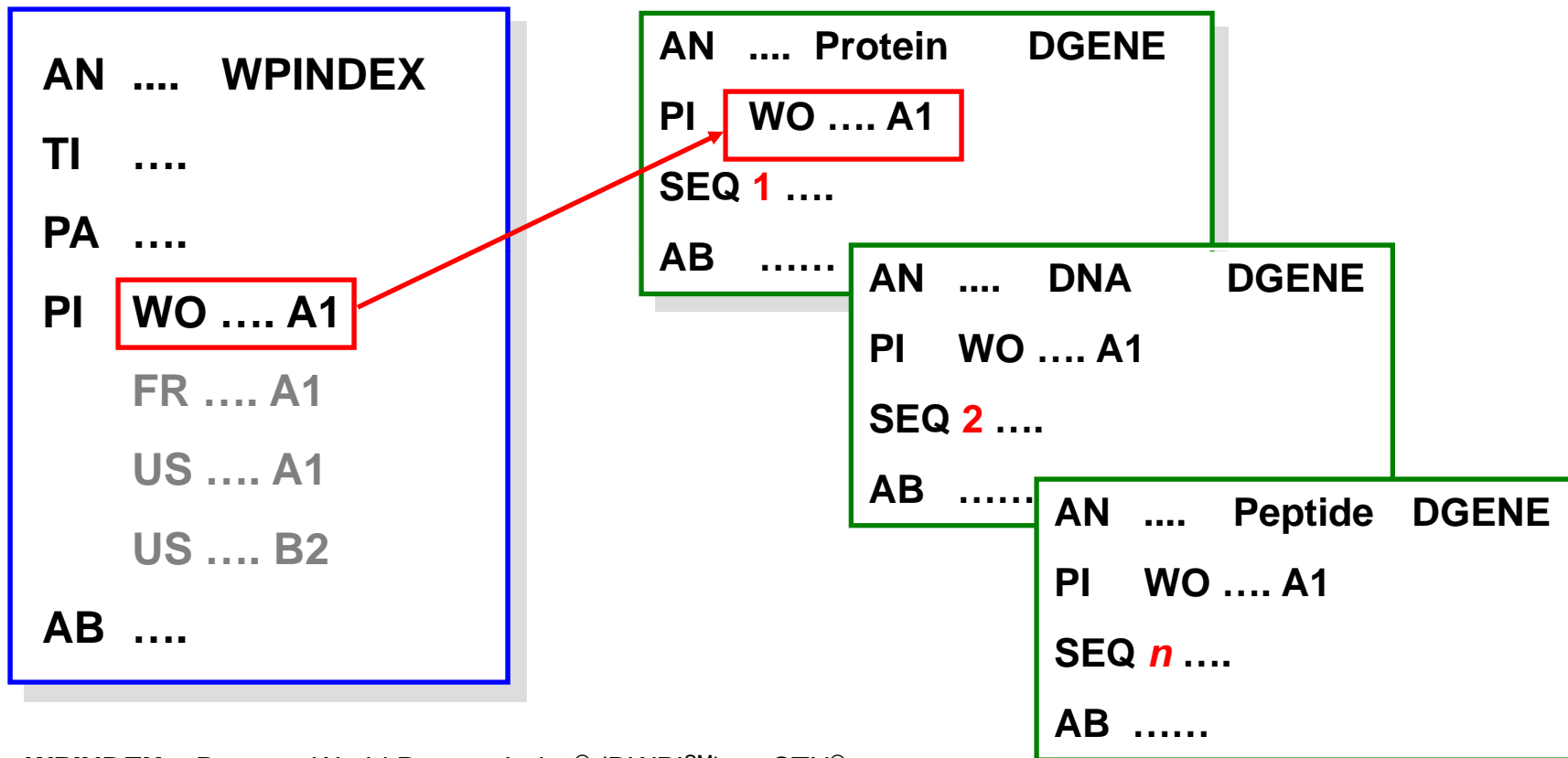
- **DGENE**
 - Thomson Reuters GENESEQ[™]
 - Value-added patent sequence data from around the globe
- **PCTGEN**
 - WIPO/PCT Patent Application Biosequences
 - The complete collection of e-published sequences from WIPO
- **USGENE**
 - The USPTO Genetic Sequence Database
 - A new and unique access point to USPTO sequence data
- **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 41 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

Relationship between DWPI patent family and DGENE sequence database



WPINDEX = Derwent World Patents Index® (DWPISM) on STN®

DGENE = GENESEQTM on STN®

Biosequences in DGENE

- Polypeptide sequences of >3 peptide residues
- Nucleic acid sequences of >9 nucleotides
- PCR primer and probe sequences of any length
- Mutated sequences derivable by Indexers
- Clear and concise sequence description and descriptive keyword Indexing
- Feature tables for annotations or modifications of the sequence
- Patent sequence location (claim, example, etc.)

DGENE records also provide

- Enhanced patent titles from DWPI
- Enhanced English abstracts per sequence written by Thomson Reuters experts
 - Including foreign translations (~20% of the database)
- Basic patent bibliographic details from DWPI
- Cross-reference Accession Numbers from corresponding DWPI records
- Displayable DWPI patent family and INPADOC Legal Status information

DGENE does not include. . .

- Sequences in which fewer than 2 residues can be presented by a letter other than N (nucleic acid sequences) or X (amino acid sequences)
- Known sequences, e.g., homology comparison sequences, unless mentioned in the claims or assigned an SEQ ID number
- Non-PCR primer sequences, nucleic acid linker or adaptor sequences, unless mentioned in claims, or described via a SEQ ID number

Each DGENE sequence record includes patent bibliography, description, and indexing

ACCESSION NUMBER: AAE10698 Protein DGENE
TITLE: An isolated polypeptide (I) possessing beta-(1,3)
(1) exoglucanase activity for improvement of plant resistance to fungal phytopathogens and to promote growth
INVENTOR: Frick M M; Huang T Y; Cheng K J; Lu Z; Laroche A J; Huang H C
PATENT ASSIGNEE: (MIAC) CANADA MIN AGRIC & AGRI-FOOD CANADA.
PATENT INFO: CA 2325774 A1 20010610 (2) 86p
APPLICATION INFO: CA 2000-2325774 20001208
PRIORITY INFO: US 1999-170168P 19991210
PAT. SEQ. LOC: Claim 8; fig 2 (3)
DATA ENTRY DATE: 10 DEC 2001 (first entry)
DOCUMENT TYPE: Patent
LANGUAGE: English
OTHER SOURCE: 2001-409063 [44] (4)
CROSS REFERENCES: N-PSDB: AAD18016
DESCRIPTION: Coniothyrium minitans beta-(1,3) exoglucanase, cbeg1.
KEYWORD: (5) Beta-(1,3) exoglucanase gene; cbeg1; laminarin; plant resistance; antifungal; growth promoter; EC 3.2.1.58. (6)
ORGANISM: Coniothyrium minitans.

Each DGENE sequence record includes a Thomson enhanced abstract

ABSTRACT:

(5) The invention relates to nucleotide sequence of a novel beta-(1,3) exoglucanase gene denoted as cbeg1 of the soil borne fungus *Coniothyrium minitans*. Beta-(1,3) exoglucanase (EC 3.2.1.58) is an enzyme that catalyses the successive hydrolysis of beta-D-glucose units from the non-reducing ends of 1,3-beta-D-glucans, releasing alpha-glucose. cbeg1 is specific for the substrate laminarin. cbeg1 sequences are useful for improvement of plant resistance to fungal phytopathogens or use in ruminant microbial transgenic strategies to improve feed digestion and nutritive carbohydrate availability from forage feed. cbeg1 is also useful for use in high temperature industrial applications such as bleaching of pulp. cbeg1 is useful as an antifungal in dicots and to promote plant growth in monocots and dicots. The present sequence is *Coniothyrium minitans* cbeg1 protein.

AMINO ACID COUNTS: 73 A; 19 R; 61 N; 39 D; 0 B; 11 C; 33 Q; 18 E; 0 Z; 77
(7) G; 13 H; 52 I; 50 L; 31 K; 14 M; 25 F; 42 P; 68 S; 59 T;
16 W; 29 Y; 55 V; 0 Others

SEQUENCE LENGTH: 785

All DGENE sequences are provided in STN standardized format

SEQUENCE

(8) 1 mrlsffscl laagppasal alpspianda tsapleerqa sswleniqh
 51 qgraafnanp agykvfrnvk dygakgdgvt ddsaaainai adgnrcapwv

 701 nvlllygegfy sffisnnsnc skntnsvrdc qnrmvsiegs stvrayslne
 751 vgalqmltvd gvdkadwmpn lsgyantigy fsyni

FEATURE TABLE:

(9)

| Key | Location | Qualifier | |
|---------|----------|-----------|--|
| Region | 1..337 | note | "N-terminal region" |
| Peptide | 1..21 | label | Signal_peptide |
| Protein | 22..785 | note | "Coniothyrium minitans mature cbeg1 protein" |
| Domain | 63..82 | label | GAK_box |
| Region | 76..82 | note | "Region targetted by Gf1 semidegenerate primer" |
| Region | 338..785 | note | "C-terminal region" |
| Domain | 425..434 | label | GAX_box |

DGENE sample record annotations

- 1) Enhanced title from DWPI for the overall invention
- 2) Bibliographic information for the DWPI basic patent including the patent assignee name and code (PACO)
- 3) Patent Sequence Location (PSL), either *Claim*, *Example*, or *Disclosure* is indexed here
- 4) Other source (OS) is the Accession Number of the corresponding DWPI record, cross References (CR) are related DGENE records from the same patent
- 5) Abstract, description (DESC), and keyword (KW) fields describe the nature and uses of the specific sequence
- 6) Organism name (ORGN) providing the name of the species from which the sequence derives (where given); normally in the Latin form: genus species

DGENE sample record annotations (cont.)

- 7) Sequence Length (SQL) and count of the individual amino acids in a polypeptide (AA) or individual nucleotides in a nucleic acid sequence (NA), these are numeric search fields
- 8) The sequence (SEQ) represented with one letter codes, uncommon amino acids are indicated with X, non-standard nucleotides are indicated with N
- 9) Feature table (FEAT) describing the modifications and features of the sequence – provided by the Thomson Indexer and/or the patent applicant

DGENE patent family display

=> FILE DGENE

There are 17 sequence records
in DGENE for CA2325774.

=> S CA2325774/PN

L1 17 CA 2325774/PN

The US member of the family
was granted on May 11th, 2004.

=> D FAM

L1 ANSWER 1 OF 17 DGENE COPYRIGHT 2008 THOMSON REUTERS on STN
PI CA 2325774 A1 20010610 (200144) 79p C12N015-10
US 2003115627 A1 20030619 (200341) A01H001-00
US 6734344 B2 20040511 (200431) C12N015-56
ADT CA 2325774 A1 CA 2000-2325774 20001208; US 2003115627 A1
Provisional US 1999-170168P 19991210, US 2000-733643
20001208; US 6734344 B2 Provisional US 1999-170168P
19991210, US 2000-733643 20001208
PRAI US 1999-170168P 19991210; US 2000-733643 20001208

DGENE Legal Status display

=> FILE DGENE

=> S WO2002079175/PN

L1 4 WO2002079175/PN

=> D PI LS

L1 ANSWER 1 OF 4 DGENE COPYRIGHT 2008 THOMSON REUTERS on STN
PI WO 2002079175 A1 20021010 67p

LEGAL STATUS INPADOCDB COPYRIGHT 2005 EPO on STN

. . .

20021204 WO121 EP: THE EPO HAS BEEN INFORMED BY WIPO THAT EP WAS
DESIGNATED IN THIS APPLICATION

20040219 WOREG REFERENCE TO NATIONAL CODE
DE8642 - DE: IMPACT ABOLISHED FOR DE

There are 4 sequence records in
DGENE for WO2002079175.

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 nlfayysaaa ycgknndapa gtnitctgna cpevekadat
        51  flysfedsgv gdvtgflald ntnklivlsf rgsrsienwi glnlnfdlkei
        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" |



PCTGEN is the WIPO/PCT World Patent Applications Biosequence Database

- Produced by FIZ Karlsruhe and WIPO
- Sequences submitted & published electronically as a formal part of PCT patent applications
- Publication number and date, patent applicant name(s), and the original publication title are provided for each sequence
- Sequence length, SEQ ID, organism name, and molecule type are included for each sequence
- Updated weekly – within **24 hours** of publication
- August 2001 – present

Relationship between PCTFULL and PCTGEN databases

AN ... PCTFULL

TI

PA

PI WO A1

AB

DETD

CLM

AN Protein PCTGEN

PI WO A1

SEQ 1

AN DNA PCTGEN

PI WO A1

SEQ 2

AN Peptide PCTGEN

PI WO A1

SEQ 3

PCTFULL = WIPO/PCT patent applications full-text on STN

PCTGEN = WIPO/PCT patent application biosequences on STN

Biosequences in PCTGEN

- Information as given by the patent applicant
- Sequence length, SEQ ID number, Organism name, and Molecule Type
- Feature tables for features or modifications
- Original PCT patent application title
- Patent assignee (applicant) names
- Publication, application, and related application numbers and dates

Each PCTGEN sequence record includes publication title and bibliography

```
L1 ANSWER 1 OF 1 PCTGEN COPYRIGHT 2008 WIPO on STN
AN 2006069200.16112 PRT (1) PCTGEN
TI Group B Streptococcus (2)
PA Tettelin, Herve (2)
  Massignani, Vega (3)
PI WO 2006069200 (20060629) (3)
RLI US 2004-638943P 20041222; US 2004-640438P 20041230
ED (20060630)
DT Patent
ORGN Streptococcus agalactiae (4)
SQL 302 (5)
SEQ
    1 mflmplasll gnltvwhhkl heiikipfsr ldilihlrpt lmlflpqitm
    51 qiylslnksm lgamdsvsva gyfdqsdkii rilftivsai ggvflprlss
      . . . .
    251 atlsgavlyy intqmsvslv nyviqslvav tiyvgivfit kapviql1XX
    301 Xn
```

Sequences are typically added to PCTGEN within 24 hours of publication by WIPO.

FEATURE TABLE:

| Key | Location | (7) |
|-------------------|---------------|----------------------|
| =====+=====+===== | | |
| VARIANT | 299, 300, 301 | Xaa = Any Amino Acid |

PCTGEN sample record annotations

- 1) Accession Number (AN), this includes the sequence (SEQ ID) number, for example, AN 2006069200.16112 is SEQ ID 16112 from WO2006069200
- 2) PCT publication title for the overall invention
- 3) Patent bibliographic information: Patent Assignee (PA), Publication Number (PN), and, where given, Related Application Number (RLN) and/or Application Number (AP)

PCTGEN sample record annotations (cont.)

- 4) Organism name (ORGN) providing the name of the species from which the sequence derives
- 5) Sequence Length (SQL), searchable/sortable
- 6) The sequence (SEQ) represented with one letter codes (following WIPO standard ST.25), non-standard nucleotides are indicated with N, uncommon amino acids are indicated with X
- 7) Feature table (FEAT) describing modifications and features of the sequence, as given by the patent applicant

USGENE is the USPTO Genetic Sequence Database

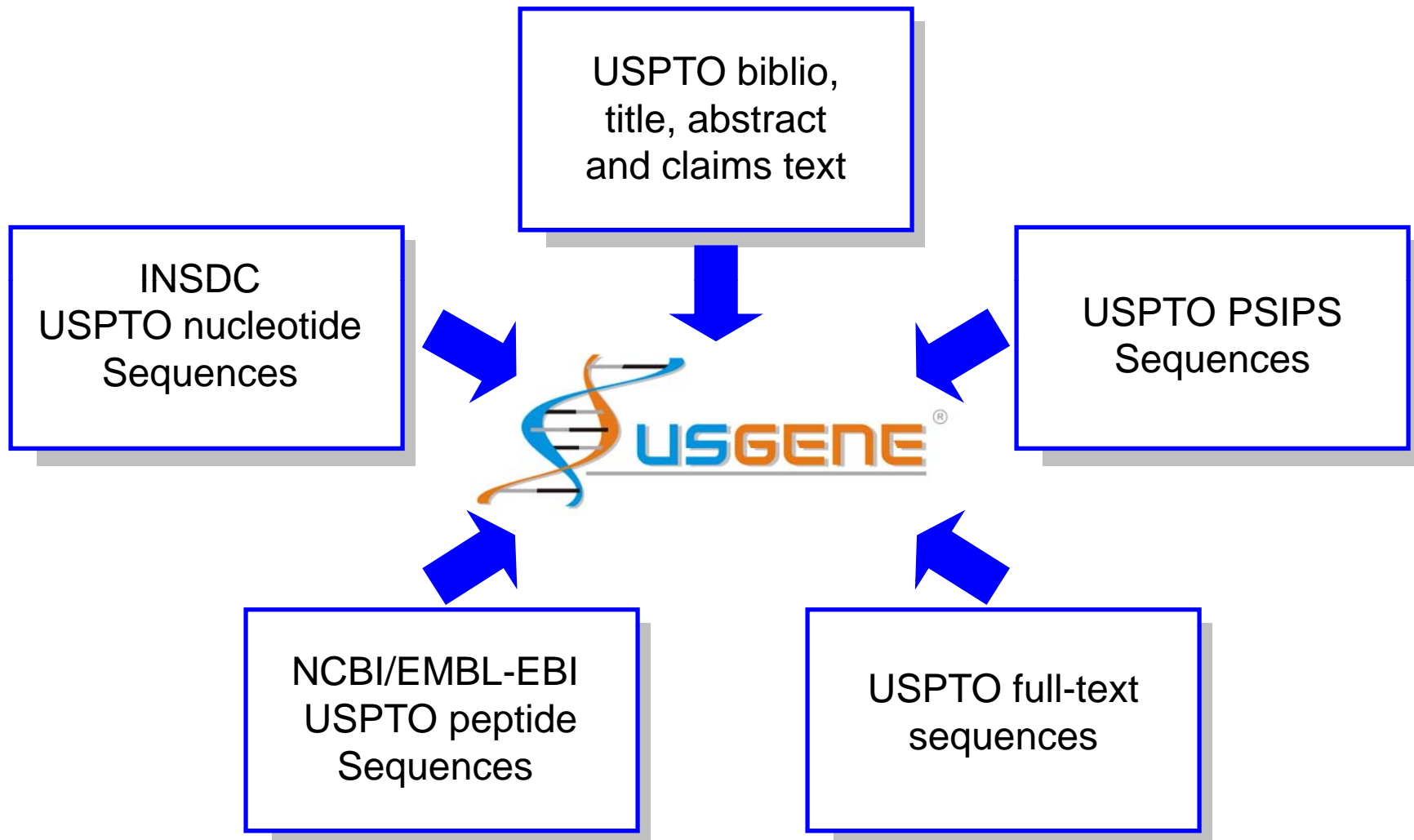
- Sequences from all relevant USPTO published patent applications and granted (issued) patents
- Assignee and full inventor names; publication, application and parent case PCT numbers and dates; original publication **title**, **abstract**, and **claims**
- Organism name, sequence length, Molecule Type, SEQ ID, and feature tables for features/annotations
- Produced by the SequenceBase Corporation
- Updated weekly – within **3 days** of publication
- 1982 – present

