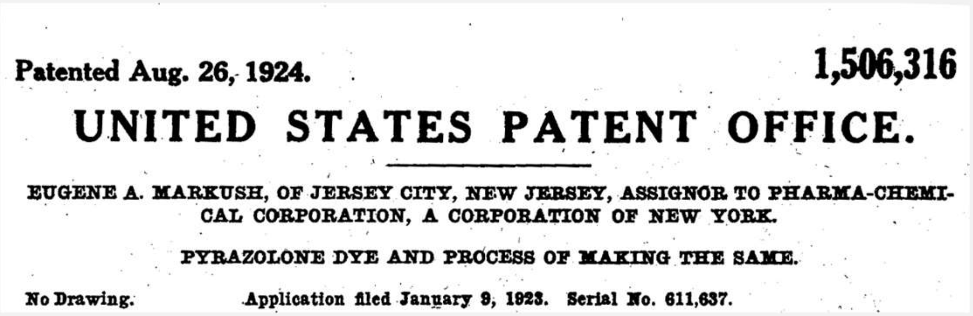
Sequence Searches and Databases



Patented Aug. 26, 1924. 1,506,316

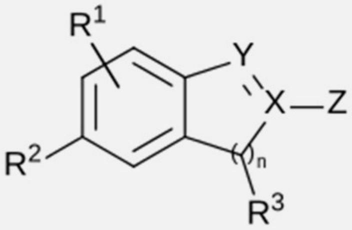
UNITED STATES PATENT OFFICE.

EUGENE A. MARKUSH, OF JERSEY CITY, NEW JERSEY, ASSIGNOR TO PHARMA-CHEMICAL CORPORATION, A CORPORATION OF NEW YORK.

PYRAZOLONE DYE AND PROCESS OF MAKING THE SAME.

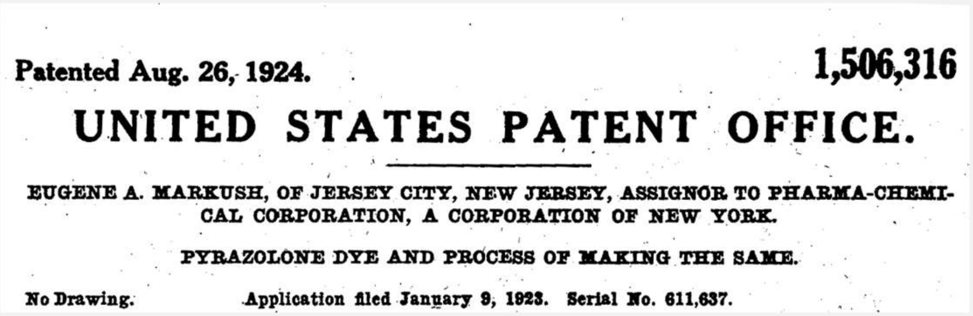
No Drawing. Application filed January 9, 1923. Serial No. 611,637.

«Markush Formula»



- multiple independently variable groups, such as **R groups**
- generic chemical structure patent filing
- protect whole classes of compounds with common properties
- **Selection inventions …**

Sequence Searches and Databases



Patented Aug. 26, 1924. **1,506,316**

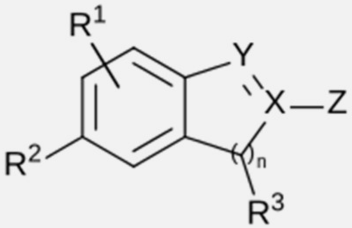
UNITED STATES PATENT OFFICE.

EUGENE A. MARKUSH, OF JERSEY CITY, NEW JERSEY, ASSIGNOR TO PHARMA-CHEMICAL CORPORATION, A CORPORATION OF NEW YORK.

PYRAZOLONE DYE AND PROCESS OF MAKING THE SAME.

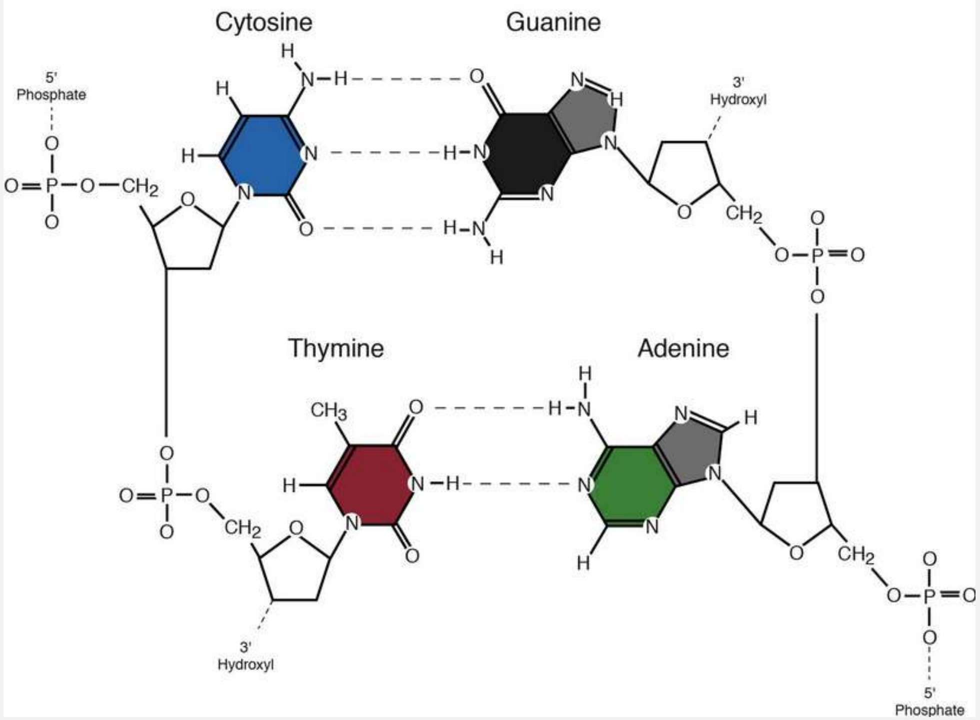
No Drawing. Application filed January 9, 1923. Serial No. 611,637.

«Markush Formula»



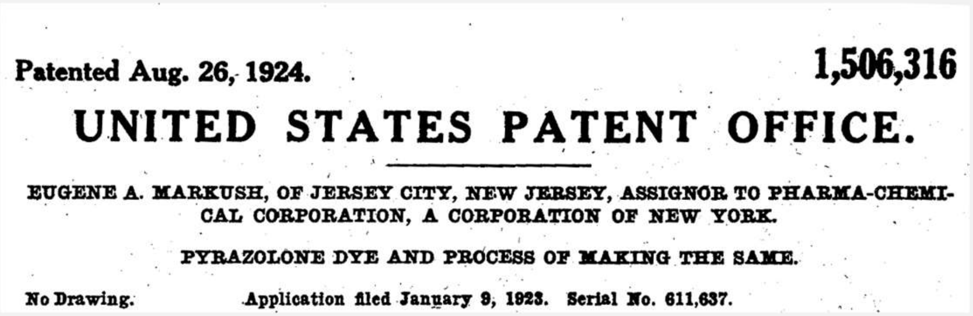
- multiple independently variable groups, such as **R groups**
- generic chemical structure patent filing
- protect whole classes of compounds with common properties
- **Selection inventions …**

… and biological compounds?



- **Large chemical compounds described as sequences**

Sequence Searches and Databases



Patented Aug. 26, 1924. 1,506,316

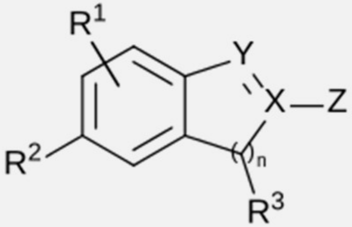
UNITED STATES PATENT OFFICE.

EUGENE A. MARKUSH, OF JERSEY CITY, NEW JERSEY, ASSIGNOR TO PHARMA-CHEMICAL CORPORATION, A CORPORATION OF NEW YORK.

PYRAZOLONE DYE AND PROCESS OF MAKING THE SAME.

No Drawing. Application filed January 9, 1923. Serial No. 611,637.

«Markush Formula»

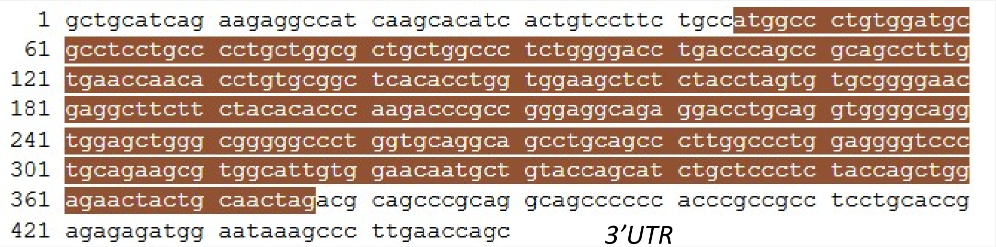


- multiple independently variable groups, such as **R groups**
- generic chemical structure patent filing
- protect whole classes of compounds with common properties
- **Selection inventions …**

- **Nucleotide sequences → absolute compound protection,** but the protection is limited to the **sequence segments that perform the function** specifically described in the patent (Swiss patent law).

5'UTR

Coding region



```
  1 gctgcatcag aagaggccat caagcacatc actgtccttc tgccatggcc ctgtggatgc
 61 gcctcctgcc cctgctggcg ctgctggccc tctggggacc tgacccagcc gcagcctttg
121 tgaaccaaca cctgtgcggc tcacacctgg tggaagctct ctacctagtg tgcggggaac
181 gaggcttctt ctacacaccc aagacccgcc gggaggcaga ggacctgcag gtggggcagg
241 tggagctggg cgggggccct ggtgcaggca gcctgcagcc cttggccctg gaggggtccc
301 tgcagaagcg tggcattgtg gaacaatgct gtaccagcat ctgctccctc taccagctgg
361 agaactactg caactagacg cagcccgcag gcagcccccc acccgccgcc tcctgcaccg
421 agagagatgg aataaagccc ttgaaccagc
```

3'UTR

- **Proteins → absolute compound protection**

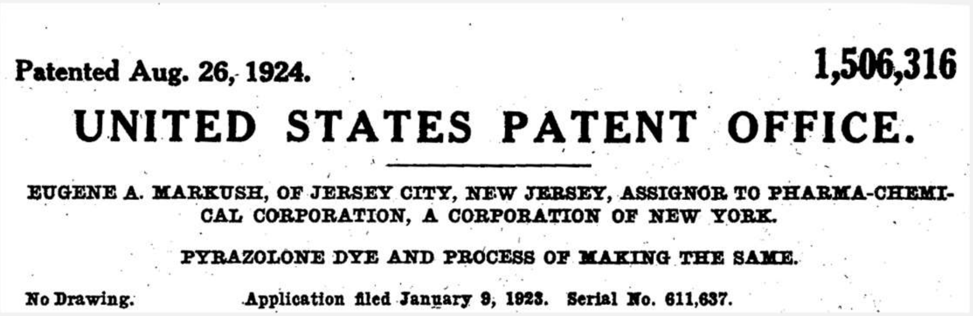
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCG
ERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQK
RGIVEQCCTSICSLYQLENYCN

insulin

https://www.ncbi.nlm.nih.gov/

IngridB.Mueller@ipi.ch

Sequence Searches and Databases



Patented Aug. 26, 1924. 1,506,316

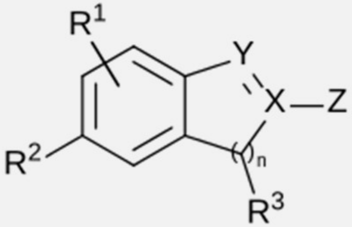
UNITED STATES PATENT OFFICE.

EUGENE A. MARKUSH, OF JERSEY CITY, NEW JERSEY, ASSIGNOR TO PHARMA-CHEMICAL CORPORATION, A CORPORATION OF NEW YORK.

PYRAZOLONE DYE AND PROCESS OF MAKING THE SAME.

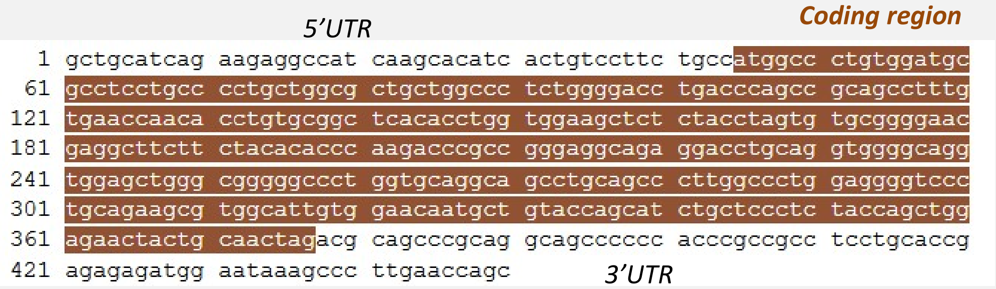
No Drawing. Application filed January 9, 1923. Serial No. 611,637.

«Markush Formula»



- multiple independently variable groups, such as **R groups**
- generic chemical structure patent filing
- protect whole classes of compounds with common properties
- **Selection inventions …**

- **Nucleotide sequences → absolute compound protection,** but the protection is limited to the **sequence segments that perform the function** specifically described in the patent (Swiss patent law).



```
  1 gctgcatcag aagaggccat caagcacatc actgtccttc tgccatggcc ctgtggatgc
 61 gcctcctgcc cctgctggcg ctgctggccc tctggggacc tgacccagcc gcagcctttg
121 tgaaccaaca cctgtgcggc tcacacctgg tggaagctct ctacctagtg tgcggggaac
181 gaggcttctt ctacacaccc aagacccgcc gggaggcaga ggacctgcag gtggggcagg
241 tggagctggg cgggggccct ggtgcaggca gcctgcagcc cttggccctg gaggggtccc
301 tgcagaagcg tggcattgtg gaacaatgct gtaccagcat ctgctccctc taccagctgg
361 agaactactg caactagacg cagcccgcag gcagcccccc acccgccgcc tcctgcaccg
421 agagagatgg aataaagccc ttgaaccagc
```

- **Proteins → absolute compound protection**

MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCG
ERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQK
RGIVEQCCTSICSLYQLENYCN

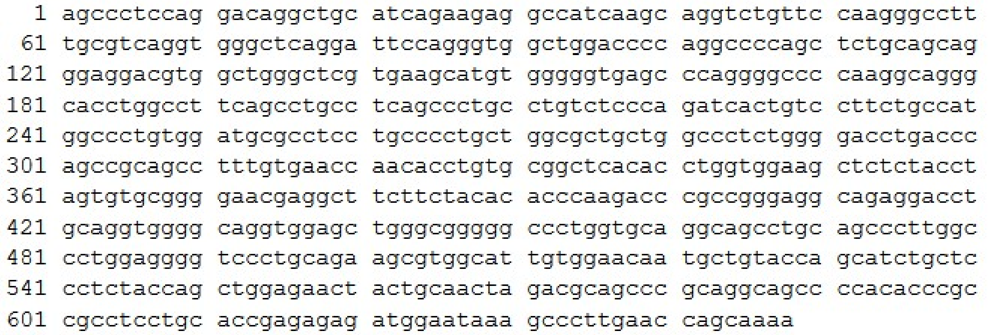
insulin

https://www.ncbi.nlm.nih.gov/

IngridB.Mueller@ipi.ch

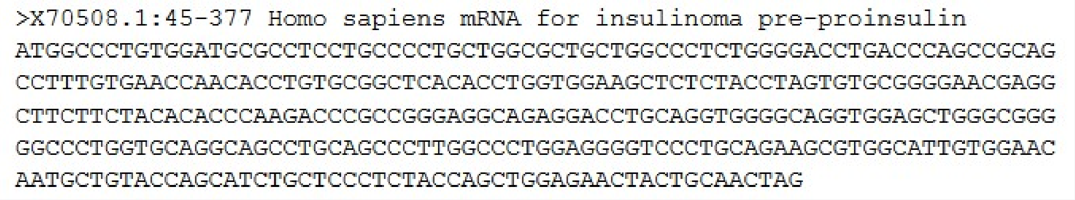
Sequence Searches and Databases

DNA



```
  1 agccctccag gacaggctgc atcagaagag gccatcaagc aggtctgttc caagggcctt
 61 tgcgtcaggt gggctcagga ttccagggtg gctggacccc aggccccagc tctgcagcag
121 ggaggacgtg gctgggctcg tgaagcatgt gggggtgagc ccaggggccc caaggcaggg
181 cacctggcct tcagcctgcc tcagccctgc ctgtctccca gatcactgtc cttctgccat
241 ggccctgtgg atgcgcctcc tgcccctgct ggcgctgctg gccctctggg gacctgaccc
301 agccgcagcc tttgtgaacc aacacctgtg cggctcacac ctggtggaag ctctctacct
361 agtgtgcggg gaacgaggct tcttctacac acccaagacc cgccgggagg cagaggacct
421 gcaggtgggg caggtggagc tgggcggggg ccctggtgca ggcagcctgc agcccttggc
481 cctggagggg tccctgcaga agcgtggcat tgtggaacaa tgctgtacca gcatctgctc
541 cctctaccag ctggagaact actgcaacta gacgcagccc gcaggcagcc ccacacccgc
601 cgcctcctgc accgagagag atggaataaa gcccttgaac cagcaaaa
```

mRNA



```
>X70508.1:45-377 Homo sapiens mRNA for insulinoma pre-proinsulin
ATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGCAG
CCTTTGTGAACCAACACCTGTGCGGCTCACACCTGGTGGAAGCTCTCTACCTAGTGTGCGGGGAACGAGG
CTTCTTCTACACACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGGTGGGGCAGGTGGAGCTGGGCGGG
GGCCCTGGTGCAGGCAGCCTGCAGCCCTTGGCCCTGGAGGGGTCCCTGCAGAAGCGTGGCATTGTGGAAC
AATGCTGTACCAGCATCTGCTCCCTCTACCAGCTGGAGAACTACTGCAACTAG
```

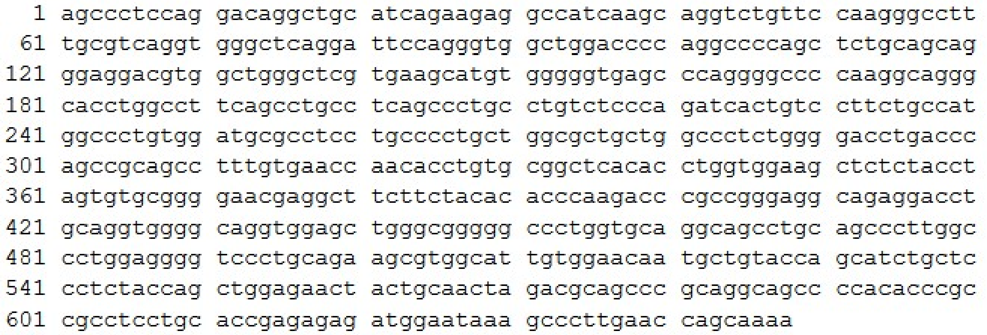
Transcription from DNA to pre-mRNA

- Processing into mRNA

- here shown as cDNA sequence (T not U)

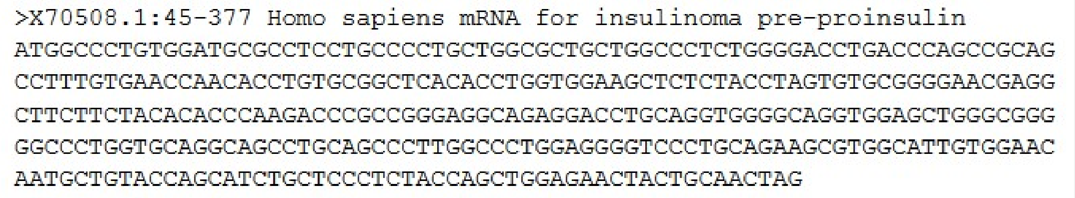
Sequence Searches and Databases

DNA



```
  1 agccctccag gacaggctgc atcagaagag gccatcaagc aggtctgttc caagggcctt
 61 tgcgtcaggt gggctcagga ttccagggtg gctggacccc aggccccagc tctgcagcag
121 ggaggacgtg gctgggctcg tgaagcatgt gggggtgagc ccaggggccc caaggcaggg
181 cacctggcct tcagcctgcc tcagccctgc ctgtctccca gatcactgtc cttctgccat
241 ggccctgtgg atgcgcctcc tgcccctgct ggcgctgctg gccctctggg gacctgaccc
301 agccgcagcc tttgtgaacc aacacctgtg cggctcacac ctggtggaag ctctctacct
361 agtgtgcggg gaacgaggct tcttctacac acccaagacc cgccgggagg cagaggacct
421 gcaggtgggg caggtggagc tgggcggggg ccctggtgca ggcagcctgc agcccttggc
481 cctggagggg tccctgcaga agcgtggcat tgtggaacaa tgctgtacca gcatctgctc
541 cctctaccag ctggagaact actgcaacta gacgcagccc gcaggcagcc ccacacccgc
601 cgcctcctgc accgagagag atggaataaa gcccttgaac cagcaaaa
```

mRNA



```
>X70508.1:45-377 Homo sapiens mRNA for insulinoma pre-proinsulin
ATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGCAG
CCTTTGTGAACCAACACCTGTGCGGCTCACACCTGGTGGAAGCTCTCTACCTAGTGTGCGGGGAACGAGG
CTTCTTCTACACACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGGTGGGGCAGGTGGAGCTGGGCGGG
GGCCCTGGTGCAGGCAGCCTGCAGCCCTTGGCCCTGGAGGGGTCCCTGCAGAAGCGTGGCATTGTGGAAC
AATGCTGTACCAGCATCTGCTCCCTCTACCAGCTGGAGAACTACTGCAACTAG
```

Protein

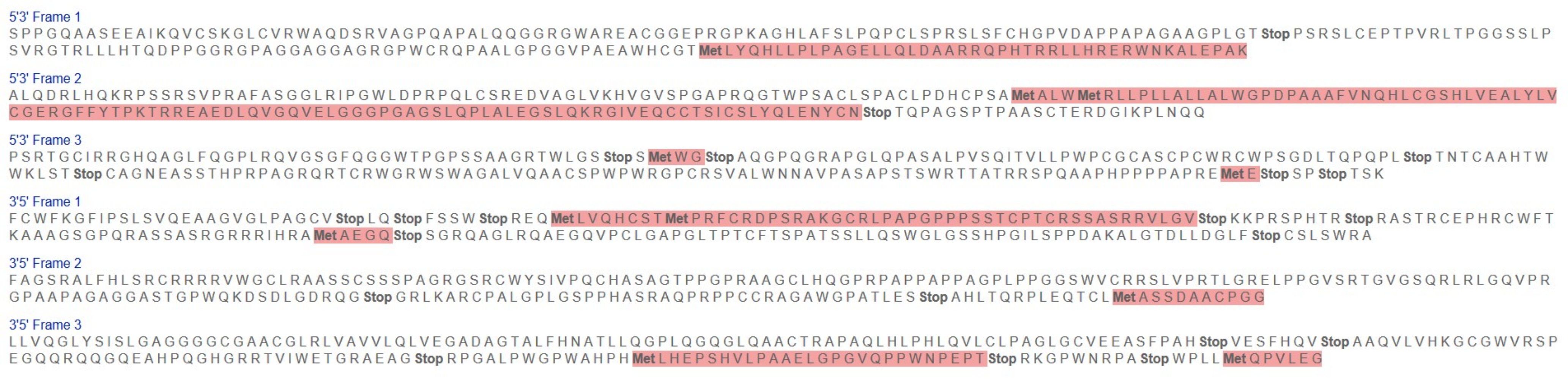
Transcription from DNA to pre-mRNA

- Processing into mRNA
- here shown as cDNA sequence (*T not U*)

Translation into protein (3 sense, 3 antisense frames)

- Which one is the right reading frame?
- Protein always starts with Methionine !

https://web.expasy.org/translate/



5'3' Frame 1

S P P G Q A A S E E A I K Q V C S K G L C V R W A Q D S R V A G P Q A P A L Q Q G G R G W A R E A C G G E P R G P K A G H L A F S L P Q P C L S P R S L S F C H G P V D A P P A P A G A A G P L G T **Stop** P S R S L C E P T P V R L T P G G S S L P
S V R G T R L L L H T Q D P P G G R G P A G G A G G A G R G P W C R Q P A A L G P G G V P A E A W H C G T **Met** L Y Q H L L P L P A G E L L Q L D A A R R Q P H T R R L L H R E R W N K A L E P A K

5'3' Frame 2

A L Q D R L H Q K R P S S R S V P R A F A S G G L R I P G W L D P R P Q L C S R E D V A G L V K H V G V S P G A P R Q G T W P S A C L S P A C L P D H C P S A **Met** A L W **Met** R L L P L L A L L A L W G P D P A A A F V N Q H L C G S H L V E A L Y L V
C G E R G F F Y T P K T R R E A E D L Q V G Q V E L G G G P G A G S L Q P L A L E G S L Q K R G I V E Q C C T S I C S L Y Q L E N Y C N **Stop** T Q P A G S P T P A A S C T E R D G I K P L N Q Q

5'3' Frame 3

P S R T G C I R R G H Q A G L F Q G P L R Q V G S G F Q G G W T P G P S S A A G R T W L G S **Stop** S **Met** W G **Stop** A Q G P Q G R A P G L Q P A S A L P V S Q I T V L L P W P C G C A S C P C W R C W P S G D L T Q P Q P L **Stop** T N T C A A H T W
W K L S T **Stop** C A G N E A S S T H P R P A G R Q R T C R W G R W S W A G A L V Q A A C S P W P W R G P C R S V A L W N N A V P A S A P S T S W R T T A T R R S P Q A A P H P P P P A P R E **Met** E **Stop** S P **Stop** T S K

3'5' Frame 1

F C W F K G F I P S L S V Q E A A G V G L P A G C V **Stop** L Q **Stop** F S S W **Stop** R E Q **Met** L V Q H C S T **Met** P R F C R D P S R A K G C R L P A P G P P P S S T C P T C R S S A S R R V L G V **Stop** K K P R S P H T R **Stop** R A S T R C E P H R C W F T
K A A A G S G P Q R A S S A S R G R R R I H R A **Met** A E G Q **Stop** S G R Q A G L R Q A E G Q V P C L G A P G L T P T C F T S P A T S S L L Q S W G L G S S H P G I L S P P D A K A L G T D L L D G L F **Stop** C S L S W R A

3'5' Frame 2

F A G S R A L F H L S R C R R R R V W G C L R A A S S C S S S P A G R G S R C W Y S I V P Q C H A S A G T P P G P R A A G C L H Q G P R P A P P A P P A G P L P P G G S W V C R R S L V P R T L G R E L P P G V S R T G V G S Q R L R L G Q V P R
G P A A P A G A G G A S T G P W Q K D S D L G D R Q G **Stop** G R L K A R C P A L G P L G S P P H A S R A Q P R P P C C R A G A W G P A T L E S **Stop** A H L T Q R P L E Q T C L **Met** A S S D A A C P G G

3'5' Frame 3

L L V Q G L Y S I S L G A G G G G C G A A C G L R L V A V V L Q L V E G A D A G T A L F H N A T L L Q G P L Q G Q G L Q A A C T R A P A Q L H L P H L Q V L C L P A G L G C V E E A S F P A H **Stop** V E S F H Q V **Stop** A A Q V L V H K G C G W V R S P
E G Q Q R Q Q G Q E A H P Q G H G R R T V I W E T G R A E A G **Stop** R P G A L P W G P W A H P H **Met** L H E P S H V L P A A E L G P G V Q P P W N P E P T **Stop** R K G P W N R P A **Stop** W P L L **Met** Q P V L E G

Sequence Searches and Databases

Performing Keyword Searches in Life Science using patent examinar tools such as

STN, Epoquenet

Patent databases:

