

# STN<sup>®</sup>

Sequence code match searching  
on STN<sup>®</sup>

Martine MICHEL  
CAPADOC

# Agenda

- REGISTRY<sup>®</sup> database content
- USGENE<sup>®</sup> database content
- Sequence Code Match searching (GETSEQ)
  - Exact, family and subsequence
  - Variability searching
- Multifile patent sequence search example

# REGISTRY

- Produced by the Chemical Abstracts Service
- Sequences from >3000 life science journals and the basic patents of the 59 patent issuing authorities of the CAplus<sup>SM</sup> file on STN
- Standardized nomenclature provided for each sequence and also patent number and location from 1999
- >61 million sequence records
- Updated daily
- 1907 - present

Learn more at: [www.cas.org](http://www.cas.org)

# Relationship between CAPLUS patent family and CAS Registry database

**AN .... CAPLUS**

**TI ....**

**PA ....**

**PI WO .... A1**

**FR .... A1**

**US .... A1**

**US .... B2**

**AB ....**

**IT RN ....**

**RN .... Protein REGISTRY**

**PI WO .... A1**

**SEQ 1 ....**

**RN .... DNA REGISTRY**

**PI WO .... A1**

**SEQ 2 ....**

**RN .... Peptide REGISTRY**

**PI WO .... A1**

**SEQ 3 ....**

# REGISTRY sequences also come from >3000 life science journal titles

- *Biochemistry*
- *Cell*
- *EMBO Journal*
- *Gene*
- *Journal of Biological Chemistry*
- *Journal of Cell Biology*
- *Journal of Molecular Biology*
- *Nature*
- *Nucleic Acid Research*
- *Nature Genetics*
- *Proceedings of the National Academy of Sciences*
- *Science*

# Sequences are indexed from patents

- 59 patent-issuing authorities are monitored
  - Including WIPO, EPO, USPTO, JP, DE, GB, FR, RU and CA
- For the major patent-issuing authorities listed above, currency is top notch
  - Bibliographic information is available in CAPLUS<sup>SM</sup> within 2 days
  - Sequences are available in REGISTRY within 1 month

# Unique sequence types are indexed in the CAS REGISTRY file

- Naturally occurring proteins and nucleotides
- Chemically modified peptides and proteins
- Sequences deduced from gene translations
- GenBank sequences and translations
- Multichain proteins
- Cyclic peptides
- Fusion proteins
- Peptide-metal complexes
- Sequences containing uncommon amino acids
- Protein-nucleotide sequences (PNA sequences)

# Sequence searching in the CAS REGISTRY file has many advantages

- Links to CAPLUS and other databases for references
  - Add date restrictions
  - Add additional keywords for relevance
- Links to electronic full text of journal articles and patents

# Agenda

- REGISTRY<sup>®</sup> database content
- USGENE<sup>®</sup> database content
- Sequence Code Match searching (GETSEQ)
  - Exact, family and subsequence
  - Variability searching
- Multifile patent sequence search example