USGENE consolidates unique USPTO sequence data from different sources

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

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

USGENE combines these sequences with bibliographic data and claims text



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

| | |
|---|---|
| (12) United States Patent Cuttitta et al. | (10) Patent No.: US 7,364,719 B2 |
| | (45) Date of Patent: Apr. 29, 2008 |
| (54) VASOREGULATING COMPOUNDS AND METHODS OF THEIR USE | 6,440,421 B1 8/2002 Cornish et al. 2002/0055615 A1 5/2002 Cuttitta et al. |
| (75) Inventors: Frank Cuttitta , Adamstown, MD (US); Alfredo Martinez , Bethesda, MD (US); William G. Stetler-Stevenson , Kensington, MD (US); Edward J. Unsworth , Kensington, MD (US); Juan M. Saavedra , Bethesda, MD (US) | FOREIGN PATENT DOCUMENTS EP 0 845 036 6/1999 EP 0 926 238 A2 11/2000 EP 0 926 238 A3 11/2000 WO 97/07214 2/1997 WO WO 01/18550 3/2001 WO WO 2004/043383 5/2004 |
| (73) Assignee: The United States of America as represented by the Department of Health and Human Services , Washington, DC (US) | OTHER PUBLICATIONS Corti et al., "Vasopeptidase Inhibitors: A New Therapeutic Concept in Cardiovascular Disease," <i>Cardiovascular Drugs</i> 104:1856-1862 (Oct. 9, 2001). Fernandez-Patron, "Vascular Matrix Metalloproteinase-2-Dependent Cleavage of Calcitonin Gene-Related Peptide Promotes Vasoconstriction," <i>Circ Res.</i> 87:670-676 (2000). Kitamura et al., "Cloning and characterization of cDNA encoding a precursor for human adrenomedullin," <i>Biochem. Biophys. Res. Comm.</i> 194:720-725 (1993). Kitamura et al., "Adrenomedullin (11-26): a novel endogenous hypertensive peptide isolated from bovine adrenal medulla," <i>Peptides</i> 22:1713-1718 (2001). Lewis et al., "Degradation of human adrenomedullin (1-52) by plasma membrane enzymes and identification of metabolites," <i>Peptides</i> 18(5):733-739 (1997). Watanabe et al., "Vasopressor activity of N-terminal fragments of adrenomedullin in anesthetized rat," <i>Biochem. Biophys. Res. Comm.</i> 219:59-63 (1996). Belloni et al., "Proadrenomedullin N-Terminal 20 Peptide (PAMP), Acting Through PAMP(12-20)-Sensitive Receptors, Inhibits Ca ²⁺ -Dependent, Agonist-Stimulated Secretion of Human Adrenal Glands," <i>Hypertension</i> 33:1185-1189 (1999). Calvo et al., "Adrenomedullin and proadrenomedullin N-terminal 20 peptide in the normal prostate and in prostate carcinoma," <i>Microsc. Res. Tech.</i> 57(2):98-104 (Apr. 2002) <i>Abstract Only</i> . Champion et al., "Proadrenomedullin NH2-terminal 20 peptide has direct vasodilator activity in the cat," <i>Am. J. Physiol.</i> 272(4 Pt 2):R1047-54 (Apr. 1997) <i>Abstract Only</i> . Champion et al., "Tone-dependent vasodilator responses to proadrenomedullin NH2-terminal 20 peptide in the hindquarters vascular bed of the rat," <i>Peptides</i> 18(4):513-519 (1997) <i>Abstract Only</i> . |
| (*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 179 days. | |
| (21) Appl. No.: 10/529,118 | |
| (22) PCT Filed: Oct. 3, 2003 | |
| (86) PCT No.: PCT/US03/31400 | |
| § 371 (c)(1), (2), (4) Date: Mar. 24, 2005 | |
| (87) PCT Pub. No.: WO2004/032708 | |
| PCT Pub. Date: Apr. 22, 2004 | |
| (65) Prior Publication Data US 2005/0261179 A1 Nov. 24, 2005 | |
| Related U.S. Application Data | |
| (60) Provisional application No. 60/416,291, filed on Oct. 4, 2002. | |
| (51) Int. Cl. A61K 49/00 (2006.01) | |

AN Protein USGENE
PI US B2
SEQ 1

AN DNA USGENE
PI US B2
SEQ 2

•
•
•

AN cDNA USGENE
PI US B2
SEQ n

Each USGENE sequence record includes full patent bibliography, title, and abstract

| | | | |
|-----|---|------------------------|---------------------|
| L1 | ANSWER 1 OF 1 USGENE COPYRIGHT 2008 SEQUENCEBASE | | ALL display format. |
| AN | 7364719.3 | (1) Protein (2) USGENE | |
| TI | Vasoregulating compounds and methods of their use (Patent) (3) | | |
| IN | Cuttitta Frank (Adamstown, MD); Martinez Alfredo (Bethesda, MD) (4) . | | |
| PA | The United States of America as represented by the Department of Health and Human Services (Washington DC) (5) | | |
| PI | US 7364719 | B2 | 20080429 |
| | US 20050261179 | A1 | 20051124 |
| | WO 2004032708 | A | 20040422 |
| AI | US 2003-529118 | | 20031003 |
| RLI | WO 2003-US31400 | | 20031003 |
| ED | 20080502 | | |
| AB | <p>Methods and compounds are described for regulating blood pressure in a subject. Methods for reversing vasodilation by administering to a subject a vasodilator (e.g., nitroglycerin) in a dosage of 0.1 to 10 mg/kg body weight (AM(11-22)). The methods are useful for a variety of purposes, including hemostasis or the treatment of shock, for example vasodilatory shock syndromes such as septic shock. Other specific embodiments are methods for reversing vasoconstriction of blood vessels, by</p> | | |

(6)

See (1) - (7) on slide 32.

(7) Granted patent sequences are typically available within 3 days publication by the USPTO.

Each USGENE sequence record includes patent or published application claims text

CLM US7364719 B2: 1. A method of vasoconstricting blood vessels in a subject, comprising:(a) selecting a subject in need of vasoconstriction; and(b) administering to the subject a therapeutically effective amount of peptide consisting of the peptide AM(11-22) (SEQ ID NO: 4) sufficient to induce vasoconstriction, thereby vasoconstricting blood vessels in the subject.

(8)

2. The method of claim 1, wherein the method of vasoconstricting blood vessels comprises administering the peptide consisting of the peptide AM(11-22) (SEQ ID NO: 4) to a subject experiencing or at risk of experiencing shock.

4. The method of claim 1, wherein the method comprises administering the peptide consisting of the peptide AM(11-22) (SEQ ID NO: 4) to a subject experiencing or at risk of experiencing septic shock.

5. A pharmaceutical composition comprising a therapeutically effective amount of the peptide AM(11-22) (SEQ ID NO: 4).

. . . .

Note: USGENE record AN 7364719.3 is SEQ ID NO: 3 from US7364719 and is displayed here in full, using the **ALL** format.

All USGENE sequences are provided in STN standardized format

SSO PROTEIN; USPTO; GRANTED (9)

ALL display format (cont.)

ORGN Homo Sapiens (10)

SQL 52 (11)

SEQ

(12) 1 yrqsmnnfqq lrsfgcrfgt ctvqklahqi yqftdkdkdn vaprskispq
51 gy

FEATURE TABLE:

(13)

Key | Location |

=====+=====+=====

mat_peptide | (1)..(52) | Mature adrenomedullin,
| | corresponding to
| | positions 95-146 of
| | preproadrenomedullin (SEQ ID
| | NO 2)

See (8) - (13)
on slide 33.

USGENE sample record annotations

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

USGENE sample record annotations (cont.)

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

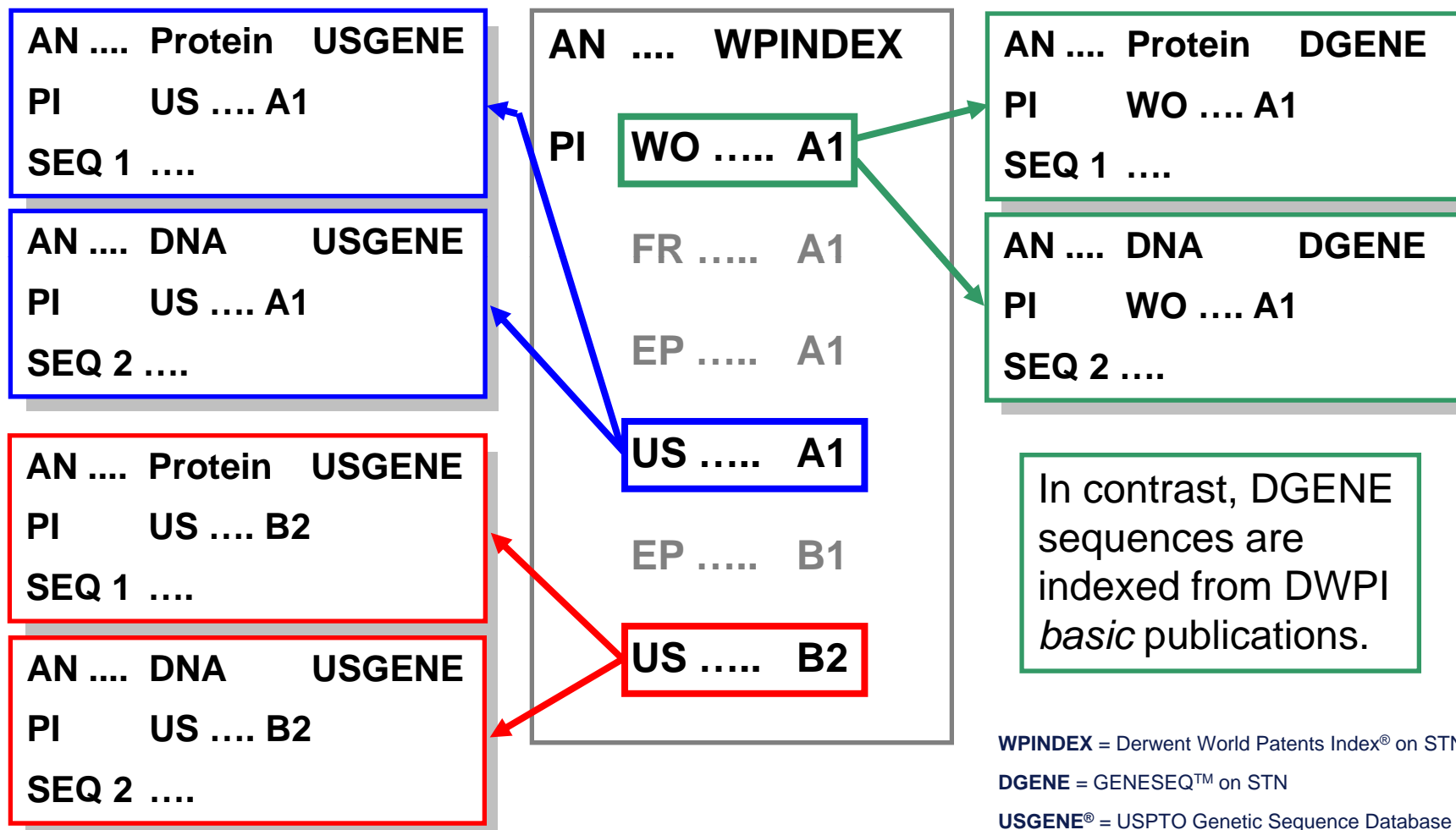
USGENE is an essential additional tool for tackling business critical searches

- 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
 - Updated weekly, within **3 days** of USPTO publication
 - Both **U.S. patents** and **U.S. published applications**
- Sequence listing variation often occurs between PCT and U.S. granted patent publication stages
 - Especially important, e.g., for freedom-to-operate

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

L1 ANSWER 1 OF 1 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
AN 20080069867.4 Protein USGENE
TI Diagnostic and/or Remedy for Ovarian Cancer (Patent) AN 20080069867.4 is
IN Shimada Kaoru (Chiba, JP); Matsumura Yasuhiro (Tokyo, JP) SEQ ID NO: 4 from
Toshiaki (Tokyo, JP) US20080069867.
PA Mitsubishi Pharma Corporation (Osakashi JP)
PI US 20080069867 A1 20080320
AI US 2005-575406 20050819
RLI WO 2005-JP15119 20050819
ED 20080321
DT Patent
AB It was found that a protein sequence, as described in the references, is typically available within 1 day of publication. The protein is an antibody having high specificity for ovarian cancer, as well as the cancer species known so far such as gastric cancer, colon cancer and breast cancer, and was useful as an ovarian cancer tissue-specific diagnostic and/or therapeutic agent.
ECLM US20080069867 A1: 1-14. (canceled) 15: A diagnostic and/or therapeutic agent for ovarian cancer, which comprises an antibody containing the amino acid sequence of any one of SEQ ID NOS: 1 to 6 described in the sequence listing.
SSO PROTEIN; USPTO; APPLICATION
ORGN Homo Sapiens
SQL 17
SEQ
1 kssqsvlyns nnkkyla

USGENE provides sequences from both USPTO published applications and **granted patents**



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

```
L1 ANSWER 1 OF 1 WPINDEX COPYRIGHT 2008 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 1994
FI 9505224 A 1995
EP 698101 A1 1996
JP 09500009 W 1997
AU 695259 B 1998
EP 698101 B1 2004
DE 69434116 E 2004
US 20050032047 A1 20050210 (200512) EN
US 6881821 B2 20050419 (200527) EN
. . . . .
ADT WO 9425602 A1 WO 1994-GB957 19940505 . . . . .
PRAI GB 1994-263 19940107
GB 1993-9237 19930505
```

In this example the patent family has:

- 9 sequences from [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 |
| 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) |

Overview of timeliness of the various sources of patent sequence data

| | Update Frequency | Typical Timeliness | Value added |
|-----------|------------------|--------------------|---|
| PCTGEN | Weekly | 24 hours | |
| USGENE | Weekly | 3 days | |
| REGISTRY | Daily | 27 days |  |
| DGENE | Biweekly | 65 days |  |
| NCBI/EMBL | Daily | 1-6 months | |


Comparing STN databases...

- **DGENE**
 - The most comprehensive patent sequence database
 - Implemented in-house at major patent offices
- **PCTGEN**
 - The most timely database (24 hours)
 - Sequences from equivalent WIPO/PCT publications
- **USGENE**
 - More timely than DGENE and REGISTRY (3 days)
 - Sequences from equivalent USPTO applications and patents
- **REGISTRY**
 - More timely than DGENE; complementary indexing
 - Unique non-patent literature coverage

Agenda

- DGENE, PCTGEN, USGENE database content
- **The 7 basic steps of BLAST**
- BLAST and Patent Family SORT (FSORT)
- Post-processing BLAST search results
- Similarity searching GETSIM (FASTA)
- Offline BATCH search mode
- Sequence Code Match searching (GETSEQ)
- Multifile patent sequence search example

DGENE, PCTGEN, and USGENE offer exactly the same sequence search options

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

Note: In this workshop we will be practicing these skills in USGENE.

The 7 basic steps of USGENE BLAST

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

Similarity searching in USGENE using BLAST

Search Question:

Find relevant U.S. published application and patent references for this protein sequence:

```
1 vqtvplsrlf dhamleahra helaidtyqe feetyipkdq kysflhdsqt
51 sfcfsdsipt psnmeetqk snlellrisl llieswlepv rflrsmfann
101 lvydtsdsdd yhllkdleeg iqtlmgrled gsrrtgqilk qtyskfdtns
151 hnhdallkny gllycfrkdm dkvetflrmv qcrsvegscg f
```

1) SAVE, UPLOAD and VERIFY

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

- Upload options
 - STN Express[®]: Use UPLOAD command or Upload Query Wizard (STN Express 8.x)
 - STN[®] on the WebSM: Use Upload feature or Sequence Assistant (link below)
- Verify the sequence with D LQUE

STN on the Web Sequence Search Assistant:

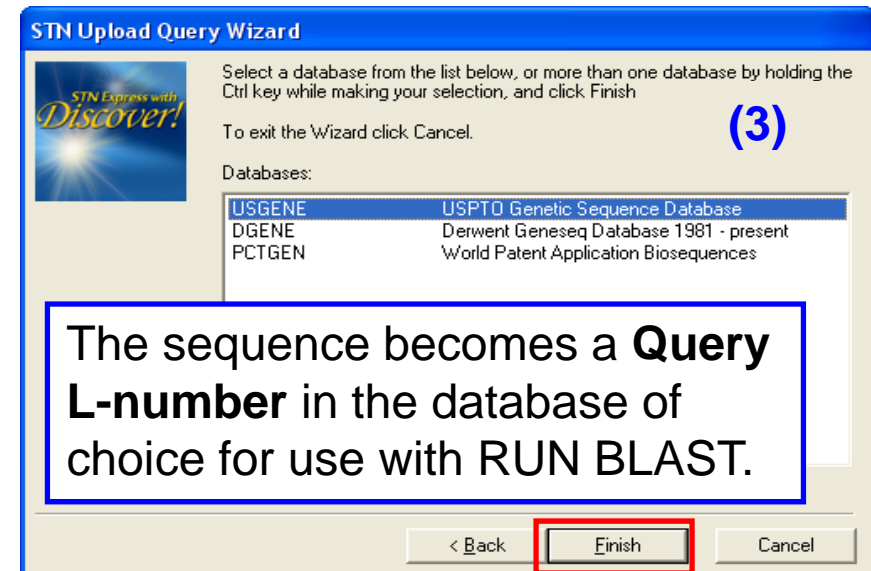
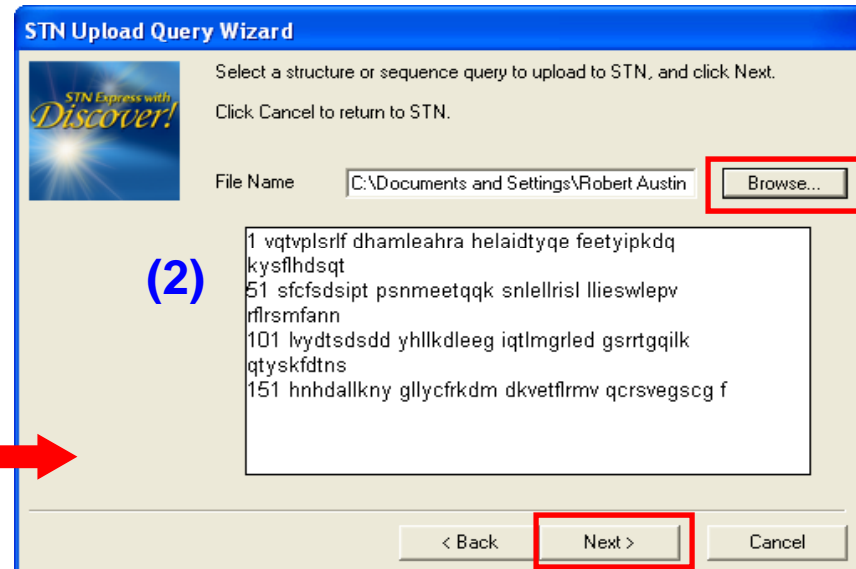
www.stn-international.com/training_center/bioseq/seq_se_ass.pdf

UPLOAD the sequence via STN Express

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



From the *Discover!* button menu.



1) SAVE, UPLOAD and VERIFY (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 the sequence was uploaded successfully with **D LQUE**.

```
L1 ANSWER 1 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
LQUE vqtvplsrlfdhamleahrahelaidtyqefeetyipkdqkysflhdsqtsfcfsdsi
ptpsnmeetqqksnllellrislllieswlepvrflrsmfannlvdytdsdsddyhllkd
leegiqtlmgrledgsrrtgqilkqtyskfdtnshnhdallknygllycfrkdmdkve
tflrmvqcrsvegscgf
```

```
=>
```

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

The 7 basic steps of USGENE BLAST

2) RUN the BLAST search

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

2) RUN the USGENE BLAST search

=> FILE USGENE

FILE 'USGENE' ENTERED AT 12:09:16 ON 02
COPYRIGHT (C) 2008 SEQUENCEBASE CORP

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

FILE LAST UPDATED: 2 MAY 2008 <20080502/UP>
MOST RECENT PUBLICATION DATE: 1 MAY 2008 <20080501/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
with the syntax... /SQP -F F

BLAST Version 2.2

The BLAST software is used herein with permission of the
National Center for Biotechnology Information (NCBI) of
the National Library of Medicine (NLM). See also,

BLAST SEARCHING

RUN BLAST command syntax

Similarity Searching with BLAST (protein/polypeptides)

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

/SQP (protein) (default)

-e (Expect-value)

-f (Filter) (on by default)

-w (Word size)

-m (Matrix)

-g (Gap penalty)

-x (Gap extension)

BATCH (offline)

ALERT (Alert/SDI)

RUN BLAST command syntax

Similarity Searching with BLAST (Nucleotide sequences)

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

/SQN (nucleotide)

SIN (single strand)

COM (complementary strand)

BOTH (both strands) (default)

-e (Expect-value)

-f (Filter)

-w (Word size)

-g (Gap penalty)

-x (Gap extension)

-q (penalty for mismatch)

-r (reward for match)

BATCH (offline)

ALERT (Alert/SDI)

RUN BLAST advanced options

Expectation Value (-E)

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

Word Size (-W)

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

RUN BLAST advanced options (cont.)

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

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

Peptide similarity matrices (-M)

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

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

| <u>Query Length</u> | <u>Matrix</u> | <u>Gap costs</u> |
|---------------------|---------------|------------------------|
| <35 | PAM-30 | (9,1) |
| 35 – 50 | PAM-70 | (10,1) |
| 50 – 85 | BLOSUM-80 | (10,1) |
| >85 | BLOSUM-62 | (11,1) (BLAST default) |

Tip: Type [HELP OPTIONS](#) in USGENE for more information on using BLAST advanced options.

