SciFinder Discovery Platform

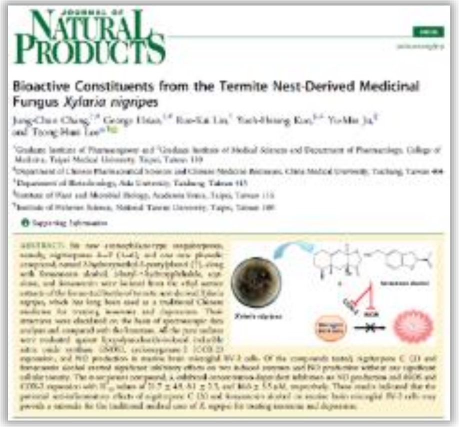
TRAINING

Karin Färber KFaerber@acs-i.org
Ernst Aichinger
ACS International / CAS
March 2022

CAS
A division of the
American Chemical Society

Content curation & Digitization

Delivering decades of paper-to-digital scientific content creation expertise



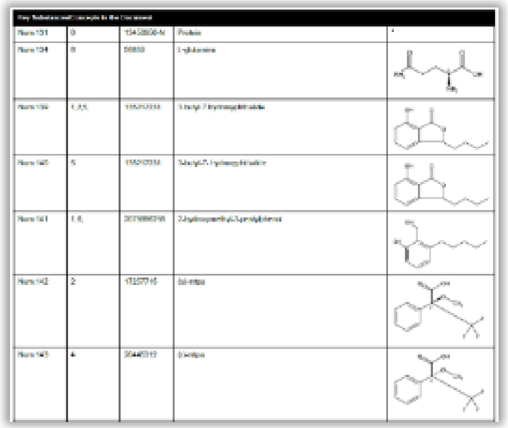
JOURNAL OF NATURAL PRODUCTS

Bioactive Constituents from the Termite Nest-Derived Medicinal Fungus *Xylaria nigripes*

[illegible]

Supporting Information

ABSTRACT: [illegible]

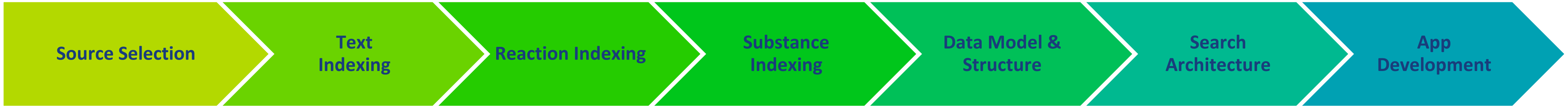


Key Substances/Concepts in the Document

[illegible]	0	[illegible]	Protein	
[illegible]	0	[illegible]	L-glutamine	
[illegible]	1,2,5,	[illegible]	[illegible]	
[illegible]	5	[illegible]	[illegible]	
[illegible]	1,6,	[illegible]	[illegible]	
[illegible]	2	[illegible]	[illegible]	
[illegible]	4	[illegible]	[illegible]	

```
<?xml version="1.0" encoding="UTF-8"?>
- <document dateTimeStamp="2018-01-03T18:38:02" dataDomain="CASEditorial">
  - <document-identity-group>
      <source-identity>
        <document-identifier>Example1_20172801[illegible]</document-identifier>
      </source-identity>
    - <cas-identity>
        <cas-document-identifier>[illegible]</cas-document-identifier>
      </cas-identity>
    </document-identity-group>
  - <document-source>
      <provider>[illegible]Examples</provider>
    </document-source>
  - <document-file-group>
      <source-file file-format="PDF">[illegible]</source-file>
      <cas-source-file file-format="PDF">[illegible].pdf</cas-source-file>
      <cas-extracted-embeddings/>
    </document-file-group>
    <pages>7</pages>
    <cas-section-number>10</cas-section-number>
  - <findings>
    - <finding-set>
      - <finding num="1" finding-type="substance">
          <regno>[illegible]</regno>
          <text-in-document>γ-lactone</text-in-document>
          <cas-display-name>γ-Butyrolactone</cas-display-name>
        - <cas-substance-name>
          - <name>
              2(3
              <hi>H</hi>
              )-Furanone, dihydro-
            </name>
            <preferred-substance-name-flag/>
            <name-type>[illegible]</name-type>
            <collective-index-code-cl>[illegible]</collective-index-code-cl>
            <collective-index-code-cl>[illegible]</collective-index-code-cl>
            <collective-index-code-cl>ACI</collective-index-code-cl>
```