# USGENE is the USPTO Genetic Sequence Database

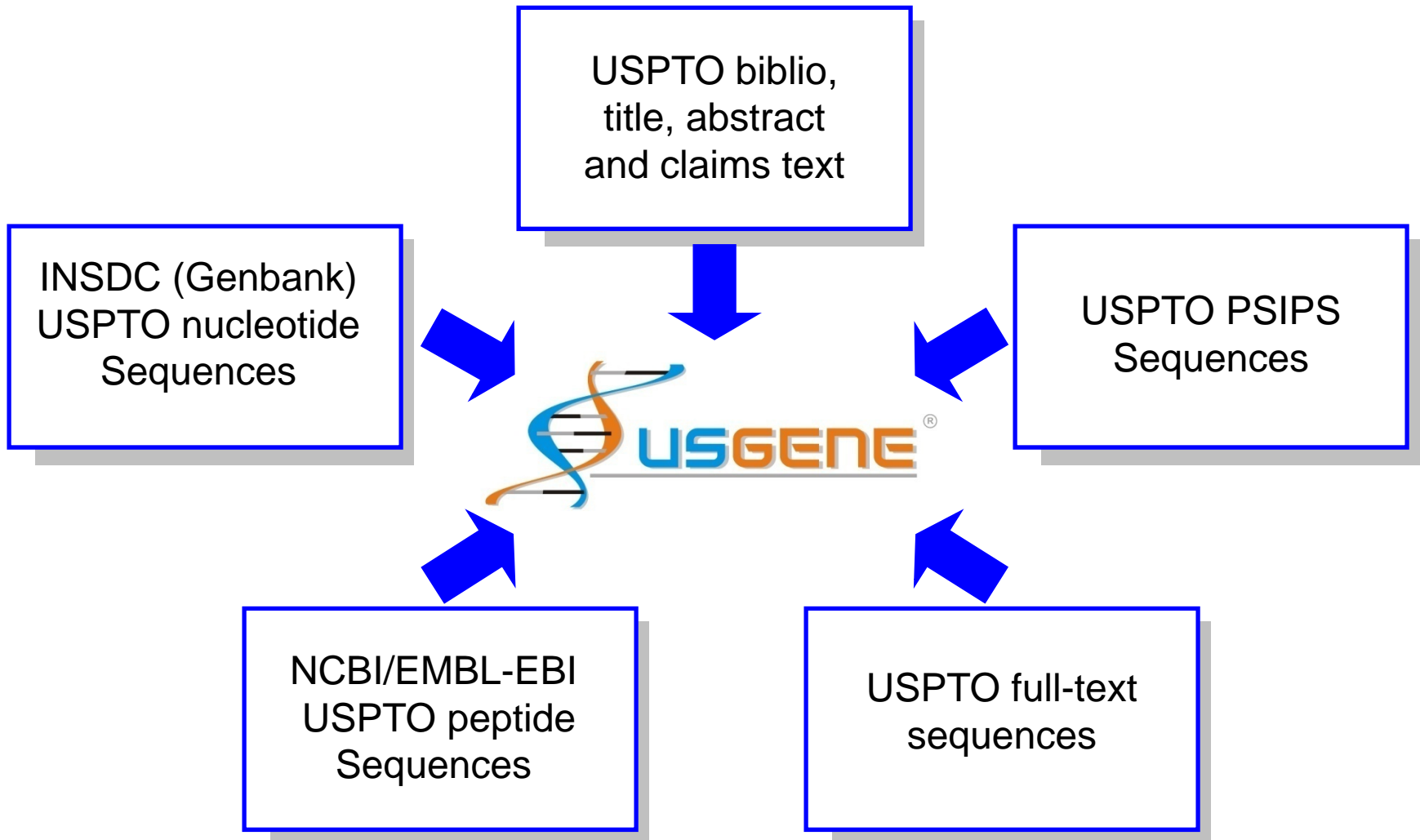
- Sequences captured from **all relevant USPTO published patent applications and granted (issued) patents**
- Assignee and full inventor names; publication, application and parent case PCT numbers and dates; **original publication title, abstract and claims**
- Organism name, sequence length, Molecule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

# USGENE consolidates unique USPTO sequence data from different sources

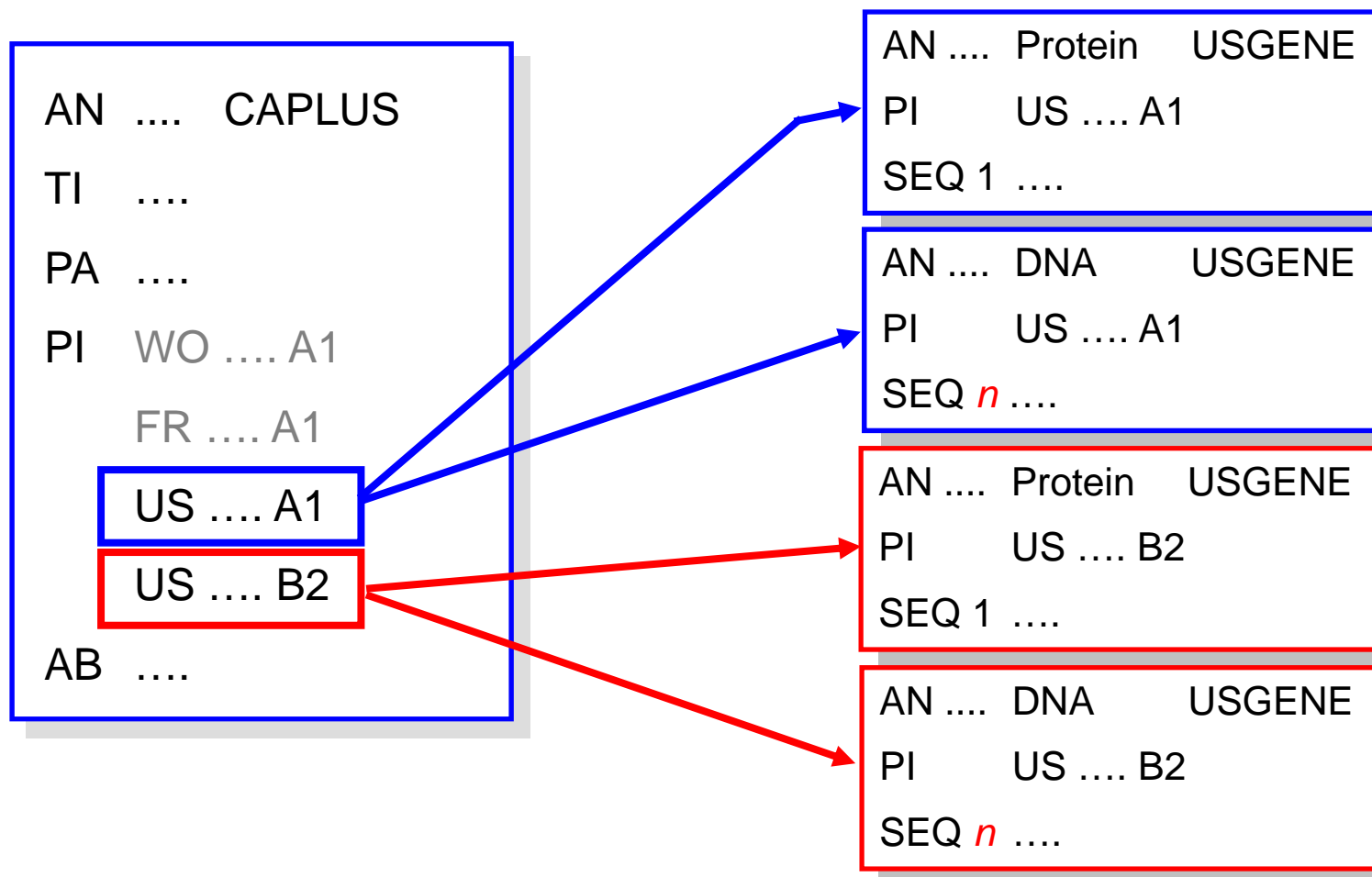
- USPTO Publication Site for Issued and Published Sequences (PSIPS)
  - The official mega-publication download site, 2001-date
- International Nucleotide Sequence Database Collaboration (INSDC) (NCBI/EMBL/DDBJ, Genbank)
  - U.S. granted patent nucleotide sequences, 1982-date
- USPTO Protein Database (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 sequences with bibliographic data and claims text