The 7 basic steps of USGENE BLAST

3) Decide how many answers to keep (L2)

- After the BLAST search, STN provides a chart summarizing the results, and asks this question:

ENTER EITHER THE NUMBER OF ANSWERS YOU WISH TO KEEP OR ENTER MINIMUM PERCENT OF SELF SCORE FOLLOWED BY %

(BEST ANSWER PERCENTAGE IS nnn%)

ENTER (ALL) OR ? :



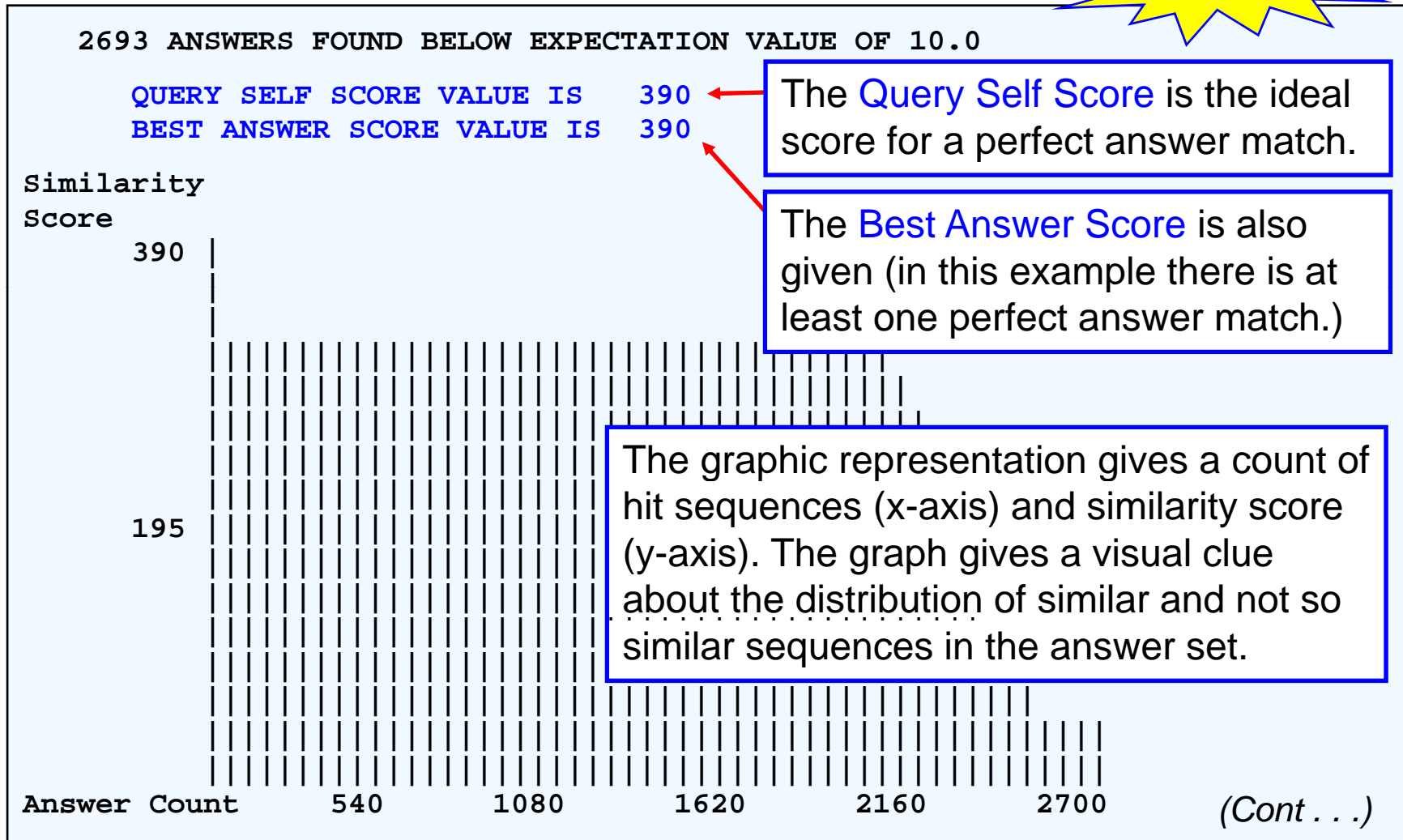
- General recommendation: Keep **ALL** answers*

(* Or use BATCH mode to enable multiple retrievals – more on that later in the workshop!)

The 7 basic steps of USGENE BLAST

- 4) SORT by SCORE descending (L3)
 - Sort the BLAST results answer set:
=> **SOR L2 SCORE D**
 - Option: Limit using text terms and/or dates (L4)
 - Remember to SORT L4 SCORE D !! (L5)

3) Decide how many answers to keep



4) SORT by SCORE descending

New!

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 ? : 85%

In this example, 85% of the Query Self Score is used to select out just the most relevant results (L2).

L2 RUN STATEMENT CREATED

L2 153 VQTVPLSRLFDHAMLEAHRAHELAIPTYQEFEEYIPKDQKYSFLHDSQT
SFCFSDSIPTPSNMEETQOKSNLELLRISLLLLIESWLEPVRFRLRSMFANN
LVYDTSDDYHLLKDLLEGIQTLMGRLEDGSRRTGQILKQTYSKFDTNS
HNHDALLKNYGLLYCFRDKMDKVETFLRMVQCRSVEGSCGF/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 153 SOR L2 SCORE D

Use SORT SCORE D to sort by descending BLAST score.

The 7 basic steps of USGENE BLAST

5) Review answers using a *free-of-charge* format including alignment (ALIGN), while “parked” in STNGUIDESM

- D L3 TRI ORGN SCORE ALIGN 1-
- FILE STNGUIDE



New !

Note: The SCORE display field also includes the percentage of the [Query Self Score](#) (maximum possible BLAST score).

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

=> D L3 TRI ORGN SCORE ALIGN 1-153; FILE STNGUIDE

L3 ANSWER 1 OF 153 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN

TI Recombinant DNA transfer vectors (Patent)

MTY Protein

SQL 191

ORGN Unknown

SCORE 390

100% of query self score 390

New!

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

BLASTALIGN

Query = 191 letters

Length = 191

Score = 390 bits (1001), Expect = e-113

Identities = 191/191 (100%), Positives = 191/191 (100%)

Query: 1 VQTVPLSRLFDHAMLEAHRAHEL AIDTYQEF EETYIPKDQKYSFLHDSQTSFCFSDSIPT

VQTVPLSRLFDHAMLEAHRAHEL AIDTYQEF EETYIPKDQKYSFLHDSQTSFCFSDSIPT

Sbjct: 1 VQTVPLSRLFDHAMLEAHRAHEL AIDTYQEF EETYIPKDQKYSFLHDSQTSFCFSDSIPT

Query: 61 PSNMEETQQKSNLELLRISLLLI ESWLEPVRFLRSMFANNLVYDTS DSDDYHLLKDLEEG

PSNMEETQQKSNLELLRISLLLI ESWLEPVRFLRSMFANNLVYDTS DSDDYHLLKDLEEG

Sbjct: 61 PSNMEETQQKSNLELLRISLLLI ESWLEPVRFLRSMFANNLVYDTS DSDDYHLLKDLEEG

. . . .

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

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

L3 ANSWER 5 OF 153 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
TI Novel antiangiogenic peptide agents and their therapeutic and
diagnostic use ([PublishedApplication](#))

MTY Protein

SQL 192

ORGN Homo Sapiens

SCORE 387 99% of query self score 390

BLASTALIGN

Query = 191 letters

Length = 192

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

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

Query: 1 VQTVPLSRLFDHAMLEAHRAHELAIPTYQEFEEITYIPK
VQTVPLSRLFDHAML+AHRAH+LAIDTYQEFEEITYIPK

Sbjct: 2 VQTVPLSRLFDHAMLQAHAHQLAIDTYQEFEEITYIPKDQKYSFLHDSQTSFCFSDSIPT

Query: 61 PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDDSDDYHLLKDLEEG
PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDDSDDYHLLKDLEEG

Sbjct: 62 PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDDSDDYHLLKDLEEG

. . . .

The 5th from top hit comes from a U.S. published application.

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

Understanding BLAST alignments

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

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

USGENE provides text search options for refining sequence searches

- The USGENE default text search index – known on STN as the *Basic Index (/BI)* – comprises
 - Original publication Title (/TI) and abstract (/AB)
 - Organism name (/ORGN) and Molecule Type (/MTY)
- The Exemplary Claim (/ECLM) and Feature Table (/FEAT) can also be added to a search
 - Either specify the fields: => **S VIRUS/BI,FEAT**
 - Or use SET SFIELDS: => **SET SFIELDS BI ECLM**
- The Basic Index and Feature Table both offer simultaneous left and right truncation (**SLART**)

USGENE provides bibliographic search options for refining sequence searches

- Patent Assignee (/PA) and Inventor (/IN)
 - Examples: GLAXO/PA, SMITH JOHN/IN
- Granted or application Sequence Source (/SSO)
 - Examples: APPLICATION/SSO, GRANTED/SSO
- Publication date (/PD) or publication year (/PY)
 - Examples: PY < 2001, PD < 1 Mar 1995
- Application date (/AD) or application year (/AY)
 - Examples: AY < 2002, AD < 1 Mar 1998
- WO application date (/RLD) or year (/RLY)
 - Examples: RLY < 1993, RLD < 1 Aug 1986

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

```
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 ? : 85%
```

In this example, 85% of the **Query Self Score** is used to select out just the most relevant results (L2).

```
L2 RUN STATEMENT CREATED
```

```
L2 153 VQTVPLSRLFDHAMLEAHRAHELAI  
SFCFSDSIPTPSNMEETQQKSNLELI  
LVYDTSDDYHLLKDLLEGIQTLMGRLEDGSRRTGQILKQTYSKFDTNS  
HNHDALLKNYGLLYCFRKDMDKVETFLRMVQCRSVEGSCGF/SQP.-F F
```

```
Answer set arranged by accession number  
similarity score, enter at an arrow
```

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

```
=> SOR SCORE D
```

```
PROCESSING COMPLETED FOR L2
```

```
L3 153 SOR L2 SCORE D
```

```
=> S L2 AND SOMATOMAMMOTROPIN/BI,ECLM AND AY<1996 AND GRANTED/SSO
```

```
L4 2 L2 AND SOMATOMAMMOTROPIN/BI,ECLM AND AY<1996 AND GRANTED/SSO
```

```
=> SOR SCORE D
```

```
PROCESSING COMPLETED FOR L4
```

```
L5 2 SOR L4 SCORE D
```

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

The 7 basic steps of USGENE BLAST

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

6) Display selected USGENE answers in a preferred bibliographic format

=> D BIB AB ECLM ORGN SSO SCORE ALIGN 1-2

L5 ANSWER 1 OF 2 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN

AN 4363877.1 Protein USGENE

TI Recombinant DNA transfer vectors (Patent

IN Goodman Howard M. (San Francisco, CA); S

CA); Seeburg Peter H. (San Francisco, CA

PA The Regents of the University of California

PI US 4363877 A 19821214

AI US 1978-897710 19780419

AB Recombinant DNA transfer vectors containing codons for human **somatomammotropin** and for human growth hormone.

ECLM US4363877 A: What is claimed is:

1. A recombinant DNA transfer vector comprising codons for human chorionic **somatomammotropin** comprising the nucleotide

ORGN Unknown

SSO PROTEIN; EMBL; **GRANTED**

SCORE 390 100% of query self score 3

BLASTALIGN

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

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

Useful USGENE display fields/formats

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

(* Free of charge display formats in USGENE.)

The importance of using the correct BLAST advanced options

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

```
. . . .
```

```
NO ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0
```

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

```
. . . .
```

```
1107 ANSWERS FOUND BELOW EXPECTATION VALUE OF 1000.0
```

```
QUERY SELF SCORE VALUE IS 38
```

```
BEST ANSWER SCORE VALUE IS 38
```

```
. . . .
```

```
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 ? : ALL
```

```
L1 RUN STATEMENT CREATED
```

```
L1 1107 GSSFLSPEHQR/SQP.-M PAM30 -W 2 -E 1000 -F F
```

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

Changing BLAST options is especially important for short sequence queries!

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

=> SOR L1 SCORE D

PROCESSING COMPLETED FOR L1
L2 1107 SOR L1 SCORE D

Correct use of BLAST options
finds relevant sequence hits.

=> D TRI ORGN SCORE ALIGN

L2 ANSWER 1 OF 1107 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
TI Antibodies against the PRO1754 polypeptides (Patent)
MTY Protein
SQL 117
ORGN Homo Sapiens
SCORE 38 100% of query self score 38

BLASTALIGN

Query = 11 letters

Length = 117

Score = 37.5 bits (81), Expect = 4e-09

Identities = 11/11 (100%), Positives = 11/11 (100%)

Query: 1 GSSFLSPEHQR 11

GSSFLSPEHQR

Sbjct: 24 GSSFLSPEHQR 34

Reminder: Type **HELP OPTIONS**
in USGENE for more information
on using BLAST advanced options.

Review: 7 steps of USGENE BLAST

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

Agenda

- DGENE, PCTGEN, USGENE database content
- The 7 basic steps of BLAST
- **BLAST and Patent Family SORT (FSORT)**
- Post-processing BLAST search results
- Similarity searching GETSIM (FASTA)
- Offline BATCH search mode
- Sequence Code Match searching (GETSEQ)
- Multifile patent sequence search example

USGENE answer sets may be grouped by source publications using Family SORT (FSORT)

- FSORT gathers multiple sequence hits from the same applications together via publication, application and/or WO/PCT related application numbers
- FSORT organizes answers into two subgroups: multiple sequence hit (multi-record) families and single sequence hit (individual-record) families
- When FSORT is used on an answer set previously sorted by similarity SCORE, the two FSORT subgroups each separately retain their similarity sort order
- FSORT makes it possible to review, e.g., just the most similar sequence answer for each application retrieved or all the sequences from a single application

USGENE answer sets may be grouped by source publications using Family SORT (FSORT)

Search Question:

Find all relevant U.S. published application and patent references with sequences similar to the *Banana Bunchy Top Virus (BBTV) Replication Initiation Protein* (NCBI: AAG44003).

Banana Bunchy Top Virus (BBTV) Replication Initiation Protein (NCBI: AAG44003)

NCBI Sequence Viewer v2.0 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?query_key=14&db=protein&qty=

STN International STN/CAS Home Page FIZ Karlsruhe Inc STN on the Web STN Viewer CRS Summary Sheets DWPI Reference DWPI user guides

NCBI Entrez Protein My NCBI [Sign In] [Register]

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books

Search Protein for [Go] [Clear]

Limits Preview/Index History Clipboard Details

Display FASTA Show 5 Send to

Range: from begin to end [Refresh]

1: [AAG44003](#). Reports replication initi... [gi:12004326] BLink, Conserved Domains, Links

>gi|12004326|gb|AAG44003.1|AF216221_1 replication initiation protein [Banana bunchy top virus]
MSSFKWCFTLNYSAAEREDFLALLKKEELNYAVVGDEVAPSSGQKHLQGYLSLKKSIKLGGLKKKYSSR
AHWERARGSDDEDNAKYCSKETLIL
KEEFVHPCLDRPWQIQLTEAIDEE
DEGSEKHIVFDIPRCNQDYLNYDV
IIYC

bbtv protein - Notepad

File Edit Format View Help

```
MSSFKWCFTLNYSAAEREDFLALLKKEELNYAVVGDEVAPSSGQKHLQGYLSLKKSIKLGGLKKKYSSR
AHWERARGSDDEDNAKYCSKETLILELGFPPASQGSNRRKLSEMVSRSPEMRIEQPEIYHRYTSVKLKKF
KEEFVHPCLDRPWQIQLTEAIDEEPDDRSIIWVYGPNGNEGKSTYAKSLMKKDWFYTRGGKKENILFSYV
DEGSEKHIVFDIPRCNQDYLNYDVIEALKDRVIESTKYPKIKLVELINIHVIVMANFMPEFCKISEDRIK
IIYC
```

Done

SAVE, UPLOAD and VERIFY

=> FILE USGENE

=> UPL R BLAST

These commands are automatically run by the STN Express Sequence Query Upload wizard (slide 46).

UPLOAD SUCCESSFULLY COMPLETED

L1 GENERATED

=> D L1 LQUE

L1 ANSWER 1 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
LQUE MSSFKWCFTLNYSSAAEREDFLALLKEEELNYAVVGDEVAPSSGQKHLQGYLSLKKS
LGGLKKKYSSRAHWERARGSDENAKYCSKETLILELGFASQGSNRRKLSEMVSRSPE
RMRIEQPEIYHRYTSVKKLKFKKEEFVHPCLDRPWQIQLTEAIDEEPDDRSIIWVYGPN
GNEGKSTYAKSLMKKDFYTRGGKKNILFSYVDEGSEKHIVFDIPRCNQDYLNVDVIE
ALKDRVIESTKYKPIKLVELINIHVIVMANFMPEFCKISEDRIKIIYC

=>

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

RUN the USGENE BLAST search

=> **FILE USGENE**

FILE 'USGENE' ENTERED AT 14:19:01 ON 02
COPYRIGHT (C) 2008 SEQUENCEBASE CORP

FILE LAST UPDATED: 2 MAY 2008 <20080502/UP>
MOST RECENT PUBLICATION DATE: 1 MAY 2008 <20080501/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**

BLAST Version 2.2

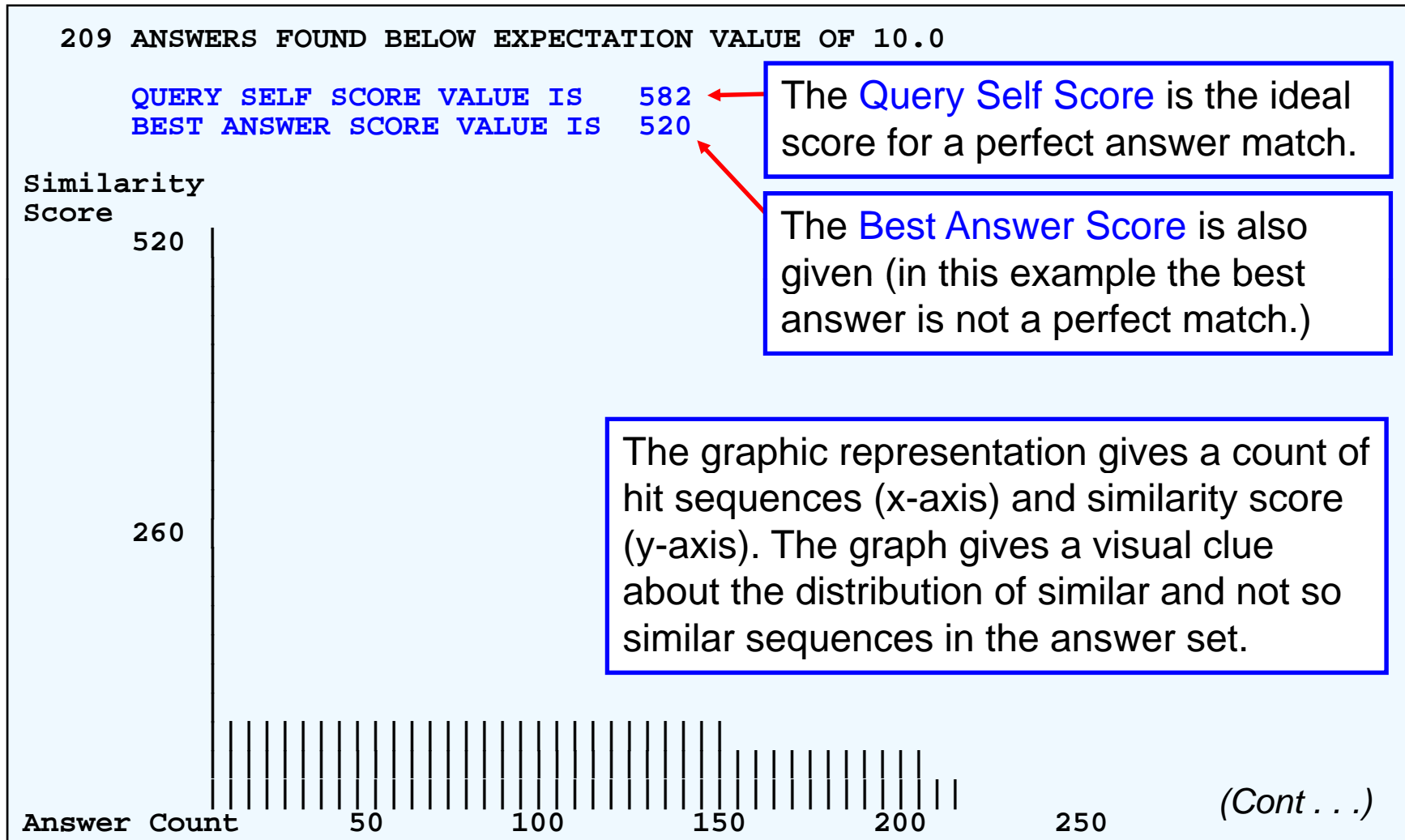
The BLAST software is used herein with permission of the
National Center for Biotechnology Information (NCBI) of
the National Library of Medicine (NLM). See also,

BLAST SEARCHING

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

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

Decide how many answers to keep



SORT by SCORE descending

```
ENTER EITHER THE NUMBER OF ANSWERS YOU WISH TO KEEP
OR ENTER MINIMUM PERCENT OF SELF SCORE FOLLOWED BY %
(BEST ANSWER PERCENTAGE IS 89%)
```

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

```
L2 RUN STATEMENT CREATED
```

```
L2 209 MSSFKWCFTLNYSAAEREDFLALLKEE
    YLSLKKSILKGLKKKYSSRAHWERARGSDENAKYCSKETLILELGFPA
    . . .
    IKLVELINIHVIVMANFMPEFCKISEDRIKIIYC/SQP.-F F
```

In this example, ALL answers have been kept (L2).