- Abstract DB (Epodoc, DWPI)
- Full-text
- Machine Translation

Non patent literature abstract databases

- BIOSIS, MEDLINE, EMBASE…

Sequence Searches and Databases

Performing Keyword Searches in Life Science using Patent examinar tools such as

STN, Epoquenet

Patent databases:

- Abstract DB (Epodoc, DWPI)
- Full-text
- Machine Translation

Non patent literature abstract databases

- BIOSIS, MEDLINE, EMBASE…

Search comprising:

> features of the inventive concept

> Proteins, genes (names, text terms, synonymes, CAS numbers, chemical identifiers)

> Function of the biomolecule (enzymatic reaction, antibody)

> Application (pharma, food, agriculture…)

> Specific patent classes

Sequence Searches and Databases

Performing Keyword Searches in Life Science using Patent examinar tools such as

STN, Epoquenet

Patent databases:

- Abstract DB (Epodoc, DWPI)
- Full-text
- Machine Translation

Non patent literature abstract databases

- BIOSIS, MEDLINE, EMBASE…

Search comprising:

> features of the inventive concept
>
> Proteins, genes (names, text terms, synonymes, CAS numbers, chemical identifiers)
>
> Function of the biomolecule (enzymatic reaction, antibody)
>
> Application (pharma, food, agriculture…)
>
> Specific patent classes

Combining keyword with classes…

… and if necessary with *sequence searches*

Sequence Searches and Databases

Performing the search using peptide or nucleic acid sequences as:

- Exact sequence
- Subsequence (fragment within a longer context)
- Motifs (e.g. repeats, alternatives, spacer…)
- Uncommon sequences

} **identity** (100% ID)

- **BLAST** (**B**asic **L**ocal **A**lignment **S**earch **T**ool) → **similarity** (in % of sequence identity, ID)

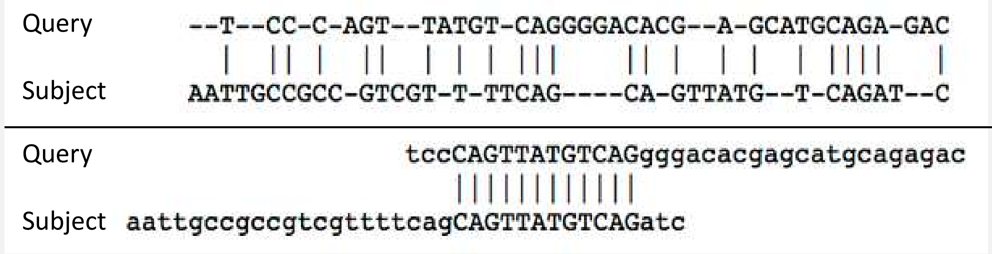
Sequence Searches and Databases

Performing the search using peptide or nucleic acid sequences as:

- Exact sequence
- Subsequence (fragment within a longer context)
- Motifs (e.g. repeats, alternatives, spacer…)
- Uncommon sequences

} **identity** (100% ID)

- **BLAST** (**B**asic **L**ocal **A**lignment **S**earch **T**ool) → **similarity** (in % of sequence identity, ID)



```
Query    --T--CC-C-AGT--TATGT-CAGGGGACACG--A-GCATGCAGA-GAC
           |  || |  ||  | | | |||    || |  | |  | ||||   |
Subject  AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG--T-CAGAT--C
```

Global alignment

```
Query                      tccCAGTTATGTCAGgggacacgagcatgcagagac
                              ||||||||||||
Subject  aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

Local alignment

Query (reference sequence)

Subject (retrieved sequence from DB)

IngridB.Mueller@ipi.ch

Sequence Searches and Databases

Performing the search using peptide or nucleic acid sequences as:

- Exact sequence
- Subsequence (fragment within a longer context)
- Motifs (e.g. repeats, alternatives, spacer…)
- Uncommon sequences

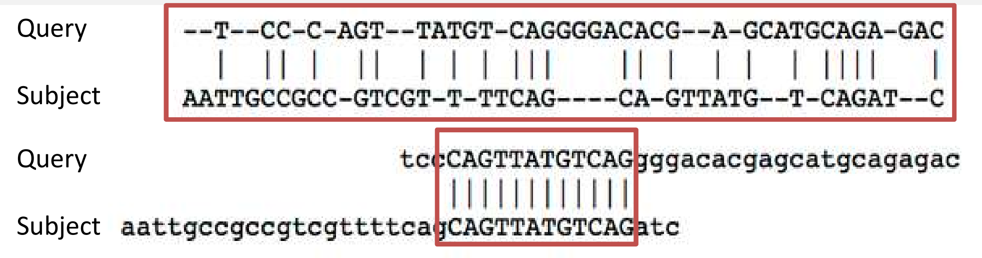
} identity (100% ID)

- **BLAST** (**B**asic **L**ocal **A**lignment **S**earch **T**ool) → similarity (in % of sequence identity, ID)

Global

vs

Local



```
Query                         --T--CC-C-AGT--TATGT-CAGGGACACG--A-GCATGCAGA-GAC
                                |  || |  ||  | | | |||    || |  | |  | ||||   |
Subject                       AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG--T-CAGAT--C

Query                      tccCAGTTATGTCAGgggacacgagcatgcagagac
                              ||||||||||||
Subject   aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

alignment ID = 46.9%

Query coverage is 100%

alignment ID = 100%

Query coverage is < 50%

IngridB.Mueller@ipi.ch

Sequence Searches and Databases

Performing the search using peptide or nucleic acid sequences as:

- Exact sequence
- Subsequence (fragment within a longer context)
- Motifs (e.g. repeats, alternatives, spacer…)
- Uncommon sequences
- **BLAST** (**B**asic **L**ocal **A**lignment **S**earch **T**ool)

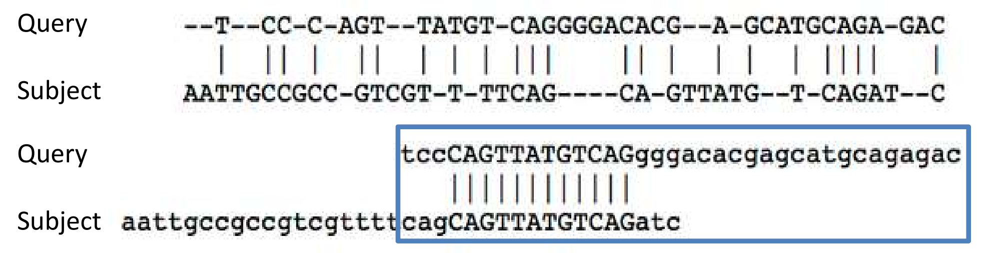
} identity (100% ID)

→ similarity (in % of sequence identity, ID)

Global

vs

Local



```
Query       --T--CC-C-AGT--TATGT-CAGGGGACACG--A-GCATGCAGA-GAC
              |  || |  ||  | | | |||    || |  | |  | ||||   |
Subject     AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG--T-CAGAT--C

Query                     tccCAGTTATGTCAGgggacacgagcatgcagagac
                             ||||||||||||
Subject aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

Query ID < 50%

-> matches over the length of the query

Query coverage is 100%

Sequence Searches and Databases

Performing the search using peptide or nucleic acid sequences as:

- Exact sequence
- Subsequence (fragment within a longer context)
- Motifs (e.g. repeats, alternatives, spacer…)
- Uncommon sequences

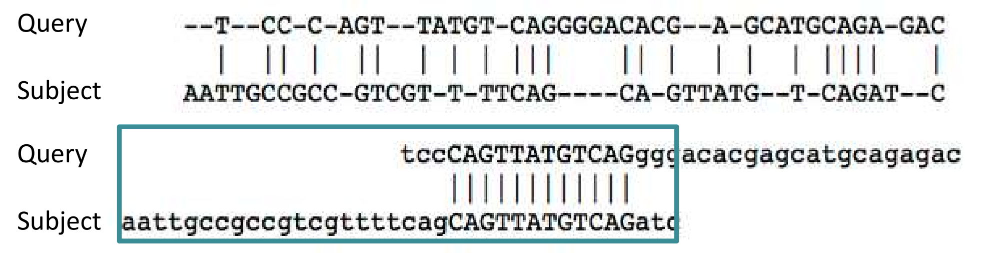
} identity (100% ID)

- **BLAST** (**B**asic **L**ocal **A**lignment **S**earch **T**ool) → similarity (in % of sequence identity, ID)

Global

vs

Local



```
Query    --T--CC-C-AGT--TATGT-CAGGGGACACG--A-GCATGCAGA-GAC
           |  || |  ||  | | | |||    || |    |  | ||||   |
Subject  AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG--T-CAGAT--C

Query                      tccCAGTTATGTCAGgggacacgagcatgcagagac
                              ||||||||||||
Subject  aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

Subject ID < 50%

-> matches over the length of the subject

Query coverage is < 50%

IngridB.Mueller@ipi.ch

Sequence Searches and Databases

Performing the search using peptide or nucleic acid sequences as:

- Exact sequence
- Subsequence (fragment within a longer context)
- Motifs (e.g. repeats, alternatives, spacer…)
- Uncommon sequences

} identity (100% ID)

- **BLAST** (**B**asic **L**ocal **A**lignment **S**earch **T**ool) → similarity (in % of sequence identity, ID)

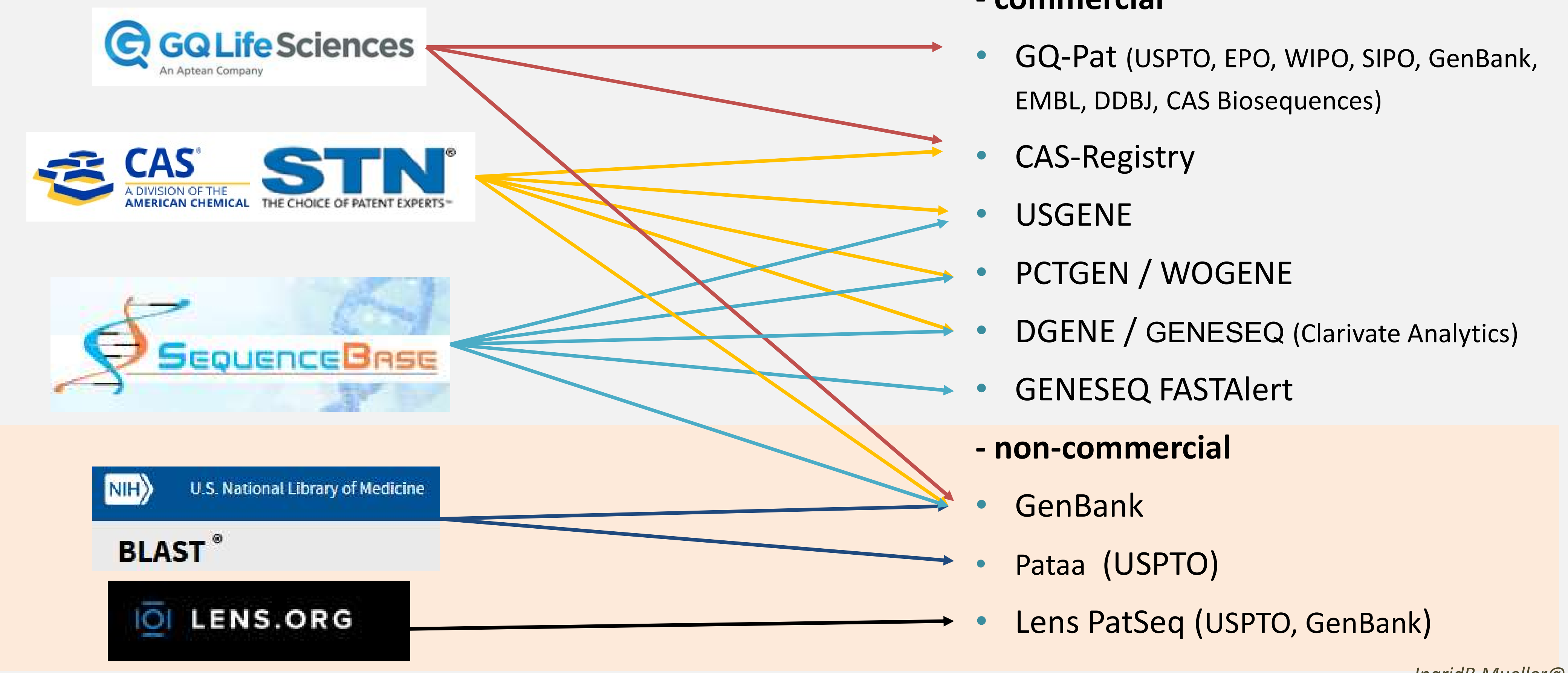
FAQ:

- *When to use which approach ?*
- *How to combine several sequences (e.g. CDRs of an antibody) ?*
- *Protein or DNA or both ?*
- *Long / short sequences ?*
- *No sequence search possible, because only mutated residues (e.g. Y47V) disclosed ?*
- *In which database(s) ?*

Sequence Searches and Databases

Search Platforms:

Patent Sequence Databases:


GQ Life Sciences
An Aptean Company
CAS
A DIVISION OF THE AMERICAN CHEMICAL
STN
THE CHOICE OF PATENT EXPERTS
SequenceBase
NIH
U.S. National Library of Medicine
BLAST
LENS.ORG

- commercial

- GQ-Pat (USPTO, EPO, WIPO, SIPO, GenBank, EMBL, DDBJ, CAS Biosequences)
- CAS-Registry
- USGENE
- PCTGEN / WOGENE
- DGENE / GENESEQ (Clarivate Analytics)
- GENESEQ FASTAlert

- non-commercial

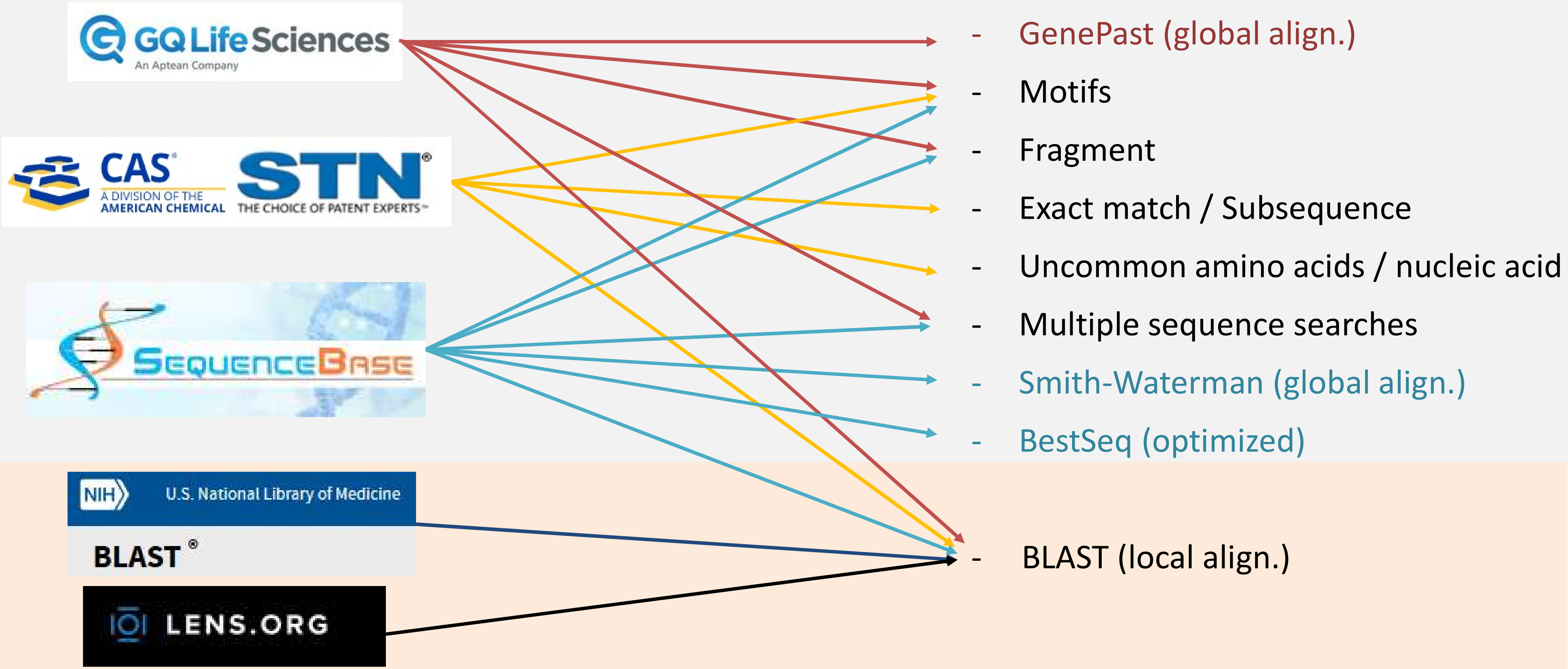
- GenBank
- Pataa (USPTO)
- Lens PatSeq (USPTO, GenBank)

IngridB.Mueller@ipi.ch

Sequence Searches and Databases

Search Platforms:

Used Algorithms for Identity / Similarity


GQ Life Sciences
An Aptean Company
CAS
A DIVISION OF THE AMERICAN CHEMICAL
STN
THE CHOICE OF PATENT EXPERTS
SequenceBase
NIH
U.S. National Library of Medicine
BLAST
LENS.ORG

- GenePast (global align.)
- Motifs
- Fragment
- Exact match / Subsequence
- Uncommon amino acids / nucleic acid
- Multiple sequence searches
- Smith-Waterman (global align.)
- BestSeq (optimized)

- BLAST (local align.)

IngridB.Mueller@ipi.ch

Sequence Searches and Databases

Advantages:

CAS Registry
BLAST / Motif
Exact / Subsequence search
Uncommon aa / nt

powerfull but complex → *presentation Topic 7*
combination with CAS numbers, controlled term
exact or fragment in context (/sqep or /sqsp)
e.g. selenocysteine (U), pyrrolysine (O), labelled aa, nt

IGE | IPI

Sequence Searches and Databases

STN® THE CHOICE OF PATENT EXPERTS™

CAS Registry
BLAST / Motif
Exact / Subsequence search
Uncommon aa / nt

SequenceBase

GENSEQ FASTAlert
BLAST / Motif / **MSS**
Smith-Waterman / BestSeq
Filtering

Advantages:

unique numbers (SBN) easy to handle results
multiple seq search (MSS) for antibody search
adapted algorithms (optimized BLAST)

IGE | IPI

Sequence Searches and Databases

STN THE CHOICE OF PATENT EXPERTS

CAS Registry
BLAST / Motif
Exact / Subsequence search
Uncommon aa / nt

SequenceBase

GENSEQ FASTAlert
BLAST / Motif / **MSS**
Smith-Waterman / BestSeq
Filtering

GQ Life Sciences An Aptean Company

GQ special database + **CAS Registry**
GenePast / Fragment
BLAST / Motif / **MSS**
Filtering, grouping

→

Advantages:

powerfull but less complicated than STN
adapted algorithm (optimized BLAST)
VENN diagramm (good for antibody search)

Sequence Searches and Databases

CAS Registry
BLAST / Motif
Exact / Subsequence search
Uncommon aa / nt

GENSEQ FASTAlert
BLAST / Motif / **MSS**
Smith-Waterman / BestSeq
Filtering

GQ special database + **CAS Registry**
GenePast / Fragment
BLAST / Motif / **MSS**
Filtering, grouping

https://blast.ncbi.nlm.nih.gov/Blast.cgi

Open access

- **NCBI BLAST**

Patent / non-patent documents
(Accession number)
Cross-linked with Lens and Pubmed

- **Lens**

PatSeq finder
Cross-linked with Pubmed and other scholarly works

Sequence Searches and Databases

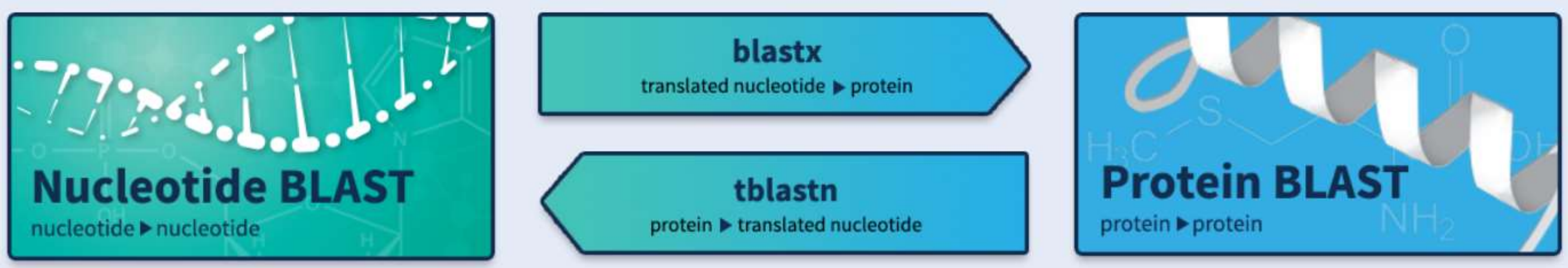
U.S. National Library of Medicine

BLAST®

Open Access Sequence Databases (DB)

NCBI BLAST (Basic Logic Algorithm Sequence Tools):

BlastN	⇨	nucleotide query (nt)	in nucleotide DB
BlastP	⇨	protein query (aa)	in protein DB
TblastN	⇨	translated protein query (aa)	in nucleotide DB
BlastX	⇨	translated nt query (nt)	in protein DB
TBlastX	⇨	translated nt query (nt)	in translated nt DB



NIH U.S. National Library of Medicine

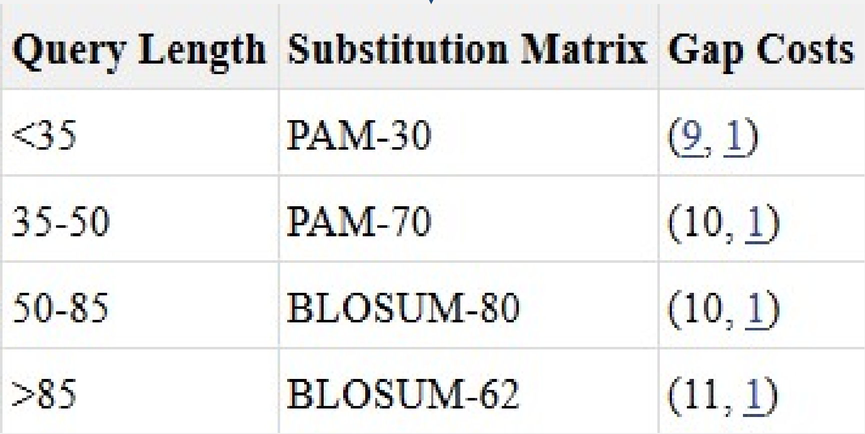
BLAST®

Open Access Sequence Databases (DB)

NCBI BLAST (Basic Logic Algorithm Sequence Tools):

BlastN	⇨	nucleotide query (nt)	in nucleotide DB
BlastP	⇨	protein query (aa)	in protein DB
TblastN	⇨	translated protein query (aa)	in nucleotide DB
BlastX	⇨	translated nt query (nt)	in protein DB
TBlastX	⇨	translated nt query (nt)	in translated nt DB

Use different **substitution matrices** depending on **length of peptide query**

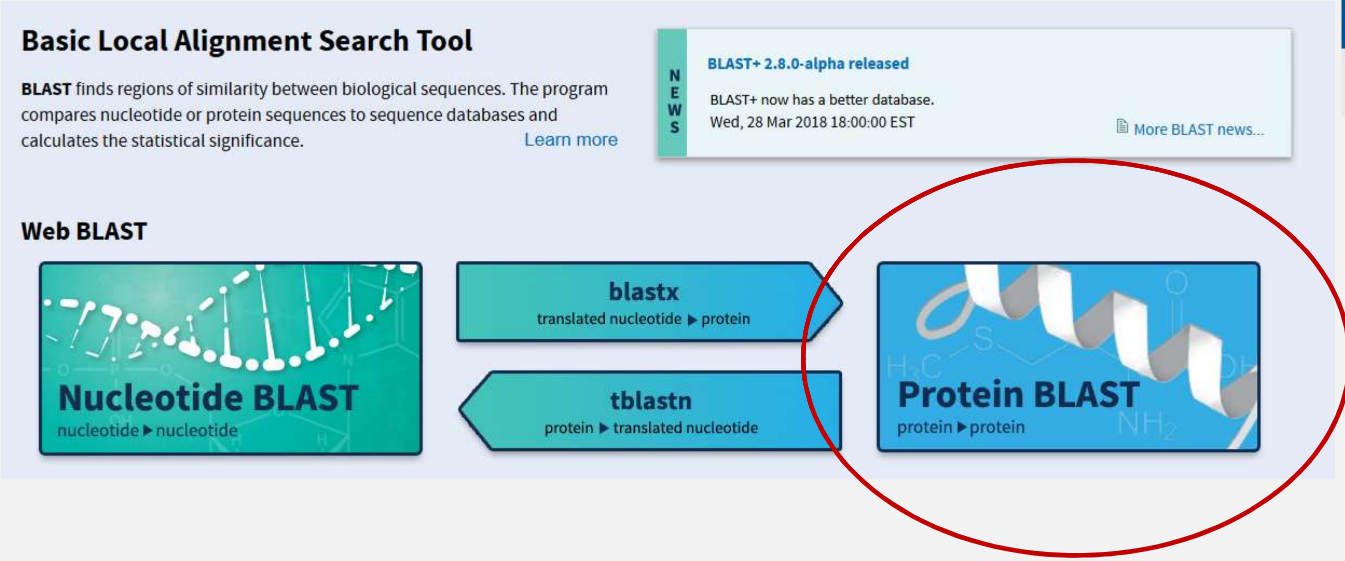


Query Length	Substitution Matrix	Gap Costs
<35	PAM-30	(9, 1)
35-50	PAM-70	(10, 1)
50-85	BLOSUM-80	(10, 1)
>85	BLOSUM-62	(11, 1)

NCBI BLAST has many more BLASTs

e.g. megaBLAST, PSI-BLAST, smartBLAST…

Sequence Searches and Databases



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

BLAST+ 2.8.0-alpha released

BLAST+ now has a better database.
Wed, 28 Mar 2018 18:00:00 EST

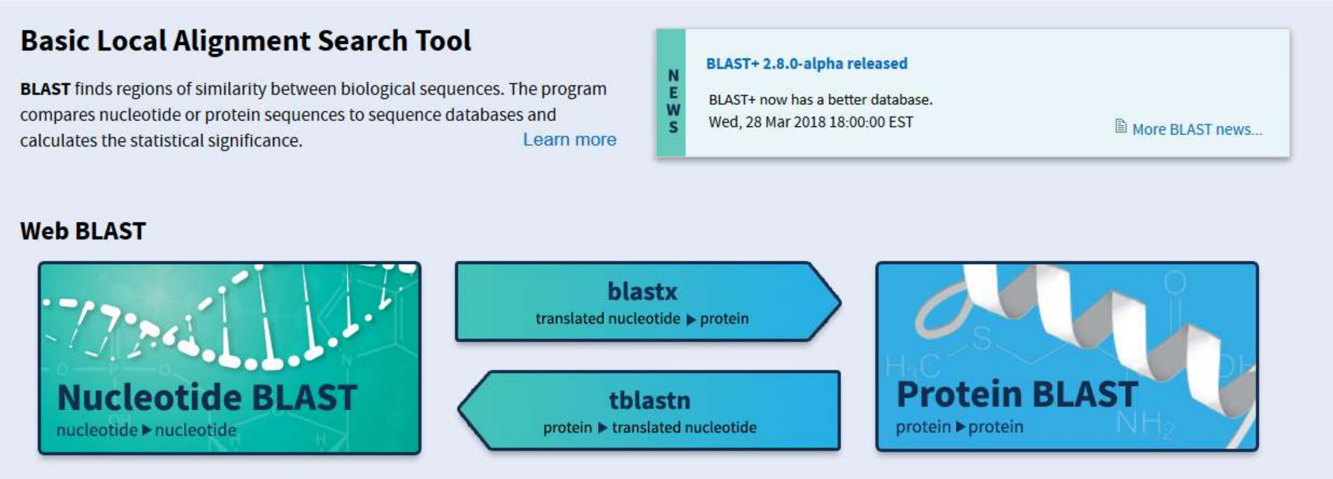
More BLAST news...

