CAS STNEXT® WEBINAR

USING THE SEQUENCE CODE MATCH SEARCH TOOL TO EFFECTIVELY SEARCH CDRS

Jim Brown, FIZ Inc.

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Agenda

- STN sequence searchable databases and search methods
- Sequence Code Match (SCM) search options
- SCM variability: Motif searching symbols
- Antibody/CDR search example
- Hypothetical multiple CDR search discussion

CAS STNext sequence searchable databases

GENESEQ™

- Produced by Clarivate Analytics

USGENE

- Produced by Clarivate Analytics

PATGENE

- Produced by FIZ Karlsruhe

CAS REGISTRY®

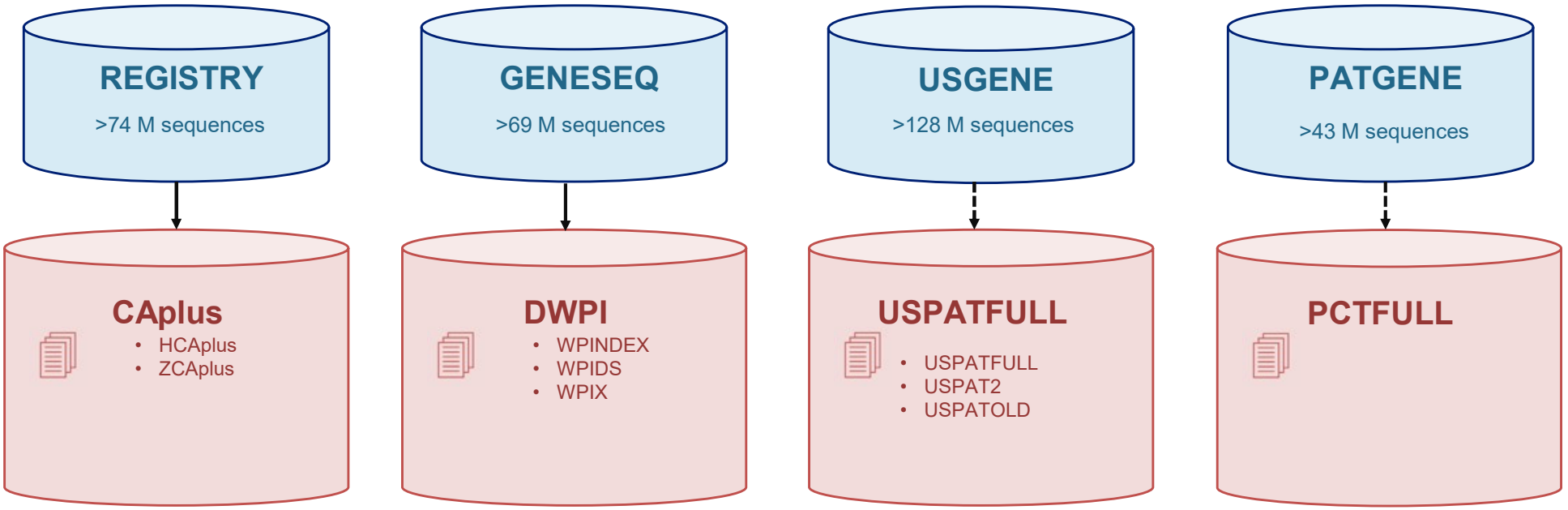
- Produced by CAS

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Linking to bibliographic information

Sequence files can be linked to bibliographic files

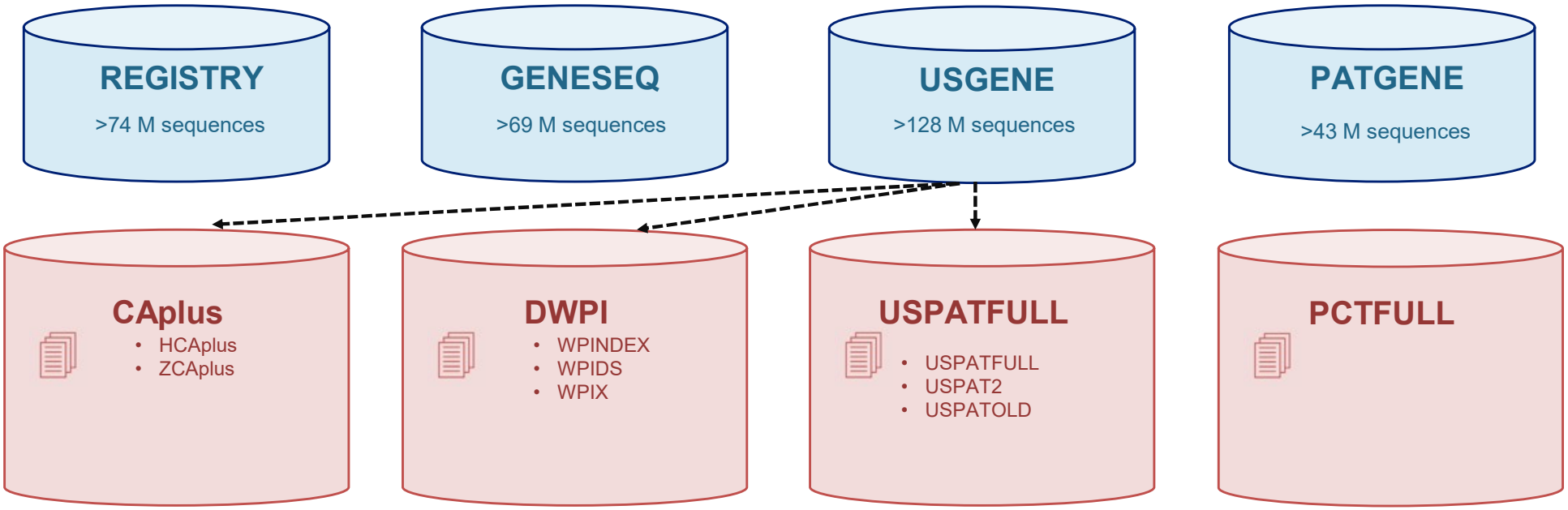


FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Linking to bibliographic information

Sequence files can be linked to bibliographic files

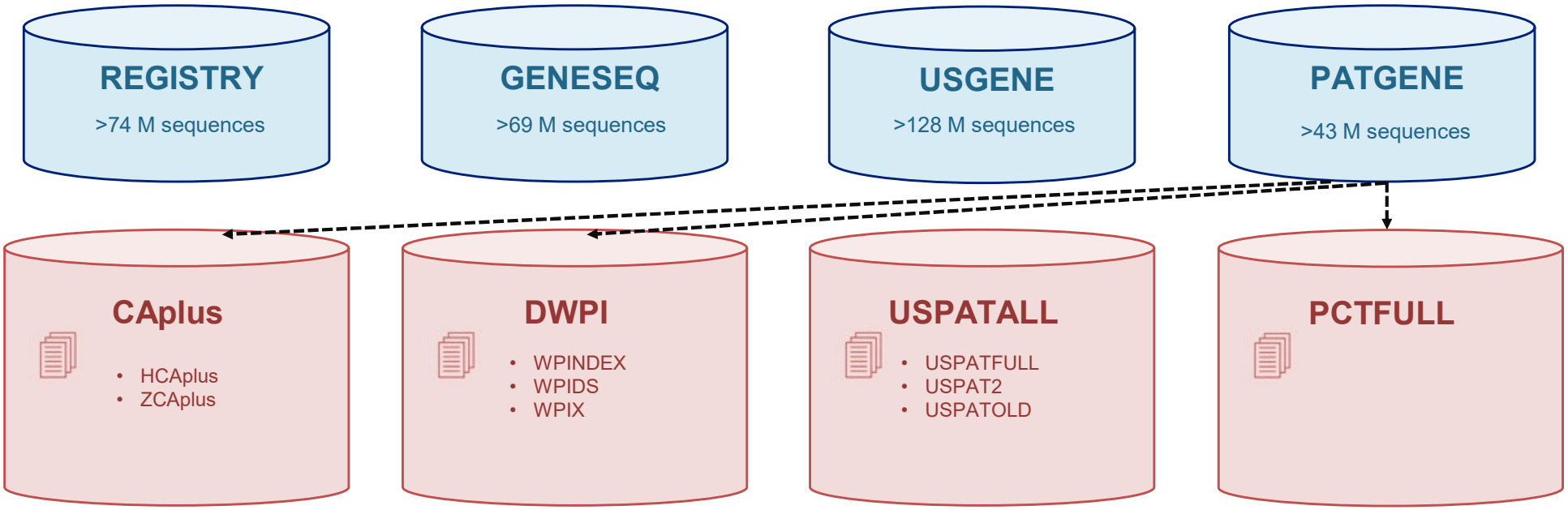


FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Linking to bibliographic information

Sequence files can be linked to bibliographic files



FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Sequence search methods

- **BLAST** NCBI BLAST for advanced similarity searching

- **GETSIM** FASTA based search algorithm; comparison of entire sequence length; more computational time required. Not available in REGISTRY

- **GETSEQ** Sequence Code Match (SCM) for simple sequence queries to find exact sequences or with controlled variations

Agenda

- STN sequence searchable databases and search methods
- Sequence Code Match (SCM) search options
- SCM variability: Motif searching symbols
- Antibody/CDR search example
- Hypothetical multiple CDR search discussion

Sequence Code Match Searching (SCM)

- Search the sequence query as is or with some variation using special symbols
- Exact (the same sequence and length)
- Subsequence (query sequence embedded in a larger sequence)
- Types of SCM Search
  - Exact
  - Exact Family (peptide searches only)
  - Subsequence
  - Subsequence Family (peptide searches only)

Sequence Code Match Searching on CAS STNext

- In REGISTRY
  - **S *sequence/search qualifier***
- In GENESEQ, USGENE and PATGENE
  - **RUN GETSEQ *sequence/search qualifier***

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Sequence Code Match: Exact search

- Exact search
  - Matches the sequence query as entered
  - Identical sequences and exact length
- Search qualifiers identify the type of sequence to search, e.g.,

  /SQEN = SeQuence Exact Nucleotide

  /SQEP = SeQuence Exact Peptide

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Exact nucleic acid search (/SQEN) and DISPLAY in REGISTRY

```
=> S AGAAGCCAGA/SQEN

            3 AGAAGCCAGA/SQEN
       418978 SQL=10
L1          3 AGAAGCCAGA/SQEN
               (AGAAGCCAGA/SQEN AND SQL=10)

=> D SQIDE 3


L1   ANSWER 3 OF 3  REGISTRY  COPYRIGHT 2024 ACS on STN
RN   253294-09-6 REGISTRY
ED   Entered STN:  21 Jan 2000
CN   DNA, d(A-G-A-A-G-C-C-A-G-A), double-stranded complementary (9CI)  (CA
     INDEX NAME)
OTHER CA INDEX NAMES:
CN   DNA, d(T-C-T-G-G-C-T-T-C-T), double-stranded complementary (9CI)
OTHER NAMES:
CN   1328: PN: WO9965924 SEQID: 1327 claimed sequence
FS   NUCLEIC ACID SEQUENCE
SQL  10
NA   5 a   2 c   3 g

PATENT ANNOTATIONS (PNTE):
Sequence |Patent
Source   |Reference
=========+=============
Not Given|WO9965924
         |claimed SEQID
         |1327
```

```
SEQ      1 agaagccaga
           ==========
HITS AT:   1-10

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
MF   Unspecified
CI   MAN
SR   CA
LC   STN Files:   CA, CAPLUS, USPATFULL

DT.CA  CAplus document type:  Patent
RL.P   Roles from patents:  BIOL (Biological study); PRP (Properties); USES
       (Uses)
              1 REFERENCES IN FILE CA (1907 TO DATE)
              1 REFERENCES IN FILE CAPLUS (1907 TO DATE)
```

Search the CAS Registry number in CAplus to retrieve relevant patent and-or non-patent literature records.