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 209 SOR L2 SCORE D
```

```
=> SET FORMAT .MYUSGENE BIB AB ECLM ORGN SQL SCORE ALIGN
```

```
SET COMMAND COMPLETED
```

```
=> SET DFORMAT .MYUSGENE
```

```
SET COMMAND COMPLETED
```

Option: Set a customized display format with SET FORMAT. The new format may be set as the file default with SET DFORMAT.

Display selected USGENE answers using the new customized default display format

=> D 1-2

```
L3 ANSWER 1 OF 209 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
AN 5846705.16 Protein USGENE
TI Nucleotide sequence of two circular SSDNA associated with banana
   bunchy top virus and method for detection of banana bunchy top virus
   (Patent)
IN Wu Rey-Yuh (Taipei, TW); You Li-Ru (Taipei, TW); Soong Tai-Seng
   (Taipei,TW)
PA Development Center for Biotechnology (Taipei TW)
PI US 5846705 A 19981208
AI US 1995-418071 19950406
AB Nucleotide sequences of two circular single-stranded DNAs . . . .
ECLM US5846705 A: 1. An isolated DNA molecule comprising a . . . .
ORGN Unknown
SQL 286
SCORE 520 89% of query self score 582
BLASTALIGN
   Query = 284 letters
   Length = 286
   Score = 520 bits (1338), Expect = e-152
   Identities = 247/282 (87%), Positives = 268/282 (94%)
Query: 3 SFKWCFTLNYSSAAEREDFLALLKEEELNYAVVGDEVAPSSGQKHLQGYLSLKKSIKLG
        S KWCFTLNYSSAAERE+FL+LLKEE+++YAVVGDEVAP++GQKHLQGYLSLKK I+LGG
Sbjct: 5 SLKWCFTLNYSSAAERENFLSLLKEEDVHYAVVGDEVAPATGQKHLQGYLSLKKRIRLGG
        . . . . .
```

The top hit is SEQ ID 16 from US5846705.

The second hit sequence comes from the same U.S. patent as the top hit

```
L3 ANSWER 2 OF 209 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
AN 5846705.17 Protein USGENE
TI Nucleotide sequence of two circular SSDNA associated with banana
   bunchy top virus and method for detection of banana bunchy top virus
   (Patent)
IN Wu Rey-Yuh (Taipei, TW); You Li-Ru (Taipei, TW); Soong Tai-Seng
   (Taipei, TW)
PA Development Center for Biotechnology(Taipei TW)
PI US 5846705 A 19981208
AI US 1995-418071 19950406
AB Nucleotide sequences of two circular single-stranded DNAs . . . .
ECLM US5846705 A: 1. An isolated DNA molecule comprising a nucleotide
     sequence encoding a polypeptide comprising amino acid . . . .
ORGN Unknown
SQL 285
SCORE 340
BLASTALIGN
    Query = 284 letters
    Length = 285
    Score = 340 bits (872), Expect = 2e-98
    Identities = 171/288 (59%), Positives = 217/288 (74%), Gaps = 7/288
Query: 1 MSSFKWCFTLNYSAAEREDFLALLKKEEELNYAVVGDEVAPSSGQKHLQGYLSLKKSIKL
        MSSFKWCFTLNYSAAEREDFLALLKKEE+++Y+VVGDEVAP++GQKHL GYLSLKKSI+L
Sbjct: 1 MSSFKWCFTLNYSAAEREDFLALLKEEDVHYSVVGDEVAPATGQKHLGGYLSLKKSIRL
        . . . . .
```

The 2nd hit is SEQ ID 17 from US5846705.

USGENE answer sets may be grouped by source publications using Family SORT (FSORT)

=> FSORT L3

. . . .

L4 209 FSO L5

36 Multi-record Families

Family 1

Family 2

Family 3

Family 4

Family 5

Family 6

Family 7

Family 8

Family 9

. . . .

Family 31

Family 32

Family 33

Family 34

Family 35

Family 36

7 Individual Records

0 Non-patent Records

Answers 1-202

Answers 1-3

Answers 4-5

Answers 6-7

Answers 8-9

Answers 10-13

Answers 14-15

Answers 16-17

Answers 18-19

Answers 20-21

Answers 178-183

Answers 184-189

Answers 190-195

Answers 196-197

Answers 198-200

Answers 201-202

Answers 203-209

The 209 sequence hits belong to 36 multi-hit and 7 individual-hit source publications.

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

The top answer is the same as before....

=> D PFAM=1-2

The first record from families 1 & 2 in default format.

```
L4 ANSWER 1 OF 209 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN FAMILY1
AN 5846705.16 Protein USGENE
TI Nucleotide sequence of two circular SSDNA associated with banana
   bunchy top virus and method for detection of banana bunchy top virus
   (Patent)
IN Wu Rey-Yuh (Taipei, TW); You Li-Ru (Taipei, TW); Soong Tai-Seng
   (Taipei, TW)
PA Development Center for Biotechnology (Taipei TW)
PI US 5846705 A 19981208
AI US 1995-418071 19950406
AB Nucleotide sequences of two circular single-stranded DNAs . . . .
ECLM US5846705 A: 1. An isolated DNA molecule comprising a . . . .
ORGN Unknown
SQL 286
SCORE 520 89% of query self score 582
BLASTALIGN
   Query = 284 letters
   Length = 286
   Score = 520 bits (1338), Expect = e-152
   Identities = 247/282 (87%), Positives = 268/282 (94%)
Query: 3 SFKWCFTLNYSAAEREDFLALLKEEELNYAVVGDEVAPSSGQKHLQGYLSLKKSIKLG
        S KWCFTLNYSAAERE+FL+LLKEE+++YAVVGDEVAP++GQKHLQGYLSLKK I+LGG
sbjct: 5 SLKWCFTLNYSAAERENFLSLLKEEDVHYAVVGDEVAPATGQKHLQGYLSLKKRIRLGG
        . . . . .
```

The top hit is SEQ ID
16 from US5846705.

...but the second answer displayed is now
the best answer from the 2nd family

```
L4 ANSWER 4 OF 209 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN FAMILY2
AN 5756708.26 Protein USGENE
TI DNA sequences of banana bunchy top virus (Patent)
IN Karan Mirko (Holland Park, AU); Burns Thomas Michael (Herston, AU);
   Dale James Langham (Moggill, AU); Harding Robert Maxwell(Lawnton, AU)
PA Queensland University of Technology(Brisbane AU)
PI US 5756708 A 19980526
AI US 1994-202186 19940224
DT Patent
AB The invention provides DNA molecules consisting essentially of a
   nucleotide sequence or part thereof which are associated . . . .
ECLM US5756708 A: 1. An isolated DNA molecule derived from banana bunchy
   top virus, consisting of a nucleotide sequence selected . . . .
ORGN Unknown
SQL 290
SCORE 243 41% of query self score 582
BLASTALIGN
   Query = 284 letters
   Length = 290
   Score = 243 bits (621), Expect = 3e-69
   Identities = 117/282 (41%), Positives = 183/282 (64%), Gaps = 6/282
Query: 5 KWCFTLNYSSAAEREDFLALLKEEELNYAVVGDEVAPSSGQKHLQGYLSLKSIKLGGLK
        +WCFTLNY + E + + ++ L YA+VGDEVAPS+GQ+HLQG++ LK +L GLK
Sbjct: 7 RWCFTLNYETEEEEANVVRRIESLNLVYAIVGDEVAPSTGQRHLQGFHHLKTGRRLQGLK
        . . . . .
```

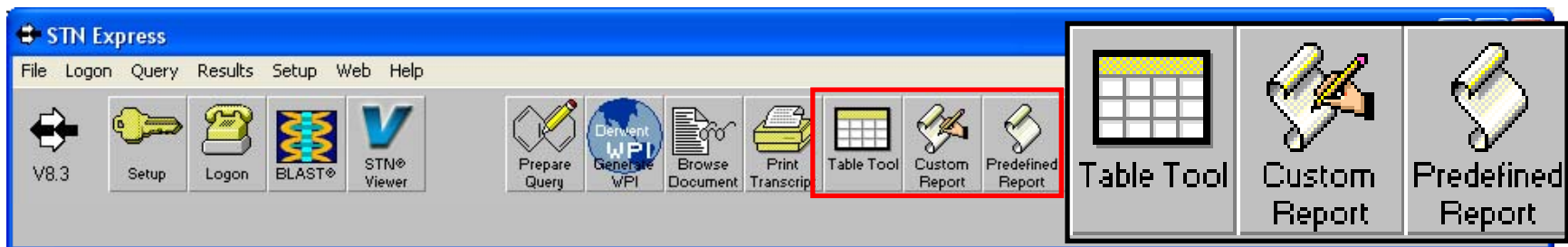
The 2nd hit is now SEQ
ID 26 from US5756708.

Agenda

- DGENE, PCTGEN, USGENE database content
- The 7 basic steps of BLAST
- BLAST and Patent Family SORT (FSORT)
- **Post-processing BLAST search results**
- Similarity searching GETSIM (FASTA)
- Offline BATCH search mode
- Sequence Code Match searching (GETSEQ)
- Multifile patent sequence search example

STN Express 8.x post-processing tools

- **Table Tool** to create tabulated results
 - Good for scanning/reviewing search results
- **Predefined Report Tool** for a report using a Standard Patent Record layout
 - Easy way to tidy-up your patent results for a client
- **Customized Report Tool** to control all options
 - e.g., fonts, cover page, which data fields to include



USGENE results may be tabulated using STN Express 8.x Table Tool

Search Question:

Find all relevant U.S. granted patent references, applied for prior to 2001, with sequences similar to the *Human osteoprotegerin (OPG) mRNA, complete CDS* (NCBI: U94332).

Human osteoprotegerin (OPG) mRNA, complete CDS (NCBI: U94332)

The screenshot shows the NCBI Sequence Viewer interface in Mozilla Firefox. The search criteria are set to 'Nucleotide' for 'U94332'. The 'Display' dropdown is set to 'FASTA', which is circled in red. The 'Range' is set from 'begin' to 'end'. A Notepad window titled 'U94332 - Notepad' is open, displaying the full nucleotide sequence in FASTA format. The sequence starts with >gi|2072184|gb|U94332.1|HSU94332 Human osteoprotegerin (OPG) mRNA, complete cds and ends with AATTGGCGAGATCCCATGGATGATAA.

```
>gi|2072184|gb|U94332.1|HSU94332 Human osteoprotegerin (OPG) mRNA, complete cds
GTATATATAACGTGATGAGCGTACGGGTGCGGAGACGCCACCGGAGCGCTCGCCCAGCCGCGYCTCCAAG
CCCCTGAGGTTTCCGGGACCACAATGAACAAGTTGCTGTGCTGCGCGCTCGTGTTCGGACATCTCCA
TTAAGTGGACCACCCAGGAAACGTTTCTCCAAAAGTACCTTCATTATGACGAAGAACCCTCATCAGT
GTTGTGTGACAAATGTCCTCCTGGTACCTACCTAAAACAACACTGTACAGCAAAAGTGGAAAGACCGT
GCCCCTTGCCCTGACCACTACTACACAGACAGCTGGCACACCAGTGCAGTGTCTATACTGCAGCCCG
TGTGCAAGGAGCTGCAGTACGTCAAGCAGGAGTGC AATCGCACCCACAAAGTGGAAAGACCGTGTGC
AGGGCGCTACCTTGAGATAGAGTTCTGCTTGAACAATAGGAGCTGCCCTGCTGCGAATGCAAGGA
GCTGGAACCCAGAGCGAAATACAGTTTGC AAAAGATGTCCAGATGGGTTTGGTGTGCAAGGAGTGC
CTAAAAGCACCCCTGTAGAAAACACACAAATTGCAGTGTCTTTGGTCTCCTGCTAACTCAGAAAGGAA
AACACACGACAAACATATGTCCGGAACAGTGAATCAACTCAAAAATGTGCTGGAACCCAGAGCGAAAT
GAGGAGGCATTCTCAGGTTTGTCTTCTACAAAAGTTTACGCCTAACTCAAAAGATGTTGGAATGTTG
ATTGTCCTGGCACCAAAAGTAAACGCAGAGAGTGTAGAGAGGATAAAAACGCTGTTGTTGTTGTTGTT
GACTTTCCAGCTGCTGAAATTATGGAACAATCAAAAACAAAAGCCCAAGATGTTGTTGTTGTTGTTG
GATATTGACCTCTGTGAAAACAGCGTGCAGCGGCACATTGGACATGCTAACTGTTGTTGTTGTTGTTG
GTAGCTTGATGGAAGCTTACCGGGAAGAAAAGTGGGAGCAGAAGACATTTGTTGTTGTTGTTGTTGTT
CAAAACCCAGTACCTGAAAGCTGCTCAGTTTGTGGCGAATAAAAAGTGGTGGTGGTGGTGGTGGTGGT
AAGGGCCTAATGCACGCCTAAAAGCACTCAAAAGCAGTACCACCTTTCCCAAGTGGTGGTGGTGGTGG
AGAAGACCATCAGGTTCTTCACAGCTTACAATGTACAAATTGTATCAGGTTGTTGTTGTTGTTGTTG
AGGTAACCAAGTCCAATCAGTAAAAAATAGCTGCTTATAACTGGAAAATGTTGTTGTTGTTGTTGTTG
AATTGGCGAGATCCCATGGATGATAA
```

Ensure you capture your STN session

STN Online and Results - [STN-C]

File Edit Online Query Results Preferences! Web Window Help

NEWS 19 APR 04 STN AnaVist, Version 1, to be discontinued

NEWS with new

NEWS hanced

NEWS nts

NEWS V8.3,

NEWS FEBRUARY 2008

NEWS ila

NEWS imp

NEWS ne

Enter

spec

AT

ag

re

of

re

STN Customer

use to scientific

r implementation

hibited and may

es.

***** STN Columbus *****

FILE 'HOME' ENTERED AT 15:22:58 ON 13 MAY 2008

=>

Discover! Transcript: HOME INS Hold Off Print Off Online 00:01:57

Record your session as a Transcript (.TRN) file or as an RTF file.

SAVE, UPLOAD and VERIFY

```
=> FILE USGENE
```

```
=> UPL R BLAST
```

These commands are automatically run by the STN Express Sequence Query Upload wizard (slide 46).

```
UPLOAD SUCCESSFULLY COMPLETED
```

```
L1 GENERATED
```

```
=> D L1 LQUE
```

```
L1 ANSWER 1 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
LQUE gtatatataacgtgatgagcgtacgggtgcggagacgcaccggagcgctcgcccagccg
ccgctccaagcccctgaggtttccggggaccacaatgaacaagttgctgtgctgcgcgc
tcgtgtttctggacatctccattaagtggaccaccaggaacgtttcctccaaagtac
. . . .
tggccattgagctgtttcctcacaattggcgagatcccatggatgataa
```

```
=>
```

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

RUN the USGENE BLAST search

=> **FILE USGENE**

FILE 'USGENE' ENTERED AT 19:53:46 ON 13
COPYRIGHT (C) 2008 SEQUENCEBASE CORP

FILE LAST UPDATED: 9 MAY 2008 <20080509/UP>
MOST RECENT PUBLICATION DATE: 8 MAY 2008 <20080508/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 /SQN -F F**

BLAST Version 2.2

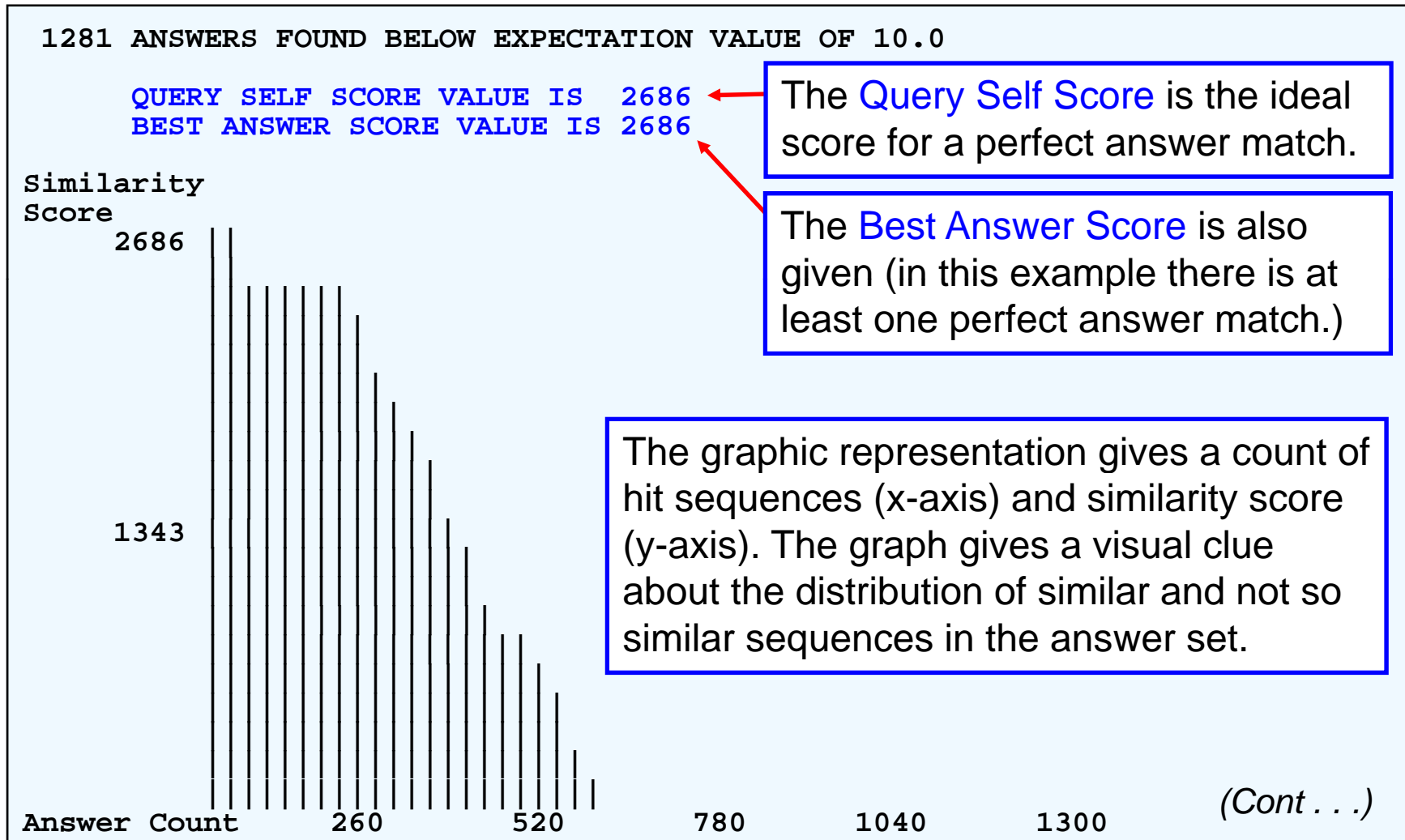
The BLAST software is used herein with permission of the
National Center for Biotechnology Information (NCBI) of
the National Library of Medicine (NLM). See also,

BLAST SEARCHING

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

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

Decide how many answers to keep



SORT by SCORE descending

```
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 ? : ALL
```

```
L2 RUN STATEMENT CREATED
```

```
L2 1281 GTATATATAACGTGATGAGCGTACGGGTC
```

```
. . . .
```

```
TGATAA/SQN.-F F
```

In this example, ALL answers have been kept (L2).

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

```
=> SET SFIELDS BI ECLM PERM
```

```
SET COMMAND COMPLETED
```

Use SET SFIELDS to change the USGENE default search index.

