



The Genesis of USGENE®

Martin Goffman, President and CEO, SequenceBase Corporation

Robert Austin, Regional Sales Manager, FIZ Karlsruhe, Inc.



From Concept to Content: The Genesis of USGENE

www.sequencebase.com/images/pdf/USGENE_Reprint_press.pdf



The Genesis of USGENE

3

- Who needs another database?
 - We do!
- Searchers viewpoints
 - Data quality is paramount
 - Availability of data ASAP
 - Comprehensive
 - Low cost

USGENE – Why a new database?

4

- My role and experience as a searcher
 - Recognize the current need to consult several disparate sources for a comprehensive search
 - Time consuming
 - Costly
 - Prone to errors
 - Currently there are
 - Data quality issues
 - Lack of Timeliness
 - Constant check for additions and corrections

USGENE

Questions an entrepreneur asks

- Why a new database? (business opportunity)
 - Identify Need
 - Is there a need?
 - Is there a defined user base?
 - Is there a suitable host?
 - Can we do it?
 - Can we provide value to our users?
 - Can we make money?

USGENE – Is there a need?

6



- Is there a need?
- Enter Robert Austin (co-author) in early 2004
 - Identified a need for a single source of US Patent Sequence Data
 - Current sources not complete
 - NCBI has no published application sequence data
 - Missing the majority of mega publications
 - Missing many relevant US Patents
 - Not timely
 - Developed a “cumbersome methodology for sequence searching involving numerous government and commercial sources”

- “In Austin's training sessions, users consistently requested a single source of sequence data”
 - As a searcher I could immediately identify with this need since I needed it myself
 - Market quickly defined
 - Patent Offices
 - Pharmaceutical Companies
 - Drug Discovery
 - Biotechnology
 - Law Firms

USGENE – is there a suitable host?

- Is there a suitable database host?
 - No desire to become an online database service a la STN[®], Questel, Dialog, Thomson Reuters
 - Reinvent the wheel
 - Takes time, resources, and money
- Enter FIZ Karlsruhe and STN International
 - Hosts the existing industry standard GENESEQ[™]
 - STN also features CAS Registry and PCTGEN
 - Familiar working with and searching in STN databases
 - Fits into our goal of providing complementary rather than competing database to those existing and in use
 - USGENE completes the remaining piece of the jigsaw puzzle for STN International and its customers

USGENE – can we do it?

- Can we do it?
 - Need for lots of creative and brilliant developers
 - Need to consolidate many disparate data sources
 - Need to clean up data and bring it all together (later slide shows this)
 - Need for great working partners like FIZ Karlsruhe
 - Like any dream – a lot more time and work involved than one would guess
- BUT, FINAL ANSWER
 - **YES!**
 - **SequenceBase** Corporation formed

- Can we provide value to our users?
 - We believe that we provide exceptional value.
 - Payback is in quality of the data and time saved in performing a quality search.
 - No need to download and search disparate sources
 - Data delivered by Friday of each week
 - Subscribers get data Thursday evening
 - Consistent availability of data within 3 days of publication
 - One day within publication of applications (same day for subscribers)

Formation of SequenceBase Corporation

11

Since the answers to all of the foregoing questions were in the affirmative, we decided to go ahead!

- Decision was made to proceed
- Team was assembled
 - Robert Austin agreed to serve as technical advisor
- Agreement with FIZ Karlsruhe signed in 2005
- Database production initialized
- Goals were established
 - Data Quality **FIRST** and **FOREMOST**
 - **TIMELY DELIVERY OF DATA**
 - Consistent delivery of USGENE by **Friday** of the week of publication (within 3 days of publication)
 - Subscribers have the data **Thursday evening** (same day as applications are published)

- Can we make money?
 - Established a **financially sound business model**
 - **Mutually beneficial agreement** with FIZ Karlsruhe
 - USGENE has been a **commercial success** on STN
 - Very rapid acceptance
 - Higher than anticipated level of usage
 - Subscription sales for in-house use taking off
 - Financially secure position
 - Future of SequenceBase assured



How is the database constructed?