Web BLAST

Sequence Searches and Databases

U.S. National Library of Medicine

BLAST®



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

BLAST+ 2.8.0-alpha released

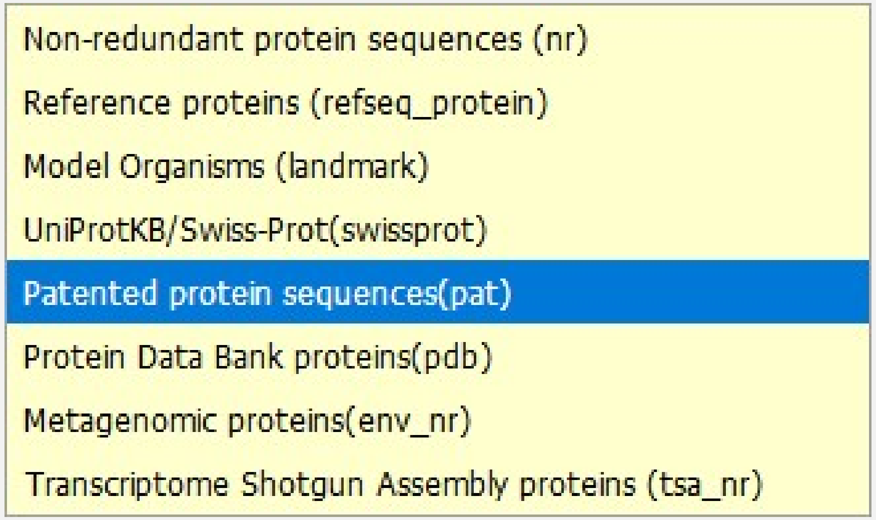
BLAST+ now has a better database.
Wed, 28 Mar 2018 18:00:00 EST

More BLAST news...

Web BLAST

EXAMPLE search for insulin (fragment 1-30 aa)

>ins30
MALWMRLLPLLALLALWGPDPAAAFVNQHL



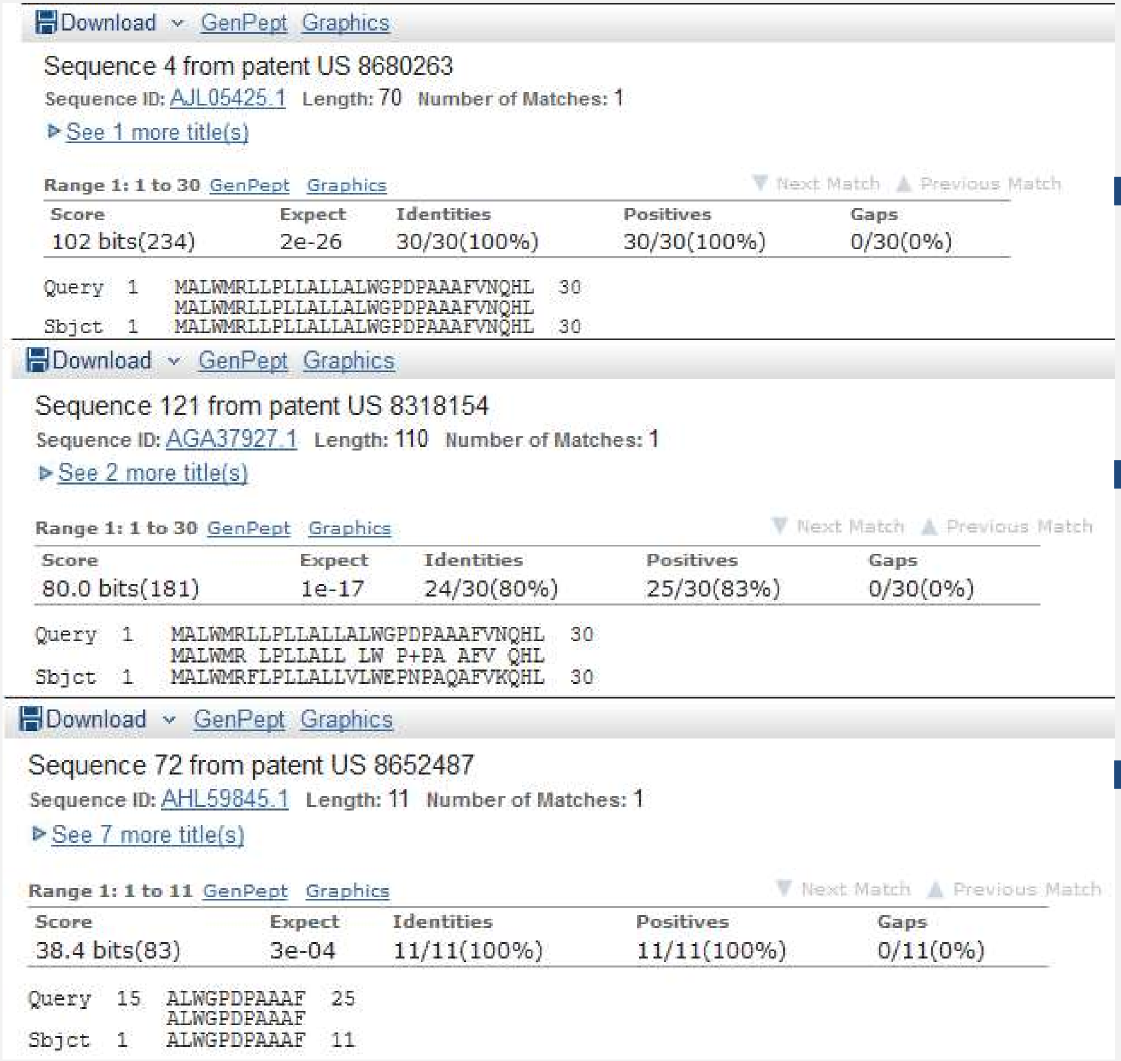
- Non-redundant protein sequences (nr)
- Reference proteins (refseq_protein)
- Model Organisms (landmark)
- UniProtKB/Swiss-Prot(swissprot)
- Patented protein sequences(pat)
- Protein Data Bank proteins(pdb)
- Metagenomic proteins(env_nr)
- Transcriptome Shotgun Assembly proteins (tsa_nr)

https://blast.ncbi.nlm.nih.gov/Blast.cgi

IngridB.Mueller@ipi.ch

IGE | IPI

Sequence Searches and Databases



Download GenPept Graphics

Sequence 4 from patent US 8680263

Sequence ID: AJL05425.1 Length: 70 Number of Matches: 1

▷ See 1 more title(s)

Range 1: 1 to 30 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps
102 bits(234)	2e-26	30/30(100%)	30/30(100%)	0/30(0%)

```
Query  1    MALWMRLLPLLALLALWGPDPAAAFVNQHL  30
            MALWMRLLPLLALLALWGPDPAAAFVNQHL
Sbjct  1    MALWMRLLPLLALLALWGPDPAAAFVNQHL  30
```

Download GenPept Graphics

Sequence 121 from patent US 8318154

Sequence ID: AGA37927.1 Length: 110 Number of Matches: 1

▷ See 2 more title(s)

Range 1: 1 to 30 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps
80.0 bits(181)	1e-17	24/30(80%)	25/30(83%)	0/30(0%)

```
Query  1    MALWMRLLPLLALLALWGPDPAAAFVNQHL  30
            MALWMR LPLLALL LW P+PA AFV QHL
Sbjct  1    MALWMRFLPLLALLVLWEPNPAQAFVKQHL  30
```

Download GenPept Graphics

Sequence 72 from patent US 8652487

Sequence ID: AHL59845.1 Length: 11 Number of Matches: 1

▷ See 7 more title(s)

Range 1: 1 to 11 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps
38.4 bits(83)	3e-04	11/11(100%)	11/11(100%)	0/11(0%)

```
Query  15   ALWGPDPAAAF  25
            ALWGPDPAAAF
Sbjct  1    ALWGPDPAAAF  11
```

NIH U.S. National Library of Medicine

BLAST®

US8680263 → SEQ ID No. 4
Identity 100 % E-value 2e-26
30 of 30 aa

US8318154 → SEQ ID No. 121
Identity 80 % E-value 1e-17
25 of 30 aa

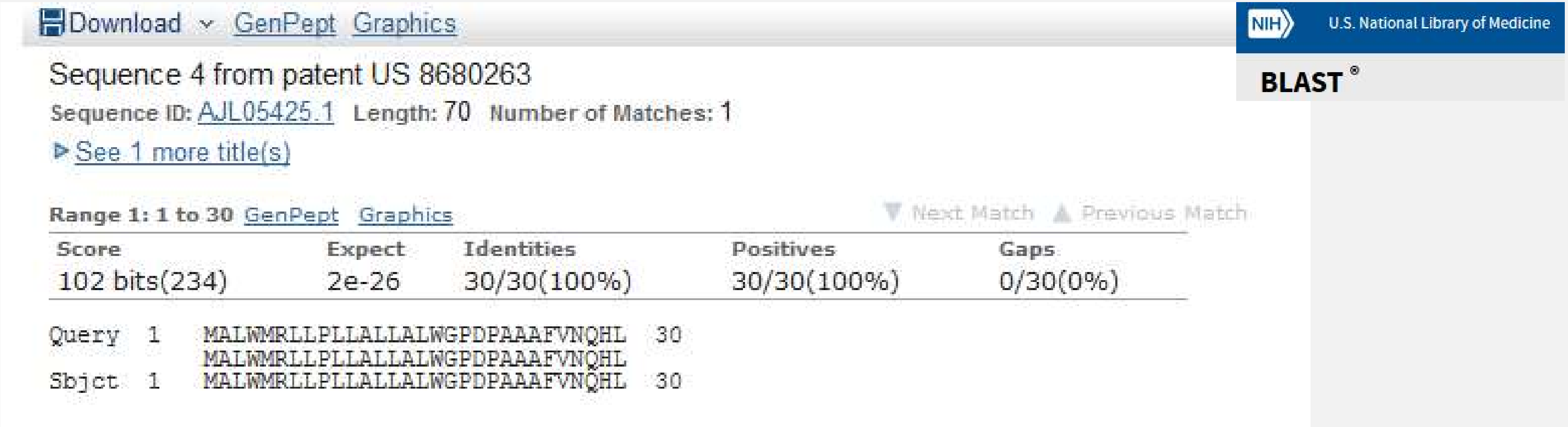
US8652487 → SEQ ID No. 72
Identity 100% E-value 3e-04
11 of 11 aa

! Attention: alignment ID 100%, not query ID

Different E-values, the smaller the better !

IngridB.Mueller@ipi.ch

Sequence Searches and Databases



NIH
U.S. National Library of Medicine
BLAST ®
Download
GenPept
Graphics

Sequence 4 from patent US 8680263

Sequence ID: AJL05425.1 Length: 70 Number of Matches: 1

▶ See 1 more title(s)

Range 1: 1 to 30 GenPept Graphics

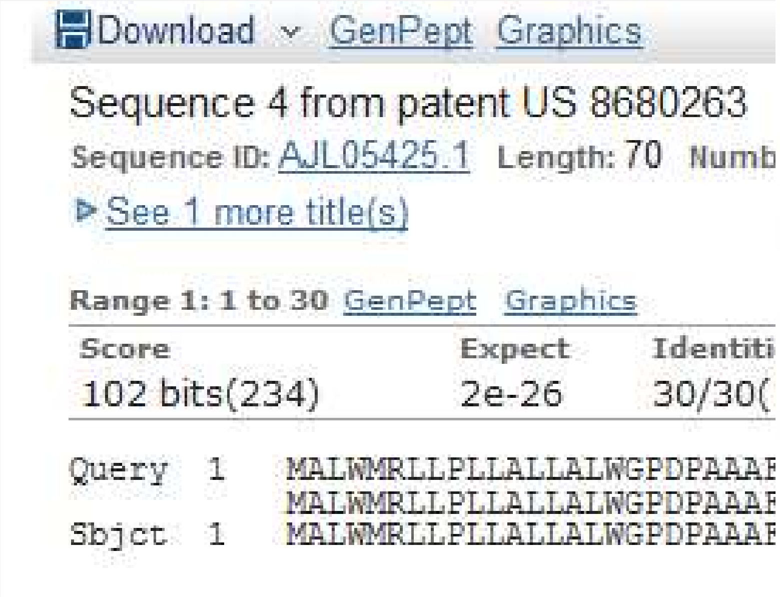
▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps
102 bits(234)	2e-26	30/30(100%)	30/30(100%)	0/30(0%)

```
Query  1    MALWMRLLPLLALLALWGPDPAAAFVNQHL  30
            MALWMRLLPLLALLALWGPDPAAAFVNQHL
Sbjct  1    MALWMRLLPLLALLALWGPDPAAAFVNQHL  30
```

IGE | IPI

Sequence Searches and Databases



Sequence 4 from patent US 8680263

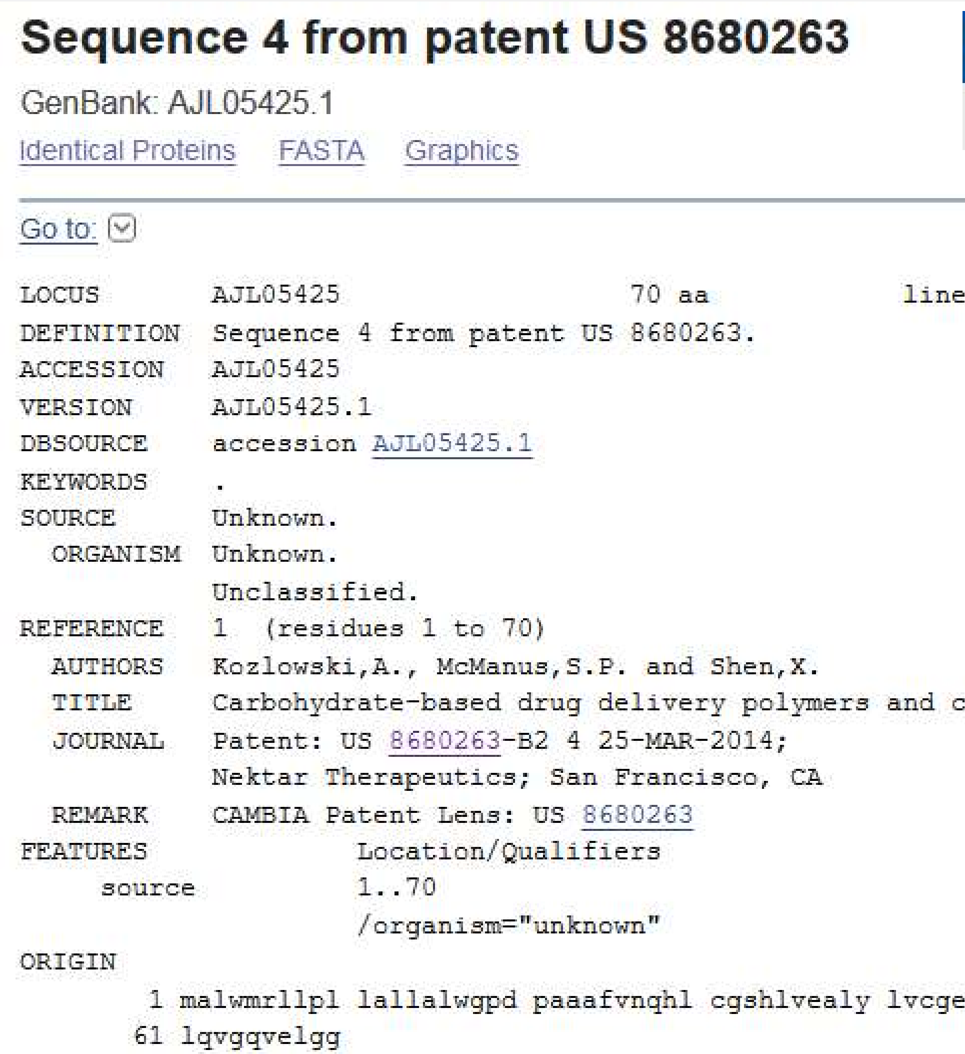
Sequence ID: AJL05425.1 Length: 70 Numb

▶ See 1 more title(s)

Range 1: 1 to 30 GenPept Graphics

Score	Expect	Identiti
102 bits(234)	2e-26	30/30(

```
Query  1    MALWMRLLPLLALLALWGPDPAAAF
            MALWMRLLPLLALLALWGPDPAAAF
Sbjct  1    MALWMRLLPLLALLALWGPDPAAAF
```



NIH U.S. National Library of Medicine

BLAST ®

Sequence 4 from patent US 8680263

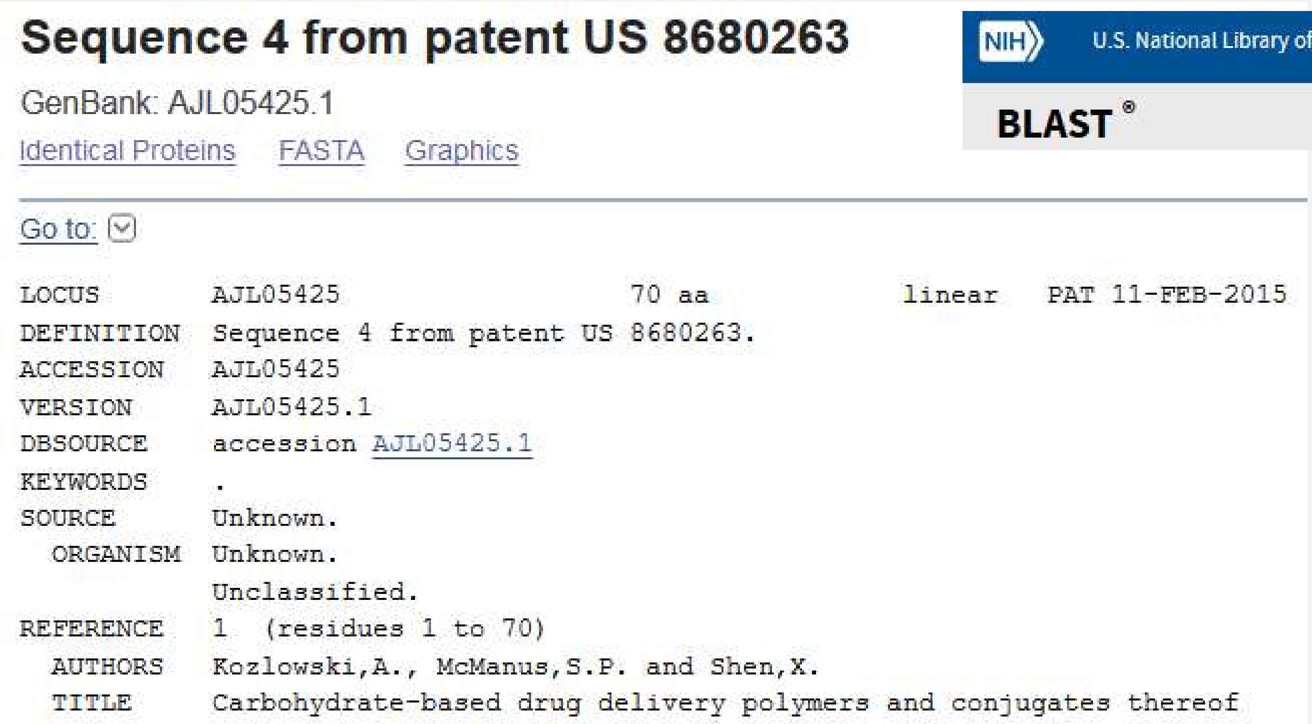
GenBank: AJL05425.1

```
LOCUS       AJL05425                  70 aa            linear   PAT 11-FEB-2015
DEFINITION  Sequence 4 from patent US 8680263.
ACCESSION   AJL05425
VERSION     AJL05425.1
DBSOURCE    accession AJL05425.1
KEYWORDS    .
SOURCE      Unknown.
  ORGANISM  Unknown.
            Unclassified.
REFERENCE   1  (residues 1 to 70)
  AUTHORS   Kozlowski,A., McManus,S.P. and Shen,X.
  TITLE     Carbohydrate-based drug delivery polymers and conjugates thereof
  JOURNAL   Patent: US 8680263-B2 4 25-MAR-2014;
            Nektar Therapeutics; San Francisco, CA
  REMARK    CAMBIA Patent Lens: US 8680263
FEATURES             Location/Qualifiers
     source          1..70
                     /organism="unknown"
ORIGIN
        1 malwmrllpl lallalwgpd paaafvnqhl cgshlvealy lvcgergffy tpktrreaed
       61 lqvgqvelgg
```

IGE | IPI

Sequence Searches and Databases

Repeat BLAST search with AJL05425.1
In **non-redundant GenBank** (NPL)



Sequence 4 from patent US 8680263

GenBank: AJL05425.1

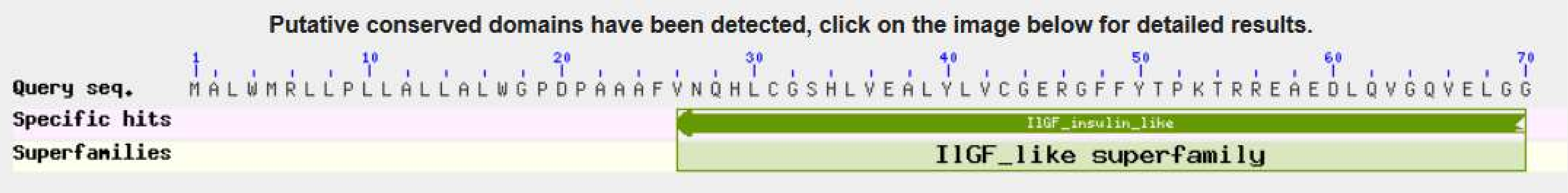
Identical Proteins FASTA Graphics

Go to:

```
LOCUS       AJL05425                70 aa            linear   PAT 11-FEB-2015
DEFINITION  Sequence 4 from patent US 8680263.
ACCESSION   AJL05425
VERSION     AJL05425.1
DBSOURCE    accession AJL05425.1
KEYWORDS    .
SOURCE      Unknown.
  ORGANISM  Unknown.
            Unclassified.
REFERENCE   1  (residues 1 to 70)
  AUTHORS   Kozlowski,A., McManus,S.P. and Shen,X.
  TITLE     Carbohydrate-based drug delivery polymers and conjugates thereof
```

NIH U.S. National Library of Medicine

BLAST®



Putative conserved domains have been detected, click on the image below for detailed results.

1
10
20
30
40
50
60
70
Query seq.
M A L W M R L L P L L A L L A L W G P D P A A A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T P K T R R E A E D L Q V G Q V E L G G
Specific hits
IIGF_insulin_like
Superfamilies
IIGF_like superfamily

Sequence Searches and Databases

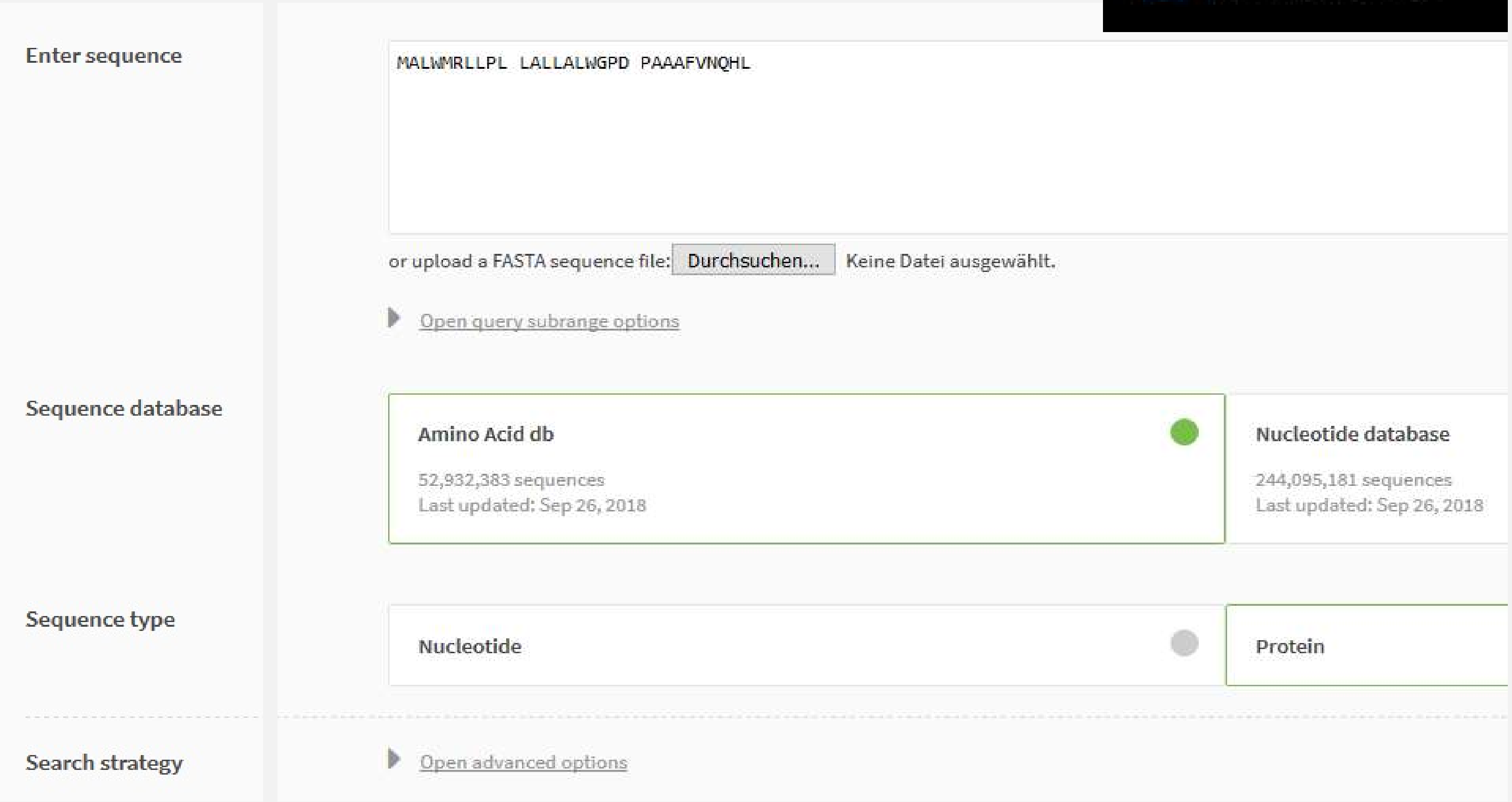
Open Access Sequence Databases (DB)

PatSeq Finder

LENS.ORG

EXAMPLE

>ins30
MALWMRLLPLLALLALWGPDPA
AAFVNQHL



Enter sequence

```
MALWMRLLPL LALLALWGPD PAAAFVNQHL
```

or upload a FASTA sequence file: Durchsuchen... Keine Datei ausgewählt.