```
=> S L2 AND (OSTEO? OR BONE#) AND GRANTED/SSO AND AY<2001
```

```
L3 310 L2 AND (OSTEO#/BI,ECLM OR BONE#/BI,ECLM) AND GRANTED/SSO AND  
AY<2001
```

```
=> SOR SCORE D
```

```
PROCESSING COMPLETED FOR L3
```

```
L4 310 SOR L3 SCORE D
```

After refining using date and text terms remember to SOR SCORE D.

USGENE answer sets may be grouped by source publications using Family SORT (FSORT)

```
=> FSORT L4
```

```
. . . .
```

```
L5          310 FSO L4
```

```
14 Multi-record Families
```

```
Answers 1-309
```

```
Family 1
```

```
Answers 1-11
```

```
Family 2
```

```
Answers 12-22
```

```
Family 3
```

```
Answers 23-33
```

```
Family 4
```

```
Answers 34-44
```

```
Family 5
```

```
Answers 45-71
```

```
Family 6
```

```
Answers 72-83
```

```
Family 7
```

```
Answers 84-118
```

```
Family 8
```

```
Answers 119-179
```

```
Family 9
```

```
Answers 180-240
```

```
Family 10
```

```
Answers 241-301
```

```
Family 11
```

```
Answers 302-303
```

```
Family 12
```

```
Answers 304-305
```

```
Family 13
```

```
Answers 306-307
```

```
Family 14
```

```
Answers 308-309
```

```
1 Individual Record
```

```
Answer 310
```

```
0 Non-patent Records
```

The 310 sequence hits belong to 14 multi-hit and 1 individual-hit source publications.

Reviewing the SCORE display can be one way to identify answers of interest

=> D PFAM=1- SCORE

The SCORE for the best answer from each family.

L5 ANSWER 1 OF 310 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN FAMILY1
SCORE 2686 100% of query self score 2686

L5 ANSWER 12 OF 310 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN FAMILY2
SCORE 2686 100% of query self score 2686

L5 ANSWER 23 OF 310 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN FAMILY3
SCORE 2686 100% of query self score 2686

L5 ANSWER 34 OF 310 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN FAMILY4
SCORE 2686 100% of query self score 2686

. . . .

L5 ANSWER 241 OF 310 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STNFAMILY10
SCORE 2375 88% of query self score 2686

L5 ANSWER 302 OF 310 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STNFAMILY11
SCORE 46 1% of query self score 2686

. . . .

L5 ANSWER 310 OF 310 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STNFAMILY12
SCORE 2686

100% of query self score 2686

Note: The FSORT individual-hit record also has a top score.

Use the PFAM feature to display selective records from an FSORT L-number

=> D PFAM=1-10,15

← Using the default display format (see slide 79).

```

L5      ANSWER 1 OF 310 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN FAMILY1
AN      6284740.5  cDNA          USGENE
TI      Osteoprotegerin (Patent)
IN      Boyle William J. (Moorpark, CA); Lacey David L. (Thousand Oaks, CA);
        Calzone Frank J. (Westlake Village, CA); . . . .
PA      Amgen Inc (Thousand Oaks CA)
PI      US 6284740          B1      20010904
AI      US 1997-974186          19971118
AB      The present invention discloses a novel secreted polypeptide, termed
        Osteoprotegerin, which is a member of the tumor necrosis . . . .
ECLM    US6284740 B1: What is claimed is:1. A method of increasing levels of
        osteoprotegerin in a mammal comprising administering to . . . .
ORGN    not provided
SQL     1355
SCORE   2686
BLASTALIGN
        Query = 1355 letters
        Length = 1355
        Score = 2686 bits (1355), Expect = 0.0
        Identities = 1355/1355 (100%)
        Strand = Plus / Plus

        Query: 1      gtatatataacgtgatgagcgtacgggtgcggagacgcaccggagcgctcgcccagccgc
                   |||
        Sbjct: 1      gtatatataacgtgatgagcgtacgggtgcggagacgcaccggagcgctcgcccagccgc
    
```

The top hit is SEQ ID 5 from US6284740.

After logging off from STN select the table tool from the main STN Express tool bar

The most recent Transcript is automatically selected.

STN Express

File Logon Query Results Setup Web Help

V8.3 Setup Logon BLAST® STN® Viewer

Prepare Query Derivent WPI Generate WPI Browse Document Print Transcript Table Tool Custom Report Predefined Report R₁₁ R-group Analysis Edit Transcript

Prefs Help Exit

Table Tool

Transcript(s)

Template

Content

Highlighting

Cover Page

Header/Footer

Fields

Statistics

All transcripts listed in the box below will be merged into your report or table. Select transcripts to add to this list by clicking the 'Browse to Add' button and locating the transcripts of interest.

Selected Transcripts :

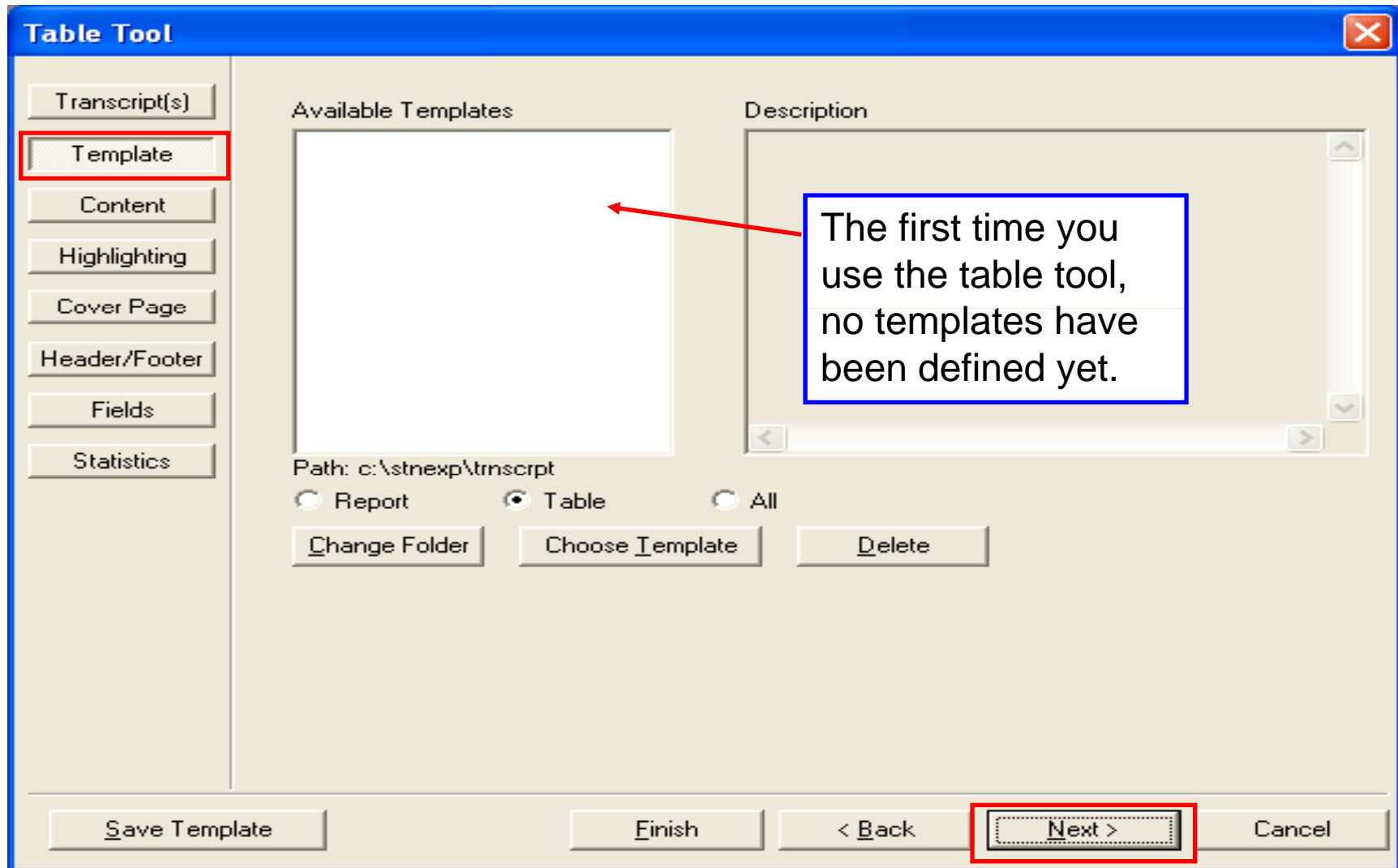
C:\STNEXP\Transcript\BLAST.DPG.trn

Browse to Add

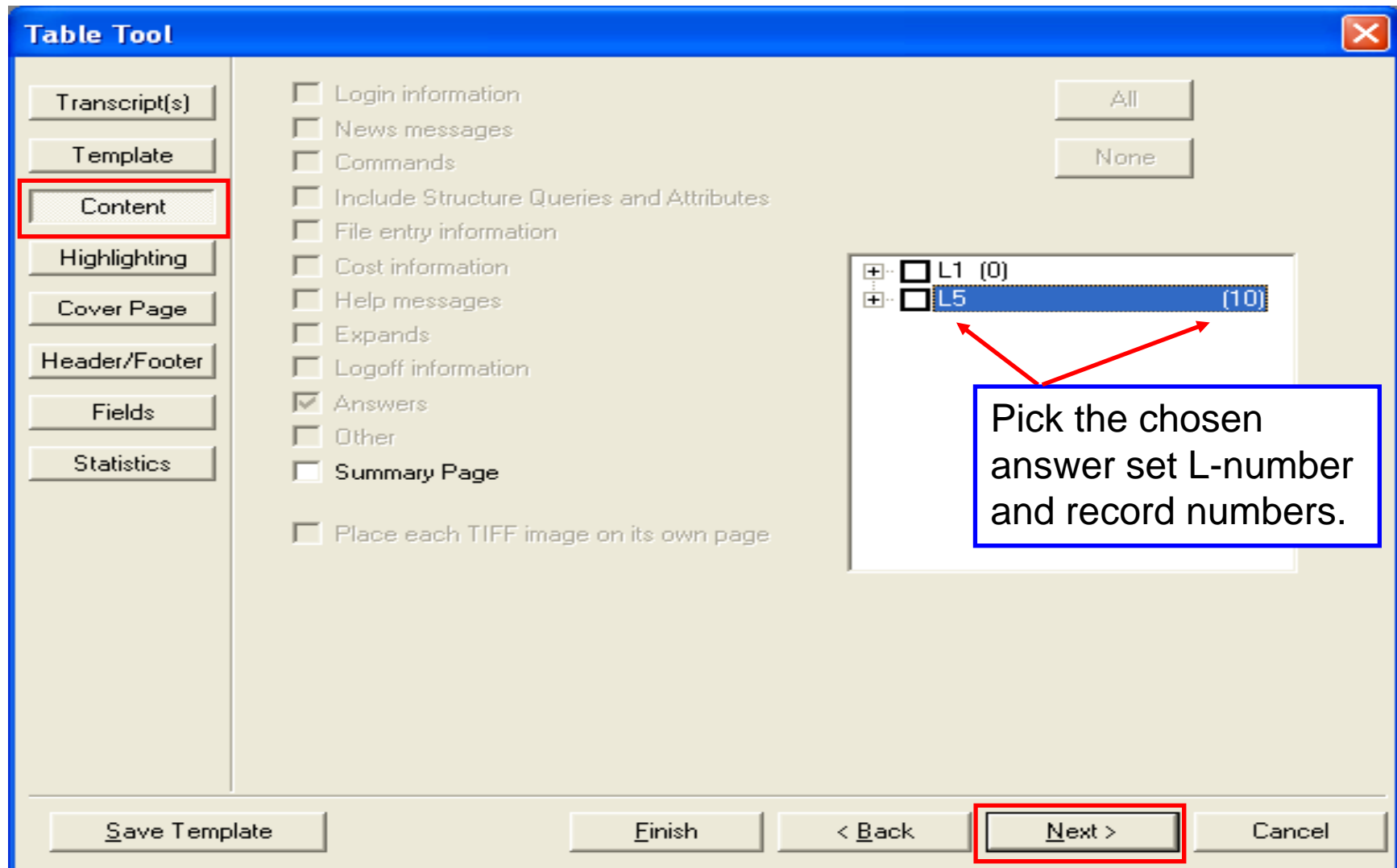
Remove

Save Template Finish < Back Next > Cancel

If available choose any template you have defined previously



Choose the L-number and records you wish to include in your tabular report



Set highlighting preferences

The screenshot shows the 'Table Tool' dialog box with the 'Highlighting' tab selected. The dialog has a sidebar on the left with buttons for 'Transcript(s)', 'Template', 'Content', 'Highlighting' (highlighted with a red box), 'Cover Page', 'Header/Footer', 'Fields', and 'Statistics'. The main area contains three sections:

- Highlight Hit Terms
Format: Red
 Don't Highlight the Following Hit Terms
- Highlight the Following Terms
Text: Amgen
Format: Lime
- Highlight Terms in this Text File
Format: Blue
Browse

A red arrow points from a text box to the 'Amgen' text in the second section. The text box contains the text: 'Extra terms that were not originally searched may be highlighted.' The 'Next >' button at the bottom right is also highlighted with a red box.

Set up report cover page

Table Tool

Transcript(s)
Template
Content
Highlighting
Cover Page
Header/Footer
Fields
Statistics

No Cover Page
 STN Express Cover Page Center
 Attach File(s)

Title
 Prepared by
 For (Client)
 Date of Search
 Strategy
 Update Info
 Cost
 Comments

| Date of Search | Strategy | Update Info | Cost | Comments |
|------------------|----------------|-------------|------|----------|
| Order | Title | Prepared By | | For |
| select the order | | | | |
| - | Title | | | |
| - | Prepared By | | | |
| - | For | | | |
| - | Date of Search | | | |
| - | Strategy | | | |
| - | Update Info | | | |
| - | Cost | | | |
| - | Comments | | | |

Select fields, fonts, colors, change field order, customize field names, and save templates

The screenshot shows the 'Table Tool' dialog box with the following components:

- Left Panel:** A vertical list of buttons: Transcript(s), Template, Content, Highlighting, Cover Page, Header/Footer, **Fields** (highlighted with a red box), and Statistics.
- Field Name List:** A list of available fields: Answer, Application Information, Cleaned CS, Comments, Copyright, Document Type, Patent Information, Row.
- Selected Fields List:** A list of fields currently selected, enclosed in a red box: Accession Number, Title, Patent Assignee, Inventor, Individual PI, Individual AI, Abstract, Exemplary Claim, BLAST Score, Sequence Length, BLAST Alignment, Organism.
- Buttons:** Insert >, Insert All >>, Rename, < Remove, << Remove All, Format, Change Order (up/down arrows).
- Options:** Remove duplicate fields within an answer, Ignore case when sorting, Autofit.
- Right Panel:** Individual AI | Individual PI tabs, Application Number, Kind Code, Date, All button, Number of PRAI rows per answer (radio buttons for All (maxm is 1) and First [] members of family), By country (US), Remove blank cells, Merge family members into a single row.
- Bottom Panel:** **Save Template** (highlighted with a red box), Finish, < Back, **Next >** (highlighted with a red box), Cancel.

Annotations:

- A blue box with the text "Choose fields, field order, formats and personalized names." has a red arrow pointing to the Selected Fields list.
- A blue box with the text "Save the template for future use." has a red arrow pointing to the Save Template button.

STN Express Table Tool output can be edited and adjusted as needed

The screenshot displays the 'STN Online and Results - [Table Output - BLAST OPG.tb]' window. The main table contains the following data:

| Accession Number | Title | Patent Assignee | Inventor | Abstract | Exemplary Claim | BLAST Score | Sequence Length | BLAST Alignment |
|-----------------------|--------------------------|------------------------------|---|--|--|-------------|-----------------|--|
| 6284740.5 cDNA USGENE | Osteoprotegerin (Patent) | Amgen Inc (Thousand Oaks CA) | Boyle William J. (Moorpark, CA); Lacey David L. (Thousand Oaks, CA); Calzone Frank J. (Westlake Village, CA); Chang Ming-Shi (Newbury Park, CA) | The present invention discloses a novel secreted polypeptide, termed Osteoprotegerin , which is a member of the tumor necrosis factor receptor superfamily and is involved in the regulation of bone metabolism. Also disclosed are nucleic acids encoding Osteoprotegerin , polypeptides, recombinant vectors and host cells for expression, antibodies which bind Osteoprotegerin , and pharmaceutical compositions. The polypeptides are used to treat bone diseases characterized by increased resorption such as osteoporosis . | US6284740 B1: What is claimed is:1. A method of increasing levels of osteoprotegerin in a mammal comprising administering to the mammal a nucleic acid encoding osteoprotegerin , wherein the administration results in an increase in the level of osteoprotegerin and wherein the increase in the level of osteoprotegerin in the mammal results in increased bone density. | 2686 | 1355 | Query = 1355 letters Length = 1355 Score = 2686 bits (1355) Identities = 1355/1355 (100%) Strand = Plus / Plus Query: 1 qtatatataacgtg Sbjct: 1 qtatatataacgtg Query: 61 cactccaagccctt Sbjct: 61 cactccaagccctt Query: 121 gtgtttctggacat Sbjct: 121 gtgtttctggacat Query: 181 cattatgacgaaga Sbjct: 181 cattatgacgaaga Query: 241 ctaaaacaacactg Sbjct: 241 ctaaaacaacactg Query: 301 tacacagacagctg Sbjct: 301 tacacagacagctg Query: 361 ctgacgacagctg Sbjct: 361 ctgacgacagctg Query: 421 gggcgtacacttga Sbjct: 421 gggcgtacacttga Query: 481 gtggtgcaagctgg Sbjct: 481 gtggtgcaagctgg |

An 'Edit' menu is overlaid on the table, showing options such as 'Undo', 'Cut', 'Copy', 'Paste', 'Clear', 'Select All', 'Create Query File', 'Show Clipboard', 'Install Dictionary', 'Edit Dictionary', 'Edit Table', 'Autofit', 'Find', 'Replace', 'Goto', 'Next Answer', 'Previous Answer', 'Next Marked Answer', 'Previous Marked Answer', 'Transcript(s)', 'Template', 'Content', 'Highlighting', 'Cover Page', 'Header/Footer', and 'Fields'. The 'Fields' option is highlighted with a red box.

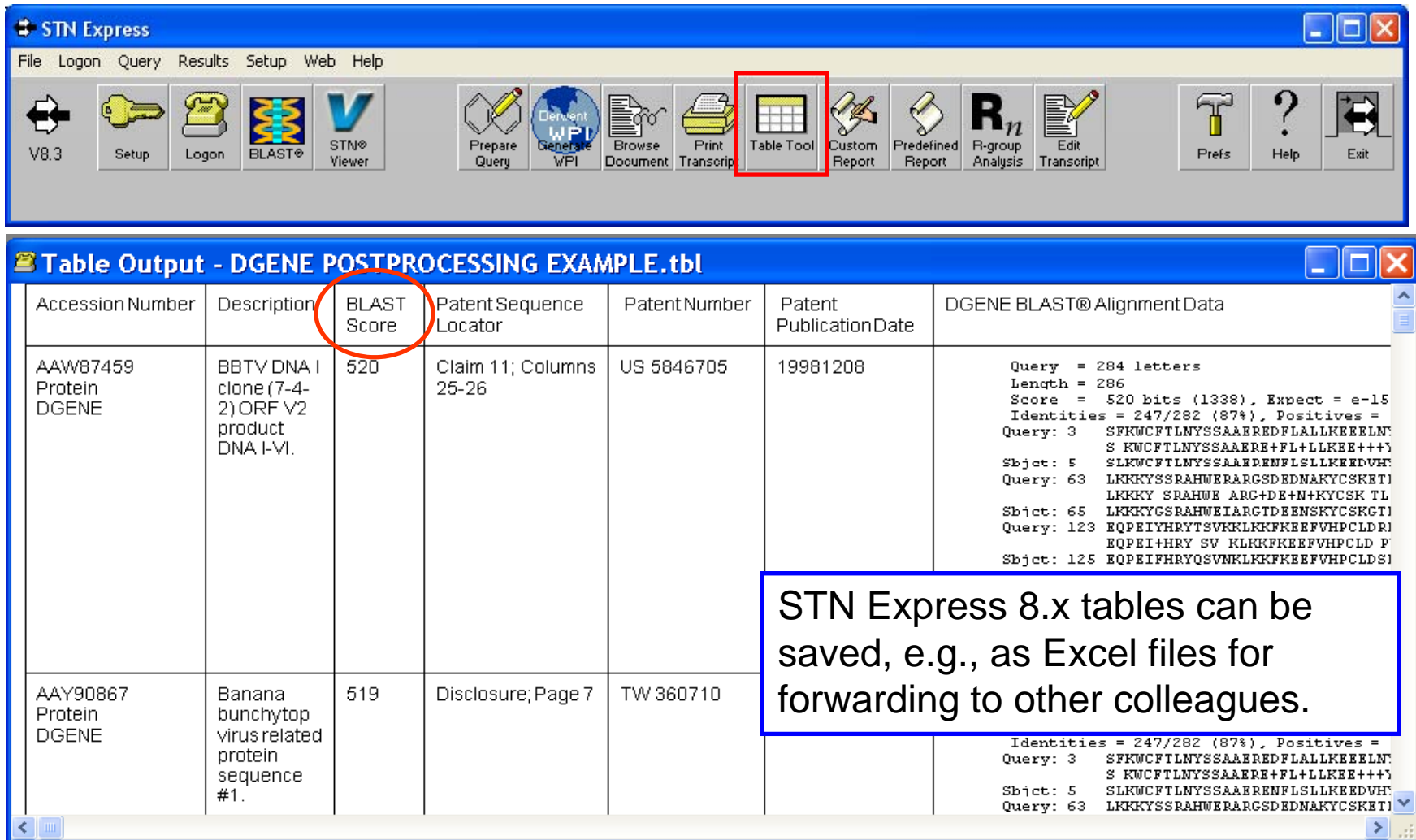
If needed, go back and edit choices fields, formats, etc.

STN Express Table Tool output can be edited, adjusted and saved in Excel® format

| A2 | A | B | C | D | E | F | G | H | I |
|----|-----------------------------|-----------------------------|------------------------------------|---|---|--|-------------|-----------------|--|
| 1 | Accession Number | Title | Patent Assignee | Inventor | Abstract | Exemplary Claim | BLAST Score | Sequence Length | BLAST Alignment |
| | 6284740.5 cDNA USGENE | Osteoprotegerin (Patent) | Amgen Inc (Thousand Oaks CA) | Boyle William J. (Moorpark, CA); Lacey David L. (Thousand Oaks, CA); Calzone Frank J. (Westlake Village, CA); Chang Ming-Shi (Newbury Park, CA) | The present invention discloses a novel secreted polypeptide, termed Osteoprotegerin , which is a member of the tumor necrosis factor receptor superfamily and is involved in the regulation of bone metabolism. Also disclosed are nucleic acids encoding Osteoprotegerin , polypeptides, recombinant vectors and host cells for expression, antibodies which bind Osteoprotegerin , and pharmaceutical compositions. The polypeptides are used to treat bone diseases characterized by increased resorption such as osteoporosis . | US6284740 B1: What is claimed is:1. A method of increasing levels of osteoprotegerin in a mammal comprising administering to the mammal a nucleic acid encoding osteoprotegerin , wherein the administration results in an increase in the level of osteoprotegerin and wherein the increase in the level of osteoprotegerin in the mammal results in increased bone density. | 2686 | 1355 | Query = 1355 letters Length = 1355 Score = 2686 bits (1355), Expect = 0.0 Identities = 1355/1355 (100%) Strand = Plus / Plus Query: 1 gtatatataaacgtgatgagcgtacgggtc Sbjct: 1 gtatatataaacgtgatgagcgtacgggtc Query: 61 cgctccaagccctgaggttccggggac Sbjct: 61 cgctccaagccctgaggttccggggac Query: 121 gtgttctggacatctccattaagtggac Sbjct: 121 gtgttctggacatctccattaagtggac Query: 181 cattatgacgaagaacctctcatcagct Sbjct: 181 cattatgacgaagaacctctcatcagct Query: 241 ctaaaacaacactgtacagcaaagtggac Sbjct: 241 ctaaaacaacactgtacagcaaagtggac Query: 301 tacacagacagctggcacaccagtgacg Sbjct: 301 tacacagacagctggcacaccagtgacg |
| | 6284728.5 cDNA USGENE | Osteoprotegerin (Patent) | Amgen Inc (Thousand Oaks CA) | Boyle William J. (Moorpark, CA); Lacey David L. (Thousand Oaks, CA); Calzone Frank J. (Westlake Village, CA); Chang Ming-Shi (Newbury Park, CA) | The present invention discloses a novel secreted polypeptide, termed Osteoprotegerin , which is a member of the tumor necrosis factor receptor superfamily and is involved in the regulation of bone metabolism. Also disclosed are nucleic acids encoding Osteoprotegerin , polypeptides, recombinant vectors and host cells for expression, antibodies which bind Osteoprotegerin , and pharmaceutical compositions. The polypeptides are used to treat bone diseases | US6284728 B1: What is claimed is:1. An isolated polypeptide consisting of the amino acid sequence as shown in FIG. 9B (SEQ ID NO:6) from residues 22 to 401 inclusive. | 2686 | 1355 | Query = 1355 letters Length = 1355 Score = 2686 bits (1355), Expect = 0.0 Identities = 1355/1355 (100%) Strand = Plus / Plus Query: 1 gtatatataaacgtgatgagcgtacgggtc Sbjct: 1 gtatatataaacgtgatgagcgtacgggtc Query: 61 cgctccaagccctgaggttccggggac Sbjct: 61 cgctccaagccctgaggttccggggac Query: 121 gtgttctggacatctccattaagtggac Sbjct: 121 gtgttctggacatctccattaagtggac Query: 181 cattatgacgaagaacctctcatcagct Sbjct: 181 cattatgacgaagaacctctcatcagct Query: 241 ctaaaacaacactgtacagcaaagtggac Sbjct: 241 ctaaaacaacactgtacagcaaagtggac |

Note: See separate appendix for the full printout of this table.

DGENE and PCTGEN results also can be post-processed into tables using STN Express 8.x



STN Express 8.x interface showing a table of DGENE postprocessing results. The 'Table Tool' icon in the toolbar is highlighted with a red box. A blue box highlights the 'BLAST Score' column in the table. A text box on the right states: 'STN Express 8.x tables can be saved, e.g., as Excel files for forwarding to other colleagues.'

| Accession Number | Description | BLAST Score | Patent Sequence Locator | Patent Number | Patent Publication Date | DGENE BLAST@ Alignment Data |
|------------------------------|---|-------------|-------------------------|---------------|-------------------------|---|
| AAW87459 Protein DGENE | BBTV DNA I clone (7-4-2) ORF V2 product DNA I-VI. | 520 | Claim 11; Columns 25-26 | US 5846705 | 19981208 | Query = 284 letters Length = 286 Score = 520 bits (1338), Expect = e-15 Identities = 247/282 (87%), Positives = Query: 3 SFKWCFTLNYSSAAEREDFLALLKKEELN S KWCFTLNYSSAAERE+FL+LLKKE+++ Sbjct: 5 SLKWCFTLNYSSAAERENFLSLLKKEEDVH Query: 63 LKKKYSSRAHWERARGSDEDNAKYCSKETI LKKKY SRAHWE ARG+DE+N+KYCSK TL Sbjct: 65 LKKKYSSRAHWELARGTDEENSKYCSKGTI Query: 123 EQPEIYHRYTSVKKLKKFKKEEFVHPCLDRI EQPEI+HRY SV KLEKFKKEEFVHPCLD P Sbjct: 125 EQPEIFHRYQSVNKLKKFKKEEFVHPCLDRI |
| AAY90867 Protein DGENE | Banana bunchytop virus related protein sequence #1. | 519 | Disclosure; Page 7 | TW 360710 | | Identities = 247/282 (87%), Positives = Query: 3 SFKWCFTLNYSSAAEREDFLALLKKEELN S KWCFTLNYSSAAERE+FL+LLKKE+++ Sbjct: 5 SLKWCFTLNYSSAAERENFLSLLKKEEDVH Query: 63 LKKKYSSRAHWERARGSDEDNAKYCSKETI |

Agenda

- DGENE, PCTGEN, USGENE database content
- The 7 basic steps of BLAST
- BLAST and Patent Family SORT (FSORT)
- Post-processing BLAST search results
- **Similarity searching GETSIM (FASTA)**
- **Offline BATCH search mode**
- Sequence Code Match searching (GETSEQ)
- Multifile patent sequence search example

Similarity searching in USGENE using FASTA-based RUN GETSIM

- GETSIM was originally developed by FIZ Karlsruhe for DGENE, and it has since been implemented in both PCTGEN and USGENE
- It is based on the industry standard FASTA methodology, and offers the same basic search modes as BLAST (/SQP, /SQN and /TSQN)
- Since GETSIM requires more computational time than BLAST, it is usually a good idea to make use of the offline BATCH search mode

General differences between FASTA (GETSIM) and BLAST algorithms

| BLAST | FASTA (GETSIM) |
|---|---|
| Faster than FASTA | Slower than BLAST |
| Equivalent for highly similar sequences | |
| Misses some less similar sequences | Better for less similar sequences |
| Comparison of shorter sequence parts | Comparison of entire sequence length |
| Less sensitive when using default settings | More sensitive, misses less homologs |
| Less separation between true homologs and random hits | More separation between true homologs and random hits |
| Calculates probabilities | Calculates significance “on the fly” from the given dataset |

RUN GETSIM command syntax

Similarity Searching with GETSIM (protein/polypeptides)

=> RUN GETSIM L1 (sequence or L-number)

/SQP (protein) (default)

BATCH (offline)

ALERT (current awareness)

Note: Unlike RUN BLAST, RUN GETSIM does not have any user-defined advanced options to consider. The optimum FASTA search settings are selected automatically by the GETSIM software, depending upon the sequence query.

RUN GETSIM command syntax

Similarity Searching with GETSIM (nucleotide sequences)

=> **RUN GETSIM L1** (sequence or L-number)
/SQN (nucleotide)

SIN (single strand) (default)

COM (complementary strand)

BOTH (both strands)

BATCH (offline)

ALERT (current awareness)

Note: To automatically search the nucleotide sequence *and* its complement specify **BOTH**:

=> **RUN GETSIM . . . /SQN BOTH**

GETSIM and BLAST similarity searches can both be run offline in BATCH search mode

- Multiple BATCH requests may be queued, to run sequentially one after another
 - A maximum of 16 requests can be queued per STN Login ID
- BATCH request results may be collected in an online session up to 3 months from initiation
 - Results that have been collected may be re-retrieved multiple times at no additional cost, up to 8 days from the initial retrieval
 - For example, multiple times each at a different score percent (%)
- BATCH is most useful for GETSIM queries, as these can take considerable computational time when run online
 - Also a higher query length limit of 2,000 characters is permitted

Similarity searching in USGENE using GETSIM in offline BATCH mode

Search Question:

Find sequences in U.S. published applications and patents which are similar to this specific cholinesterase protein (NCBI: AAA98113):