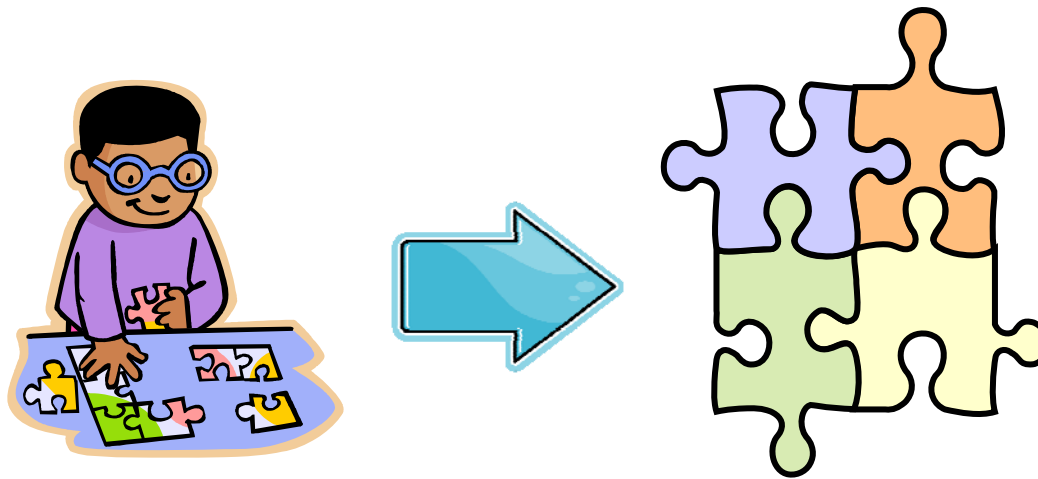


- Sequences from all relevant USPTO published patent applications and granted (issued) patents
- Assignee and full inventor names; publication, application and parent case PCT numbers and dates; original publication **title**, **abstract** and **claims**
- Organism name, sequence length, Molecule Type, SEQ ID and feature tables for features/annotations
- Updated weekly – on Thursday to subscribers and available each Friday on STN
- 1982 – present



