CAS
A division of the
American Chemical Society

STNext®

REGISTRY ファイル
配列検索

JAICI

＊ 目次 ＊

A REGISTRY ファイル 配列検索

REGISTRY ファイルの配列に関する収録内容を紹介し，ホモロジー検索以外の配列検索方法（完全配列検索，完全配列ファミリー検索，部分配列検索，部分配列ファミリー検索），およびその他の配列関連情報の検索方法について説明します．

REGISTRY ファイルの概要

■ REGISTRY ファイルは CAS が作成する化学物質と配列の辞書ファイルである.

（2022 年 2 月）

ファイル名	REGISTRY
製作者	Chemical Abstracts Service (CAS)
収録源	・ CAplus/CA ファイルに収録されているベーシック特許および雑誌論文 ・ GenBank 由来の配列データ （2005 年以降は出典のあるもののみ）
特長	・ 特許だけでなく, 雑誌からも配列情報も収録している. ・ GenBank 由来の配列情報も収録している. ・ 配列レコードを, CAplus/CA ファイルにクロスオーバー検索すると, その配列に関する文献（特許・非特許）情報が簡単に得られる. ・ 配列に関する規制情報, 安全性情報, 供給業者などの情報が, REGISTRY ファイルから他のファイルへクロスオーバー検索することで簡単に得られる.
収録年	配列の収録は 1950 年代前半～
レコード構成	物質単位（配列単位）
収録件数	核酸 : 60,889,100 件 タンパク質 : 12,361,900 件 合計 : 73,251,000 件
CAS RN® 付与率	100 %
更新頻度	毎日
タイムラグ	27 日以内（主要国特許）

REGISTRY ファイルの配列情報

■ REGISTRY ファイルの配列レコード

・ 核酸・タンパク質の配列が約 7,000 万件以上収録されている.

- 収録されている配列レコードの約 91 % は 2000 年以降の登録である.

・ 配列に関する文献情報（特許，非特許文献）が必要な場合は，CAplus/CA ファイルへクロスオーバー検索する.

■ 配列の収録源

・ CAplus/CA ファイルに収録された文献

- 56 ヶ国の特許発行機関と 5 国際機関（EAPO，EPO，WIPO，ARIPO，GCC）から発行された特許で，CAplus/CA ファイルのベーシック特許*
- 雑誌論文
- 会議録
- 技術レポート

・ GenBank

- 特許 ：USPTO（米国），EPO（欧州），JPO（日本）などから提供されたデータ
- 研究者から受領したデータ ： 出典のあるデータのみ（2005 年 3 月以降）
  それ以前は未発表データも収録

＊ ベーシック特許とは，一つの特許ファミリーのうち，CAS が最初に入手した特許である.

■ 収録基準

<table>
<tr><th></th><th>核酸</th><th>タンパク質・ペプチド*1</th></tr>
<tr><td>配列数</td><td>塩基数が 9 以上の配列</td><td>アミノ酸が 4 以上の配列</td></tr>
<tr><td rowspan="3">特許由来の
配列</td><td>1999 年 9 月以前
・新規性に関与する完全配列*2</td><td>1987 年以前
・新規性に関与する完全配列*2
1988 年～1998 年
・新規性に関与する完全配列と部分配列（ギャップを含まない配列）</td></tr>
<tr><td colspan="2">1999 年（核酸は 1999 年 10 月）～ 2014 年 4 月
・新規性の有無や記載位置に関係なく，特許に記載されたすべての配列*3</td></tr>
<tr><td colspan="2">2014 年 5 月以降
・請求項および実施例中の請求項に関連した主要な配列*3</td></tr>
<tr><td rowspan="2">非特許由来の
配列</td><td>1991 年以前
・新規性に関与する完全配列*2

1992 年以降
・新規性に関与する完全配列と部分配列（ギャップを含まない配列）</td><td>1998 年以前
・新規性に関与する完全配列*2 と選択された雑誌由来の新規性に関与する部分配列（20 以上のアミノ酸配列でギャップを含まない配列）
1999 年以降
・新規性に関与するすべての完全配列および部分配列</td></tr>
<tr><td>GenBank
由来の配列*4</td><td colspan="2">2005 年 2 月以前
・全配列（未発表，不完全な残基を含む）
2005 年 3 月以降
・CAplus/CA ファイルの収録対象である雑誌や特許などに記載された配列のみ*5</td></tr>
</table>

*1 REGISTRY ファイルでは，アミノ酸 50 個未満をペプチド，50 個以上をタンパク質と定義している．

*2 完全配列とは，① 著者が完全な核酸分子であると報告したもの，② より大きな核酸分子に含まれタンパク質または RNA を産生するためのすべてのコード情報を有する領域（遺伝子），③ 開始と終了を有する遺伝子領域（プロモーター，調整領域）など．

*3 2005 年以降は，請求項に 4,000 以上の配列を含む特許の配列は収録していない．

*4 GenBank は核酸配列が収録されている．また，核酸配列の仮コーディング領域からタンパク質に翻訳した配列も収録されており，CN フィールドに翻訳された核酸の GenBank 番号が含まれている．

*5 2007 年後半以降は 500 以上の GenBank アクセッション番号を持つ論文の配列は収録しない．

■ 登録ルール

・ 一つでも核酸塩基が異なるものは別の配列として収録される．

・ 配列が同じでも，化学修飾された配列や側鎖の置換基の異なる配列，同位体で置換された配列は，別配列として収録される．

・ GenBank 由来の配列は，1 GenBank 番号につき 1 レコードとして収録される．
ただし，文献・特許由来の配列と GenBank 由来の配列が，同じ配列の場合は，一つのレコードにまとまる．

＊ 注意 ： 登録ルールの例外
- 2002 年以降の特許・雑誌由来の配列は，修飾基も含めてまったく同一の配列であっても同一レコードにならず特許・雑誌ごとに別レコードとして収録される場合がある．
- GenBank 由来の配列と文献・特許由来の配列が同じ配列であっても同一レコードにならず，別レコードとして収録される場合がある．

■ 核酸のレコード例（SQIDE 表示形式）

```
CAS RN®  RN   2128329-99-5  REGISTRY
CA 索引名,  CN   DNA (synthetic clone WO2017-140839 oligonucleotide)
化学物質名    (CA INDEX NAME)
             OTHER NAMES:
             CN   1: PN: WO2017140839 SEQID: 2 claimed DNA
ファイルセグメント FS   NUCLEIC ACID SEQUENCE
配列長       SQL  115
核酸タイプと数  NA   34 a   27 c   27 g   27 t
特徴表*1     NTE  modified
             ------------------------------------------------------------------
              type            ------ location ------         description
             ------------------------------------------------------------------
             modified base    c-1                          5'-phosphate
             ------------------------------------------------------------------

特許情報     PATENT ANNOTATIONS (PNTE):
             Sequence |Patent
             Source   |Reference
             =========+=============
             Not Given|WO2017140839
                      |claimed SEQID
                      |2

配列*4       SEQ      1 catgttcgat gaggcacgat agatgtacgc tttgacatac gctttgacaa
                     51 tacttgagca gtcggcagat ataggatgtt gcaagctccg tgagtcccac
                    101 aaaccaaaaa cctc

             **RELATED SEQUENCES AVAILABLE WITH SEQLINK**
分子式       MF   Unspecified
クラス識別子  CI   MAN
収録源       SR   CA
CAS RN® 所在  LC   STN Files:   CA, CAPLUS
CAplus ファイルの資料書類  DT.CA  CAplus document type:  Patent
CAS ロール   RL.P   Roles from patents:  BIOL (Biological study); PRP (Properties)
                    1 REFERENCES IN FILE CA (1907 TO DATE)
                    1 REFERENCES IN FILE CAPLUS (1907 TO DATE)
```

特許番号と配列の記載位置，配列番号*2 *3

核酸に限定（NS/FS）

配列長で限定（/SQL）

特許番号で限定（/PN）
記載位置や配列番号で限定（/PNTE）

*1　クレームされている配列についての注釈が NTE（特徴表）フィールドに収録される.

*2　1999 年 10 月以降，CAS が特許から索引した配列のレコードには，収録源である特許の特許番号と配列の記載位置，配列番号などが CN（化学物質名称）フィールドに収録される.

*3　1999 年 10 月以降，CAS が特許から索引した配列のレコードには，配列の特許情報が PNTE（特許情報）フィールドに収録される.

- 収録源の特許の特許番号（STN 形式）
- クレームされているかどうか
- 特許明細書中の記載位置（配列番号など）
- 特許についている配列表のアノテーション情報（WIPO の公表した規格の用語が使われている.）

*4　配列長が 700,000 を超えるとシステムリミットのために，SEQ フィールドに配列は表示されず，DISPLAY LENGTH EXCEEDS SYSTEM LIMITS のメッセージが表示される.

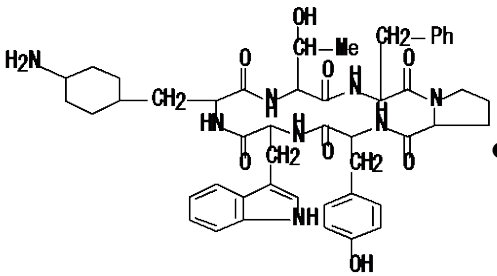
■ タンパク質のレコード例（SQIDE 表示形式）

```
RN   216259-64-2  REGISTRY
CN   Cyclo[3-(trans-4-aminocyclohexyl)-L-alanyl-L-threonyl-L-phenylalanyl-L-
     prolyl-L-tyrosyl-D-tryptophyl] (9CI)  (CA INDEX NAME)
OTHER NAMES:
CN   70: PN: US20020042374 PAGE: 10 claimed protein
CN   74: PN: US6268342 SEQID: 80 claimed protein
FS   PROTEIN SEQUENCE; STEREOSEARCH
SQL  6
特徴表*5 NTE  cyclic
     modified (modifications unspecified)
-------------------------------------------------------------------------
 type            ------ location ------         description
-------------------------------------------------------------------------
stereo           Trp-6           -         D
-------------------------------------------------------

PATENT ANNOTATIONS (PNTE):
Sequence |Patent
Source   |Reference
=========+=============
Not Given|US2002042374
         |claimed PAGE
         |10
---------+-------------
         |US6268342
         |claimed SEQID
         |80

SEQ      1 ATFPYW

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
MF   C47 H58 N8 O8
SR   CA
LC   STN                                    AT2, USPATFULL
DT.CA  CAplus document type:  Patent
RL.P   Roles from patents:  BIOL (Biological study); PRP (Properties); USES
       (Uses)
RLD.P  Roles for non-specific derivatives from patents:  BIOL (Biological
       study); PREP (Preparation); USES (Uses)
```




```
7 REFERENCES IN FILE CA (1907 TO DATE)
1 REFERENCES TO NON-SPECIFIC DERIVATIVES IN FILE CA
7 REFERENCES IN FILE CAPLUS (1907 TO DATE)
```

特許番号と配列の記載位置，配列番号*2 *3

タンパク質に限定（PS/FS）

配列長で限定（/SQL）

配列の形態や修飾の情報で限定（/NTE）

特許番号で限定（/PN）
記載位置や配列番号で限定（/PNTE）

SQIDE3 表示形式の場合 ：アミノ酸が 3 文字コードで表示される
SEQ3　1 Ala-Thr-Phe-Pro-Tyr-Trp

SEQLINK EXACT コマンドで同じ配列をまとめることができる

構造（水素以外の元素数が 252 以下の場合は構造検索も可能）

*5 クレームされている配列についての注釈が NTE（特徴表）フィールドに収録される.
ただし，特殊・未定義のアミノ酸コードを有するタンパク質の配列の場合は，クレームの有無に関わらず，特殊・未定義のアミノ酸コードの定義が特徴表に収録される.

回答表示形式

■ 表示形式（ 　　 は利用頻度の高い表示形式）

・ デフォルトの IDE 表示形式では，配列情報が表示されない．

・ 配列情報を表示するには，配列専用の定型表示形式（SQIDE 表示形式など）を指定する．

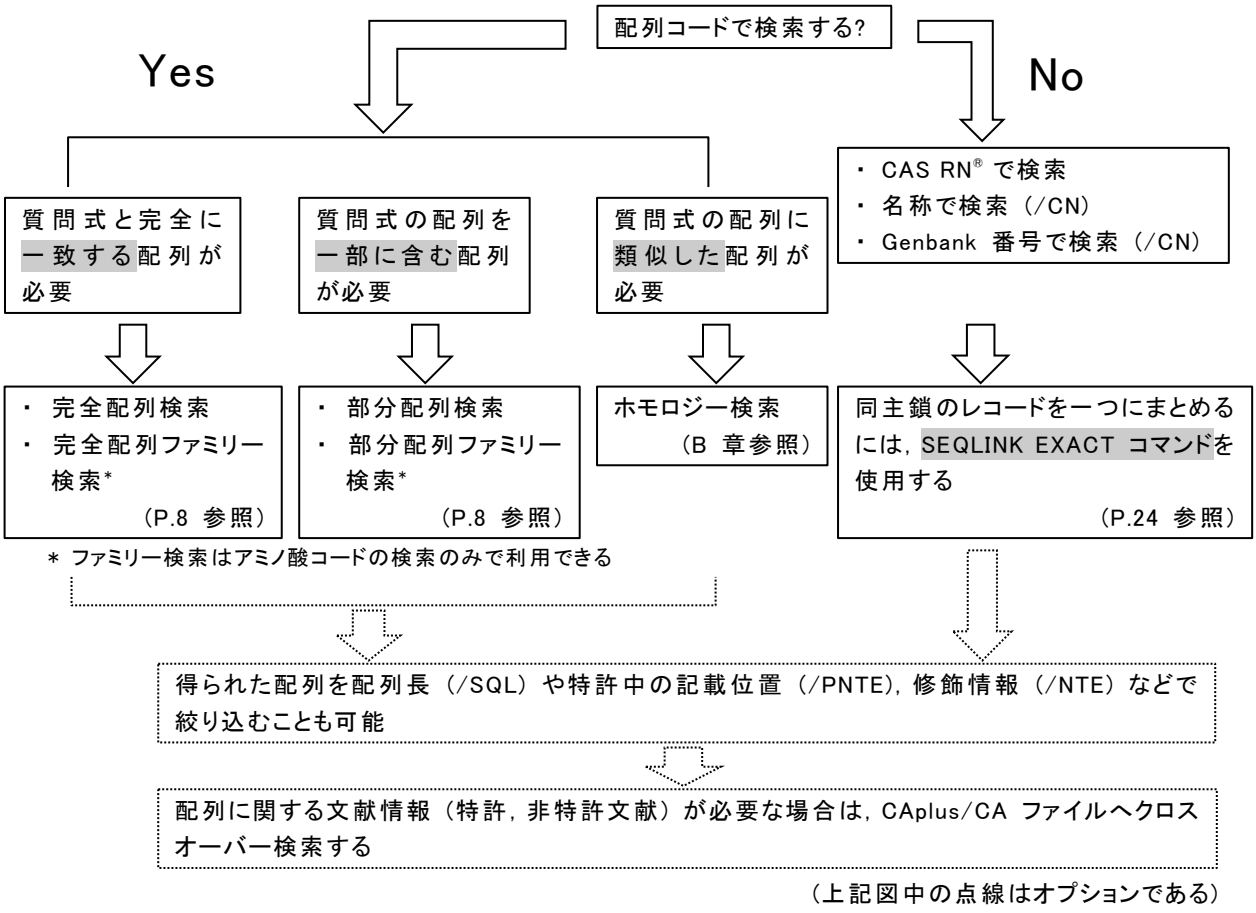
・ SCAN 表示形式では，配列長を確認することは出来るが，配列は表示されない．

表示形式		内容
一般	IDE	FIDE 表示形式から環系データ（RSD）と物性データ（PROP）を除いた物質情報．ただし，50 名称まで．デフォルト表示形式
	SCAN	SAM 表示形式と同じ．ただし，回答番号は表示されず，回答順序はランダム CA 索引名と配列長（配列は表示されない）
	SAM	CA 索引名，分子式，クラス識別子コード，構造，配列長，組成
	REG	CAS RN® (RN, DR, AR, PR, RR)
	PPROP	すべての計算物性値
	EPROP	すべての実測物性値
	PROP	すべての物性値（EPROP, ETAG, PPROP）
	FIDE	タンパク質・核酸の配列データを除くすべての物質情報
	ALL*1, IALL*2, MAX	すべてのフィールド，および CA ファイルの最新 10 件分の文献情報 (BIB, ABS, IND)
配列関連	SQD	CAS RN® (RN, AR, DR, PR, RR)，ファイルセグメント（FS），配列長（SQL）， 特徴表（NTE），特許情報（PNTE），1 文字コードの配列データ（SEQ）
	SQD3	SQD 表示形式と同じ．ただし，配列データ（SEQ3）は 3 文字コードで表示される．
	SQN	CAS RN® (RN, AR, DR, PR, RR)，化学物質名称（CN），ファイルセグメント（FS），配列長（SQL），CAplus/CA/CAOLD ファイルの文献数（REF）
	SQIDE	IDE 表示形式，および配列長（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 表示形式と同じ．ただし，配列データ（SEQ3）は 3 文字コードで表示される．
ほか	HIT*3	ヒットした配列およびヒットタームを含むフィールド
	KWIC*3	ヒットした配列の列およびヒットタームの前後 20 語
	QRD*3	IDE 表示形式とヒットした配列およびヒットタームを含むフィールド