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Exact nucleic acid search (/SQEN) and DISPLAY in GENESEQ

```
=> RUN GETSEQ AGAAGCCAGA/SQEN


RUN GETSEQ AT 11:53:14 ON 2024-07-13
COPYRIGHT (C) 2024 FIZ KARLSRUHE
Algorithm: GetSeq motif search. Version: 1.0.0

GetSeq motif search by FIZ Karlsruhe

GENESEQ
Query time:          238
L2    RUN STATEMENT CREATED
L2           2 AGAAGCCAGA/SQEN
```

```
=> D BIB SEQ 2


L2     ANSWER 2 OF 2 GENESEQ COPYRIGHT 2024 CLARIVATE on STN.
AN     AAF39527   GENESEQ  ED 20211030 UP 20211030
       DED 20010323 Full-text
TI     Yeast gene coding sequences comprising NORF genes with serial analysis
       of gene expression (SAGE) tags, useful for studying, monitoring and
       affecting phases of the cell cycle.
IN     Velculescu V; Vogelstein B; Kinzler K
PA     UNIV JOHNS HOPKINS (UYJO)
LA     English
DT     Patent
PI     WO 2000077214   A2        20001221
PIT    WOA2 INTERNATIONAL APPLICATION PUBLISHED WITHOUT INTERNATIONAL SEARCH
       REPORT or INTERNATIONAL APPLICATION PUBLISHED WITH DECLARATION UNDER
       ARTICLE 17 (2) (A) [FROM 20090101 ONWARDS]
AI     WO 2000-US16223          20000614
PRAI   US 1999-335032           19990616
FS     NUCLEIC; NS
OS     2001-061874 [07]
MTY    DNA
PSL    Example; Page 223; 419pp
DESC   Yeast NORF gene SAGE tag oligonucleotide SEQ ID NO:6266.

SEQ
               1 agaagccaga
```

Searches run in USGENE and PATGENE are run the same way as GENESEQ.

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Exact peptide search (/SQEP) and DISPLAY in REGISTRY

```
=> S DSDGP/SQEP

             1 DSDGP/SQEP
        156228 SQL=5
L1           1 DSDGP/SQEP
                 (DSDGP/SQEP AND SQL=5)

=> D SEQ SEQ3


L1   ANSWER 1 OF 1  REGISTRY  COPYRIGHT 2024 ACS on STN

SEQ      1 DSDGP
           =====
HITS AT:   1-5

SEQ3     1 Asp-Ser-Asp-Gly-Pro
           === === === === ===
HITS AT:   1-5
```

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

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Exact peptide search (/SQEP) and DISPLAY in GENESEQ

```
=> RUN GETSEQ DSDGP/SQEP


RUN GETSEQ AT 13:36:39 ON 2024-07-13
COPYRIGHT (C) 2024 FIZ KARLSRUHE
Algorithm: GetSeq motif search. Version: 1.0.0

GetSeq motif search by FIZ Karlsruhe

GENESEQ
Query time:           55
L2    RUN STATEMENT CREATED
L2           1 DSDGP/SQEP


=> D SEQ SEQ3


L2     ANSWER 1 OF 1 GENESEQ COPYRIGHT 2024 CLARIVATE on STN.

SEQ
               1 dsdgp

SEQ3
               1 Asp-Ser-Asp-Gly-Pro
```

Searches run in USGENE and PATGENE are run the same way as GENESEQ.

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Sequence Code Match

SCM

- Exact
- Exact Family
- Subsequence
- Subsequence Family

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

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, not nucleic acids

/SQEFP = SeQuence Exact Family Peptide

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Amino acid family substitutions

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 peptide search (/SQEFP) in REGISTRY

```
=> S DSDGP/SQEFP

          166 DSDGP/SQEFP
       156228 SQL=5
L1        166 DSDGP/SQEFP
                (DSDGP/SQEFP AND SQL=5)

=> D SEQ


L1   ANSWER 1 OF 166  REGISTRY  COPYRIGHT 2024 ACS on STN

SEQ      1 DTDAS
           =====
HITS AT:   1-5
```

Possible family substitutions for DSDGP:

D	S	D	G	P
Q	P	Q	P	A
N	A	N	A	G
E	G	E	S	S
B	T	B	T	T

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Exact Family peptide search (/SQEFP) in GENESEQ

```
=> RUN GETSEQ DSDGP/SQEFP


RUN GETSEQ AT 13:55:34 ON 2024-07-13
COPYRIGHT (C) 2024 FIZ KARLSRUHE
Algorithm: GetSeq motif search. Version: 1.0.0

GetSeq motif search by FIZ Karlsruhe

GENESEQ
Query time:           65
L2    RUN STATEMENT CREATED
L2         355 DSDGP/SQEFP


=> D SEQ


L2     ANSWER 1 OF 355 GENESEQ COPYRIGHT 2024 CLARIVATE on STN.

SEQ
               1 qgqsg
```

Possible family substitutions for DSDGP:

D	S	D	G	P
Q	P	Q	P	A
N	A	N	A	G
E	G	E	S	S
B	T	B	T	T

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Sequence Code Match

SCM

- Exact
- Exact Family
- Subsequence
- Subsequence Family

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Sequence Code Match: Subsequence search

Subsequence search

- Retrieves exact answers plus sequences that are embedded in longer sequence

/SQSN = SeQuence Subsequence Nucleotides

/SQSP = SeQuence Subsequence Peptides

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Subsequence nucleotide search (/SQSN) in REGISTRY

```
=> S ACCCTGCAAA TAGCA/SQSN

  IS NOT A VALID NUCLEIC ACID SYMBOL

=> S ACCCTGCAAATAGCA/SQSN

L4          176 ACCCTGCAAATAGCA/SQSN

=> D SEQ


L4   ANSWER 1 OF 176  REGISTRY  COPYRIGHT 2024 ACS on STN

SEQ      1 atggctgcag cttcatatga tcagttgtta aagcaagttg aggcactgaa
        51 gatggagaac tcaaatcttc gacaagagct agaagataat tccaatcatc
       101 ttacaaaact ggaaactgag gcatctaata tgaaggaagt acttaaacaa
       151 ctacaaggaa gtattgaaga tgaagctatg gcttcttctg gacagattga
       201 tttattagag cgtcttaaag agcttaactt agatagcagt aatttccctg
       251 gagtaaaact gcggtcaaaa atgtccctcc gttcttatgg aagccgggaa
       301 ggatctgtat caagccgttc tggagagtgc agtcctgttc ctatgggttc
       351 atttccaaga agagggtttg taaatggaag cagagaaagt actggatatt
       401 tagaagaact tgagaaagag aggtcattgc ttcttgctga tcttgacaaa
```

In REGISTRY, the spaces in a sequence must be removed.

. . .

```
3551 ccacagatat tccttcatca cagaaacagt cattttcatt ctcaaagagt
3601 tcatctggac aaagcagtaa aaccgaacat atgtcttcaa gcagtgagaa
3651 tacgtccaca ccttcatcta atgccaagag gcagaatcag ctccatccaa
3701 gttctgcaca gagtagaagt ggtcagcctc aaaaggctgc cacttgcaaa
3751 gtttcttcta ttaaccaaga aacaatacag acttattgtg tagaagatac
3801 tccaatatgt ttttcaagat gtagttcatt atcatctttg tcatcagctg
3851 aagatgaaat aggatgtaat cagacgacac aggaagcaga ttctgctaat
3901 accctgcaaa tagcagaaat aaaagaaaag attggaacta ggtcagctga
     ========== =====
3951 agatcctgtg agcgaagttc cagcagtgtc acagcaccct agaaccaaat
4001 ccagcagact gcagggttct agtttatctt cagaatcagc caggcacaaa
4051 gctgttgaat tttcttcagg agcgaaatct ccctccaaaa gtggtgctca
4101 gacacccaaa agtccacctg aacactatgt tcaggagacc ccactcatgt
4151 ttagcagatg tacttctgtc agttcacttg atagttttga gagtcgttcg
4201 attgccagct ccgttcagag tgaaccatgc agtggaatgg taagtggcat
```

. . .

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Subsequence nucleotide search (/SQSN) in GENESEQ

```
=> RUN GETSEQ ACCCTGCAAA TAGCA/SQSN


RUN GETSEQ AT 15:29:17 ON 2024-07-13
COPYRIGHT (C) 2024 FIZ KARLSRUHE
Algorithm: GetSeq motif search. Version: 1.0.0

GetSeq motif search by FIZ Karlsruhe

GENESEQ
Query time:         6926
L3    RUN STATEMENT CREATED
L3         240 ACCCTGCAAA TAGCA/SQSN


=> D ALIGN


L3     ANSWER 1 OF 240 GENESEQ COPYRIGHT 2024 CLARIVATE on STN.
ALIGN
  Sequence Length: 6631;
  Strand: Plus / Plus;
  Hits at: 1980-1994
       1 GTGACCTTAA TTTTGTGATC TCTTGATTTT ATTTCAGGCA AATCCTAAGA GAGAACAACT
      61 GTCTACAAAC TTTATTACAA CACTTAAAAT CTCATAGTTT GACAATAGTC AGTAATGCAT
     121 GTGGAACTTT GTGGAATCTC TCAGCAAGAA ATCCTAAAGA CCAGGAAGCA TTATGGGACA
     181 TGGGGGCAGT TAGCATGCTC AAGAACCTCA TTCATTCAAA GCACAAAATG ATTGCTATGG
     241 GAAGTGCTGC AGCTTTAAGG AATCTCATGG CAAATAGGCC TGCGAAGTAC AAGGATGCCA
```

In GENESEQ, USGENE and PATGENE, spaces do not need to be removed from the sequence.

. . .

```
1501 CCAATTATAG TGAACGTTAC TCTGAAGAAG AACAGCATGA AGAAGAAGAG AGACCAACAA
1561 ATTATAGCAT AAAATATAAT GAAGAGAAAC GTCATGTGGA TCAGCCTATT GATTATAGTT
1621 TAAAATATGC CACAGATATT CCTTCATCAC AGAAACAGTC ATTTTCATTC TCAAAGAGTT
1681 CATCTGGACA AAGCAGTAAA ACCGAACATA TGTCTTCAAG CAGTGAGAAT ACGTCCACAC
1741 CTTCATCTAA TGCCAAGAGG CAGAATCAGC TCCATCCAAG TTCTGCACAG AGTAGAAGTG
1801 GTCAGCCTCA AAAGGCTGCC ACTTGCAAAG TTTCTTCTAT TAACCAAGAA ACAATACAGA
1861 CTTATTGTGT AGAAGATACT CCAATATGTT TTTCAAGATG TAGTTCATTA TCATCTTTGT
1921 CATCAGCTGA AGATGAAATA GGATGTAATC AGACGACACA GGAAGCAGAT TCTGCTAATA
                                                                     =
1981 CCCTGCAAAT AGCAGAAATA AAAGAAAAGA TTGGAACTAG GTCAGCTGAA GATCCTGTGA
     ========== ====
2041 GCGAAGTTCC AGCAGTGTCA CAGCACCCTA GAACCAAATC CAGCAGACTG CAGGGTTCTA
2101 GTTTATCTTC AGAATCAGCC AGGCACAAAG CTGTTGAATT TTCTTCAGGA GCGAAATCTC
2161 CCTCCAAAAG TGGTGCTCAG ACACCCAAAA GTCCACCTGA ACACTATGTT CAGGAGACCC
2221 CACTCATGTT TAGCAGATGT ACTTCTGTCA GTTCACTTGA TAGTTTTGAG AGTCGTTCGA
2281 TTGCCAGCTC CGTTCAGAGT GAACCATGCA GTGGAATGGT AAGTGGCATT ATAAGCCCCA
2341 GTGATCTTCC AGATAGCCCT GGACAAACCA TGCCACCAAG CAGAAGTAAA ACACCTCCAC
```

. . .

