CAS STNext®

核酸・タンパク質配列検索

化学情報協会　情報事業部

本日の内容

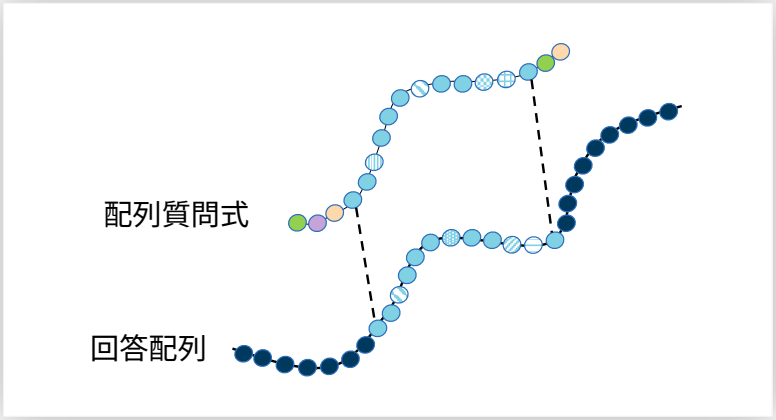
- CAS STNext の配列検索
- 完全配列検索・部分配列検索
- ホモロジー検索
  - REGISTRY
  - GENESEQ
- CAS Sequences

CAS STNEXT の配列検索

配列検索とは

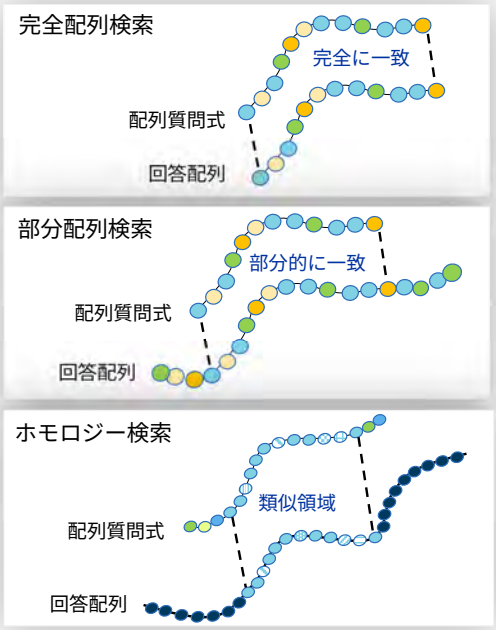
配列検索とは、塩基コード、アミノ酸コードを用いた検索のことである

- 化学物質名検索や構造検索*ではヒットしない核酸・タンパク質が、配列検索でヒットする場合がある
(*水素以外の元素数が 252 以下の場合は構造検索も可能)

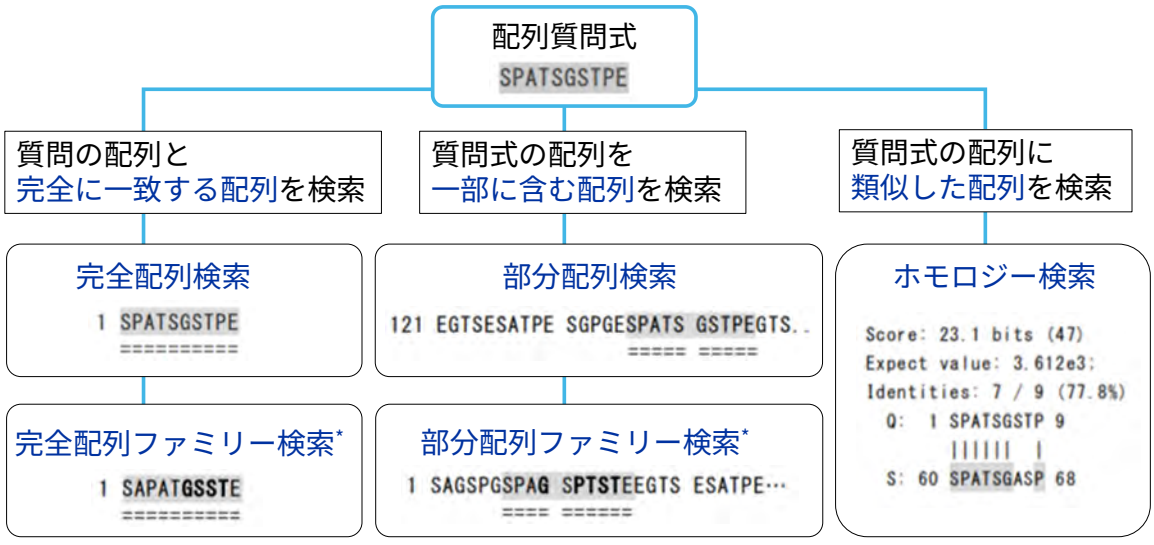


配列検索の種類

- 完全配列検索
  - 質問式と完全に一致する配列を検索
- 部分配列検索
  - 質問式を一部に含む配列を検索
- ホモロジー検索
  - 質問式と類似した配列を検索



配列検索の選択



* ファミリー検索：等価なアミノ酸も含めた配列検索

CAS STNext の配列関連のファイルと機能

ファイル名	内容
REGISTRY*	世界の雑誌や特許から抽出した配列を収録
GENESEQ*	世界の特許から抽出した配列を収録
USGENE	米国特許に記載された配列を収録
PATGENE	PCT 出願に記載された配列を収録
GENBANK	米国国立衛生研究所作成の核酸配列データベース

機能	内容
CAS Sequences*	下記から配列を収録 – REGISTRY ファイル由来の配列 – 73 特許発行機関の特許から抽出した配列 – NCBI 由来の配列

* 本講習会で説明するファイルや機能

検索方法・コマンドの違い

ファイル名	完全配列検索 部分配列検索	ホモロジー検索	
		BLAST	GETSIM
REGISTRY	コマンドで検索* – SEARCH	独立したソフトウェアで検索* – コマンド不要	―
GENESEQ USGENE PATGENE	コマンドで検索* – RUN GETSEQ	コマンドで検索* – RUN BLAST	コマンドで検索* – RUN GETSIM
GENBANK	配列検索はできない		

機能	完全配列検索 部分配列検索	ホモロジー検索			
		BLAST			GETSIM
CAS Sequences – コマンド不要	―	BLAST*	CDR	Motif	―

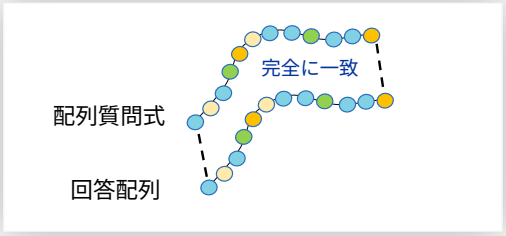
* は本講習会で説明するファイルや機能

完全配列検索・部分配列検索

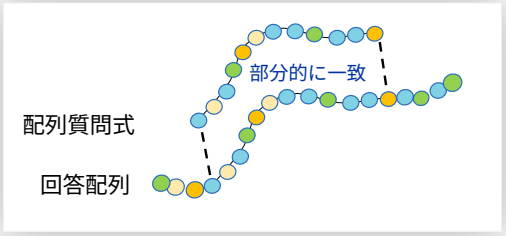
完全配列検索・部分配列検索

質問式と一致する配列を検索したい場合は完全配列検索または部分配列検索を実行する

完全配列検索



部分配列検索



完全配列検索・部分配列検索の検索式

REGISTRY では SEARCH コマンド、GENESEQ では RUN コマンドを使う

- REGISTRY： => S コード/検索フィールド
- GENESEQ： => RUN GETSEQ コード/検索フィールド

	検索タイプ	検索フィールド
核酸	完全配列	/SQEN
	部分配列	/SQSN
タンパク質	完全配列	/SQEP
	完全配列ファミリー*	/SQEFP
	部分配列	/SQSP
	部分配列ファミリー*	/SQSFP

* 等価なアミノ酸も含めた配列がヒットする

塩基コード

核酸の検索では下記の塩基コードを利用できる

- 塩基コードは 5’ 末端から 3’ 末端の順で入力する
- 各ファイルで利用可能な核酸の塩基コードの詳細は、=> HELP NUC 参照。下記以外に曖昧コードも利用可能

塩基	塩基名	完全配列検索でヒット		部分配列検索でヒット	
		REGISTRY	GENESEQ	REGISTRY	GENESEQ
A	アデニン	A	A	A	A
C	シトシン	C	C	C	C
G	グアニン	G	G	G	G
T	チミン (DNA)	T	T, U	T, U	T, U
U	ウラシル (RNA)	U	T, U	T, U	T, U

アミノ酸コード

タンパク質の検索では下記の 20 種類のアミノ酸コードと次のページに示すコードも利用できる

- アミノ酸コードは N 末端 (NH2) から C 末端 (COOH) の順で入力する

1文字コード	3文字コード	アミノ酸名	完全・部分配列検索でヒット	
			REGISTRY	GENESEQ
A	ALA	アラニン	A	
C	CYS	システイン	C	
D	ASP	アスパラギン酸	D	
E	GLU	グルタミン酸	E	
F	PHE	フェニルアラニン	F	
G	GLY	グリシン	G	
H	HIS	ヒスチジン	H	
I	ILE	イソロイシン	I	
K	LYS	リシン	K	
L	LEU	ロイシン	L	

1文字コード	3文字コード	アミノ酸名	完全・部分配列検索でヒット	
			REGISTRY	GENESEQ
M	MET	メチオニン	M	
N	ASN	アスパラギン	N	
P	PRO	プロリン	P	
Q	GLN	グルタミン	Q	
R	ARG	アルギニン	R	
S	SER	セリン	S	
T	THR	トレオニン	T	
V	VAL	バリン	V	
W	TRP	トリプトファン	W	
Y	TYR	チロシン	Y	

アミノ酸コード (続き)

1文字コード	3文字コード	アミノ酸名	完全配列検索でヒット		部分配列検索でヒット	
			REGISTRY	GENESEQ	REGISTRY	GENESEQ
O	PYL	ピロリシン	O	O	O	O
U	SCY	セレノシステイン	U	U	U	U
B	ASX	アスパラギン アスパラギン酸	B	B,D,N	D, N	B,D,N
J	XLE	イソロイシン ロイシン	J	I,L	I,L	I,L
Z	GLX	グルタミン酸 グルタミン	Z	Z,E,Q	E,Q	Z,E,Q
X	XXX	特殊・未定義	X	A-Y, X*	X	A-Y, X*

- 各ファイルで利用できるアミノ酸コードの詳細は、=> HELP AAC 参照
- REGISTRY ファイルでは， X (特殊・未定義のアミノ酸コード) の詳細を => HELP AAU で確認できる
- GENESEQ ファイルの X の詳細は、 => HELP ACC 参照

コードの詳細は「<詳細版> REGISTRY ファイル配列検索」「<詳細版> GENESEQ ファイル配列検索」参照
https://www.jaici.or.jp/stn-ip-protection-suite/cas-stnext/documents/

検索例 - REGISTRY

WTLNSAGYLLGPH を一部に含むタンパク質を REGISTRY ファイルで調査する

[検索条件]

- 配列長は 30 以下で限定する
- 特許でクレームされている配列に限定する

絞り込みに便利な検索フィールド - REGISTRY

検索項目	フィールド	入力例
配列長	/SQL	– 数値検索フィールド => S L1 AND 100/SQL => S L3 AND 10-20/SQL => S L5 AND 200=<SQL
特許情報	/PNTE または /FEAT	– クレームされている配列に限定 => S L2 AND CLAIM?/PNTE – クレームされていない配列に限定 => S L3 AND UNCLAIM?/PNTE
特徴表	/NTE	– クレームされている配列の形態や修飾の情報を収録 => S L1 AND CYCLIC/NTE => S L3 AND MODIFIED/NTE – 特殊・未定義のアミノ酸コードの定義を収録

その他の配列関連検索フィールドおよび特徴表についての詳細は 「<詳細版> REGISTRY ファイル配列検索」 参照
https://www.jaici.or.jp/stn-ip-protection-suite/cas-stnext/documents/

配列レコードを表示するときのポイント - REGISTRY

SQIDE 表示形式などの配列情報が表示される表示形式を使用する

– デフォルトの IDE 表示形式では配列情報は表示されない

表示形式	内容
SCAN	CA 索引名、分子式、クラス識別子、構造図、配列長 * 配列は表示されない
SQIDE	IDE 表示形式 (基本的な物質情報)、配列長 (SQL)、核酸 (NA)、特徴表 (NTE)、特許情報 (PNTE)、1 文字コードの配列データ (SEQ)
SQIDE3	SQIDE と同じ * 配列データ (SEQ3) は 3 文字コードで表示される