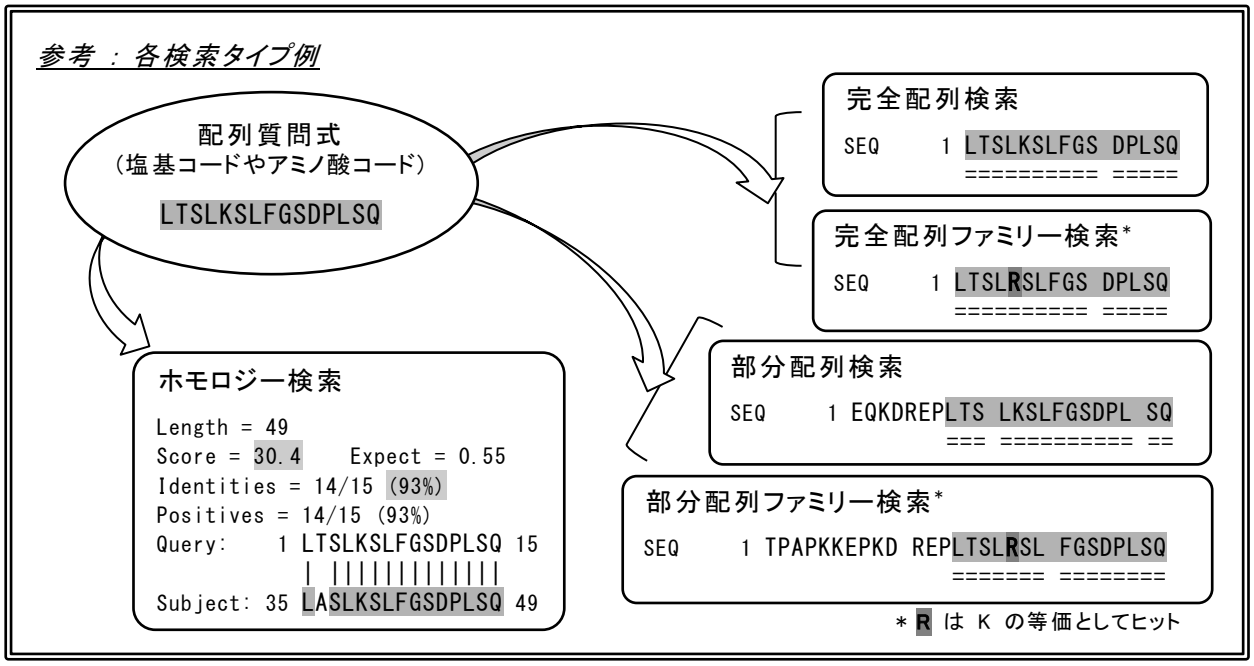
*1　データ量が非常に多い場合がある．CA ファイルの文献（10 件以下）を表示できる．
*2　インデント型 ALL 表示形式．CA ファイルの文献情報部分が完全なフィールド名で表示される．
*3　完全配列検索した場合は，FS, SQL フィールドも表示される．

配列検索の指針

■ 核酸・タンパク質を調査する際には，配列コードから検索したいのか，CAS RN® や名称，Genbank 番号から検索したいのかにより，検索方法が異なる．



（上記図中の点線はオプションである）



完全配列検索・部分配列検索

■ 完全配列検索・部分配列検索の入力方法

=> S コード/検索タイプ

・ 配列検索は，核酸は塩基コード，タンパク質はアミノ酸コードを使って検索する.

・ コード間にスペースは使用しない.

■ 検索タイプ

核酸	/SQEN	完全配列	/SQSN	部分配列
タンパク質	/SQEP	完全配列	/SQSP	部分配列
	/SQEFP	完全配列ファミリー	/SQSFP	部分配列ファミリー

・ 完全配列検索　：配列質問式に完全に一致した配列がヒットする.
　部分配列検索　：配列質問式を一部に含む配列がヒットする.
　ファミリー検索　：等価なアミノ酸も含めた配列がヒットする（アミノ酸の等価表 P.13 参照）.

・ 部分配列検索および部分配列ファミリー検索では，ギャップ記号・特殊記号を利用することができる.

■ 入力例

=> S GCCAAGCTGGCATCCGTCA/SQEN　← *核酸の完全配列検索*
=> S LTSLKSLF/SQEP　← *タンパク質の完全配列検索*
=> S L1 AND CLAIM?/PNTE　← *クレームされている配列に限定*
=> S L1 AND SQL=<300　← *配列長で限定*
=> S (TPKTR(L)ALEGSL)/SQSP　← *同一鎖中に含まれている配列を検索*
=> S MALWMRLL.{1-5}LALWP/SQSP　← *ギャップ記号を使った検索*
=> S (TGACGGAT|ATCCGTCA)/SQSN　← *特殊記号を使った検索*

■ 配列のコーディング（コードの並び順）

・ 核酸 ：5’ 末端から 3’ 末端の順

・ タンパク質 ： N 末端（NH2）から C 末端（COOH）

■ 核酸の相補鎖検索

・ 核酸の完全配列（/SQEN），部分配列検索（/SQSN）では，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.

■ 完全配列検索では，配列長 200 までの配列を EXPAND で確認することができる．

・ 入力例

=> E TTTGGGGTTT/SQEN

=> E TALKR/SQEP

■ 配列の抽出

・ SELECT コマンドにより，レコードから配列を抽出できるが，200 コードまでである．

■ 配列検索に利用できる演算子

・ 部分配列検索（/SQSN, /SQSP），部分配列ファミリー検索（/SQSFP）で（L）演算子と（NOTL）演算子を利用できる．

－（L）：指定した複数の部分配列が同一鎖中にあるレコードを検索する．
例 ：=> S (GIVEQCCTSI(L)GERGFFYTPK)/SQSP

・ （L）演算子を 2 個以上利用する場合は「コード/検索タイプ」の式を（L）演算子で組み合わせる必要がある．
例 ：=> S GIVEQCCTSI/SQSP(L)DFVNQHLCGS/SQSP(L)GERGFFYTPK/SQSP

－（NOTL）：指定した部分配列が同一鎖中にないレコードを検索する．

■ アラート（自動 SDI 検索）

・ REGISTRY ファイルの完全配列検索・部分配列検索，BLAST ホモロジー検索は，アラートを登録することができる．

塩基コード・アミノ酸コード

■ 核酸の塩基コード

・ 塩基コード（Symbol）と曖昧コード（Ambiguity Codes）がある.
=> HELP NUC で確認できる.

塩基コード	塩基名	完全配列でヒットするコード	部分配列でヒットするコード
A	adenine	A	A
C	cytosine	C	C
G	guanine	G	G
I	inosine	I	I
T	thymine (in DNA)	T	T, U
U	uracil (in RNA)	U	T, U
曖昧コード	**定義**	**完全配列でヒットするコード**	**部分配列でヒットするコード**
Y	pyrimidine	Y	A, G, R, X 以外
R	purine	R	C, T, U, X, Y 以外
M	amino	M	G, T, U, K, X 以外
K	keto	K	A, C, M, X 以外
S	strong interaction (3 H bonds)	S	A, T, U, W, X 以外
W	weak interaction (2 H bonds)	W	G, C, S, X 以外
B	not-A	B	A, X 以外
V	not-T, not-U	V	T, U, X 以外
D	not-C	D	C, X 以外
H	not-G	H	G, X 以外
N	unknown nucleotide	N	すべて (any)
X	uncommon nucleotide	X	X
Z	non-specific nucleotide	使用不可	A, C, G, T, U, I 以外

・ 曖昧コードは GenBank 由来の核酸配列のレコードで主に使用されている.

・ 完全配列検索 : => S 曖昧コードを含む配列/SQEN

- 曖昧コードを完全配列で検索すると, 特定のコードのみが検索される.
（例 : Y/SQEN で Y がヒットする）
そのため, ヒットする核酸配列のレコードは主に GenBank 由来である.

・ 部分配列検索 : => S 曖昧コードを含む配列/SQSN

- 曖昧コードを部分配列で検索すると, 定義に沿うコードがヒットする.
（例 : Y は A, G, R, X 以外のコードでヒット）
そのため, GenBank 由来以外の CAplus/CA ファイル由来の配列レコードもヒットする.

- ただし, 曖昧コードを [ ] で囲むと, 特定のコードのみが検索される.
（例 : [B][R][Y]/SQSN とすると BRY がヒットする）
この場合には, ヒットする核酸配列のレコードは主に GenBank 由来である.

■ アミノ酸コード

・ 一般アミノ酸コードと特殊・未定義のアミノ酸コードがある.

・ => HELP AAC (一般アミノ酸コード), => HELP AAU (特殊・未定義のアミノ酸コード) で確認できる.

■ 一般アミノ酸コード

・ 1 文字コードと 3 文字コードが利用できる.

1 文字コード	3 文字コード	アミノ酸名	完全配列でヒットするコード	部分配列でヒットするコード
A	ALA	Alanine	A	A
B *	ASX	Aspartic acid or Asparagine	B	D, N
C	CYS	Cysteine	C	C
D	ASP	Aspartic acid	D	D
E	GLU	Glutamic acid	E	E
F	PHE	Phenylalanine	F	F
G	GLY	Glycine	G	G
H	HIS	Histidine	H	H
I	ILE	Isoleucine	I	I
J *	XLE	Isoleucine or Leucine	J	I, L
K	LYS	Lysine	K	K
L	LEU	Leucine	L	L
M	MET	Methionine	M	M
N	ASN	Asparagine	N	N
O	PYL	Pyrrolysine	O	O
P	PRO	Proline	P	P
Q	GLN	Glutamine	Q	Q
R	ARG	Arginine	R	R
S	SER	Serine	S	S
T	THR	Threonine	T	T
U	SCY	Selenocysteine	U	U
V	VAL	Valine	V	V
W	TRP	Tryptophan	W	W
X	XXX	Uncommon or Unspecified	X	X
Y	TYR	Tyrosine	Y	Y
Z *	GLX	Glutamic acid or Glutamine	Z	E, Q

＊ B, J, Z は /SQSP, /SQSFP の検索で利用する.

	ヒットするコード	
	/SQSP*1	/SQEFP,/SQSFP
B	D, N	B, D, N, E, Q, Z
Z	E, Q	B, D, N, E, Q, Z
J	I, L	J, I, L, M, V

*1 /SQSP 検索では B, Z, J そのもののコードはヒットしない.

・ 3 文字コードを利用するときは，以下の 2 通りの方法がある.
=> S 'Ala-Thr-Phe-Pro-Tyr-Trp'/SQEP
=> S 'Ala''Thr''Phe''Pro''Tyr''Trp'/SQEP

・ 1 文字コードと 3 文字コードの組み合わせも利用できる.
=> S VYTA'ABU'EM/SQSP

・ 3 文字コードは REGISTRY BLAST では利用できない.

■ 特殊・未定義のアミノ酸コード

・ 3 文字コードで検索する.

- SQIDE3 表示形式では 3 文字コードが表示される.

- SQIDE 表示形式では特殊・未定義のアミノ酸コードは，すべて X として表示される.
ただし，NTE フィールドで X に対応する特殊・未定義のアミノ酸が判別できる.

- X (XXX) で検索すると，特殊・未定義のアミノ酸コードをまとめて検索できる.

コード	アミノ酸名
AAA *1	alpha-amino acid
AAD	2-aminoadipic acid (2-aminohexanedioic acid)
AAN	alpha-asparagine
ABU	2-aminobutanoic acid
ACA	2-aminocapric acid (2-aminodecanoic acid)
AGN	alpha-glutamine
AIB	alpha-aminoisobutyric acid (2-aminoalanine)
APM	2-aminopimelic acid (2-aminoheptanedioic acid)
APP	gamma-amino-beta-hydroxybenzenepentanoic acid
ASU	2-aminosuberic acid (2-aminooctanedioic acid)
AZE	2-carboxyazetidine
BAL	beta-alanine
BAS	beta-aspartic acid
BLY	3,6-diaminohexanoic acid (beta-lysine)
BUA	butanoic acid
BUX	4-amino-3-hydroxybutanoic acid
CAP	gamma-amino-beta-hydroxycyclohexanepentanoic acid
CHA *2	3-cyclohexylalanine
CIT	N5-aminocarbonylornithine
CYA	3-sulfoalanine
DAB	2,4-diaminobutanoic acid
DPM	diaminopimelic acid
DPR	2,3-diaminopropanoic acid
DSU	2,7-diaminosuberic acid (2,7-diaminooctanedioic acid)
EDC	S-ethylthiocysteine
GGU	gamma-glutamic acid
GLA	gamma-carboxyglutamic acid
GLC	hydroxyacetic acid (glycolic acid)
GLP	pyroglutamic acid
HAR	homoarginine
HCY	homocysteine
HHS	homohistidine
HIV	2-hydroxyisovaleric acid

*1 一般および特殊アミノ酸以外の alpha アミノ酸は AAA

*2 2006 年に導入されたコード

（続き）

コード	アミノ酸名
HSE	homoserine
HVA	2-hydroxypentanoic acid
HYL	5-hydroxylysine
HYP	4-hydroxyproline
IVA	isovaline
LAC	2-hydroxypropanoic acid (lactic acid)
MAA	mercaptoacetic acid
MBA	mercaptobutanoic acid
MHP	4-methyl-3-hydroxyproline
MPA	mercaptopropanoic acid
NAL *2	3-naphtylalanine
NLE	norleucine
NTY	nortyrosine
NVA	norvaline
OAA *3	omega-amino acid
OIC	2-carboxyoctahydroindole
ORN	ornithine
PEN	penicillamine (3-mercaptovaline)
PHG	2-phenylglycine
PIP	2-carboxypiperidine
SAR	sarcosine (N-methylglycine)
SPG	1-amino-1-carboxycyclopentane
STA	statine (4-amino-3-hydroxy-6-methylheptanoic acid)
THI	3-thienylalanine
TIC	1,2,3,4-tetrahydro-3-isoquinolinecarboxylic acid
TLE *2	3-methylvaline
TML	epsilon-N-trimethyllysine
TZA	3-thiazolylalanine
UND *4	undefined
WIL	alpha-amino-2,4-dioxopyrimidinepropanoic acid

*2 2006 年に導入されたコード

*3 骨格中にヘテロ原子が含まれているアミノ酸

*4 ペプチド鎖中に挿入されている非アミノ酸グループとして定義される

■ アミノ酸の等価表

・ ファミリー検索では，等価のアミノ酸を含めて検索できる.

等価クラス	アミノ酸コード
親水性，酸・アミド	D, E, N, Q, B*, Z*
親水性，塩基	H, K, R
弱疎水性，中性	A, G, P, S, T
疎水性	I, L, M, V, J*
疎水性，芳香族	F, Y, W
橋かけ形状	C

＊B, J, Z にする注意点は P.11 表の下の ＊ 参照

ギャップ記号・特殊記号

■ 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.

・ ギャップ記号・特殊記号は部分配列検索（/SQSN, /SQSP）と部分配列ファミリー検索（/SQSFP）でのみ利用することができる.

– ただし & 記号は完全配列検索（/SQEN, /SQEP）と完全配列ファミリー検索（/SQEFP）でも利用できる.

・ 同一質問式中の実行順序

?, *, +（数の指定）> { }（繰り返し数の指定）> &（複数配列の結合）> |（代替配列指定）

■ ギャップ記号

記号	定義	入力例	ヒット例*
.	1 残基のギャップ	=> S GT.C..A/SQSN => S SY...RPG/SQSP	GTGCCCA SYYGLRPG
.{m} または [m.]	m 残基のギャップ	=> S CG.{3}GG/SQSN => S RGP[5.]SA/SQSP	CGTCAGG RGPEAQAESA
.{m,n} または .{m-n}	m から n 残基のギャップ	=> S AA.{2,5}GGC/SQSN => S RGP.{3-4}ES/SQSFP	AAGCTGGC RGPEAQAES
: または .? または .{0,1} または .{0-1}	ゼロまたは 1 残基のギャップ	=> S CCG:CAG/SQSN => S CCG.?CAG/SQSN => S AQAE.{0,1}SAA/SQSP => S AQAE.{0-1}SAA/SQSP	CCGTCAG CCGCAG AQAEESAA AQAESAA
.* または .{0,} または .{0-}	ゼロ残基以上のギャップ	=> S GGCCT.*GT/SQSN => S TACCGT.{0-}CGTA/SQSN => S SAA.{0,}AREAE/SQSFP	GGCCTACCGT TACCGTCGTA SAAEAREAE
.+ または .{1,} または .{1-}	1 残基以上のギャップ	=> S TGCC.{1-}AG/SQSN => S CGGCC.{1,}ACCGT/SQSN => S SG.+MEH/SQSFP	TGCCCAAG CGGCCTACCGT SGPYKMEH

＊ 網掛け部分はギャップ記号でヒットしたコード

■ 特殊記号

記号	定義	入力例	ヒット例
[ ]	代替残基の指定	=> S TT[GAT]CA/SQSN	TTGCA TTACA TTTCA
[-]	特定の代替残基の除外	=> S AAR[-AE]LEY/SQSP	AARVLEY AARGLEYA
{m}	直前の配列*1 を m 回繰り返す	=> S GG(AT){2}CG/SQSN => S GG(FL){3}FI/SQSP	GGATATCGC GGFLFLFLFI
{m,n} または {m-n}	直前の配列*1 を m 回から n 回繰り返す	=> S CTCTA{0-2}GGAT/SQSN => S G(LS){1-2}RI/SQSP	CTCTGGATA CTCTAGGAT CTCTAAGGAT GLSRIGR GLSLSRI
? または {0,1} または {0-1}	直前の配列*1 をゼロまたは 1 回繰り返す	=> S CTAGG?TTA/SQSN => S FLRRI(RP){0-1}K/SQSP	CTAGTTA CTAGGTTAA FLRRIKFR FLRRIRPK
* または {0,} または {0-}	直前の配列*1 をゼロ回以上 繰り返す	=> S GCAT(CTG)*TTAA/SQSN => S ALAS(GG){0,}/SQSP	GCATTTAAG GCATCTGTTAA ALASGGA ALASGGGGS
+ または {1,} または {1-}	直前の配列*1 を 1 回以上繰り 返す	=> S CAT(CTG){1-}TT/SQSN => S KLK(DL)+AL/SQSFP	CATCTGTT CATCTGCTGTTC KLKDLAL KLKDLDLAL
&*2	配列質問式 (L#) を結合する	=> S L1&L2&L3 => S L1&L2&L3/SQSP	
|	代替配列の指定	=> S S(ACD|KLM)F/SQSP	SACDF SKLMF
^	最初，または最後の配列コード の指定	=> S ^MCGIL/SQSP => S VCDS^/SQSFP	MCGILAVF RKSCGVCDS

*1　繰り返す配列は直前の配列コード，または括弧内の配列，または配列質問式の L 番号
*2　完全配列 (/SQEN, /SQEP)，完全配列ファミリー検索 (/SQEFP) でも利用できる

■ 検索例 1：核酸の配列（GCCCAAGCTGGCATCCGTCA）を相補鎖を含めて検索する.

