



Enhancements for exact sequence searching with RUN GETSEQ in DGENE, USGENE and PCTGEN

Maximum answer limit increased to 250,000, new free-of-charge HIT (synonym: ALIGN) display available

Exponentially growing numbers of sequences in the biosequence databases DGENE, USGENE and PCTGEN call for an expansion of the answer number limit for sequence code match (SCM) searches and adaptations for a better review of large answer sets resulting from this search option. For this purpose the answer number limit for a GETSEQ search has been increased to 250,000 answers in the three biosequence files DGENE, USGENE and PCTGEN, and an adapted free-of-charge HIT (syn. ALIGN) display format has been made available. You may now benefit from extended system limits and a possibility for a cost effective way to scan large answer sets for relevant hits after sequence code match (SCM) searching.

As of May 3rd 2009, the general answer set limit for GETSEQ searches in DGENE, USGENE and PCTGEN has been increased to 250,000. For answer sets comprising between 25,000 and 250,000 answers now multiple L-numbers containing 25,000 hit

sequences each will be created. A system message will inform about the expected number of answers and L-numbers. In a second step, the resulting L-numbers then need to be merged via the SEARCH command generating the complete answer set with up to 250,000 answers. Subsequently, the complete answer set may be refined with text and/or date terms if desired.

A free-of-charge HIT (synonym: ALIGN) display format allows for a cost effective review of GETSEQ answers. The HIT (ALIGN) format contains the part of the hit sequence with the matching residues which are highlighted with double underlining. In addition, the information HITS AT: will give the residue numbers of the start and end point of the matching part of the hit sequence. If multiple matching parts are found in one hit sequence the HIT (ALIGN) display will only show the first appearance of a matching part within the hit sequence (with matching residues underlined) plus the residue numbers of start and end points of all matching parts of the hit sequence with HITS AT:. To display the complete hit sequence with all matching parts highlighted with double underlining plus all HITS AT: please use the SEQ display format in each file.

=> **FIL DGENE**

FILE 'DGENE' ENTERED AT 11:34:40 ON 22 APR 2009
 COPYRIGHT (C) 2009 THOMSON REUTERS

=> **RUN GETSEQ ARG/SQSP**

RUN GETSEQ AT 11:34:48 ON 22 APR 2009
 COPYRIGHT (C) 2009 FIZ KARLSRUHE GMBH

Number of answers 229895 will create 10 Answer Sets

```
L1 RUN STATEMENT CREATED
L1 25000 ARG/SQSP
L2 RUN STATEMENT CREATED
L2 25000 ARG/SQSP
L3 RUN STATEMENT CREATED
L3 25000 ARG/SQSP
L4 RUN STATEMENT CREATED
L4 25000 ARG/SQSP
L5 RUN STATEMENT CREATED
L5 25000 ARG/SQSP
L6 RUN STATEMENT CREATED
L6 25000 ARG/SQSP
L7 RUN STATEMENT CREATED
L7 25000 ARG/SQSP
L8 RUN STATEMENT CREATED
L8 25000 ARG/SQSP
L9 RUN STATEMENT CREATED
L9 25000 ARG/SQSP
L10 RUN STATEMENT CREATED
L10 4895 ARG/SQSP
```

For answer sets >25,000 hits, multiple L-numbers with 25,000 answers each will be generated. A system message informs about the expected number of hit documents and resulting L-numbers.

In the next step, merge the resulting L-numbers with the SEARCH command to retrieve the complete answer set with all hit sequences.

=> **S L1-L10**

L11 229895 (L1 OR L2 OR L3 OR L4 OR L5 OR L6 OR L7 OR L8 OR L9 OR L10)

=> **D 11 HIT**

```
L11 ANSWER 11 OF 229895 DGENE
SEQ
      arg
      ===
HITS AT: 121-123; 191-193
```

A free-of-charge HIT display format allows for a cost effective review of answers. The HIT display format contains only the **first** appearance of a matching part within the retrieved hit sequence plus information about the sites of all matching parts within the complete hit sequence with HITS AT: . For a display of the complete hit sequence with all matching parts please use the SEQ format.

=> **D 11 SEQ**

```
L11 ANSWER 11 OF 229895 DGENE COPYRIGHT 2009 THOMSON REUTERS on STN
SEQ
      1 vkgivlaggs gtrlhpltva fskqllplyd kpmiyyplst lmlggvrefl
      51 iistpadlpl frkllgtgae lglrfsyaeq qrpagiaeaf rigadfvvdp
      101 pvslilgdni fhspqlpql argmaevdgc alfghtvadp rpygvvekda
           ===
      151 egrlvgieek parprsseiv tglyvysadv velahrirps argeleitdv
           ===
      201 nrhylaqgra rlhslgpdst wldagtydgl ldaaafvrse qrrgiriapc
      251 eeiafrmgyi dadalyrlgs rrqnsygyry lmdisrgave agvga
HITS AT: 121-123; 191-193
```

=> **S L11 and claim?/PSL and uroathic/KW and PY>2007**

```
12154735 CLAIM?/PSL
268551 UROPATHIC/KW
1278706 PY>2007
      (PY>2007)
L12 688 L11 AND CLAIM?/PSL AND UROPATHIC/KW AND PY>2007
```

The complete GETSEQ answer set may be refined with text and/or date terms if desired using the search fields available in the files.