検索例 - REGISTRY

```
=> FILE REGISTRY                    ← REGISTRY ファイルに入る

=> S WTLNSAGYLLGPH/SQSP             ← 部分配列検索を実行する
L1          345 WTLNSAGYLLGPH/SQSP

=> S L1 AND SQL=<30                 ← 配列長で限定
L2          123 L1 AND SQL=<30

=> S L2 AND CLAIM?/PNTE             ← クレームで限定
L3           12 L2 AND CLAIM?/PNTE

=> D L3 1-12 SQIDE                  ← 配列情報を表示する
:
L3   ANSWER 1 OF 12  REGISTRY  COPYRIGHT 2024 ACS on STN
RN   1381842-73-4  REGISTRY
ED   Entered STN:  06 Jul 2012
CN   L-Serine, glycyl-L-tryptophyl-L-threonyl-L-leucyl-L-
asparaginyl-L-seryl-L-
     alanylglycyl-L-tyrosyl-L-leucyl-L-leucylglycyl-L-prolyl-L-
histidyl-L-
          :
FS   PROTEIN SEQUENCE; STEREOSEARCH
SQL  30
```

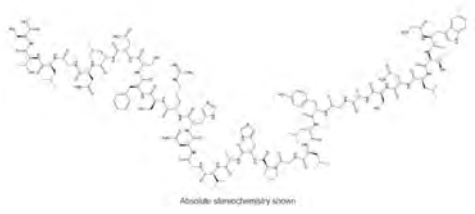
配列長

```
PATENT ANNOTATIONS (PNTE):
Sequence |Patent
Source   |Reference
=========+=============
Not Given|AU2012200046
         claimed SEQID
         |89
---------+-------------
         |US2012156186
            :
SEQ      1 GWTLNSAGYL LGPHAVGNHR SFSDLNGLTS
            ========= ====
HITS AT:   2-14
MF   C139 H209 N41 O43
SR   CA
```

特許番号と配列の記載位置、配列番号

ヒットしたコードには二重線 (=) が付く

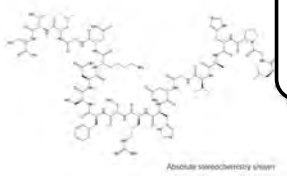
ヒット位置



参考：CAplus ファイルへクロスオーバー検索

```
=> FILE CAPLUS            ← CAplus ファイルに入る

=> S L3                   ← REGISTRY ファイルの L3 を
L4          151 L3          クロスオーバー検索する

=> D 5 BIB ABS HITSEQ     ← BIB ABS HITSEQ 表示形式で表示する

AN   2021:2774180  CAPLUS Full-text
DN   178:51304
TI   Site-selective itaconation of complex peptides by photoredox
     catalysis
AU   Wang, Siyao; Zhou, QingQing; Zhang, Xiaheng; Wang, Ping
CS   Shanghai Key Laboratory for Molecular Engineering of Chiral
     Drugs, School of Chemistry and Chemical Engineering,
           :
SO   Angewandte Chemie, International Edition (2022), 61(5),
     e202111388
     CODEN: ACIEF5; ISSN: 1433-7851
DOI  10.1002/anie.202111388
PB   Wiley-VCH Verlag GmbH & Co. KGaA
DT   Journal; (online computer file)
LA   English
OS   CASREACT 178:51304
```

```
AB   Site-selective peptide functionalization provides a
     straightforward and cost-effective access to diversify
     peptides for biol. studies.  Among many existing non-invasive
     peptide conjugations methodologies, photoredox
           :
IT   136024-41-4
     RL: RCT (Reactant); THU (Therapeutic use); BIOL (Biological
     study); RACT (Reactant or reagent); USES (Uses)
        (synthesis of itaconated peptides and their derivs. Through
        combination of transamination and photoredox conditions)
RN   136024-41-4  CAPLUS
CN   L-Serine, glycyl-L-tryptophyl-L-threonyl-L-leucyl-L-
     asparaginyl-L-seryl-L-alanylglycyl-L-tyrosyl-L-leucyl-L-
           :
SEQ      1 GWTLNSAGYL LGPHAVGNHR SFSDKNGLTS
```



HITSEQ
ヒットした CAS RN®、 そのロールとテキスト説明句、CA 索引名、配列

使用できる記号

ギャップ記号・特殊記号を利用し，配列質問式に柔軟な条件付けを指定することができる

- 部分配列検索 (/SQSN, /SQSP) と部分配列ファミリー検索 (/SQSFP) で利用できる
- REGISTRY ファイルのギャップ記号、特殊記号利用例
  - . (ピリオド) は 1 残基のギャップを指定する
    => S SY...RPG/SQSP → . (ピリオド) に 1 残基が入った配列がヒット
    SYTYYRPG、AFWSYKRLRPG など
  - [ ] は代替残基を指定する
    => S QS[ILM]SSW/SQSP → [ ] 内で指定したいずれかの残基が入った配列がヒット
    QSLSSW、QSISSWLA など
- ギャップ記号、特殊記号は => HELP SQQ 参照

その他記号の利用例は「<詳細版> REGISTRY ファイル配列検索」 参照
https://www.jaici.or.jp/stn-ip-protection-suite/cas-stnext/documents/

配列長の制限値

配列質問式の長さの制限

- REGISTRY ファイル

入力方法	核酸	タンパク質
配列コードを直接入力し一回で検索	250 コード	
QUERY コマンドで作成した検索フィールド付きの質問式の L 番号を & 記号でつなげる	1,000 コード	完全配列 1,000 コード 部分配列 2,400 コード
REGISTRY BLAST ソフトウェア	50,000 コード	

入力方法の詳細は「<詳細版> REGISTRY ファイル配列検索」 参照
https://www.jaici.or.jp/stn-ip-protection-suite/cas-stnext/documents/

- GENESEQ ファイルは => HELP QLMITS 参照

SEQLINK EXACT コマンド - REGISTRY

- REGISTRY ファイルでは同主鎖の配列を持っていても、個別の CAS RN® を持つ場合がある
- SEQLINK EXACT コマンドを使用すると同主鎖で あっても別のレコードになった核酸、タンパク質をまとめることができる
- 配列検索 (完全配列、部分配列、ホモロジー検索) の結果に対しては行う必要はない

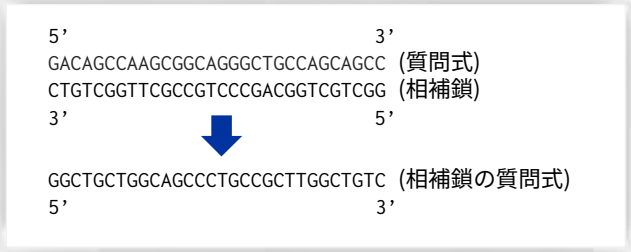
```
=> SEQ  CAS RN®
=> SEQ  L#                    ← L# は配列レコードを含む回答セットの L 番号
```

検索例 - GENESEQ

GACAGCCAAGCGGCAGGGCTGCCAGCAGCC と完全に一致する核酸配列を GENESEQ ファイルで調査する

[検索条件]

- 相補鎖も含める (デフォルト)



- 特許でクレームされている配列に限定する

核酸の相補鎖の検索

- REGISTRY ファイル
  - 相補鎖は自動的には検索されないので、必要に応じて相補鎖も別途検索する

    => S GCCCAAGCTGGC/SQSN ← 入力した配列コード (一本鎖) のみ検索 (デフォルト)

    => S GCCAGCTTGGGC/SQSN ← 相補鎖の配列コードを 5’→3’の順で作成し検索
- GENESEQ ファイル
  - デフォルトで自動的に相補鎖を含めて検索される
  - 相補鎖を含めるかはオプションで変更できる

    => RUN GETSEQ GCCCAAGCTGGC/SQSN ← 相補鎖を含めて検索 (デフォルト)

    => RUN GETSEQ GCCCAAGCTGGC/SQSN -S BOTH (上記と同じ検索)

    => RUN GETSEQ GCCCAAGCTGGC/SQSN -S SIN ← 入力した配列コード (一本鎖) のみ検索

    => RUN GETSEQ GCCCAAGCTGGC/SQSN -S COM ← 入力した配列コードの相補鎖のみ検索

絞り込みに便利な検索フィールド - GENESEQ

検索項目	フィールド	入力例
配列長	/SQL	– 数値検索フィールド => S L1 AND 100/SQL => S L3 AND 10-20/SQL => S L5 AND 200=<SQL
特許中の配列の記載位置	/PSL	– クレームされている配列に限定 => S L2 AND CLAIM?/PSL – クレームされていない配列に限定 => S L3 AND DISCLOSURE/PSL
特徴表	/FEAT	– 配列の特徴を収録 => S L1 AND (DISULFIDE(W)BOND)/FEAT => S L3 AND MRNA/FEAT – 特殊・未定義のアミノ酸コード (X) に対応する情報を収録 =>S (HSE OR HOMOSERINE)/FEAT
WPI レコード番号	/OS	– WPI ファイルのレコード番号を検索 => S 94-151326/OS

その他の配列関連検索フィールドおよび特徴表についての詳細は「<詳細版> GENESEQ ファイル配列検索」参照　https://www.jaici.or.jp/stn-ip-protection-suite/cas-stnext/documents/

配列レコードを表示するときのポイント - GENESEQ

– 配列の概要やアライメント情報など、回答の適合性を確認したい時は、 TRIAL や ALIGN 表示形式を使用する

– 書誌情報や配列に関する詳細な情報を確認したい場合は、BIB や ALL 表示形式を使用する

表示形式	内容
TRIAL	レコード番号 (AN)、標題 (TI)、分子式タイプ (MTY)、配列の説明 (DESC)、キーワード (KW)、配列長 (SQL)
ALIGN	完全配列、部分配列検索結果の場合 – 全配列情報　(ヒット位置には二重下線が付く) ホモロジー検索結果の場合 – ヒットした配列と質問配列のアライメント情報 (一致/不一致/ギャップ)
BIB	書誌情報、配列の説明 (DESC)、クロスレファレンス (CR)
ALL	書誌情報、抄録、配列情報およびキーワード (KW)

検索例 - GENESEQ

```
=> FILE GENESEQ                          ←GENESEQ ファイルに入る

=> RUN GETSEQ GACAGCCAAGCGGCAGGGCTGCCAGCAGCC/SQEN
L1     RUN STATEMENT CREATED
L1             3 GACAGCCAAGCGGCAGGGCTGCCAGCAGCC/SQEN

=> S L1 AND CLAIM?/PSL                   ←クレームで限定
L2             3 L1 AND CLAIM?/PSL
```

核酸の配列検索では、デフォルトで自動的に相補鎖を含めて検索される

```
=> D L2 1-3 TRIAL ALIGN     ←TRIAL ALIGN 表示形式を用いて標題、
                                配列長、ヒットしたコードを確認する
L2     ANSWER 1 OF 3 GENESEQ COPYRIGHT 2024 CLARIVATE on STN.
AN     BKQ82934   GENESEQ
TI     Diagnosing subject with benign, pre-malignant, or malignant
       hyperproliferative cell used for detecting cancer cell in
       subject, involves detecting presence, absence, and/or
       quantity of non-coding RNA or its functional fragment in
       sample.
MTY    cDNA
DESC   Human orphan non-coding RNA (oncRNA), SEQ ID 6810.
KW     cancer; cytostatic; diagnostic test; oncRNA; orphan
       noncoding RNA; rna detection; ss; therapeutic; tumor marker
SQL    30
ALIGN
  ALIGNMENT FROM L-NUMBER L1
  Sequence Length: 30;
  Strand: Plus / Minus;
  Hits at: 30-1
     30 GACAGCCAAG CGGCAGGGCT GCCAGCAGCC
        ========== ========== ==========
```

Minus の場合は相補鎖でヒットしている

ヒット位置

ヒットしたコードには二重線 (=) が付く

検索例 - GENESEQ

```
L2     ANSWER 2 OF 3 GENESEQ COPYRIGHT 2024 CLARIVATE on STN.
AN     BKQ82933   GENESEQ
TI     Diagnosing subject with benign, pre-malignant, or malignant
       hyperproliferative cell used for detecting cancer cell in
       subject, involves detecting presence, absence, and/or
       quantity of non-coding RNA or its functional fragment in
       sample.
MTY    cDNA
DESC   Human orphan non-coding RNA (oncRNA), SEQ ID 6809.
KW     cancer; cytostatic; diagnostic test; oncRNA; orphan
       noncoding RNA; rna detection; ss; therapeutic; tumor marker
SQL    30
ALIGN
  ALIGNMENT FROM L-NUMBER L1
  Sequence Length: 30;
  Strand: Plus / Plus;
  Hits at: 1-30
     1 GACAGCCAAG CGGCAGGGCT GCCAGCAGCC
       ========== ========== ==========
```