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Subsequence peptide search (/SQSP) in REGISTRY

```
=> S DSDGP/SQSP

L1         1274 DSDGP/SQSP

=> D SEQ


L1   ANSWER 1 OF 1274  REGISTRY  COPYRIGHT 2024 ACS on STN

SEQ       1 MSLGKSDGQC AASTSRCEGG AQGPARTPPL QRGTPPPAQV TPGLSHSTRD
         51 NPEVICLDTP PAMPATSLPA TLTTTTVTTV ACGGPIMSTL GLPTGGHITS
        101 PRPSPASDQQ TPPGAVWVTP MERTPIPSGE VERRQQGAVT PSPPPPQETS
        151 RDEGDSDGPW RVRLGRRARR RLSSGSSGGK SPPTTSADNN NNSLNLQNNL
                 =====
        201 INNNVEVSMS NNILQQANAF LVLGTSVASG PSPLDDGGAP GDMRAEDHSA
        251 GHPSVPVSVC GVMGRPRGAR ASVDSGANEQ TPSEPQRGGG GHHRRPRRVE
        301 VAPVQPFRGQ PAPAVAARQR RRQRVVARDA LVGRAEAVAS LADLEEFAAS
        351 VALFFGEEAV GVARGDAPGA RDRPVRLRAA RGRRGEGRVP ARDGANVERG
        401 PVPAGPQQRQ LGEGRGDWTR EAKRIQALYR VNRRRAIREV LQGPAEFCRV
        451 PTRRVQAYFE DLYRGGERLD NAGAEAERVD PPRREDVELL MAPFTEREVD
        501 CRIRRMNNSA PGPDGLTYRD LRAADPGARL LAAFFNACYR LEAVPASWKT
        551 SNTVLVHKKD DPGLLENWRP LALGDTTPKL FAALVADRLT GWAINNNKLS
```

. . .

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Subsequence peptide search (/SQSP) in GENESEQ

```
=> RUN GETSEQ DSDGP/SQSP


RUN GETSEQ AT 15:25:52 ON 2024-07-13
COPYRIGHT (C) 2024 FIZ KARLSRUHE
Algorithm: GetSeq motif search. Version: 1.0.0

GetSeq motif search by FIZ Karlsruhe

GENESEQ
Query time:          231
L2    RUN STATEMENT CREATED
L2        1505 DSDGP/SQSP


=> D ALIGN


L2     ANSWER 1 OF 1505 GENESEQ COPYRIGHT 2024 CLARIVATE on STN.
ALIGN
  Sequence Length: 225;

  Hits at: 177-181
       1 THTCPPCPAP ELLGGPSVFL FPPKPKDTLM ISRTPEVTCV VVDVSHEDPE VKFNWYVDGV
      61 EVHNAKTKPR EEQYNSTYRV VSVLTVLHQD WLNGKEYKCK VSNKALPVPI EKTISKAKGQ
     121 PREPQVYTLP PSREEMTKNQ VSLTCLVKGF YPSDIAVEWE SNGQPENNYK TTPPVLDSDG
                                                                      ====
     181 PFFLYSKLTV DKSRWQQGNV FSCSVMHEAL HNHYTQKSLS LSPGK
           =
```

Use the ALIGN display option to show where the match is.

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Sequence Code Match

SCM

- Exact
- Exact Family
- Subsequence
- Subsequence Family

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Sequence Code Match: Subsequence Family search

Subsequence Family search

– Retrieves exact sequence match, subsequence match, and sequences that contain family substitutions of amino acids

/SQSFP = SeQuence Subsequence Family Peptides

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Subsequence Family peptide search in REGISTRY

```
=> S DSDGP/SQSFP

L5      2191142 DSDGP/SQSFP

=> D SEQ


L5   ANSWER 1 OF 2191142  REGISTRY  COPYRIGHT 2024 ACS on STN

SEQ      1 MSQTQPPAPV GPGDPDVYLK GVPSAGMHPR GVHAPRGHPR MISGPPQRGD
             =====
        51 NDQAAGQCGD SGLLRVGADT TISKPSEAVR PPTIPRTPRV PREPRVPRPP
       101 REPREPRVPR DPRDPRVPRD PRDPRQPRSP REPRSPREPR TPRTPREPRT
       151 ARGSV
HITS AT:   3-7
```

Possible family substitutions for DSDGP:

D	**S**	**D**	**G**	**P**
Q	P	Q	P	A
N	A	N	A	G
E	G	E	S	S
B	T	B	T	T

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Subsequence Family peptide search in GENESEQ

```
=> RUN GETSEQ DSDGP/SQSFP


RUN GETSEQ AT 16:16:54 ON 2024-07-13
COPYRIGHT (C) 2024 FIZ KARLSRUHE
Algorithm: GetSeq motif search. Version: 1.0.0

GetSeq motif search by FIZ Karlsruhe

GENESEQ
Query time:           30
L7    RUN STATEMENT CREATED
L7      250000 DSDGP/SQSFP

The maximum number of hits has been reached.
Please specify your search sequence.
```

```
=> D ALIGN


L7     ANSWE
ALIGN
  Sequence L

  Hits at: 3
       1 QVQ
      61 NEK
     121 FPLAPSSKST SGGTAALGCL VKDYFPEPVT VSWNSGALTS GVHTFPAVLQ SSGLYSLSSV
     181 VTVPSSSLGT QTYICNVNHK PSNTKVDKKV EPKSCDKTHT CPPCPAPELL GGPSVFLFPP
     241 KPKDTLMISR TPEVTCVVVD VSHEDPEVKF NWYVDGVEVH NAKTKPREEQ YNSTYRVVSV
     301 LTVLHQDWLN GKEYKCKVSN KALPAPIEKT ISKAKGQPRE PQVYTLPPSR DELTKNQVSL
     361 TCLVKGFYPS DIAVEWESNG QPENNYKTTP PVLDSDGSFF LYSKLTVDKS RWQQGNVFSC
                                                  =====
     421 SVMHEALHNH YTQKSLSLSP GK
```

Possible family substitutions for DSDGP:

D	S	D	G	P
Q	P	Q	P	A
N	A	N	A	G
E	G	E	S	S
B	T	B	T	T

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Summary of Sequence Code Match options

Search Type	Proteins	Nucleic Acids
Exact	/SQEP	/SQEN
Exact Family	/SQEFP	Not Applicable
Subsequence	/SQSP	/SQSN
Subsequence Family	/SQSFP	Not Applicable

Agenda

- STN sequence searchable databases and search methods
- Sequence Code Match (SCM) search options
- SCM variability: Motif searching symbols
- Antibody/CDR search example
- Hypothetical multiple CDR search discussion

Special motif symbols allow flexibility in sequence searching

- Specify motif patterns that consist of different amino acid(s) at one or more locations in the sequence
- Ability to specify specific amino acid sequences between unknown number of amino acids (gaps)
- Ability to search for sequence patterns at either beginning or the end of the sequence
- Specify the number or range of repeats for amino acid(s) or gaps

Motif searching symbols and characters

Symbols	Functions	Examples	Possible answers
^	Search at the beginning or the end of a sequence	^MCGIL/SQSP VCDS^/SQSFP	"MCGIL……“ "……VCDS"
[ ]	Specify alternate residues	LGP[VL]/SQSP	LGPV LGPL
[-] or [~]	Exclude one or more residues	PTGK[-H]/SQSP PTGK[~H]/SQSP	PTGKACCD
{#,#} {# - #} {#}	Repeat preceding residue(s)	GG(FL){1,3}/SQSP GG(FL){1-3}/SQSP GG(FL){3}/SQSP	GGFL GGFLFL GGFLFLFL
.	Specify gap(s) in the sequence	SY.RPG/SQSP SY…RPG/SQSP	SYARPG SYAAARPG
|	Specify alternate residues	ACD|KLM/SQSP	ACD KLM

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Motif searching symbols and characters

```
=> S ^MCGIL/SQSP

L1          263 ^MCGIL/SQSP

=> D SEQ


L1    ANSWER 1 OF 263  REGISTRY  COPYRIGHT 2024 ACS on STN

SEQ       1 MCGILAVLGV AEVSLAKRSR IIELSRRLRH RGPDWSGLHC HEDCYLAHQR
            =====
         51 LAIIDPTSGD QPLYNEDKTV VVTVNGEIYN HEELKAKLKT HEFQTGSDCE
        101 VIAHLYEEYG EEFVDMLDGM FSFVLLDTRD KSFIAARDAI GICPLYMGWG
        151 LDGSVWFSSE MKALSDDCER FITFPPGHLY SSKTGGLRRW YNPPWFSETV
        201 PSTPYNALFL REMFEKAVIK RLMTDVPFGV LLSGGLDSSL VASVASRHLN
        251 ETKVDRQWGN KLHTFCIGLK GSPDLKAARE VADYLSTVHH EFHFTVQEGI
        301 DALEEVIYHI ETYDVTTIRA STPMFLMSRK IKSLGVKMVI SGEGSDEIFG
        351 GYLYFHKAPN KKEFHEETCR KIKALHLYDC LRANKATSAW GVEARVPFLD
        401 KSFISVAMDI DPEWKMIKRD LGRIEKWVIR NAFDDDERPY LPKHILYRQK
        451 EQFSDGVGYS WIDGLKDHAS QHVSDSMMMN AGFVYPENTP TTKEGYYYRM
        501 IFEKFFPKPA ARSTVPGGPS VACSTAKAVE WDASWSKNLD PSGRAALGVH
        551 DAAYEDTAGK TPASADPVSD KGLRPAIGES LGTPVASATA V
HITS AT:    1-5

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
```

```
=> RUN GETSEQ ^MCGIL/SQSP


RUN GETSEQ AT 14:04:31 ON 2024-07-15
COPYRIGHT (C) 2024 FIZ KARLSRUHE
Algorithm: GetSeq motif search. Version: 1.0.0

GetSeq motif search by FIZ Karlsruhe

GENESEQ
Query time:          195
L2    RUN STATEMENT CREATED
L2        1057 ^MCGIL/SQSP


=> D ALIGN


L2     ANSWER 1 OF 1057  GENESEQ  COPYRIGHT 2024 CLARIVATE on STN.
ALIGN
  Sequence Length: 65;

  Hits at: 1-5
      1 MCGILAVLGC ADEAKGSSKR SRVLELSRRL KHRGPDWSGL RQVGDCYLSH QRLAIIDPAS
        =====
     61 GDQPL
```


FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Motif searching symbols and characters (cont.)