検索のポイント

・ 完全配列，部分配列検索では，相補鎖は自動的に含まれないので別途検索する必要がある．また，相補鎖も 5’末端から 3’末端の並びで入力して検索する必要がある.

今回の検索例の場合

5’ 3’
gcccaagctggcatccgtca（質問式 1）

cgggttcgaccgtaggcagt（相補鎖） ⟹ tgacggatgccagcttgggc（相補鎖の質問式 2）
3’ 5’ 5’ 3’

<相補鎖も含めた完全配列検索>

=> S 質問式 1/SQEN OR 質問式 2/SQEN

- 完全配列検索では，配列長 200 以内の配列については EXPAND で確認できる.

<相補鎖も含めた部分配列検索>

=> S（質問式 1|質問式 2)/SQSN

- 特殊記号 |（代替配列の指定）は REGISTRY ファイルの完全配列検索では利用できない.

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

=> E GCCCAAGCTGGCATCCGTCA/SQEN 5          ← 塩基コードを /SQEN フィールドで EXPAND する
E1            1     GCCCAAGCTGGCATCCGT/SQEN
E2            4     GCCCAAGCTGGCATCCGTC/SQEN
E3          239 --> GCCCAAGCTGGCATCCGTCA/SQEN
E4            1     GCCCAAGCTGGCATCCGTCAG/SQEN
E5           16     GCCCAAGCTGGCAUCCGUCA/SQEN

=> S E3                                   ← E3 を検索する（完全配列検索）
          239 GCCCAAGCTGGCATCCGTCA/SQEN
       931765 SQL=20                      ● 配列長（SQL）も自動的に検索される
L1        239 (GCCCAAGCTGGCATCCGTCA)/SQEN
                (GCCCAAGCTGGCATCCGTCA/SQEN AND SQL=20)
```

```
=> E TGACGGATGCCAGCTTGGGC/SQEN 5          ← 相補鎖の塩基コードを EXPAND する
E1            1     TGACGGATGATTACGCCGAGATTGGTTTATTGGAGGGCGAGGGTGATTACTCTACACCTA
                          :
                    GCTCCCTAAAGAATGGCAAA/SQEN
E2            1     TGACGGATGCATTAAACTTGCCAAGTGA/SQEN
E3            1 --> TGACGGATGCCAGCTTGGGC/SQEN
E4            2     TGACGGATGCCAGCTTGGGCT/SQEN
E5            1     TGACGGATGCCAGGAGGTGAGTTTCATG/SQEN

=> S E3                                   ← E3 を検索する
            1 TGACGGATGCCAGCTTGGGC/SQEN
       931765 SQL=20
L2          1 (TGACGGATGCCAGCTTGGGC)/SQEN
              (TGACGGATGCCAGCTTGGGC/SQEN AND SQL=20)

=> S L1 OR L2                             ← 相補鎖も含めた検索結果
L3        240 L1 OR L2

=> D L3 5 SQIDE                           ← SQIDE 表示形式で表示する

L3   ANSWER 5 OF 240  REGISTRY  COPYRIGHT 2018 ACS on STN
RN   1523583-51-8  REGISTRY
CN   DNA, d(G-C-C-C-A-A-G-C-T-G-G-C-A-T-C-C-G-T-C-A)  (CA INDEX NAME)
OTHER NAMES:
CN   22: PN: WO2013192233 SEQID: 22 unclaimed DNA
FS   NUCLEIC ACID SEQUENCE
SQL  20                                   ← 配列長
NA   4 a   8 c   5 g   3 t

PATENT ANNOTATIONS (PNTE):                ← 特許情報
Sequence |Patent
Source   |Reference
=========+============
Not Given|WO2013192233
         |unclaimed
         |SEQID 22

SEQ      1 gcccaagctg gcatccgtca
           ========== ==========
HITS AT:   1-20
```

配列，配列長，特徴表（NTE），特許情報（PNTE）を確認するには，SQIDE 表示形式を使用する

完全に一致した配列
ヒットしたコードには二重下線（=）が付く

ヒット位置

```
**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
MF   Unspecified
CI   MAN
SR   CA
LC   STN Files:   CA, CAPLUS, USPATFULL
DT.CA  CAplus document type:  Patent
RL.P   Roles from patents:  PRP (Properties)
               1 REFERENCES IN FILE CA (1907 TO DATE)
               1 REFERENCES IN FILE CAPLUS (1907 TO DATE)
```

■ 検索例 2：VEALY と GIVEQ を同一鎖中に含むタンパク質を検索する.

検索のポイント

・ 複数の塩基コード，アミノ酸コードが同一鎖中に存在する配列に限定するには，近接演算子（L）を使用する.

=> S（アミノ酸コード(L)アミノ酸コード）/SQSP　← *タンパク質の部分配列検索*

－（L）演算子は，部分配列検索（/SQSN, /SQSP），部分配列ファミリー検索（/SQSFP）で利用できる.

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

=> S (VEALY(L)GIVEQ)/SQSP
         11876 VEALY/SQSP
         10231 GIVEQ
L1        2034 (VEALY(L)GIVEQ)/SQSP

=> D 1 SQIDE                    ← SQIDE 表示形式で表示する

L1   ANSWER 1 OF 2034  REGISTRY  COPYRIGHT 2018 ACS on STN
RN   2229062-30-8  REGISTRY
CN   Insulin (synthetic human) fusion protein with immunoglobulin G1 (synthetic
     human Fc region containing fragment isoform)  (CA INDEX NAME)
OTHER NAMES:
CN   5: PN: WO2018107117 SEQID: 5 claimed protein
FS   PROTEIN SEQUENCE
SQL  288

PATENT ANNOTATIONS (PNTE):
Sequence |Patent
Source   |Reference
=========+=============
Not Given|WO2018107117
         |claimed SEQID
         |5


SEQ      1 FVNQHLCGSH LVEALYLVCG EEGFFYTPKA AKGIVEQCCT SICSLYQLEN
                     =====                  =====
        51 YCNGGGGAGG GGDKTHTCPP CPAPELLGGP SVFLFPPKPK DTLMISRTPE
       101 VTCVVVDVSH EDPEVKFNWY VDGVEVHNAK TKPREEQYNS TYRVVSVLTV
       151 LHQDWLNGKE YKCKVSNKAL PAPIEKTISK AKGQPREPQV YTLPPSRDEL
       201 TKNQVSLTCL VKGFYPSDIA VEWESNGQPE NNYKTTPPVL DSDGSFFLYS
       251 KLTVDKSRWQ QGNVFSCSVM HEALHNHYTQ KSLSLSPG
HITS AT:   12-16, 33-37
MF   Unspecified
CI   MAN
SR   CA
LC   STN Files:   CA, CAPLUS
DT.CA  CAplus document type:  Patent
RL.P   Roles from patents:  BIOL (Biological study); PREP (Preparation); PRP
       (Properties); USES (Uses)
              1 REFERENCES IN FILE CA (1907 TO DATE)
              1 REFERENCES IN FILE CAPLUS (1907 TO DATE)
```

2 つのアミノ酸コードが同一鎖中に存在する配列を検索する

(L) 演算子は，入力した順序に関係なく，VEALY と GIVEQ が同一鎖中に存在するレコードがヒットする

■ 配列関連検索フィールド

フィールド	検索項目		句/単語	前方一致	後方・中間一致[*1]	入力例
なし または /BI	基本索引	CAS RN®	単語	○	×	S 91449-61-5
		名称セグメント				S INSULIN S PET191 REDUCED
		成分分子式				S C62H108N12O14
		特許関連情報				S WO0154474 S CLAIMED S UNCLAIMED
/CN	化学物質完全名称		句	○	×	S INTERFERON?/CN S GENBANK M12334/CN
/CNS	自然名称セグメント		単語	○	○	S ?INSULIN?/CNS
/FS	ファイルセグメント		句	○	×	S L2 AND PS/FS S L2 AND NS/FS
/ED[*2]	入力日		—	○	×	S L1 AND 19990101<=ED
/SQL[*2]	配列長		—	×	×	S L3 AND 4-20/SQL S L5 AND SQL<=500
/NA[*2]	塩基の種類（と数）		—	×	×	S 12-42 A/NA[*3] S L7 AND G/NA
/NA.CNT[*2]	特定の種類の塩基数		—	×	×	S 13-15/NA.CNT
/NTE[*4]	特徴表		単語	○	○	S L8 AND CYCLIC/NTE S L8 AND CHLORO?/NTE S OAA 17/NTE
/PNTE[*5] または /FEAT[*5]	特許情報		単語	○	×	S L7 AND CLAIM?/PNTE S L7 AND UNCLAIM?/PNTE S SEQID 104/PNTE
/PN[*5]	特許番号		句	○	×	S JP11243959/PN
/PC[*5]	特許発行国		句	○	×	S L7 AND JP/PC
/MF	完全分子式		句	○	×	S C62H108N12O14/MF
/SR	収録源		句	○	×	S L7 AND GENBANK/SR
/LC	CAS RN® 所在		句	○	×	S L1 AND GENBANK/LC

*1 中間一致・後方一致検索を利用する場合は，入力語は最低 4 文字は必要．

*2 数値演算子または範囲指定による検索が可能な数値検索フィールド

*3 =>S 12-42 A/NA の検索は，=>S A/NA(S)12-42/NA.CNT が実行される．

*4 /NTE フィールドではスペースは（P）演算子となる．

*5 1999 年 10 月以降，CAS が特許から索引した配列レコードに収録される．

特徴表（NTE）および特許情報（PNTE）フィールド

■ 特徴表（NTE）フィールド

・ クレームされている配列について，配列の形態や修飾の情報が特徴表（NTE）フィールドに収録される.

- ただし，特殊・未定義のアミノ酸コードを有するタンパク質配列が収録される際には，クレームの有無に関わらず，特殊・未定義のアミノ酸コードの定義が特徴表に収録される.

・ 特徴表の情報は /NTE フィールドで検索する*.

- 例 : 化学的修飾を受けている配列
  => S MODIFIED/NTE

・ 特徴表（NTE）フィールドに収録される主な用語は P.23 参照

■ 特許情報（PNTE）フィールド

・ 1999 年 10 月以降，CAS が特許から索引した配列のレコードには，配列の特許情報が特許情報（PNTE）フィールドに収録される.

・ PNTE フィールドに収録されている情報および検索フィールド

収録情報	検索フィールド
収録源の特許の特許番号（STN 形式）	/PN（特許番号），/PC（特許発行国）
クレームされているかどうか 特許明細書中の記載位置（配列番号ほか） 特許についている配列表のアノテーション情報*	/PNTE または /FEAT

＊ WIPO（世界知的所有権機関）の公表した規格に従った用語が使われている.

- 例 : クレームされている配列に限定　　=> S CLAIM?/PNTE
  クレームされていない配列に限定　　=> S UNCLAIM?/PNTE

＊ 化学修飾が多い配列の /NTE フィールドの検索について

システム制限により、スペースを含む用語や近接演算子を使った NTE フィールドの検索では 31 番目以降の修飾情報を同一行に限定することができません.

（例）2612159-69-8 の場合

```
      NTE  modified
      ------------------------------------------------------------------
行数   type            ------ location ------         description
      ------------------------------------------------------------------
  1   modified base     u-1                          2'-fl
  :     :
 30   modified link     u-2          - a-3           P-deoxy
 31   modified link     u-2          - a-3           P-substituted
 32   modified link     a-21         - t-22          P-thio
```

=> S (U (P) THIO)/NTE の検索でヒットしてしまう

■ 特徴表（NTE）フィールドで使用されている主な用語

NTE 中の用語		内容
一般的用語	modified	化学的修飾を受けている配列； DNA-RNA, RNA-PNA（ペプチド核酸）のように互いに異なる分子種と鎖を形成している場合にも用いる.
	metal salt	金属塩として存在している配列（相手金属も記載される）
	complex	非金属物質あるいは非核酸物質との複合体として存在している配列
	conjugated	少なくとも 2 つ以上のアミノ酸物質（ペプチドやタンパク質も含む）と化学的に結合している核酸配列
	singlestranded	一本鎖構造の核酸配列（2001 年まで利用）
	doublestranded (DS)	二本鎖構造の核酸配列（2001 年まで利用）
	multistranded (#)	二本以上の多鎖構造の核酸配列（カッコ内の数字は鎖の数を表す）
鎖関連の用語	linear	直線状の配列 ；5'- 末端と 3'- 末端はフリー
	cyclic	5'- 末端と 3'- 末端が化学的に結合し, 環状構造を形成している配列; 多鎖構造の配列の場合には, 鎖番号と関連付けされている.
	copolymer	二鎖以上の多鎖構造の核酸配列からできたポリマー; 多鎖構造の配列の場合には, 鎖番号と関連付けされている.
	homopolymer	複数回反復する配列が存在する核酸配列; 反復回数は不明; 多鎖構造の配列の場合には, 鎖番号と関連付けされている.
修飾タイプ関連の用語	modified base	置換, エステル化, 変換反応などによって化学的修飾を受けた塩基や糖を持つ配列 例 ： プリンやピリミジン環に置換基が結合した結果, 縮合, スピロ結合, 架橋結合などで環が新たに形成された塩基 プリンやピリミジン環の炭素や窒素が他元素に置換された塩基
	uncommon base	a, c, g, t, u などの通常の塩基を修飾することでは表現できない塩基; 配列中のコードは X で表される. 例 ： プリンやピリミジン環の一部またはすべてが欠けている塩基 プリンやピリミジン環とは全く異なる環が含まれている塩基 ヘキソースなど通常ではない糖分子を含む塩基
	DNA-containing	DNA 残基を含む RNA 配列
	RNA-containing	RNA 残基を含む DNA 配列
	PNA-containing	DNA あるいは RNA 残基を含む PNA（ペプチド核酸）配列
	modified link	リン酸による置換やエステル化によりリン酸ジエステル結合を形成している配列 ； 結合位置は隣り合う 2 つの塩基の位置番号で示される.
	uncommon link	正常なリン酸結合部位に変換や延長などが起こっている配列; 結合位置は隣り合う 2 つの塩基の位置番号で示される.
	stereoisomer	1 つ以上の糖残基が, α-D-erythro-, β-D-xylo- などの通常にはない立体構造を持つ配列
	metal complex	金属に配位し, 複合体を形成している配列 ； 金属および配位している塩基の位置番号が記載される ； 結合位置不明の場合は ? と表示される
	labeled	配列, 置換基, エステル上のあらゆる原子のラベル状態
	covalent bridge	複数の鎖の間で 2 つ以上の核酸が架橋構造を形成している配列; 鎖番号と架橋結合している位置番号が表示される.
	radical ion	エステル化や置換などによりラジカルイオンを含む配列

■ 検索例 3:「CCGAAT(2～5 の残基)GGC(ゼロ または 1 の残基)CA」の配列を一部に含み,
配列長が 100 以下で,クレームされている核酸配列の特許を調査する.

検索のポイント

・ 配列質問式に条件を指定する場合は,ギャップ記号を利用する.

.{m,n} : m から n 残基のギャップ
:(コロン) : ゼロまたは 1 残基のギャップ

・ 配列の長さは配列長 (/SQL) で指定する.

・ 特許に関する情報は特許情報 (/PNTE) で指定する.