入力した質問式のコードでヒット

```
L2     ANSWER 3 OF 3 GENESEQ COPYRIGHT 2024 CLARIVATE on STN.
AN     BKQ82932   GENESEQ
TI     Diagnosing subject with benign, pre-malignant, or malignant
       hyperproliferative cell used for detecting cancer cell in
       subject, involves detecting presence, absence, and/or
       quantity of non-coding RNA or its functional fragment in
       sample.
MTY    RNA
DESC   Human orphan non-coding RNA (oncRNA), SEQ ID 6808.
KW     cancer; cytostatic; diagnostic test; oncRNA; orphan
noncoding RNA; rna
       detection; ss; therapeutic; tumor marker
SQL    30
ALIGN
  ALIGNMENT FROM L-NUMBER L1
  Sequence Length: 30;
  Strand: Plus / Plus;
  Hits at: 1-30
     1 GACAGCCAAG CGGCAGGGCU GCCAGCAGCC
       ========== ========== ==========
```

検索例 - GENESEQ

```
=> D L2 1 BIB ALIGN          ← 書誌情報とヒットしたコードを
                               組み合わせて表示する
L2      ANSWER 1 OF 3 GENESEQ COPYRIGHT 2024 CLARIVATE on STN.
AN      BKQ82934    GENESEQ  ED 20220325 UP 20220708
        DED 20220324 DUPD 20220707 Full-text
TI      Diagnosing subject with benign, pre-malignant, or malignant
        hyperproliferative cell used for detecting cancer cell in
        subject, involves detecting presence, absence, and/or
        quantity of non-coding RNA or its functional fragment in
        sample.
IN      Goodarzi H
PA      UNIV CALIFORNIA (REGC)
LA      English
DT      Patent
PI      WO 2022040106    A2         20220224
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 2021-US46186          20210816
PRAI    US 2020-66269P           20200816
FS      NUCLEIC; NS
OS      2022-27666X [020]
```

BIB
書誌情報、配列の説明 (DESC)

```
MTY     cDNA
PSL     Claim 64; SEQ ID NO 6810; 916pp
DESC    Human orphan non-coding RNA (oncRNA), SEQ ID 6810.
ALIGN
  ALIGNMENT FROM L-NUMBER L1
  Sequence Length: 30;
  Strand: Plus / Minus;
  Hits at: 30-1
     30 GACAGCCAAG CGGCAGGGCT GCCAGCAGCC
        ========== ========== ==========
```

特許中の配列の記載位置

ALIGN

ヒットしたコードには
二重線 (=) が付く

実習 1

下記の配列を含むタンパク質を REGISTRY ファイルと GENESEQ ファイルで調べる

CRHKPMRTVTNFYIANLAATDVTFLLCCVPFTALLYPLPGWVLGDFMCKFVNYI

[検索条件]

- 400 以下の配列長でクレームに記載された配列に限定する

検索	REGISTRY	GENESEQ	参照スライド
タンパク質の部分配列検索	=> **S** 質問式/SQSP	=> **RUN** GETSEQ 質問式/SQSP	11
配列長で限定	SQL 検索フィールドを利用する		16, 25
クレームに限定	CLAIM?/PNTE	CLAIM?/PSL	
表示形式	SQIDE	BIB ALIGN	17, 26

実習１の回答

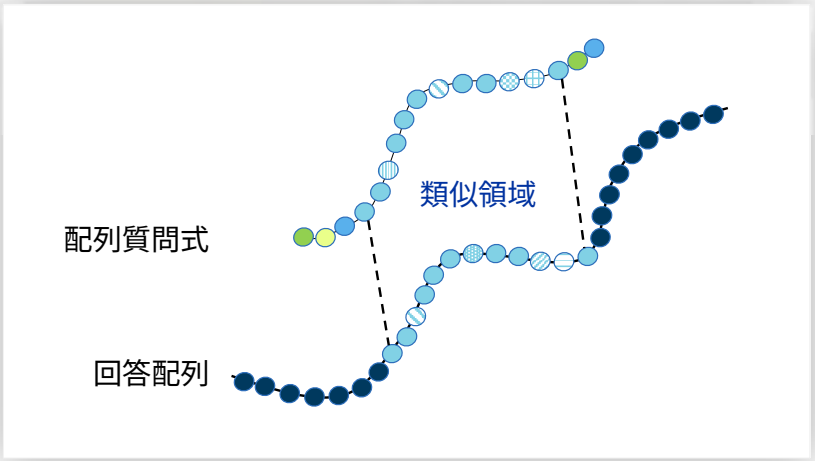
```
=> FILE REGISTRY                        ← REGISTRY ファイルに入る
                                        ↓ 部分配列検索を実行する
=> S CRHKPMRTVTNFYIANLAATDVTFLLCCVPFTALLYPLPGWVLGDFMCKFVNYI/SQSP
L1    46 CRHKPMRTVTNFYIANLAATDVTFLLCCVPFTALLYPLPGWVLGDFMCKFVNYI/SQSP

=> S L1 AND 400>=SQL                    ← 配列長で限定
      36441342 400>=SQL
L2          46 L1 AND 400>=SQL

=> S L2 AND CLAIM?/PNTE                 ← クレームで限定
      17169830 CLAIM?/PNTE
L3          26 L2 AND CLAIM?/PNTE

=> D SQIDE 1-26                         ← 配列情報を表示する

L3   ANSWER 1 OF 26  REGISTRY  COPYRIGHT 2024 ACS on STN
RN   1070931-36-0  REGISTRY
ED   Entered STN:  05 Nov 2008
CN   G protein-coupled receptor (human gene GPR54)  (CA INDEX NAME)
OTHER NAMES:
CN   511: PN: US20080260744 SEQID: 514 claimed protein
FS   PROTEIN SEQUENCE
SQL  398
  :
```

```
=> FILE GENESEQ          ← GENESEQ ファイルに入る

=> RUN GETSEQ            ← 部分配列検索を実行する
CRHKPMRTVTNFYIANLAATDVTFLLCCVPFTALLYPLPGWVLGDFMCKFVNYI/SQSP
:
GENESEQ
Query time:          211
L4    RUN STATEMENT CREATED
L4          67 CRHKPMRTVTNFYIANLAATDVTFLLCCVPFTALLYPLPGWVLGDFMCKF
               VNYI/SQSP

=> S L4 AND 400>=SQL     ← 配列長で限定
      56076893 400>=SQL
L5          66 L4 AND 400>=SQL

=> S L5 AND CLAIM?/PSL   ← クレームで限定
      36962361 CLAIM?/PSL
L6          27 L5 AND CLAIM?/PSL

=> D BIB ALIGN 1-27      ← 書誌情報とヒットしたコードを
                           組み合わせて表示する
L6     ANSWER 1 OF 27 GENESEQ COPYRIGHT 2024 CLARIVATE on STN.
AN     BOW29084   GENESEQ  ED 20240426 UP 20240426
          :
```


ホモロジー検索

ホモロジー検索

配列質問式と類似した配列を検索したい場合はホモロジー検索を実行する



ホモロジー検索のプログラム

CAS STNext では BLAST と GETSIM の 2 種類のホモロジー検索を実行できる

プログラム	BLAST	GETSIM
概要	Basic Local Alignment Search Tool 最もよく利用されている配列検索プログラム。他のプログラムに比べて高速処理できる。ギャップをあまり考慮しないため、検出感度や選択性が低いと考えられがちだが、実際には他と比べてそれほど遜色はない。	FASTA 系列のプログラム データベース中のすべての配列との間で忠実にアラインメントを行ってホモロジースコアを算定する。BLAST ホモロジー検索で回答が得られない場合でも、GETSIM ホモロジー検索で回答が得られることがある。
処理速度	速い	遅い
類似性の高い配列	○	○
類似性の低い配列	△	○
比較方法	短い部分配列を比較 ギャップはあまり考慮されない	配列全体を比較 ギャップも考慮される
感受性	デフォルトの設定では低い	高い
データベース	REGISTRY，GENESEQ，USGENE，PATGENE, CAS Sequences	GENESEQ, USGENE, PATGENE

REGISTRY ホモロジー検索

REGISTRY BLAST 検索の流れ

準備
① ソフトウェアのインストール

↓

BLAST 検索 (独立したソフトウェアで実施)
② 検索
- ソフトウェアを起動
- 配列質問式の入力
- 検索タイプの選択
- パラメータの設定

③ 回答表示
- 結果の確認

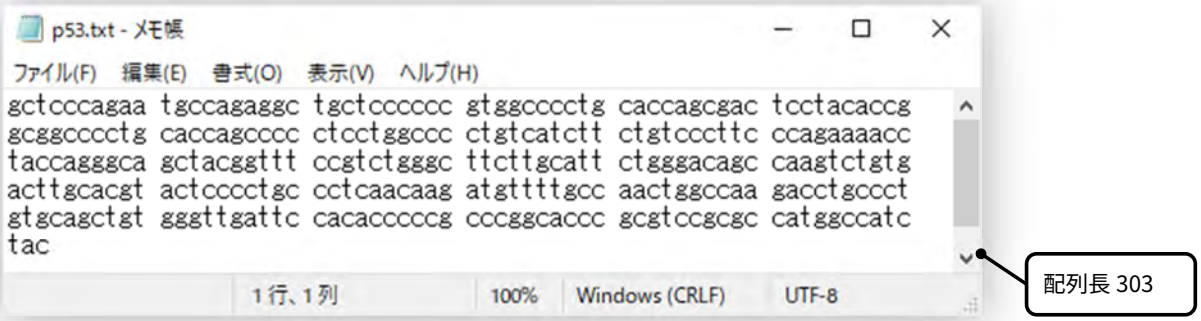
④ STN 移行のための準備
- スクリプトファイル保存
- アライメントデータ保存

↓

CAS STNext 移行
⑤ 検索と表示
- スクリプトを実行、回答を表示
- レポート作成 (オプション)

検索例 – REGISTRY BLAST

下記のがん抑制遺伝子 p53 に類似する配列を REGISTRY BLAST 検索で調査する



① 準備：ソフトウェアのインストール

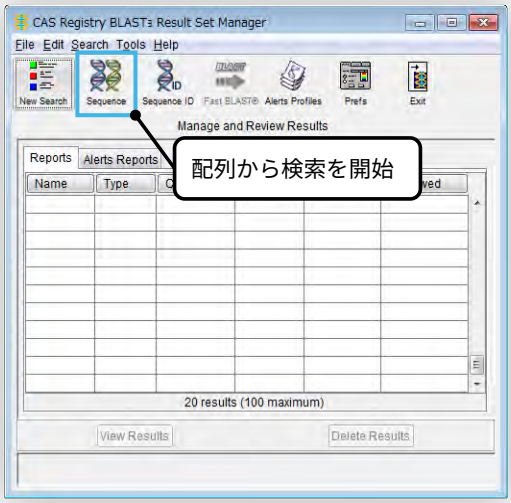
REGISTRY BLAST はソフトウェアで実行するため、ソフトウェアをインストールする

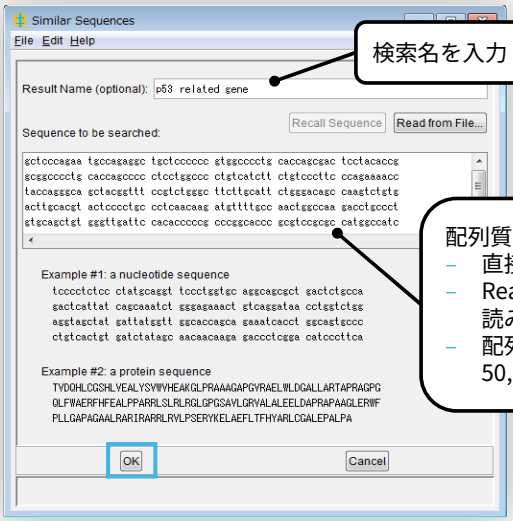
- ダウンロードサイトにアクセスしてソフトウェア (.exe) をダウンロードする
  https://www.stn.org/stn/downloads/blast-download.html

- .exe ファイルを実行してインストールする

② REGISTRY BLAST 検索

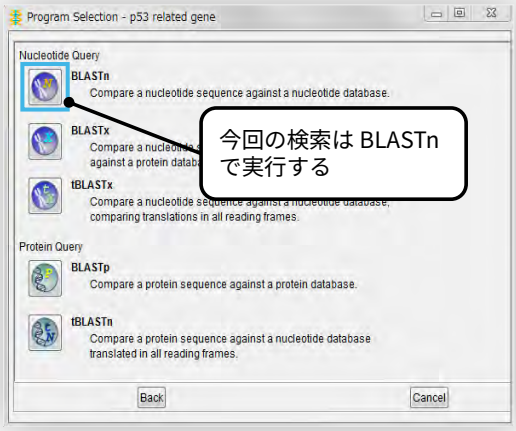
ソフトウェアをクリックして REGISTRY BLAST を起動する。次に、Sequence ボタンをクリックして配列質問式を入力する (スライド 35 の配列を使用)