Symbols	Functions	Examples	Possible answers
^	Search at the beginning or the end of a sequence	^MCGIL/SQSP VCDS^/SQSFP	"MCGIL……" "……VCDS"
[ ]	Specify alternate residues	LGP[VL]/SQSP	LGPV LGPL
[-] or [~]	Exclude one or more residues	PTGK[-H]/SQSP PTGK[~H]/SQSP	PTGKACCD
{#,#} {# - #} {#}	Repeat preceding residue(s)	GG(FL){1,3}/SQSP GG(FL){1-3}/SQSP GG(FL){3}/SQSP	GGFL GGFLFL GGFLFLFL
.	Specify gap(s) in the sequence	SY.RPG/SQSP SY…RPG/SQSP	SYARPG SYAAARPG
|	Specify alternate residues	ACD|KLM/SQSP	ACD KLM

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Motif searching symbols and characters (cont.)

```
=> S LGP[VL]/SQSP

L3      115695 LGP[VL]/SQSP

=> D SEQ


L3   ANSWER 1 OF 115695 REGISTRY  COPYRIGHT 2024 ACS on STN

SEQ      1 MRHIICHGGV ITEEMAASLL DQLIEEVLAD NLPPPSHFEP PTLHELYDLD
        51 VTAPEDPNEE AVSQIFPESV MLAVQEGIDL FTFPPAPGSP EPPHLSRQPE
       101 QPEQRALGPV SMPNLVPEVI DLTCHEAGFP PSDDEDEEGE EFVLDYVEHP
                ====
       151 GHGCRSCHYH RRNTGDPDIM CSLCYMRTCG MFVYSPVSEP EPEPEPEPEP
       201 ARPTRRPKLV PAILRRPTSP VSRECNSSTD SCDSGPSNTP PEIHPVVPLC
       251 PIKPVAVRVG GRRQAVECIE DLLNESGQPL DLSCKRPRP
HITS AT:    107-110

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
```

```
=> RUN GETSEQ LGP[VL]/SQSP


RUN GETSEQ AT 14:19:38 ON 2024-07-15
COPYRIGHT (C) 2024 FIZ KARLSRUHE
Algorithm: GetSeq motif search. Version: 1.0.0

GetSeq motif search by FIZ Karlsruhe

GENESEQ
Query time:          242
L4    RUN STATEMENT CREATED
L4      147002 LGP[VL]/SQSP


=> D ALIGN


L4     ANSWER 1 OF 147002 GENESEQ COPYRIGHT 2024 CLARIVATE on STN.
ALIGN
  Sequence Length: 195;

  Hits at: 123-126
       1 MAFVLSLLMA LVLVSYGPGR SLGCYLSEDH MLGARENLRL LARMNRLSPH PCLQDRKDFG
      61 LPQEMVEGSQ LQKDQAISVL HEMLQQCFNL FHIEHSSAAW NTTLLEQLCT GLQQQLEDLD
     121 ACLGPVMGEK DSDMGRMGPI LTVKKYFQDI HVYLKEKEYS DCAWEIIRVE MMRALSSSTT
           ====
     181 LQKRLRKMGG DLNSL
```


FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Motif searching symbols and characters (cont.)

Symbols	Functions	Examples	Possible answers
^	Search at the beginning or the end of a sequence	^MCGIL/SQSP VCDS^/SQSFP	"MCGIL……“ "…..……VCDS"
[ ]	Specify alternate residues	LGP[VL]/SQSP	LGPV LGPL
[-] or [~]	Exclude one or more residues	PTGK[-H]/SQSP PTGK[~H]/SQSP	PTGKACCD
{#,#} {# - #} {#}	Repeat preceding residue(s)	GG(FL){1,3}/SQSP GG(FL){1-3}/SQSP GG(FL){3}/SQSP	GGFL GGFLFL GGFLFLFL
.	**Specify gap(s) in the sequence**	**SY.RPG/SQSP** **SY…RPG/SQSP**	**SYARPG** **SYAAARPG**
|	Specify alternate residues	ACD|KLM/SQSP	ACD KLM

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Motif searching symbols and characters (cont.)

```
=> S SY...RPG/SQSP

L6          920 SY...RPG/SQSP

=> D SEQ


L6   ANSWER 1 OF 920  REGISTRY  COPYRIGHT 2024 ACS on STN

SEQ       1 MESPSAPPHR WCIPWQRLLL TASLLTFWNP PTTAKLTIES TPFNVAEGKE
         51 VLLLVHNLPQ HLFGYSWYKG ERVDGNRQII GYVIGTQQAT PGPAYSGREI
        101 IYPNASLLIQ NIIQNDTGFY TLHVIKSDLV NEEATGQFRV YPELPKPSIS
        151 SNNSKPVEDK DAVAFTCEPE TQDATYLWWV NNQSLPVSPR LQLSNGNRTL
        201 TLFNVTRNDT ASYKCETQNP VSARRSDSVI LNVLYGPDAP TISPLNTSYR
        251 SGENLNLSCH AASNPPAQYS WFVNGTFQQS TQELFIPNIT VNNSGSYTCQ
        301 AHNSDTGLNR TTVTTITVYA EPPKPFITSN NSNPVEDEDA VALTCEPEIQ
        351 NTTYLWWVNN QSLPVSPRLQ LSNDNRTLTL LSVTRNDVGP YECGIQNKLS
        401 VDHSDPVILN VLYGPDDPTI SPSYTYYRPG VNLSLSCHAA SNPPAQYSWL
                                 ========
        451 IDGNIQQHTQ ELFISNITEK NSGLYTCQAN NSASGHSRTT VKTITVSAEL
        501 PKPSISSNNS KPVEDKDAVA FTCEPEAQNT TYLWWVNGQS LPVSPRLQLS
        551 NGNRTLTLFN VTRNDARAYV CGIQNSVSAN RSDPVTLDVL YGPDTPIISP
        601 PDSSYLSGAN LNLSCHSASN PSPQYSWRIN GIPQQHTQVL FIAKITPNNN
        651 GTYACFVSNL ATGRNNSIVK SITVSASGTS PGLSAGATVG IMIGVLVGVA
        701 LI
HITS AT:   423-430

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
```

```
=> RUN GETSEQ SY...RPG/SQSP


RUN GETSEQ AT 14:23:27 ON 2024-07-15
COPYRIGHT (C) 2024 FIZ KARLSRUHE
Algorithm: GetSeq motif search. Version: 1.0.0

GetSeq motif search by FIZ Karlsruhe

GENESEQ
Query time:          241
L5    RUN STATEMENT CREATED
L5        1442 SY...RPG/SQSP


=> D ALIGN


L5     ANSWER 1 OF 1442 GENESEQ COPYRIGHT 2024 CLARIVATE on STN.
ALIGN
  Sequence Length: 702;

  Hits at: 423-430
       1 MESPSAPPHR WCIPWQRLLL TASLLTFWNP PTTAKLTIES TPFNVAEGKE VLLLVHNLPQ
      61 HLFGYSWYKG ERVDGNRQII GYVIGTQQAT PGPAYSGREI IYPNASLLIQ NIIQNDTGFY
     121 TLHVIKSDLV NEEATGQFRV YPELPKPSIS SNNSKPVEDK DAVAFTCEPE TQDATYLWWV
     181 NNQSLPVSPR LQLSNGNRTL TLFNVTRNDT ASYKCETQNP VSARRSDSVI LNVLYGPDAP
     241 TISPLNTSYR SGENLNLSCH AASNPPAQYS WFVNGTFQQS TQELFIPNIT VNNSGSYTCQ
     301 AHNSDTGLNR TTVTTITVYA EPPKPFITSN NSNPVEDEDA VALTCEPEIQ NTTYLWWVNN
     361 QSLPVSPRLQ LSNDNRTLTL LSVTRNDVGP YECGIQNELS VDHSDPVILN VLYGPDDPTI
     421 SPSYTYYRPG VNLSLSCHAA SNPPAQYSWL IDGNIQQHTQ ELFISNITEK NSGLYTCQAN
           ========
     481 NSASGHSRTT VKTITVSAEL PKPSISSNNS KPVEDKDAVA FTCEPEAQNT TYLWWVNGQS
     541 LPVSPRLQLS NGNRTLTLFN VTRNDARAYV CGIQNSVSAN RSDPVTLDVL YGPDTPIISP
     601 PDSSYLSGAN LNLSCHSASN PSPQYSWRIN GIPQQHTQVL FIAKITPNNN GTYACFVSNL
     661 ATGRNNSIVK SITVSASGTS PGLSAGATVG IMIGVLVGVA LI
```


FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Motif searching symbols and characters (cont.)

Symbols	Functions	Examples	Possible answers
?	Include residue(s) zero or one time	FLRR(RP)?K/SQSP	FLRRK FLRRRPK
*	Include residue(s) zero or more times	KLK(WD)*N/SQSP	KLKN KLKWDN KLKWDWDN KLKWDWDWDN
+	Repeat residue(s) one or more times	DAQP+/SQSP	DAQP DAQPP DAQPPP DAQPPPP
		D(AQP)+/SQSP	DAQP DAQPAQP DAQPAQPAQP DAQPAQPAQPAQP
&	Join multiple sequence fragments together as one	ACDKLM & KLKWDN/SQSP	ACDKLMKLKWDN

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Motif searching symbols and characters (cont.)

```
=> S KSSQSLWDSGNQKNFLT(.)+WTSYRES/SQSP

L1          19 KSSQSLWDSGNQKNFLT(.)+WTSYRES/SQSP

=> d seq


L1   ANSWER 1 OF 19  REGISTRY  COPYRIGHT 2024 ACS on STN

SEQ      1 DIVMTQSPDS LAVSLGERAT INCKSSQSLW DSGNQKNFLT WYQQKPGQPP
                                  ======= ========== ==========
        51 KLLIYWTSYR ESGVPDRFSG SGSGTDFTLT ISSLQAEDVA VYYCQNDYFY
           ==========  ==
       101 PHTFGGGTKV EIXTVAAPSV FIFPPSDEQL KSGTASVVCL LNNFYPREAK
       151 VQWKVDNALQ SGNSQESVTE QDSKDSTYSL SSTLTLSKAD YEKHKVYACE
       201 VTHQGLSSPV TKSFNRGEC
HITS AT:   24-62
```

```
=> RUN GETSEQ KSSQSLWDSGNQKNFLT(.)+WTSYRES/SQSP


RUN GETSEQ AT 11:15:13 ON 2024-07-18
COPYRIGHT (C) 2024 FIZ KARLSRUHE
Algorithm: GetSeq motif search. Version: 1.0.0

GetSeq motif search by FIZ Karlsruhe

GENESEQ
Query time:          213
L2    RUN STATEMENT CREATED
L2        120 KSSQSLWDSGNQKNFLT(.)+WTSYRES/SQSP


=> D ALIGN


L2     ANSWER 1 OF 120  GENESEQ  COPYRIGHT 2024 CLARIVATE on STN.
ALIGN
  Sequence Length: 219;

  Hits at: 24-62
       1 DIVMTQSPDS LAVSLGERAT INCKSSQSLW DSGNQKNFLT WYQQKPGQPP KLLIYWTSYR
                                 ======= ========== ========== ==========
      61 ESGVPDRFSG SGSGTDFTLT ISSLQAEDVA VYYCQNDYFY PHTFGGGTKV EIXTVAAPSV
         ==
     121 FIFPPSDEQL KSGTASVVCL LNNFYPREAK VQWKVDNALQ SGNSQESVTE QDSKDSTYSL
     181 SSTLTLSKAD YEKHKVYACE VTHQGLSSPV TKSFNRGEC
```


FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Motif searching symbols and characters (cont.)