CAS SciFinder^n



Source Selection

Text Indexing

Reaction Indexing

Substance Indexing

Data Model & Structure

Search Architecture

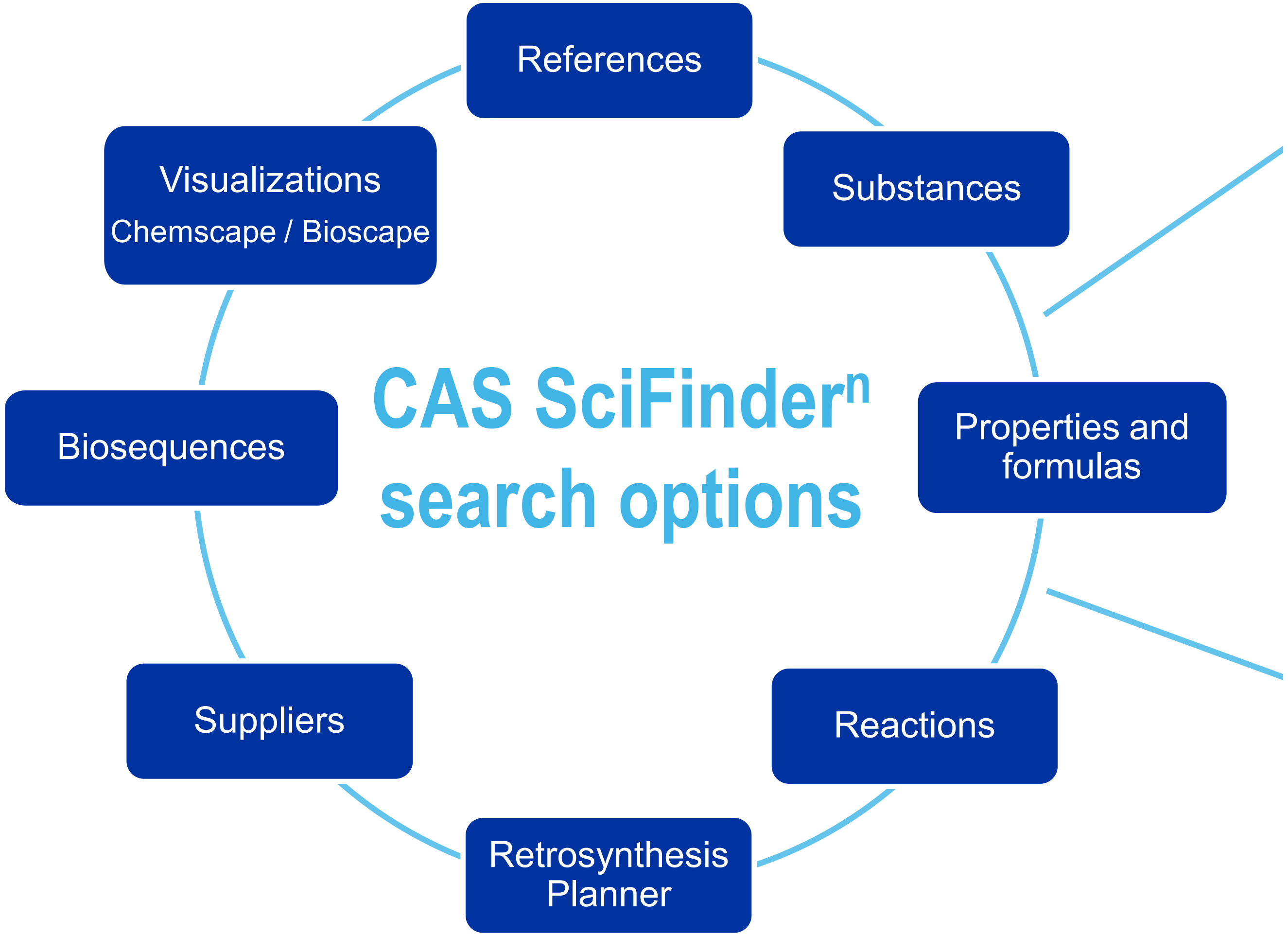
App Development

CAS

A division of the American Chemical Society

Components and Search Types

SciFinder Discovery Platform



Detailed Formulations

CAS Formulus

Quick access to the detailed formulation information for all active ingredients and excipients.
Main areas: pharma, agrochemistry and cosmetics.

Detailed Analytical Methods

CAS Analytical Methods

A single source for searching and comparing the latest published scientific analytic methods. Search and compare hundreds of thousands of methods across multiple scientific fields of study.

CAS
A division of the
American Chemical Society

CAS CONTENT COLLECTION

Curated, connected data analyzed by scientific experts fluent in over 50 languages

CAS REGISTRY

CAS Reactions

CAS References

CAS Markush

CAS Commercial Sources

CAS Formulations

Medline (largest portion of PubMed)

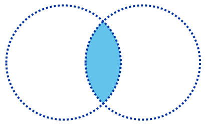
CAS Biosequences

CAS
A division of the
American Chemical Society

Boolean operators

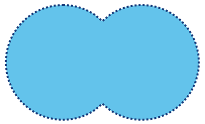
CAS SciFinder^n searching

AND requires both concepts to be present within the document



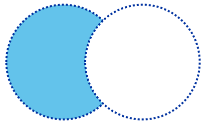
A AND B

OR requires either one or both concepts to be present



A OR B

NOT excludes documents from an answer set



A NOT B

Example: menthol **and** (food **or** candy) **not** cigarette

CAS
A division of the
American Chemical Society

Phrase search and wildcards

Terms enclosed in double quotes will be searched in the input order

- E.g.: "transcription factor"

Wildcards for internal and right-hand truncation

- * replaces any number of characters | plasticiz* → plasticizer, plasticized, plasticizing
- ? replaces 0 or 1 character | 1,?-Dibromobutane → 1,1-Dibromobutane ... 1,4-Dibromobutane

SEARCHING WITH SCIFINDER^{N}

BIOSEQUENCES

CAS

A division of the
American Chemical Society

Biosequence Search Options

- **Compounds with max. 252 non-H atoms**
  - Indexed as “usual” small molecules
  - Can be searched by sequence, structure, name, molecular formula, etc.
- **Defined compounds with more than 252 non-H atoms**
  - Are not indexed with structure: NOT searchable by structure
  - Might be indexed with molecular formula
  - Can thus be searched by name/identifier, molecular formula and sequence
- **Unspecific biomolecules**
  - Can be searched by name/identifier
  - Include reference search with broad/generic definitions

Biosequence Search Options

- **Compounds with max. 252 non-H atoms**
  - Indexed as “usual” small molecules
  - Can be searched by sequence, structure, name, molecular formula, etc.

Search shorter sequences for modifications and uncommon amino acids

- **Defined compounds with more than 252 non-H atoms**
  - Are not indexed with structure: NOT searchable by structure
  - Might be indexed with molecular formula
  - Can thus be searched by name/identifier, molecular formula and sequence

BLAST
CDR
Motif

- **Unspecific biomolecules**
  - Can be searched by name/identifier
  - Include reference search with broad/generic definitions

E.g. for classes of sequences or trade-names

Retrieve sequence-related information via text search

Biosequence Searching

New Data, New Options

Comprehensive collection:

- Over 600M patent-sequence relationships from more than 1.1 M patents and 62+ patent authorities
- Over 500M sequences submitted to the non-redundant NCBI protein and nucleotide databases
- Manually curated sequences not found in electronic sequence listings and other databases

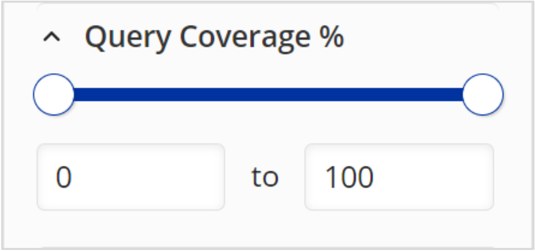
New biosequence search option:

- **BLAST** Search nucleotides and proteins
- **CDR** Search complementarity-determining regions of antibodies and T-Cell receptors
- **Motif** Sequence matching with wild cards and other features