```
MPSSVSWGILLLAGLCCLVPVSLAEDPQGDAAQKTDTSHHQDHPNFKITPN  
LAEFAFSLYRQLASTNIFFSVSIATAFAMLSLGTKADTHDEILEGLNFNLTE  
IPEAQIHEGFQELLRTLNPDSQLQLTTGNGLFLSEGLKLVDFLEDVKKLYH  
SEAFVNFVGDTEEAQKQINDYVEKGTQGKIVDLVKELDRDTVFALVNYIFFKG  
KWERPFVVDTEEEDFHVDQVTTVKVPMKRLGMFNIQHCKKLSWVLLMKYL  
GNATAIFFLPDEGKLQHLENELTHDIITKFLNEDRRSASLHLPKLSITGTID  
LKSVLGQLGITKVFVSGADLSGVTEEAFLKLSKAVHKAVLTIDEKGTAAAGAM  
FLEAIPMSIPPEVKFNKPFVFLMIEQNTKSPLFMGKVVNPTQK
```

SAVE, UPLOAD, and VERIFY the query text file for the GETSIM BATCH search

```
=> FILE USGENE
=> UPL R BLAST
```

These commands are automatically run by the STN Express Sequence Query Upload wizard (slide 46).

```
UPLOAD SUCCESSFULLY COMPLETED
```

```
L1 GENERATED
```

```
=> D L1 LQUE
```

Verify the sequence was uploaded successfully with **D LQUE**.

```
L1 ANSWER 1 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
LQUE MPSSVSWGILLLAGLCCLVPVSLAEDPQGDAAQKTDTSHHDQDHPTFNKITPNLAE
FAFSLYRQLASTNIFFSVSIATAFAMLSLGTKADTHDEILEGLNFNLTEIPEAQI
HEGFQELLRTLNQPSQLQLTTGNGLFLSEGLKLVDFLEDVKKLYHSEAFTVNFV
DTEEAKKQINDYVEKGTQGKIVDLVKELDRDTVFALVNYIFFKWKWERPFVVDTE
EEDFHVDQVTTVKVPMKRLGMFNIQHCKKLSWVLLMKYLGNATAIFFLPDEGKL
QHLENELTHDIITKFLNEDRRSASLHLPKLSITGTYDLKSVLGQLGITKVFVSN
DLSGVTEEA KPFVF
LMIEQNTKSI
```

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

```
=>
```

RUN the GETSIM search in BATCH mode

=> FILE USGENE

FILE 'USGENE' ENTERED AT 17:32:27 ON 08
COPYRIGHT (C) 2008 SEQUENCEBASE CORP

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

FILE LAST UPDATED: 6 JUN 2008 <20080606/UP>
MOST RECENT PUBLICATION DATE: 5 JUN 2008 <20080605/PD>

FILE COVERS 1982 TO DATE

Add BATCH for BATCH mode.

=> RUN GETSIM L1 /SQP BATCH

PLEASE ENTER BATCH IDENTIFIER (MAX. 8 CHARS): EXAMPLE4

Name the
BATCH search.

RUN GETSIM AT 17:32:48 ON 08 JUN 2008
COPYRIGHT (C) 2008 FIZ KARLSRUHE GMBH

PREVIOUS BATCH REQUEST STILL RUNNING
BATCH PROCESSING QUEUED FOR EXAMPLE4

In this example, there is already
a BATCH search running, so
this request has been Queued.

=> LOG H

SESSION WILL BE HELD FOR 120 MINUTES
STN INTERNATIONAL SESSION SUSPENDED AT 17:33:06 ON 08 JUN 2008

Use RUN GETBATCH to retrieve and manage the results of BATCH searches

```
* * * * * RECONNECTED TO STN INTERNATIONAL * *  
SESSION RESUMED IN FILE 'USGENE' AT 18:06:27 ON 0  
FILE 'USGENE' ENTERED AT 18:06:27 ON 08 JUN 2008
```

Login within 2 hours if you want to reconnect to your previous STN session.

```
=> RUN GETBATCH
```

```
Please enter your batch identifier  
or enter # for batch id list  
or enter * for batch id at top of list  
or enter - before batch id to delete  
or enter . for (end)
```

Enter # for a BATCH ID list.

```
BATCH REQUEST: #
```

```
Batch result files remaining:
```

```
EXAMPLE1 Retrieved (blast)  
EXAMPLE2 Retrieved (getsim)  
EXAMPLE3 Completed (blast)  
EXAMPLE4 Completed (getsim)
```

BATCH results file status can be: Queued, Running, Completed or Retrieved.

```
-----  
Please enter your batch identifier  
or enter # for batch id list  
or enter * for batch id at top of list  
or enter - before batch id to delete  
or enter . for (end)
```

Enter the name of the BATCH search results to retrieve.

```
BATCH REQUEST: EXAMPLE4
```

Decide how many answers to keep

New!

2194 ANSWERS FOUND ABOVE A THRESHOLD OF 172

QUERY SELF SCORE VALUE IS 2692
BEST ANSWER SCORE VALUE IS 2692

The **Query Self Score** is the ideal score for a perfect answer match.

The **Best Answer Score** is also given (in this example there is at least one perfect answer match.)

Similarity
Score

2692

1346

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

Answer Count

440

880

1320

1760

2200

(Cont...)

After BATCH collection all search, sort and display options are the same as in online search mode

New!

```
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 ? :80%
```

In this example, 80% of the Query Self Score is used to select out just the most relevant results (L2).

```
L2 RUN STATEMENT CREATED
L2 142 MPSSVSWGILLLAGLCCLVPVSLAE
```

```
TPNLAEFAFSLYRQLASTNIFPVSIAATAFAMLSLGTKADTHDEILEGL
NFNLTEIPEAQIHEGFQELLRTLNPDS
LEDVKKLYHSEAFTVNFQDTEEAKKQIN
VFALVNYIFFKGKWERPFVVDTEEEDF
HCKKLSSWVLLMKYLGNATAIFFLPDEC
RRSASLHLPKLSITGTDLKSVLGLGI
KAVHKAVLTIDEKGTAAAGAMFLEAIPM
SPLFMGKVVNPTQK/SQP
```

Reminder: BATCH results that have been collected, may be re-retrieved multiple times at no additional cost, up to 8 days from the initial retrieval.

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

```
=> SOR SCORE D
```

```
PROCESSING COMPLETED FOR L2
L3 142 SOR L2 SCORE D
```

As with a BLAST search, the initial GETSIM search answer set should be sorted by similarity score descending, to bring the best answers to the top.

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

=> D L3 TRI ORGN SCORE ALIGN 1-142; FILE STNGUIDE

L3 ANSWER 1 OF 142 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
TI Inhibitors of serine protease activity, methods and compositions for
treatment of viral infections (Patent)

MTY Protein

SQL 414

SCORE 2692

100% of query self score 2692

ORGN Unknown

ALIGN Smith-Waterman score: 2692

414 aa overlap starting at 1

mpssvswgillllaglcclvpvs laedpqqdaaqktdtshhdqdhptfnkitpnlaefafs

.....

mpssvswgillllaglcclvpvs laedpqqdaaqktdtshhdqdhptfnkitpnlaefafs

. . . .

L3 ANSWER 142 OF 142 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN

TI Anti-Cd28 Antibody (Published Application)

MTY Protein

SQL 669

SCORE 2344 87% of query self score 2692

ORGN ARTIFICIAL SEQUENCE

ALIGN Smith-Waterman score: 2344

372 aa overlap starting at 277

fnkitpnlaefafslyrqla____stniffspvsiatat

.....

fnkitpnlaefafslyrqlahqsnstniffspvsiatat

. . . .

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

The GETSIM ALIGN display:

- First line: Portion of query with similarity
- Second line: Similarity (identical- 2 dots, no match-blank, one dot- family match)
- Third line: Portion of retrieved sequence with similarity

Display selected USGENE answers in a preferred bibliographic format

```
=> D L3 BIB AB ECLM ORGN SQL SCORE ALIGN
```

```
L3 ANSWER 1 OF 142 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
AN 6849605.19 Protein USGENE
TI Inhibitors of serine protease activity, methods and compositions for
   treatment of viral infections (Patent)
IN Shapiro Leland (Denver, CO)
PA The Trustees of University Technology Corporation(Boulder CO)
PI US 6849605 B1 20050201
AI US 2000-518098 20000303
DT Patent
AB A novel method of treating and preventing viral infections is
   provided. In particular a method of blocking viral replication is
   facilitated by a serine proteolytic (SP) inhibitor. The method
   consists of administering to a subject suffering from viral infection
   a therapeutically effective amount of . . .
ECLM US6849605 B1: What is claimed is:1. A method for inhibiting human
     immunodeficiency virus (HIV) replication in a patient harboring said
     HIV comprising administering to the patient a combination
     comprising:at least one first compound exhibiting al-antitrypsin
     (AAT)-like protease inhibiting activity, wherein said compound . . .
ORGN Unknown
SQL 414
SCORE 2692 100% of query self score 2692
ALIGN Smith-Waterman score: 2692
      414 aa overlap starting at 1
      mpssvswgillllagllcclvpvsllaedpqqda
      ::::::::::::::::::::::::::::::::::::
      mpssvswgillllagllcclvpvsllaedpqqdaa qktatcsmnqqanptlnktpnaeraia . . .
```

USGENE records can be displayed in a wide variety of customized formats.

100% of query self score 2692

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

Sequence code match (SCM) searching in USGENE using RUN GETSEQ

- GETSEQ is designed to retrieve either exact matches to a sequence query or answers with conservative variation using special symbols
- It can also be used to retrieve exact length matches or subsequence hits, e.g. where the query is a small part of a larger hit sequence
- GETSEQ can prove to be a fast, precise and effective alternative to BLAST for very short sequence queries, e.g., DNA probes and primers
- Remember that an SCM search may also be run in REGISTRY, but the SEARCH (=> S) command is used instead of RUN GETSEQ

Agenda

- DGENE, PCTGEN, USGENE database content
- The 7 basic steps of BLAST
- BLAST and Patent Family SORT (FSORT)
- Post-processing BLAST search results
- Similarity searching GETSIM (FASTA)
- Offline BATCH search mode
- **Sequence Code Match searching (GETSEQ)**
- Multifile patent sequence search example

RUN GETSEQ command syntax

Sequence Code Match (SCM) searching with GETSEQ

=> **RUN GETSEQ L1** (sequence or query L-number)
/SQEP (**exact protein**) (**default**)
/SQEFP (exact family protein)
/SQSP (subsequence protein)
/SQSFP (subsequence family protein)
/SQEN (exact nucleotide)
/SQSN (subsequence nucleotide)

Reminder: USGENE, DGENE and PCTGEN all use the same search command for SCM: **RUN GETSEQ**.

EXACT (/SQEN) and SUBSEQUENCE (/SQSN) nucleic acid searching

```
=> RUN GETSEQ GCCGCCGT/SQEN
L1 RUN STATEMENT CREATED
L1 2 GCCGCCGT/SQEN
```

```
=> D L1 1 SEQ SQL
L1 ANSWER 1 OF 2 USGENE COPYRIGHT 2008
```

```
SEQ 1 gccgccgt
=====
HITS AT: 1-8
SQL 8
```

The SEQ display in USGENE shows the entire sequence with the hit nucleic acids underlined and identified by "HITS AT".