Symbols	Functions	Examples	Possible answers
?	Include residue(s) zero or one time	FLRR(RP)?K/SQSP	FLRRK FLRRRPK
*	Include residue(s) zero or more times	KLK(WD)*N/SQSP	KLKN KLKWDN KLKWDWDN KLKWDWDWDN
+	Repeat residue(s) zero or more times	DAQP+/SQSP	DAQP DAQPP DAQPPP DAQPPPP
		(AQP)+/SQSP	AQPAQP AQPAQPAQP AQPAQPAQPAQP AQPAQPAQPAQPAQP
&	**Join multiple sequence fragments together as one**	**L1&L2/SQSP**	

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Motif searching symbols and characters (cont.)

```
=> run getseq mgmdcdiegk dgkqyesvlm vsidqlldsm keigsnclnn efnffkrhic    dankegmflf raarklrqfl kmnstgdfdl hllkvsegtt illnctgqvk    grkpaalgea qptksleenk slkeqkklnd lcflkrllqe iktcwnkilm    gtkehrntgr ggeekkkeke keeqeeretk tpecpshtqp
lgvflfppkp/sqsp


RUN GETSEQ AT 17:54:29 ON 2024-07-15
COPYRIGHT (C) 2024 FIZ KARLSRUHE
Algorithm: GetSeq motif search. Version: 1.0.0

GetSeq motif search by FIZ Karlsruhe

GENESEQ
Query time:          197
L1    RUN STATEMENT CREATED
L1          18 MGMDCDIEGK DGKQYESVLM VSIDQLLDSM KEIGSNCLNN EFNFFK
               RHIC    DANKEGMFLF RAARKLRQFL KMNSTGDFDL HLLKVSEGT
               T ILLNCTGQVK    GRKPAALGEA QPTKSLEENK SLKEQKKLND L
               CFLKRLLQE IKTCWNKILM    GTKEHRNTGR GGEEKKKEKE KEEQ
               EERETK TPECPSHTQP LGVFLFPPKP/SQSP


=> run getseq kdtlmisrtp evtcvvvdvs qedpevqfnw yvdgvevhna ktkpreeqfn  styrvvsvlt vlhqdwlngk eykckvsnkg lpssiektis kakgqprepq  vytlppsqee mtknqvsltc lvkgfypsdi avewesngqp ennykttppv  ldsdgsffly srltvdksrw qegnvfscsv mhealhnhyt
qkslslslgk/sqsp


RUN GETSEQ AT 17:54:46 ON 2024-07-15
COPYRIGHT (C) 2024 FIZ KARLSRUHE
Algorithm: GetSeq motif search. Version: 1.0.0

GetSeq motif search by FIZ Karlsruhe

GENESEQ
Query time:          208
L2    RUN STATEMENT CREATED
L2       19941 KDTLMISRTP EVTCVVVDVS QEDPEVQFNW YVDGVEVHNA KTKPRE
               EQFN  STYRVVSVLT VLHQDWLNGK EYKCKVSNKG LPSSIEKTIS
               KAKGQPREPQ  VYTLPPSQEE MTKNQVSLTC LVKGFYPSDI AVEWE
               SNGQP ENNYKTTPPV  LDSDGSFFLY SRLTVDKSRW QEGNVFSCSV
                MHEALHNHYT QKSLSLSLGK/SQSP
```

The character limit in a single command line is 292, which includes all spaces, search qualifiers, etc. Longer sequences may need to be broken up into multiple searches.

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Motif searching symbols and characters (cont.)

```
=> run getseq l1&l2/sqsp


RUN GETSEQ AT 17:55:32 ON 2024-07-15
COPYRIGHT (C) 2024 FIZ KARLSRUHE
Algorithm: GetSeq motif search. Version: 1.0.0

GetSeq motif search by FIZ Karlsruhe

GENESEQ
Query time:          346
L3    RUN STATEMENT CREATED
L3          18 (MGMDCDIEGK DGKQYESVLM VSIDQLLDSM KEIGSNCLNN EFNFF
               KRHIC    DANKEGMFLF RAARKLRQFL KMNSTGDFDL HLLKVSEG
               TT ILLNCTGQVK    GRKPAALGEA QPTKSLEENK SLKEQKKLND
               LCFLKRLLQE IKTCWNKILM    GTKEHRNTGR GGEEKKKEKE KEE
               QEERETK TPECPSHTQP LGVFLFPPKP)&KDTLMISRTP EVTCVVVD
               VS QEDPEVQFNW YVDGVEVHNA KTKPREEQFN  STYRVVSVLT VL
               HQDWLNGK EYKCKVSNKG LPSSIEKTIS KAKGQPREPQ  VYTLPPS
               QEE MTKNQVSLTC LVKGFYPSDI AVEWESNGQP ENNYKTTPPV  L
               DSDGSFFLY SRLTVDKSRW QEGNVFSCSV MHEALHNHYT QKSLSLS
               LGK/SQSP


=> d seq


L3     ANSWER 1 OF 18 GENESEQ COPYRIGHT 2024 CLARIVATE on STN.

SEQ
                1 mgmdcdiegk dgkqyesvlm vsidqlldsm keigsnclnn efnffkrhic
               51 dankegmflf raarklrqfl kmnstgdfdl hllkvsegtt illnctgqvk
              101 grkpaalgea qptksleenk slkeqkklnd lcflkrllqe iktcwnkilm
              151 gtkehrntgr ggeekkkeke keeqeeretk tpecpshtqp lgvflfppkp
              201 kdtlmisrtp evtcvvvdvs qedpevqfnw yvdgvevhna ktkpreeqfn
              251 styrvvsvlt vlhqdwlngk eykckvsnkg lpssiektis kakgqprepq
              301 vytlppsqee mtknqvsltc lvkgfypsdi avewesngqp ennykttppv
              351 ldsdgsffly srltvdksrw qegnvfscsv mhealhnhyt qkslslslgk
```

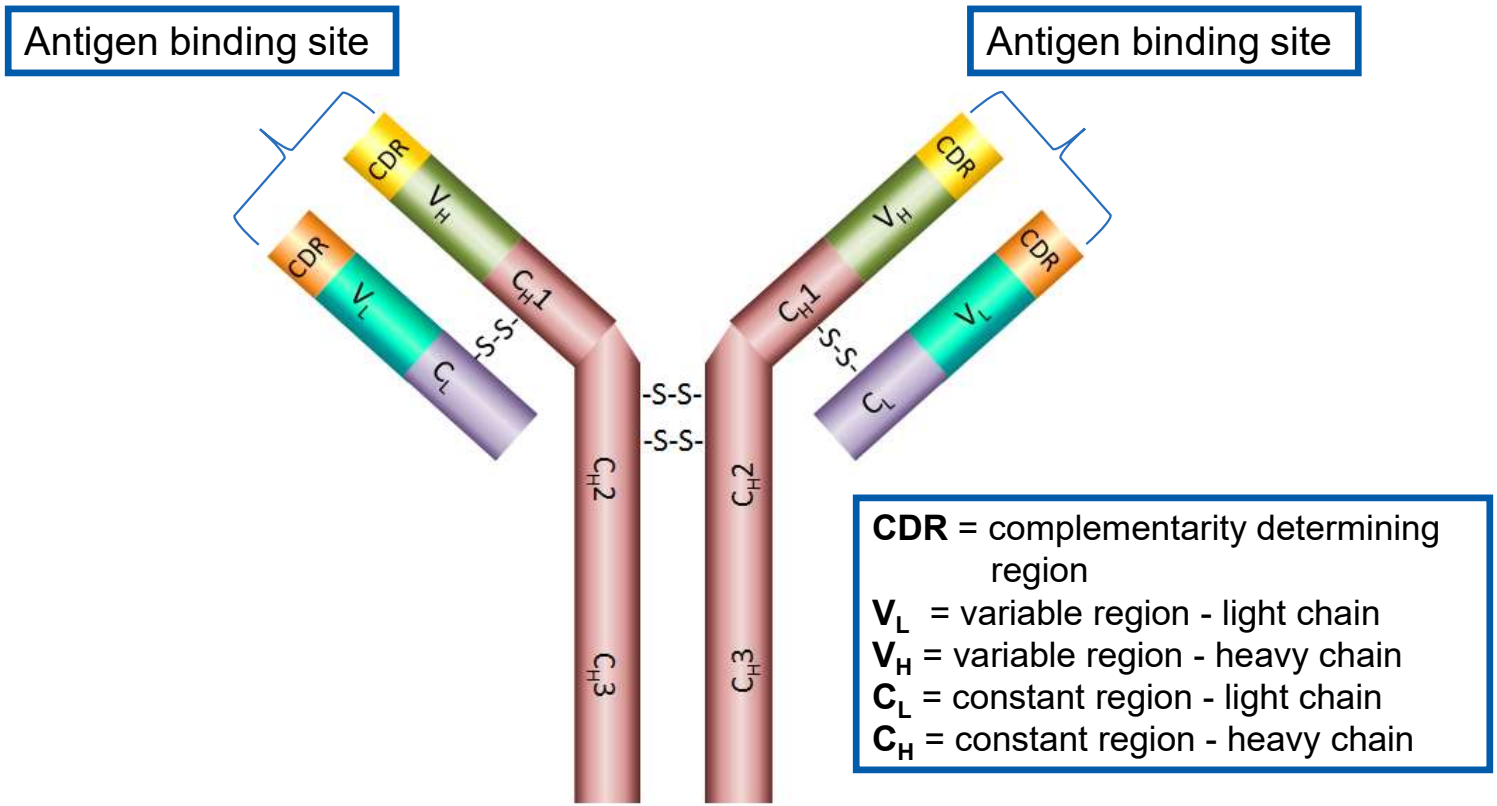

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Agenda

- STN sequence searchable databases and search methods
- Sequence Code Match (SCM) search options
- SCM variability: Motif searching symbols
- Antibody/CDR search example
- Hypothetical multiple CDR search discussion

Antibodies have a Y-shaped structure



CDR = complementarity determining region
$\mathbf{V_L}$ = variable region - light chain
$\mathbf{V_H}$ = variable region - heavy chain
$\mathbf{C_L}$ = constant region - light chain
$\mathbf{C_H}$ = constant region - heavy chain

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Search example: Multi-file sequence search on STNext

Search Question:

Find sequences in patents with priority dates earlier than 2005 which are similar (≥75%) to this antibody variable light (VL) chain region:

DVLMTQTPLSLPVSLGDQASISCRSSHYIVHSDGNTYLEWYLQKPGQSPKLLIY
KVSNRFSGVPDRFSGSGSGTDFTLKISRVEAEDLGIYYCFQGSHVP

and within which this CDR must be present:

KVSNRFSGVPDRFSGSG

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Antibody search methodology - Part 1

Save sequence in text file

Open CAS BLAST Client tool, run BLAST search for the variable light chain sequence

Save script and alignment files

Go onto CAS STNext, run script file generated from CAS BLAST client tool in REGISTRY

Run SCM search on CDR, AND the two sets together

Go into (H)CAplus, find corresponding patent records from REGISTRY search, limit by date

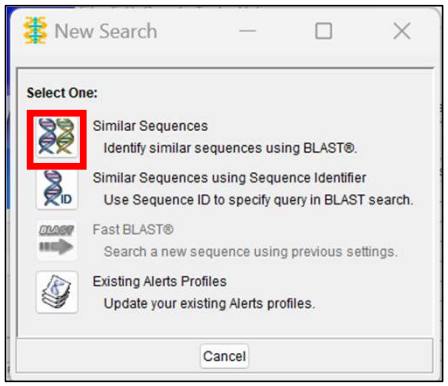
Display records of interest including HITRN

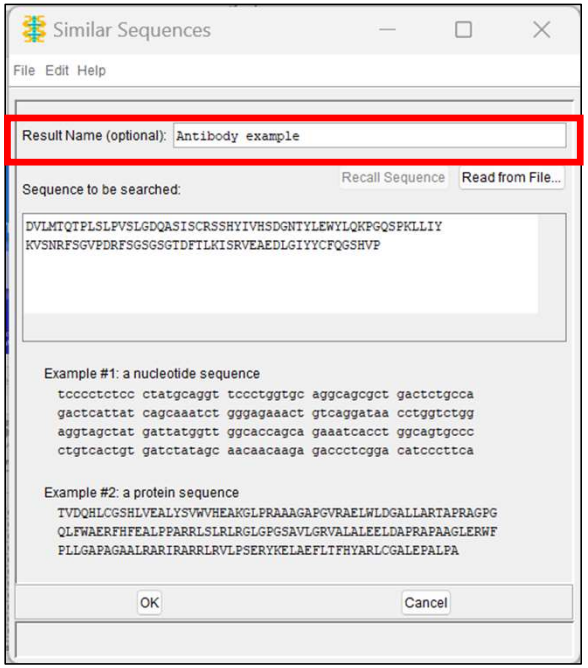
For more information on using the CAS BLAST Client tool, see Using CAS Registry BLAST® – CAS STNext (zendesk.com)