Bioscape visualization tool provides additional analysis options

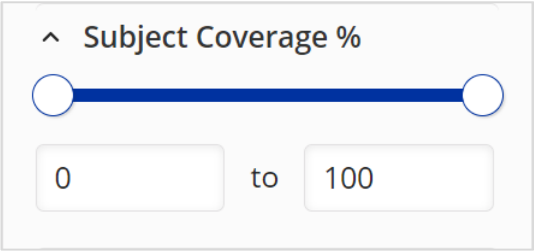
Alignment parameters

Coverage and Sequence Identity percentages



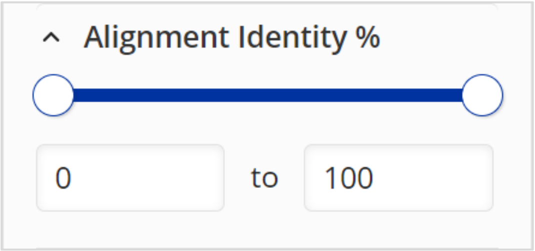
Query Coverage %

0 to 100



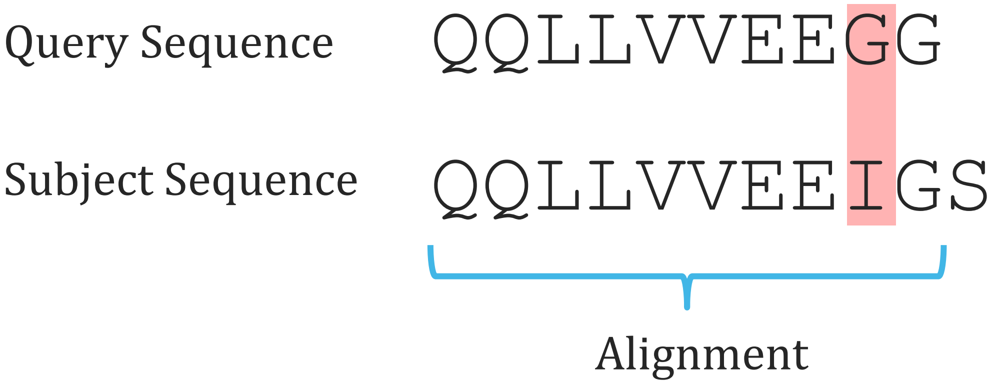
Subject Coverage %

0 to 100



Alignment Identity %

0 to 100



Query Sequence QQLLVVEEGG

Subject Sequence QQLLVVEEIGS

Alignment

$$\text{Query Coverage (100\%)} = \frac{\mathit{Alignment\ Length}}{\mathit{Query\ Length}} = \frac{10}{10}$$

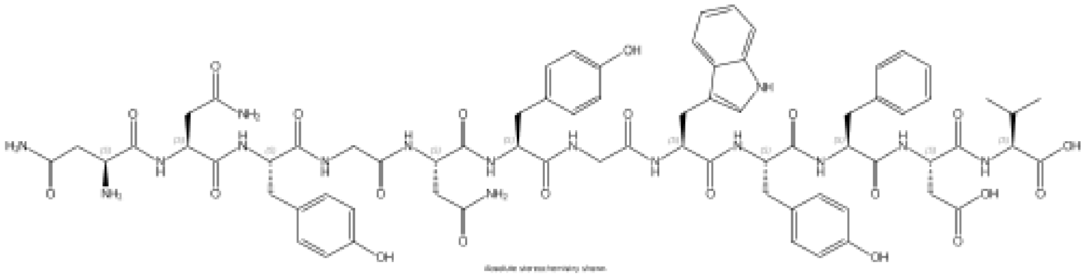
$$\text{Subject Coverage (91\%)} = \frac{\mathit{Alignment\ Length}}{\mathit{Subject\ Length}} = \frac{10}{11}$$

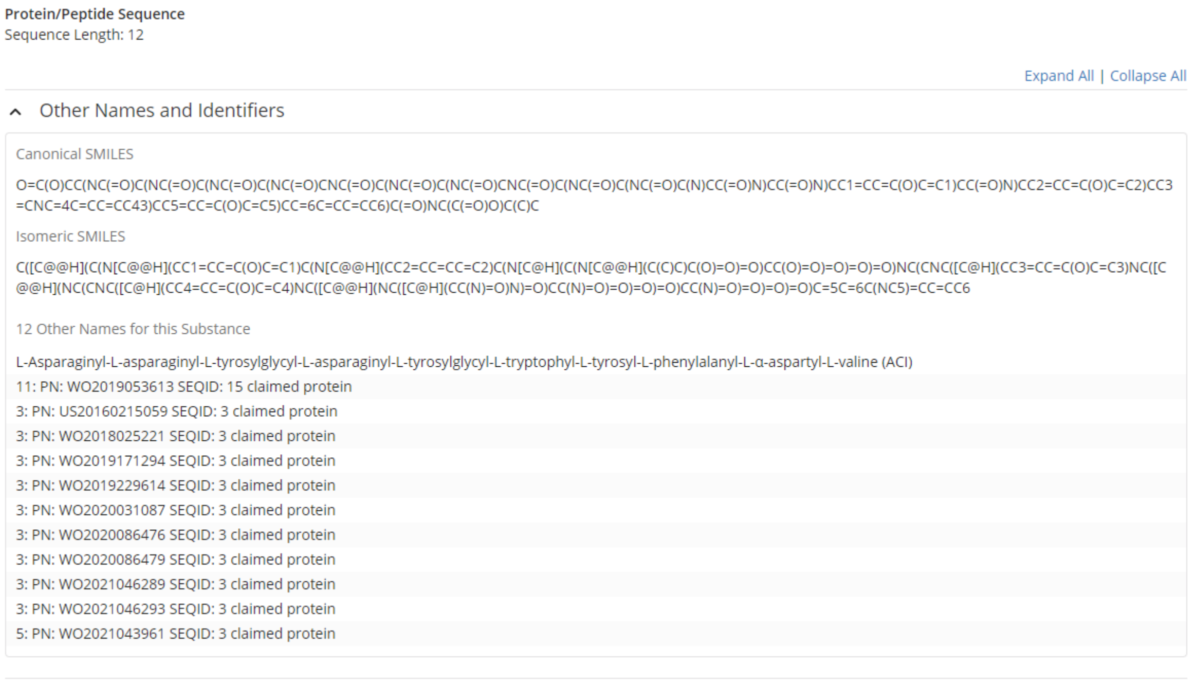
$$\text{Alignment Identity (90\%)} = \frac{\mathit{Number\ of\ Matches}}{\mathit{Alignment\ Length}} = \frac{9}{10}$$

Biosequences search example

CAS RN: 1967777-53-2

NNYGNYGWYFDV





Protein/Peptide Sequence
Sequence Length: 12

Expand All | Collapse All

Other Names and Identifiers

Canonical SMILES

O=C(O)CC(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)CNC(=O)C(NC(=O)C(NC(=O)CNC(=O)C(NC(=O)C(NC(=O)C(N)CC(=O)N)CC(=O)N)CC1=CC=C(O)C=C1)CC(=O)N)CC2=CC=C(O)C=C2)CC3=CNC=4C=CC=CC43)CC5=CC=C(O)C=C5)CC=6C=CC=CC6)C(=O)NC(C(=O)O)C(C)C