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

=> S CCGAAT.{2,5}GGC:CA/SQSN
L1       35793 CCGAAT.{2,5}GGC:CA/SQSN

=> S L1 AND SQL=<100
L2         105 L1 AND SQL=<100

=> S L2 AND CLAIM?/PNTE
L3          35 L2 AND CLAIM?/PNTE

=> D L3 15 SQIDE                          ← SQIDE 表示形式で表示する

L3   ANSWER 15 OF 35  REGISTRY  COPYRIGHT 2018 ACS on STN
RN   942242-17-3  REGISTRY
CN   DNA, d(C-C-C-G-A-A-T-T-C-G-G-C-A-C-A-A-G-C-A-T-A-A-G-C)  (CA INDEX NAME)
OTHER NAMES:
CN   1: PN: CN1974788 PAGE: 3 claimed sequence
FS   NUCLEIC ACID SEQUENCE
SQL  24
NA   8 a   8 c   5 g   3 t

PATENT ANNOTATIONS (PNTE):
Sequence |Patent
Source   |Reference
=========+==============
Not Given|CN1974788
         |claimed PAGE 3


SEQ      1 cccgaattcg gcacaagcat aagc
           ========= =====
HITS AT:   2-15

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
MF   Unspecified
CI   MAN
SR   CA
LC   STN Files:   CA, CAPLUS
DT.CA  CAplus document type:  Patent
RL.P   Roles from patents:  BIOL (Biological study); PRP (Properties); USES
       (Uses)
               1 REFERENCES IN FILE CA (1907 TO DATE)
               1 REFERENCES IN FILE CAPLUS (1907 TO DATE)
```

.{m,n} は m から n 残基のギャップ,
: は ゼロまたは 1 残基のギャップを検索できる

配列長 (SQL) を 100 以下に限定

クレームされているものに限定する
＊1999 年 10 月以降に索引した特許の配列レコードに
PNTE フィールドが収録されている

.{2,5} の検索式に対して 2 残基の TC がヒットし,
: の検索式に対しては 1 残基でヒットしている

・ 文献情報を得るには，CAplus ファイルへクロスオーバー検索する．

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

=> S L3                         ← L3 をクロスオーバー検索する
L4          24 L3

=> D L4 2 BIB ABS HITSEQ        ← BIB ABS HITSEQ 表示形式で表示する

L4   ANSWER 2 OF 24  CAPLUS  COPYRIGHT 2018 ACS on STN
PatentPak PDF | PatentPak PDF+ | PatentPak Interactive
AN   2016:66460  CAPLUS Full-text
DN   164:170145
TI   Methods for identification and quantification of RNA transcript variants
     in samples analyzed by RNA sequencing, microarray or quantitative PCR
IN   Paul, Lukas; Kubala, Petra; Reda, Torsten
PA   Lexogen GmbH, Austria
SO   PCT Int. Appl., 172pp.
     CODEN: PIXXD2
DT   Patent
LA   English
FAN.CNT 1
PPPI
     PATENT NO.          KIND  DATE      LANGUAGE   PatentPak
     ----------------    ----  --------  ---------- ------------------------
     WO 2016005524        A1   20160114  English    PDF | PDF+ | Interactive
     CN 106471134         A    20170301  Chinese    PDF
     KR 2017028383        A    20170313  Korean     PDF
            :
PI
     PATENT NO.          KIND  DATE        APPLICATION NO.          DATE
     ----------------    ----  --------    ---------------------    --------
     WO 2016005524        A1   20160114    WO 2015-EP65756          20150709
         :
     CN 106471134         A    20170301    CN 2015-80035408         20150709
     KR 2017028383        A    20170313    KR 2017-7002619          20150709
     EP 3167076           A1   20170517    EP 2015-736490           20150709
         :
PRAI EP 2014-176417       A    20140709
     WO 2015-EP65756      W    20150709
ASSIGNMENT HISTORY FOR US PATENT AVAILABLE IN LSUS DISPLAY FORMAT
AB   The present invention relates to the field of transcriptomics and provides
     a method for the controlled identification and/or quantification of
     transcript variants in samples, comprising providing a reference set of
         :
IT   1854162-48-3   1854162-49-4, DNA (synthetic cDNA fragment)
     RL: BUU (Biological use, unclassified); PRP (Properties); BIOL (Biological
     study); USES (Uses)
        (nucleotide sequence; methods for identification and quantification of
        RNA transcript variants in samples analyzed by RNA sequencing,
        microarray or quant. PCR)
RN   1854162-48-3  CAPLUS
CN   DNA, d(G-A-T-A-C-C-G-A-A-T-T-T-A-G-A-G-G-C-C-A-T-A-G-G-T-T-A-T-G-G-A-A-A-A-
     A-G-T-C-A-G-T-G)  (CA INDEX NAME)

SEQ      1 gataccgaat ttagaggcca taggttatgg aaaaagtcag tg


RN   1854162-49-4  CAPLUS
CN   DNA (synthetic cDNA fragment)  (CA INDEX NAME)

SEQ      1 gtggagaagc aaatacttgg ataccgaatt tagaggccat aggttatgga
        51 aaaagtcagt g
```

BIB 書誌情報

ABS 抄録

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

SEQLINK EXACT コマンド

■ 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．

・ 配列の主鎖は同じであっても，化学修飾，側鎖の置換基の異なるもの，同位体で置換された配列は別配列として別レコードに収録される．（化学修飾などの情報は特徴表に表記される）

・ GenBank 由来の更新前と更新後の核酸配列は，同じ配列であっても別レコードとして収録される．

・ P.3 に記載した登録ルールの例外によって，同じ配列であっても別レコードとして収録されている場合がある．

■ SEQLINK EXACT コマンド

・ SEQLINK EXACT コマンドは，上記のような同じ配列（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て，一つの L 番号を作成するコマンドである．

- SEQLINK EXACT は SEQ と省略形で入力できる．

・ CAS RN® や GenBank 番号，名称からの検索結果に対して有効である．

- 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．

・ 入力方法

・ CAS RN® 例 : => SEQLINK EXACT 445115-49-1
=> SEQ 445115-49-1

・ 配列レコードを含む回答セットの L 番号 例 : => SEQ L1

・ CAS RN® に相当する E 番号 例 : => SEQ E3

・ ANALYZE コマンドでまとめた CAS RN® の L 番号
例 : => SEQ L2

■ 検索例 4：CAS RN® 1352257-43-2 の配列と，同じ配列だが別の CAS RN® を持つレコードを含めて検索する.

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

=> S 1352257-43-2                   ← CAS RN® で検索する
L1            1 1352257-43-2
                (1352257-43-2/RN)

=> D SQIDE                          ← SQIDE 表示形式で表示する

L1   ANSWER 1 OF 1  REGISTRY  COPYRIGHT 2018 ACS on STN
RN   1352257-43-2  REGISTRY
CN   DNA (Bacillus amyloliquefaciens strain 13563-0 60-kilodalton molecular
     chaperone GroEL gene cpn60)  (CA INDEX NAME)
OTHER NAMES:
CN   GenBank GU937107
FS   NUCLEIC ACID SEQUENCE
SQL  552
NA   161 a   122 c   136 g   133 t

SEQ      1 gcgactgtgc ttgcacaggc tatgatccgc gaaggcctta aaaacgtaac
        51 tgcgggagct aatcctgtcg gcgtgcgtaa aggtatggaa caagccgtaa
       101 ccgtggcaat cgaaaactta aaagaaattt ctaagccgat cgaaggcaaa
       151 gagtctatcg ctcaggttgc tgcaatctct gctgctgatg aggaagtcgg
       201 aagccttatc gctgaagcaa tggagcgcgt aggaaacgac ggcgttatca
       251 caatcgaaga gtctaaaggc ttcacaactg agcttgaagt ggttgaaggt
       301 atgcaattcg accgcggata tgcgtctcct tacatggtga ctgactctga
       351 taagatggaa gcggttcttg ataatcctta catcttaatc acagacaaaa
       401 aaatcacaaa cattcaagaa atccttcctg tgcttgagca agttgtacag
       451 caaggcaaac cattgcttct gatcgctgaa gatgttgaag gtgaagctct
       501 tgctacactc gttgtcaaca aacttcgcgg cacattcaac gctgttgccg
       551 tt

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
MF   Unspecified
CI   MAN
SR   GenBank
LC   STN Files:   CA, CAPLUS, GENBANK, TOXCENTER
DT.CA  CAplus document type:  Journal
RL.NP  Roles from non-patents:  BIOL (Biological study); PRP (Properties)
               1 REFERENCES IN FILE CA (1907 TO DATE)
               1 REFERENCES IN FILE CAPLUS (1907 TO DATE)


=> SEQ L1

L2            3 SEQLINK EXACT L1

=> S L2 NOT L1                      ← SEQLINK EXACT を実行して，新たに得られた 2 件を
L3            2 L2 NOT L1             確認する
```

GENBANK 由来の配列

当レコードと同じ配列を有する別レコードが存在することを示している

SEQLINK EXACT コマンドを実行すると，同じ配列を有する別レコードをまとめることができる

SEQLINK EXACT コマンドを実行する

=> D 1-2 SQIDE　　← *SQIDE 表示形式で表示する*

```
L3   ANSWER 1 OF 2  REGISTRY  COPYRIGHT 2018 ACS on STN
RN   1352257-33-0  REGISTRY
CN   DNA (Bacillus subtilis strain ATCC 55405 60-kilodalton molecular chaperone
     GroEL gene cpn60)  (CA INDEX NAME)
OTHER NAMES:
CN   GenBank GU937102
FS   NUCLEIC ACID SEQUENCE
SQL  552
NA   161 a   122 c   136 g   133 t

SEQ      1 gcgactgtgc ttgcacaggc tatgatccgc gaaggcctta aaaacgtaac
        51 tgcgggagct aatcctgtcg gcgtgcgtaa aggtatggaa caagccgtaa
       101 ccgtggcaat cgaaaactta aaagaaattt ctaagccgat cgaaggcaaa
       151 gagtctatcg ctcaggttgc tgcaatctct gctgctgatg aggaagtcgg
                 :
       451 caaggcaaac cattgcttct gatcgctgaa gatgttgaag gtgaagctct
       501 tgctacactc gttgtcaaca aacttcgcgg cacattcaac gctgttgccg
       551 tt

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
MF   Unspecified
CI   MAN
SR   GenBank
LC   STN Files:   CA, CAPLUS, GENBANK, TOXCENTER
DT.CA  CAplus document type:  Journal
RL.NP  Roles from non-patents:  BIOL (Biological study); PRP (Properties)
               1 REFERENCES IN FILE CA (1907 TO DATE)
               1 REFERENCES IN FILE CAPLUS (1907 TO DATE)

L3   ANSWER 2 OF 2  REGISTRY  COPYRIGHT 2018 ACS on STN
RN   869754-75-6  REGISTRY
CN   92: PN: US20050260619 SEQID: 92 unclaimed DNA (9CI)  (CA INDEX NAME)
FS   NUCLEIC ACID SEQUENCE
SQL  552
NA   161 a   122 c   136 g   133 t

PATENT ANNOTATIONS (PNTE):
Sequence |Patent
Source   |Reference
=========+=============
Not Given|US20050260619
         |unclaimed
         |SEQID 92

SEQ      1 gcgactgtgc ttgcacaggc tatgatccgc gaaggcctta aaaacgtaac
        51 tgcgggagct aatcctgtcg gcgtgcgtaa aggtatggaa caagccgtaa
       101 ccgtggcaat cgaaaactta aaagaaattt ctaagccgat cgaaggcaaa
       151 gagtctatcg ctcaggttgc tgcaatctct gctgctgatg aggaagtcgg
                 :
       451 caaggcaaac cattgcttct gatcgctgaa gatgttgaag gtgaagctct
       501 tgctacactc gttgtcaaca aacttcgcgg cacattcaac gctgttgccg
       551 tt

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

GENBANK 由来の配列

特許由来の配列

配列質問式の長さの制限値と入力方法

■ 配列質問式の長さの制限値

入力方法	核酸*1	タンパク質*1*2
配列コードを直接入力し，一回で検索*3	250 コード	
QUERY コマンドを使用し，一回で入力できるコード数*3	250 コード	
QUERY コマンドで作成した質問式の L 番号（検索フィールドあり）を & 記号でつなげた場合*4	1,000 コード	1,000 コード （/SQEP, /SQEFP） 2,400 コード （/SQSP, /SQSFP）

*1 質問式のコードが短いとシステムリミットに達することがある．その場合にはコードを増やす．
制限値を超えて長い配列を検索する場合は，部分配列検索を実行して，配列長（/SQL）で限定する．
*2 タンパク質は 3 文字コードを使用した場合は 3 コードとカウントするため，3 文字コードを利用すると検索式に含めるアミノ酸は各上限の 3 分の 1 となる．
*3 コード間にスペースは使用しない．
*4 アラートを登録する場合は，& 記号でつなげた検索式のコード数を 240 コード以下に抑える．

■ 配列質問式の入力方法

入力方法	入力例
① 直接入力する．	=> S TTTGGGG/SQEN
② 長い配列を検索したい場合は，1 行あたり 250 コード以下になるように分割した質問式を QUERY コマンドで作成し，最後に & でつないで検索に使用する．QUERY コマンドで作成する質問式には，必ず検索フィールドを含める	=> QUE AGATTATGAG/SQEN L1 QUE AGATTATGAG/SQEN => QUE GACCAGATGA/SQEN L2 QUE GACCAGATGA/SQEN => S L1&L2
③ 同じ配列について異なる検索タイプを実行する場合は，QUERY コマンドで配列コードのみの質問式を作成し，その L 番号に検索フィールドを付与して検索する．	=> QUE TTTGGGGTTT L1 QUE TTTGGGGTTT => S L1/SQEN => S L1/SQSN
④ 他のデータベースの配列検索結果である L 番号を，質問式として利用する．	=> FILE GENESEQ => RUN GETSEQ GACC/SQEN L1 1 GACC/SQEN => FILE REGISTRY => S L1
⑤ /SQEN または /SQEP フィールドで EXPAND した配列コードの E 番号を利用する．検索フィールドを変更すれば，他の検索タイプも実行できる．	=> E AGATT…GATGA/SQEN : E3 1 --> AGATT…GATGA/SQEN => S E3 ← /SQEN で検索 => S E3/SQSN ← 他の検索タイプで検索
⑥ SELECT コマンドで抽出した配列コード（200 コードまで）の E 番号を利用する．検索フィールドを変更すれば，他の検索タイプも実行できる．	=> SEL L1 1 SEQ E1 THROUGH E1 ASSIGNED => S E1 ← /SQSN で検索 => S E1/SQEN ← 他の検索タイプで検索

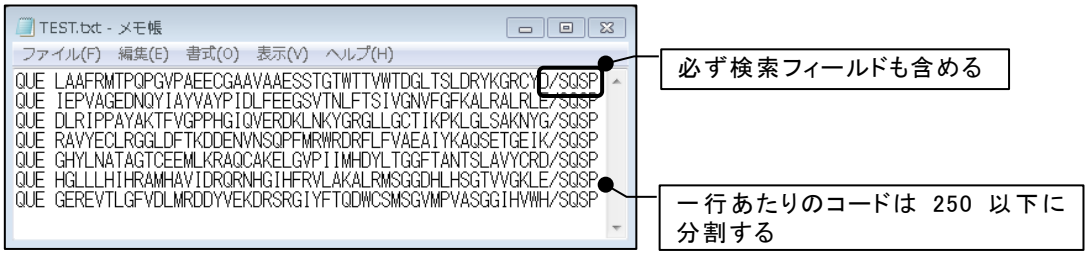
■ 入力例 ：REGISTRY ファイルで，350 コードのアミノ酸配列について部分配列検索を行う．

・ 長い配列を検索したい場合は，1 行あたり 250 コード以下になるように分割した質問式を QUERY コマンドで作成し，最後に & でつないで検索に使用する．

・ QUERY コマンドを直接入力してもよいが，あらかじめテキストファイル（.txt）を用意しておくと，スムーズに検索を実行することができる．

配列質問式のテキストファイルの作成

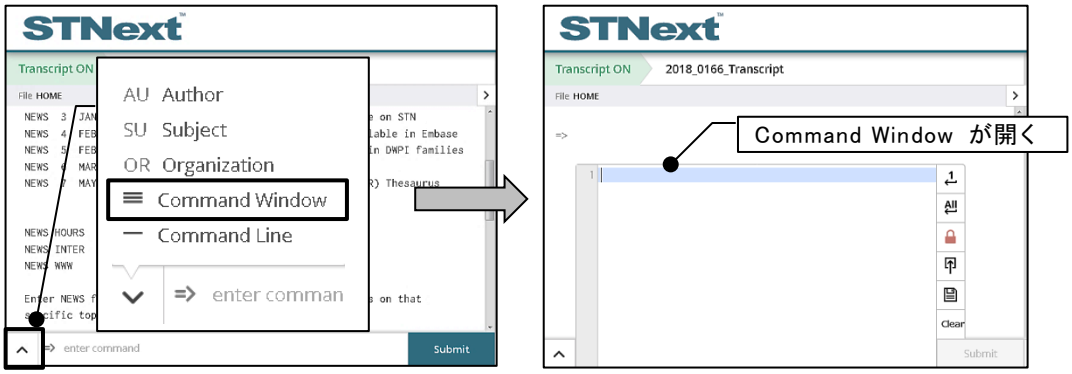
① テキストエディタを起動して，QUERY コマンドで質問式を作成する．（ここでは 1 行あたり 50 コードずつに分割）この時，必ず検索フィールドも入力する．