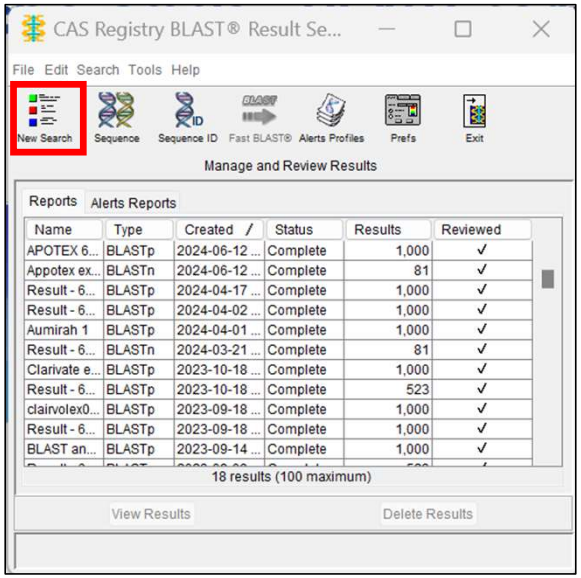
FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

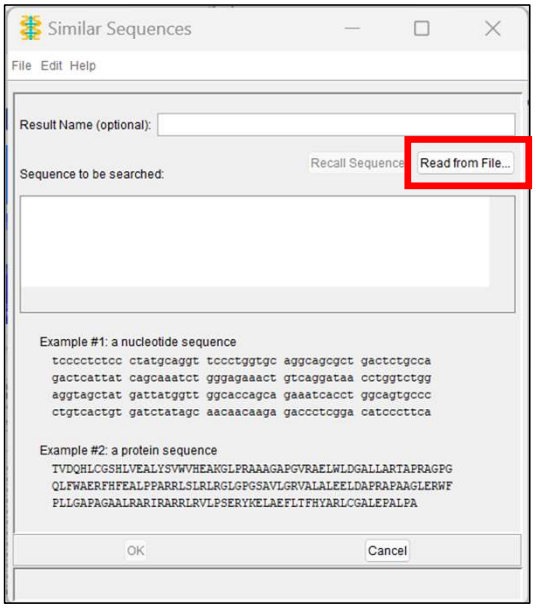
CAS
A division of the American Chemical Society

Antibody search example – REGISTRY BLAST portion





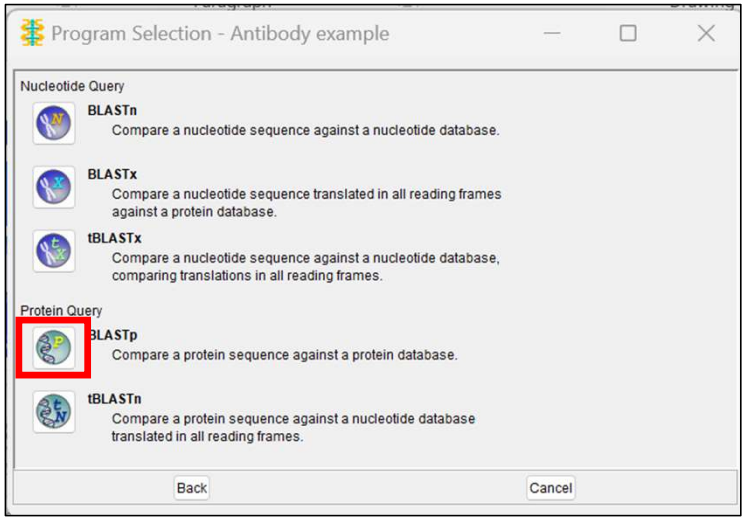


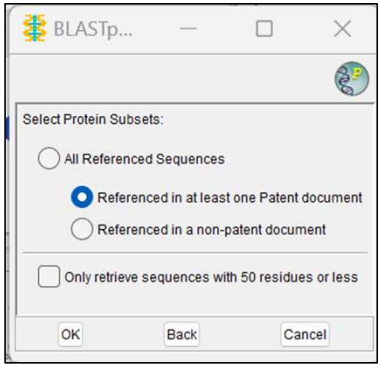


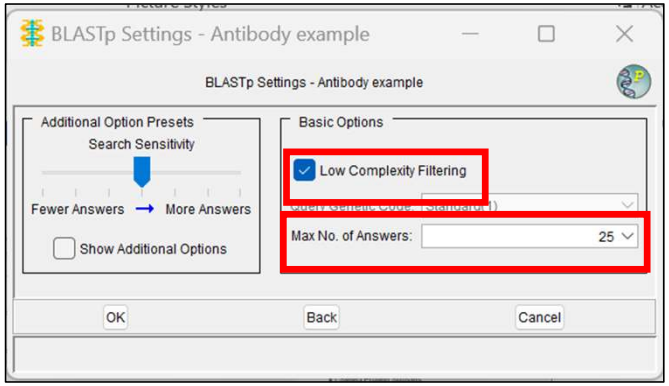
FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

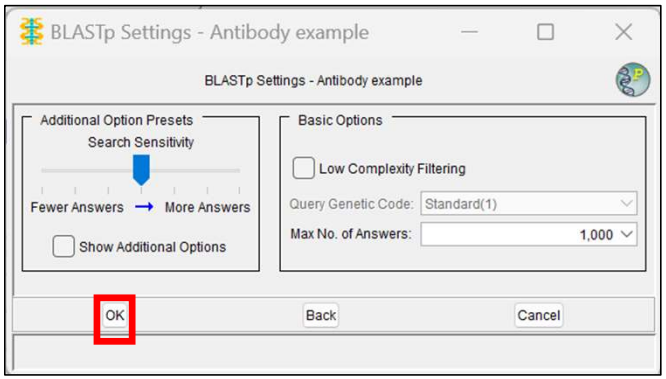
CAS
A division of the American Chemical Society

Antibody search example – REGISTRY BLAST portion





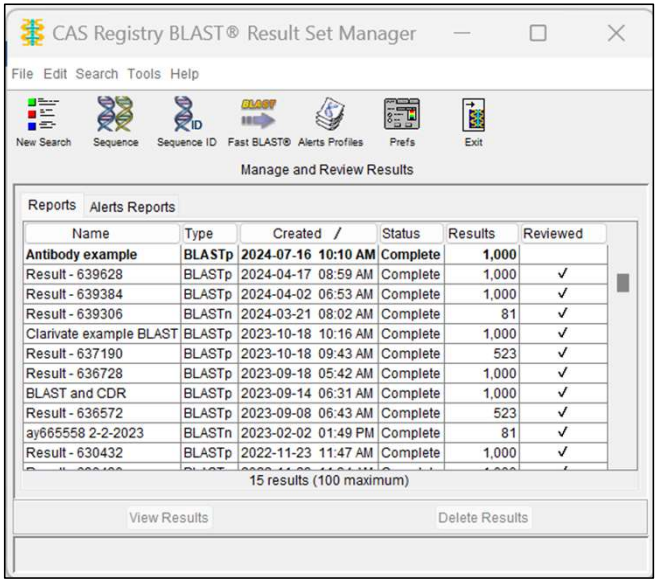


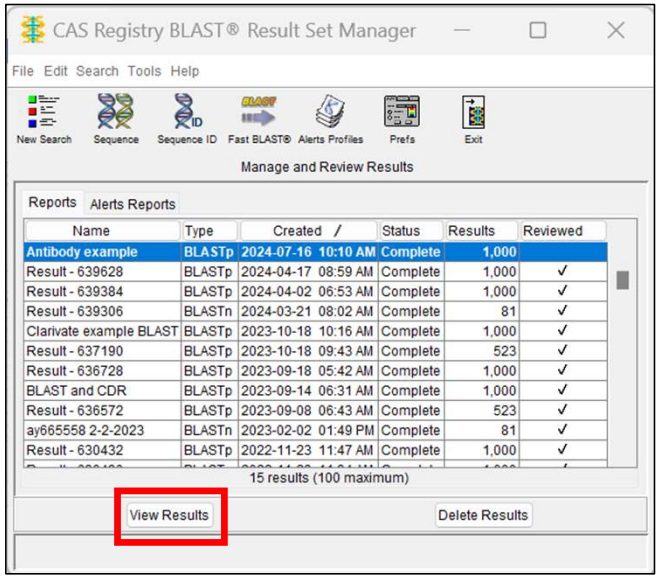


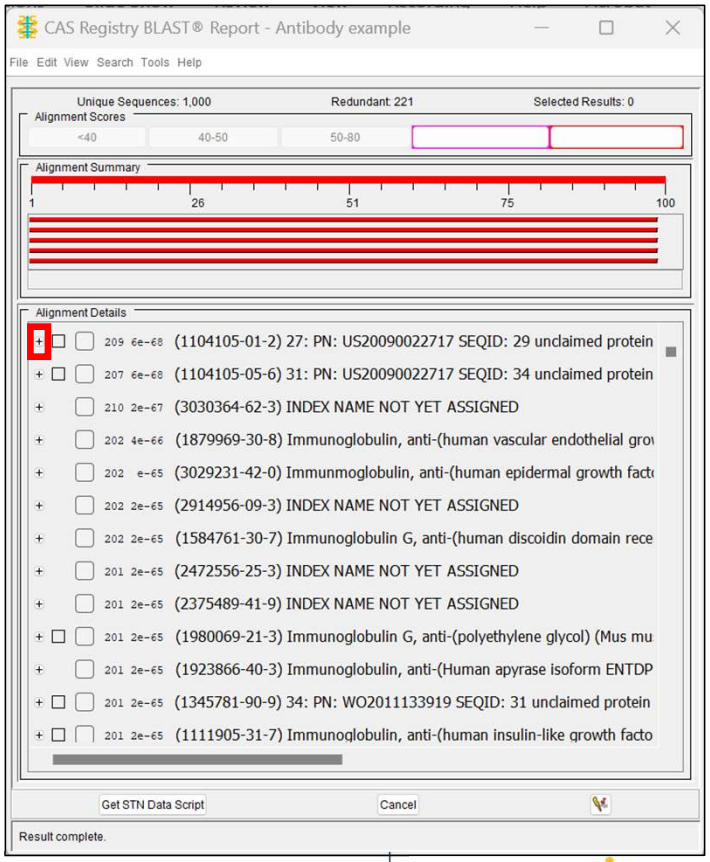
FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Antibody search example – REGISTRY BLAST portion