=> D TRIAL ALIGN

L12 ANSWER 1 OF 688 DGENE COPYRIGHT 2009 THOMSON REUTERS on STN
 AN AWH85252 protein DGENE

TI New isolated anti-Tweak receptor (TweakR) Binding fragment, useful for treating a cancer, breast cancer, colorectal cancer, pancreatic cancer, in a subject.

After refinement of the complete answer set with text and/or date terms the ALIGN display format may be used, which is synonymous to HIT after a GETSEQ search.

DESC Human TNF superfamily receptor 12A protein sequence, SEQ ID 2.

KW tumor suppressor; antibody therapy; therapeutic; solid tumor; cytostatic; bladder cancer; uropathic; colorectal tumor; gastrointestinal-gen.; lung tumor; respiratory-gen.; melanoma; dermatological; pancreas tumor; ovary tumor; endocrine-gen.; gynecological; renal tumor; nephrotropic; head and neck tumor; esophagus tumor; uterus tumor; stomach tumor; uterine cervix tumor; glioblastoma; sarcoma; TNF superfamily receptor 12A; TNFRSF12A; TweakR; Fn14; BOND_PC; type I transmembrane protein Fn14; TNFRSF12A; FN14; TWEAKR; CD266; tumor necrosis factor receptor superfamily, member 12A; tumor necrosis factor receptor superfamily, member 12A, isoform CRA b; type I transmembrane protein; GO1525; GO4872; GO5515; GO6915; GO6928; GO7155; GO7275; GO16020; GO16021; GO30154.

SQL 129
 SEQ

Please note that the resulting hit sequences after a GETSEQ search are not relevance-sorted and cannot be sorted according to a SCORE value (as known from the similarity search options of BLAST and GETSIM).

arg
 ===
 HITS AT: 2-4

=> FSORT L12

SET SMARTSELECT ON
 SET COMMAND COMPLETED
 SET HIGHLIGHTING OFF
 SET COMMAND COMPLETED

You may sort the answer set into patent families with FSORT bringing together hit sequence documents from the same invention. In this example 688 documents are grouped into 45 multi-record families and 25 individual record families.

SET AUDIT OFF
 SET COMMAND COMPLETED

SEL L12 1- PN,APPS
 L13 SEL L12 1- PN APPS : 342 TERMS

'L13' DELETED
 L13 688 FSO L12

45 Multi-record Families	Answers 1-663
Family 1	Answers 1-7
Family 2	Answers 8-13
Family 3	Answers 14-15
Family 4	Answers 16-25
Family 5	Answers 26-36
Family 6	Answers 37-38
Family 7	Answers 39-42
Family 8	Answers 43-61
Family 9	Answers 62-67
Family 10	Answers 68-71
Family 11	Answers 72-73
Family 12	Answers 74-115
Family 13	Answers 116-117
Family 14	Answers 118-125
Family 15	Answers 126-128
Family 16	Answers 129-131
Family 17	Answers 132-137
Family 18	Answers 138-149
Family 19	Answers 150-161
Family 20	Answers 162-164
Family 21	Answers 165-198
Family 22	Answers 199-204
Family 23	Answers 205-373
Family 24	Answers 374-375
Family 25	Answers 376-421
Family 26	Answers 422-431
Family 27	Answers 432-435
Family 28	Answers 436-441
Family 29	Answers 442-443
Family 30	Answers 444-447
Family 31	Answers 448-449
Family 32	Answers 450-464

```

Family 33           Answers 465-475
Family 34           Answers 476-501
Family 35           Answers 502-503
Family 36           Answers 504-538
Family 37           Answers 539-542
Family 38           Answers 543-544
Family 39           Answers 545-585
Family 40           Answers 586-592
Family 41           Answers 593-607
Family 42           Answers 608-611
Family 43           Answers 612-613
Family 44           Answers 614-645
Family 45           Answers 646-663
25 Individual Records  Answers 664-688
0 Non-patent Records