```
=> RUN GETSEQ ACCCTGCAAATAGCA/SQSN
L2 RUN STATEMENT CREATED
L2 49 ACCCTGCAAATAGCA/SQSN
```

A **SUBSEQUENCE** search also includes answers which are longer than the query sequence.

```
=> D L2 30 SEQ SQL
L2 ANSWER 30 OF 49 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
SEQ 1 tgtagttcat tatcatcttt gtcacagct gaagatgaaa taggatgtaa
51 tcagacgaca caggaagcag attctgctaa taccctgcaa atagcaga
=====
HITS AT: 82-96
SQL 98
```

EXACT (/SQEP) and SUBSEQUENCE (/SQSP) protein searching

```
=> RUN GETSEQ SMAEP/SQEP
```

```
L3 RUN STATEMENT CREATED
```

```
L3 3 SMAEP/SQEP
```

```
=> D L3 1 SQL SEQ
```

```
L3 ANSWER 1 OF 3 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
```

```
SQL 5
```

```
SEQ 1 smaep
```

```
=====
```

```
HITS AT: 1-5
```

```
=> RUN GETSEQ KGPSYSLR/SQSP
```

```
L4 RUN STATEMENT CREATED
```

```
L4 102 KGPSYSLR/SQSP
```

```
=> D L4 11 SQL SEQ
```

```
L4 ANSWER 11 OF 102 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
```

```
SQL 19
```

```
SEQ 1 kgpsyslrst tmmirpldf
```

```
=====
```

```
HITS AT: 1-8
```

In all sequence databases, the typed order of the display fields will be the order that the fields are displayed.

A **SUBSEQUENCE** search also includes answers which are longer than the query sequence.

EXACT (/SQEFP) and SUBSEQUENCE (/SQSFP) FAMILY protein searching

```
=> RUN GETSEQ SMAEP/SQEFP
L5 RUN STATEMENT CREATED
L5 23 SMAEP/SQEFP
```

SMAEP/SQEP retrieved 3 records (L3).
SMAEP/SQEFP retrieved 23 records.

```
=> D L5 2-3 SQL SEQ
L5 ANSWER 2 OF 23 USGENE COPYRIGHT 2008 SE
```

```
SQL 5
SEQ 1 gites
=====
```

HITS AT: 1-5

Possible amino acid family substitutions for SMAEP:

| S | M | A | E | P |
|---|---|---|---|---|
| P | I | G | Q | A |
| A | L | T | N | G |
| G | V | P | D | S |
| T | | S | B | T |

```
=> RUN GETSEQ KGPSYSLR/SQSFP
L6 RUN STATEMENT CREATED
L6 2384 KGPSYSLR/SQSFP
```

KGPSYSLR/SQSP retrieved 102 records (L4).
KGPSYSLR/SQSFP retrieved 2384 records.

```
=> D L6 73 SEQ SQL
L6 ANSWER 73 OF 2384 USGENE C
```

```
SQL 43
SEQ 1 hfrgkfcgki apppvvssgp flfikfvtsy ethgagfsir yei
=====
```

HITS AT: 33-40

Amino acid families for RUN GETSEQ SQEFP and QSFP search options

| GROUP | AMINO ACIDS |
|---------------------------|------------------|
| Neutral-Weak Hydrophobics | P, A, G, S, T |
| Acid Amines-Hydrophilic | Q, N, E, D, B, Z |
| Basic-Hydrophilic | H, K, R |
| Hydrophobics | I, M, L, V |
| Aromatic | F, W, Y |
| Cross-Linking | C |

Special variability symbols allow flexibility in sequence motif searching

- Variability symbols (pattern matching)
 - Allow users to specify motif patterns that consist of different amino acid(s) at one location of a sequence
 - Provide the ability to specify sequences separated by an unknown number of amino acids (gaps)
 - Provide the ability to search for sequence patterns at either beginning or the end of the sequence
 - Allow users to specify the number or range of repeats for amino acid(s) or gaps

Note: A complete table of all variability symbols, with search examples, is given in the USGENE, DGENE and PCTGEN database summary sheets:
www.stn-international.com/stndatabases/databases/onlin_db.html

Variability symbols for RUN GETSEQ sequence code match searches

| <u>Symbol</u> | <u>Function</u> |
|---------------|---|
| [] | Specify alternate residues |
| [-] | Exclude a specific residue or alternate residues |
| { } | Repeat the preceding symbol(s) (number or range) |
| ? | Repeat the preceding symbol(s) zero or one time |
| * | Repeat the preceding symbol(s) zero or more times |
| + | Repeat the preceding symbol(s) one or more times |
| ^ | Query appears at the beginning or the end of a sequence |
| | Alternate sequence expressions |
| . | A gap of one residue |
| : | A gap of zero or one residues |
| & | Concatenate (join together) sequence queries |

Use SCM variability symbols to search USGENE* and REGISTRY

Search Question:

Find patent references* disclosing one or more of the sequences represented by this Markush peptide sequence formula:

LGPX₁QLCX₂LVX₃CAP

X₁ = V or L

X₂ = any amino acid except, G or H

X₃ = any amino acid

(* DGENE and PCTGEN should also be included, but have been omitted simply to save on presentation time.)

RUN GETSEQ SCM search strategy

=> **RUN GETSEQ LGP[VL]QLC[-GH]LV.CAP/SQSP**

– Possible sequence retrieval

- *LGPVQLCALVHCAP*
- *LGPVQLCSLVVCAP*
- *LGPLQLCVLVACAP*
- *LGPLQLCPLVTCAP*

Reminder: An SCM search may also be run in REGISTRY, but the SEARCH (=> S) command is used instead of RUN GETSEQ.

Run the USGENE GETSEQ SCM search

=> FILE USGENE

=> RUN GETSEQ LGP[VL]QLC[-GH]LV.CAP/SQSP

RUN GETSEQ AT 21:42:25 ON 13 MAY 2008
COPYRIGHT (C) 2008 FIZ KARLSRUHE GMBH

L1 RUN STATEMENT CREATED
L1 32 LGP[VL]QLC[-GH]LV.CAP/SQSP

32 sequence hits (L1) have been found in USGENE containing the sequence fragment(s) of interest.

=> D TRI SEQ

L1 ANSWER 1 OF 32 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
TI Nucleotide and amino acid sequences, and assays and methods of use thereof for diagnosis of prostate cancer (Patent)

MTY Protein

SQL 417

SEQ

```
1 mrfawtvlll gplqlcalvh cappaagqqq  
= =====  
51 ngqvfsllsl gsqyqpqrrr dpgaavpgaa nasaqqprtp illirdnrta  
.  
.  
.  
401 rytghhayas gctispy
```

The hit portion of the answer sequence is highlighted with double underlining.

HITS AT: 10-23

Repeat the USGENE search in REGISTRY and combine all results in CPlusSM

```
=> FILE REGISTRY
=> S L1
L2          38 LGP[VL]QLC[-GH]LV.CAP/SQSP

=> FIL HCPLUS

=> S L2 AND P/DT
L3          28 L2 AND P/DT

=> TRA PN L1
L4          TRANSFER L1 1- PN :      30 TERMS
L5          65 L4

=> S L3 OR L5
L6          75 L3 OR L5

=> S L6 AND (ANTIBOD### OR IMMUNOGLOBULIN#) AND DIAGNOS? AND
    PROSTAT? AND (CANCER? OR TUMOR? OR NEOPLAS?)
L7          4 L6 AND (ANTIBOD
    PROSTAT? AND
```

To repeat an SCM search
in REGISTRY simply
SEARCH the answer set
L-number from USGENE.

L3 = CPlus patent records
found using REGISTRY.
L5 = CPlus patent records
found using USGENE.
L6 = CPlus records found
using both USGENE and
REGISTRY in combination.

The CPlus search may be further refined
using CAS value-added abstracts and indexing.

Use USGENE and REGISTRY in combination to locate relevant CPlus records

=> D L7 BIB ABS HITIND

L7 ANSWER 1 OF 4 HCAPLUS COPYRIGHT
 AN 2007:463771 HCAPLUS
 TI Detection of tissue-derived glycoproteins in blood serum in **diagnosis** and monitoring of disease
 IN Zhang, Hui; Aebersold, Rudolf H.
 PA Institute for Systems Biology, USA

This example CPlus record was uniquely retrieved by the combination of a USGENE GETSEQ search and CPlus value-added indexing search.

FAN.CNT 1

| | PATENT NO. | KIND | DATE | APPLICATION NO. | DATE |
|------|-----------------|------|----------|-----------------|--------------|
| PI | WO 2007047796 | A2 | 20070426 | WO 2006-US40784 | 20061017 |
| | US 20070099251 | A1 | 20070503 | US 2006-582861 | 20061017 <-- |
| PRAI | US 2005-728044P | P | 20051017 | | |

AB A method of detecting tissue-derived glycoproteins in blood serum that is useful in the **diagnosis** of disease and in monitoring

IT Bladder, **neoplasm**
 Ovary, **neoplasm**
 Prostate gland, disease
 Prostate gland, **neoplasm**

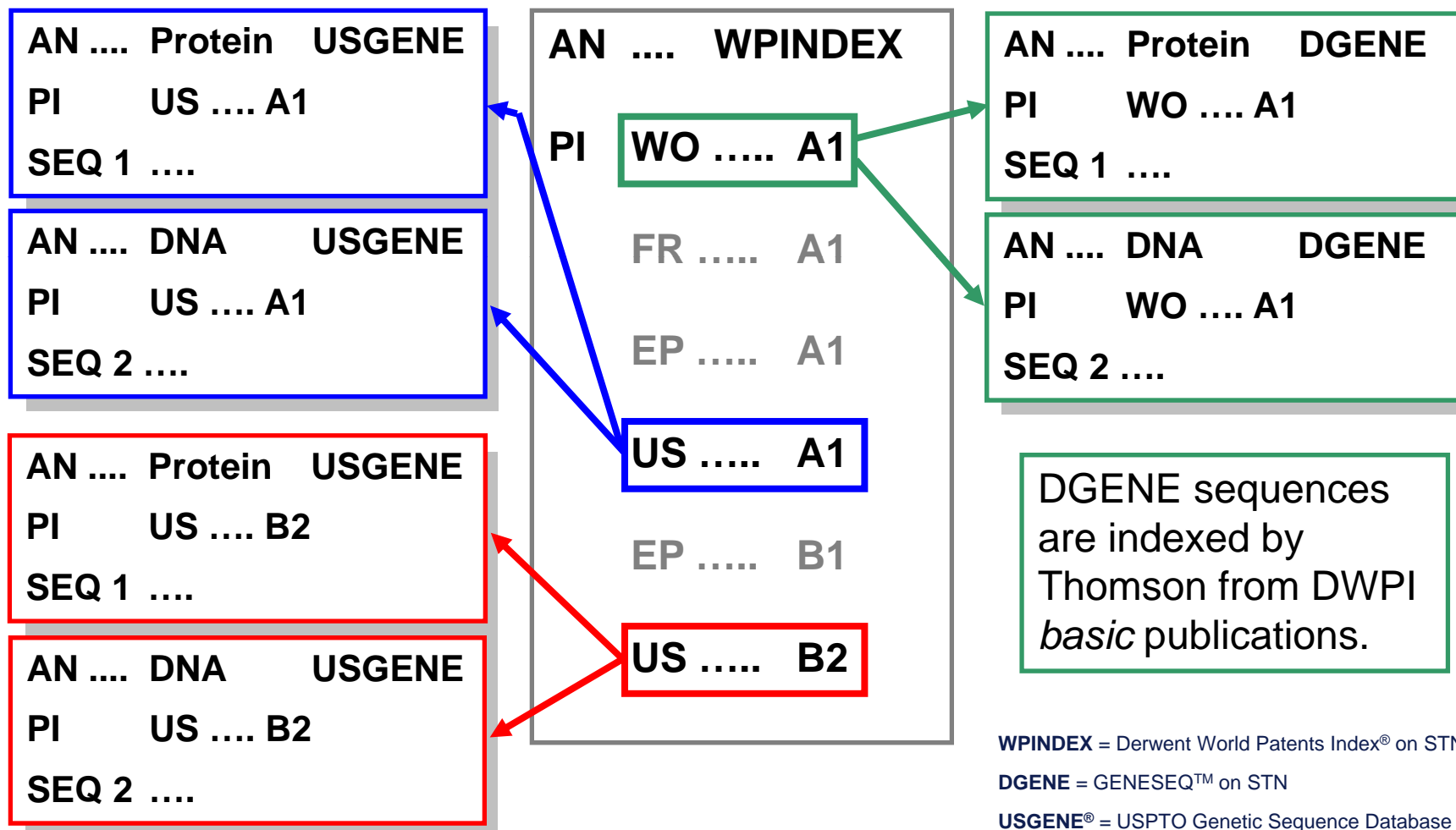
Tip: This arrow indicates the family member that was retrieved in the USGENE RUN GETSEQ search (L1).

(glycoprotein shedding into blood in **diagnosis** of; detection of tissue-derived glycoproteins shed into blood serum in diagnosis and monitoring of disease)

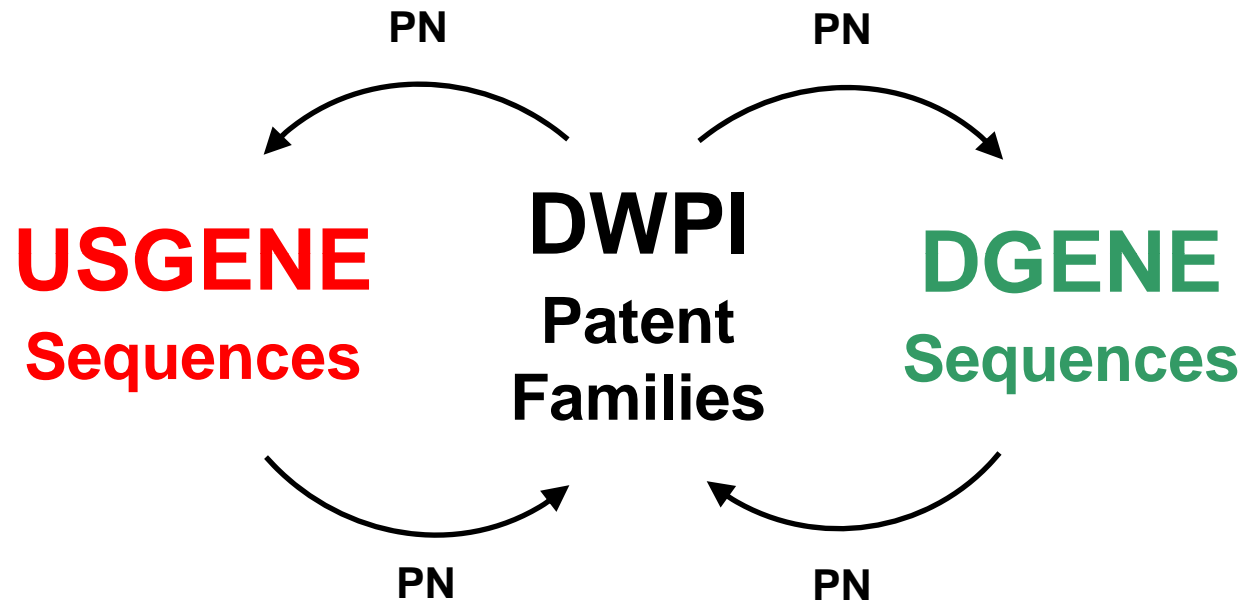
Agenda

- DGENE, PCTGEN, USGENE database content
- The 7 basic steps of BLAST
- BLAST and Patent Family SORT (FSORT)
- Post-processing BLAST search results
- Similarity searching GETSIM (FASTA)
- Offline BATCH search mode
- Sequence Code Match searching (GETSEQ)
- **Multifile patent sequence search example**

Reminder: USGENE and DGENE often capture sequences from different patent family members



The “best-practice” recipe for multifile 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 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 results: D PFAM=1- 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  
ATENAKEEVRRILGLLDAYLKTRTFLVGERVTLADITVVCTLLWLYKQVLEPSFRQAFPNTNR  
WFLTCINQPQFRAVLGEVKLCEKMAQFDAKKFAETQPKKDTPRKEKGSREEKQKPQAERKEEK  
KAAAPAPEEEMDECEQALAAEPKAKDPFAHLPKSTFVLDEFKRKYSNEDTLSVALPYFWEHFD  
KDGWSLWYSEYRFPEELTQTFMSCNLITGMFQRLDKLRKNAFASVILFGTNNSSSISGVWVFR  
GQELAFPLSPDWQVDYESYTWKLDPGSEETQTLVREYFSWEGAFQHVGKAFNQKIFK
```

(Search conducted on May 14th, 2008.)