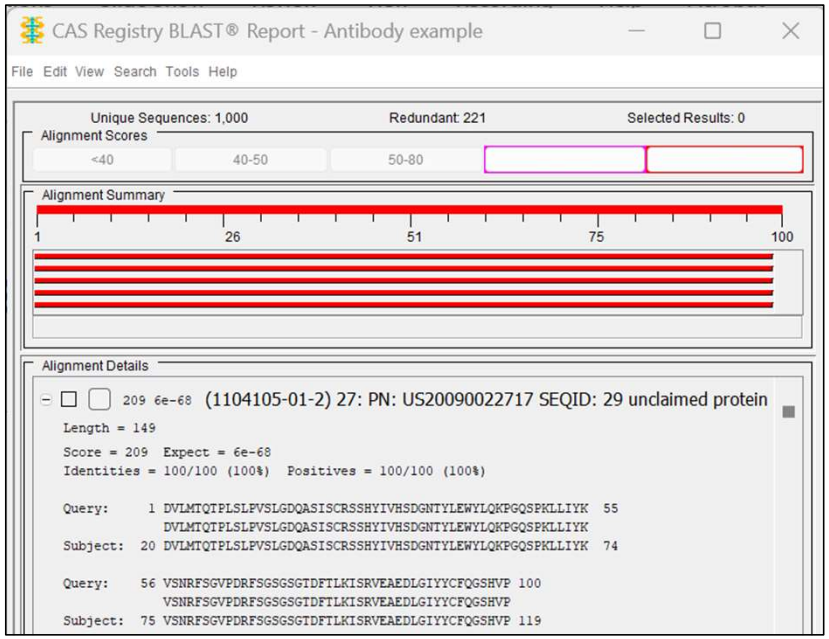




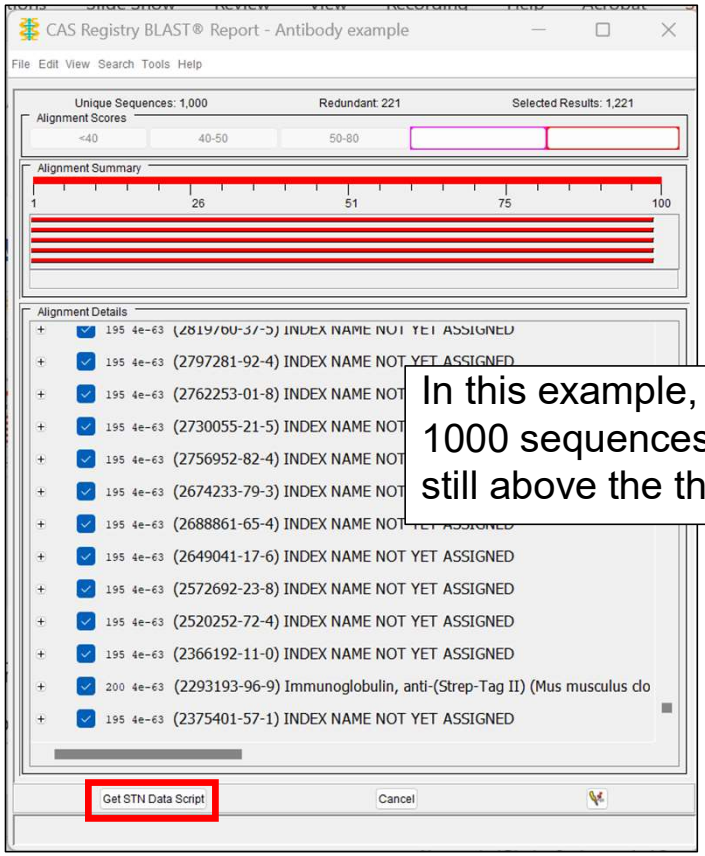
FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Antibody search example – REGISTRY BLAST portion



In this example, the score 209 represents a 100% match. Therefore 75% percent or better would be all scores above 156.

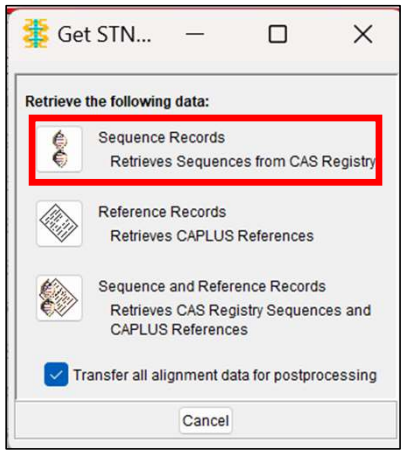


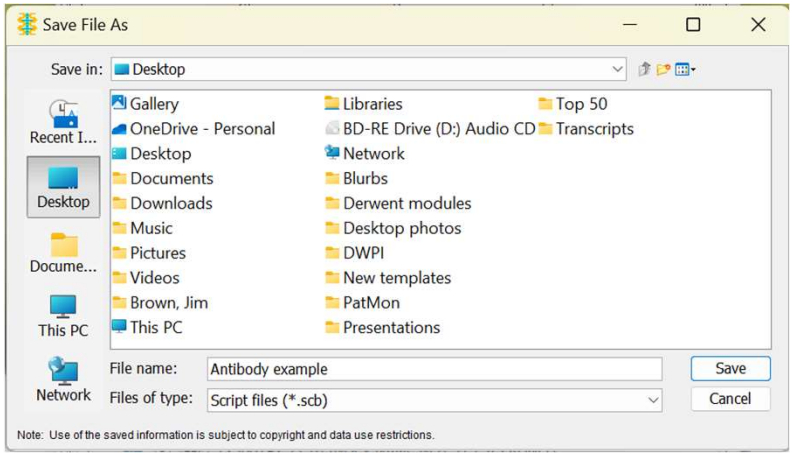
In this example, the limit of 1000 sequences was hit while still above the threshold of 75%.

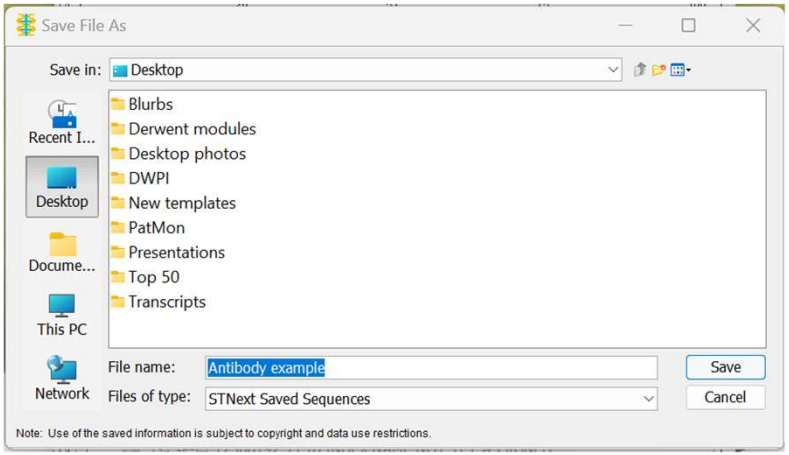
FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Antibody search example – REGISTRY BLAST portion





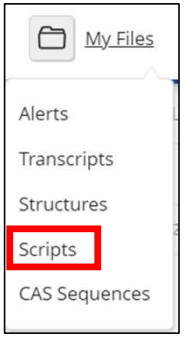


The REGISTRY Script file and the corresponding Saved Sequence file have been created and saved.

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Antibody search example – REGISTRY BLAST portion



FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Antibody search example – REGISTRY BLAST portion

. . .

```
              1 2674233-79-3/RN
              1 2688861-65-4/RN
              1 2649041-17-6/RN
              1 2572692-23-8/RN
              1 2520252-72-4/RN
              1 2366192-11-0/RN
              1 2293193-96-9/RN
              1 2375401-57-1/RN
L102         38 L92 OR L93 OR L94

=> S L95 OR L96 OR L97 OR L98 OR L99 OR L100 OR L101 OR L102

L103       1221 L95 OR L96 OR L97 OR L98 OR L99 OR L100 OR L101 OR L102
```


FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Antibody search example – REGISTRY SCM portion and (H)CAplus search

```
=> S KVSNRFSGVPDRFSGSG/SQSP

L104       8273 KVSNRFSGVPDRFSGSG/SQSP

=> S L103 AND L104

L105       1060 L103 AND L104

=> FILE HCAPLUS
```

. . .

```
=> S L105

L106        636 L105

=> S L106 AND PRY<2005

       4716397 PRY<2005
L107         66 L106 AND PRY<2005
```

```
=> D BIB HITRN 1-66


L107 ANSWER 1 OF 66  HCAPLUS  COPYRIGHT 2024 ACS on STN
PatentPak PDF | PatentPak PDF+ | PatentPak Interactive
AN   2009:642872  HCAPLUS Full-text
DN   150:561968
TI   Murine anti-IGF-I receptor antibody EM164 and humanized antibodies for
     diagnosis and treatment of IGF-I receptor-expressing cancer
IN   Singh, Rajeeva; Tavares, Daniel J.; Dagdigian, Nancy E.
PA   ImmunoGen Inc., USA
UO   ABBVIE INC
UOS  AbbVie
SO   U.S., 76 pp.
     CODEN: USXXAM
DT   Patent
LA   English
FAN.CNT 3
PPPI
     PATENT NO.          KIND  DATE      LANGUAGE   PatentPak
     ----------------    ----  --------  ---------- ------------------------
     US 7538195          B2    20090526  English    PDF | PDF+ | Interactive
     WO 2003106621       A2    20031224  English    PDF | PDF+ | Interactive
     AU 2003241580       A1    20031231  English    PDF
     AU 2003241580       B2    20091119  English    PDF
     CN 1678633          A     20051005  Chinese    PDF
     CN 1678633          B     20120829  Chinese    PDF
     JP 4699754          B2    20110615  Japanese   PDF
     CN 101693742        A     20100414  Chinese    PDF
```

. . .

```
     US 8268617          B2    Dead           20221013
     US 20100260678      A1    Dead           20201121
     JP 2011041564       A     Alive          20201120
     AU 2011201809       A1    Dead           20201121
     US 20130295050      A1    Dead           20201121
ASSIGNMENT HISTORY FOR US PATENT AVAILABLE IN LSUS DISPLAY FORMAT
IT   1156551-90-4 DP, humanized or chimeric derivs.
     1156551-97-1 DP, humanized or chimeric derivs.
     RL: BPN (Biosynthetic preparation); BSU (Biological study, unclassified);
     DGN (Diagnostic use); PRP (Properties); THU (Therapeutic use); BIOL
     (Biological study); PREP (Preparation); USES (Uses)
        (amino acid sequence; murine anti-IGF-I receptor antibody EM164 and
        humanized antibodies for diagnosis and treatment of IGF-I
        receptor-expressing cancer)
IT   1156554-02-7   1156554-10-7   1156554-13-0
     RL: PRP (Properties)
        (unclaimed protein sequence; murine anti-IGF-I receptor antibody EM164
        and humanized antibodies for diagnosis and treatment of IGF-I
        receptor-expressing cancer)
OSC.G   8     THERE ARE 8 CAPLUS RECORDS THAT CITE THIS RECORD (14 CITINGS)
RE.CNT  29    THERE ARE 29 CITED REFERENCES AVAILABLE FOR THIS RECORD
              ALL CITATIONS AVAILABLE IN THE RE FORMAT
```


FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Antibody search methodology – part 2

Go into GENESEQ

Click on My Files

Select Structures option

Click on Import Biosequence

Click on Browse to find variable light chain file

Upload sequence, run BLAST search

Run SCM search on CDR portion, AND the two sets together

TRANSFER PNs from (H)CAplus set, NOT from GENESEQ set

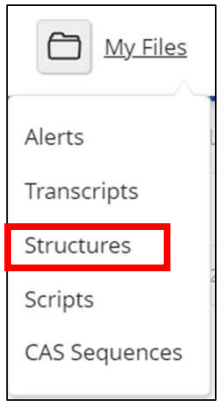
Display unique records with alignment

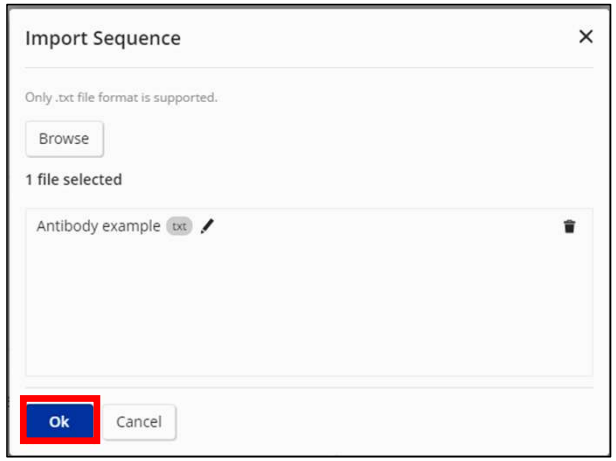
FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Antibody search example – GENESEQ BLAST portion

```
=> FILE GENESEQ
. . .
```