```

```

SET SMARTSELECT OFF
SET COMMAND COMPLETED

```

```

SET HIGHLIGHTING ON
SET COMMAND COMPLETED

```

```

SET AUDIT ON
SET COMMAND COMPLETED

```

```

=> D PFAM 1- 1 TRIAL ALIGN

```

```

L13 ANSWER 1 OF 688 DGENE COPYRIGHT 2009 THOMSON REUTERS on STN FAMILY 1
AN AWH94009 protein DGENE
TI New isolated antibody and its antigen-binding fragment that bind to
specific neublastin peptide sequence, useful for detecting and
quantifying neublastin polypeptides from samples like serum and saliva,
and for antagonizing neublastin.
DESC Human mature neublastin (R48E/R51E) mutant protein, SEQ ID NO:9.
KW antibody production; antibody therapy; therapeutic; protein
quantitation; protein detection; cell growth; neuron; signal
transduction; neuropathy; neuroprotective; neuropathic pain; analgesic;
breast tumor; testis tumor; cytostatic; endocrine-gen.; uropathic;
esophagus tumor; gastrointestinal-gen.; gastrointestinal tumor; colon
tumor; connective tissue neoplasm; musculoskeletal-gen.renal tumor;
nephrotropic; lung tumor; respiratory-gen.; small-cell lung cancer;
ovary tumor; gynecological; skin cancer; dermatological; stomach tumor;
uterus tumor; Artemin ligand; ARTN; mutein.
SQL 113
SEQ
arg
===
HITS AT: 13-15

```

```

L13 ANSWER 8 OF 688 DGENE COPYRIGHT 2009 THOMSON REUTERS on STN FAMILY 2
AN AVA92966 protein DGENE
TI New targeted binding agent that specifically binds to kinase insert
domain-containing receptor (KDR) and inhibits receptor dimerization,
useful for treating a malignant tumor in an animal, and for treating a
non-neoplastic disease.
DESC Human anti-KDR antibody germline heavy chain variable region, SEQ ID
138.
KW therapeutic; protein therapy; dimerization; melanoma; cytostatic;
dermatological; small-cell lung cancer; respiratory-gen.; non-small-cell
lung cancer; breast tumor; gynecological; prostate tumor; uropathic;
glioma; hepatocellular carcinoma; gastrointestinal-gen.; hepatotropic;
thyroid tumor; endocrine-gen.; stomach tumor; ovary tumor; bladder
cancer; lung tumor; glioblastoma; endometroid carcinoma; renal tumor;
nephrotropic; colon tumor; pancreas tumor; esophagus tumor; head and
neck tumor; mesothelioma; sarcoma; biliary tumor; ocular disease;
ophthalmological; inflammatory disease; cardiovascular disease;
cardiovascular-gen.; sepsis; antimicrobial-gen.; KDR gene; VEGF-2
receptor; heavy chain variable region.
SQL 116
SEQ
arg
===
HITS AT: 97-99

```

With the display patent family (D PFAM) command you may display bibliographic information from each invention plus hit sequence information that triggered the respective hit. Please note that unlike after BLAST and GETSIM similarity searches the first hit in a patent family after GETSEQ is not necessarily the most relevant one and display of more than the first hit sequence may be favourable.

```
L13 ANSWER 688 OF 688 DGENE COPYRIGHT 2009 THOMSON REUTERS on STN
AN ARZ09526 protein DGENE
TI New fusion protein comprising a structure domain of human F-box protein
and A-B box structure domain of human RB protein, useful for preventing
or treating disease caused by HPV 18 infection, preferably cervical
carcinoma.
DESC Human F-box protein Fbw1A-RB1 fusion protein, SEQ ID 2.
KW fusion protein; vector; prophylactic to disease; protein therapy;
therapeutic; papillomavirus infection; virucide; uterine cervix tumor;
cytostatic; gynecological; uropathic; F-box only protein gene; RB1 gene.
SQL 691
SEQ
      arg
      ===
HITS AT: 161-163
```

The presented search strategy started with a quite unspecific sequence code match search in DGENE generating 229,895 answers. Using text and date terms for refinement of the answer set and the STN FSORT command to group together patent inventions lead to 70 patent families. The new free-of-charge HIT (ALIGN) display format was used for a cost effective review and relevance check. Likewise, such search strategies can be used in the USGENE and the PCTGEN files, taking advantage of the enhanced answer limit and the free-of-charge HIT (ALIGN) display format.

FIZ Karlsruhe

STN Europe
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen, Germany

E-mail: helpdesk@fiz-karlsruhe.de
Phone: +49 7247 808 555
Fax: +49 7247 808 259
www.fiz-karlsruhe.de

CAS

STN North America
P.O.Box 3012
Columbus, Ohio 43210-0012
U.S.A.

E-mail: help@cas.org
Phone: 1 614 447 3700
Fax: 1 614 447 3798
www.cas.org

JAICI

STN Japan
Nakai Building
6-25-4 Honkomagome, Bunkyo-ku
Tokyo 113-0021, Japan

Phone: +81-3-5978-3621
Fax: +81-3-5978-3600
E-mail: cas-stn@jaici.or.jp
www.jaici.or.jp

Imprint

Editors:
FIZ Karlsruhe – Gerhard Herlan
CAS – Alena Miller

STNews is written and produced cooperatively by
CAS and FIZ Karlsruhe and printed in three separate
editions.

For the European edition © FIZ Karlsruhe 2008

Quoting or republishing of material from the STNews
is encouraged provided that acknowledgement is
made of the STNews as the source. FIZ Karlsruhe
requests that a copy of the reproduced material be
sent to FIZ Karlsruhe.