② 名前を付けて，テキスト文書（.txt）で保存する．

STNext での操作

① REGISTRY ファイルに入り，画面左下の ^ をクリックし，Command Window を選択する．



② Upload script ボタン（ ）をクリックし，Browse より事前作成した配列質問式のテキストファイルを選択し，OK をクリックすると，質問式がアップロードされる．

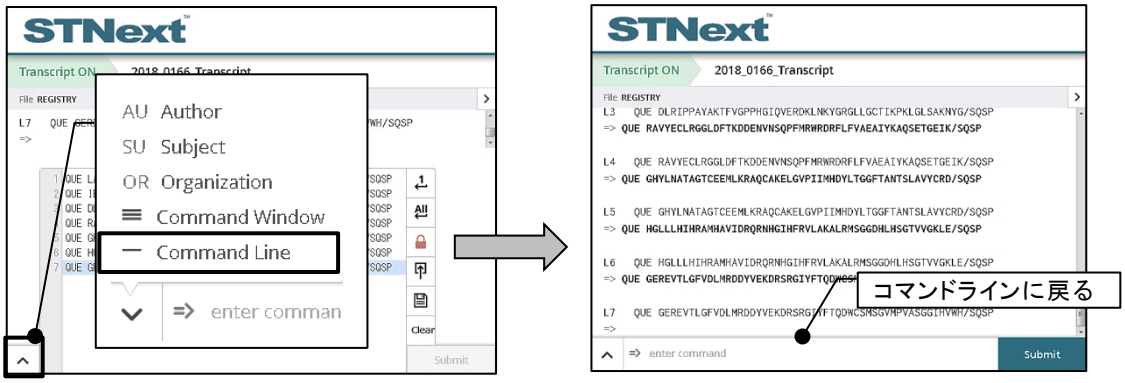
③ Send all lines ボタン（ All ）をクリックすると自動的にコマンドが送信され，L 番号が作成される．

（自動的に送信された内容）

```
=> QUE LAAFRMTPQPGVPAEECGAAVAAESSTGTWTTVWTDGLTSLDRYKGRCYD/SQSP
L1    QUE LAAFRMTPQPGVPAEECGAAVAAESSTGTWTTVWTDGLTSLDRYKGRCYD/SQSP
=> QUE IEPVAGEDNQYIAYVAYPIDLFEEGSVTNLFTSIVGNVFGFKALRALRLE/SQSP
L2    QUE IEPVAGEDNQYIAYVAYPIDLFEEGSVTNLFTSIVGNVFGFKALRALRLE/SQSP
          :
=> QUE HGLLLHIHRAMHAVIDRQRNHGIHFRVLAKALRMSGGDHLHSGTVVGKLE/SQSP
L6    QUE HGLLLHIHRAMHAVIDRQRNHGIHFRVLAKALRMSGGDHLHSGTVVGKLE/SQSP
=> QUE GEREVTLGFVDLMRDDYVEKDRSRGIYFTQDWCSMSGVMPVASGGIHVWH/SQSP
L7    QUE GEREVTLGFVDLMRDDYVEKDRSRGIYFTQDWCSMSGVMPVASGGIHVWH/SQSP
```

自動的にテキストファイルの内容が送信され，L 番号が作成される

④ 画面左下の ^ をクリックし，Command Line を選択すると，コマンドラインに戻る．



⑤ & 記号で L 番号をつなげて，検索を実行する．

```
=> S L1&L2&L3&L4&L5&L6&L7
L8              1 (LAAFRMTPQPGVPAEECGAAVAAESSTGTWTTVWTDGLTSLDRYKGRCYD)(IEPVAGEDNQY
                  IAYVAYPIDLFEEGSVTNLFTSIVGNVFGFKALRALRLE)(DLRIPPAYAKTFVGPPHGIQVER
                  DKLNKYGRGLLGCTIKPKLGLSAKNYG)(RAVYECLRGGLDFTKDDENVNSQPFMRWRDRFLFV
                  AEAIYKAQSETGEIK)(GHYLNATAGTCEEMLKRAQCAKELGVPIIMHDYLTGGFTANTSLAVY
                  CRD)(HGLLLHIHRAMHAVIDRQRNHGIHFRVLAKALRMSGGDHLHSGTVVGKLE)(GEREVTL
                  GFVDLMRDDYVEKDRSRGIYFTQDWCSMSGVMPVASGGIHVWH)/SQSP

=> D SQIDE
      :
```

まとめ

- 核酸の完全配列・部分配列検索
  - 完全配列検索　　: => S コード/SQEN
  - 部分配列検索　　: => S コード/SQSN
    - 核酸の配列検索では，相補鎖が自動的に検索されないので，必要の場合には相補鎖も含めた式を考慮する．
- タンパク質の完全配列・部分配列検索
  - 完全配列検索　　: => S コード/SQEP
  - 部分配列検索　　: => S コード/SQSP
- 同主鎖の配列を持っていても，別レコードになっていることがある．そのため，CAS RN®，名称検索などの辞書検索で配列レコードを得た場合には，SEQLINK EXACT コマンドを実行し，同主鎖のレコードをまとめる．

B REGISTRY BLAST ホモロジー検索

REGISTRY ファイルの BLAST ホモロジー検索は，STN とは切り離された独立したソフトウェアで実行されます．この章では，STNext による REGISTRY BLAST ホモロジー検索の方法を説明します．

REGISTRY ファイルのホモロジー検索

■ REGISTRY ファイルでは，BLAST ホモロジー検索が実行できる．

・ ホモロジー検索の主なプログラム

プログラム	概要	REGISTRY	DGENE/PCTGEN /USGENE
BLAST (ブラスト)	Basic Local Alignment Search Tool 最もよく利用されている．他のプログラムに比べて高速処理できる．ギャップを考慮しないため，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．	○	○
		独立したソフトウェアで検索	RUN BLAST で検索
FASTA (ファストエー)	BLAST と異なりギャップを考慮したアラインメントを行ってくれる．ギャップ付きのアライメントを行うとは言っても，データベースの中から候補を絞り込む段階ではある種の近似が行われており，これによって高速化が図られている．	×	×
GETSIM	FASTA 系列のプログラム．FASTA のように近似を行うことなく，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．このため，計算量は他のプログラムに比べて膨大になる．近似を排して厳密に比較を行うため，進化的に離れた配列であっても，それが統計的に優位である限り見落とすことはないという安心感がある．	×	○
			RUN GETSIM で検索

■ REGISTRY BLAST ホモロジー検索の検索タイプ

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

REGISTRY BLAST ホモロジー検索

■ REGISTRY ファイルでは独立したソフトウェアを利用するため，下記の手順で BLAST ソフトウェアをインストールする．

① ダウンロードサイト https://next.stn.org/stn/downloads/blast-download.html へアクセスし，STN の ID とパスワードを入力して「Login」ボタンをクリックする．

② 「Download」ボタンをクリックし，ソフトウェア（.exe ファイル）をダウンロードする．

③ ダウンロードした .exe ファイルを実行してインストールする．

■ REGISTRY BLAST ホモロジー検索時の配列入力

・ 配列コードのみを入力する．/検索フィールドは付与しない．

・ ギャップ記号・特殊記号は利用できない．

・ アミノ酸の 3 文字コードは利用できないので，1 文字コードを利用する．

・ 曖昧コードも利用できる．

・ 質問式にスペースや行番号は含まれていてもよい．一行の配列コードはスペースを含めて 300 文字以内にする．

■ REGISTRY BLAST ホモロジー検索時の相補鎖

・ 相補鎖も含めて検索される．(「相補鎖を検索しない」設定はできない．)

■ システム制限値

配列質問式の長さ	50,000 コードまで
回答レコードの上限	1,000 件まで（デフォルトは 25 件）
回答セットの上限	自動的に 100 セットまで保存される 101 個目の回答セットが作成される場合は，最も古い回答セットが削除される
アラートの回答セットの上限	登録ごとに最大 20 セットまで，合計 100 セットまで自動的に保存される．101 個目の回答セットが作成される場合は，最も古い回答セットが削除される

REGISTRY BLAST ソフトウェア

■ 検索の流れ（表中の番号は，P.34 以降の検索例に沿う）

準備

初回のみ	REGISTRY BLAST ソフトウェアをインストールする
①	検索したい配列質問式を準備する

BLAST 検索

②～③	REGISTRY BLAST ソフトウェアを起動する
④～⑤	配列質問式を入力する (Genbank 番号または CAS RN® でも入力できる)
⑥～⑧	ホモロジー検索のタイプやパラメータを指定する
⑨	検索を実行する
⑩	結果を確認する（⑪ 印刷・保存）

STN での検索

⑫～⑭	配列検索で得られた結果を STN へ移行するための Script を作成する
⑮	アライメントを含めたレポートを作成するために，アライメントデータ（STNext Saved Sequences (.xss)）を保存する
⑯	STNext で Script をインポートする
⑰	インポートした Script を実行する

レポート作成（オプション）

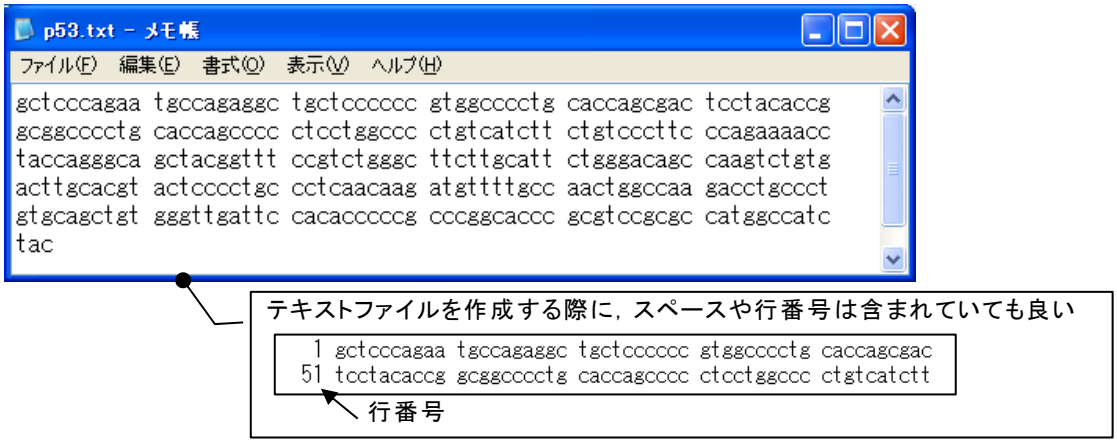
P.46	検索履歴に CAS RN® に対応するアライメントを含めたレポートを作成する

■ 検索例 ： 下記の癌抑制遺伝子 p53 に類似する核酸および関連する文献を調べる.

```
gctcccagaa tgccagaggc tgctcccccc gtggcccctg caccagcgac tcctacaccg
gcggcccctg caccagcccc ctcctggccc ctgtcatctt ctgtcccttc ccagaaaacc
taccagggca gctacggttt ccgtctgggc ttcttgcatt ctgggacagc caagtctgtg
acttgcacgt actcccctgc cctcaacaag atgttttgcc aactggccaa gacctgccct
gtgcagctgt gggttgattc cacacccccg cccggcaccc gcgtccgcgc catggccatc
tac
```

準備

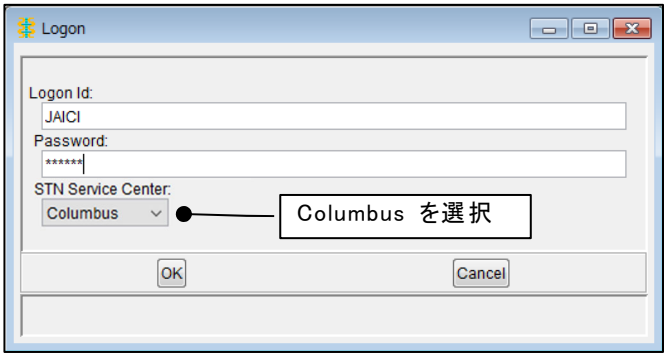
① 長い配列質問式の場合には，テキストファイルを作成しておく.（.txt ファイル）



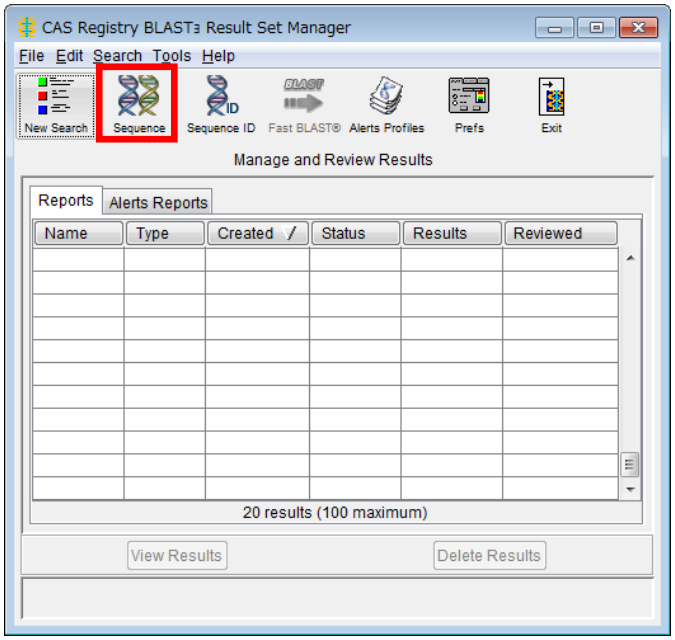
BLAST 検索

② REGISTRY BLAST を起動する.

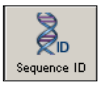
③ STN のログイン ID とパスワードを入力して OK をクリックする.



④ Result Set Manager ダイアログボックスが表示される.

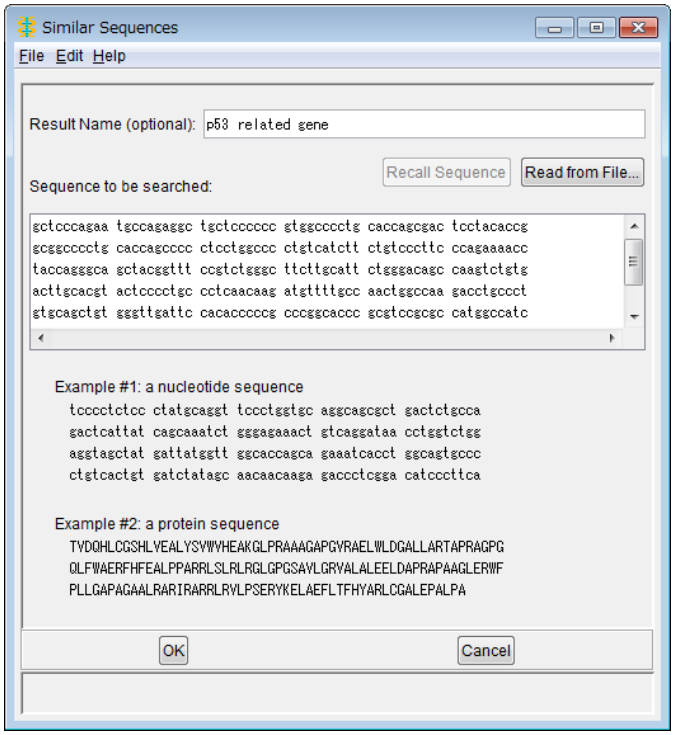


配列から検索を開始する場合は Sequence アイコンをクリックする.



配列の GenBank 番号や CAS RN® から配列質問式を呼び出して検索する場合は Sequence ID アイコンをクリックする.

⑤ ④ で Sequence アイコンをクリックすると, Similar Sequences ダイアログボックスが表示される. ここでは,「Read from File」ボタンをクリックして, ① で用意したテキストファイルを呼び出し,「OK」ボタンをクリックする.

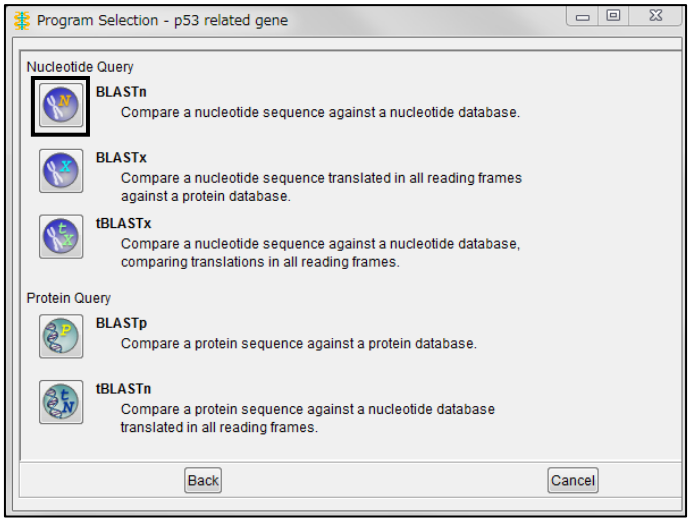


Result Name: に検索名を入力する.

Sequence to be searched: に配列質問式（50,000 コードまで）を以下のいずれかの方法で入力し「OK」ボタンをクリックする.