Sequence database

Amino Acid db ●
52,932,383 sequences
Last updated: Sep 26, 2018

Nucleotide database
244,095,181 sequences
Last updated: Sep 26, 2018

Sequence type

Nucleotide ○

Protein

Search strategy

https://www.lens.org/lens/bio/patseqfinder

IngridB.Mueller@ipi.ch

IGE | IPI

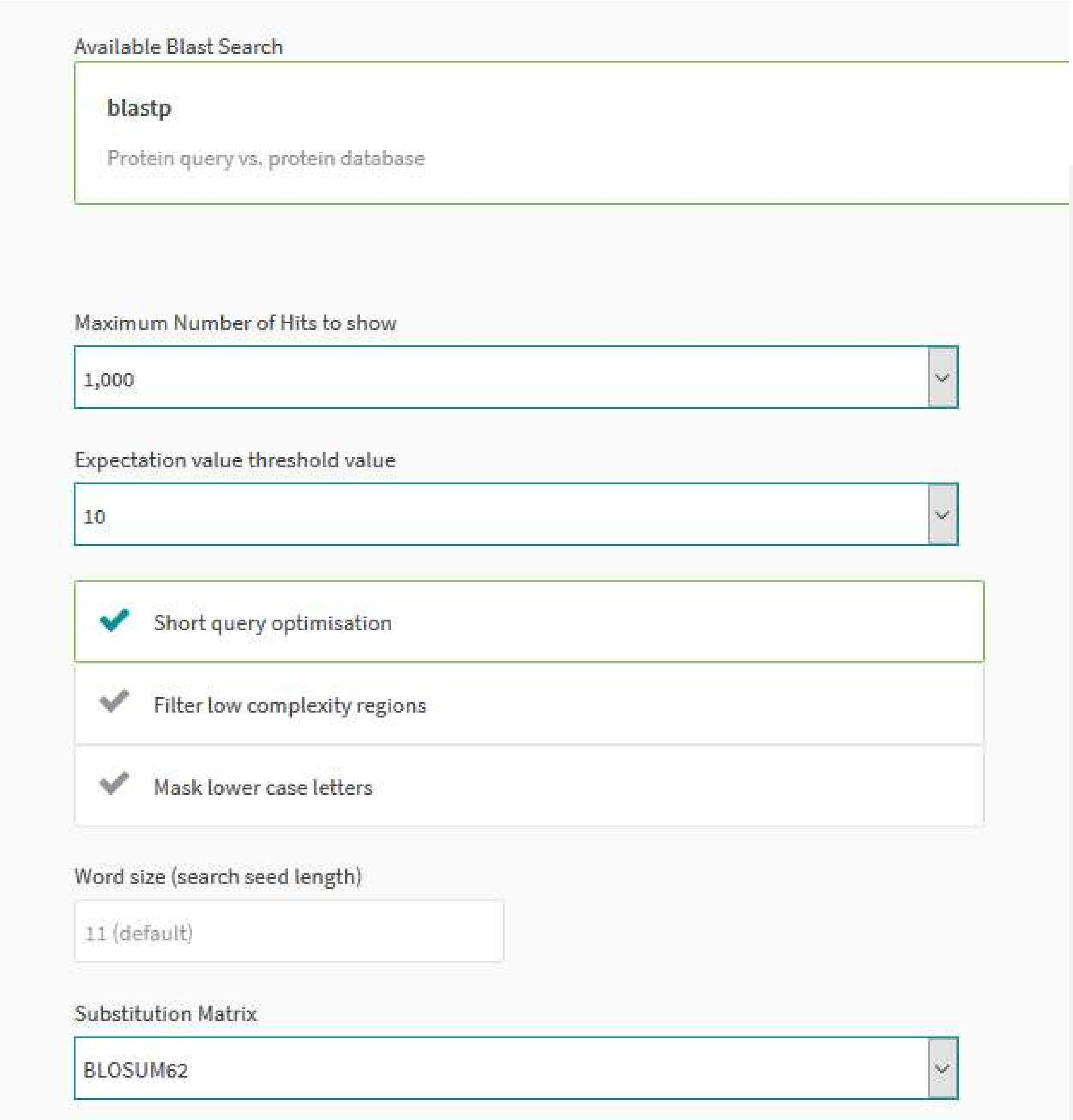
Open Access Sequence Databases (DB)

EXAMPLE

>ins30
MALWMRLLPLLALLALWGPDPAAAFVNQHL

Substitution Matrix: **BLOSUM62**

https://www.lens.org/lens/bio/patseqfinder



Available Blast Search

blastp

Protein query vs. protein database

Maximum Number of Hits to show

1,000

Expectation value threshold value

10

☑ Short query optimisation

☐ Filter low complexity regions

☐ Mask lower case letters

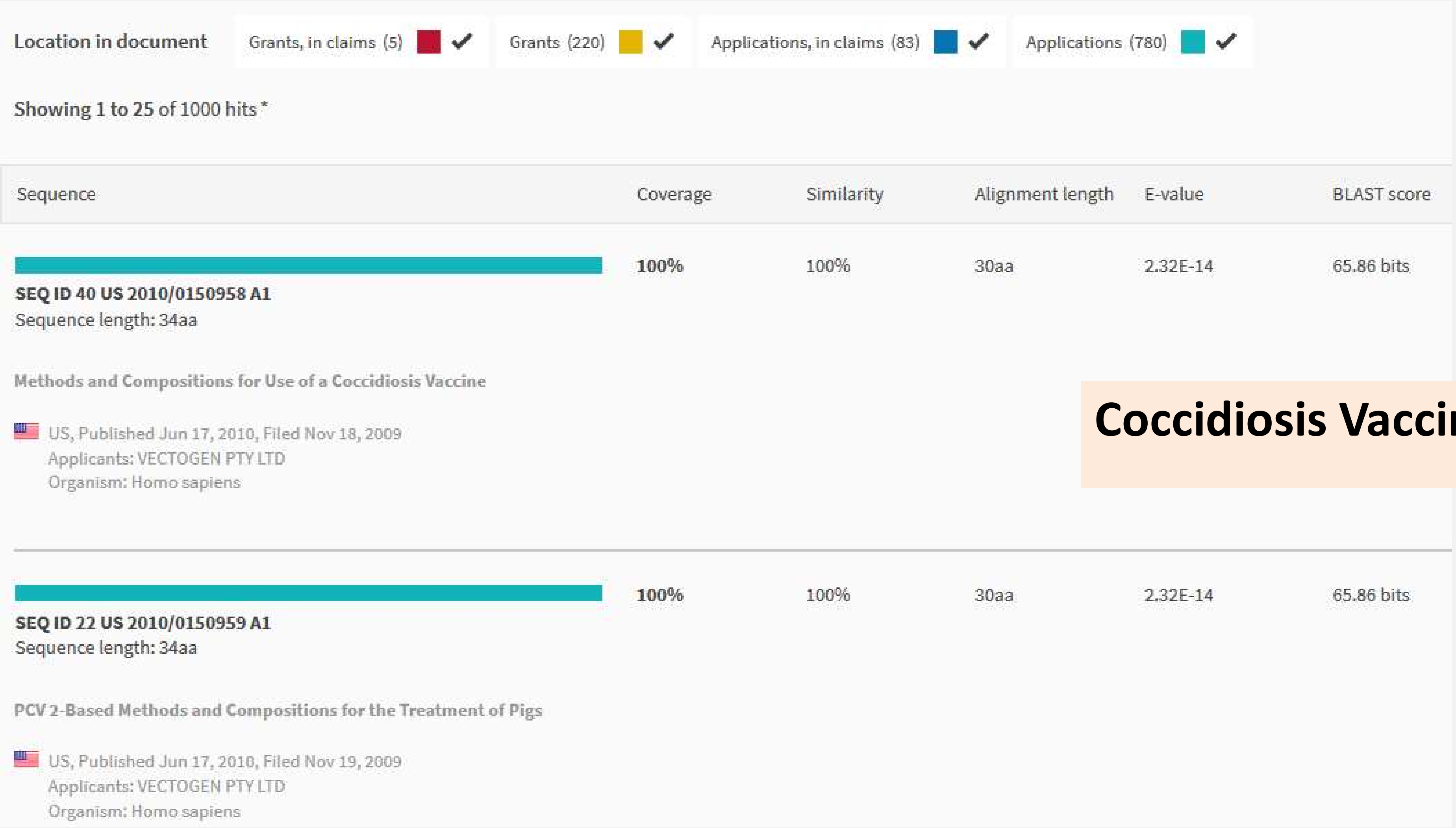
Word size (search seed length)

11 (default)

Substitution Matrix

BLOSUM62

Sequence Searches and Databases



LENS.ORG

Location in document Grants, in claims (5) 🟥 ✓ Grants (220) 🟨 ✓ Applications, in claims (83) 🟦 ✓ Applications (780) 🟦 ✓

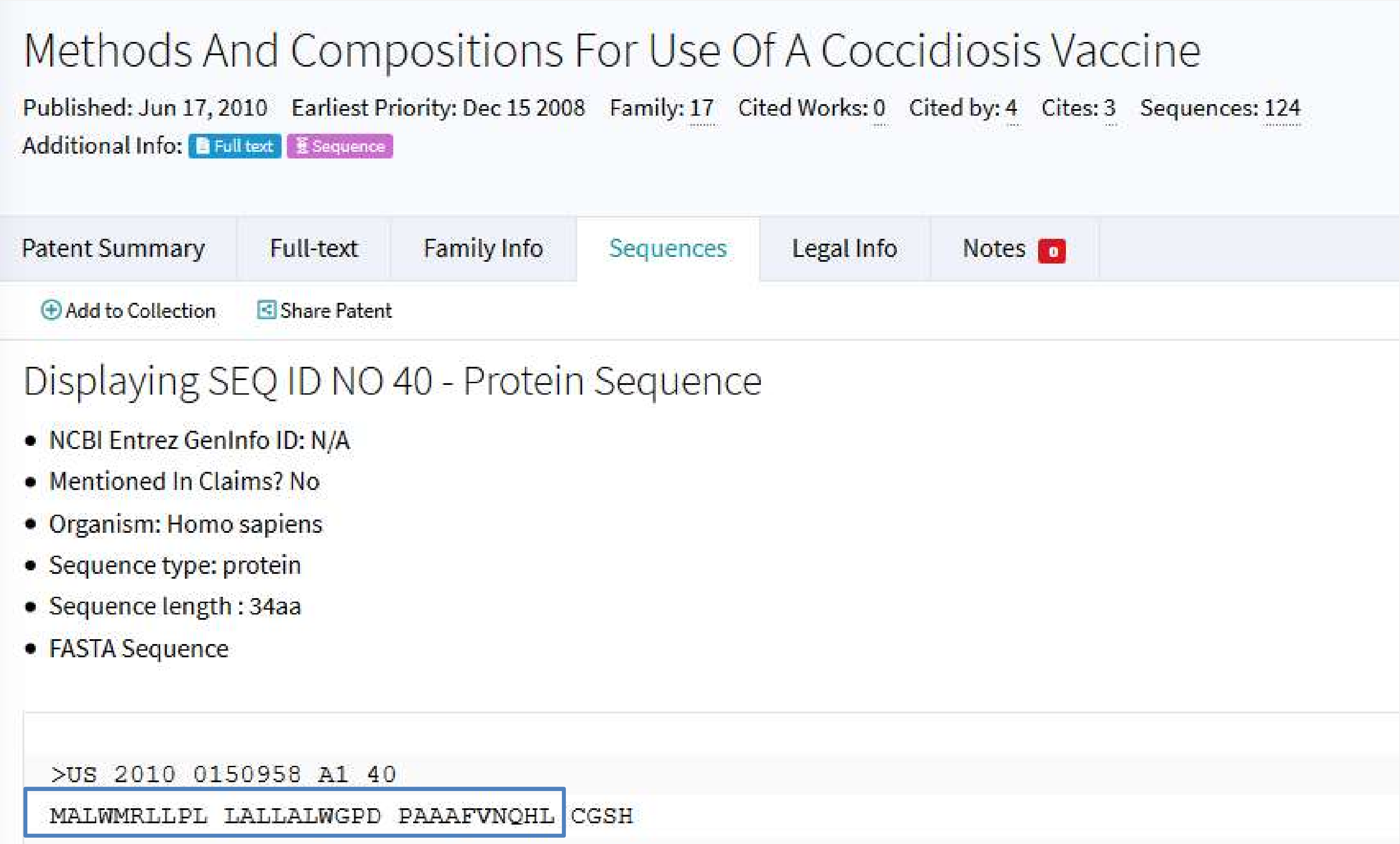
Showing 1 to 25 of 1000 hits *

Sequence	Coverage	Similarity	Alignment length	E-value	BLAST score
SEQ ID 40 US 2010/0150958 A1 Sequence length: 34aa Methods and Compositions for Use of a Coccidiosis Vaccine US, Published Jun 17, 2010, Filed Nov 18, 2009 Applicants: VECTOGEN PTY LTD Organism: Homo sapiens	**100%**	100%	30aa	2.32E-14	65.86 bits
SEQ ID 22 US 2010/0150959 A1 Sequence length: 34aa PCV 2-Based Methods and Compositions for the Treatment of Pigs US, Published Jun 17, 2010, Filed Nov 19, 2009 Applicants: VECTOGEN PTY LTD Organism: Homo sapiens	**100%**	100%	30aa	2.32E-14	65.86 bits

Coccidiosis Vaccine

Sequence Searches and Databases

LENS.ORG



Methods And Compositions For Use Of A Coccidiosis Vaccine

Published: Jun 17, 2010 Earliest Priority: Dec 15 2008 Family: 17 Cited Works: 0 Cited by: 4 Cites: 3 Sequences: 124

Additional Info: Full text Sequence

Displaying SEQ ID NO 40 - Protein Sequence

- NCBI Entrez GenInfo ID: N/A
- Mentioned In Claims? No
- Organism: Homo sapiens
- Sequence type: protein
- Sequence length : 34aa
- FASTA Sequence

```
>US 2010 0150958 A1 40
MALWMRLLPL LALLALWGPD PAAAFVNQHL CGSH
```

IngridB.Mueller@ipi.ch

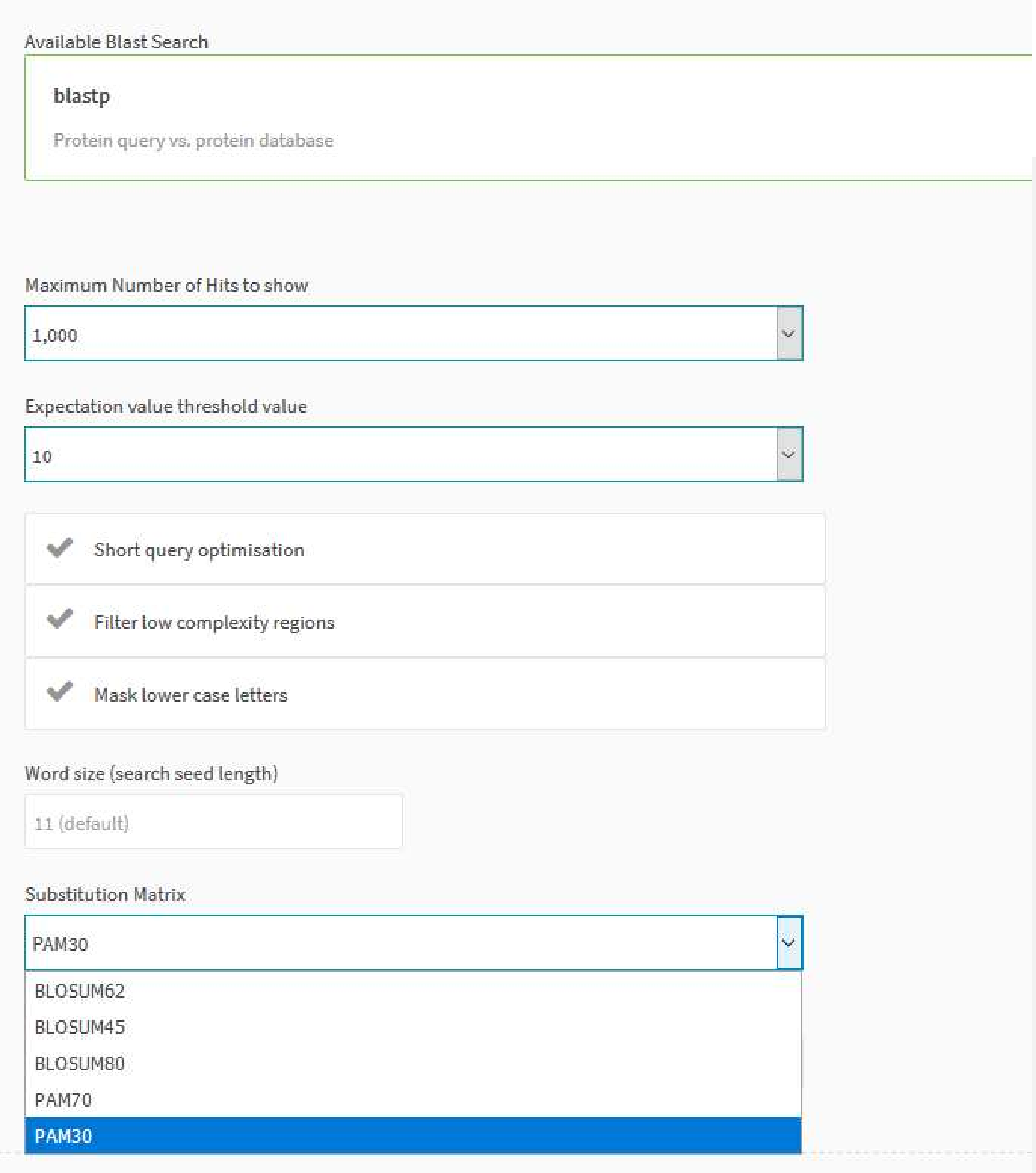
IGE | IPI

Open Access Sequence Databases (DB)

EXAMPLE

>ins30
MALWMRLLPLLALLALWGPDPAAAFVNQHL

Substitution Matrix: **PAM30**



Available Blast Search

blastp

Protein query vs. protein database

Maximum Number of Hits to show: 1,000

Expectation value threshold value: 10

✓ Short query optimisation

✓ Filter low complexity regions

✓ Mask lower case letters

Word size (search seed length): 11 (default)

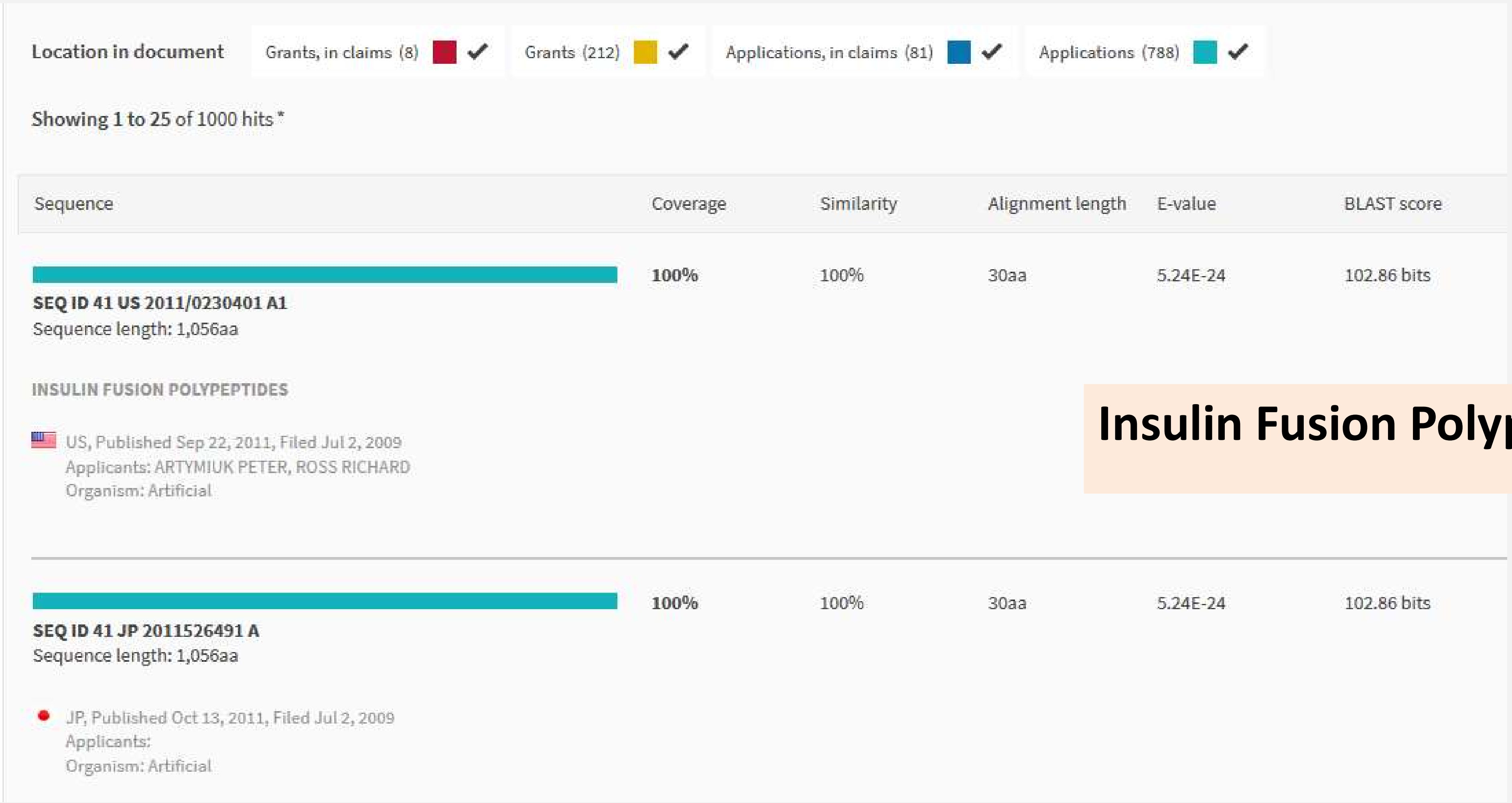
Substitution Matrix: PAM30

LENS.ORG

https://www.lens.org/lens/bio/patseqfinder

Sequence Searches and Databases

LENS.ORG



Location in document Grants, in claims (8) 🟥 ✓ Grants (212) 🟨 ✓ Applications, in claims (81) 🟦 ✓ Applications (788) 🟦 ✓

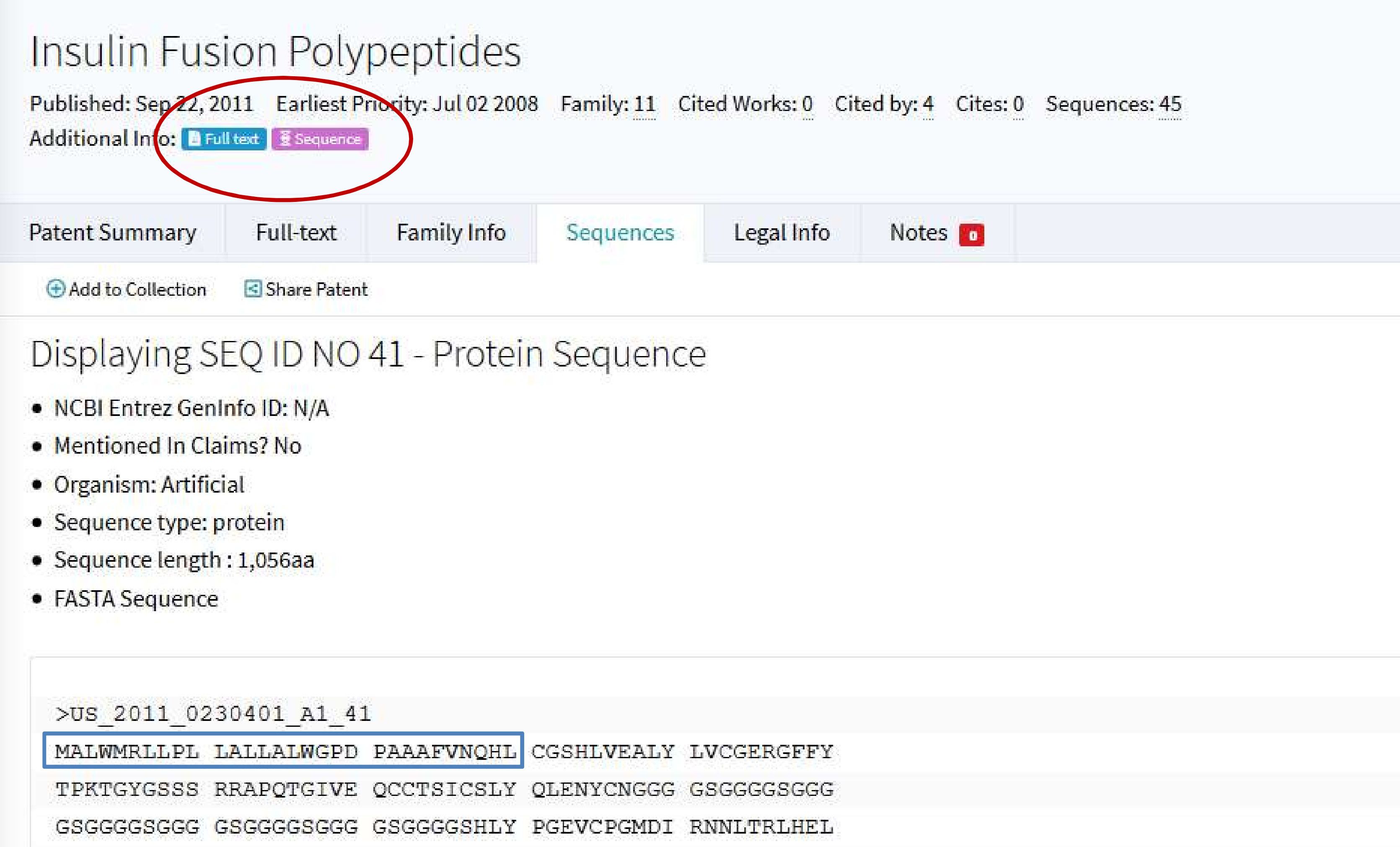
Showing 1 to 25 of 1000 hits *

Sequence	Coverage	Similarity	Alignment length	E-value	BLAST score
SEQ ID 41 US 2011/0230401 A1 Sequence length: 1,056aa INSULIN FUSION POLYPEPTIDES US, Published Sep 22, 2011, Filed Jul 2, 2009 Applicants: ARTYMIUK PETER, ROSS RICHARD Organism: Artificial	**100%**	100%	30aa	5.24E-24	102.86 bits
SEQ ID 41 JP 2011526491 A Sequence length: 1,056aa ● JP, Published Oct 13, 2011, Filed Jul 2, 2009 Applicants: Organism: Artificial	**100%**	100%	30aa	5.24E-24	102.86 bits

Insulin Fusion Polypetides

Sequence Searches and Databases

LENS.ORG



Insulin Fusion Polypeptides

Published: Sep 22, 2011 Earliest Priority: Jul 02 2008 Family: 11 Cited Works: 0 Cited by: 4 Cites: 0 Sequences: 45

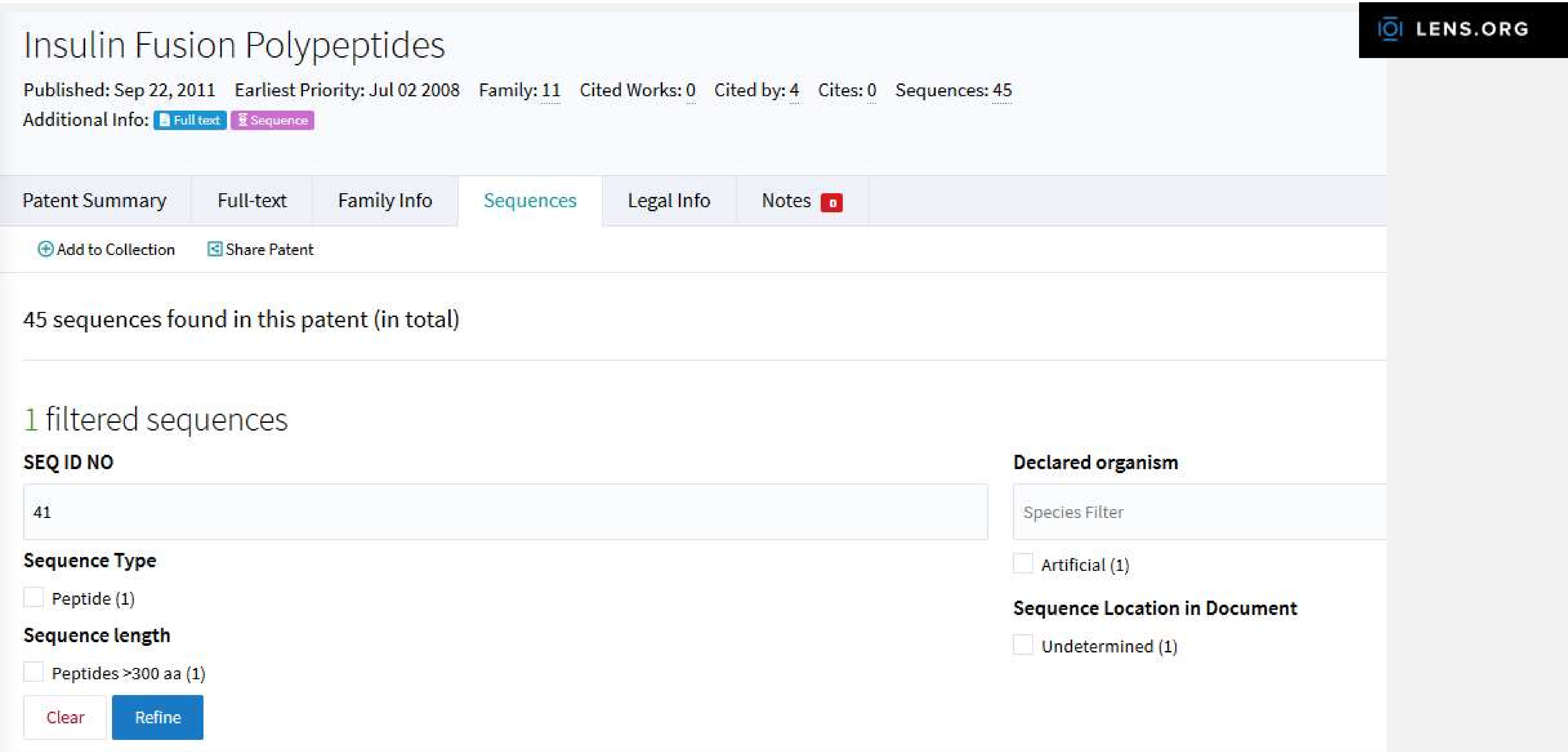
Additional Info:

Displaying SEQ ID NO 41 - Protein Sequence

- NCBI Entrez GenInfo ID: N/A
- Mentioned In Claims? No
- Organism: Artificial
- Sequence type: protein
- Sequence length : 1,056aa
- FASTA Sequence

```
>US_2011_0230401_A1_41
MALWMRLLPL LALLALWGPD PAAAFVNQHL CGSHLVEALY LVCGERGFFY
TPKTGYGSSS RRAPQTGIVE QCCTSICSLY QLENYCNGGG GSGGGGSGGG
GSGGGGSGGG GSGGGGSGGG GSGGGGSHLY PGEVCPGMDI RNNLTRLHEL
```

Sequence Searches and Databases



LENS.ORG

Insulin Fusion Polypeptides

Published: Sep 22, 2011 Earliest Priority: Jul 02 2008 Family: 11 Cited Works: 0 Cited by: 4 Cites: 0 Sequences: 45

Additional Info: Full text Sequence

45 sequences found in this patent (in total)

1 filtered sequences

SEQ ID NO

41

Sequence Type

□ Peptide (1)

Sequence length

□ Peptides >300 aa (1)

Declared organism

Species Filter

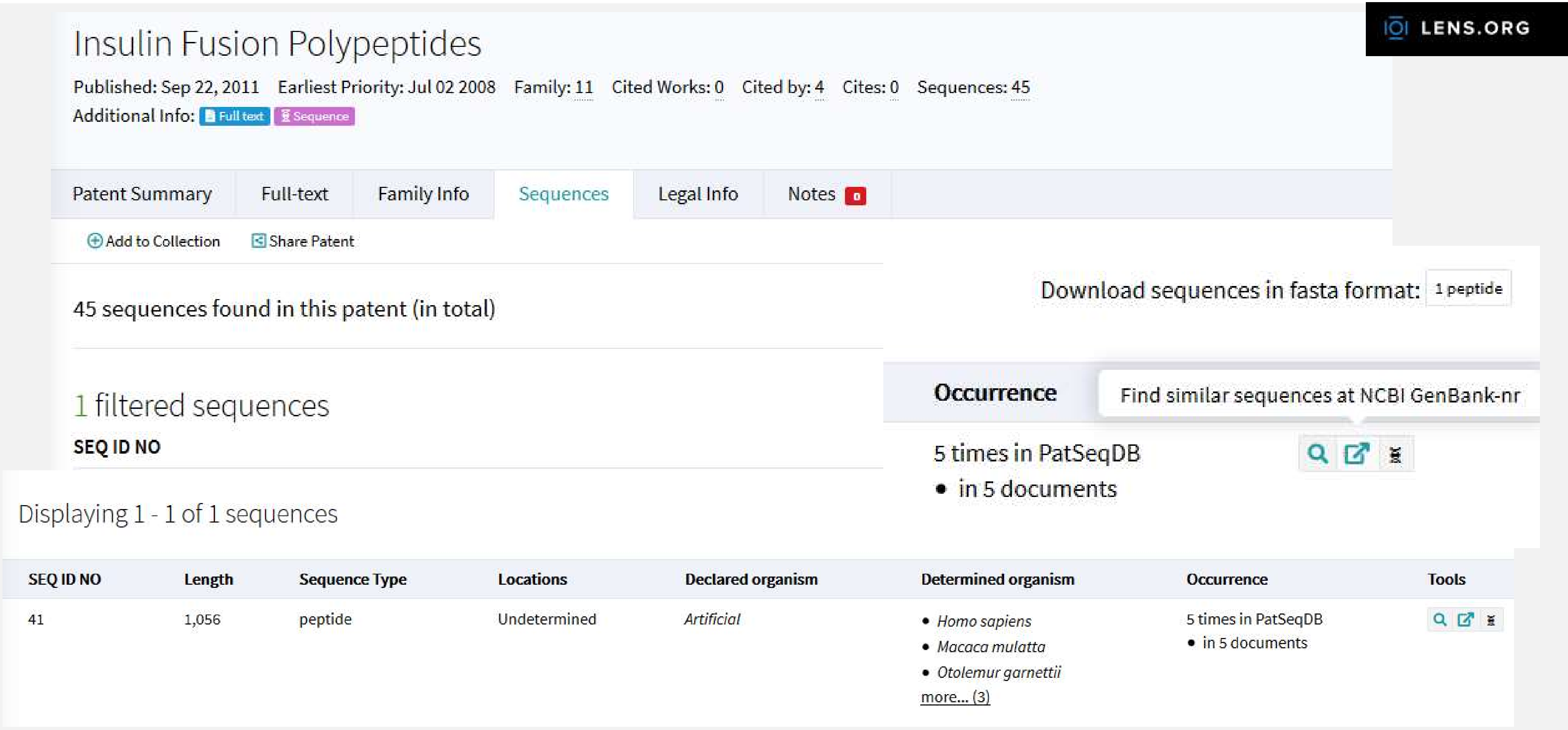
□ Artificial (1)

Sequence Location in Document

□ Undetermined (1)

IGE | IPI

Sequence Searches and Databases



LENS.ORG

Insulin Fusion Polypeptides

Published: Sep 22, 2011 Earliest Priority: Jul 02 2008 Family: 11 Cited Works: 0 Cited by: 4 Cites: 0 Sequences: 45

Additional Info: Full text Sequence

45 sequences found in this patent (in total)

Download sequences in fasta format: 1 peptide

1 filtered sequences

SEQ ID NO

Occurrence

Find similar sequences at NCBI GenBank-nr

5 times in PatSeqDB

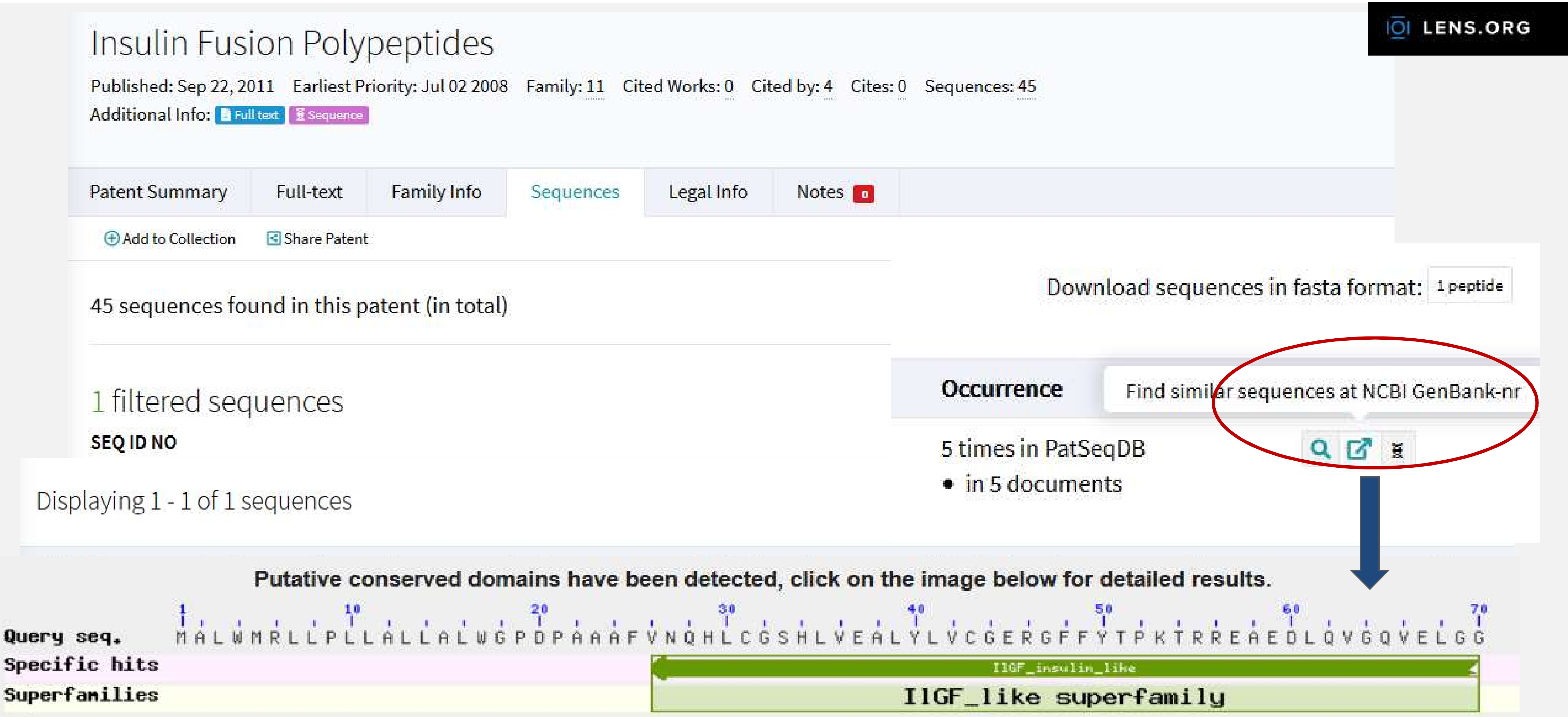
- in 5 documents

Displaying 1 - 1 of 1 sequences

SEQ ID NO	Length	Sequence Type	Locations	Declared organism	Determined organism	Occurrence	Tools
41	1,056	peptide	Undetermined	*Artificial*	• *Homo sapiens* • *Macaca mulatta* • *Otolemur garnettii* more... (3)	5 times in PatSeqDB • in 5 documents	

IngridB.Mueller@ipi.ch

Sequence Searches and Databases



LENS.ORG

Insulin Fusion Polypeptides

Published: Sep 22, 2011 Earliest Priority: Jul 02 2008 Family: 11 Cited Works: 0 Cited by: 4 Cites: 0 Sequences: 45

Additional Info: Full text Sequence

45 sequences found in this patent (in total)

Download sequences in fasta format: 1 peptide

1 filtered sequences

SEQ ID NO

Displaying 1 - 1 of 1 sequences

Occurrence

Find similar sequences at NCBI GenBank-nr

5 times in PatSeqDB

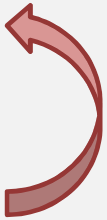
- in 5 documents

Putative conserved domains have been detected, click on the image below for detailed results.

Query seq. M A L W M R L L P L L A L L A L W G P D P A A A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T P K T R R E A E D L Q V G Q V E L G G

Advice 1: remember how to search…

1) Understanding
2) Search tools
3) Search strategy
4) Databases



adapt and repeat if necessary

Advice 2: … when searching for peptides chose the right parameters

- optimize your search by using **different algorithms, substitution matrix**
- **exact search vs similarity search**
- check the **application / use** of the peptide sequence

EXERCISE 1

Independent claims:

Claim 1: A peptide capable of inhibiting the binding of trimeric influenza hemagglutinin protein to its receptor, wherein the peptide is consisting of the sequence SEQ ID NO: 1

Claim 4: A pharmaceutical composition comprising the peptide according to claim 1

Claim 6: A peptide according to claim 1 for use in treating an influenza virus infection

Claim 8: A method for the preparation of a peptide according to claim 1

SEQ ID NO: 1: Pro Tyr Asp Val Pro Asp Tyr Glu

1) Transform the above listed sequence in single letter code
2) Search the sequence using
   - BLAST (NCBI)
   - Lens.org

Hint: Definition of Terms

Comprising: **inclusive**, open ended
(e.g. " a polypeptide comprising SEQ ID NO: 1")

Consisting of: **exclusive**
(e.g. "a peptide consisting of SEQ ID NO: 1")

https://blast.ncbi.nlm.nih.gov/Blast.cgi https://www.lens.org/lens/bio/patseqfinder

IngridB.Mueller@ipi.ch

Sequence Searches and Databases

EXERCISE 2

Claim 1. A polypeptide having antimicrobial activity selected from the group consisting of:

(a) a polypeptide comprising an amino acid sequence which has at least 65% identity with amino acids 1 to 40 of SEQ ID NO:2;

(b) a polypeptide which is encoded by a nucleotide sequence which hybridizes under medium stringency conditions using 0.2 x SSC at 42°C for washing with a polynucleotide probe selected from the group consisting of:
(i) the complementary strand of nucleotides 166 to 285 of SEQ ID NO:1,
(ii) the complementary strand of nucleotides 70 to 285 of SEQ ID NO:1 and
(iii) the complementary strand of nucleotides 1 to 285 of SEQ ID NO:1; and

(c) a fragment of (a) or (b) that has antimicrobial activity.

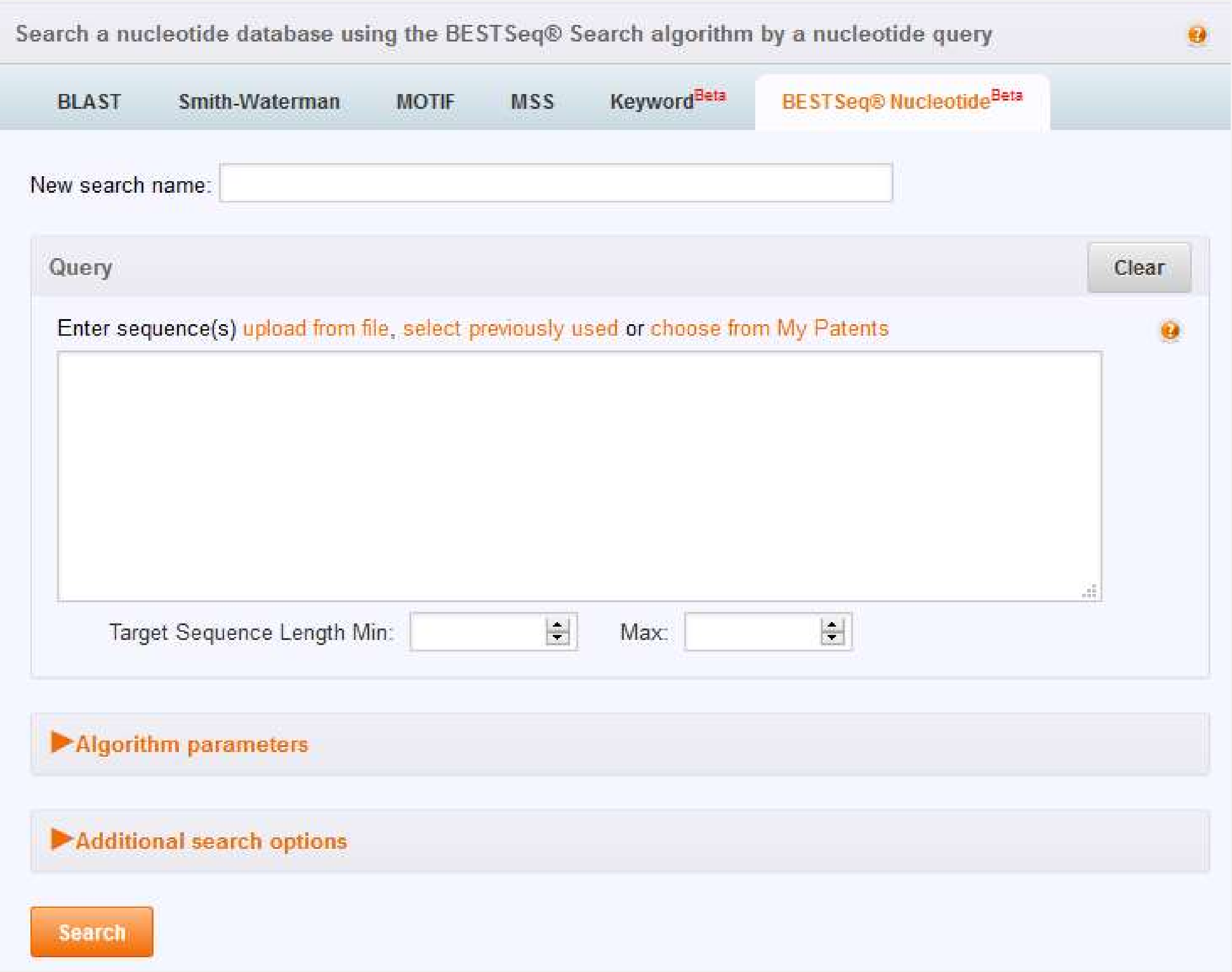
1) Get the sequences from the family member: US2005124064 using Lens.org

1) Search the claimed sequences in
   - ○ BLAST (NCBI)
   - ○ Lens.org

3) Find relevant documents…

https://blast.ncbi.nlm.nih.gov/Blast.cgi https://www.lens.org/lens/bio/patseqfinder

Sequence Searches and Databases



Search a nucleotide database using the BESTSeq® Search algorithm by a nucleotide query

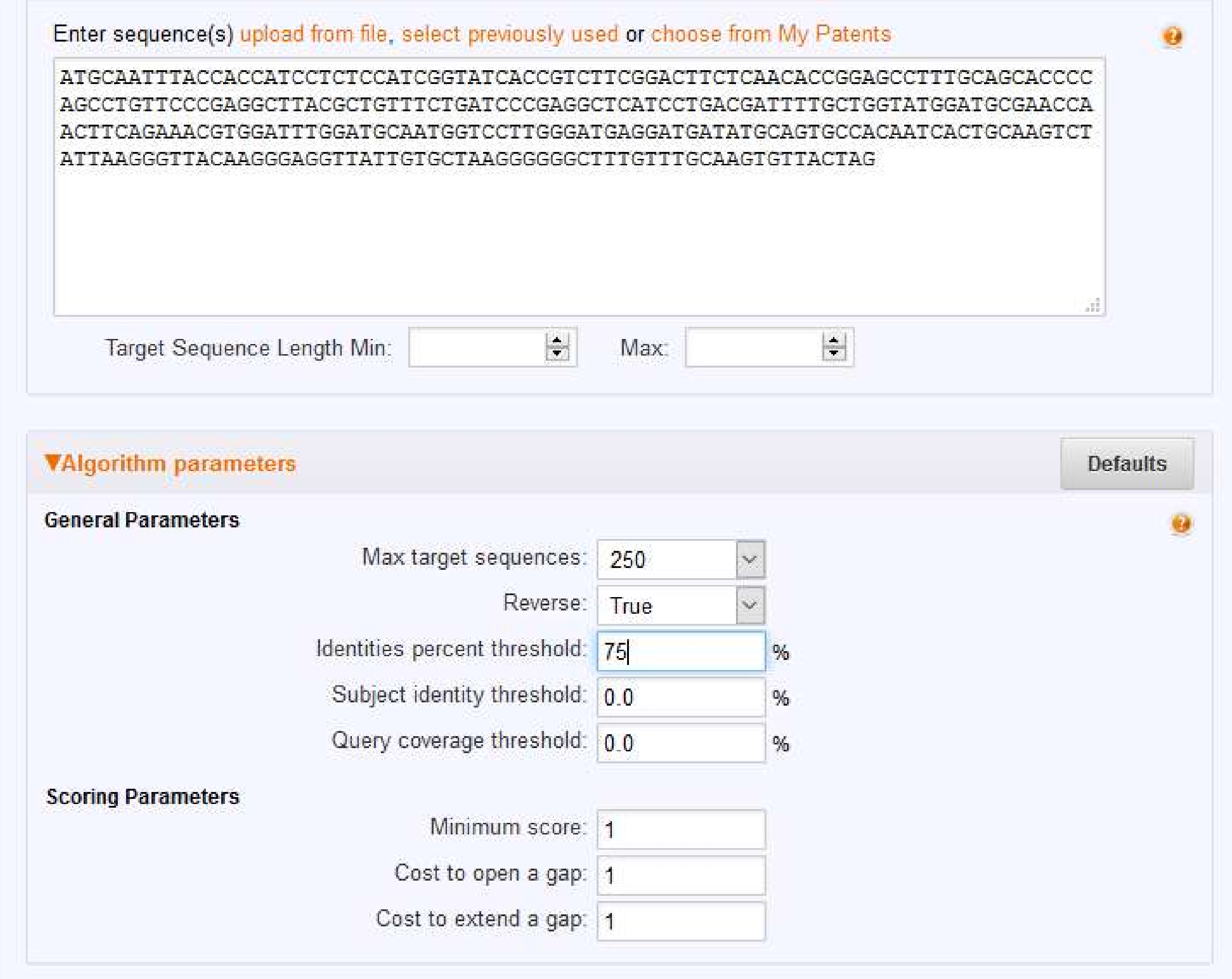
BLAST
Smith-Waterman
MOTIF
MSS
KeywordBeta
BESTSeq® NucleotideBeta
New search name:
Query
Clear
Enter sequence(s) upload from file, select previously used or choose from My Patents
Target Sequence Length Min:
Max:
▶Algorithm parameters
▶Additional search options
Search

EXAMPLE: search SEQ ID NO:1 using BESTSeq of SequenceBase

US2005124064 A1
Applicant: **Novozymes**
Priority: 20 Nov 2001

```
>US_2005_0124064_A1_1
ATGCAATTTA CCACCATCCT CTCCATCGGT ATCACCGTCT
TCGGACTTCT CAACACCGGA GCCTTTGCAG CACCCCAGCC
TGTTCCCGAG GCTTACGCTG TTTCTGATCC CGAGGCTCAT
CCTGACGATT TTGCTGGTAT GGATGCGAAC CAACTTCAGA
AACGTGGATT TGGATGCAAT GGTCCTTGGG ATGAGGATGA
TATGCAGTGC CACAATCACT GCAAGTCTAT TAAGGGTTAC
AAGGGAGGTT ATTGTGCTAA GGGGGGCTTT GTTTGCAAGT
GTTACTAG
```

Sequence Searches and Databases



Enter sequence(s) upload from file, select previously used or choose from My Patents

```
ATGCAATTTACCACCATCCTCTCCATCGGTATCACCGTCTTCGGACTTCTCAACACCGGAGCCTTTGCAGCACCCC
AGCCTGTTCCCGAGGCTTACGCTGTTTCTGATCCCGAGGCTCATCCTGACGATTTTGCTGGTATGGATGCGAACCA
ACTTCAGAAACGTGGATTTGGATGCAATGGTCCTTGGGATGAGGATGATATGCAGTGCCACAATCACTGCAAGTCT
ATTAAGGGTTACAAGGGAGGTTATTGTGCTAAGGGGGGCTTTGTTTGCAAGTGTTACTAG
```

Target Sequence Length Min:	
Max:	

General Parameters

Max target sequences:	250
Reverse:	True
Identities percent threshold:	75 %
Subject identity threshold:	0.0 %
Query coverage threshold:	0.0 %

Scoring Parameters

Minimum score:	1
Cost to open a gap:	1
Cost to extend a gap:	1

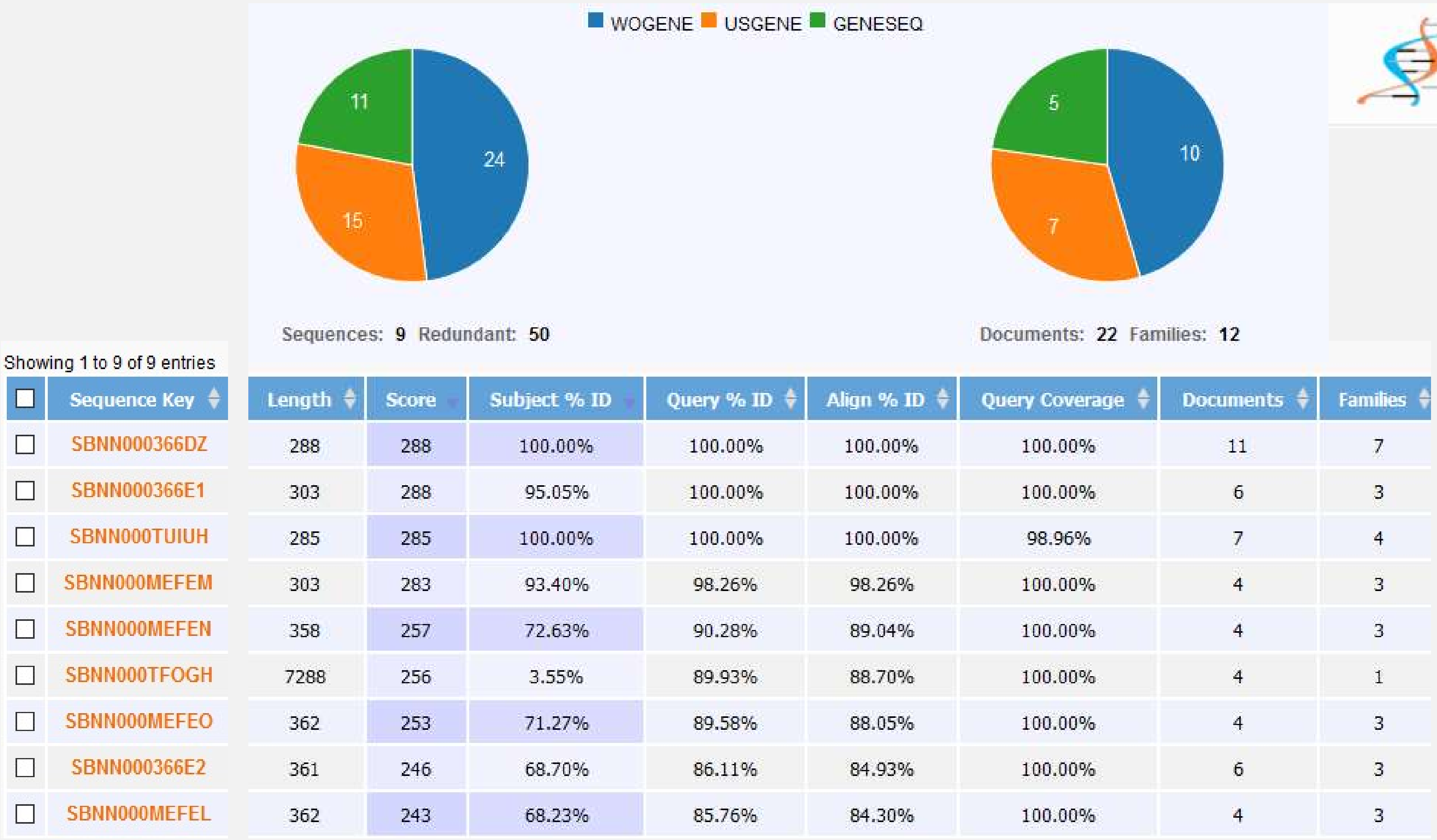
EXAMPLE: search SEQ ID NO:1 using BESTSeq of SequenceBase