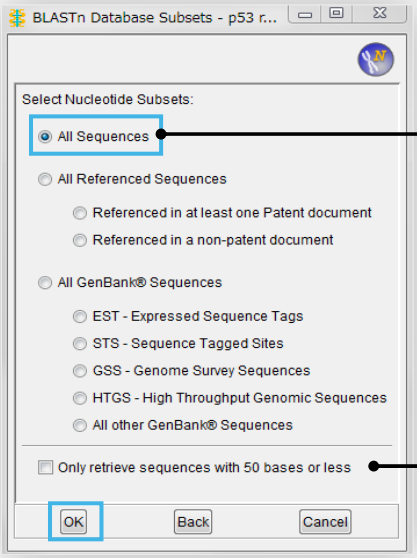


② REGISTRY BLAST 検索

検索タイプを選択する



検索対象を選択する



参考：検索タイプ

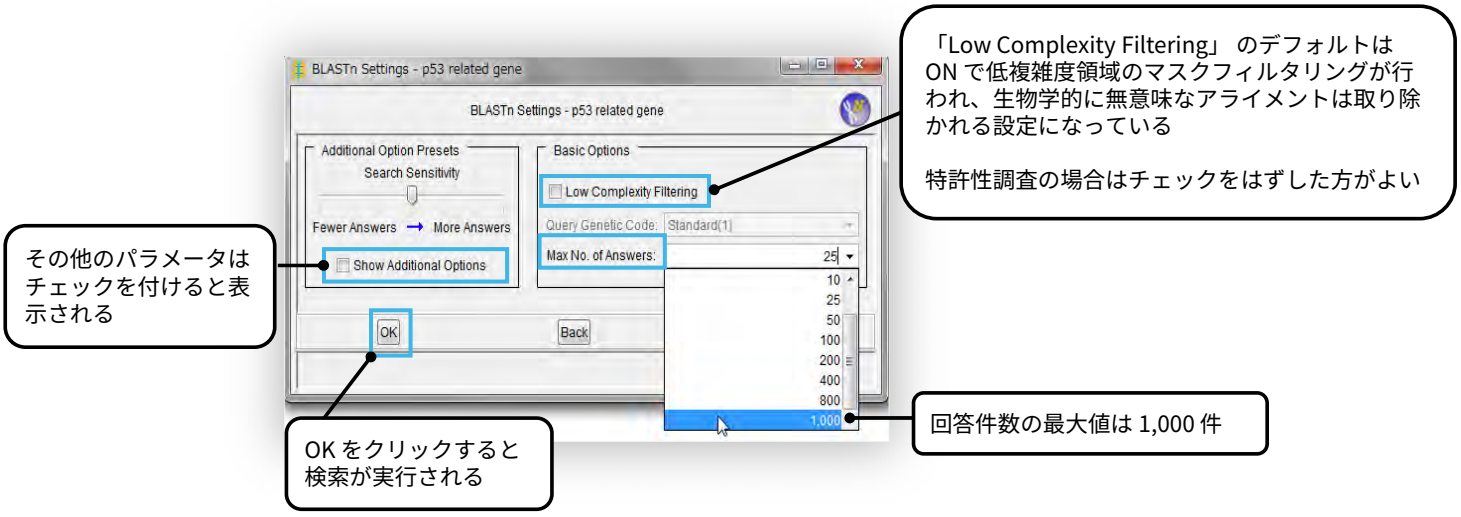
REGISTRY BLAST 検索の検索タイプ

検索タイプ	検索機能	配列質問式	回答
BLASTn	塩基配列の質問式に類似した塩基配列を検索	塩基配列	塩基配列
tBLASTn	データベース中の塩基配列をアミノ酸に翻訳した配列の中からアミノ酸配列の質問式に類似した配列を検索	アミノ酸配列	塩基配列
tBLASTx	塩基配列の質問式をアミノ酸配列に翻訳して、これに類似したアミノ酸配列に翻訳された塩基配列を検索	塩基配列	塩基配列
BLASTp	アミノ酸配列の質問式に類似したアミノ酸配列を検索	アミノ酸配列	アミノ酸配列
BLASTx	塩基配列の質問式をアミノ酸配列に翻訳して、これに類似したアミノ酸配列を検索	塩基配列	アミノ酸配列

核酸検索時は相補鎖も含めて検索される

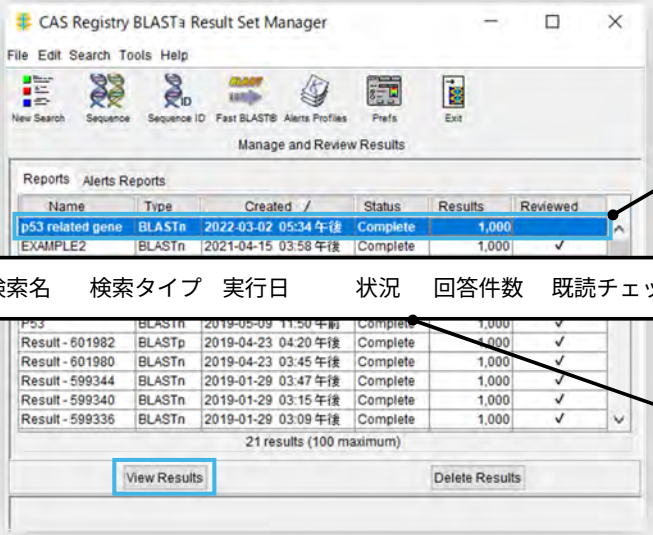
② REGISTRY BLAST 検索

パラメータと回答件数の最大値を設定する



③回答表示

検索完了後、結果を表示する



検索名　検索タイプ　実行日　状況　回答件数　既読チェック

Results は、1,000件×100 セットが最大
(101個目の回答セットを作成する際は、
最も古い回答セットが削除される)

Status (状況) 表示
- Running　：実行中
- Complete　：完了
- Queued　：実行待ち
- Failed　：失敗
  (Failed になった場合は再度検索する)

③回答表示

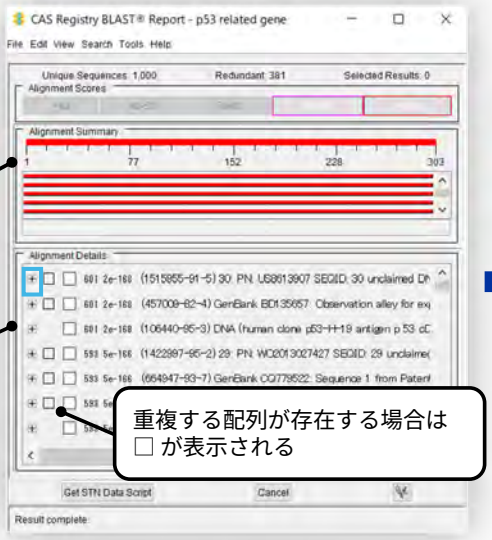
回答はスコア値の高い順に表示される

件数情報

スコア値
分布

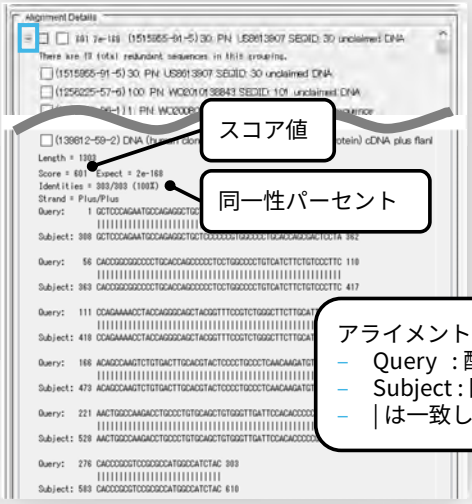
アライメント
の概略

アライメント
の詳細



重複する配列が存在する場合は
☐ が表示される

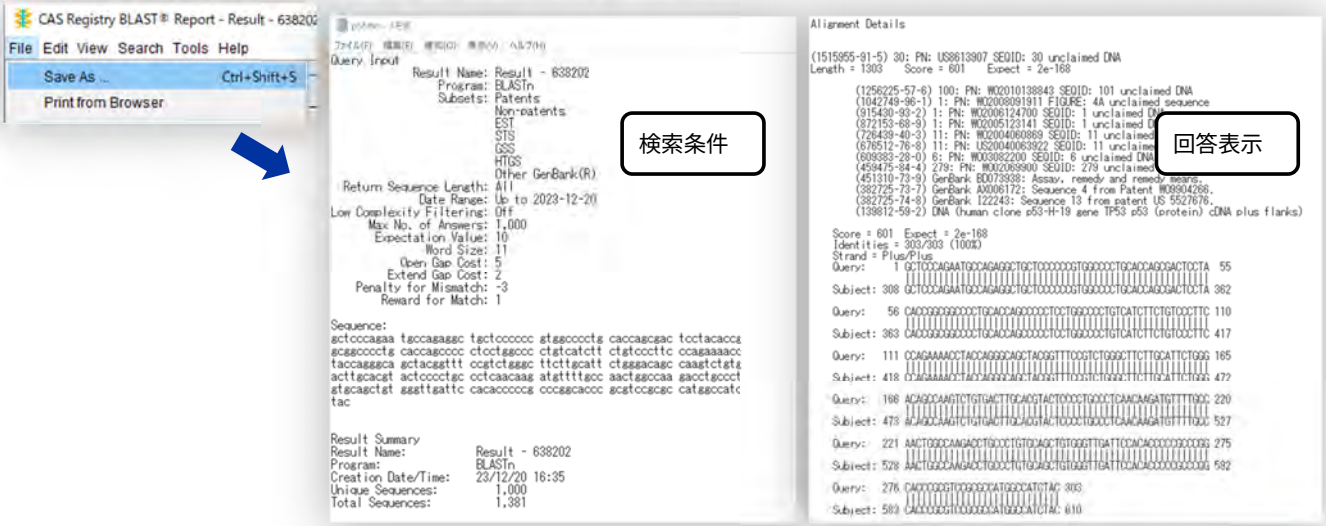
＋ をクリックすると詳しいアライメント情報が
表示される



アライメント表示
- Query　：配列質問式
- Subject：回答の配列
- |　は一致しているコード

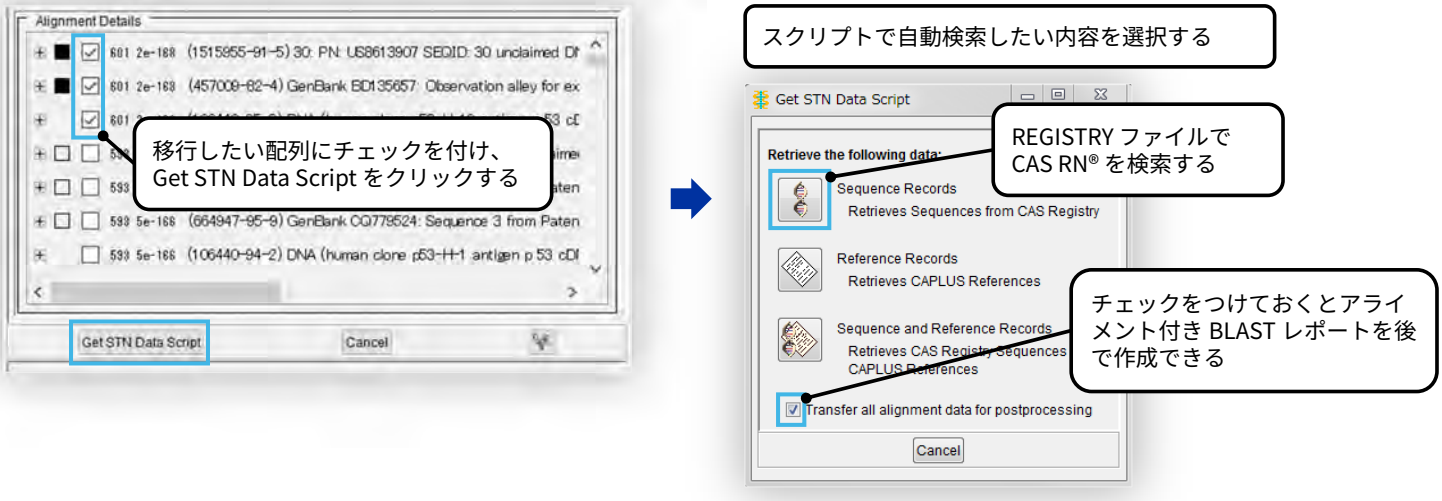
③ 回答表示

REGISTRY BLAST の回答はテキスト形式で保存可能



④ CAS STNext 移行のための準備

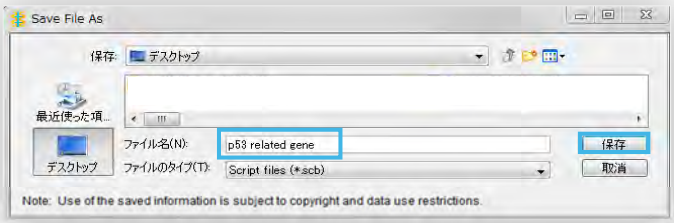
得られた配列を CAS STNext へ移行するために Script を作成する