Database Content

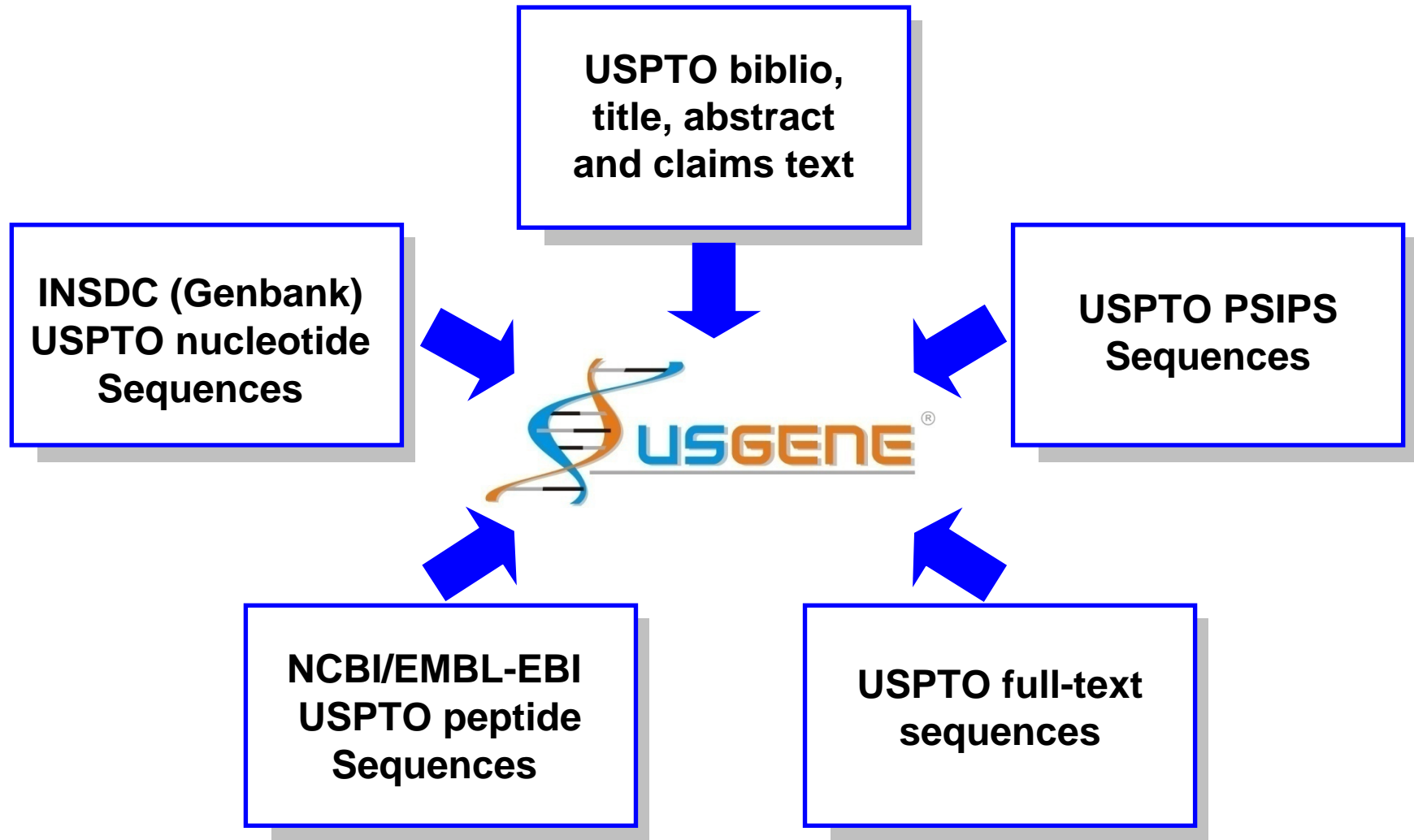
Putting the puzzle together



USGENE consolidates unique USPTO sequence data from different sources

1. The International Nucleotide Sequence Database Collaboration (INSDC) (Genbank)
 - U.S. patent nucleotide sequences, 1982-date
2. The USPTO Protein Database (NCBI/EMBL)
 - U.S. patent peptide sequences, 1982-date
3. The USPTO Publication Site for Issued and Published Sequences (PSIPS)
 - The mega-publication download site, 2001-date
4. Sequences from the USPTO granted patents and published applications full-text
 - Filling in coverage gaps and to enhance timeliness

USGENE combines sequences with bibliographic data and claims text



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

```
L1 ANSWER 1 OF 1 USGENE COPYRIGHT 2009 SEQUENCEBASE CORP on STN
AN 20090019558.6 Protein USGENE
TI PGDS AS MODIFIERS OF THE PTEN PATHWAY AND METHYLATION
(Published Application)
IN Song Chunyan (Foster City, CA); Ollmann Michael (Foster
City, CA); Bjerke Lynn Margaret (Surrey, GB)
PA EXELIXIS INC (South San Francisco CA)
PI US 20090019558 A1 20090115
AI US 2005-628637 20050620
RLI WO 2005-US22087 20050620
ED 20090116
AB Human PGD genes and methods comprising the steps of: (a) providing
thus are therapeutic agents for the treatment of cancer, and
PTEN function. Methods comprising the steps of: (a) providing
screening for agents that modulate PTEN function; (b) providing
ECLM US20090019558 A1 1
modulating agent; (c) providing the test agent; (d) providing
an assay system comprising a PGD polypeptide or nucleic acid; (b)
contacting the assay system with a test agent under conditions
whereby, but for the presence of the test agent, . . . .
SSO PROTEIN; USPTO; APPLICATION
ORGN Homo Sapiens
SQL 483
SEQ
1 maqadialig lavmgqnlil nmndhgr
51 kvvgaqslke mvsklkkpr iillvka
101 ggnseyrdtt rrcrdlkakg ilfvgs
. . . .
```

AN 20090019558.6 is SEQ ID NO: 6 from US20090019558.

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

AN 20090019558.6 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 STN Viewer

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

This USPTO example is US5210028, which was issued on May 11th, 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: CATATGGCATACCG CCCGAGCGAGACCCCTGTGCGGTGGCGAGCTCGTAGACACTCTGCAG60
TTCGTTTGTGGTGACCGTGGCTTCTACTTCTCTCGTCCTGCTAGCCGTGTATCTCGCCGT120
TCTAGAGGCATCGTTGAAGAGTGCTGTTTCCGCAGCTGTGACTACTGCGCAACTCCAGCAAATCCGAATAAGGATCC216 (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: Synthetic Human IGF-II DNA Sequence; E.coli pref (B) STRAIN: pIGF-II/3 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:3: GAATTCGACGCTTATGGCTTACAGACCATCCGAAACCTTGTACACCTTGCAATTCGTTTGTGGTGACAGAGGTTT CTA