US2005124064 A1
Applicant: **Novozymes**
Priority: 20 Nov 2001

```
>US_2005_0124064_A1_1
ATGCAATTTA CCACCATCCT CTCCATCGGT ATCACCGTCT
TCGGACTTCT CAACACCGGA GCCTTTGCAG CACCCCAGCC
TGTTCCCGAG GCTTACGCTG TTTCTGATCC CGAGGCTCAT
CCTGACGATT TTGCTGGTAT GGATGCGAAC CAACTTCAGA
AACGTGGATT TGGATGCAAT GGTCCTTGGG ATGAGGATGA
TATGCAGTGC CACAATCACT GCAAGTCTAT TAAGGGTTAC
AAGGGAGGTT ATTGTGCTAA GGGGGGCTTT GTTTGCAAGT
GTTACTAG
```

IngridB.Mueller@ipi.ch

Sequence Searches and Databases



Sequences: 9 Redundant: 50

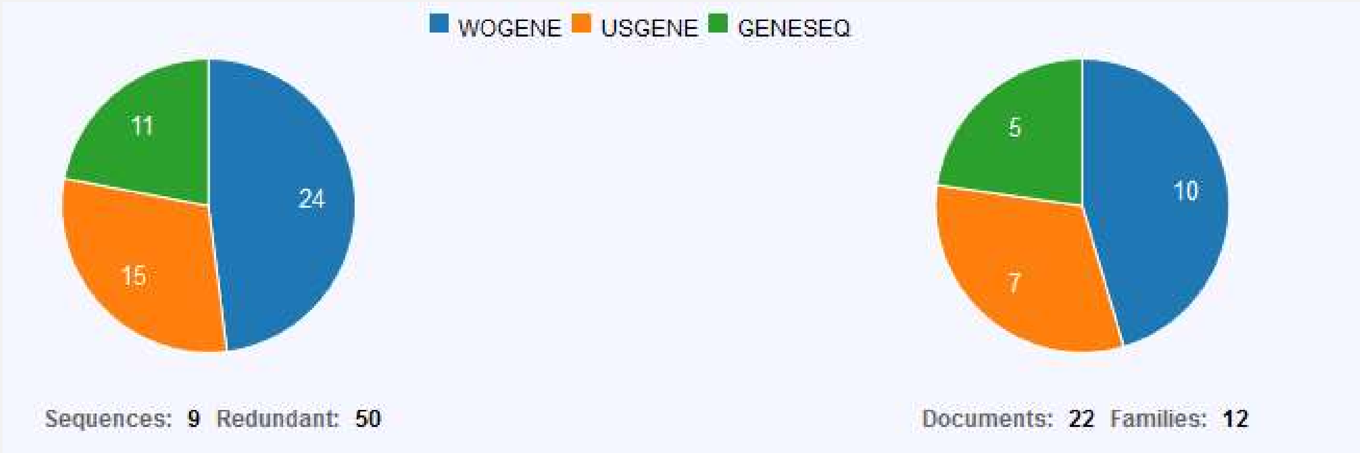
Documents: 22 Families: 12

Showing 1 to 9 of 9 entries

□	Sequence Key	Length	Score	Subject % ID	Query % ID	Align % ID	Query Coverage	Documents	Families
□	SBNN000366DZ	288	288	100.00%	100.00%	100.00%	100.00%	11	7
□	SBNN000366E1	303	288	95.05%	100.00%	100.00%	100.00%	6	3
□	SBNN000TUIUH	285	285	100.00%	100.00%	100.00%	98.96%	7	4
□	SBNN000MEFEM	303	283	93.40%	98.26%	98.26%	100.00%	4	3
□	SBNN000MEFEN	358	257	72.63%	90.28%	89.04%	100.00%	4	3
□	SBNN000TFOGH	7288	256	3.55%	89.93%	88.70%	100.00%	4	1
□	SBNN000MEFEO	362	253	71.27%	89.58%	88.05%	100.00%	4	3
□	SBNN000366E2	361	246	68.70%	86.11%	84.93%	100.00%	6	3
□	SBNN000MEFEL	362	243	68.23%	85.76%	84.30%	100.00%	4	3

IngridB.Mueller@ipi.ch

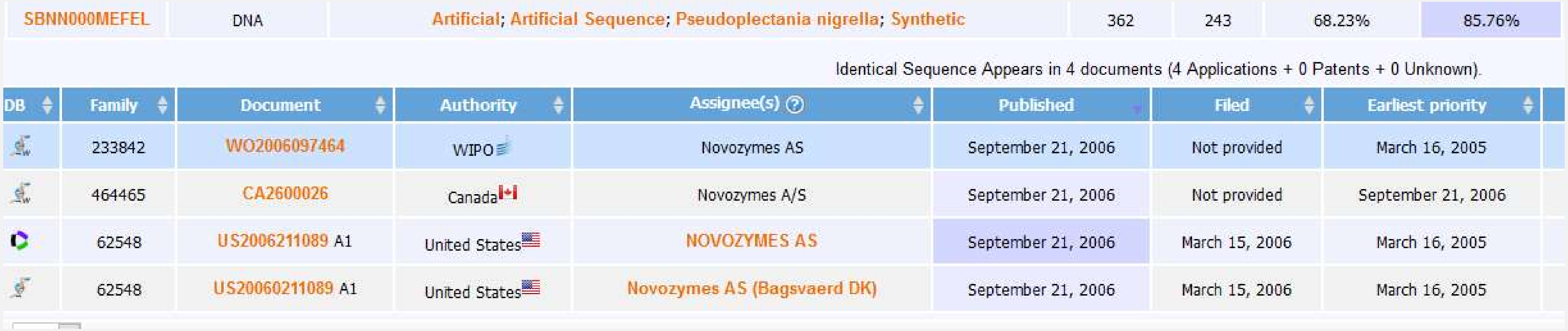
Sequence Searches and Databases



Sequences: 9 Redundant: 50

Documents: 22 Families: 12

SequenceBase



SBNN000MEFEL	DNA	Artificial; Artificial Sequence; Pseudoplectania nigrella; Synthetic	362	243	68.23%	85.76%

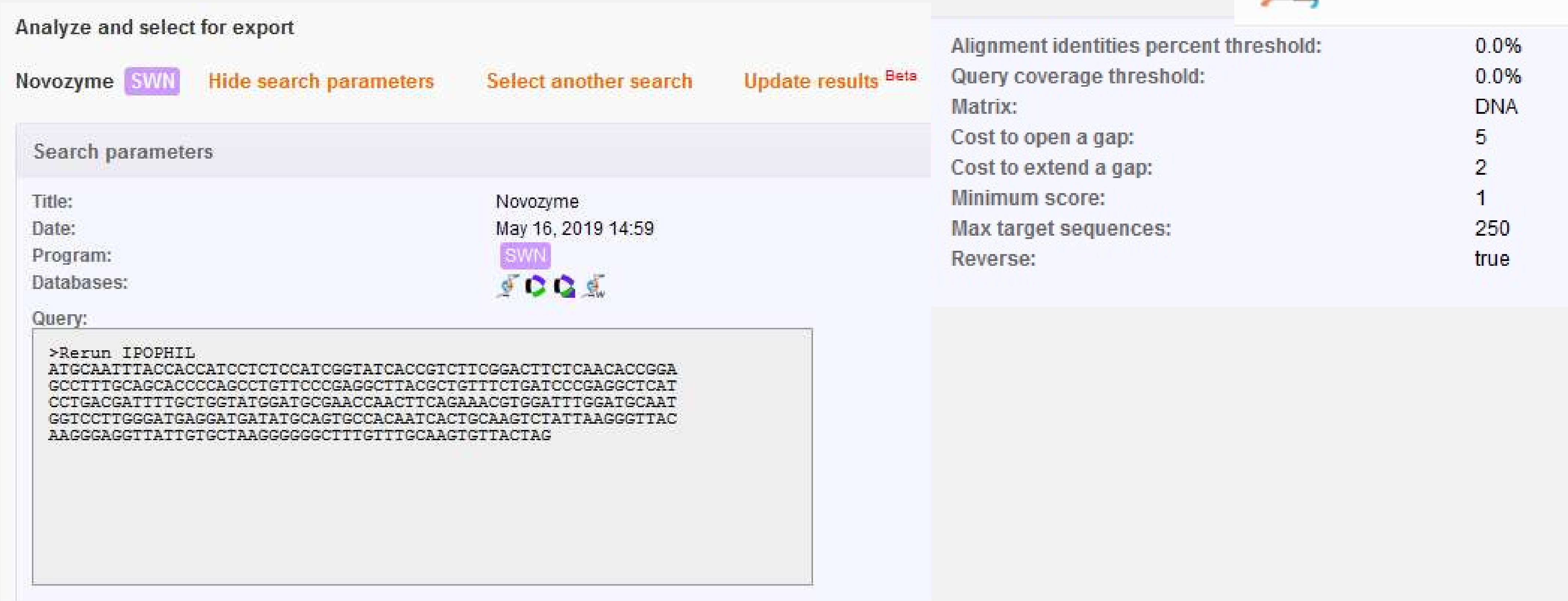
Identical Sequence Appears in 4 documents (4 Applications + 0 Patents + 0 Unknown).

DB	Family	Document	Authority	Assignee(s)	Published	Filed	Earliest priority
	233842	WO2006097464	WIPO	Novozymes AS	September 21, 2006	Not provided	March 16, 2005
	464465	CA2600026	Canada	Novozymes A/S	September 21, 2006	Not provided	September 21, 2006
	62548	US2006211089 A1	United States	NOVOZYMES AS	September 21, 2006	March 15, 2006	March 16, 2005
	62548	US20060211089 A1	United States	Novozymes AS (Bagsvaerd DK)	September 21, 2006	March 15, 2006	March 16, 2005

Novozyme 2006 → No prior art found using BESTSeq

IngridB.Mueller@ipi.ch

EXAMPLE: search SEQ ID NO:1 with Smith-Waterman for nucleotides (SWN)



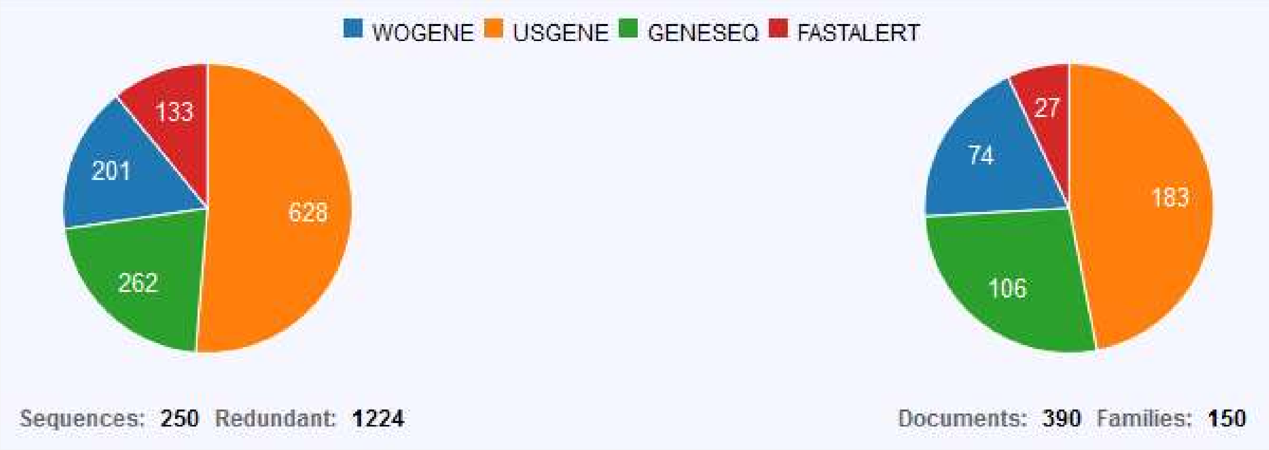
Title:	Novozyme
Date:	May 16, 2019 14:59
Program:	SWN
Databases:	

Query:

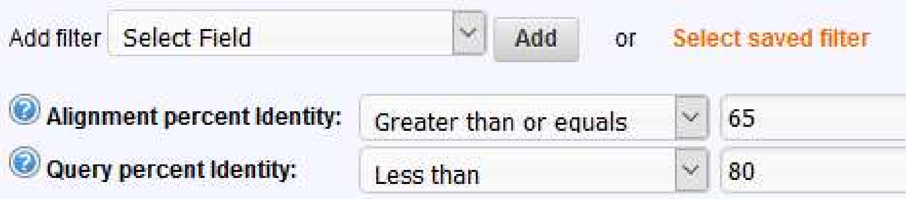
```
>Rerun IPOPHIL
ATGCAATTTACCACCATCCTCTCCATCGGTATCACCGTCTTCGGACTTCTCAACACCGGA
GCCTTTGCAGCACCCCAGCCTGTTCCCGAGGCTTACGCTGTTTCTGATCCCGAGGCTCAT
CCTGACGATTTTGCTGGTATGGATGCGAACCAACTTCAGAAACGTGGATTTGGATGCAAT
GGTCCTTGGGATGAGGATGATATGCAGTGCCACAATCACTGCAAGTCTATTAAGGGTTAC
AAGGGAGGTTATTGTGCTAAGGGGGGCTTTGTTTGCAAGTGTTACTAG
```

Alignment identities percent threshold:	0.0%
Query coverage threshold:	0.0%
Matrix:	DNA
Cost to open a gap:	5
Cost to extend a gap:	2
Minimum score:	1
Max target sequences:	250
Reverse:	true

Sequence Searches and Databases



If too many documents -> use filter



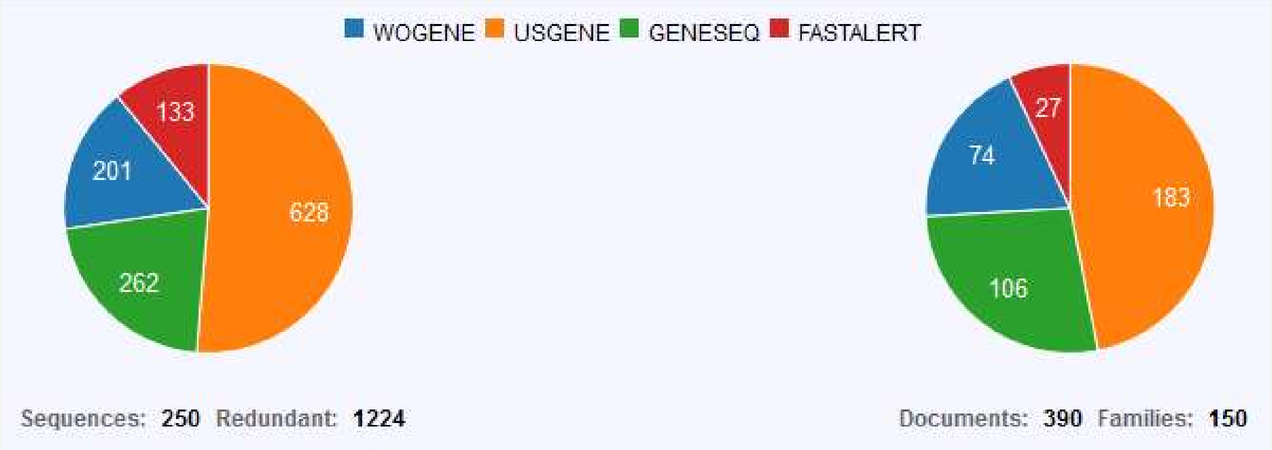
Add filter Select Field Add or Select saved filter

Alignment percent Identity: Greater than or equals 65

Query percent Identity: Less than 80

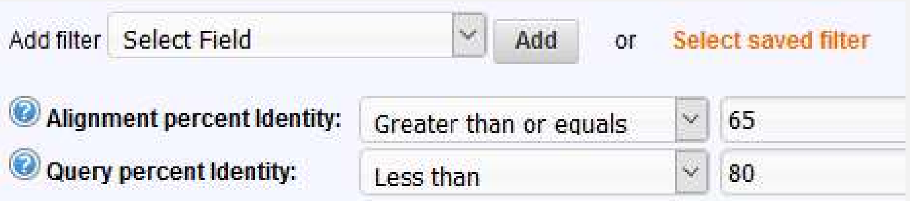
IGE | IPI

Sequence Searches and Databases



Sequences: **250** Redundant: **1224**

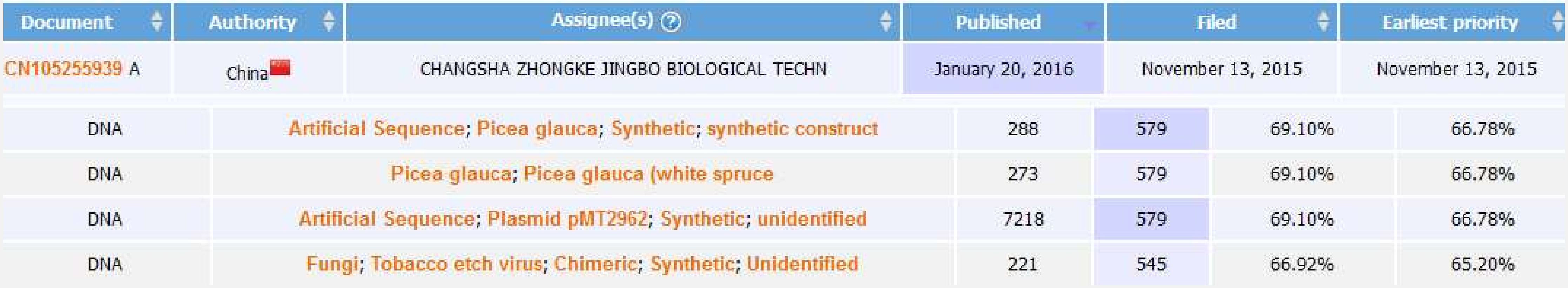
Documents: **390** Families: **150**



Add filter Select Field Add or Select saved filter

Alignment percent Identity: Greater than or equals 65

Query percent Identity: Less than 80



Document	Authority	Assignee(s) ?	Published	Filed	Earliest priority
CN105255939 A	China	CHANGSHA ZHONGKE JINGBO BIOLOGICAL TECHN	January 20, 2016	November 13, 2015	November 13, 2015

DNA	Artificial Sequence; Picea glauca; Synthetic; synthetic construct	288	579	69.10%	66.78%
DNA	Picea glauca; Picea glauca (white spruce	273	579	69.10%	66.78%
DNA	Artificial Sequence; Plasmid pMT2962; Synthetic; unidentified	7218	579	69.10%	66.78%
DNA	Fungi; Tobacco etch virus; Chimeric; Synthetic; Unidentified	221	545	66.92%	65.20%

… Not Novozymes, but 2016, thus not prior art

IngridB.Mueller@ipi.ch

Sequence Searches and Databases

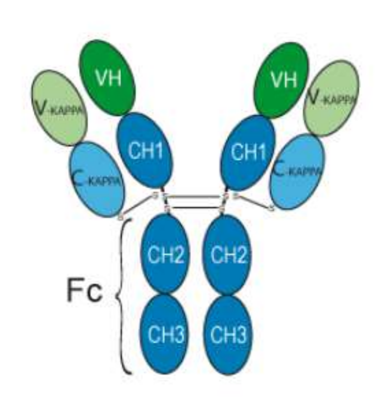
EXAMPLE: How to search a specific antibody

VH (1-120)

EVQLVESGGG LVQPGGSLRL SCAASGFNIK DTYIHWVRQA PGKGLEWVAR
IYPTNGYTRY ADSVKGRFTI SADTSKNTAY LQMNSLRAED TAVYYCSRWG
GDGFYAMDYW GQGTLVTVSS

VL (1-107)

DIQMTQSPSS LSASVGDRVT ITCRASQDVN TAVAWYQQKP GKAPKLLIYS
ASFLYSGVPS RFSGSRSGTD FTLTISSLQP EDFATYYCQQ HYTTPPTFGQ
GTKVEIK



trastuzumab

http://www.imgt.org/mAb-DB/

IGE | IPI

Sequence Searches and Databases

SequenceBase

EXAMPLE: How to search a specific antibody

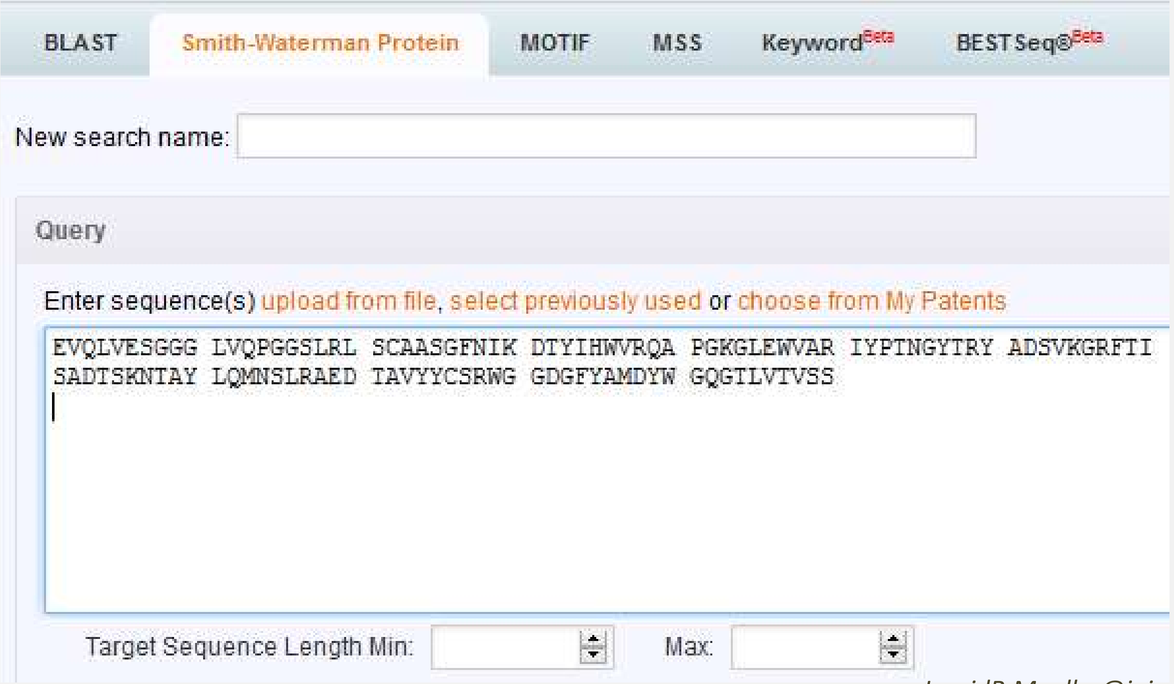
VH (1-120)

EVQLVESGGG LVQPGGSLRL SCAASGFNIK DTYIHWVRQA PGKGLEWVAR
IYPTNGYTRY ADSVKGRFTI SADTSKNTAY LQMNSLRAED TAVYYCSRWG
GDGFYAMDYW GQGTLVTVSS

VL (1-107)

DIQMTQSPSS LSASVGDRVT ITCRASQDVN TAVAWYQQKP GKAPKLLIYS
ASFLYSGVPS RFSGSRSGTD FTLTISSLQP EDFATYYCQQ HYTTPPTFGQ
GTKVEIK

Search full-length variable heavy (VH) chain using Smith-Waterman Protein (SWP) algorithm on SequenceBase



New search name:

Query

Enter sequence(s) upload from file, select previously used or choose from My Patents

```
EVQLVESGGG LVQPGGSLRL SCAASGFNIK DTYIHWVRQA PGKGLEWVAR IYPTNGYTRY ADSVKGRFTI
SADTSKNTAY LQMNSLRAED TAVYYCSRWG GDGFYAMDYW GQGTLVTVSS
```

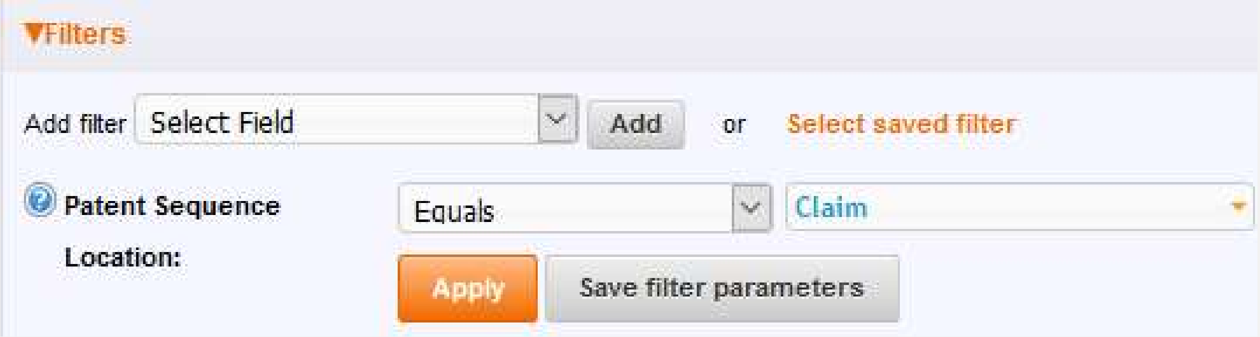
Target Sequence Length Min: Max:

IGE | IPI

Sequence Searches and Databases

If too many documents -> use filter

for e.g. claims, or DB, or both…

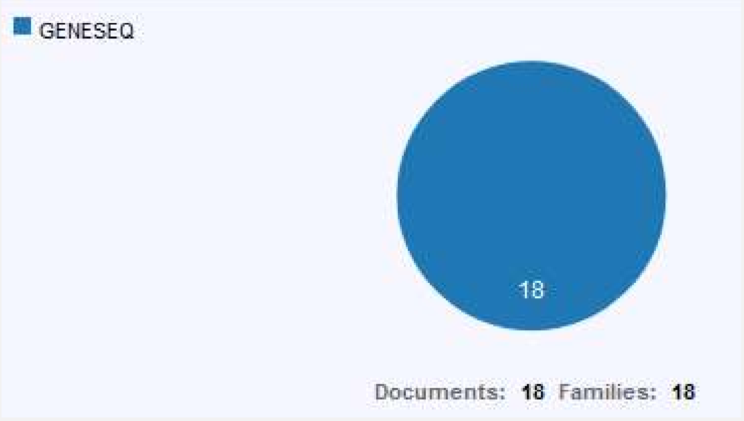


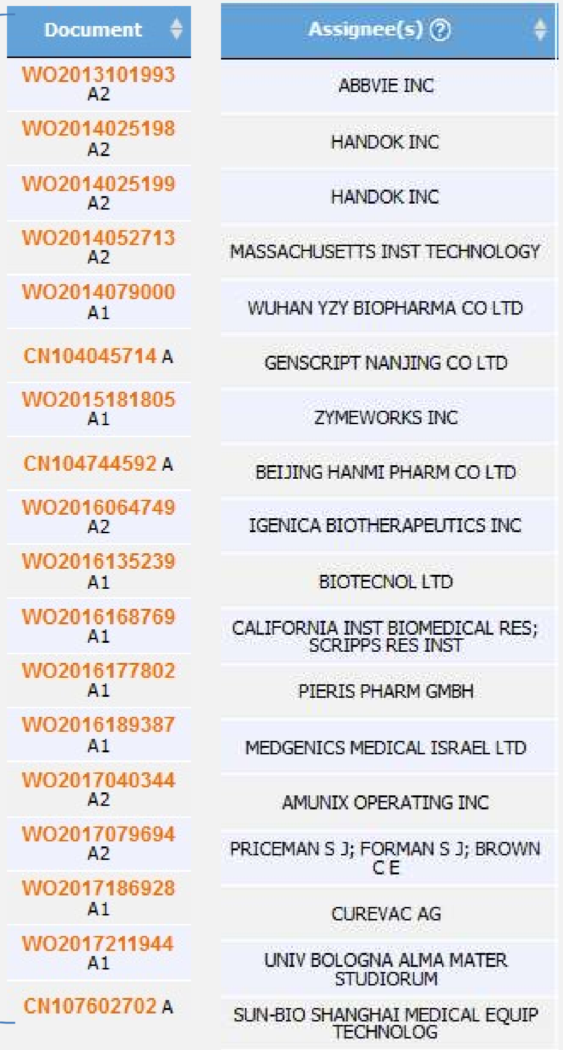
▼Filters

Add filter: Select Field ▾ Add or Select saved filter

Patent Sequence Location: Equals ▾ Claim ▾

Apply | Save filter parameters

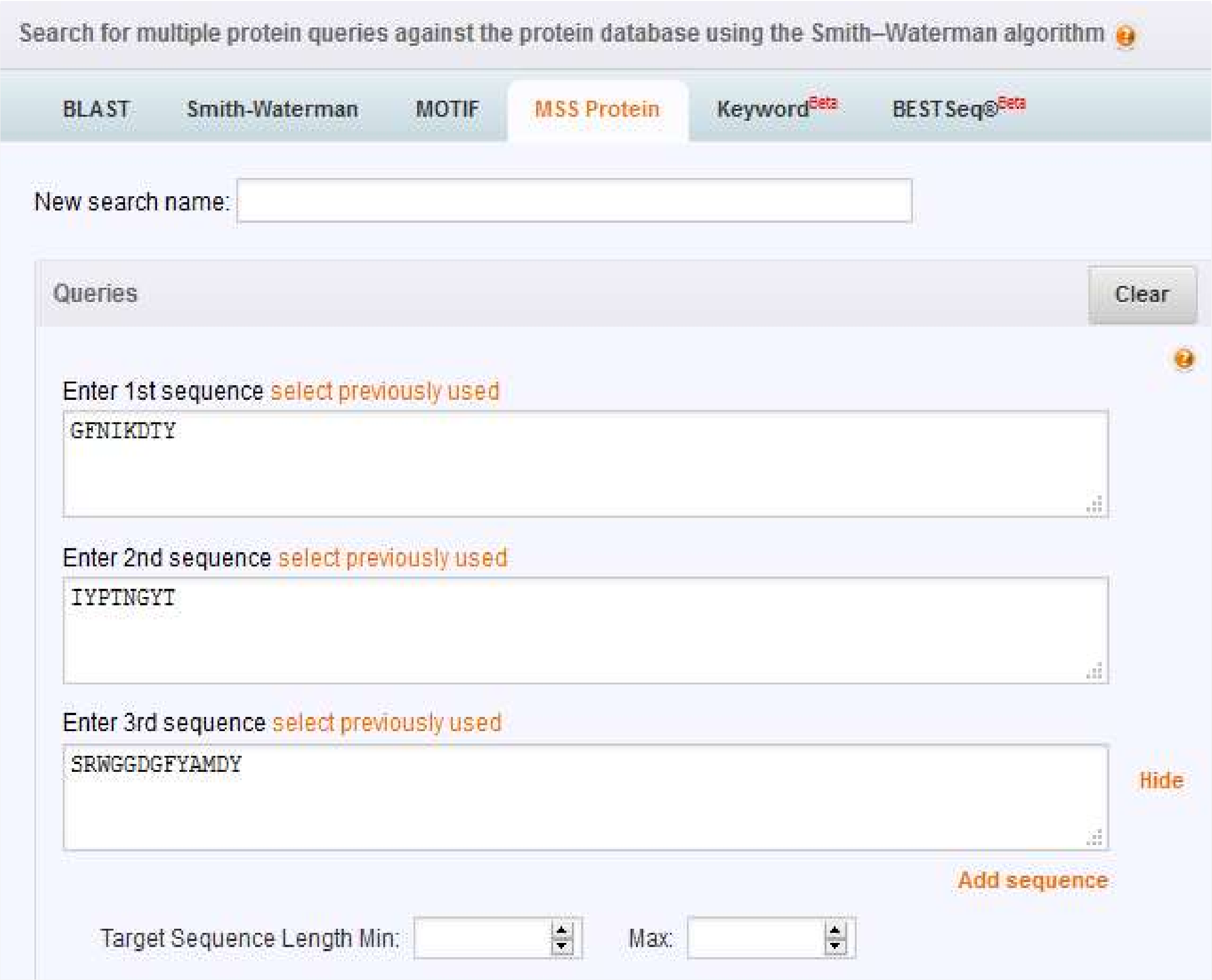




Document	Assignee(s)
WO2013101993 A2	ABBVIE INC
WO2014025198 A2	HANDOK INC
WO2014025199 A2	HANDOK INC
WO2014052713 A2	MASSACHUSETTS INST TECHNOLOGY
WO2014079000 A1	WUHAN YZY BIOPHARMA CO LTD
CN104045714 A	GENSCRIPT NANJING CO LTD
WO2015181805 A1	ZYMEWORKS INC
CN104744592 A	BEIJING HANMI PHARM CO LTD
WO2016064749 A2	IGENICA BIOTHERAPEUTICS INC
WO2016135239 A1	BIOTECNOL LTD
WO2016168769 A1	CALIFORNIA INST BIOMEDICAL RES; SCRIPPS RES INST
WO2016177802 A1	PIERIS PHARM GMBH
WO2016189387 A1	MEDGENICS MEDICAL ISRAEL LTD
WO2017040344 A2	AMUNIX OPERATING INC
WO2017079694 A2	PRICEMAN S J; FORMAN S J; BROWN C E
WO2017186928 A1	CUREVAC AG
WO2017211944 A1	UNIV BOLOGNA ALMA MATER STUDIORUM
CN107602702 A	SUN-BIO SHANGHAI MEDICAL EQUIP TECHNOLOG

IGE | IPI

Sequence Searches and Databases


Search for multiple protein queries against the protein database using the Smith–Waterman algorithm

New search name:

Enter 1st sequence select previously used

GFNIKDTY

Enter 2nd sequence select previously used

IYPTNGYT

Enter 3rd sequence select previously used

SRWGGDGFYAMDY

Target Sequence Length Min: Max:

SequenceBase

Search VH-CDRs of trastuzumab using MSS (multiple sequence search)

CDR1

SCAASGFNIKDTYIHWVRQA

CDR2

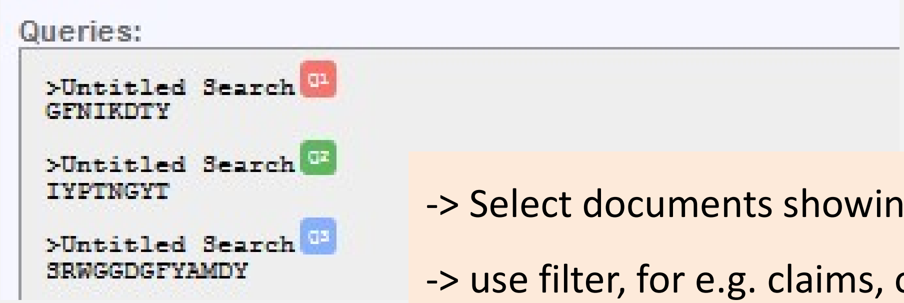
PGKGLEWVARIYPTNGYTRY

CDR3

TAVYYCSRWGGDGFYAMDYW

IngridB.Mueller@ipi.ch

Sequence Searches and Databases

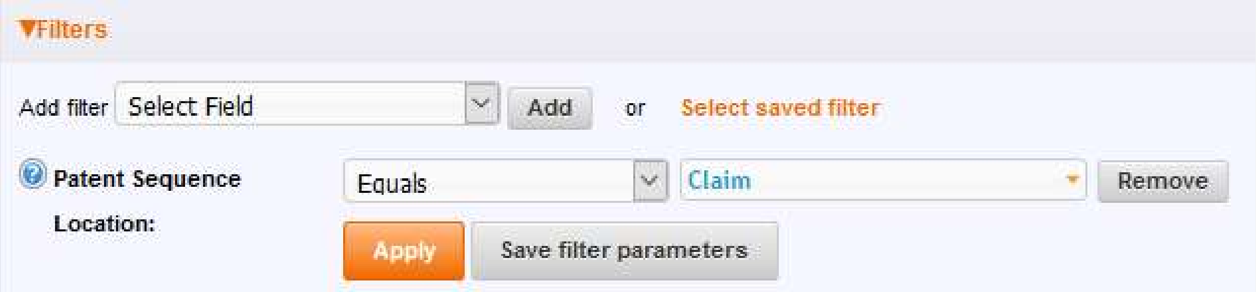


Queries:

```
>Untitled Search Q1
GFNIKDTY

>Untitled Search Q2
IYPTNGYT

>Untitled Search Q3
SRWGGDGFYAMDY
```

-> Select documents showing all three CDRs

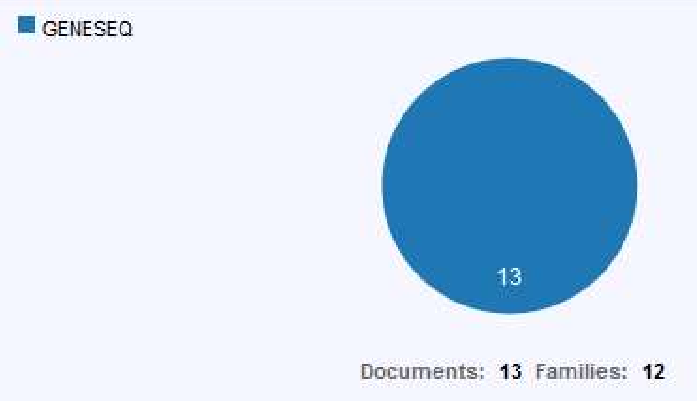
-> use filter, for e.g. claims, or DB, or both…



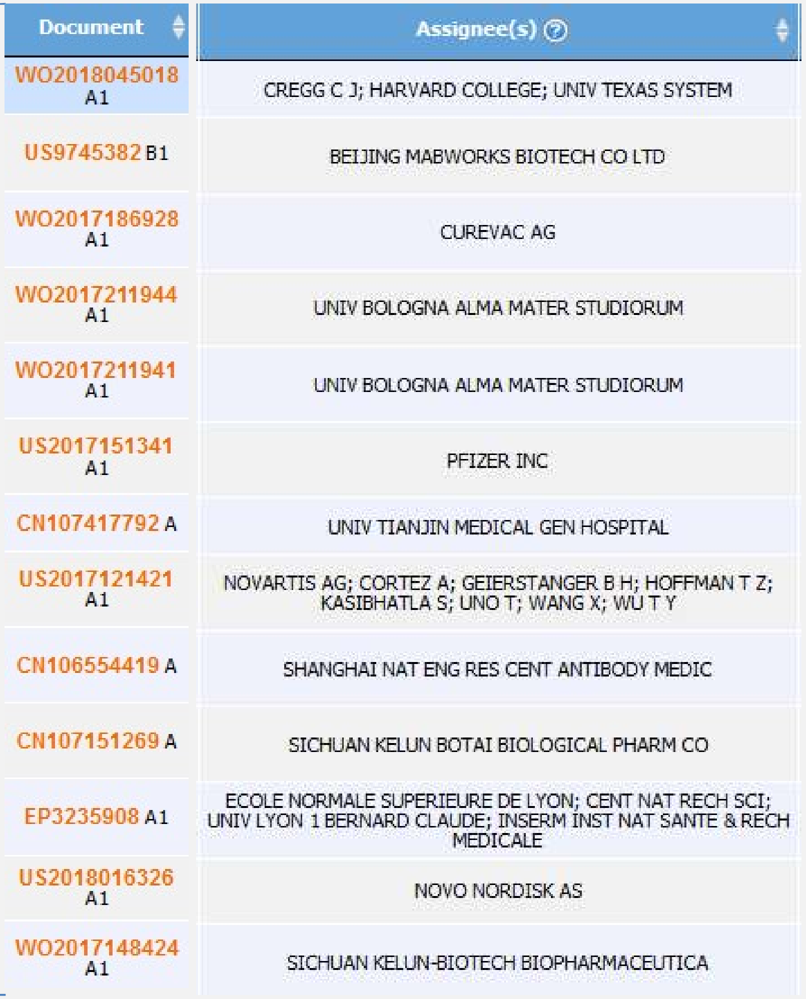
▼Filters

Add filter: Select Field | Add | or | Select saved filter

Patent Sequence Location: Equals | Claim | Remove



Documents: 13 Families: 12

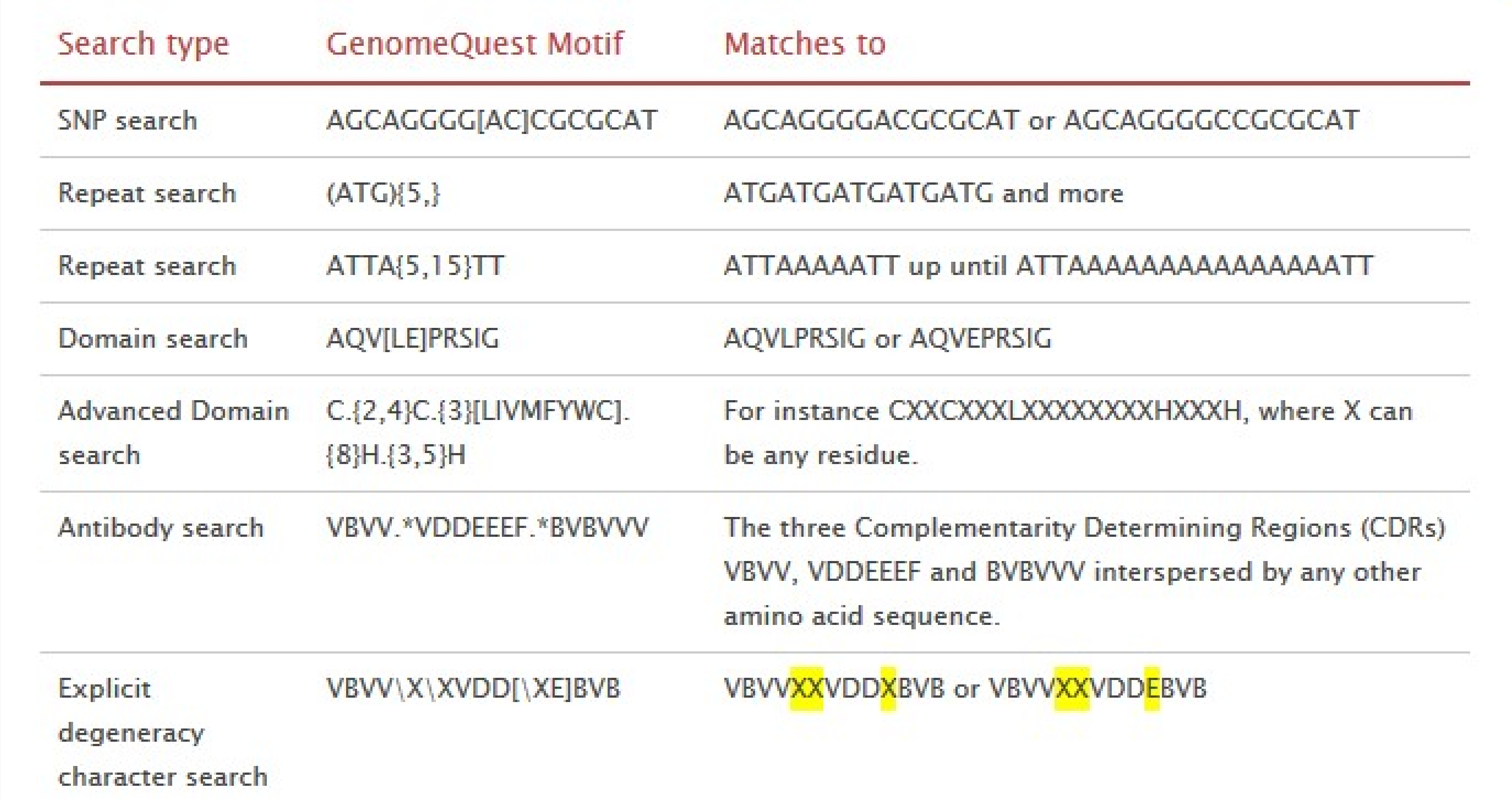


Document	Assignee(s)
WO2018045018 A1	CREGG C J; HARVARD COLLEGE; UNIV TEXAS SYSTEM
US9745382 B1	BEIJING MABWORKS BIOTECH CO LTD
WO2017186928 A1	CUREVAC AG
WO2017211944 A1	UNIV BOLOGNA ALMA MATER STUDIORUM
WO2017211941 A1	UNIV BOLOGNA ALMA MATER STUDIORUM
US2017151341 A1	PFIZER INC
CN107417792 A	UNIV TIANJIN MEDICAL GEN HOSPITAL
US2017121421 A1	NOVARTIS AG; CORTEZ A; GEIERSTANGER B H; HOFFMAN T Z; KASIBHATLA S; UNO T; WANG X; WU T Y
CN106554419 A	SHANGHAI NAT ENG RES CENT ANTIBODY MEDIC
CN107151269 A	SICHUAN KELUN BOTAI BIOLOGICAL PHARM CO
EP3235908 A1	ECOLE NORMALE SUPERIEURE DE LYON; CENT NAT RECH SCI; UNIV LYON 1 BERNARD CLAUDE; INSERM INST NAT SANTE & RECH MEDICALE
US2018016326 A1	NOVO NORDISK AS
WO2017148424 A1	SICHUAN KELUN-BIOTECH BIOPHARMACEUTICA

Sequence Searches and Databases

Using motif search by GenomeQuest

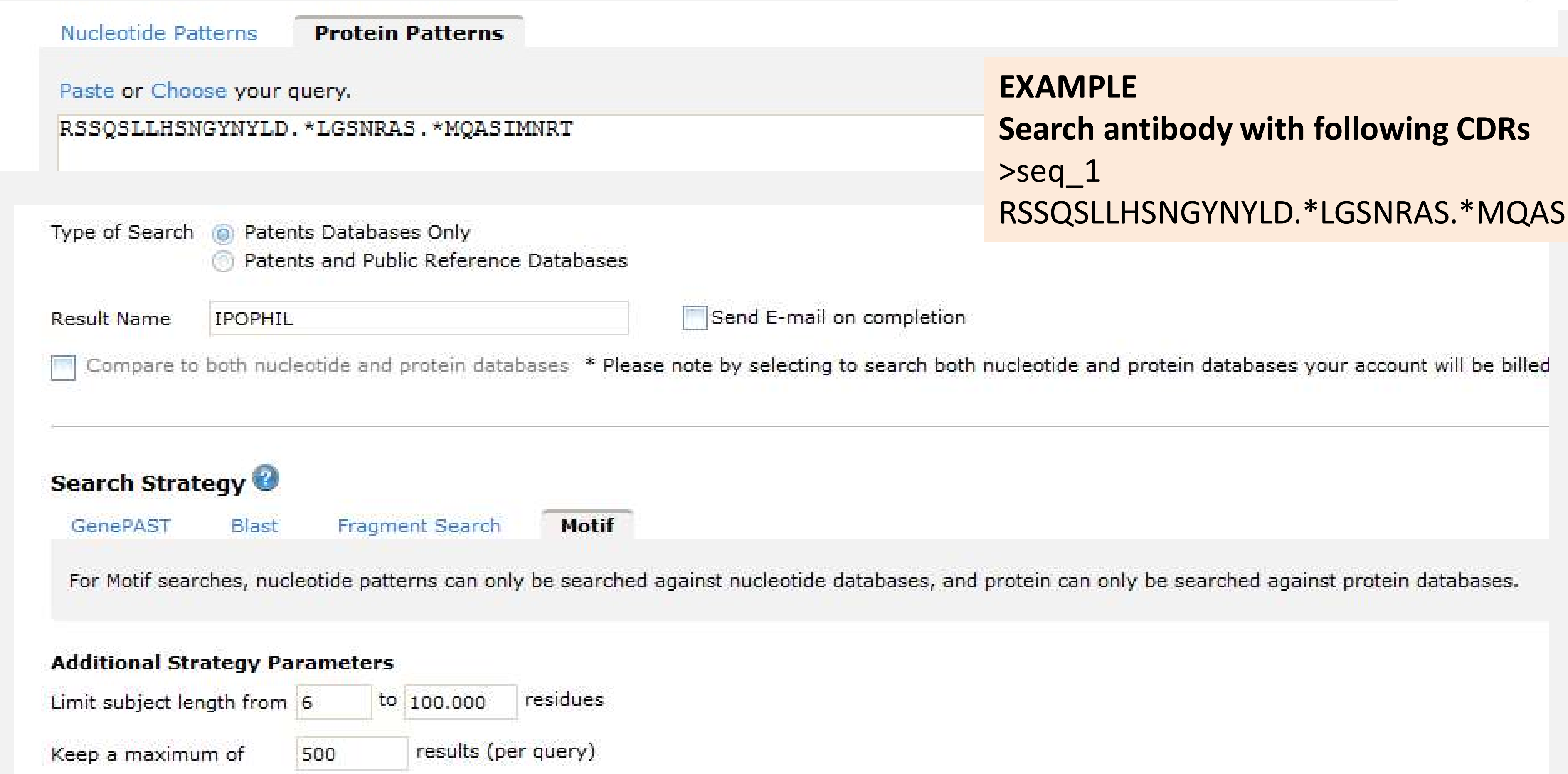
GQ Life Sciences
An Aptean Company



Search type	GenomeQuest Motif	Matches to
SNP search	AGCAGGGG[AC]CGCGCAT	AGCAGGGGACGCGCAT or AGCAGGGGCCGCGCAT
Repeat search	(ATG){5,}	ATGATGATGATGATG and more
Repeat search	ATTA{5,15}TT	ATTAAAAATT up until ATTAAAAAAAAAAAAAAATT
Domain search	AQV[LE]PRSIG	AQVLPRSIG or AQVEPRSIG
Advanced Domain search	C.{2,4}C.{3}[LIVMFYWC].{8}H.{3,5}H	For instance CXXCXXXLXXXXXXXXHXXXH, where X can be any residue.
Antibody search	VBVV.*VDDEEEF.*BVBVVV	The three Complementarity Determining Regions (CDRs) VBVV, VDDEEEF and BVBVVV interspersed by any other amino acid sequence.
Explicit degeneracy character search	VBVV\X\XVDD[\XE]BVB	VBVVXXVDDXBVB or VBVVXXVDDEBVB

Motif search for CDRs in the same sequence

GQ Life Sciences
An Aptean Company



Nucleotide Patterns | **Protein Patterns**

Paste or Choose your query.

```
RSSQSLLHSNGYNYLD.*LGSNRAS.*MQASIMNRT
```

Type of Search ● Patents Databases Only
○ Patents and Public Reference Databases

Result Name: IPOPHIL □ Send E-mail on completion

□ Compare to both nucleotide and protein databases * Please note by selecting to search both nucleotide and protein databases your account will be billed

Search Strategy

GenePAST | Blast | Fragment Search | **Motif**

For Motif searches, nucleotide patterns can only be searched against nucleotide databases, and protein can only be searched against protein databases.

Additional Strategy Parameters

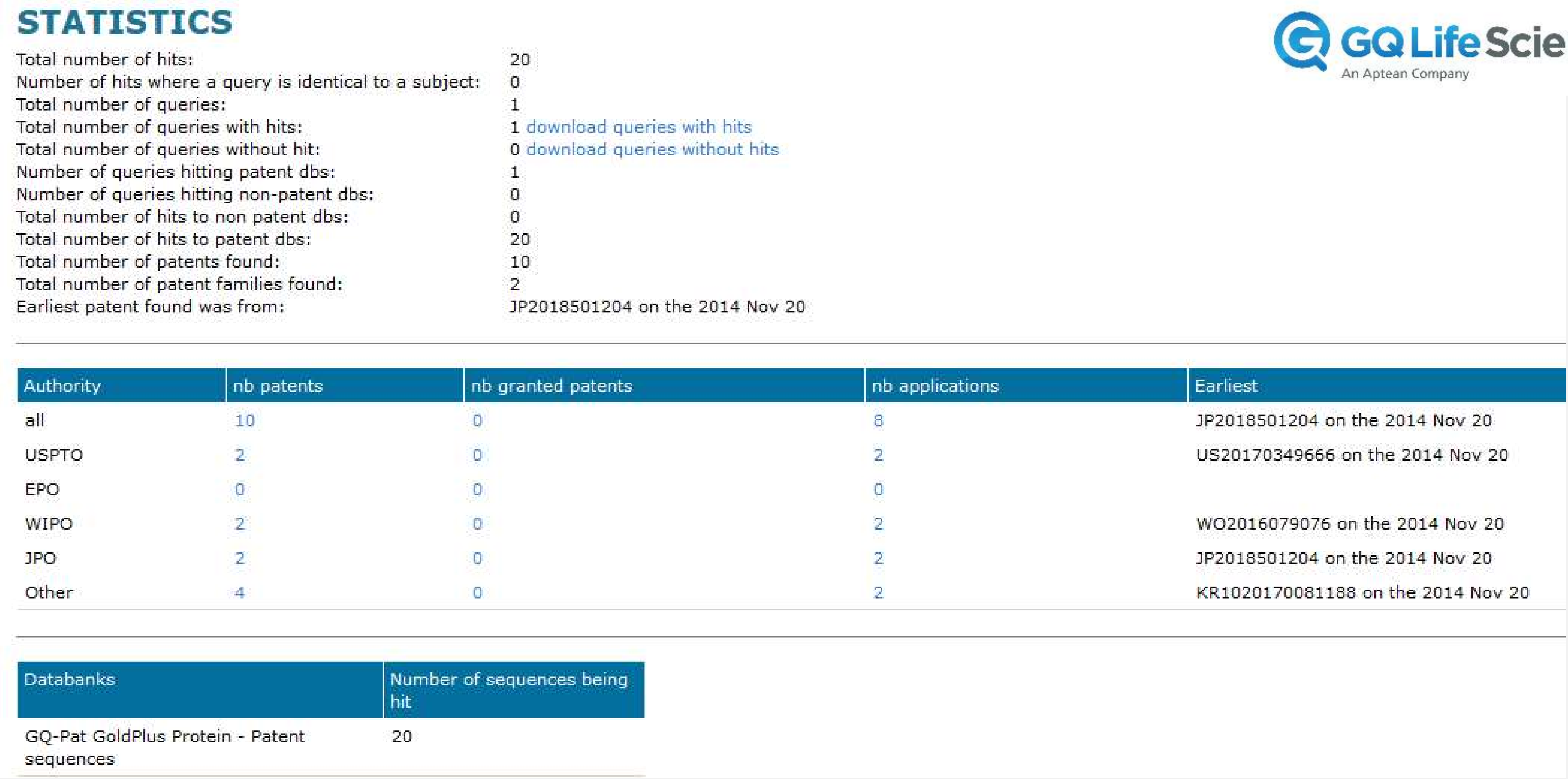
Limit subject length from 6 to 100.000 residues

Keep a maximum of 500 results (per query)

EXAMPLE
Search antibody with following CDRs
>seq_1
RSSQSLLHSNGYNYLD.*LGSNRAS.*MQASIMNRT