- 直接入力する.
- 配列データベースからコピーしたデータを貼り付ける.
- 「Read from File」ボタンをクリックしてファイルを呼び出す.
  （text ファイル, GCG または FASTA 形式）

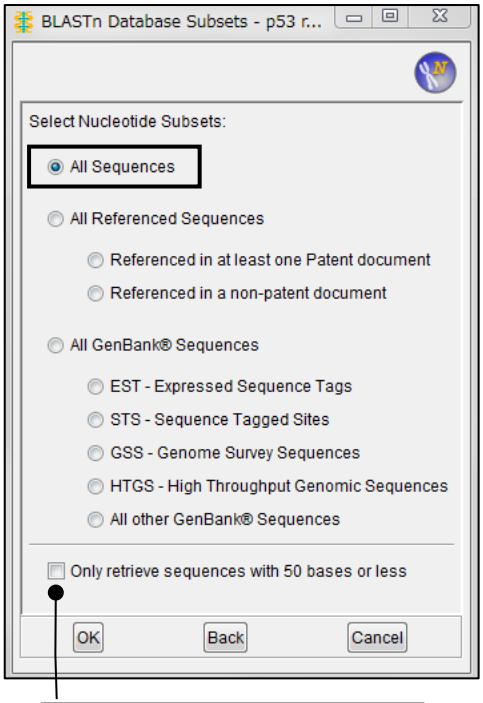
⑥ Program Selection ダイアログボックスが表示されるので，検索タイプを選択する．ここでは「BLASTn」ボタンをクリックする．



⑦ 検索対象となるデータベースのサブセットを選択する．ここで表示されるダイアログボックスは ⑥ で選択した検索タイプによって異なる．

ここでは「All Sequences」を選択し，「OK」ボタンをクリックする．

〈BLASTn〉

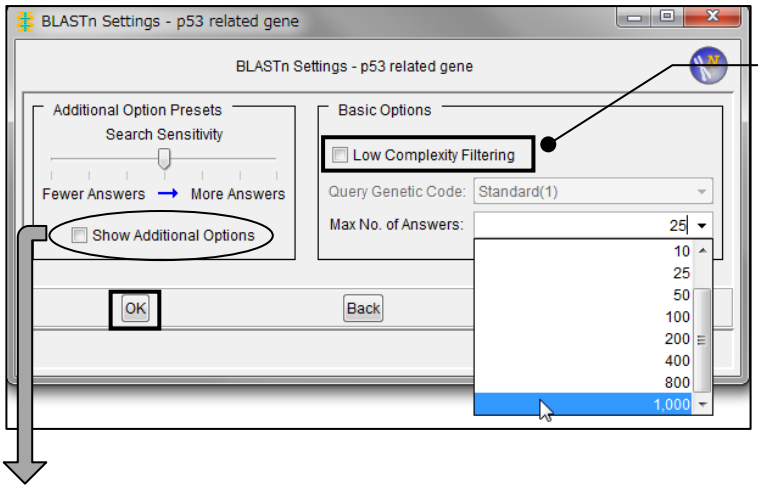


50 塩基以下の配列のみが必要な場合はチェックを付ける

- All Sequences：REGISTRY ファイル中のすべての配列
- All Referenced Sequences：文献のあるすべての配列
  - Referenced in at least one Patent document：特許に記載されている配列
  - Referenced in a non-patent document：特許以外の文献に記載されている配列
- All GenBank Sequences：GenBank ファイルから収録された配列
  - EST：mRNA（cDNA）由来の発現配列タグ
  - STS：ゲノム配列中の配列標識部位
  - GSS：ゲノム調査関連の配列
  - HTGS：High-throughput sequencing センターで発生した配列で，まだシークエンシングが完了していないもの
  - All other GenBank Sequences：GenBank ファイルから収録された配列のうち EST，STS，GSS，HTGS 以外の配列

⑧ BLASTn Settings ダイアログボックスで検索パラメータを指定する.

ここでは,「Low Complexity Filtering」のチェックをはずし,回答件数の最大値を 1,000 にして「OK」ボタンをクリックする.



「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

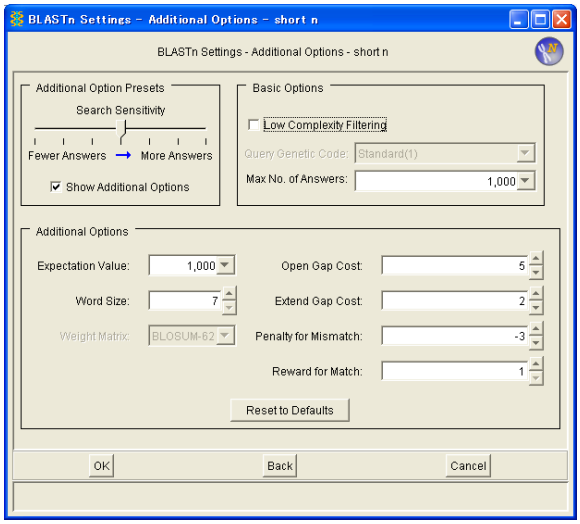
参考 : 配列質問式の配列長によるパラメータの設定

- BLAST ホモロジー検索のパラメータは「Show Additional Options」にチェックを付けると表示される.

- 短い配列長の検索

  - 配列長が 30 以下の配列質問式では下記のようにパラメータの設定を変更する.

検索配	設定値
核酸	・ フィルタリングを行わない ・ 期待値 : 1000 ・ word size : 7
タンパク質	・ フィルタリングを行わない ・ 期待値 : 20000 ・ word size : 2 ・ 置換行列 : PAM30



- 長い配列長の検索

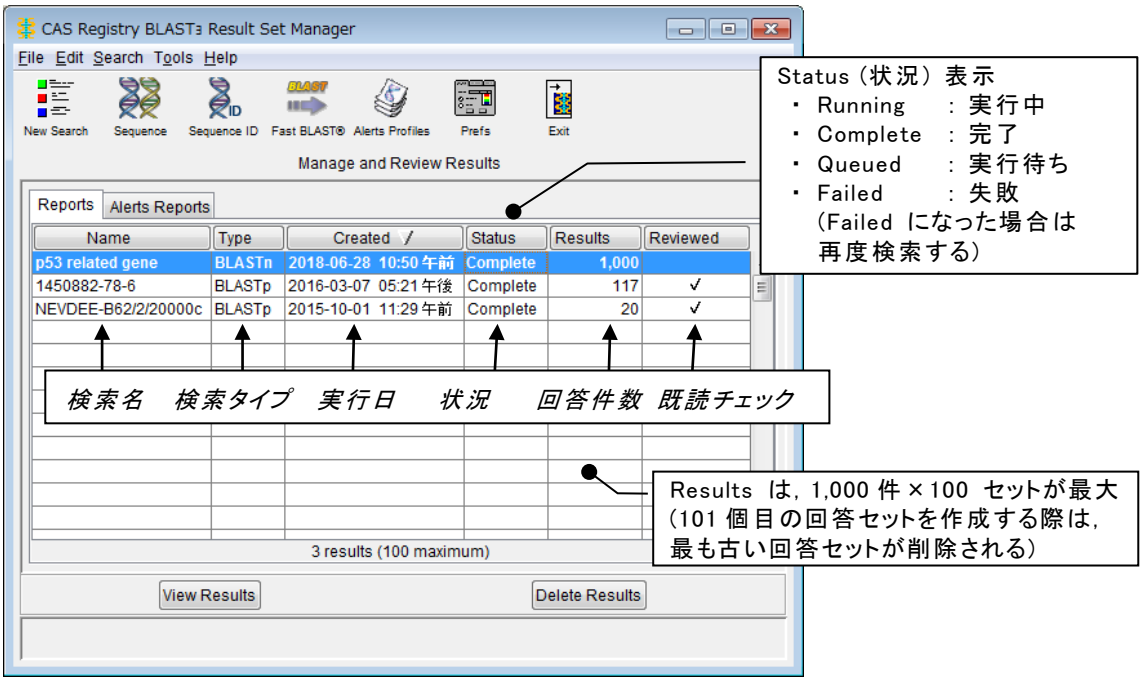
  - 長い配列長の配列質問式を用いると,回答候補が多すぎてタイムアウトが起こり,検索が完了しないことがある.(この時の Status は Failed になる.)この場合は,パラメータの「Word Size」の値を大きくする.

- 各パラメータの詳細は APPENDIX 参照

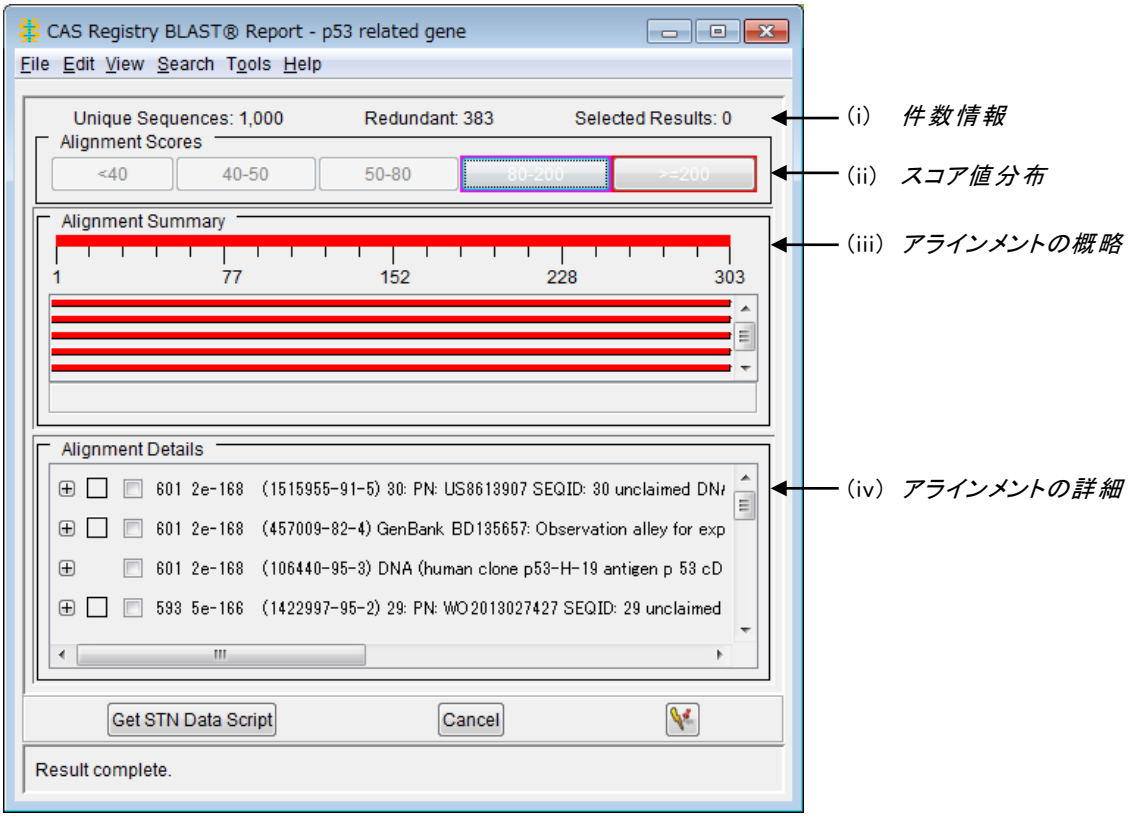
⑨ 検索が開始されると検索状況が表示される．検索実行中は Status の カラムに Running と表示される．検索が完了すると Complete と表示される．

・ Reports タブには，BLAST ホモロジー検索の回答セット（100 セットまで）がリストアップされる．

・ Alerts Reports タブには，アラートの回答セットがリストアップされる．



⑩ ⑨ で検索結果をハイライトし，「View Results」 ボタンをクリックすると，配列質問式との類似性を確認することができる．

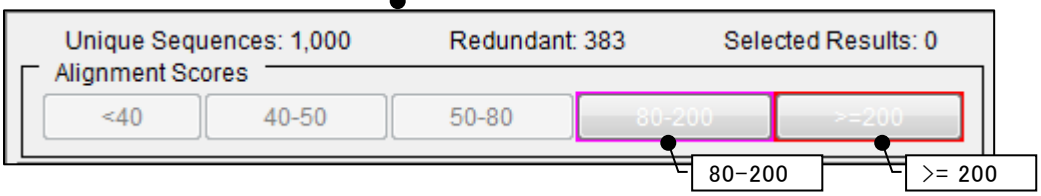


(i) 件数情報

修飾があるものも同じとみなされる

Unique Sequences : ユニークな配列
Redundant : 重複する配列 （同じ主配列であるが，別レコードになっているもの）
Selected Results : 選択した配列

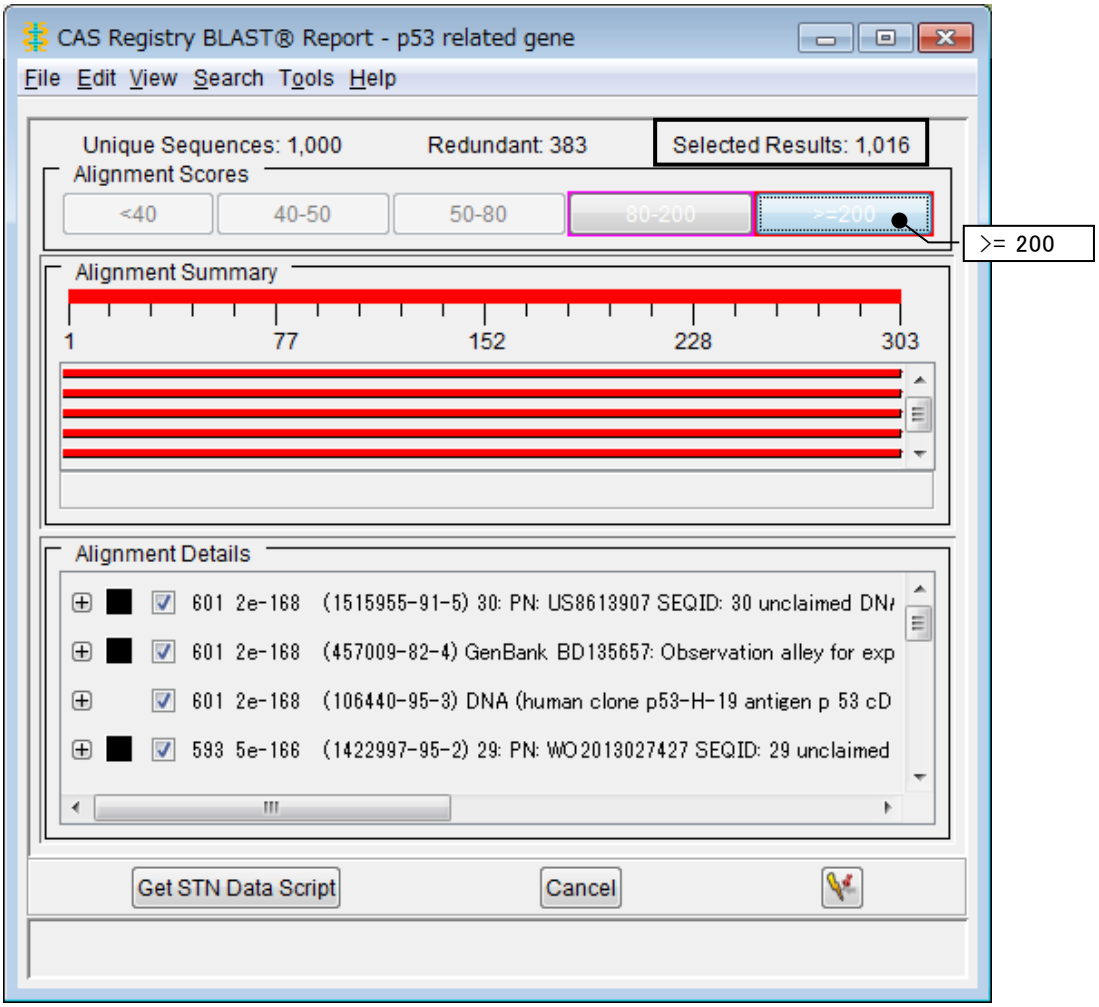
ユニークな配列 1,000 件の中には，重複する配列 383 件は含まない
つまり，実際に得られた配列数は 1,000 + 383 = 1,383 件である



(ii) スコア値分布

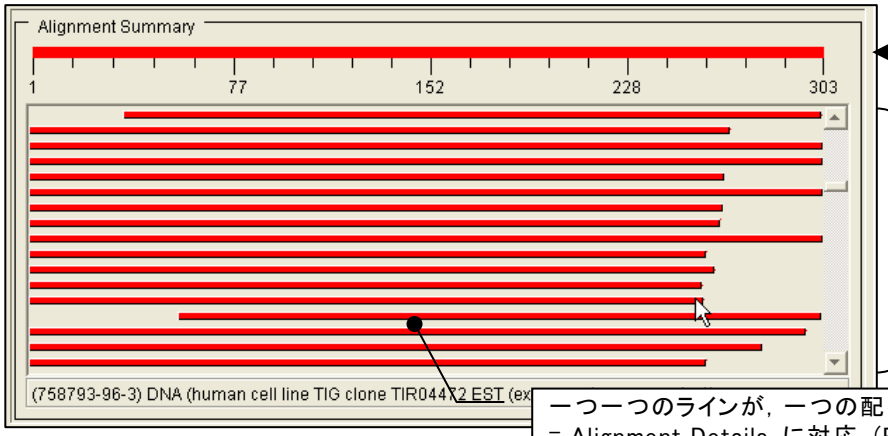
- スコア値（Alignment Score）を大きく五つの範囲に分け，各範囲の配列件数が確認できるようになっている．各範囲のボタンをクリックすると，その範囲の配列が選択され，対応する配列のチェックボックスにチェックが入り，Selected Results に件数が表示される．例えば >=200 のボタンをクリックすると，スコア値 200 以上の配列が 1,016 件であることがわかる．