# Relationship between CAPLUS patent family and USGENE sequence database



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

```
L3 ANSWER 1 OF 1 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
AN 20080256649.126 Protein USGENE
TI Novel Acetylcholinesterase Gene Responsible for
resistance and Applications Thereof (Published
IN Weill Mylene (Montpellier, FR); Fort Philippe
Raymond Michel (Montpellier, FR); Pasteur Nicc
PA CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE (
PI US 20080256649 A1 20081016
AI US 2003-518072 20030619
RLI WO 2003-FR1876 20030619
ED 20081017
AB The invention re
responsible for
mosquitoes, which
acetylcholinester
protein AchE1) ar
ECLM US20080256649 A1
that it comprises a central catalytic region which has an amino acid
sequence selected from the group consisting of the sequence SEQ ID NO
1 and the sequences exhibiting at least 60% identity or 70%
similarity with the sequence SEQ ID NO 1, with the exclusion . . . .
SSO PROTEIN; USPTO; APPLICATION
ORGN Anopheles gambiae
SQL 737
SEQ
1 meirgllmgr lrlgrrmvpl gllgvt
51 igshqlsaaa gvglssqsaq sgslas
. . . .
```

AN 20080256649.126  
is SEQ ID NO: 126  
from US20080256649.

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

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

# USGENE also has an extensive backfile

United States Patent: 5210028 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

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

Most Visited STN International STN/CAS Home Page STN on the Web STN Viewer

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

This USPTO example is US5210028, which was issued on May 11<sup>th</sup>, 1993.

SEQUENCE LISTING (1) GENERAL INFORMATION:

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

INFORMATION FOR SEQ ID NO:2: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 216 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: double (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (iii) HYPOTHETICAL: N (iv) ANTI-SENSE: N (vi) ORIGINAL SOURCE: (A) ORGANISM: Synthetic Human IGF-II DNA Sequence; E.coli pref (B) STRAIN: pBB8/IGF-II (xi) SEQUENCE DESCRIPTION: SEQ ID NO:2: CATATGGCATAACCG CCCGAGCGAGACCCCTGTGCGGTGGCGAGCTCGTAGACACTCTGCAG60  
TTCGTTTGTGGTGACCGTGGCTTCTACTTCTCTCGTCCTGCTAGCCGTGTATCTCGCCGT120  
TCTAGAGGCATCGTTGAAGAGTGCTGTTTCCGCAGCTGTG  
TACTGCGCAACTCCAGCAAATCCGAATAAGGATCC216 (2) 1