1) Ensure 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
=> SET FORMAT .MYWPINDEX BIB
=> FILE USGENE; SET DFORMAT .MYUSGENEALIGN
=> FILE DGENE; SET DFORMAT .MYDGENEALIGN
=> FILE WPINDEX; SET DFORMAT .MYWPINDEX
=> 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

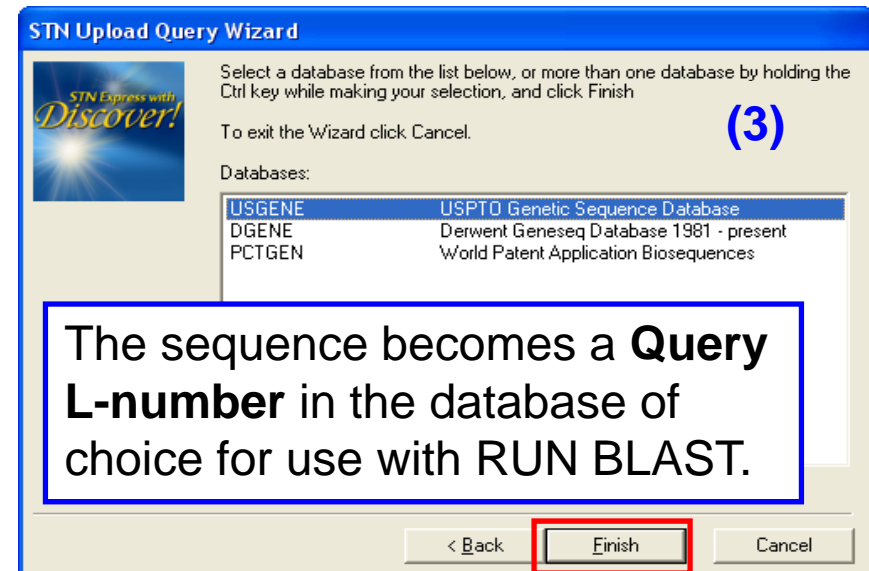
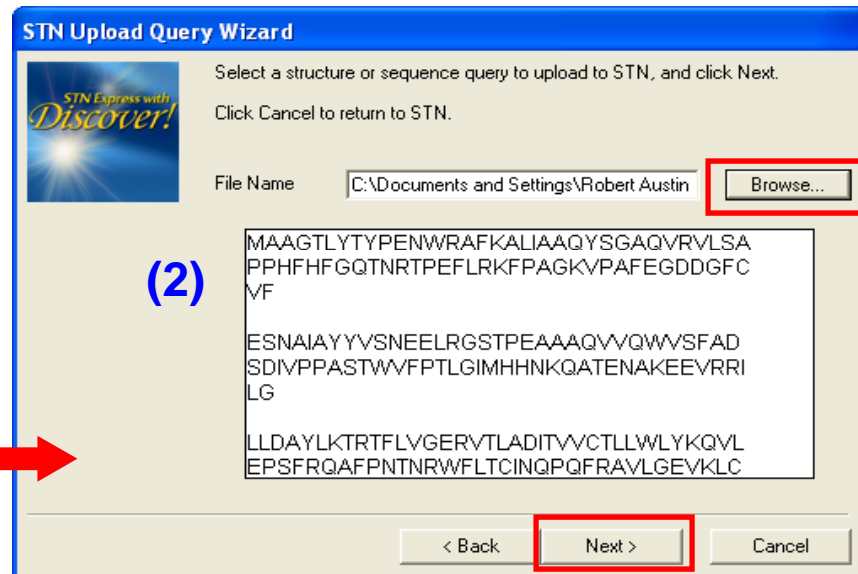
| | | |
|----------------|--------------------------------|---------|
| .MYDGENEALIGN | TRI OS SCORE ALIGN | DGENE |
| .MYUSGENEALIGN | TRI ORGN SEQN SEQC SCORE ALIGN | USGENE |
| .MYWPINDEX | BIB | WPINDEX |

2) UPLOAD the sequence query

- (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 MAAGTLYTYPENWRAFKALIAAQYSGAQVRVLSAPPHFHFGQTNRTPEFLRKFPAAGKVP
AFEGDDGFCVFESENIAIAYYSNEELRGSTPEAAAQVVQWVSFADSDIVPPASTWVFPTL
GIMHHNKQATENAKEEVRRILGLLDAYLKTRTFLVGERVTLADITVVCTLLWLKQVLE
PSFRQAFPNTNRWFLTCINQPQFRAVLGEVKLCEKMAQFDAKKFAETQPKKDTPRKEKG
SREEKQKPQAERKEEKKAAAPAPEEEMDECEQALAAEPKAKDPFAHLPKSTFVLDEFKR
KYSNEDTLSVALPYFWEHFDKDGWSLWYSEYRFPEELTQTFMSCNLITGMFQRLDKLRK
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 21:29:18 ON 14
COPYRIGHT (C) 2008 SEQUENCEBASE CORP

FILE LAST UPDATED: 9 MAY 2008 <20080509/UP>
MOST RECENT PUBLICATION DATE: 8 MAY 2008 <20080508/PD>

FILE COVERS 1982 TO DATE

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

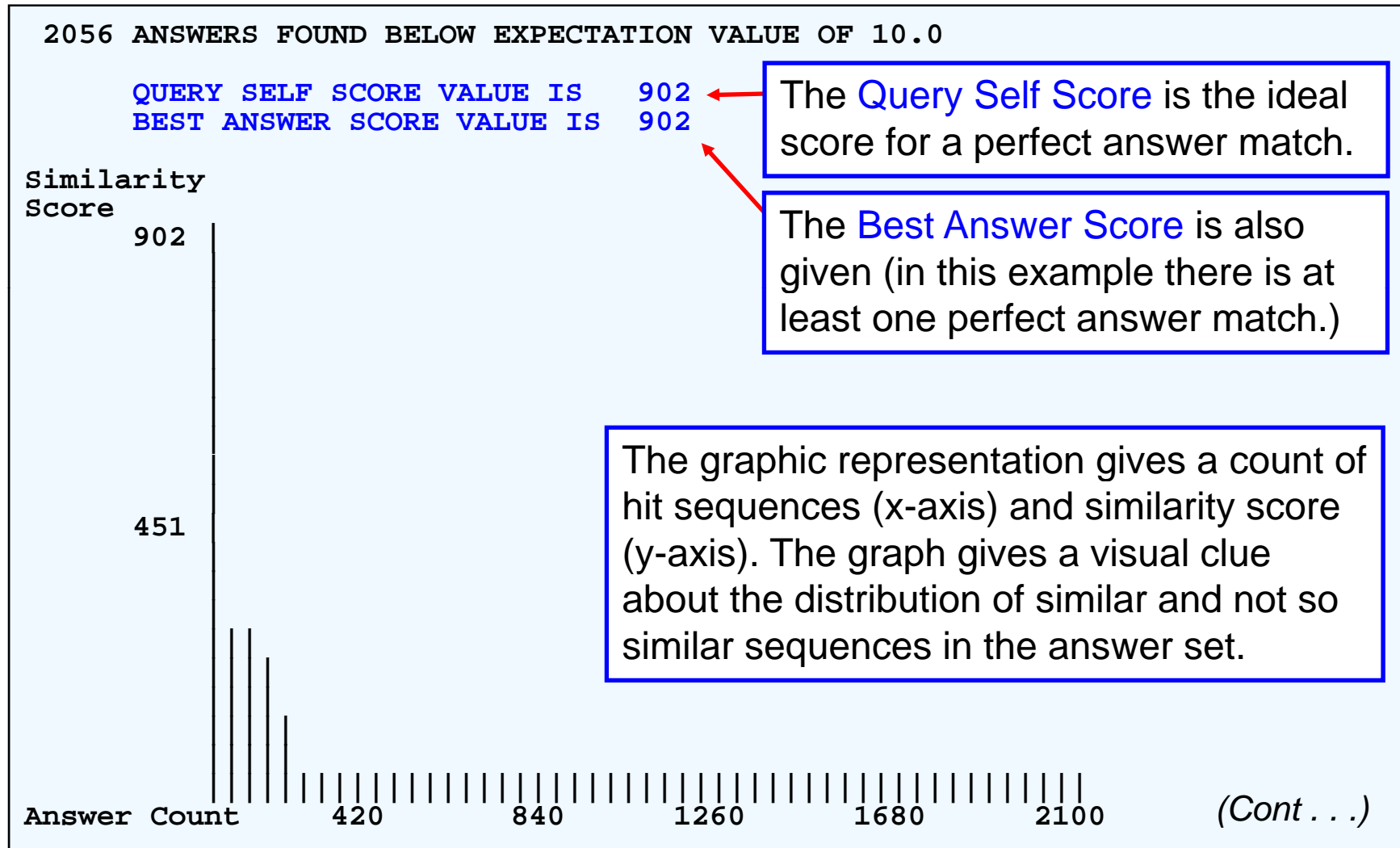
=> RUN BLAST L1 /SQP -F F

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      15 MAAGTLYTYPENWRAFKALIAAQ
      RKFPAGKVPAPAFEGDDGFCVFESNAIAYYSNEELRGSTPEAAAQVVQWVS
      FADSDIVPPASTWVFPTLGIMHHNKQATENAKEEVRRILGLLDAYLKTRT
      FLVGERVTLADITVVCTLLWLYKQVLEPSFRQAFPNTNRWFLTCINQPQF
      RAVLGEVKLCEKMAQFDAKKFAETQPKKDTPRKEKGSREEKQKPQAERKE
      EKKAAPAPEEEMDECEQALAAEPKAKDPFAHLPKSTFVLDEFKRKYSNE
      DTL SVALPYFWEHFDKDGWSLWYSEYRFPEELTQTFMSCNLITGMFQRLD
      KLRKNAFASVILFGTNNSSSISGVVFRGQELAFPLSPDWQVDYESYTW
      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      15 SOR L2 SCORE D
```

Use SORT SCORE D to sort by descending BLAST score.

3) RUN the USGENE BLAST search (cont.)

=> D 1-15

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

L3 ANSWER 1 OF 15 USGENE COPYRI

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

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

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 FEGDDGFCVFESNAIAYYVSNEELRGSTPEAAAQVVQVVSFADSDIVPPASTWVFPTLGI

. . . .

3) RUN the USGENE BLAST search (cont.)

=> D SCORE 1-15

Another way to review quickly is by BLAST SCORE.

L3 ANSWER 1 OF 15 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
SCORE 902 100% of query self score 902

. . . .

L3 ANSWER 10 OF 15 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
SCORE 788 87% of query self score 902

. . . .

L3 ANSWER 13 OF 15 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
SCORE 656 72% of query self score 902

L3 ANSWER 14 OF 15 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
SCORE 656 72% of query self score 902

L3 ANSWER 15 OF 15 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
SCORE 495 54% of query self score 902

=> SOR AN 1-14

PROCESSING COMPLETED FOR L3

L4 14 SOR L3 1-14 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 20 MAAGTLYTYPENWRAFKALIAAC
RKFPAGKVPAFEGDDGFCVFESM

. . . .

KLRKNAFASVILFGTNNSSISGVVFRGQELAFPLSPDWQVDYESYTW
KLDPGSEETQTLVREYFSWEGAFQHVGKAFNQKIFK/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 20 SOR L5 SCORE D

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

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

Use SORT SCORE D to sort
by descending BLAST score.

4) RUN the DGENE BLAST search (cont.)

=> D 1-20

L6 ANSWER 1 OF 20 DGENE COPYRI
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; GIG

SQL 437

OS 2006-747536 [76]

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

. . . .

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

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

4) RUN the DGENE BLAST search (cont.)

=> D SCORE 1-20

Another way to review quickly is by BLAST SCORE.

L6 ANSWER 1 OF 20 DGENE COPYRIGHT 2008 THE THOMSON CORP on STN
SCORE 902 100% of query self score 902

. . . .

L6 ANSWER 14 OF 20 DGENE COPYRIGHT 2008 THE THOMSON CORP on STN
SCORE 880 97% of query self score 902

. . . .

L6 ANSWER 18 OF 20 DGENE COPYRIGHT 2008 THE THOMSON CORP on STN
SCORE 656 72% of query self score 902

L6 ANSWER 19 OF 20 DGENE COPYRIGHT 2008 THE THOMSON CORP on STN
SCORE 495 54% of query self score 902

L6 ANSWER 20 OF 20 DGENE COPYRIGHT 2008 THE THOMSON CORP on STN
SCORE 495 54% of query self score 902

=> SOR AN 1-18

PROCESSING COMPLETED FOR L6

L7 18 SOR L6 1-18 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 : 14 USGENE sequence hits (L4)
L9 11 L8 found 11 DWPI records (L9).

L10 TRANSFER L7 1- PN : 18 DGENE sequence hits (L7)
L11 15 L10 found 15 DWPI records (L11).

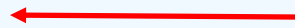
=> S L9 OR L11

L12 20 L9 OR L11

Total DWPI records is 20 (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 21:36:23 ON 14 MAY 2008
COPYRIGHT (C) 2008 SEQUENCEBASE CORP

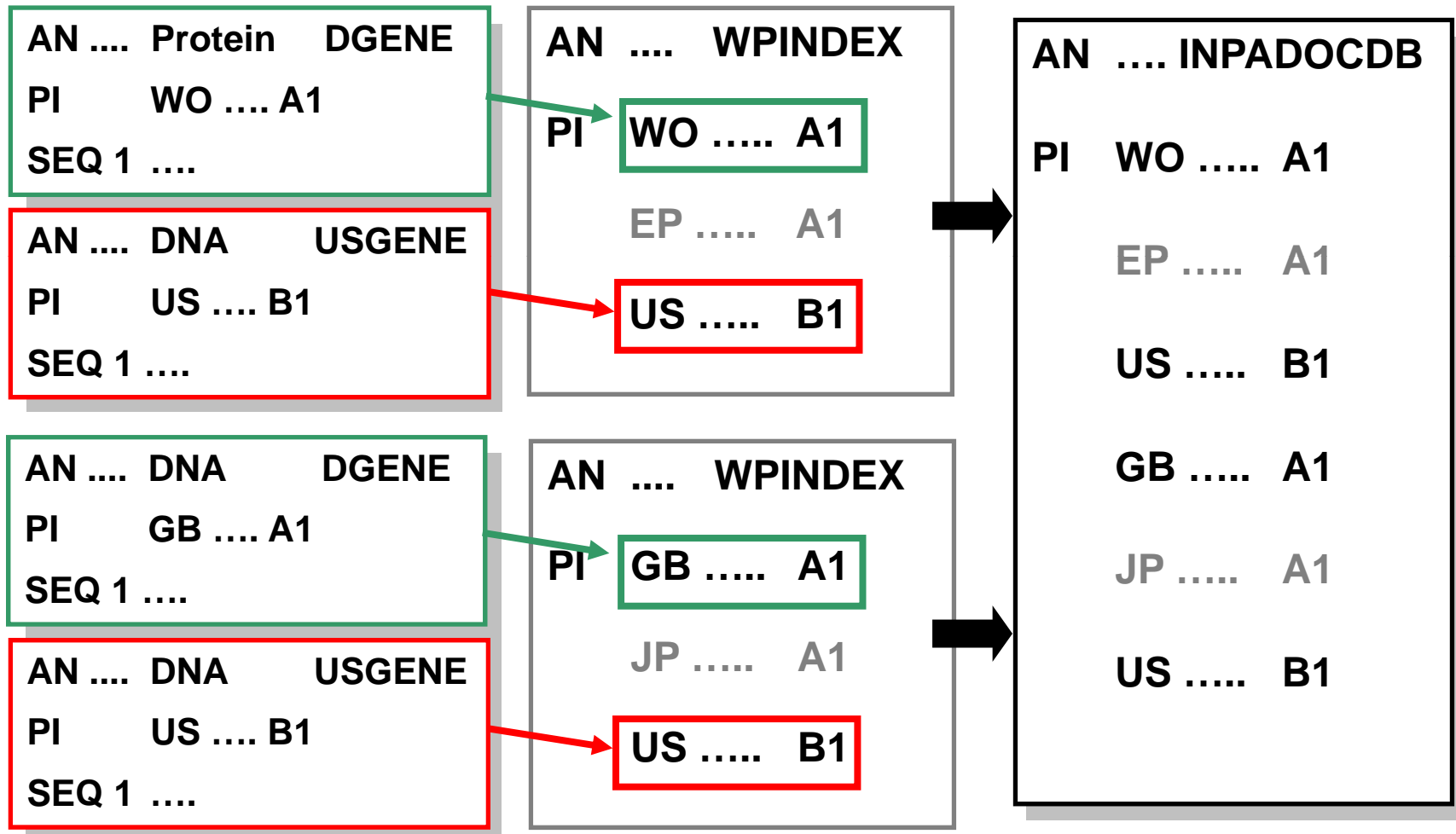
FILE 'DGENE' ENTERED AT 21:36:23 ON 14 MAY 2008
COPYRIGHT (C) 2008 THE THOMSON CORPORATION

FILE 'WPINDEX' ENTERED AT 21:36:23 ON 14 MAY 2008
COPYRIGHT (C) 2008 THOMSON REUTERS

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

L13 52 DUP IDE L4 L7 L12 (INCLUDES 0 SETS OF DUPLICATES)
 ANSWERS '1-14' FROM FILE USGENE
 ANSWERS '15-32' FROM FILE DGENE
 ANSWERS '33-52' 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

52 FSO L13

19 Multi-record Families Answers 1-52

| | |
|-----------|---------------|
| Family 1 | Answers 1-3 |
| Family 2 | Answers 4-11 |
| Family 3 | Answers 12-14 |
| Family 4 | Answers 15-16 |
| Family 5 | Answers 17-18 |
| Family 6 | Answers 19-20 |
| Family 7 | Answers 21-23 |
| Family 8 | Answers 24-26 |
| Family 9 | Answers 27-28 |
| Family 10 | Answers 29-30 |
| Family 11 | Answers 31-33 |
| Family 12 | Answers 34-36 |
| Family 13 | Answers 37-39 |
| Family 14 | Answers 40-41 |
| Family 15 | Answers 42-43 |
| Family 16 | Answers 44-45 |
| Family 17 | Answers 46-48 |
| Family 18 | Answers 49-50 |
| Family 19 | Answers 51-52 |

0 Individual Records

0 Non-patent Records

The 20 DWPI records (L12), 14 USGENE sequence hits and 18 DGENE sequence hits belong to 19 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 127)

=> D PFAM=1- TOTAL ←

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

. . . .

L14 ANSWER 12 OF 52 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN **FAMILY3**

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

MTY Protein

SQL 437

ORGN Homo Sapiens

SEQN 2421

SEQC 6355

USGENE hit sequence display(s).

SCORE 889 98% of query self score 902

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

Query: 61 FEGDDGFCVFESNAIAYYVSNEELRGSTPEAAAQVVQVVSFADSDIVPPASTWVFP TLGI

. . . .

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

```
L14 ANSWER 13 OF 52 DGENE COPYRIGHT 2008 THE THOMSON CORP on STN FAMILY3
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] DGENE hit sequence display(s).
SCORE 889 98% of query self score 902
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
. . . .
```

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

```
L14 ANSWER 14 OF 52 WPINDEX COPYRIGHT 2008 THOMSON REUTERS on STN FAMILY3
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.

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

| | SEQs | SEQs > 70% | PNs | DWPI Records | FSORT Families |
|----------------|-------------|--------------------------|------------|-------------------------|---------------------------|
| DGENE | 1957 | 18 | 15 | 15 | 14 |
| USGENE | 2056 | 14 | 13 | 11 | 11 |
| Overlap | - | - | 0 | 6 | 6 |
| Total | - | - | 28 | 20 | 19 |

Example: USGENE unique retrieval

```
L14 ANSWER 29 OF 52 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STNFAMILY10
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
SCORE 889          98% of query self score 902
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
Query: 61  FEGDDGFCVFESENAIAYYVSNEELRGSTPEAAAQVVQVVSFADSDIVPPASTWVFPTLGI
          FEGDDGFCVFESENAIAYYVSNEELRGSTPEAAAQVVQVVSFADSDIVPPASTWVFPTLGI
Sbjct: 61  FEGDDGFCVFESENAIAYYVSNEELRGSTPEAAAQVVQVVSFADSDIVPPASTWVFPTLGI
Query: 121 MHHNKQATENAKEEVRRILGLLDAYLKTRTFLVGERVTLADITVVCTLLWLKYQVLEPSF
          . . . .
```

This USGENE hit sequence uniquely retrieved the DWPI record on the following slide (as of May 14th, 2008).

Example: USGENE unique retrieval (cont.)

```
L14 ANSWER 30 OF 52 WPINDEX COPYRIGHT 2008 THOMSON REUTERS on STN FAMILY10
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 (20050922) EN
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 (as of May 14th, 2008).

Summary

- RUN BLAST, RUN GETSIM (FASTA) and RUN GETSEQ (SCM) command line search options are available for DGENE, USGENE and PCTGEN
- A new command line feature to refine BLAST and GETSIM answer sets by percent (%) is now available
- USGENE is a vital tool for business critical patent searches, providing a complete collection of all available U.S. granted patent and published application sequence data within **3 days** of publication by the USPTO
- DGENE, USGENE, PCTGEN and REGISTRY are all required for a comprehensive patent sequence search

Resources for sequence searching on STN

- More on the new percent option for BLAST & GETSIM
www.stn-international.com/New_sequence_search.html
- *Sequence Searching on STN* modular workshop
www.fiz-k.com/bostonsequenceworkshop
 - 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
- CAS REGISTRY sequence coverage and resources
www.cas.org/support/stngen/stndoc/sequences.html

STN[®]

Effective patent sequence searching
on STN[®]

Jim Brown – FIZ Karlsruhe