IGE | IPI

Sequence Searches and Databases



GQ Life Sciences
An Aptean Company

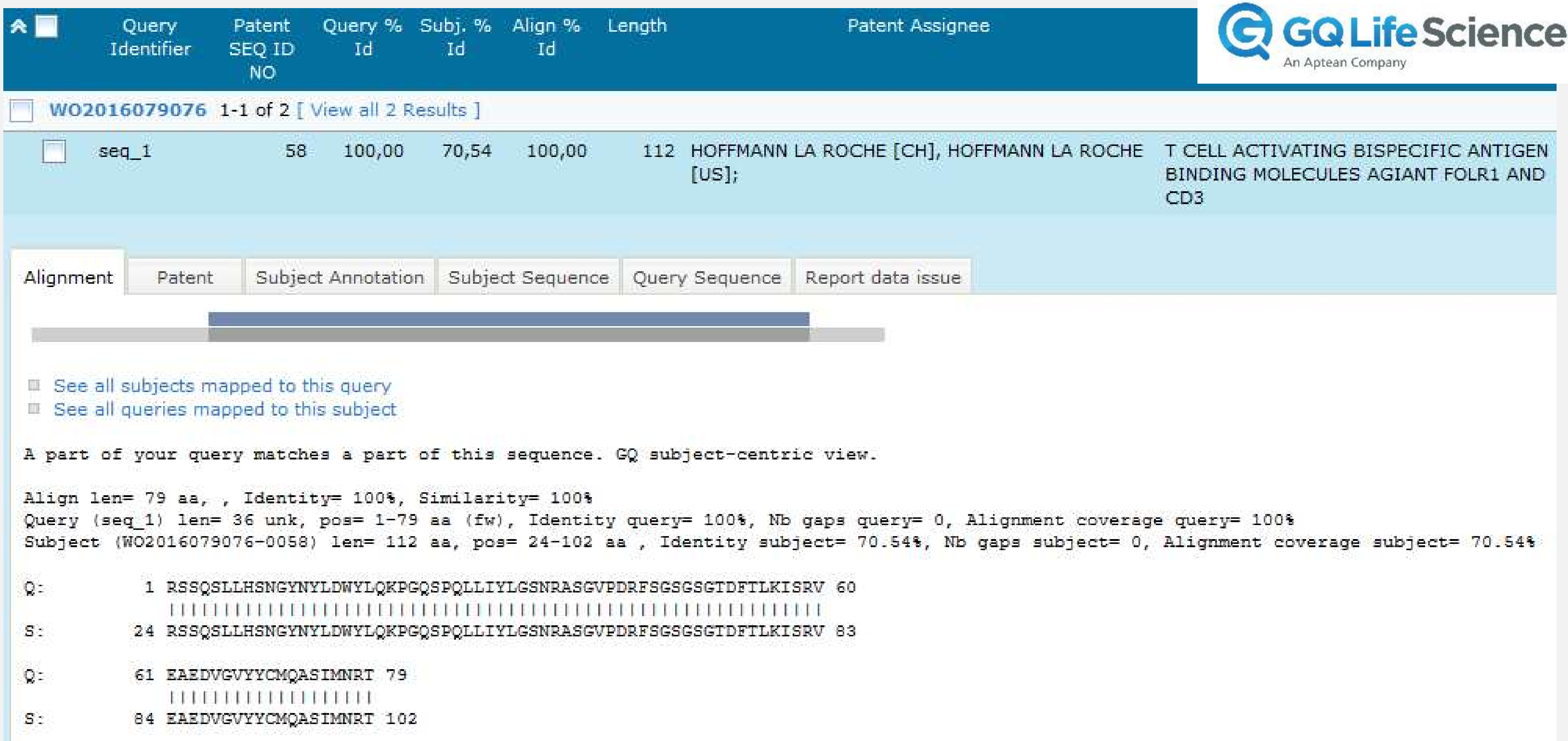
STATISTICS

Total number of hits:	20
Number of hits where a query is identical to a subject:	0
Total number of queries:	1
Total number of queries with hits:	1 download queries with hits
Total number of queries without hit:	0 download queries without hits
Number of queries hitting patent dbs:	1
Number of queries hitting non-patent dbs:	0
Total number of hits to non patent dbs:	0
Total number of hits to patent dbs:	20
Total number of patents found:	10
Total number of patent families found:	2
Earliest patent found was from:	JP2018501204 on the 2014 Nov 20

Authority	nb patents	nb granted patents	nb applications	Earliest
all	10	0	8	JP2018501204 on the 2014 Nov 20
USPTO	2	0	2	US20170349666 on the 2014 Nov 20
EPO	0	0	0	
WIPO	2	0	2	WO2016079076 on the 2014 Nov 20
JPO	2	0	2	JP2018501204 on the 2014 Nov 20
Other	4	0	2	KR1020170081188 on the 2014 Nov 20

Databanks	Number of sequences being hit
GQ-Pat GoldPlus Protein - Patent sequences	20

Sequence Searches and Databases



GQ Life Sciences
An Aptean Company

□	Query Identifier	Patent SEQ ID NO	Query % Id	Subj. % Id	Align % Id	Length	Patent Assignee	
□ WO2016079076 1-1 of 2 [View all 2 Results]								
□	seq_1	58	100,00	70,54	100,00	112	HOFFMANN LA ROCHE [CH], HOFFMANN LA ROCHE [US];	T CELL ACTIVATING BISPECIFIC ANTIGEN BINDING MOLECULES AGIANT FOLR1 AND CD3

- See all subjects mapped to this query
- See all queries mapped to this subject

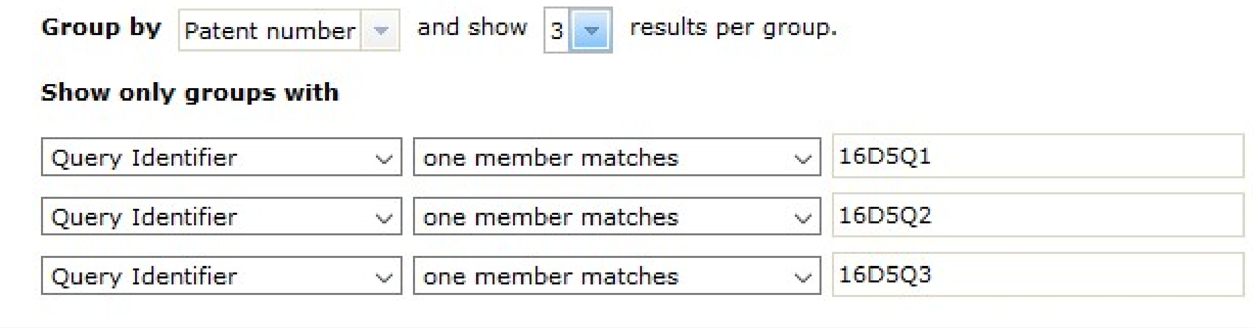
```
A part of your query matches a part of this sequence. GQ subject-centric view.

Align len= 79 aa, , Identity= 100%, Similarity= 100%
Query (seq_1) len= 36 unk, pos= 1-79 aa (fw), Identity query= 100%, Nb gaps query= 0, Alignment coverage query= 100%
Subject (WO2016079076-0058) len= 112 aa, pos= 24-102 aa , Identity subject= 70.54%, Nb gaps subject= 0, Alignment coverage subject= 70.54%

Q:         1 RSSQSLLHSNGYNYLDWYLQKPGQSPQLLIYLGSNRASGVPDRFSGSGSGTDFTLKISRV 60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
S:        24 RSSQSLLHSNGYNYLDWYLQKPGQSPQLLIYLGSNRASGVPDRFSGSGSGTDFTLKISRV 83

Q:        61 EAEDVGVYYCMQASIMNRT 79
             |||||||||||||||||||
S:        84 EAEDVGVYYCMQASIMNRT 102
```

Sequence Searches and Databases

EXAMPLE: using multiple alignments
Search antibody with following CDRs

>16D5Q1
NAWMS
>16D5Q2
RIKSKTDGGTTDYAAPVKG
>16D5Q3
PWEWSWYDY

All 3 CDRs in the same patent document !



Group by Patent number ▾ and show 3 ▾ results per group.

Show only groups with

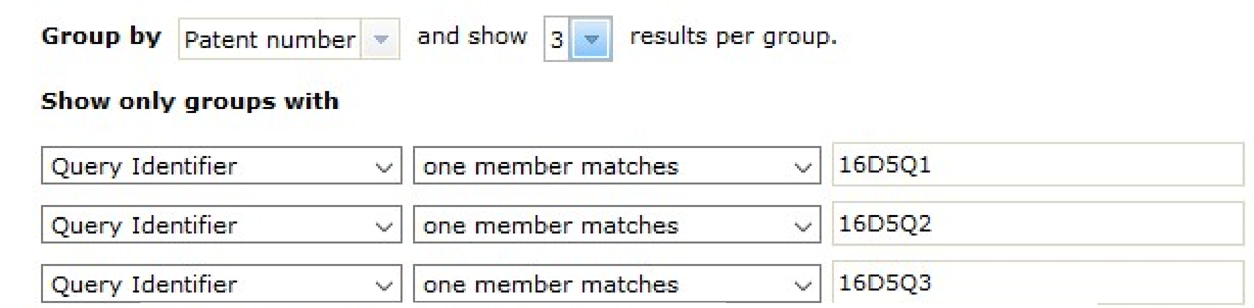
Query Identifier ▾	one member matches ▾	16D5Q1
Query Identifier ▾	one member matches ▾	16D5Q2
Query Identifier ▾	one member matches ▾	16D5Q3

Sequence Searches and Databases

EXAMPLE: using multiple alignments
Search antibody with following CDRs

>16D5Q1
NAWMS
>16D5Q2
RIKSKTDGGTTDYAAPVKG
>16D5Q3
PWEWSWYDY

All 3 CDRs in the same patent document !

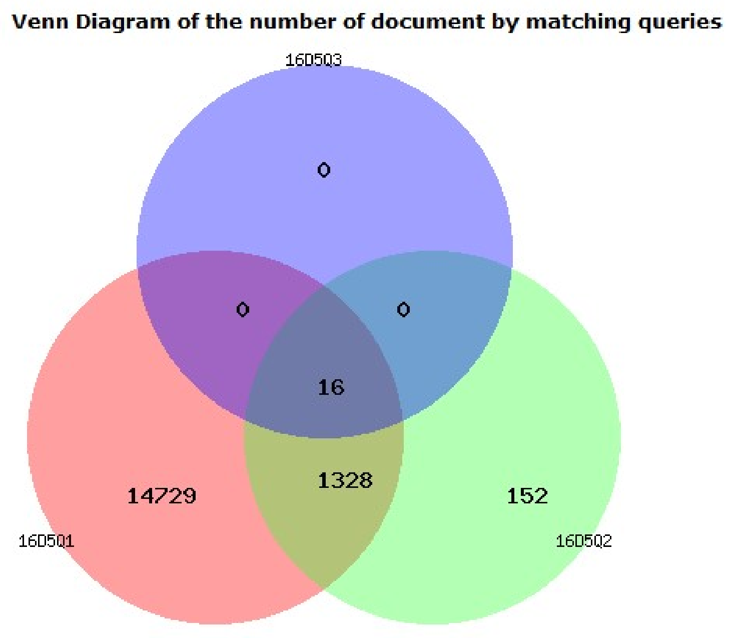


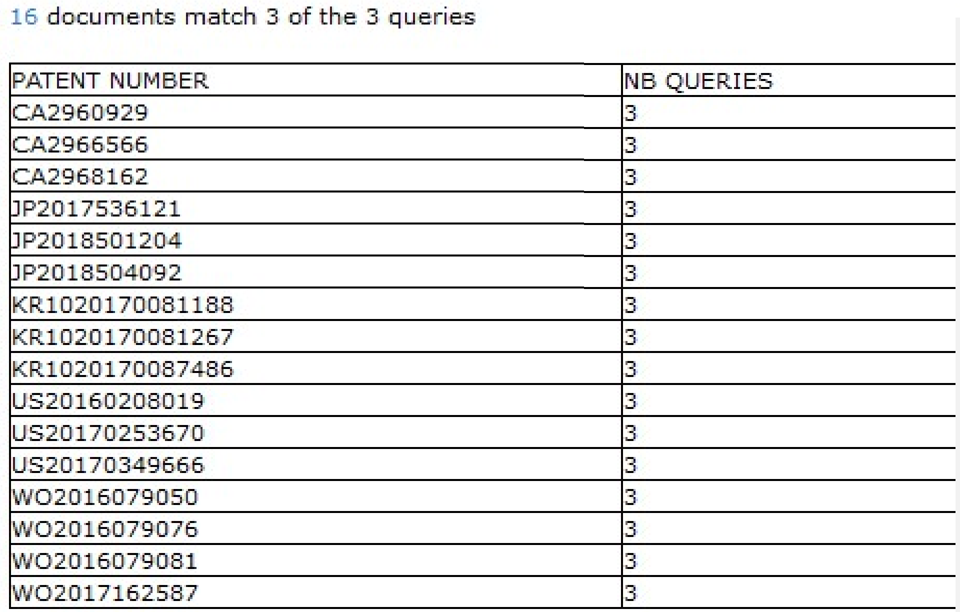
Group by Patent number ▾ and show 3 ▾ results per group.

Show only groups with

Query Identifier ▾	one member matches ▾	16D5Q1
Query Identifier ▾	one member matches ▾	16D5Q2
Query Identifier ▾	one member matches ▾	16D5Q3

Venn Diagram of the number of document by matching queries





16 documents match 3 of the 3 queries

PATENT NUMBER	NB QUERIES
CA2960929	3
CA2966566	3
CA2968162	3
JP2017536121	3
JP2018501204	3
JP2018504092	3
KR1020170081188	3
KR1020170081267	3
KR1020170087486	3
US20160208019	3
US20170253670	3
US20170349666	3
WO2016079050	3
WO2016079076	3
WO2016079081	3
WO2017162587	3

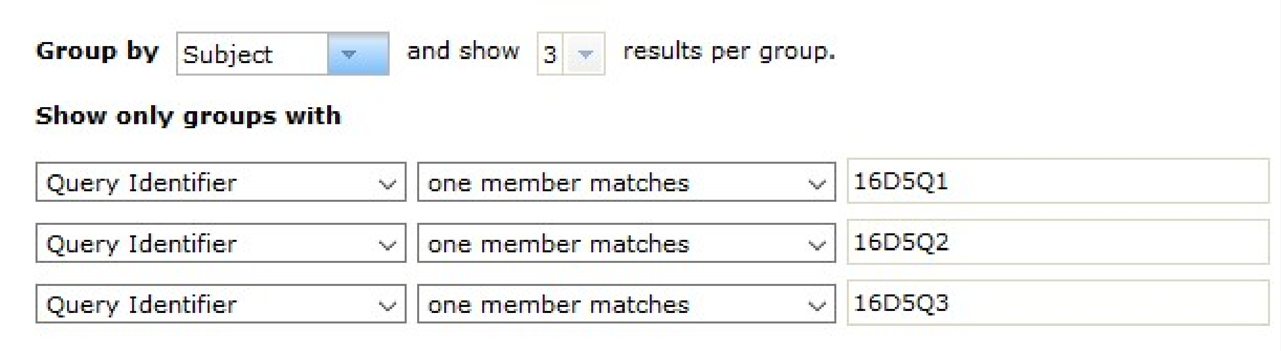
IngridB.Mueller@ipi.ch

Sequence Searches and Databases

EXAMPLE: using multiple alignments
Search antibody with following CDRs

>16D5Q1
NAWMS
>16D5Q2
RIKSKTDGGTTDYAAPVKG
>16D5Q3
PWEWSWYDY

All 3 CDRs in the same sequence !



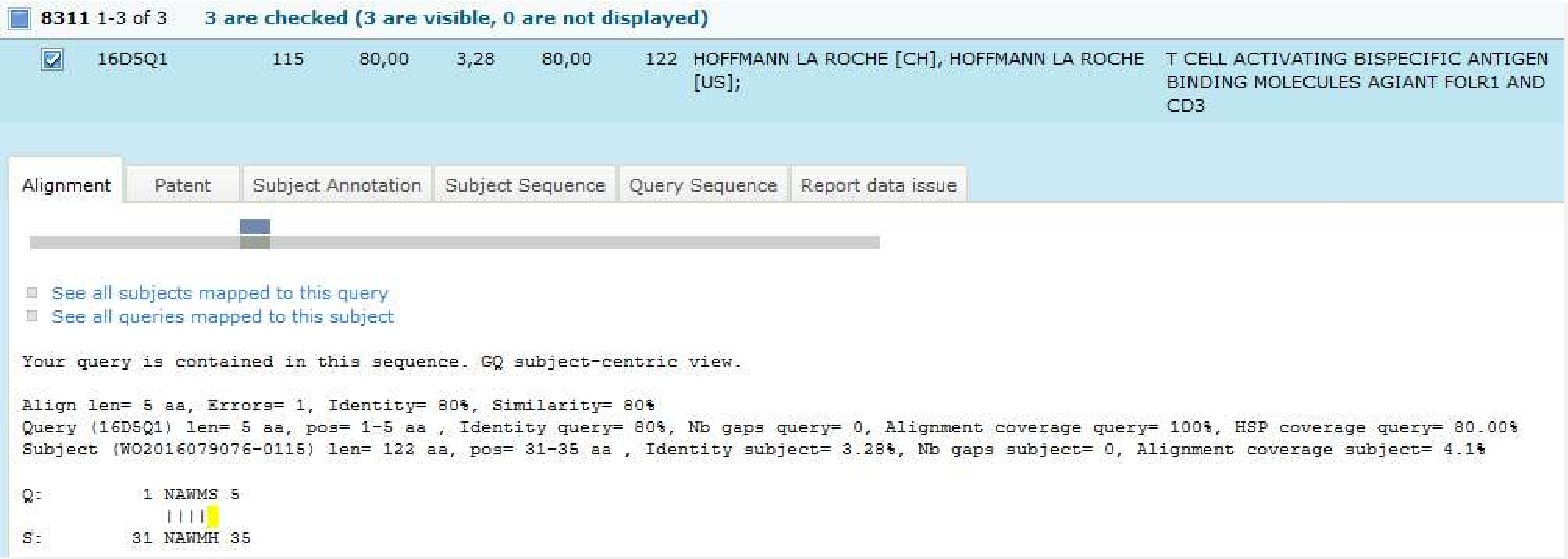
Group by Subject and show 3 results per group.

Show only groups with

Query Identifier — one member matches — 16D5Q1
Query Identifier — one member matches — 16D5Q2
Query Identifier — one member matches — 16D5Q3

IngridB.Mueller@ipi.ch

Sequence Searches and Databases



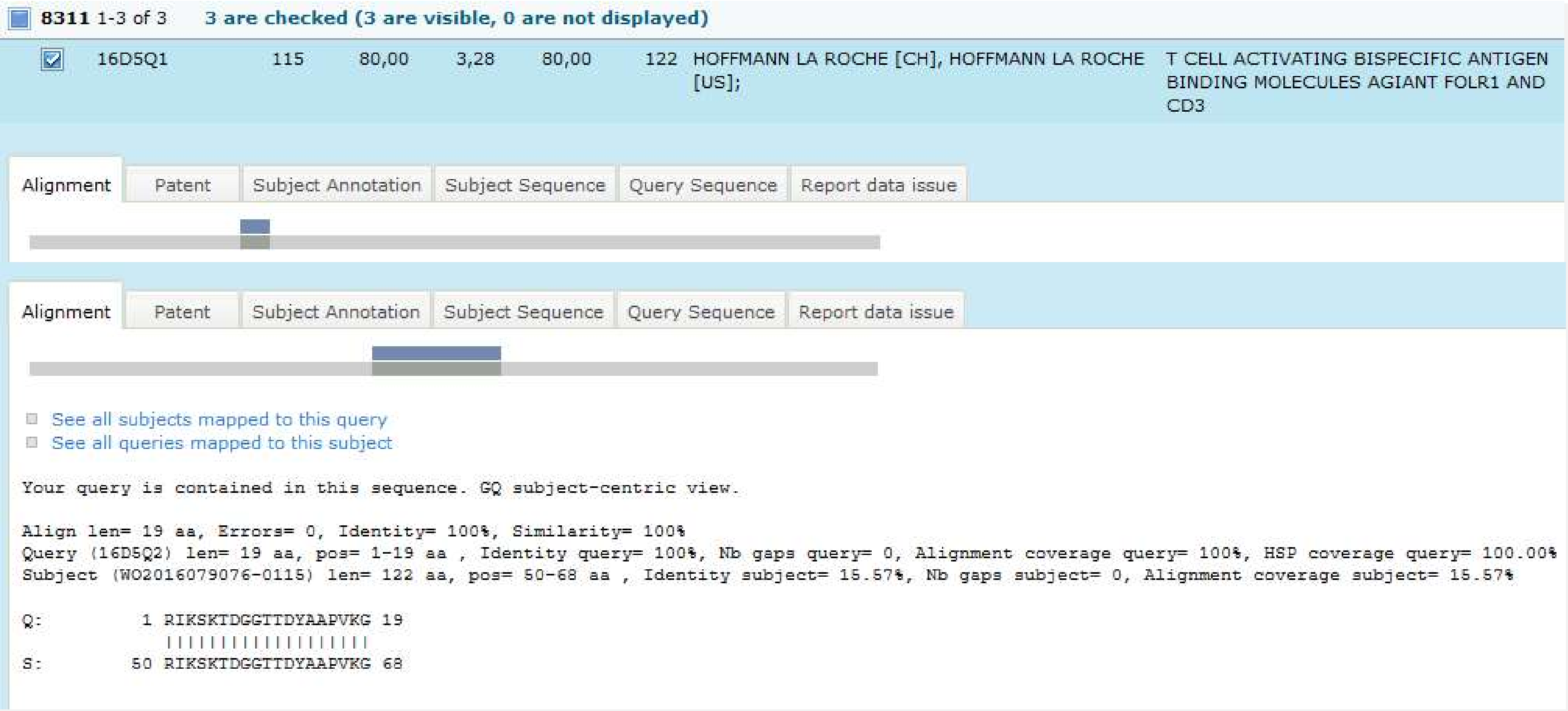
☐ **8311** 1-3 of 3 **3 are checked (3 are visible, 0 are not displayed)**

☑	16D5Q1	115	80,00	3,28	80,00	122	HOFFMANN LA ROCHE [CH], HOFFMANN LA ROCHE [US];	T CELL ACTIVATING BISPECIFIC ANTIGEN BINDING MOLECULES AGIANT FOLR1 AND CD3

☐ See all subjects mapped to this query
☐ See all queries mapped to this subject

```
Your query is contained in this sequence. GQ subject-centric view.

Align len= 5 aa, Errors= 1, Identity= 80%, Similarity= 80%
Query (16D5Q1) len= 5 aa, pos= 1-5 aa , Identity query= 80%, Nb gaps query= 0, Alignment coverage query= 100%, HSP coverage query= 80.00%
Subject (WO2016079076-0115) len= 122 aa, pos= 31-35 aa , Identity subject= 3.28%, Nb gaps subject= 0, Alignment coverage subject= 4.1%

Q:          1 NAWMS 5
              ||||
S:         31 NAWMH 35
```

Sequence Searches and Databases

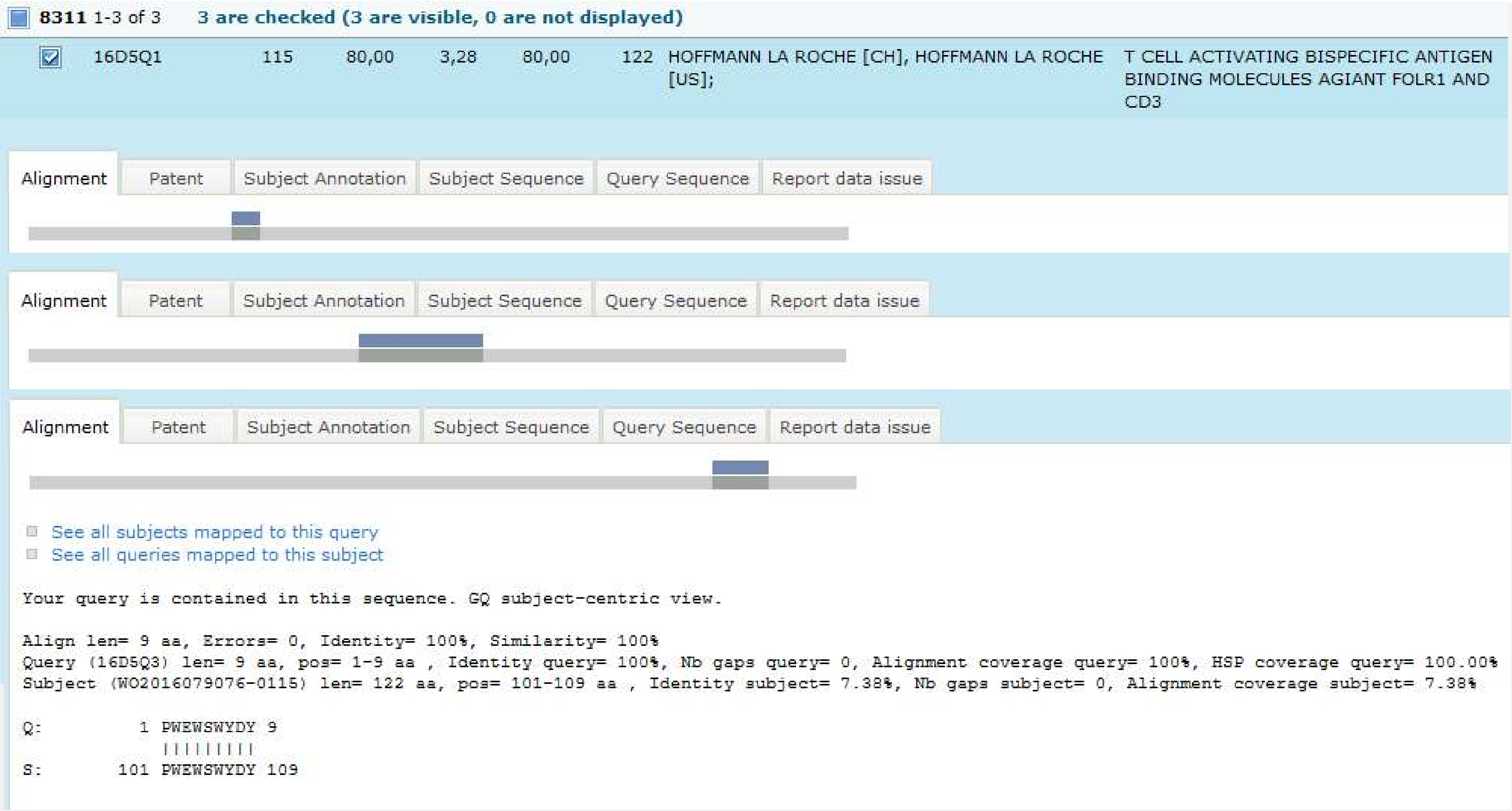


☐ **8311** 1-3 of 3 **3 are checked (3 are visible, 0 are not displayed)**

☑	16D5Q1	115	80,00	3,28	80,00	122	HOFFMANN LA ROCHE [CH], HOFFMANN LA ROCHE [US];	T CELL ACTIVATING BISPECIFIC ANTIGEN BINDING MOLECULES AGIANT FOLR1 AND CD3

```
Your query is contained in this sequence. GQ subject-centric view.

Align len= 19 aa, Errors= 0, Identity= 100%, Similarity= 100%
Query (16D5Q2) len= 19 aa, pos= 1-19 aa , Identity query= 100%, Nb gaps query= 0, Alignment coverage query= 100%, HSP coverage query= 100.00%
Subject (WO2016079076-0115) len= 122 aa, pos= 50-68 aa , Identity subject= 15.57%, Nb gaps subject= 0, Alignment coverage subject= 15.57%

Q:          1 RIKSKTDGGTTDYAAPVKG 19
                |||||||||||||||||||
S:         50 RIKSKTDGGTTDYAAPVKG 68
```

Sequence Searches and Databases



☐ **8311** 1-3 of 3 **3 are checked (3 are visible, 0 are not displayed)**

☑ 16D5Q1 115 80,00 3,28 80,00 122 HOFFMANN LA ROCHE [CH], HOFFMANN LA ROCHE [US]; T CELL ACTIVATING BISPECIFIC ANTIGEN BINDING MOLECULES AGIANT FOLR1 AND CD3

```
Your query is contained in this sequence. GQ subject-centric view.

Align len= 9 aa, Errors= 0, Identity= 100%, Similarity= 100%
Query (16D5Q3) len= 9 aa, pos= 1-9 aa , Identity query= 100%, Nb gaps query= 0, Alignment coverage query= 100%, HSP coverage query= 100.00%
Subject (WO2016079076-0115) len= 122 aa, pos= 101-109 aa , Identity subject= 7.38%, Nb gaps subject= 0, Alignment coverage subject= 7.38%

Q:         1 PWEWSWYDY 9
             |||||||||
S:       101 PWEWSWYDY 109
```