LENGTH: 233 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: double (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (iii) HYPOTHETICAL: N (iv) ANTI-SENSE: N (vi) ORIGINAL SOURCE: (A) ORGANISM: Human IGF-II (xi) SEQUENCE DESCRIPTION: SEQ ID NO:3: GAATTCGACGCTTATGGCTTACAGACCATCCGAAACCTTGT  
CACCTTGCAATTCGTTTGTGGTGACAGAGGTTT CTA

TTCTAGAAGATCCAGAGGTATCGTTGAAGAATGTTGTTTCA  
GTTGAAACCTACTGTGCTACCCAGCTAAGTCTGAATGAATGCGTCGAATTC233 (2) INFORMATION FOR SEQ ID NO:4: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 150 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: double (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (iii) HYPOTHETICAL: N (iv) ANTI-SENSE: N (vi) ORIGINAL SOURCE: (A) ORGANISM: ner gene ribosomal binding site of phage Mu (B) STRAIN: pPLMu (xi) SEQUENCE DESCRIPTION: SEQ ID NO:4: GAATT  
CTTACACTTAGTTAAATTGCTAACTTTATAGATTACAAAACCTTAGGAGGGTTTTT60  
ACCATGGTTACGAATCCCGGGGATCCGTCGACCTGCAGCCAAGCTTGGCTGCCTCGCGC120  
GTTTCGGTGATGACGGTGAAAACCTCTGAC 150

\*\*\*\*\*

Done

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

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

```
L1 ANSWER 1 OF 1 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
AN 5210028.1 protein USGENE
TI Process for the production of unfused IGF-II (Patent)
IN Schmitz Albert (Basel, CH); Marki Walter (M
PA Ciba Geigy Corporation (Ardsley NY)
PI US 5210028 A 19930511
AI US 1990-616470 19901121
AB A process for the preparation of a recombinant IGF-II (rIGF-II)
without a covalently attached foreign protein moiety and without N-
terminal attached methionine or a derivative of methionine or of a
salt of said IGF-II, rIGF-II produc
ECLM US5210028 A: We claim:1. A process
recombinant IGF-II without a covale
moiety and without N-terminal attac
methionine, said process comprising
of E. coli, said strain being a lon.sup.- and htpR.sup.- double
mutant, with a hybrid vector comprising an expression cassette
consisting of the following elements in the 5' to 3' direction, said
elements which are operably linked: an inducible promoter, a
ribosomal binding site, and the cod
linked in proper reading frame to a
IGF-II having the amino acid sequenc
SSO PROTEIN; USPTO; GRANTED
ORGN Human IGF II
SQL 67
SEQ 1 ayrpsetlcg gelvdtlqfv cgdrgyfysr pasrvsrrsr giveeccfrs
51 cdlalletyc atpakse
```

AN 5210028.1 is SEQ ID NO: 1 from US5210028.

AN 5210028.1 is displayed here in BRIEF format, which includes the Exemplary Claim (ECLM).