Isomeric SMILES

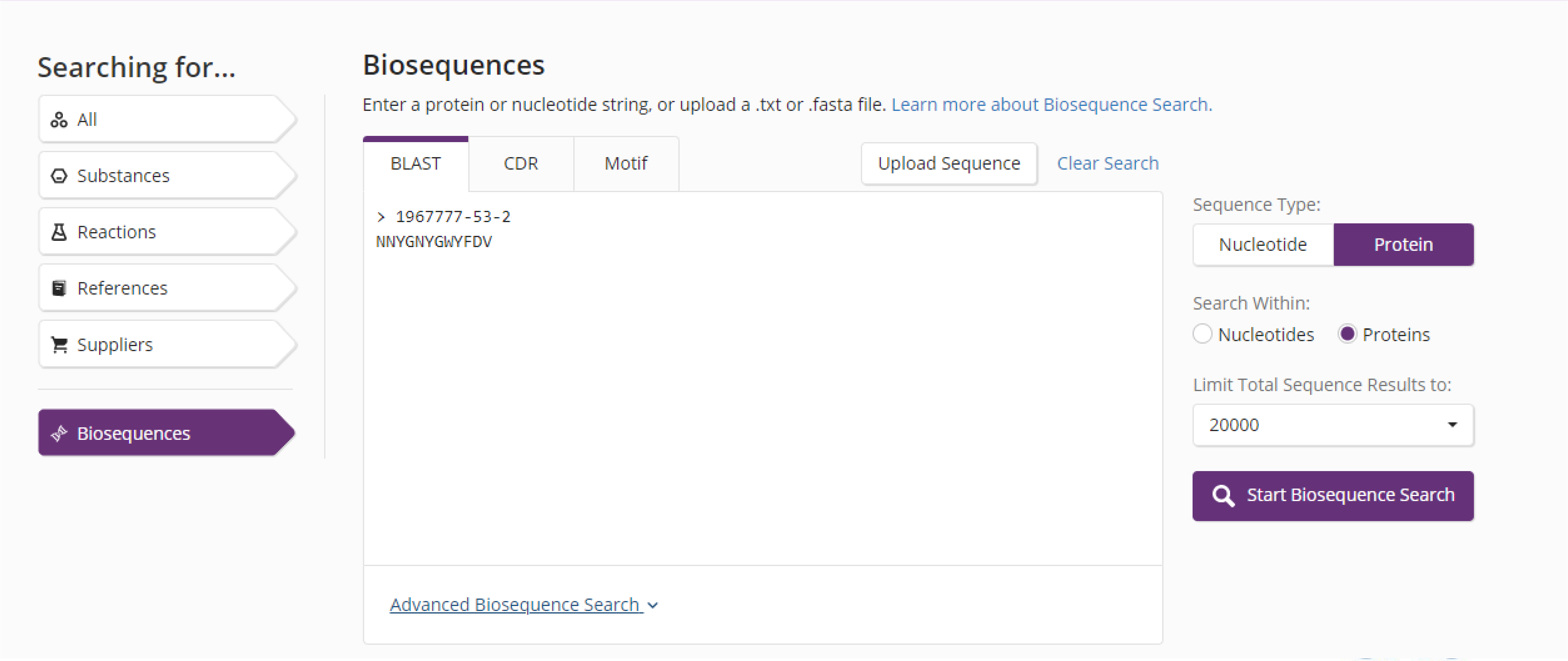
C([C@@H](C(N[C@@H](CC1=CC=C(O)C=C1)C(N[C@@H](CC2=CC=CC=C2)C(N[C@H](C(N[C@@H](C(C)C)C(O)=O)=O)CC(O)=O)=O)=O)=O)NC(CNC([C@H](CC3=CC=C(O)C=C3)NC([C@@H](NC(CNC([C@H](CC4=CC=C(O)C=C4)NC([C@@H](NC([C@H](CC(N)=O)N)=O)CC(N)=O)=O)=O)=O)CC(N)=O)=O)=O)=O)C=5C=6C(NC5)=CC=CC6

12 Other Names for this Substance

L-Asparaginyl-L-asparaginyl-L-tyrosylglycyl-L-asparaginyl-L-tyrosylglycyl-L-tryptophyl-L-tyrosyl-L-phenylalanyl-L-α-aspartyl-L-valine (ACI)
11: PN: WO2019053613 SEQID: 15 claimed protein
3: PN: US20160215059 SEQID: 3 claimed protein
3: PN: WO2018025221 SEQID: 3 claimed protein
3: PN: WO2019171294 SEQID: 3 claimed protein
3: PN: WO2019229614 SEQID: 3 claimed protein
3: PN: WO2020031087 SEQID: 3 claimed protein
3: PN: WO2020086476 SEQID: 3 claimed protein
3: PN: WO2020086479 SEQID: 3 claimed protein
3: PN: WO2021046289 SEQID: 3 claimed protein
3: PN: WO2021046293 SEQID: 3 claimed protein
5: PN: WO2021043961 SEQID: 3 claimed protein

CAS
A division of the American Chemical Society

Flexible search examples



Searching for...

Biosequences

Enter a protein or nucleotide string, or upload a .txt or .fasta file. Learn more about Biosequence Search.

```
> 1967777-53-2
NNYGNYGWYFDV
```