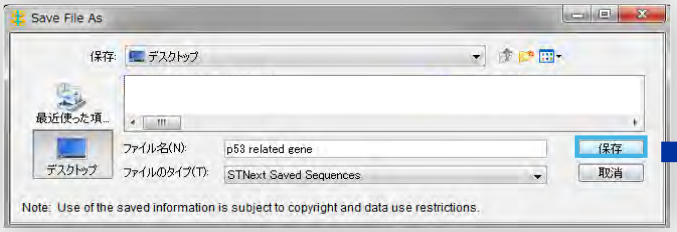
④ CAS STNext 移行のための準備

スクリプトファイルとアライメントデータを保存する

STNext へ移行するための Script (.scb) ファイルを保存する



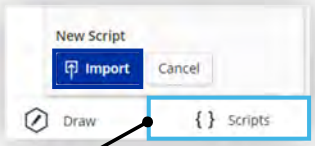
アライメント付きレポートを作成するためにアライメントデータ (.xss) を保存する (オプション)



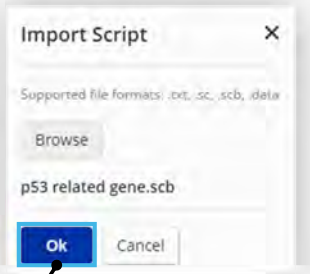
ファイルを保存したら REGISTRY BLAST を終了する

⑤ CAS STNext での検索と表示

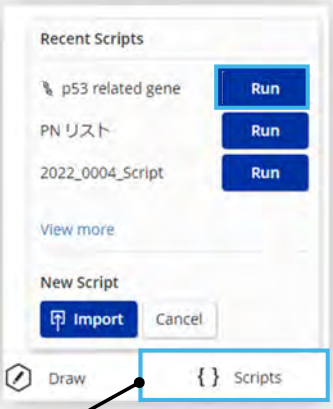
CAS STNext に接続し、Script を実行する (REGISTRY ファイルで配列レコードを検索する)



CAS STNext の画面右下の Script から Import を選択する



保存したファイル (.scb) を選択し OK をクリック



再度，Script をクリックしてインポートしたファイルの Run をクリックする

⑤ CAS STNext での検索と表示

自動的に REGISTRY ファイルに入り配列レコードが検索される

⑤ CAS STNext での検索と表示

回答を SQIDE 表示形式などで表示する

```
=> D L3 SQIDE 1-16                    ← 配列情報を表示する

L3   ANSWER 1 OF 16  REGISTRY  COPYRIGHT 2022 ACS on STN
RN   1515955-91-5  REGISTRY
ED   Entered STN:  09 Jan 2014
CN   30: PN: US8613907 SEQID: 30 unclaimed DNA  (CA INDEX NAME)
FS   NUCLEIC ACID SEQUENCE
SQL  1303                                   配列長
NA   292 a   403 c   348 g   260 t

PATENT ANNOTATIONS (PNTE):
Sequence |Patent
Source   |Reference
=========+==========
Not Given|US8613907                  特許情報 (特許番号、配列の
         |unclaimed                  記載位置、配列番号)
         |SEQID 30

SEQ      1 gtccaggagc aggtagctgc tgggctccgg ggacactttg cgttcgggct
        51 gggagcgtgc tttccacgac ggtgacacgc ttccctggat tggcagccag
       101 actgccttcc gggtcactgc catggaggag ccgcagtcag atcctagcgt
       151 cgagcccccc ctgagtcagg aaacattttc agacctatgg aaactacttc
       201 ctgaaaacaa cgttctgtcc cccttgccgt cccaagcaat ggatgatttg
                :
```

```
                :
      1101 atttcaccct tcagatccgt gggcgtgagc gcttcgagat gttccgagag
      1151 ctgaatgagg ccttggaact caaggatgcc caggctggga aggagccagg
      1201 ggggagcagg gctcactcca gccacctgaa gtccaaaaag ggtcagtcta
      1251 cctcccgcca taaaaaactc atgttcaaga cagaagggcc tgactcagac
      1301 tga                                          配列長

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
MF   Unspecified
CI   MAN
SR   CA
LC   STN Files:   CA, CAPLUS, TOXCENTER, USPAT2, USPATFULL

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

⑤ CAS STNext での検索と表示

CAplus ファイルにクロスオーバーし、配列に関する文献情報を表示する

```
=> FILE CAPLUS            ← CAplus ファイルに入る

=> S L3                   ← REGISTRY ファイルで得られた L 番号を
L4          33 L3           クロスオーバー検索する

=> D BIB HITRN 1-33       ← 書誌情報 (BIB) とヒットした CAS RN® を
                            表示する

L4   ANSWER 1 OF 33  CAPLUS  COPYRIGHT 2023 ACS on STN
AN   2013:1974020  CAPLUS Full-text
DN   160:94599
TI   Compositions that inhibit proliferation of cancer cells using
     inhibitors of laminin 5 interactions with integrin receptors
           :
PI
     PATENT NO.         KIND  DATE      APPLICATION NO.     DATE
     ---------------    ----  --------  ------------------  --------
     US 8613907          B2   20131224  US 2003-392113      20030317
     US 20030224993      A1   20031204
           :
PRAI US 2000-60239705     P    20001012
           :
```

```
OS   CASFORMULTNS 2013:1974020
IT   1515955-91-5
     RL: PRP (Properties)
        (unclaimed nucleotide sequence; compns. that inhibit
        proliferation of cancer cells using inhibitors of laminin 5
        interactions with integrin receptors)
           :
```

ヒットした配列の CAS RN®

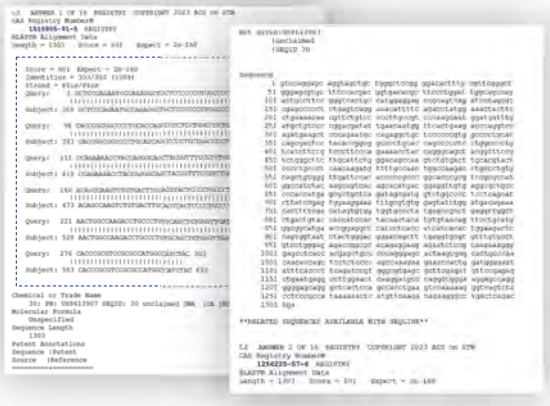
オンラインや Transcript ではアライメント情報は表示できないが、レポートにするとアライメントを組み込める。後述のアライメント付きレポートを作成する場合は、HITRN 表示形式で回答を出力しておくとよい。

他にも HITSEQ 表示形式が利用可能で、ヒットした配列の CAS RN® 自身の配列をオンラインや Transcript で表示できる。アライメント付きレポートを作成する場合は、アライメントが重複表示になるため HITRN 表示形式のほうがよい。

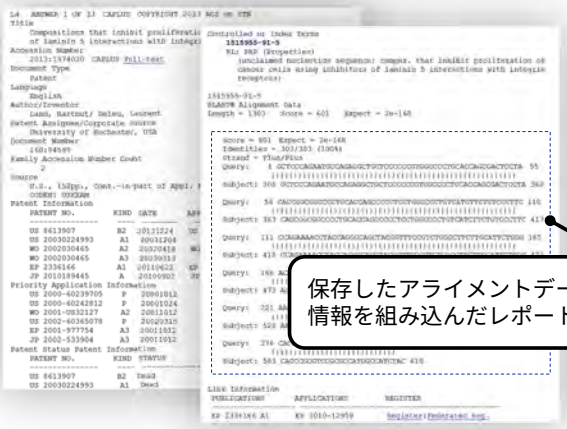
⑤ (オプション) レポート作成

CAS RN® を含む表示形式で表示した場合は、アライメント付き BLAST レポートを作成できる

(例) REGISTRY ファイルの出力結果を利用した
アライメント付き BLAST レポート



(例) CAplus ファイルの出力結果を利用した
アライメント付き BLAST レポート



保存したアライメントデータ (.xss) の情報を組み込んだレポート

レポート作成方法は下記ページの「<詳細版> REGISTRY ファイル配列検索」を参照
https://www.jaici.or.jp/stn-ip-protection-suite/cas-stnext/documents/

GENESEQ ホモロジー検索

GENESEQ ファイルのホモロジー検索

GENESEQ ファイルの配列検索には RUN コマンドを使う

=> RUN BLAST コード/検索フィールド パラメータ
=> RUN GETSIM コード/検索フィールド パラメータ

- コードは 5 つ以上を入力する
  - 核酸 は、5’ 末端から 3’ 末端の順でコードを入力する
  - タンパク質は、N 末端 (NH2) から C 末端 (COOH) の順でコードを入力する
- アミノ酸の 3 文字コード、ギャップ記号、特殊記号は利用できない
- 回答数の上限のデフォルトは 15,000 件。パラメータで最大 10 万件に変更可

GENESEQ ファイル – ホモロジー検索の検索フィールド

検索フィールド	BLAST	GETSIM	内容	質問式	回答
/SQN	○	○	塩基配列の質問式に類似した塩基配列を検索	塩基配列	塩基配列
/TSQN	○	○	データベース中の塩基配列をアミノ酸に翻訳した配列の中からアミノ酸配列の質問式に類似した配列を検索	アミノ酸配列	塩基配列
/SQP	○	○	アミノ酸配列の質問式に類似したアミノ酸配列を検索	アミノ酸配列	アミノ酸配列
/SQM	○	×	非常に類似した（種内などの）配列用に最適化された BLASTn (megaBLAST)	塩基配列	塩基配列
/SQDM	○	×	一部の塩基を無視し (多少のミスマッチを許容し)、より離れた（種間などの）配列を検索するために最適化された BLASTn (discontiguous megaBLAST)	塩基配列	塩基配列
/TSQP	○	×	塩基配列の質問式をアミノ酸配列に翻訳してこれに類似したアミノ酸配列を検索 (BLASTx)	塩基配列	アミノ酸配列
/TSQNX	○	×	塩基配列の質問式をアミノ酸配列に翻訳してこれに類似したアミノ酸配列に翻訳された塩基配列を検索 (tBLASTx)	塩基配列	塩基配列

GENESEQ ファイル - パラメータ設定例

- フィルタリングを行わず検索する

  => RUN BLAST GCTCCCAGAATGC/SQN -F F

  「Low Complexity Filtering」 のデフォルトは ON で低複雑度領域のマスクフィルタリングが行われ、生物学的に無意味なアライメントは取り除かれる設定になっている
  特許性調査の場合はチェックをはずした方がよい

- 核酸検索のデフォルトでは相補鎖も含めて検索される
  - 完全・部分配列検索と同様に相補鎖を含めるかどうかをパラメータにより変更できる
  - 例 :フィルタリングを行わず、入力したコードの鎖のみ検索する (相補鎖を含まない)

    => RUN BLAST GCTCCCAGAATGC/SQN -F F -S SIN
- 回答上限の変更
  - 例 :フィルタリングを行わず、回答上限を 10 万件に変更して検索する

    => RUN BLAST L1/SQN -F F -MAXSEQ 100000

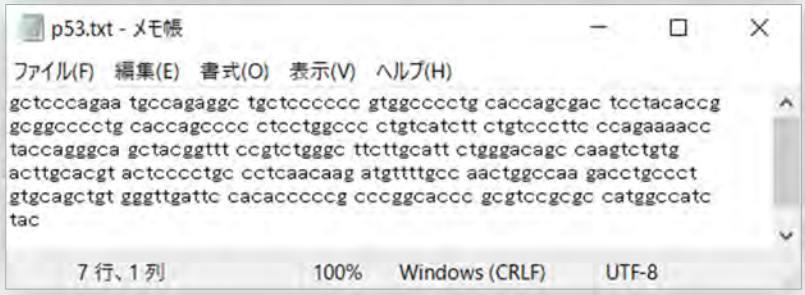
* パラメータの詳細については、=> HELP OPTIONS 参照

検索例 – GENESEQ BLAST

がん抑制遺伝子 p53 に類似する核酸を GENESEQ ファイルの BLAST 検索で調査する

[事前準備]

- 長い配列質問式の場合はテキストファイル (.txt) を作成する



アップロードを使用した => RUN BLAST コード/SQN の検索の上限は 30,000 コード。詳細は => HELP QLIMITS 参照

① 配列質問式のアップロード

=> FILE GENESEQ ← GENESEQ ファイルに入る

1) 画面右上の My Files のプルダウンメニューから Structures を選択する (あるいは画面左の アイコンをクリック)

2) Import Biosequence をクリックし，事前に作成した配列質問式のテキストファイルを呼び出して，OK をクリックする。

3) 配列質問式がインポートされたら Upload をクリックする。

4) アップロードした配列質問式の L 番号が表示される。

② ホモロジー検索の実行

```
=> RUN BLAST L1/SQN -F F              ← フィルタを外して BLAST ホモロジー検索を実行する

Algorithm: BLAST - BLASTN. Version: 2.12.0+

The BLAST software is used herein with permission of the
Center for Biotechnology Information (NCBI) of the Nation
Medicine (NLM).  See also: Zheng Zhang, Scott Schwartz, L
and Webb Miller (2000), "A greedy algorithm for aligning
sequences", J Comput Biol 2000; 7(1-2):203-14.