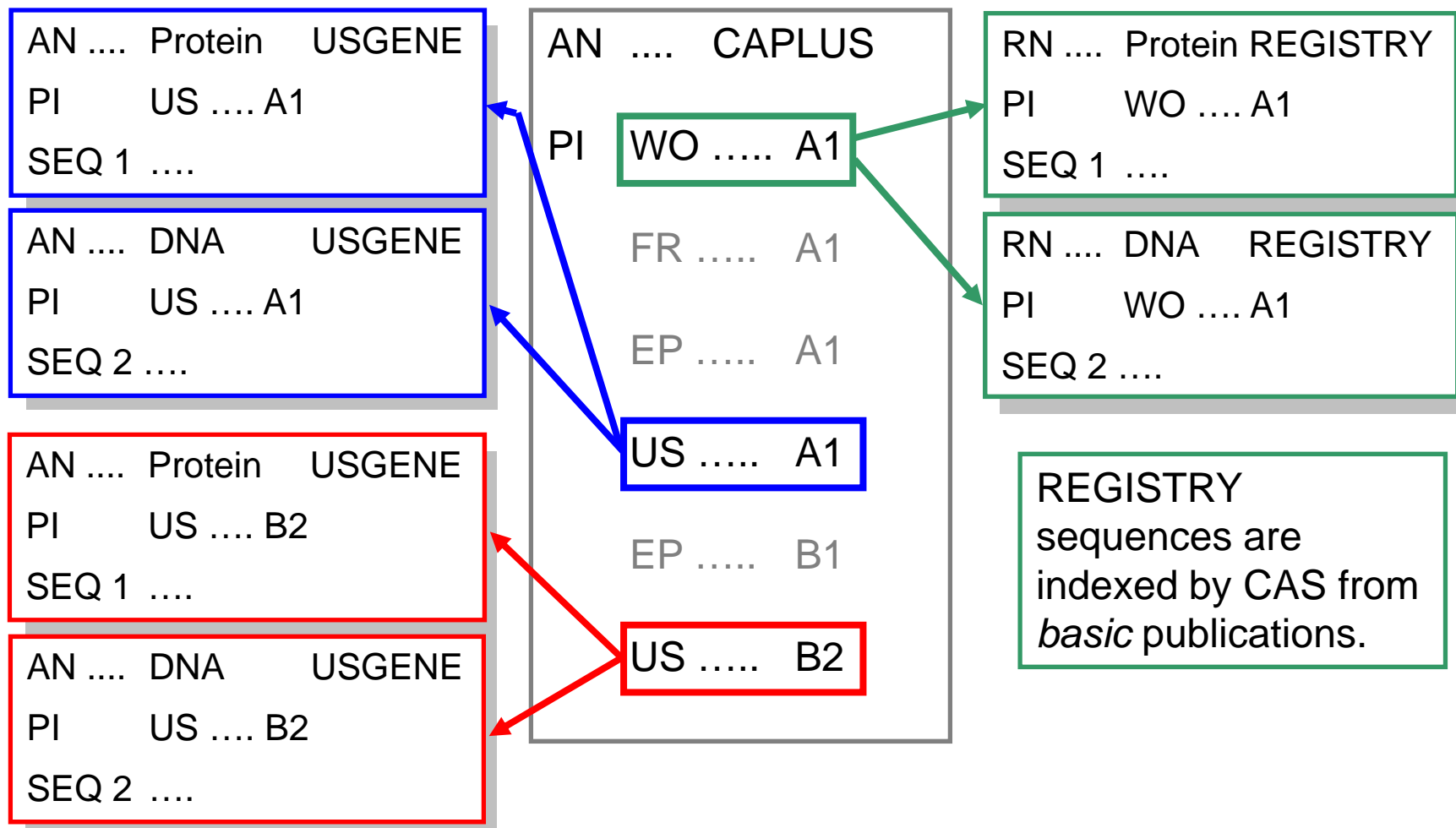
Compare the STN standardized USGENE record to the original data source on the previous slide.

1 ayrpsetlcg gelvdtlqfv cgdrgyfysr pasrvsrrsr giveeccfrs  
51 cdlalletyc atpakse

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

- REGISTRY provides patent sequence data from the CAPLUS *basic* publication
  - 61% of *basics* are WIPO/PCT published applications
  - Updated daily, typically 27 days from publication
- USGENE provides all available sequence data from the USPTO as a single merged resource
  - Both **U.S. patents** and **U.S. published applications**
  - Updated weekly, within **3 days** of USPTO publication
- Sequence listing variation often occurs between PCT and U.S. granted patent publication stages
  - Especially important, e.g. for freedom-to-operate

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





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

```
L3 ANSWER 1 OF 1 CAPLUS COPYRIGHT 2009 ACS on STN
AN 1995:394850 CAPLUS
TI Nucleic acids and peptides type-specific for hepatitis C virus types 4, 5
   and 6 and use of nucleic acids and peptides for diagnosis and in vaccines
IN Simmonds, Peter; Yap, Peng Lee; Pike, Ian Hugo
PA Common Services Agency, UK; Murex Diagnostics Ltd.
PATENT NO.          KIND    DATE          APPLICATION NO.    DATE
-----
PI WO 9425602        A1     19941110     WO 1994-GB957     19940505
   W: AU, CA, FI, J
   RW: AT, BE, CH, D
...
CA 2162134
ES 2231773
FI 9505224
FI 118688
US 6881821          B2     19940505     US 1995-537802    19951221
US 20050032047     A1     20050210
JP 2005124585      A      20050519     JP 2005-18491     20050126
JP 3895747         B2     20070322
...
```











In this example the patent family has:

- 21 sequences from [WO9425602](#) in REGISTRY
- 50 sequences from [US2005032047](#) in USGENE
- 58 sequences from [US6881821](#) in USGENE

# How does USGENE compare to other USPTO sequence data sources?

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

# How does USGENE compare to other USPTO sequence data sources? (cont.)

	USPTO Pub Apps	USPTO Patents	USPTO claims text	Value added
USGENE				
DGENE (DWPI basics)				
REGISTRY (CAplus basics)				
NCBI/EMBL				

# Agenda

- REGISTRY<sup>®</sup> database content
- USGENE<sup>®</sup> database content
- **Sequence Code Match searching (GETSEQ)**
  - Exact, family and subsequence
  - Variability searching
- Multifile patent sequence search example

# Command syntax

## Sequence Code Match (SCM)

- USGENE, DGENE, and PCTGEN use the same search command for SCM : **RUN GETSEQ**  
**=> RUN GETSEQ L1/SQEP**
- REGISTRY uses the general STN search command for SCM:  
**Search**  
**=> S L1/SQEP**