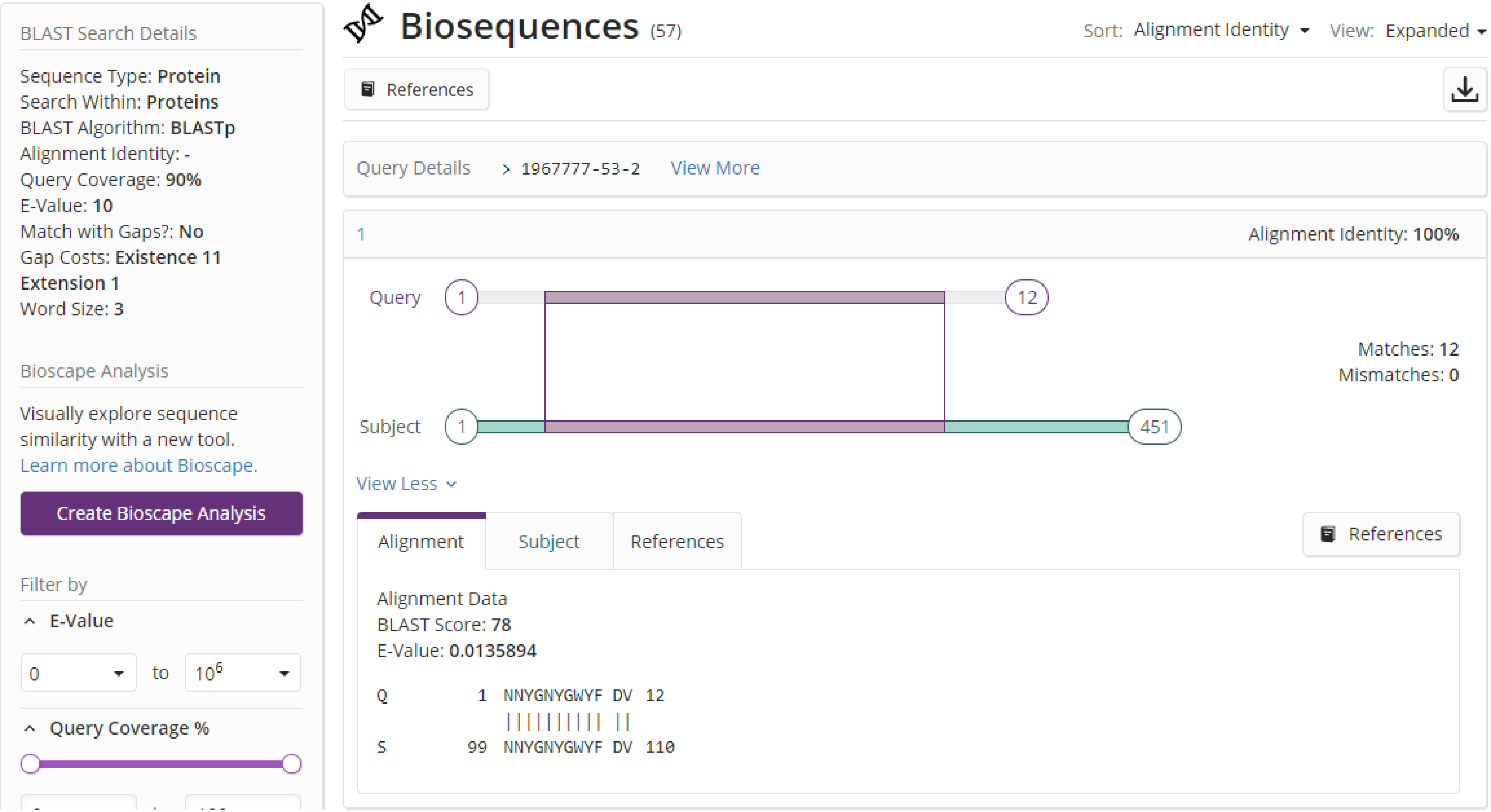
Sequence Type: Nucleotide [✓ Protein]

Search Within: ○ Nucleotides ● Proteins

Limit Total Sequence Results to: 20000

BLAST results

View alignments, filters, links to patents



BLAST Search Details

Sequence Type: **Protein**
Search Within: **Proteins**
BLAST Algorithm: **BLASTp**
Alignment Identity: -
Query Coverage: **90%**
E-Value: **10**
Match with Gaps?: **No**
Gap Costs: **Existence 11 Extension 1**
Word Size: **3**

Bioscape Analysis

Visually explore sequence similarity with a new tool.
Learn more about Bioscape.

Filter by

E-Value

0 to 10^6

Query Coverage %

Biosequences (57)

Sort: Alignment Identity
View: Expanded

Query Details > 1967777-53-2 View More

1

Alignment Identity: **100%**

Matches: **12**
Mismatches: **0**

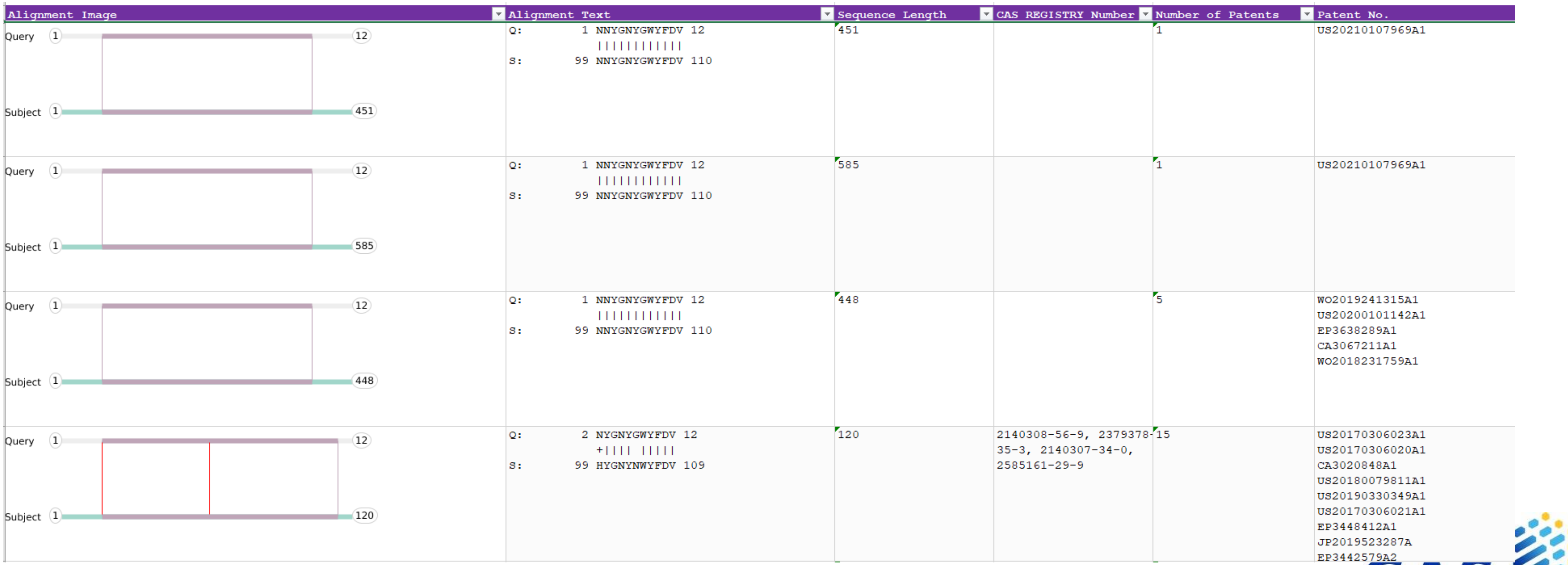
Alignment Data
BLAST Score: **78**
E-Value: **0.0135894**

```
Q        1  NNYGNYGWYF DV  12
            |||||||||| ||
S       99  NNYGNYGWYF DV  110
```