Database: GENESEQ
Posted Date: Mar 2, 2022 06:00 AM CET
Number of Hits to DB: 3430
Expectation Value: 10.0
Number of Sequences: 40160811
Length of query: 303
Length of database: 2257459437
Search space: 6840102094110
Effective search space: 5834704045646
Lambda: 0.625
Kappa: 0.41
Entropy: 0.78

Highest possible score value: 547.7
Best answer score value: 547.7

3430 ANSWERS FOUND BELOW EXPECTATION VALUE OF: 10.0
GENESEQ
```

オプション : フィルタをはずす (-F F)

Low Complexity Filtering (低複雑度領域フィルタ)
- 低複雑度領域のマスクフィルタリングを行う
- これにより統計的に有意であっても生物学的に無意味なアラインメント (例：酸性リッチ、塩基リッチ、プロリンリッチ領域) を取り除くことができる
- 配列質問式中にマスクされるコードは塩基配列の場合は N で、アミノ酸配列の場合は X で置き換えられ検索される

配列質問式の長さ

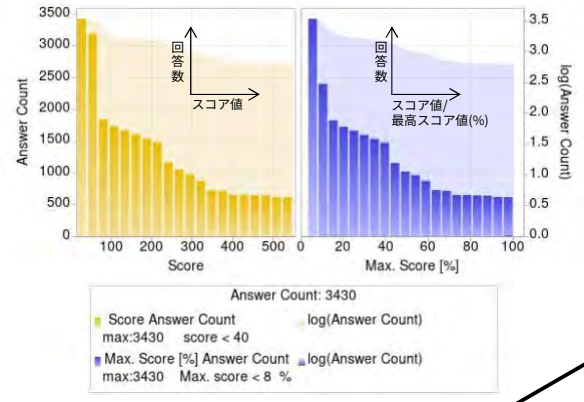
配列質問式が回答中の配列に完全に一致した場合に得られるスコア値

最も類似度が高い回答のスコア値

回答

CAS
A division of the American Chemical Society

② ホモロジー検索の実行




```
Query time:          708
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 ?:ALL
```

回答のオプションを指定する
- 回答全件 (ALL と入力)
- スコア値の最低値 (数字を入力)
  入力例：80
- スコア値パーセント (スコア値/最高スコア値) の最低値
  入力例：85% または 85% SCORE
- 同一性パーセント (一致コード数/Alignment コード数) の最低値
  入力例：100% IDENT
- スコア値パーセントの最低値と同一性パーセントの最低値
  入力例：85% SCORE 100% IDENT

CAS
A division of the American Chemical Society

② ホモロジー検索の実行

```
L2    RUN STATEMENT CREATED
L2        3430 GCTCCCAGAATGCCAGAGGCTGCTCCCCCCGTGGCCCCTGCACCAGCGAC
                TCCTACACCGCGGCCCCTGCACCAGCCCCCTCCTGGCCCCTGTCATCTT
                CTGTCCCTTCCCAGAAAACCTACCAGGCAGCTACGGTTTCCGTCTGGGC
                TTCTTGCATTCTGGGACAGCCAAGTCTGTGACTTGCACGTACTCCCCTGC
                CCTCAACAAGATGTTTTGCCAACTGGCCAAGACCTGCCCTGTGCAGCTGT
                GGGTTGATTCCACACCCCCGCCCGGCACCCGCGTCCGCGCCATGGCCATC
                TAC/SQN -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

=> SAVE L2 P53GENESEQ/A                ← 結果を保存する場合は回答を並べ替える前の L 番号を指定する (任意)
ANSWER SET L2 HAS BEEN SAVED AS 'P53GENESEQ/A'

=> SORT L2 1- SCORE D IDENT D          ← スコア値 (SCORE) の降順 (D) かつ同一性 (IDENT) の降順 (D) で回答を並べ替える

PROCESSING COMPLETED FOR L2
L3         3430 SORT L2 1- SCORE D IDENT D
```

回答はレコード番号の新しい順に並んでいる

回答集合のオプションは複数回指定できる。
終了する場合は END を入力する

③ 回答表示

```
=> D L3 1 3000 TRIAL ALIGN             ← 1 番目と 3000 番目の回答を TRIAL ALIGN 表示形式で表示する

L3     ANSWER 1 OF 3430 GENESEQ COPYRIGHT 2022 CLARIVATE on STN.
AN     AYM36275   GENESEQ
TI     Evaluating a patient with acute lymphoblastic leukemia (ALL)
       characterized by the presence of Philadelphia chromosome comprises
       generating an expression profile of ALL biomarkers from a test biological sample.
MTY    cDNA
DESC   Acute lymphoblastic leukemia prognosis determining DNA marker, SEQ 101.
KW     TP53 coding sequence; acute lymphoblastic leukemia; biomarker; coding
           :
SQL    1303
ALIGN
  Query Length: 303; Sequence Length: 1303;     ← 配列質問式の長さ ; 回答配列の配列長
  Score: 547.7 bits (606) , 100.0% of highest possible score 547.7;     ← スコア値
  Expect value: 7.759e-153;                                             ← 期待値
  Identities: 303 / 303 (100.0%);                                       ← 同一性パーセント
  Strand: Plus / Plus; Alignment Length: 303;
  Q:   1 GCTCCCAGAATGCCAGAGGCTGCTCCCCCCGTGGCCCCTGCACCAGCGACTCCTACACCG 60
         ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  S: 308 GCTCCCAGAATGCCAGAGGCTGCTCCCCCCGTGGCCCCTGCACCAGCGACTCCTACACCG 367

  Q:  61 GCGG                                    GCCCCTGTCATCTTCTGTCCCTTCCCAGAAAACC 120
         ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  S: 368 GCGGCCCCTGCACCAGCCCCCTCCTGGCCCCTGTCATCTTCTGTCCCTTCCCAGAAAACC 427
       :
```

スコア値 (類似性) が高い回答

TRIAL

ALIGN

全配列中の位置

アライメント表示
- Q (配列質問式 : Query)
- S (回答の配列 : Subject)
- | は一致しているコード

③回答表示

```
L3     ANSWER 3000 OF 3430 GENESEQ COPYRIGHT 2022 CLARIVATE on STN.
AN     BBA99493   GENESEQ
TI     Realizing efficient and fixed-point transgenosis by using transc
            :
MTY    DNA
DESC   Pig P53 target DNA.
KW     dna detection; ds; p53 gene; p53 tumor suppressor protein; rna detection
SQL    47
ALIGN
  Query Length: 303; Sequence Length: 47;
  Score: 41.9 bits (45) , 7.6% of highest possible score 547.7;
  Expect value: 1.46;
  Identities: 27 / 30 (90.0%);
  Strand: Plus / Plus; Alignment Length: 30;
  Q: 160 TCTGGGACAGCCAAGTCTGTGACTTGCACG 189
         ||||| ||||||||||||||| || ||||||
  S:   1 TCTGGAACAGCCAAGTCTGTAACCTGCACG 30

=> D L3 1 ALL ALIGN                ← 書誌情報、抄録、配列情報を表示する

L3     ANSWER 1 OF 3430 GENESEQ COPYRIGHT 2022 CLARIVATE on STN.
AN     AYM36275   GENESEQ  ED 20211030 UP 20211030
       DED 20110120 Full-text
TI     Evaluating a patient with acute lymphoblastic leukemia (ALL) that is
       characterized by the presence of Philadelphia chromosome comprises
            :
```

スコア値 (類似性) が低い回答

ALL

③回答表示

```
PI     WO 2010138843 A2        2010120
            :
OS     2010-P75161 [82]
MTY    cDNA
            :
PSL    Disclosure; SEQ ID NO 101; 43pp
            :
AB     The present          markers and a method for
       evaluating t          acute lymphoblastic leukemia
       (ALL) that is characterized by the presence of Philadelphia chromosome
       (Ph+). The method involves generating an expression profile from a
       biological sample obtained from a ALL diagnosed patient, where the
            :
SEQN   101
SQL    1303
SEQK   9102e84c616917d71fd3efb3093278c9ff90b18976b8511d45
SEQ                    ← 配列
                  1 gtccaggagc aggtagctgc tgggctccgg ggacact
                                    :
               1251 cctcccgcca taaaaaactc atgttcaaga cagaagg
               1301 tga
 NA                    ← 核酸の種類
       Code    Count  Percent
          A      292     22.4
          C      403     30.9
            :
```

ベーシック特許の特許番号のみ収録されている

WPI ファイルのレコード番号 (AN)

特許中の配列の記載位置

ALL

```
ALIGN
  Query Length: 303; Sequence Length: 1303;
  Score: 547.7 bits (606) , 100.0% of highest possible score 547.7;
  Expect value: 7.759e-153;
  Identities: 303 / 303 (100.0%);
  Strand: Plus / Plus; Alignment Length: 303;
  Q:   1 GCTCCCAGAATGCCAGAGGCTGCTCCCCCCGTGGCCCCTGCACCAGCGACTCCTACACCG 60
         ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  S: 308 GCTCCCAGAATGCCAGAGGCTGCTCCCCCCGTGGCCCCTGCACCAGCGACTCCTACACCG 367
            :
```

ALIGN

参考：WPI ファイルにクロスオーバー

```
=> D HIS                          ← 検索履歴を表示する
     :
L3          3430 SORT L2 1- SCORE D IDENT D

=> FILE WPINDEX                   ← WPI ファイルに入る

=> TRA L3 OS /AN
L4           TRANSFER L3 1- OS :      866 TERMS
L5          866 L4/AN
ALL TERMS IN L4/AN RETRIEVED.