TTCTAGAAGATCCAGAGGTATCGTTGAAGAATGTTGTTTCAAGTTTGAAACCTACTGTGCTACCCAGCTAAGTCTGAATGAATGCGTCAATTC233 (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: Synthetic Human IGF-II DNA Sequence; E.coli pref (B) STRAIN: pPLMu (xi) SEQUENCE DESCRIPTION: SEQ ID NO:4: CTTACACTTAGTTAAATTGCTAACTTTATAGATTACAAAACCTTACACCATGGTTACGAATTC

CCCGGGGATCCGTCGACCTGCAGCCAAGGTTTCGGTGATGACGGTGAAAACCTCTGAC 150

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

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

Done

To facilitate precise searching all USGENE sequences are in standardized format

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

AN 5210028.1 is SEQ ID NO: 1 from US5210028.

Each USGENE sequence record includes searchable bibliographic details.

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.

The original format of a USGENE sequence is available for display using the SEQO display

```
=> S 20070224666.21/AN  
L1          1 20070224666.21/AN
```

USGENE Accession Numbers (/AN) comprise the publication number + the sequence identity number (SEQ ID NO).

```
=> D TRI SEQO
```

```
L1 ANSWER 1 OF 1 USGENE COPYRIGHT 2009 SEQUENCEBASE CORP on STN  
TI Alleles of the zwf gene from coryneform bacteria  
(PublishedApplication)
```

```
MTY DNA  
SQL 1263  
SEQO
```

Often the SEQO original format includes the patent applicant's alignment of the nucleotide sequence coding region with its corresponding protein sequence.

```
gtg gcc ctg gtc gta cag aaa t  
Met Ala Leu Val Val Gln Lys T  
1          5          10          15          20          25          30          35          40          45          96  
gaa cgc att aga aac gtc gct gaa cgg atc gtt gcc acc aag aag gct  
Glu Arg Ile Arg Asn Val Ala Glu Arg Ile Val Ala Thr Lys Lys Ala  
20          25          30          35          40          45          96  
gga aat gat gtc gtg gtt gtc tgc tcc gca atg gga gac acc acg gat 144  
Gly Asn Asp Val Val Val Val Cys Ser Ala Met Gly Asp Thr Thr Asp  
35          40          45          96  
. . . . .
```

In contrast, NCBI/EMBL/DDBJ patent records have minimal bibliographic and text data