CAS
A division of the
American Chemical Society

BLAST results

Export sequence data to EXCEL



<table>
<tr><th>Alignment Image</th><th>Alignment Text</th><th>Sequence Length</th><th>CAS REGISTRY Number</th><th>Number of Patents</th><th>Patent No.</th></tr>
<tr><td>Query 1 12
Subject 1 451</td><td>Q: 1 NNYGNYGWYFDV 12
||||||||||||
S: 99 NNYGNYGWYFDV 110</td><td>451</td><td></td><td>1</td><td>US20210107969A1</td></tr>
<tr><td>Query 1 12
Subject 1 585</td><td>Q: 1 NNYGNYGWYFDV 12
||||||||||||
S: 99 NNYGNYGWYFDV 110</td><td>585</td><td></td><td>1</td><td>US20210107969A1</td></tr>
<tr><td>Query 1 12
Subject 1 448</td><td>Q: 1 NNYGNYGWYFDV 12
||||||||||||
S: 99 NNYGNYGWYFDV 110</td><td>448</td><td></td><td>5</td><td>WO2019241315A1
US20200101142A1
EP3638289A1
CA3067211A1
WO2018231759A1</td></tr>
<tr><td>Query 1 12
Subject 1 120</td><td>Q: 2 NYGNYGWYFDV 12
+|||| |||||
S: 99 HYGNYNWYFDV 109</td><td>120</td><td>2140308-56-9, 2379378-35-3, 2140307-34-0, 2585161-29-9</td><td>15</td><td>US20170306023A1
US20170306020A1
CA3020848A1
US20180079811A1
US20190330349A1
US20170306021A1
EP3448412A1
JP2019523287A
EP3442579A2</td></tr>
</table>

CAS
A division of the
American Chemical Society

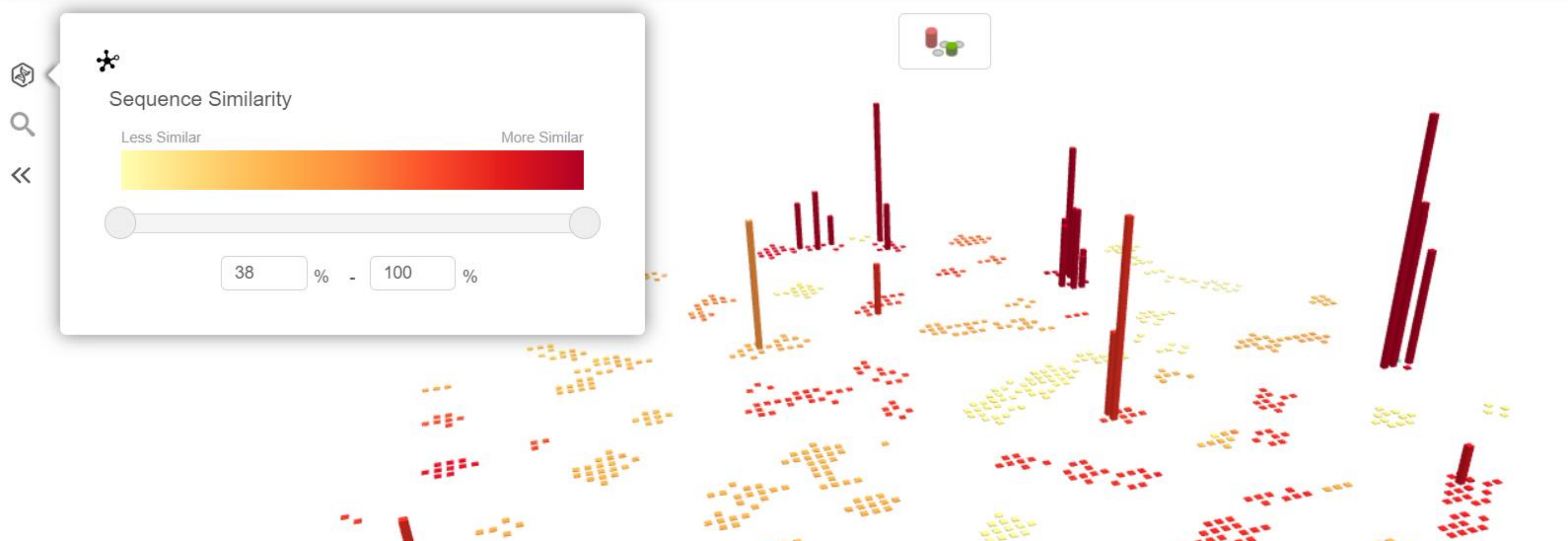
BioScape Visualization

Interactive view of sequence answer set

Position
Similarity to Query
(multiple dimensions)

Color
Similarity measured
by Alignment Identity

Query position
Indicated by
turquoise circle ○

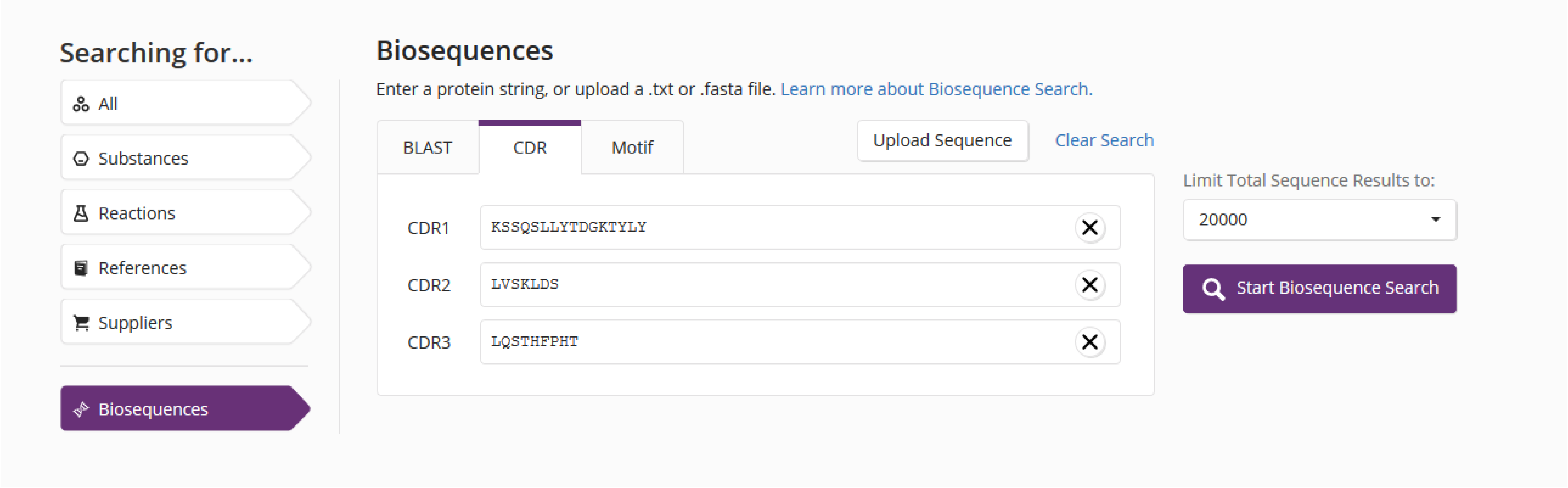


Sequence Similarity

38 % - 100 %

CDR Search

Match Three Regions Within the Protein



Biosequences

Enter a protein string, or upload a .txt or .fasta file. Learn more about Biosequence Search.