=> D L5 10                        ← 書誌情報を表示する
L5   ANSWER 10 OF 866  WPINDEX COPYRIGHT 2022   CLARIVATE ANALYTIC
AN   2021-B5534W [2021087]   WPINDEX Full-text
TI   New recombinant Newcastle disease virus obtained by replacing
     newcastle disease virus losata with the F protein of virulent
     Newcastle disease virus useful for preparing medicine for treating tumor e.g. liver cancer
DC   B04; C06; D16
IN   JIANG S; LI D; LIU T; LIU Z; WANG Z; XIAO W
PA   (JIAN-N) JIANGSU KANGYUAN RUIAO BIOMEDICAL
     KANIONREAL BIOMEDICAL TECHNOLOGY CO LTD
CYC  135
PI   WO 2021197506   A1 20211007 (2021087)* ZH 27[13]
     CN 113462658    A  20211001 (2021087)  ZH
ADT  WO 2021197506 A1 WO 2021-CN95200 20210521; CN 113462658 A CN 2020-10238162
            :
```

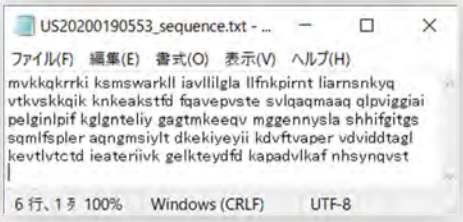
TRA L# OS /AN
GENESEQ ファイルの OS フィールドを抽出して
WPI ファイルの AN (レコード番号) フィールドで検索する

TRA L# PN を利用しない理由
WPI ファイルへクロスオーバする際に、特許番号（PN）を用いると GENESEQ ファイルでヒットした配列情報を含む特許に加え、関連レコード（分割出願，継続出願など）がヒットする可能性があるため

特許ファミリーを確認できる

実習 2

下記に類似するタンパク質を GENESEQ ファイルの BLAST ホモロジー検索で調べる



	実習の流れ	参照スライド
1	GENESEQ ファイルに入る	58
2	配列質問式のファイルをアップロードする	58
3	=> RUN BLAST L#/SQP –F F	54-56
4	全件 (ALL) を入手する	60
5	スコア値と同一性の高い順に並べ替える	61
6	ALL ALIGN 表示形式で表示する	63

実習 2 の回答

```
=> FILE GENESEQ                          ← GENESEQ ファイルに入る

(配列質問式をアップロードする)

=>
Uploading sequence file: example2

UPLOAD SUCCESSFULLY COMPLETED
L1    GENERATED

=> RUN BLAST L1/SQP -F F                 ← フィルタを外して BLAST
                                           ホモロジー検索を実行する
Algorithm: BLAST - BLASTP. Version: 2.12.0+
             :
ENTER EITHER "ALL" TO KEEP ALL ANSWERS
OR ENTER THE MINIMUM SCORE VALUE YOU WISH TO KEEP
             :
OR ENTER "END". "END" MUST BE ENTERED TO COMPLETE THE RUN COMMAND.
ENTER (ALL) OR ?:ALL                     ← ALL で全件入手する

L2     RUN STATEMENT CREATED
L2         1415 MVKKQKRRKIKSMSWARKLLIAVLLILGLALLFNKPIRNTLIARNSNKYQ
             :
```

```
ENTER EITHER "ALL" TO KEEP ALL ANSWERS
OR ENTER THE MINIMUM SCORE VALUE YOU WISH TO KEEP
             :
OR ENTER "END". "END" MUST BE ENTERED TO COMPLETE THE RUN COMMAND.
ENTER (ALL) OR ?:END                     ← END で終了する
             :
=> SORT L2 1- SCORE D IDENT D            ← スコア値 (SCORE) の降順 (D)
PROCESSING COMPLETED FOR L2                かつ 同一性 (IDENT) の降順 (D)
L3    1415 SORT L2 1- SCORE D IDENT D    で回答を並べ替える