23

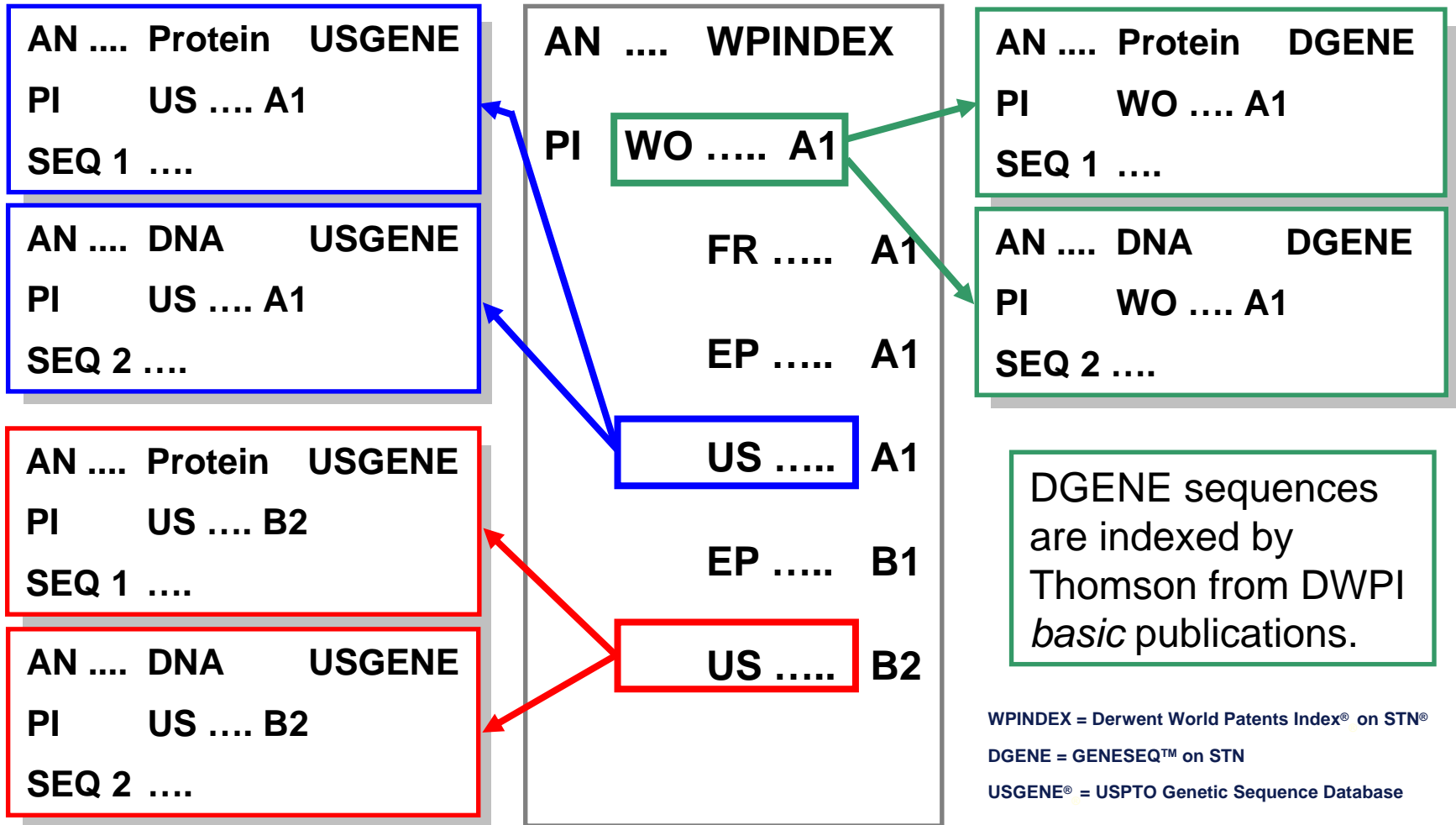
General Information			
Accession #	AAA00521		
SRS Entry ID	USPO_PRT:AAA00521		
Molecule Type	PRT		
Sequence Length	40		
Entry Data Class	STANDARD		
Sequence Version	AAA00521.1		
Creation Date	21-MAY-1993		
UniParc	UPI0000035113		
Description			
Description	Sequence 1 from Patent US 4563352.		
Organism	Unknown		
References			
1.	Rivier; J.E.F.; Spiess; J. and Vale; W.W. Jr.; Human pancreatic GRF Patent number US4563352 -A/1 07-JAN-1986; The Salk Institute For Biological Studies; San Diego, CA		
	Position	1-40	
Features			
Key	Location	Qualifier	Value
source	1..40		
Sequence			
Characteristics	Length: 40 AA		
Sequence	<pre>>uspo_prt AAA00521 AAA00521 Sequence 1 from Patent US 4563352. YADAIFTNSYRKVLGQLSARKLLQDIMSRQOGESNQERGA</pre>		

Reminder: NCBI/EMBL/DDBJ cover sequences from U.S. granted patents – sequences from U.S. published applications are not covered .

USGENE is an essential additional tool for tackling business critical searches

- Thomson Reuters GENESEQ and CAS Registry provide value-added sequence data from the DWPI or CAplus basic publication respectively
- USGENE provides sequence data from all basic and equivalent **US published applications** and **granted patents** within 3 days of publication
- Sequence listing variation often occurs between basic publication and granted patent stages
 - Especially important e.g. in Freedom to Operate searches