CDR1 KSSQSLLYTDGKTYLY

CDR2 LVSKLDS

CDR3 LQSTHFPHT

Limit Total Sequence Results to: 20000

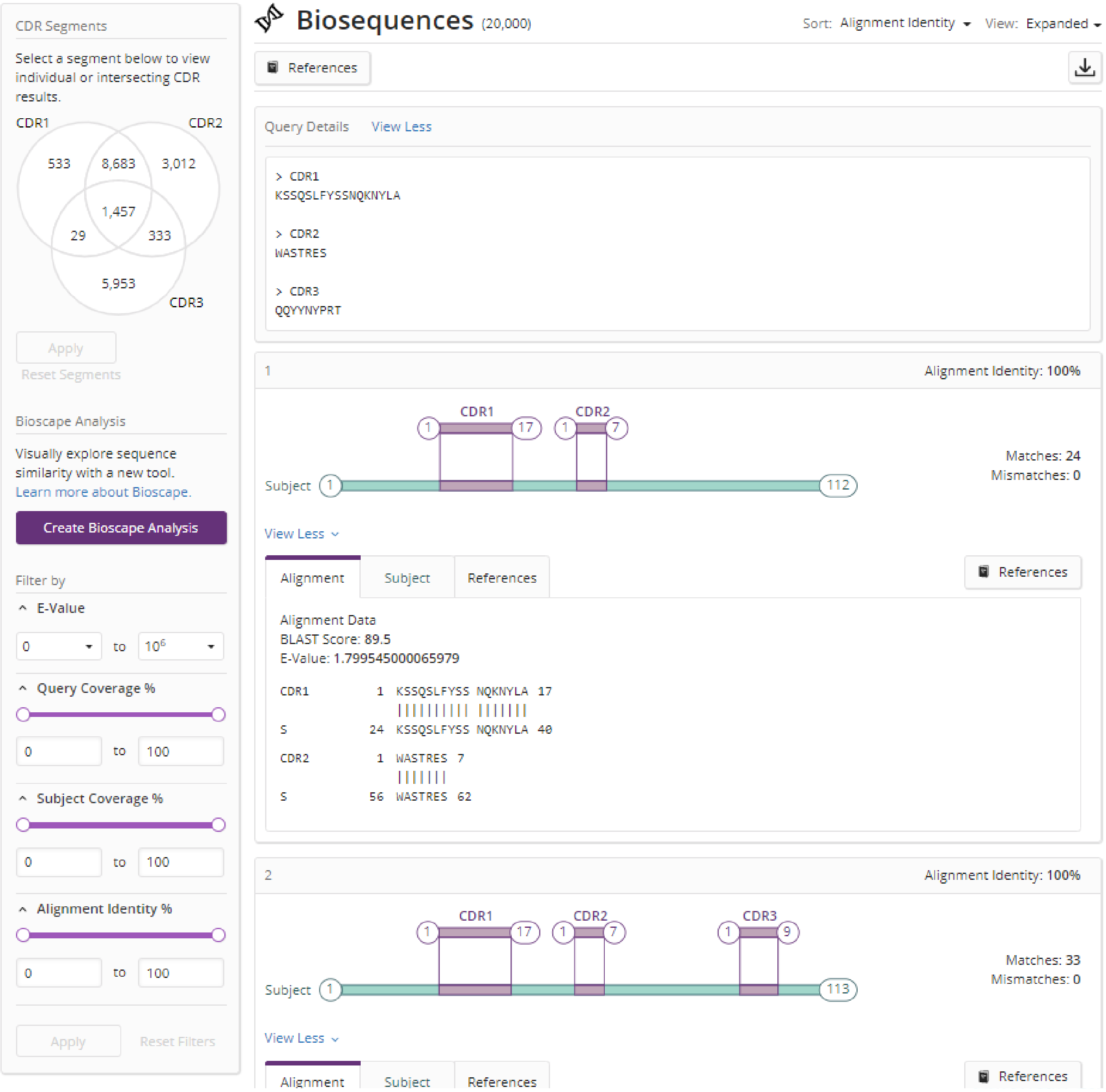
CAS
A division of the
American Chemical Society

CDR Search

Filters Options

- Alignment visualization provides subject coverage, alignment position, and alignment identity at a glance

- Visually identify alignments to one, two, or all three of the CDR queries



CDR Segments

Select a segment below to view individual or intersecting CDR results.

Bioscape Analysis

Visually explore sequence similarity with a new tool. Learn more about Bioscape.

Filter by

E-Value: 0 to 10^6

Query Coverage %: 0 to 100

Subject Coverage %: 0 to 100

Alignment Identity %: 0 to 100

Biosequences (20,000)

Sort: Alignment Identity

View: Expanded

Query Details View Less

```
> CDR1
KSSQSLFYSSNQKNYLA

> CDR2
WASTRES

> CDR3
QQYYNYPRT
```

1 Alignment Identity: 100%

Matches: 24
Mismatches: 0

Alignment Data
BLAST Score: 89.5
E-Value: 1.799545000065979

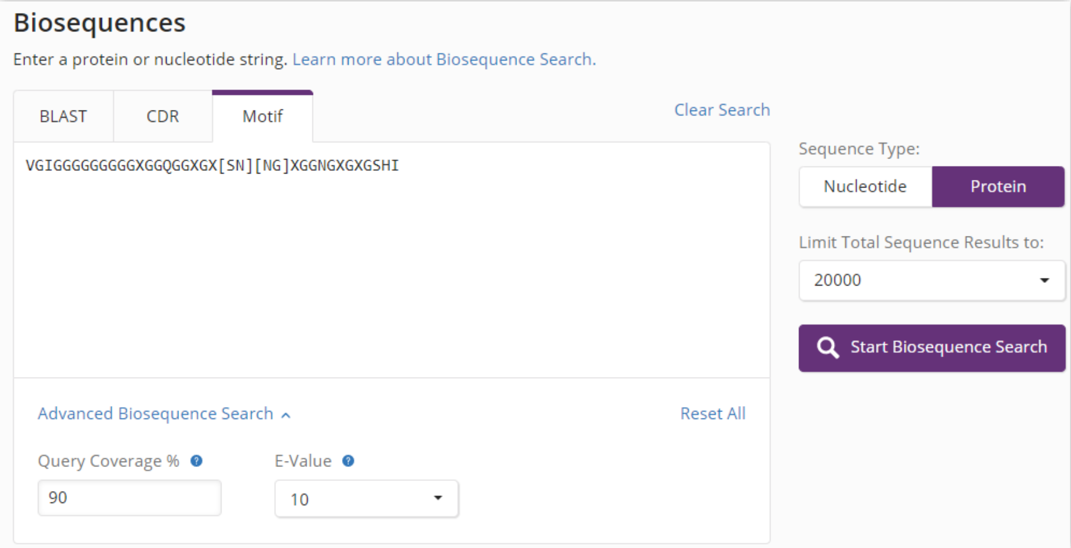
```
CDR1         1  KSSQSLFYSS NQKNYLA  17
                |||||||||| |||||||
S           24  KSSQSLFYSS NQKNYLA  40

CDR2         1  WASTRES  7
                |||||||
S           56  WASTRES  62
```

2 Alignment Identity: 100%

Matches: 33
Mismatches: 0

Motif Searching

A sequence match search



Biosequences

Enter a protein or nucleotide string. Learn more about Biosequence Search.