- 各範囲に配列が存在する場合はボタンに色がつき，スコア値の分布状態を把握しやすいようになっている．

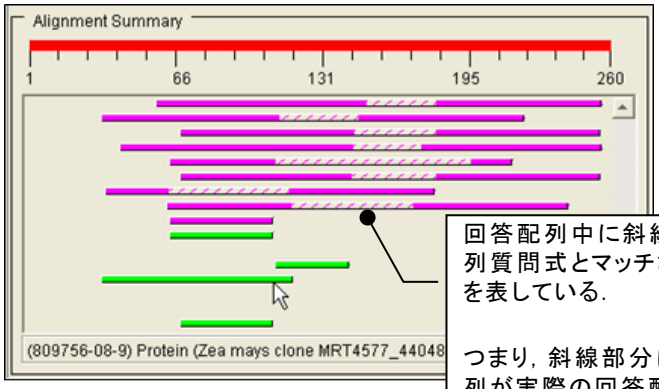


(iii) アライメントの概略

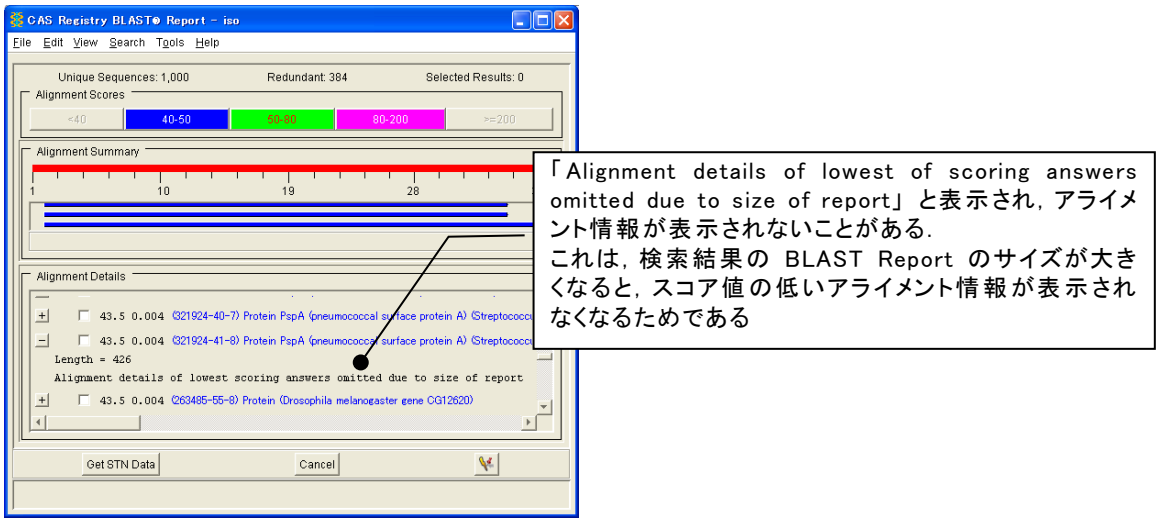
一番上の配列質問式に対して，回答の全配列が類似性の高い順にアライメント表示されている．各配列はスコア値に応じて範囲ボタンと同じ色で表示されている．また，各配列の上にポインタを置くと，下のバーに配列の名称が表示される．



参考：アライメントの概略に表示される斜線部分

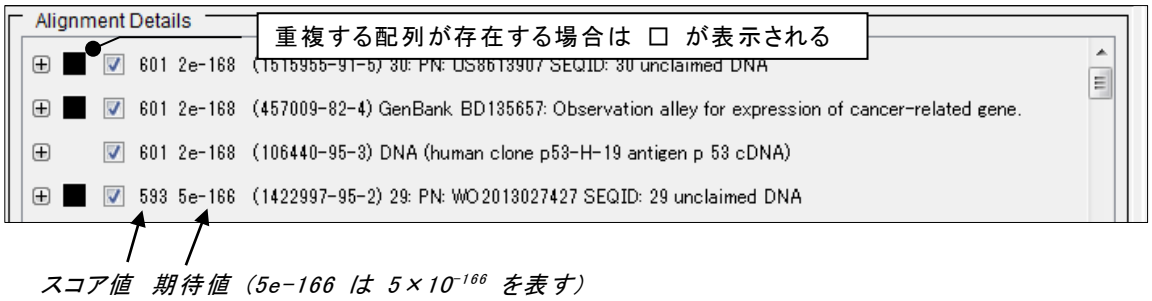


参考：アライメント情報が表示されない場合

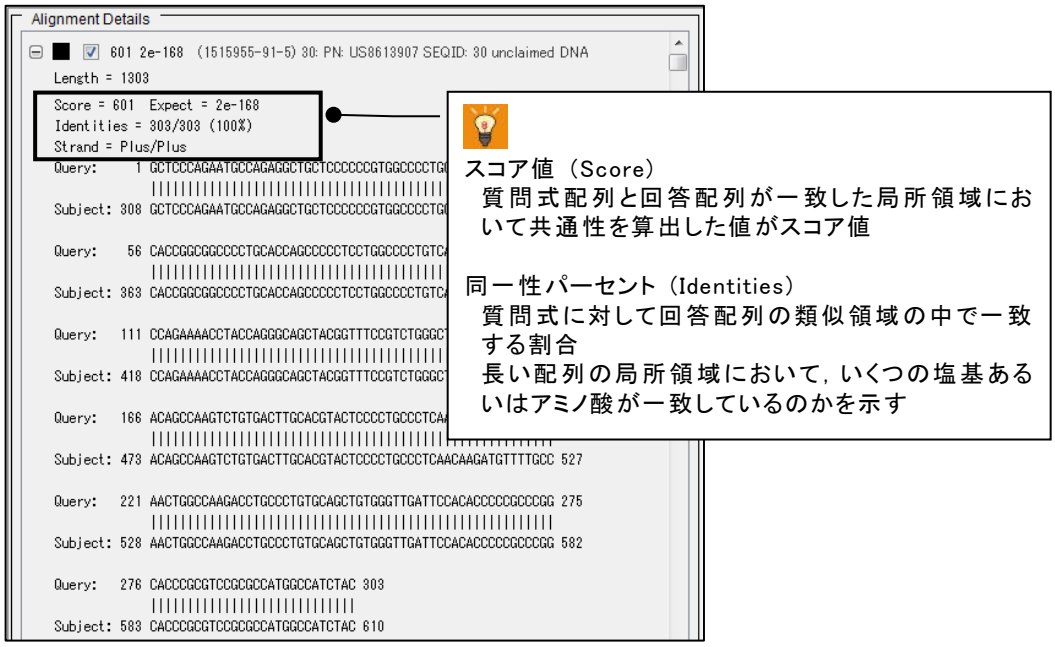


(iv) アライメントの詳細

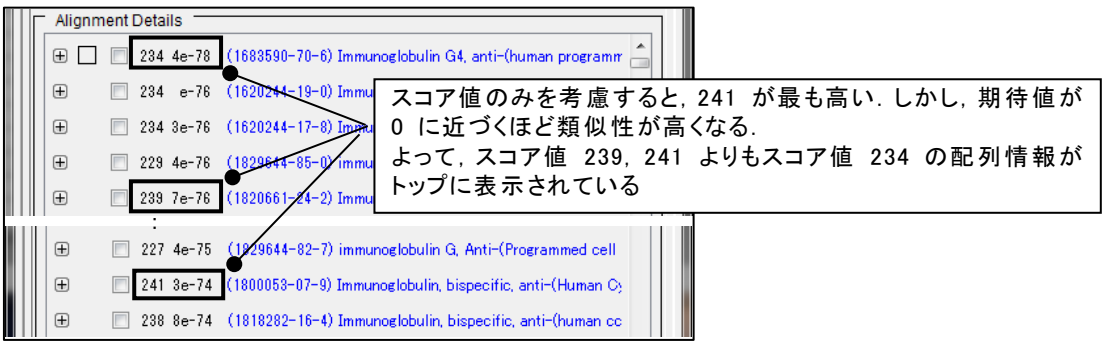
回答の全配列がスコア値の高い順に表示されている.



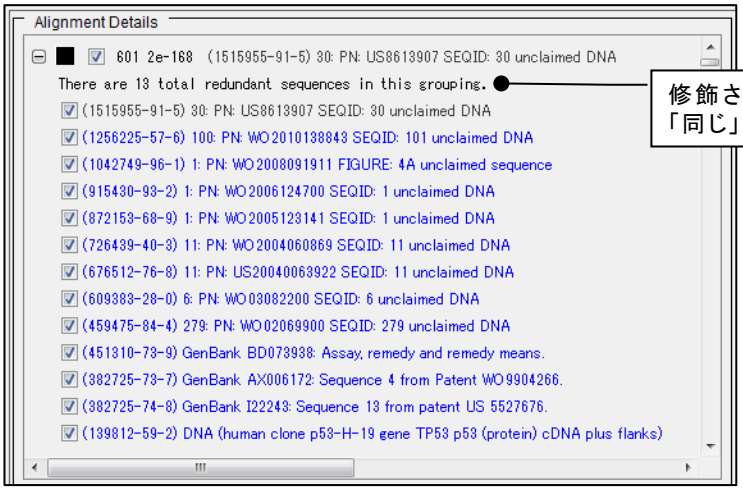
各配列の左端にある ＋ ボタンをクリックすると, 詳しいアライメント情報が表示される.



参考:スコア値が高い順に表示されない場合

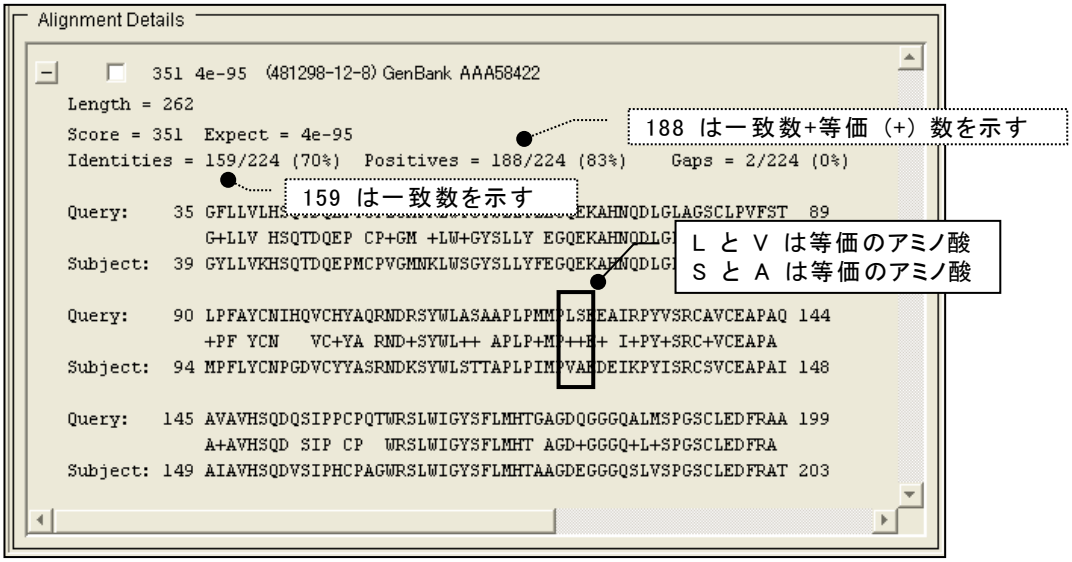


□ ボタンをクリックすると，重複する配列（同じ配列で別の CAS RN® が付与されているもの）が表示される．

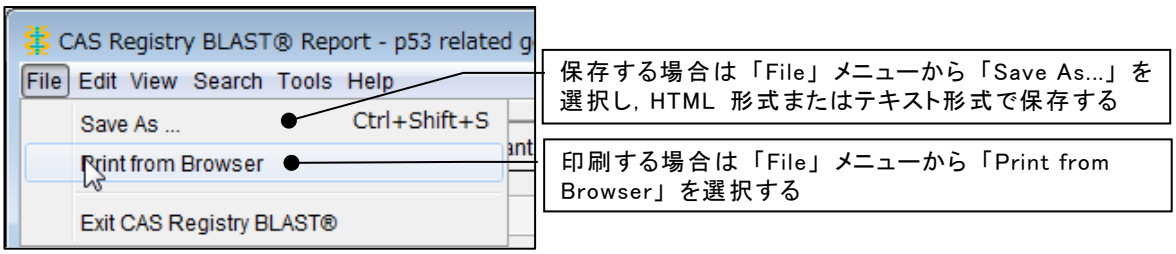


参考 ： アライメントの詳細中の ＋ 記号

アミノ酸配列を用いて BLAST ホモロジー検索する場合は，下図のように ＋ 記号が表示される場合がある．これは等価のアミノ酸であることを表している．



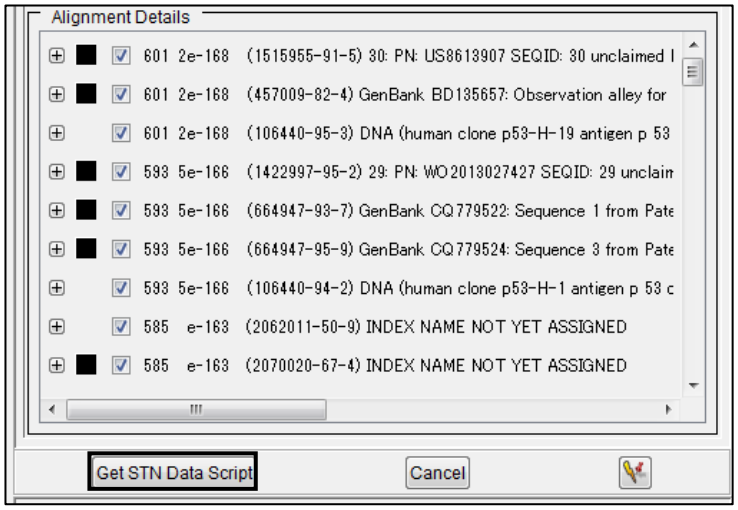
⑪ （任意）REGISTRY BLAST ホモロジー検索の回答を印刷・保存する．



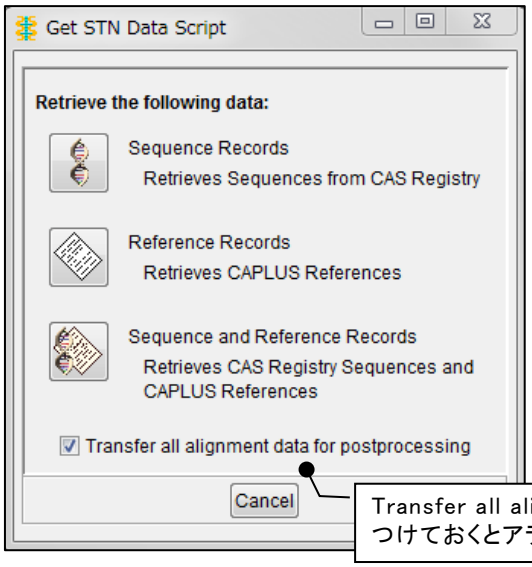
STN での検索

⑫ REGISTRY BLAST ホモロジー検索で得られた配列を STNext へ移行するために Script を作成する.

STNext へ移行したい配列（ここではスコア値の高い上位 9 配列）を選択し,「Get STN Data Script」をクリックする.



⑬ ダイアログボックスが表示される. スクリプトで自動検索したい内容を選択する.
ここでは, Sequence and Reference Records を選択する.



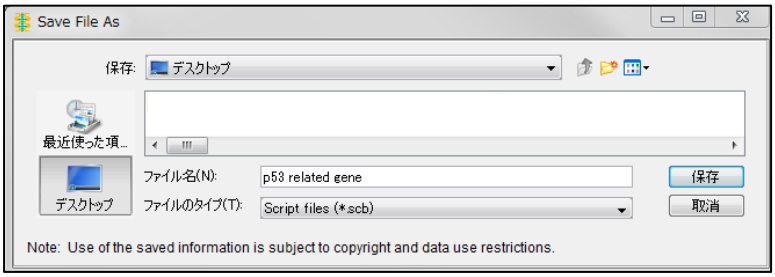
Sequence Records:
- REGISTRY ファイルで CAS RN® を検索する.

Reference Records:
- REGISTRY ファイルで検索し, CAplus ファイルにクロスオーバー検索する.