USGENE captures sequence data from all available U.S. patent family members



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

```

L1 ANSWER 1 OF 1 WPINDEX COPYRIGHT 2009 THOMSON REUTERS on STN
AN 1994-358278 [44] WPINDEX
TI New polynucleotide(s) specific for hepatitis C virus types 4, 5 and 6 -
and related antigenic peptide(s) and antibodies, useful in vaccines,
diagnosis, HCV typing and treatment
DC B04; D16; S03
IN PIKE I H; SIMMONDS P; YAP P L
PA (COMM-N) COMMON SERVICES AGENCY; (MURE-N) MUREX DIAGNOSTICS INT INC; . . .
PI WO 9425602 A1 19941110 (199444)* EN 70[5]
AU 9465797 A 19941121 (199508) EN
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 2005
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

27

Patent Kind covered in USGENE (STN field /PK)

- 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)

Results of a multifile sequence search using USGENE and GENESEQ on STN

Search Question:

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

VRSSSRTPSDKPVAVVAVNPQAEGQLQWLNRRANALLANGVELRDNQLVVPSEGLYLIYSQVLF
KGQGCPSTHVLLTHTISRIVSYQTKVNLLSAIKSPCQRETPRGAEAKPWYEPYIYLGGVFQLEK
GDRLSAEINRPDYLDFAESGQVYFGIIAL

(Search conducted on October 22nd, 2008.)

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

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

Example: USGENE unique retrieval

```
L14 ANSWER 541 OF 1859 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
    FAMILY 20
TI Fusion proteins comprising gp39 and CD8 (Patent)
MTY protein
SQL 157
ORGN Not provided
SEQN 5
SEQC 9
SCORE 313 99% of query self score 316
```

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

BLASTALIGN