*/SQEP* (exact protein)

*/SQEFP* (exact family protein)

*/SQSP* (subsequence protein)

*/SQSFP* (subsequence family protein)

*/SQEN* (exact nucleotide)

*/SQSN* (subsequence nucleotide)

# Sequence Code Match: Exact search

- Exact search
  - Matches the sequence query as entered
  - Identical sequences and exact length
- File codes identify the type of sequence to search, e.g.
  - /SQEP = SeQ uence Exa ct P eptide**
  - /SQEN = SeQ uence Exa ct N ucleotide**

# EXACT amino acid search (/SQEP) in REGISTRY

=> **S DSDGP/SQEP**

1 DSDGP/SQEP  
75132 SQL=5

L1 1 DSDGP/SQEP  
(DSDGP/SQEP AND SQL=5)

=> **D L1 1 SEQ SQL**

L1 ANSWER 1 OF 1 REGISTRY COPYRIGHT 2008 ACS on STN  
SEQ 1 DSDGP ←

HITS AT: 1-5  
SQL 5

=> **D SEQ3**

L1 ANSWER 1 OF 1 REGISTRY COPYRIGHT 2008 ACS on STN  
SEQ3 1 Asp-Ser-Asp-Gly-Pro ←

HITS AT: 1-5

Amino acids can be displayed as a single letter codes (SEQ) or three letter codes (SEQ3).

The SEQ display format shows the entire sequence, with the hit acids underlined, and identified by "HITS AT".

# EXACT nucleic acid search (/SQEN) in REGISTRY

=> S GGAATT/SQEN

3 GGAATT/SQEN

75336 SQL=6

L1 3 GGAATT/SQEN

(GGAATT/SQEN AND SQL=6)

=> D L1 1 SEQ

L1 ANSWER 1 OF 3 REGISTRY COPYRIGHT 2008 ACS on STN

SEQ 1 ggaatt ←

=====

HITS AT: 1-6

# EXACT amino acid search (/SQEP) in USGENE

```
=> RUN GETSEQ SMAEP/SQEP
L1   RUN STATEMENT CREATED
L1           3 SMAEP/SQEP
```

```
=> D L1 1 SQL SEQ
```

```
L1 ANSWER 1 OF 3 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
```

```
SQL      5
SEQ      1 smaep
        =====
HITS AT: 1-5
```

```
=> D SEQ3
```

```
L1 ANSWER 1 OF 3 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
SEQ3     1 Ser-Met-Ala-Glu-Pro
        === === === === ===
HITS AT: 1-5
```

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

Amino acids can be displayed as single letter codes or three letter codes.

# EXACT nucleic acid search (/SQEN) in USGENE

```
=> 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 SEQUENCEBASE CORP on STN
```

```
SEQ      1 gccgccgt ←  
          =====
```

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

```
SQL      8
```

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

# Sequence Code Match

- **SCM**
  - Exact
  - **Exact Family**
  - Subsequence
  - Subsequence Family

# Sequence Code Match: Exact Family search

- Exact Family search
  - Matches the sequence query as entered and allows family substitution to occur
  - Retrieves identical sequences and family sequences with exact length
  - Family substitutions only occur for proteins and not nucleic acids

**/SQEFP = SeQ uence EXact FAmily P eptide**

# Amino acid families for SQEFP and SQSFP 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

# EXACT FAMILY amino acid search (/SQEFP) in REGISTRY

=> S DSDGP/SQEFP

59 DSDGP/SQEFP  
75132 SQL=5

L1 59 DSDGP/SQEFP  
(DSDGP/SQEFP AND SQL=5)

=> D L1 3 SEQ SQL

L1 ANSWER 3 OF 59 REGISTRY COPYRIGHT 2008 ACS on STN

SEQ 1 DGDGG  
=====

HITS AT: 1-5

SQL 5

In REGISTRY:  
**DSDGP/SQEP**  
retrieved 1 record.

Possible family  
substitutions for DSDGP:

<b>D</b>	<b>S</b>	<b>D</b>	<b>G</b>	<b>P</b>
Q	P	Q	P	A
N	A	N	A	<b>G</b>
E	<b>G</b>	E	S	S
B	T	B	T	T

# EXACT FAMILY amino acid search (/SQEFP) in USGENE

```
=> RUN GETSEQ SMAEP/SQEFP
```

```
L1 RUN STATEMENT CREATED
```

```
L1 23 SMAEP/SQEFP
```

```
=> D L1 2 SEQ SQL
```

```
L1 ANSWER 2 OF 23 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
```

```
SEQ 1 gites ←
```

```
=====
```

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

```
SQL 5
```

In USGENE:

**SMAEP/SQEP** retrieved 3 records.

**SMAEP/SQEFP** retrieved 23 records.

Possible family  
substitutions for SMAEP:

<b>S</b>	<b>M</b>	<b>A</b>	<b>E</b>	<b>P</b>
P	I	G	Q	A
A	L	T	N	G
G	V	P	D	S
T		S	B	T

# Sequence Code Match

- **SCM**
  - Exact
  - Exact Family
  - **Subsequence**
  - Subsequence Family

# Sequence Code Match: Subsequence search

- Subsequence search
  - Retrieves exact sequences plus sequences that are embedded in longer sequence

**/SQSP = SeQ uence S ubsequence P eptides**

**/SQSN = SeQ uence S ubsequence N ucleotides**

# SUBSEQUENCE amino acid search (/SQSP) in REGISTRY

=> S DSDGP/SQSP

L1 794 DSDGP/SQSP

=> D L1 132 SEQ SQL

L1 ANSWER 132 OF 794 REGISTRY COPYRIGHT 2008 ACS on STN

SEQ 1 GVPCLCDSDG PSVRGNALSG IIWLAGCPSG WHNCKKHGPT  
===== =  
IGWCCKQ

HITS AT: 7-11

SQL 47

The specific 5 amino acid sequence is embedded in a sequence that is 47 amino acid long.

# SUBSEQUENCE nucleic acid search (/SQSN) in REGISTRY

```
=> S GGAATT/SQSN AND SQL<25
```

```
25201 GGAATT/SQSN
```

```
5553085 SQL<25
```

```
L1 25201 GGAATT/SQSN AND SQL<25
```

SQL : Sequence Length

```
=> D L1 1 SEQ SQL
```

```
L1 ANSWER 1 OF 25201 REGISTRY COPYRIGHT 2008 ACS on STN
```

```
SEQ 1 ggaattcaga gacaacatgg
```

```
=====
```

```
HITS AT: 1-6
```

```
SQL 20
```

# SUBSEQUENCE amino acid search (/SQSP) in USGENE

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

```
L1 RUN STATEMENT CREATED
```

```
L1 260 KGPSYSLR/SQSP
```

```
=> D HIT SQL
```

```
L1 ANSWER 1 OF 260 USGENE COPYRIGHT 2009 on STN  
SEQ
```

```
kgpsyslr
```

```
=====
```

```
HITS AT: 479-486
```

```
SQL 497
```

HIT : free-of-charge display comprising only the hit part of the subject sequence plus all information on "HITS AT  
SQL : sequence length (also free)

# SUBSEQUENCE amino acid search (/SQSP) in USGENE

=> D SEQ SQL

L1 ANSWER 1 OF 260 USGENE COPYRIGHT 2009 on STN  
SEQ

```
1 mtvflsfaff aailthigcs nqrrspengg rrynriqhgq caytfilpeh
51 dngncresate qyntnalqrd aphvetdfss qklqhlehm enytqwlqkl
101 enyivenmks emaqiqqnav qnhtatmlei gtsllsqtae qtrkltdvet
151 qvlnqtsrle iqllenslst yelekqllqq tneilkiquek nsllhekile
201 megkhkeeld tlkeekenlq glvtrqtfii qelekqlsra tsnnsvlqkq
251 qllelmdtvhn lvslctkevl lkggkreeek pfrdcadvyq agfnksgiyt
301 iyfnnmpepk kvfcnmdivne ggwtviqhre dgsldfqrqw keykmgfgnp
351 sgeywlgnef ifaitsqrqy mlrielmdwe gnraysqydr fhignqkqny
401 rlylkgghtgt agkqsslilh gadfstkdad ndncmckcal mltggwwfda
451 cgpsnlngmf ytagqnhgkl ngikwhyfkg psyslrsttm mirpldf
                                == =====
```

HITS AT: 479-486

SQL 497

# SUBSEQUENCE nucleic acid search (/SQSN) in USGENE

```
=> RUN GETSEQ ACCCTGCAAA TAGCA/SQSN
```

```
L1 RUN STATEMENT CREATED
```

```
L1 49 ACCCTGCAAATAGCA/SQSN
```

STN will ignore spaces  
within the query sequence.

```
=> D L1 30 SEQ SQL
```

```
L1 ANSWER 30 OF 49 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
```

```
SEQ 1 tgtagtcat tatcatctt gtcacagct gaagatgaaa taggatgtaa
```

```
51 tcagacgaca caggaagcag attctgctaa taccctgcaa atagcagaaa
```

```
===== =====
```

```
101 taaaagaaaa gattggaact a
```

```
HITS AT: 82-96
```

```
SQL 121
```