=> D ALL ALIGN 1

L3      ANSWER 1 OF 1415 GENESEQ COPYRIGHT 2024 CLARIVATE on STN.
AN      BHW56191    GENESEQ  ED 20211030 UP 20211030
        DED 20200806 Full-text
TI      Deep eutectic solvent used to conduct transamidation
        reactions, i.e. sortase catalyzed reactions comprises (2-
             :
IN      Boenitz-Dulat M; Schatte M
PA      HOFFMANN LA ROCHE INC (HOFF)
LA      English
DT      Patent
PI      US 20200190553  A1        20200618
             :
```


CAS SEQUENCES

CAS STNext に搭載されている配列検索ツール

CAS Sequences 機能で膨大な配列コンテンツを検索できる

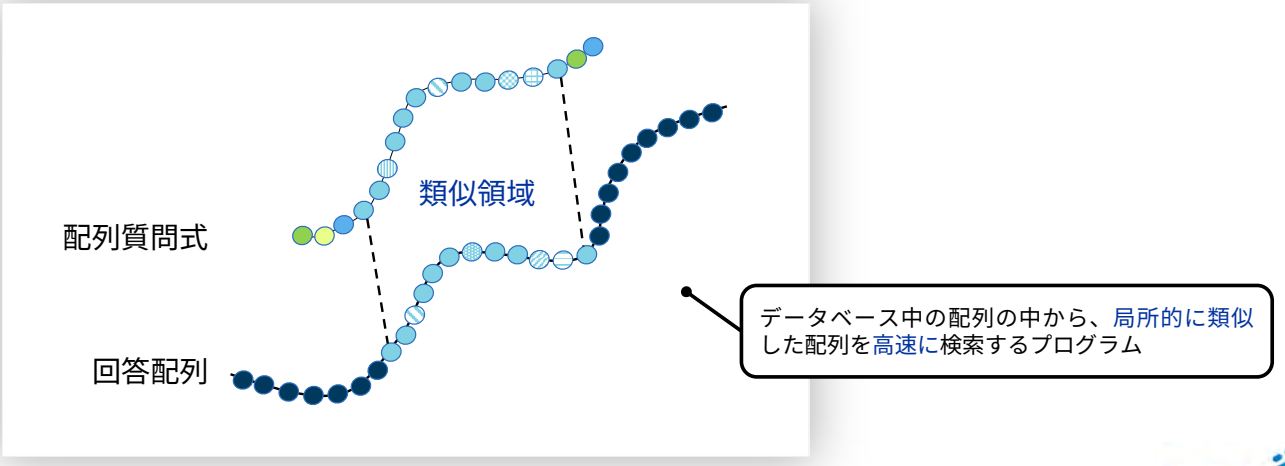
- 収録源
  - CAS が独自のルールに従い収集した REGISTRY ファイル収録の配列
  - 73 特許発行機関の特許から抽出した配列
  - NCBI 由来の配列
- 3 つの配列検索プログラム
  - BLAST ホモロジー検索
  - CDR 配列検索
  - Motif 配列検索

詳細は CAS STNext CAS Squences ガイド参照
https://www.jaici.or.jp/stn-ip-protection-suite/cas-stnext/documents/

BLAST ホモロジー検索

配列ホモロジー検索でよく使われている NCBI のプログラム

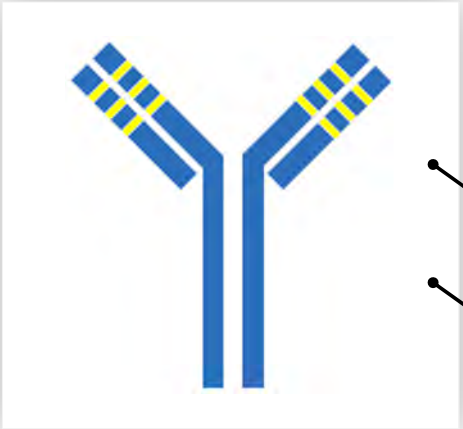
- BLAST (Basic Local Alignment Search Tool) の名が示す通り局所的に類似した配列を検索する



CDR 配列検索

CDR (相補性決定領域) を指定し検索するプログラム

- BLAST をベースにしている。CDR 配列検索は，検索対象をあらかじめ抗体や T 細胞受容体といった CDR を持つ配列を対象にしている



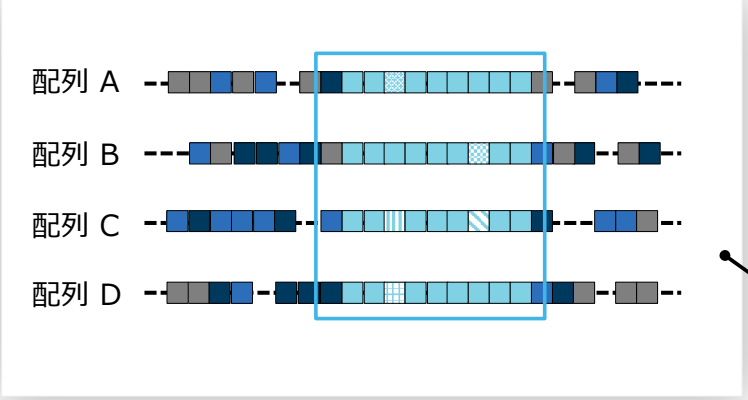
抗原に特異的な重要な部分を CDR (相補性決定領域) という

抗体や T 細胞受容体にある CDR の配列を複数指定し、検索できる

Motif 配列検索

パターン配列を検索するプログラム

- BLAST をベースにしている。[ ] などの記号を利用した検索が可能。数パターンの配列を一度に検索できる

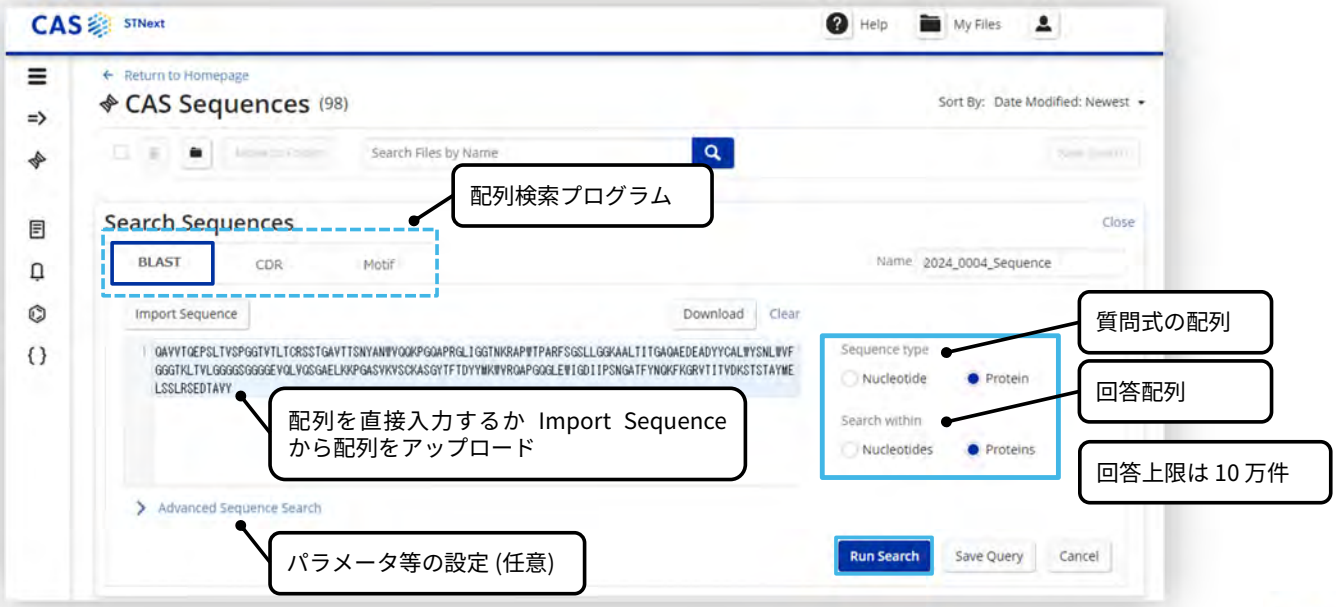


機能的に重要、立体構造と関連する保存配列パターンを Motif という

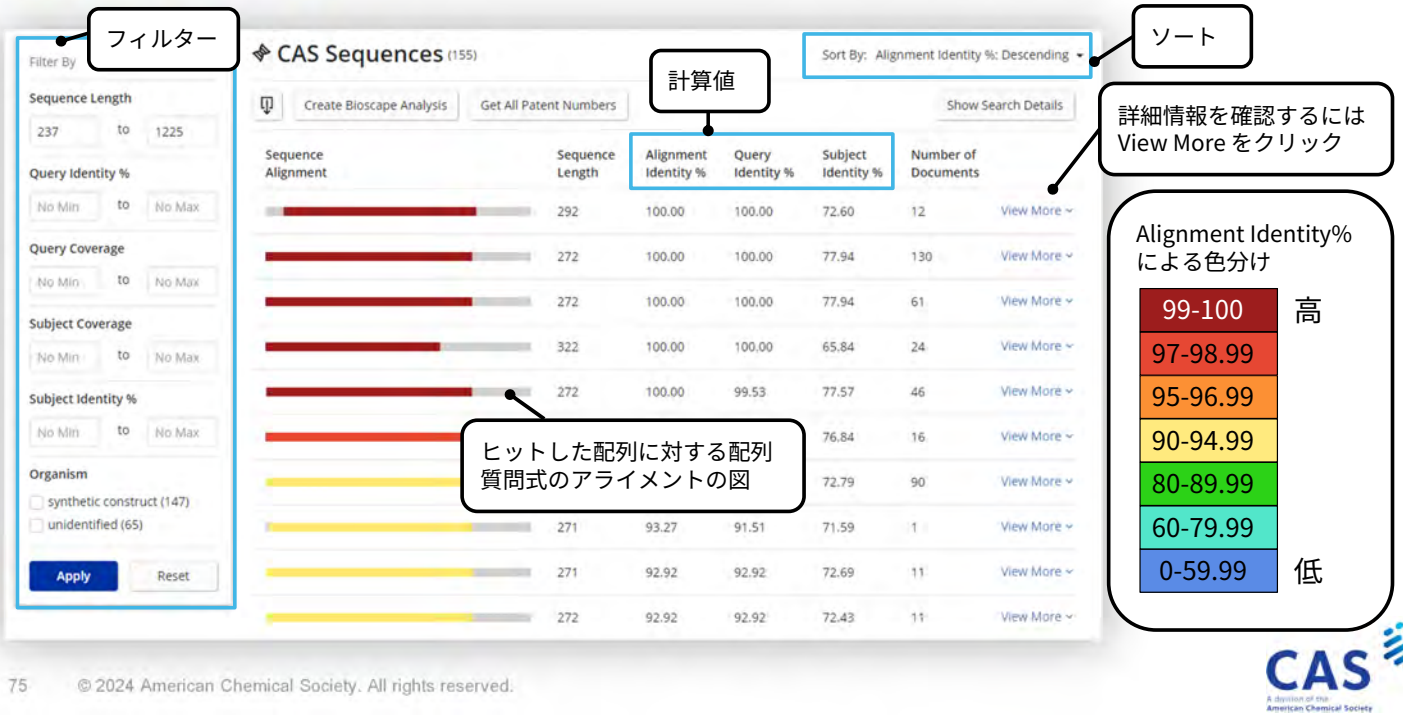
CAS Sequences の検索の流れ

左側の アイコンまたは My Files の CAS Sequences から検索をスタート

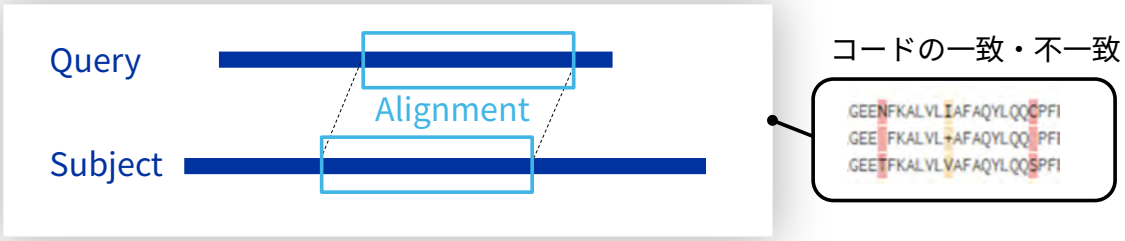
初期画面 - BLAST



検索結果一覧



用語の説明



用語	内容
Alignment (類似領域)	配列質問式と回答配列を並べてどこが類似領域か示したもの
Query (配列質問式)	検索した配列質問式
Subject (回答配列)	ヒットした回答配列

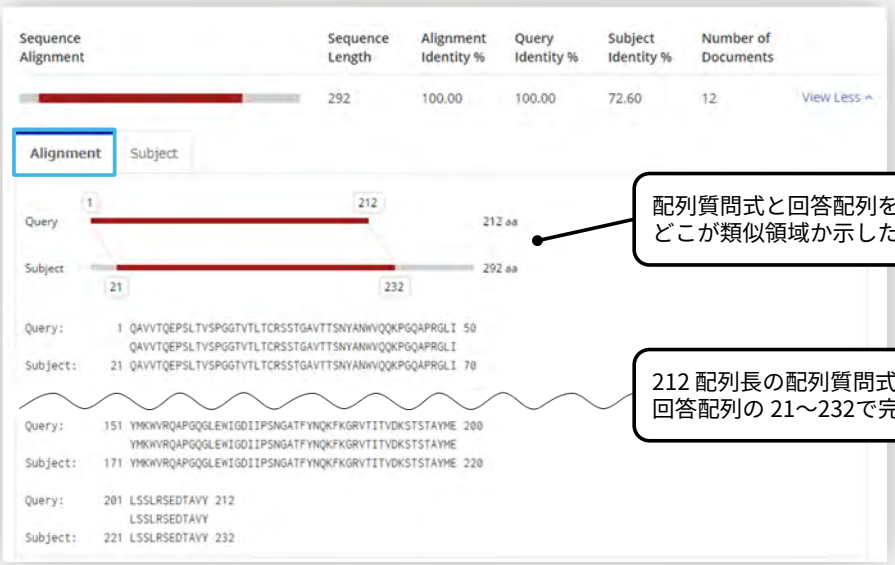
参考：各項目の計算値

– 計算方法

項目	内容	Filter By	Sort By
Alignment Identity %	一致したコード÷類似領域×100	-	○ (降順が デフォルト)
Query Identity %	一致したコード÷配列質問式×100	○	○
Query Coverage	類似領域÷配列質問式×100	○	○
Subject Identity %	一致したコード÷回答配列×100	○	○
Subject Coverage	類似領域÷回答配列×100	○	○

検索結果詳細 – Alignment タブ

Alignment% の高い回答 (100%)



配列質問式と回答配列を並べて
どこが類似領域か示したもの

212 配列長の配列質問式の 1～212 が、292 配列長の
回答配列の 21～232で完全に一致している

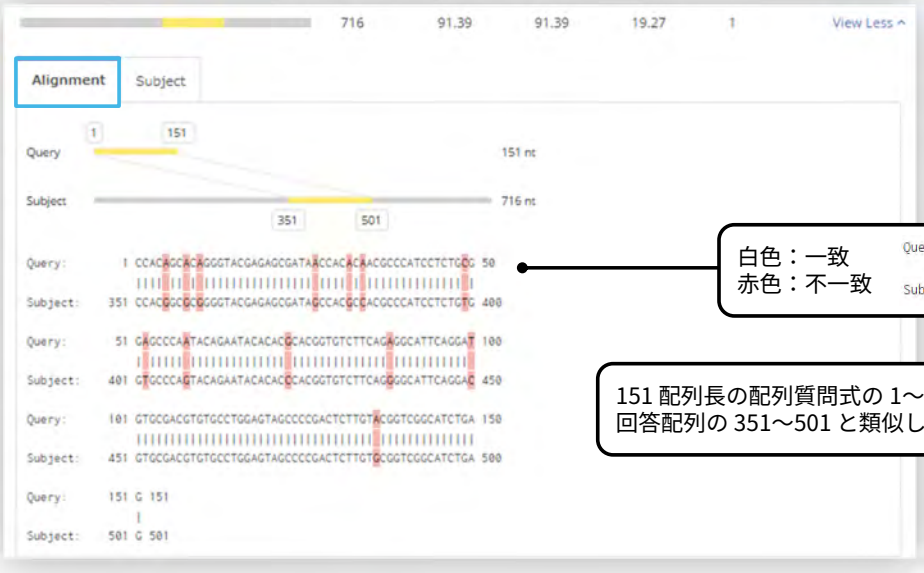
検索結果詳細 – Alignment タブ

Alignment% の低い回答 (92.79%)

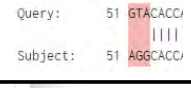
212 配列長の配列質問式の 5〜212 が、270 配列長の回答配列の 4〜211 と類似している (類似領域は 208)

白色：一致
赤色：不一致
橙色：不一致だが等価のアミノ酸でヒット

参考：核酸の Alignment 例



白色：一致
赤色：不一致

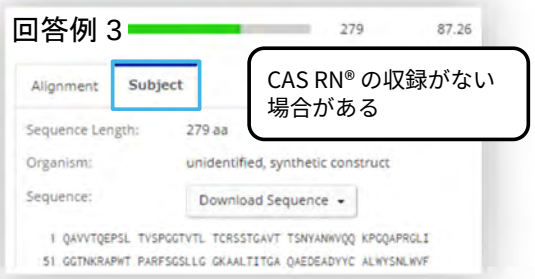


151 配列長の配列質問式の 1〜151 が、716 配列長の回答配列の 351〜501 と類似している (類似領域は 151)

検索結果詳細 – Subject タブ

回答配列に関連する CAS RN®

回答配列に関連する
GenBank Accession 番号

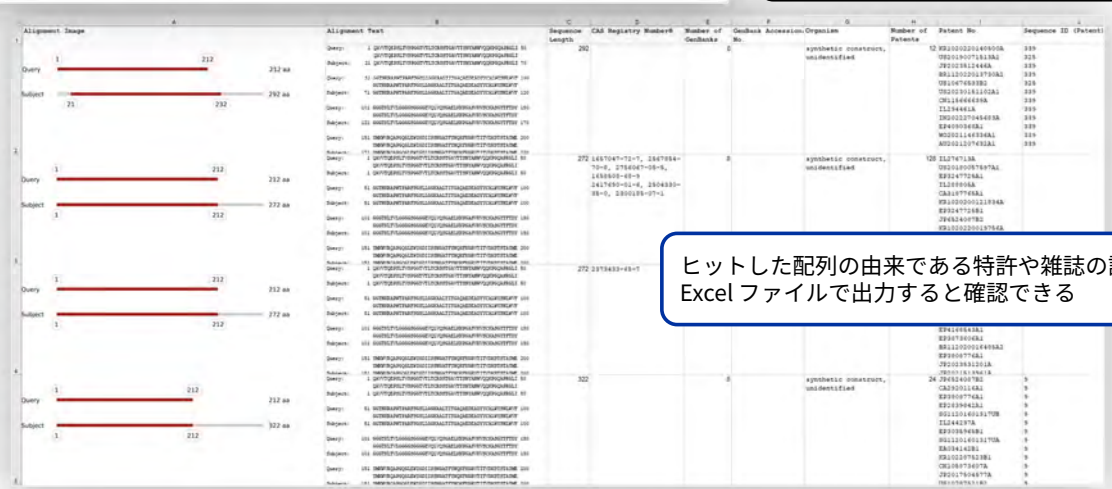


CAS RN® の収録がない
場合がある

表示している配列に関する CAS RN® や
GenBank Accession 番号を検索する
(上限 5,000)
① CAS RN® を検索
② GenBank Accession 番号を検索

検索結果のダウンロード

Excel で検索結果をダウンロード
(上限 1,000 件)



ヒットした配列の由来である特許や雑誌の詳細は
Excel ファイルで出力すると確認できる

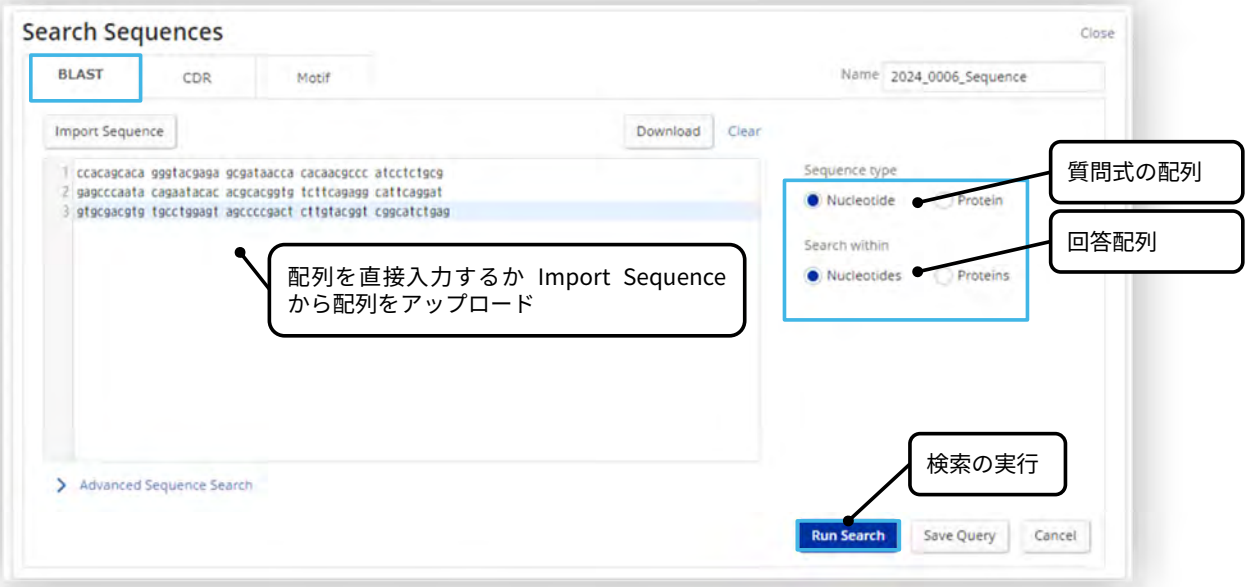
実習 3

下記の配列と類似した核酸を CAS Sequences で調査する

ccacagcaca gggtacgaga gcgataacca cacaacgccc atcctctgcg
gagcccaata cagaatacac acgcacggtg tcttcagagg cattcaggat
gtgcgacgtg tgcctggagt agccccgact cttgtacggt cggcatctgag

	実習の流れ	参照スライド
1	CAS Sequenses にアクセスする	73
2	BLAST タブを開き、配列質問式を入力する	74
3	Sequence Type、Search within で Nucleotides を選択する	
4	回答の View More をクリックして回答の詳細情報を確認する	75
5	500 以下の配列長で限定する	

実習 3（回答）：検索



実習 3（回答）：検索中～検索完了

実習 3（回答）：検索結果画面

JAICI ヘルプデスク

Tel : 0120-003-462 (平日 9:00-17:00)

Mail : support@jaici.or.jp

JAICI 化学情報協会 | CAS A division of the American Chemical Society