```
Query = 157 letters
Length = 157
Score = 313 bits (803), Expect = 6e-91
Identities = 155/157 (98%), Positives = 156/157 (98%)
Query: 1 VRSSSRTPSDKPVAHVVANPQAEGQLQWLNRRANALLANGVELRDNQLVVPSEGLYLIYS
VRSSSRTPSDKPVAHVVANPQAEGQLQWLNRRANALLANGVELRDNQLVVPSEGLYLIYS
Sbjct: 1 VRSSSRTPSDKPVAHVVANPQAEGQLQWLNRRANALLANGVELRDNQLVVPSEGLYLIYS
Query: 61 QVLFKGQGCPSTHVLLTHTISRIAVSYQTKVNLLSAIKSPCQRETPRGAEAKPWYEPIYL
QVLFKGQGCPSTHVLLTHTISRIAVSYQTKVNLLSAIKSPCQRETP GAEAKPWYEPIY+
Sbjct: 61 QVLFKGQGCPSTHVLLTHTISRIAVSYQTKVNLLSAIKSPCQRETPEGAEAKPWYEPIYI
Query: 121 GGVFQLEKGDRLSAEINRPDYLDFAESGQVYFGIIAL 157
GGVFQLEKGDRLSAEINRPDYLDFAESGQVYFGIIAL
Sbjct: 121 GGVFQLEKGDRLSAEINRPDYLDFAESGQVYFGIIAL 157
```

Example: USGENE unique retrieval (cont.)

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

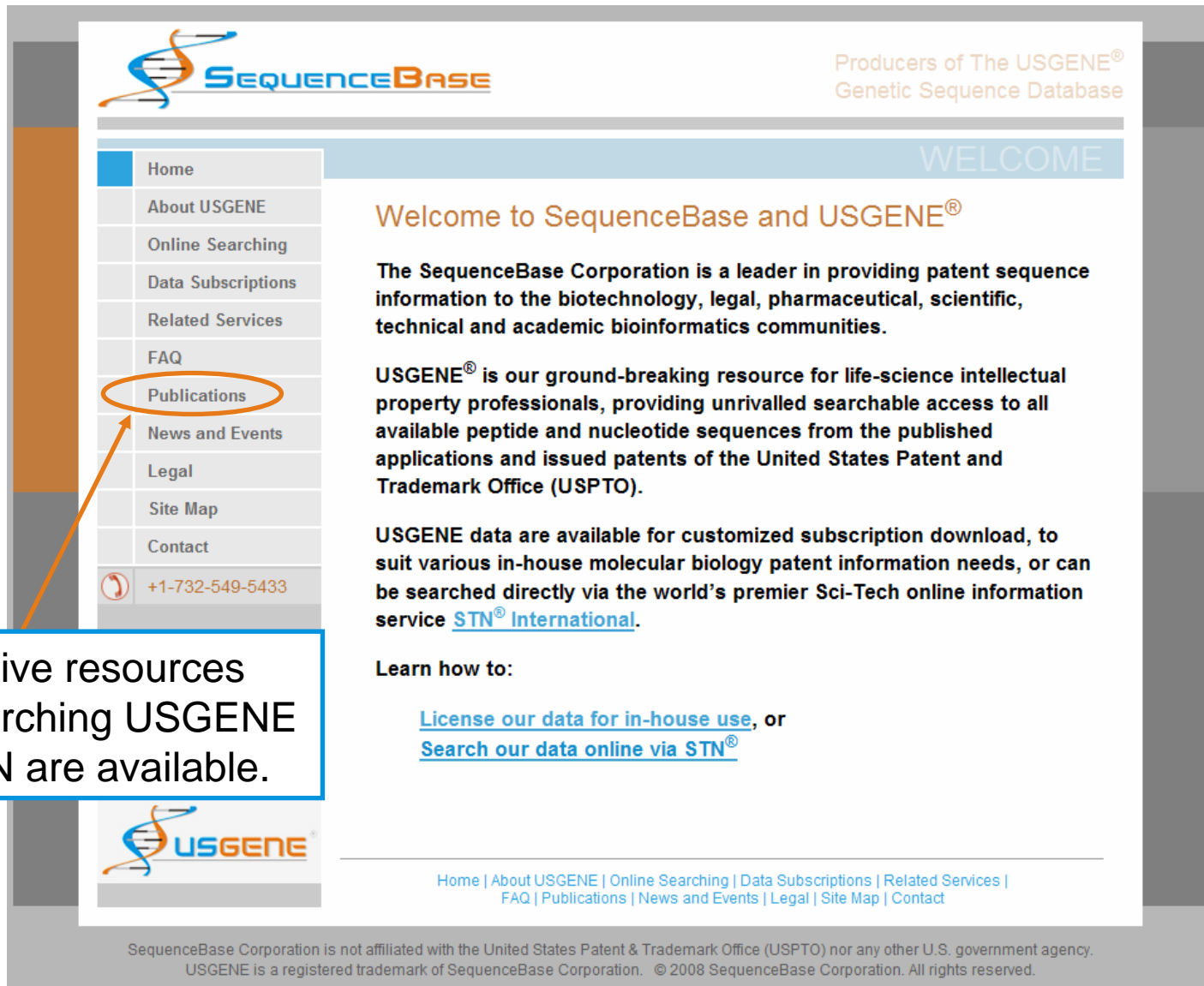
This relevant DWPI record was uniquely retrieved via a USGENE BLAST search (on Oct 22nd, 2008).

How does USGENE compare to other USPTO sequence data sources?

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



Other Distribution Channels



SEQUENCEBASE

Producers of The USGENE®
Genetic Sequence Database

WELCOME

Home
About USGENE
Online Searching
Data Subscriptions
Related Services
FAQ
Publications
News and Events
Legal
Site Map
Contact

+1-732-549-5433

Welcome to SequenceBase and USGENE®

The SequenceBase Corporation is a leader in providing patent sequence information to the biotechnology, legal, pharmaceutical, scientific, technical and academic bioinformatics communities.

USGENE® is our ground-breaking resource for life-science intellectual property professionals, providing unrivalled searchable access to all available peptide and nucleotide sequences from the published applications and issued patents of the United States Patent and Trademark Office (USPTO).

USGENE data are available for customized subscription download, to suit various in-house molecular biology patent information needs, or can be searched directly via the world's premier Sci-Tech online information service [STN® International](#).

Learn how to:

[License our data for in-house use](#), or
[Search our data online via STN®](#)

Home | About USGENE | Online Searching | Data Subscriptions | Related Services |
FAQ | Publications | News and Events | Legal | Site Map | Contact

SequenceBase Corporation is not affiliated with the United States Patent & Trademark Office (USPTO) nor any other U.S. government agency.
USGENE is a registered trademark of SequenceBase Corporation. © 2008 SequenceBase Corporation. All rights reserved.

Extensive resources
for searching USGENE
on STN are available.

Conclusions

- We are financially sound and well established in the marketplace
- Our goal of producing consistently high quality data has been met by our constantly improving proprietary algorithms
- We are dependable and constantly deliver our quality data reliably and on time
- Our subscriber base is pleased



The Genesis of USGENE®

Martin Goffman, President and CEO

SequenceBase Corporation

Tel: +1 732 549-5433

E-mail: mgoffman@sequencebase.com

www.sequencebase.com