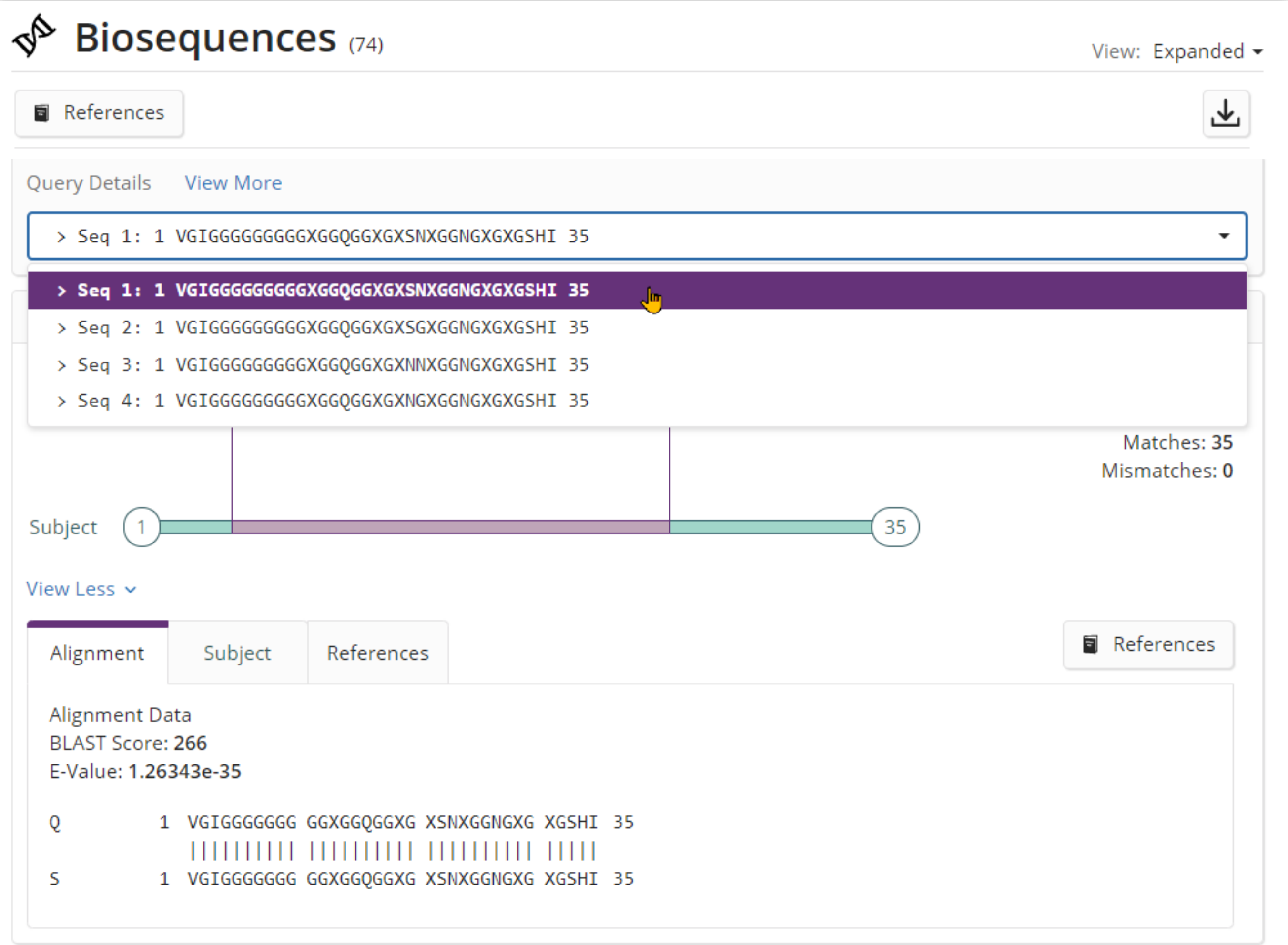
```
VGIGGGGGGGGGXGGQGGXGX[SN][NG]XGGNGXGXGSHI
```

Sequence Type: Nucleotide | Protein

Limit Total Sequence Results to: 20000

Query Coverage %: 90

E-Value: 10



Biosequences (74)

View: Expanded

```
> Seq 1: 1 VGIGGGGGGGGGXGGQGGXGXSNXGGNGXGXGSHI 35
> Seq 1: 1 VGIGGGGGGGGGXGGQGGXGXSNXGGNGXGXGSHI 35
> Seq 2: 1 VGIGGGGGGGGGXGGQGGXGXSGXGGNGXGXGSHI 35
> Seq 3: 1 VGIGGGGGGGGGXGGQGGXGXNNXGGNGXGXGSHI 35
> Seq 4: 1 VGIGGGGGGGGGXGGQGGXGXNGXGGNGXGXGSHI 35
```

Matches: 35
Mismatches: 0

Alignment Data
BLAST Score: 266
E-Value: 1.26343e-35

```
Q         1  VGIGGGGGGG GGXGGQGGXG XSNXGGNGXG XGSHI  35
             |||||||||| |||||||||| |||||||||| |||||
S         1  VGIGGGGGGG GGXGGQGGXG XSNXGGNGXG XGSHI  35
```

[**SN**] in the above query finds sequences which contain either Serine (S) or Asparagine (N) at amino acid position 22.

X will find positive mismatches (+) for any hit amino acid or a match between query and hit X amino acid.

Biosequence Motif Codes

Degenerate Code	Meaning
X	Any amino acid
B	D or N
Z	E or Q
J	I or L

Degenerate Code	Meaning
N	A or C or G or T
R	A or G
Y	C or T
M	A or C
K	G or T
S	C or G
W	A or T
H	A or C or T
B	C or G or T
V	A or C or G
D	A or G or T

Biosequence Motif Codes

Degenerate Code	Meaning
[XYZ]	Any nucleotide or amino acid listed within the square brackets
{m,n}	At least m and maximum n residues length
{n}	Exactly n length
^XYZ$	Search for the exact sequence XYZ