# Sequence Code Match

- **SCM**
  - Exact
  - Exact Family
  - Subsequence
  - **Subsequence Family**

# Sequence Code Match: Subsequence Family search

- Subsequence Family search
  - Exact sequence match, subsequence match, and sequences that contain family substitution of amino acid

⇒ **S DSDGP/SQSFP**

**/SQSFP = SeQ uence S ubsequence F amily P eptides**

# SUBSEQUENCE FAMILY amino acid search (/SQSFP) in REGISTRY

=> S DSDGP/SQSFP

L1 1284147 DSDGP/SQSFP

=> S L1 AND SQL<25

5199939 SQL<25

L2 9538 L1 AND SQL<25

=> D L2 375 SEQ SQL

L2 ANSWER 375 OF 9538 REGISTRY COPYRIGHT 2008 ACS on STN

SEQ 1 ADWPGPPELD VCVEEA **EGEA** P  
 =====

HITS AT: 17-21

SQL **21**

In REGISTRY:  
**DSDGP/SQSP** retrieved 794 records.  
**DSDGP/SQSFP** retrieved 9538 records.

Possible family substitutions for DSDGP:

<b>D</b>	<b>S</b>	<b>D</b>	<b>G</b>	<b>P</b>
Q	P	Q	P	A
N	A	N	<b>A</b>	G
<b>E</b>	<b>G</b>	<b>E</b>	S	S
B	T	B	T	T

# SUBSEQUENCE FAMILY amino acid search (/SQSFP) in USGENE

=> RUN GETSEQ KGPSYSLR/SQSFP

L1 RUN STATEMENT CREATED

L1 2384 KGPSYSLR/SQSFP

=> D L1 73 SEQ SQL

L1 ANSWER 73 OF 2384 USGENE COPYRIGHT 2008 SEQUENCEBASE COR on STN

SEQ 1 gdtikiespg yltspgyphs yhpsekcewl iqapdpyqri minfnphfdl

51 edrdckydyv evfdgeneng hfrgkfcgki apppvvssgp flfikfvscy

101 ethgagfsir yei

=====

HITS AT: 103-110

SQL 113

In USGENE:

**KGPSYSLR/SQSP**

retrieved 102 records.

**KGPSYSLR/SQSFP**

retrieved 2384 records.

Possible family substitutions for KGPSYSLR:

<u>K</u>	<u>G</u>	<u>P</u>	<u>S</u>	<u>Y</u>	<u>S</u>	<u>L</u>	<u>R</u>
H	P	A	T	F	P	I	H
R	A	G	P	W	A	M	K
	S	S	A		G	V	
	T	T	G		T		

# Summary of the Sequence Code Match options

Search Type	Polypeptides	Nucleic Acids
EXACT	/SQEP	/SQEN
EXACT FAMILY	/SQEFP	—
SUBSEQUENCE	/SQSP	/SQSN
SUBSEQUENCE FAMILY	/SQSFP	—

# Agenda

- REGISTRY<sup>®</sup> database content
- USGENE<sup>®</sup> database content
- **Sequence Code Match searching (GETSEQ)**
  - Exact, family and subsequence
  - Variability searching
- Multifile patent sequence search example

# 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, PCTGEN, and REGISTRY database summary sheets:

[www.stn-international.com/stndatabases/databases/onlin\\_db.html](http://www.stn-international.com/stndatabases/databases/onlin_db.html)

# Variability symbols for 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

# Agenda

- REGISTRY<sup>®</sup> database content
- USGENE<sup>®</sup> database content
- Sequence Code Match searching (GETSEQ)
  - Exact, family and subsequence
  - Variability searching
- **Multifile patent sequence search example**

# 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:

**LGPX<sub>1</sub>QLCX<sub>2</sub>LVX<sub>3</sub>CAP**

**X<sub>1</sub> = V or L**

**X<sub>2</sub> = any amino acid except, G or H**

**X<sub>3</sub> = any amino acid**

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

# USGENE

## 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 will also be run in REGISTRY, but the SEARCH (=> S) command will be 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

=> 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

HITS AT: 10-23

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

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

# Repeat the USGENE search in REGISTRY and combine all results in CPlus<sup>SM</sup>

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

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

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

# Resources for sequence searching on STN

- *Sequence Searching on STN* modular workshop  
[www.fiz-k.com/bostonsequenceworkshop](http://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](http://www.sequencebase.com)
- CAS REGISTRY sequence coverage and resources  
[www.cas.org/support/stngen/stndoc/sequences.html](http://www.cas.org/support/stngen/stndoc/sequences.html)

**SSTN**®

Martine MICHEL  
[martine.michel@capadoc.fr](mailto:martine.michel@capadoc.fr)

CAPADOC  
[www.capadoc.com](http://www.capadoc.com)