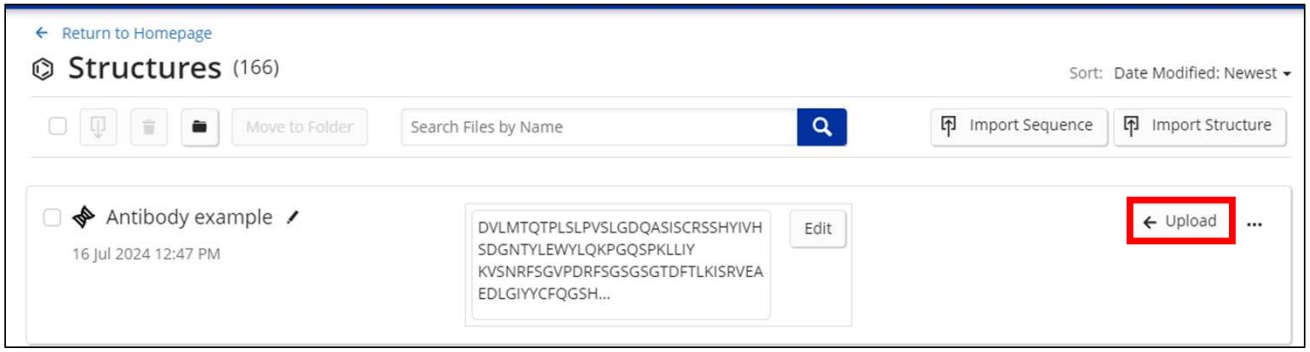




FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Antibody search example – GENESEQ BLAST portion



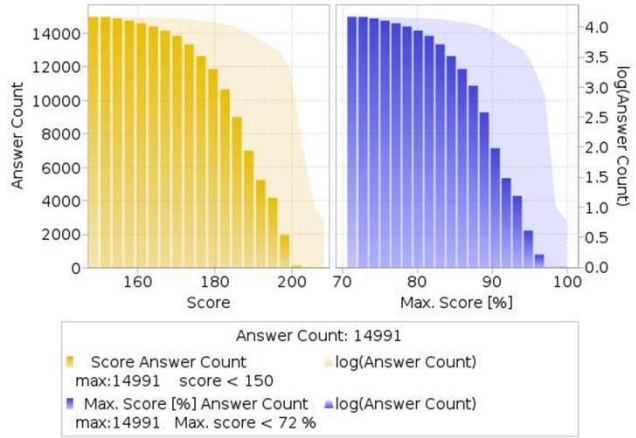

```
=>
Uploading sequence file: Antibody example


UPLOAD SUCCESSFULLY COMPLETED
L108 GENERATED

=> RUN BLAST L108/SQP -F F

Algorithm: BLAST - BLASTP. Version: 2.12.0+
```


FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Antibody search example – GENESEQ BLAST portion

```
14991 ANSWERS FOUND BELOW EXPECTATION VALUE OF: 10.0
GENESEQ
```




```
Query time:          1674

ENTER EITHER "ALL" TO KEEP ALL ANSWERS
OR ENTER THE MINIMUM SCORE VALUE YOU WISH TO KEEP
OR ENTER THE MINIMUM PERCENT OF SCORE FOLLOWED BY "% SCORE"
OR ENTER THE MINIMUM PERCENT OF IDENTITY FOLLOWED BY "% IDENT"
OR COMBINE MINIMUM PERCENT OF SCORE AND IDENTITY AS "X% SCORE Y% IDENT"
OR ENTER "END". "END" MUST BE ENTERED TO COMPLETE THE RUN COMMAND.
ENTER (ALL) OR ?:
```

```
ENTER (ALL) OR ?:ALL

L109  RUN STATEMENT CREATED
L109     14991 DVLMTQTPLSLPVSLGDQASISCRSSHYIVHSDGNTYLEWYLQKPGQSPK
               LLIYKVSNRFSGVPDRFSGSGSGTDFTLKISRVEAEDLGIYYCFQGSHVP
               /SQP -F F


ENTER EITHER "ALL" TO KEEP ALL ANSWERS
OR ENTER THE MINIMUM SCORE VALUE YOU WISH TO KEEP
OR ENTER THE MINIMUM PERCENT OF SCORE FOLLOWED BY "% SCORE"
OR ENTER THE MINIMUM PERCENT OF IDENTITY FOLLOWED BY "% IDENT"
OR COMBINE MINIMUM PERCENT OF SCORE AND IDENTITY AS "X% SCORE Y% IDENT"
OR ENTER "END". "END" MUST BE ENTERED TO COMPLETE THE RUN COMMAND.
ENTER (ALL) OR ?:75%

L110  RUN STATEMENT CREATED
L110     14759 DVLMTQTPLSLPVSLGDQASISCRSSHYIVHSDGNTYLEWYLQKPGQSPK
               LLIYKVSNRFSGVPDRFSGSGSGTDFTLKISRVEAEDLGIYYCFQGSHVP
               /SQP -F F


ENTER EITHER "ALL" TO KEEP ALL ANSWERS
OR ENTER THE MINIMUM SCORE VALUE YOU WISH TO KEEP
OR ENTER THE MINIMUM PERCENT OF SCORE FOLLOWED BY "% SCORE"
OR ENTER THE MINIMUM PERCENT OF IDENTITY FOLLOWED BY "% IDENT"
OR COMBINE MINIMUM PERCENT OF SCORE AND IDENTITY AS "X% SCORE Y% IDENT"
OR ENTER "END". "END" MUST BE ENTERED TO COMPLETE THE RUN COMMAND.
ENTER (ALL) OR ?:END
```


FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Antibody search example – GENESEQ SCM portion

```
=> RUN GETSEQ KVSNRFSGVPDRFSGSG/SQSP


RUN GETSEQ AT 13:01:10 ON 2024-07-16
COPYRIGHT (C) 2024 FIZ KARLSRUHE
Algorithm: GetSeq motif search. Version: 1.0.0

GetSeq motif search by FIZ Karlsruhe

GENESEQ
Query time:          211
L111  RUN STATEMENT CREATED
L111     22844 KVSNRFSGVPDRFSGSG/SQSP


=> S L110 AND L111

L112     10379 L110 AND L111

=> S L112 AND PRY<2005

      22359064 PRY<2005
L113       838 L112 AND PRY<2005
```

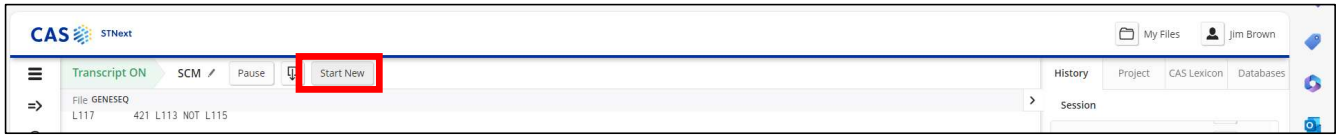

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Unique GENESEQ sequence records

```
=> TRA L107 PN 1-

L114        TRANSFER L107 1- PN :    1033 TERMS
L115     11007 L114
L116           QUE  TERMS FROM L114 WITH NO HITS:   941 TERMS

=> S L113 NOT L115

L117       421 L113 NOT L115
```

Optional: Start a new transcript for GENESEQ records.




```
=> SOR L117 SCORE D 1-

PROCESSING COMPLETED FOR L117
L118        421 SOR L117 1- SCORE D

=> D BIB ALIGN 1-
```


FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Unique GENESEQ sequence records

```
L118   ANSWER 1 OF 421 GENESEQ COPYRIGHT 2024 CLARIVATE on STN.
AN     AYN11564   GENESEQ  ED 20211030 UP 20220114
       DED 20110217 DUPD 20220113 Full-text
TI     Producing members of specific binding pairs - by expression in
       recombinant host cells with a secreting replicable genetic display
       package.
IN     Bonnert TP; Chiswell DJ; Clackson TP; Griffiths AD; Holliger KP;
       Hoogenboom HRJ; Jackson RH; Johnson KS; Marks JD; Mccafferty J; Pope AR;
       Winter GP
PA     MEDICAL RES COUNCIL (MRCX)
       MEDIMMUNE LTD (ASTR)
LA     English
DT     Patent
PI     US 20100317540  A1        20101216
PIT    USA1 FIRST PUBLISHED PATENT APPLICATION [FROM 2001 ONWARDS]
AI     US 2010-772829            20100503
PRAI   GB 1990-15198             19900710
       GB 1990-22845             19901019
       GB 1990-24503             19901112
       GB 1991-4744              19910306
       GB 1991-10549             19910515
       WO 1991-GB1134            19910710
       US 1993-971857            19930108
       US 1995-484893            19950607
       US 1999-417479            19991013
       US 2000-706507            20001103
       US 2006-555464            20061101
```

```
ALIGNMENT FROM L-NUMBER L110
Query Length: 100; Sequence Length: 249;
Score: 195.3 bits (495), 93.7% of highest possible score 208.4;
Expect value: 1.734e-48;
Identities: 93 / 100 (93.0%); Positives: 97 / 100 (97.0%);
Query Identity: 93.0%; Query Coverage: 100.0%;
Subject Identity: 37.3%; Subject Coverage: 40.2%;
Alignment Length: 100;
Q:   1 DVLMTQTPLSLPVSLGDQASISCRSSHYIVHSDGNTYLEWYLQKPGQSPKLLIYKVSNRF 60
       |+ +|||||||||||||||||||||  ||||+|||||||||||||||||||||||||||
S: 137 DIELTQTPLSLPVSLGDQASISCRSSQSIVHSNGNTYLEWYLQKPGQSPKLLIYKVSNRF 196

Q:  61 SGVPDRFSGSGSGTDFTLKISRVEAEDLGIYYCFQGSHVP                    100
       ||||||||||||||||||||||||||||||+||||||||||
S: 197 SGVPDRFSGSGSGTDFTLKISRVEAEDLGVYYCFQGSHVP                    236


ALIGNMENT FROM L-NUMBER L111
Sequence Length: 249;

Hits at: 191-207
     1 QVQLQESGGG LVQPGGSLRL SCATSGFTFS NYYMGWVRQP PGKALEWLGS VRNKVNGYTT
    61 EYSASVKGRF TISRDNFQSI LYLQINTLRT EDSATYYCAR GYDYGAWFAY WGQGTLVTVS
   121 SGGGGSGGGG SGGGGSDIEL TQTPLSLPVS LGDQASISCR SSQSIVHSNG NTYLEWYLQK
   181 PGQSPKLLIYKVSNRFSGVP DRFSGSGSGT DFTLKISRVE AEDLGVYYCF QGSHVPYTFG
               ========== =======
   241 GGTKLEIKR
```


FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the American Chemical Society

Agenda

- STN sequence searchable databases and search methods
- Sequence Code Match (SCM) search options
- SCM variability: Motif searching symbols
- Antibody/CDR search example
- Hypothetical multiple CDR search discussion

Search considerations

1. Understand the query from the patron
2. Consider how (else) this invention may be discussed in documents – it may be radically different than the patron's description of the invention
3. If using value-added databases, consider the tools that the database producer has to cover these concepts
4. Run search, evaluate results, repeat steps 2-4 until … satisfied?

Hypothetical example

Patron wants an antibody searched. It has three CDRs and the patron wants them to be in a certain order

… CDR1 … CDR2 … CDR3 …

You could build a SCM search for all sequences that have those three CDRs in that order, with appropriate gaps as part of search strategy

You get some results. Are you done?

Hypothetical example

What if patent says ‘claim an antibody with these three CDRs’, but it doesn’t specify the order?

> Search for three CDRs in three separate searches and AND the results together (REGISTRY and GENESEQ)

What if antibody is written about the CDRs but not the entire sequence?

> Search for all three CDRs in three separate searches, AND sets documents together in patent family databases ((H)CAplus and GENESEQ)

Each strategy captured different records. Which are valid???

Summary

Sequence code match is an important tool in sequence searching

This tool allows the user to define variability as specifically as needed

Can be used in conjunction with other search capabilities

Keywords, dates, assignees, etc.

Multiple databases increase the comprehensiveness of search

Between problems and progress are connections that matter

Jim Brown

CAS
help@cas.org
cas.org

FIZ Karlsruhe
EMEAhelp@cas.org

CONTACT

CAS
A division of the American Chemical Society