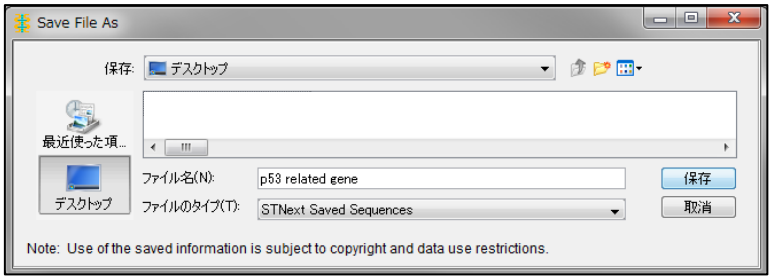
Sequence and Reference Records:
- REGISTRY ファイルで検索・回答表示して CAplus ファイルにクロスオーバー検索し, 回答を表示する.

Transfer all alignment data for postprocessing のチェックをつけておくとアライメント付き BLAST レポートが作成できる

⑭ STN へ移行するための Script (.scb) ファイルを保存する.

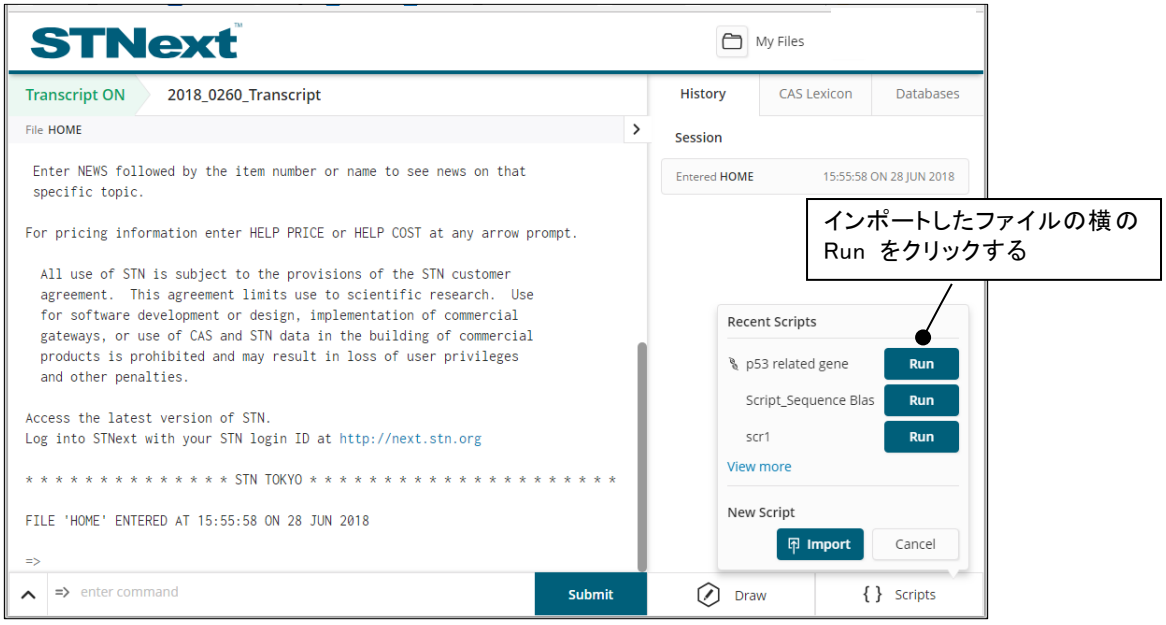


⑮ アライメント付きレポートを作成するために，アライメントデータ（STNext Saved Sequences (.xss)）を保存する．

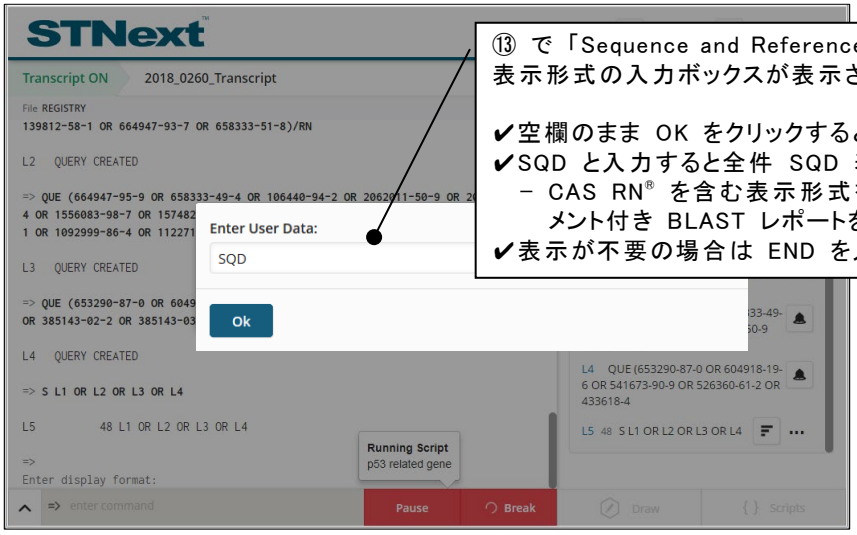


⑯ STNext で STN へログインして，Script をインポートする．

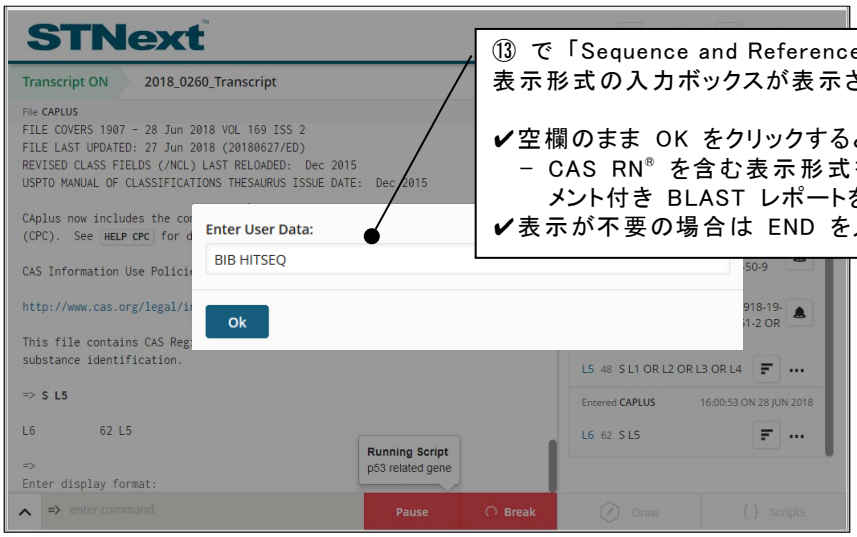
⑰ 再度，画面右下の Scripts ボタンをクリックする．インポートしたファイル名が表示されるので「Run」をクリックすると，自動検索が実行される．



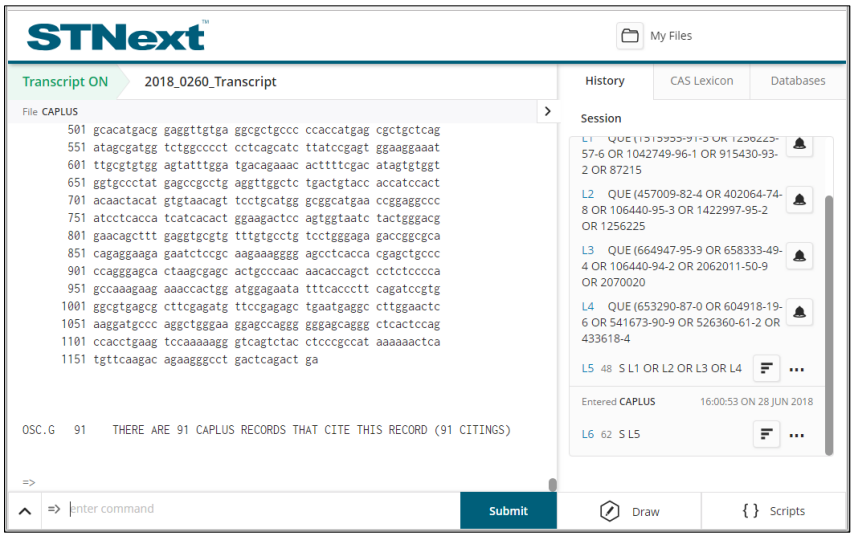
自動的に REGISTRY ファイルに入り，配列が検索される．



⑬ で「Reference Records」または「Sequence and Reference Records」を選択した場合は，自動的に CAplus ファイルに入り，クロスオーバー検索が実行される．



Script の実行が終わると，矢印プロンプトが表示される．

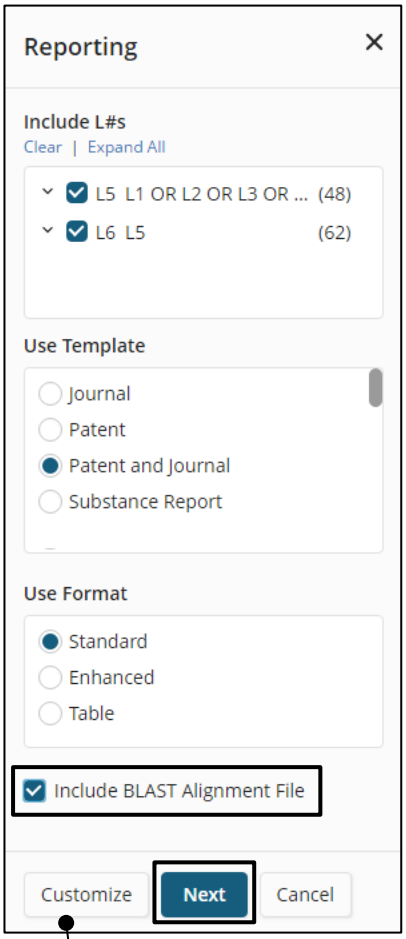


レポート作成（オプション）

■ 検索履歴に BLAST ホモロジー検索結果を組み入れたアライメント付き BLAST レポートが作成できる.

① My Files の Transcripts ページで「Report」を選択する

② Reporting 画面で，L 番号，テンプレート，フォーマット（Standard または Enhanced）を選択し「Include BLAST Alignment File」にチェックを付け，Next をクリックする.

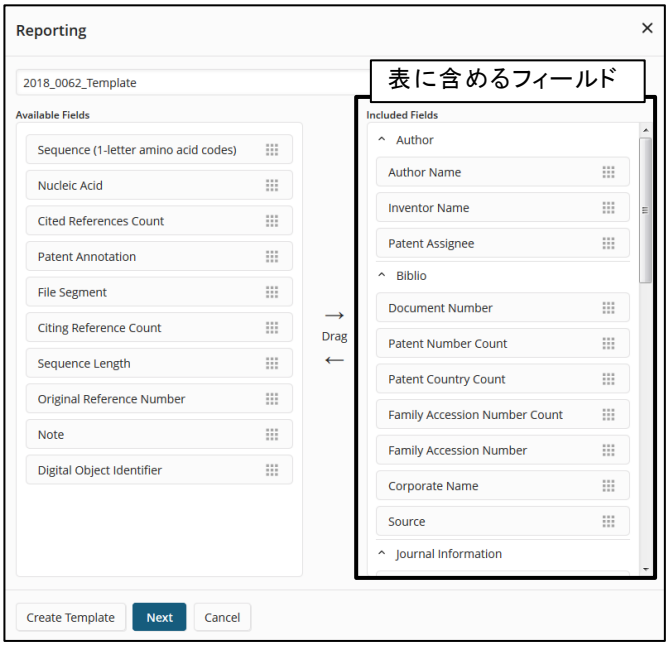


レポートに含めるフィールドを選択したい場合は，Customize をクリックする

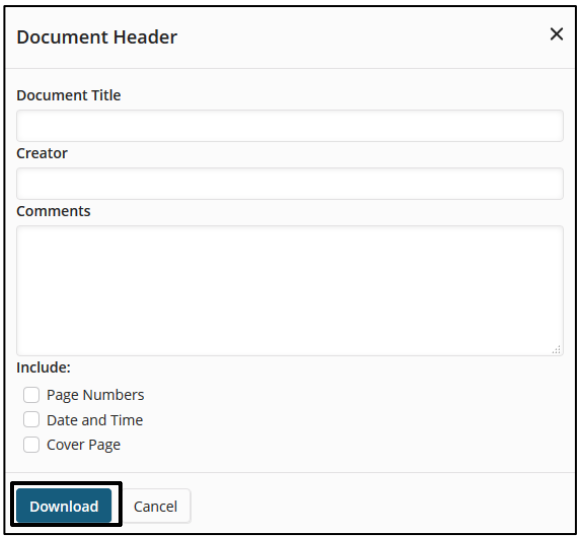
③ Import Alignment Data 画面の「Browse」より，P.44 ⑮ で保存した STN Saved Sequence ファイル（.xss）を指定し，「Next」をクリックする．

・ 前ページの ② で「Customize」を選択した場合は，レポートに含めるフィールドを選択する画面が表示される．

　– 表に含めるフィールドを右側に移す



④ Document Header 画面で，タイトル，製作者，コメント等を入力する．最後に「Download」をクリックするとレポートが完成する．



■ アライメント付き BLAST レポートの例
https://www.jaici.or.jp/stnext/news20181002_report_example.pdf を参照

REGISTRY BLAST ホモロジー検索のアラート

■ REGISTRY ファイルの BLAST ホモロジー検索はアラート登録することができる.

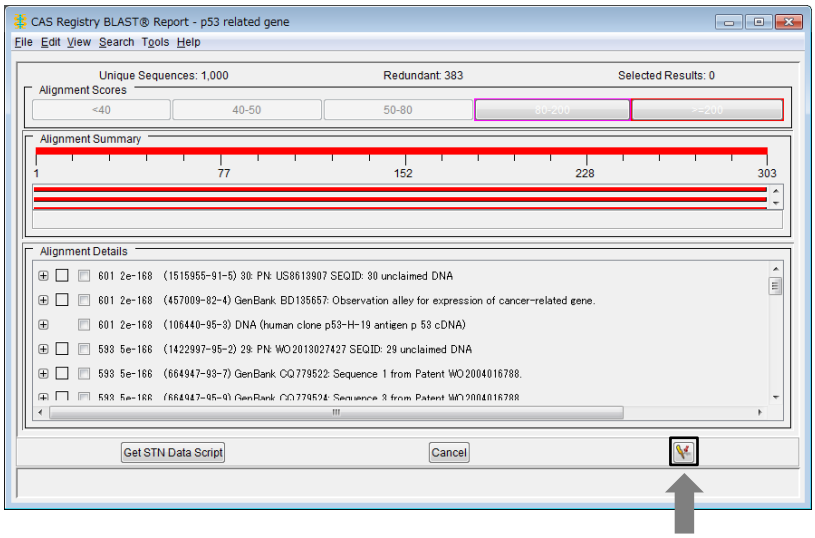
・ 実行頻度 : 毎週または隔週

・ アラートは最大 100 件まで登録できる.

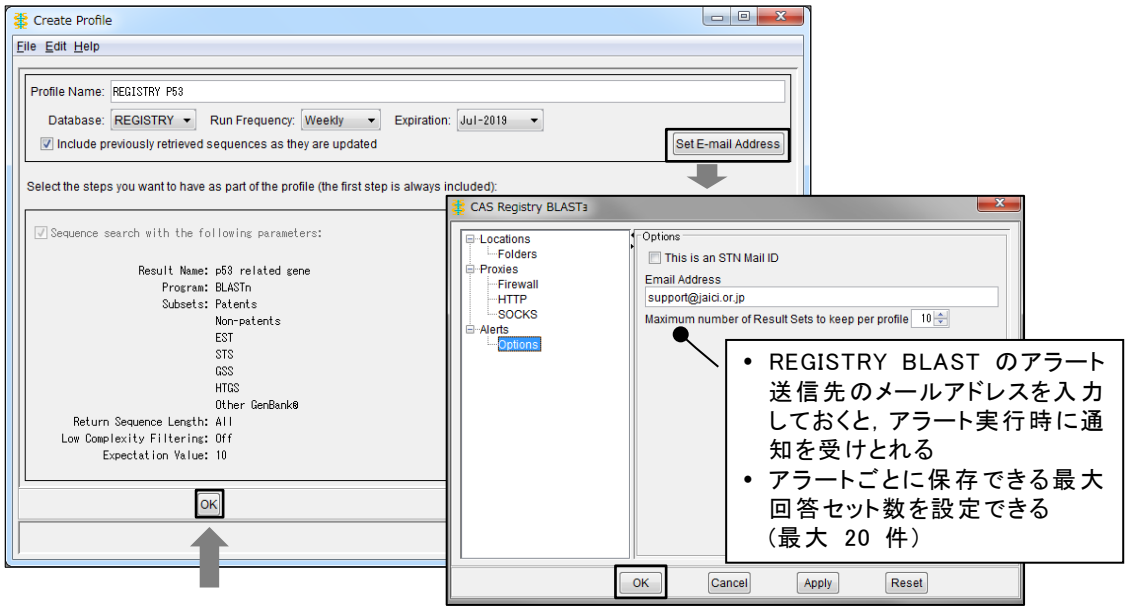
・ 回答セットはアラートごとに最大 20 件まで保存される.(デフォルトは 10 件)

■ アラート登録例

① BLAST ホモロジー検索の回答を表示し, 画面右下にある  アイコンをクリックする.



② Create Profile ダイアログボックスが表示される. アラート登録名を入力し, 実行頻度, アラート終了月を選択して, OK をクリックする.

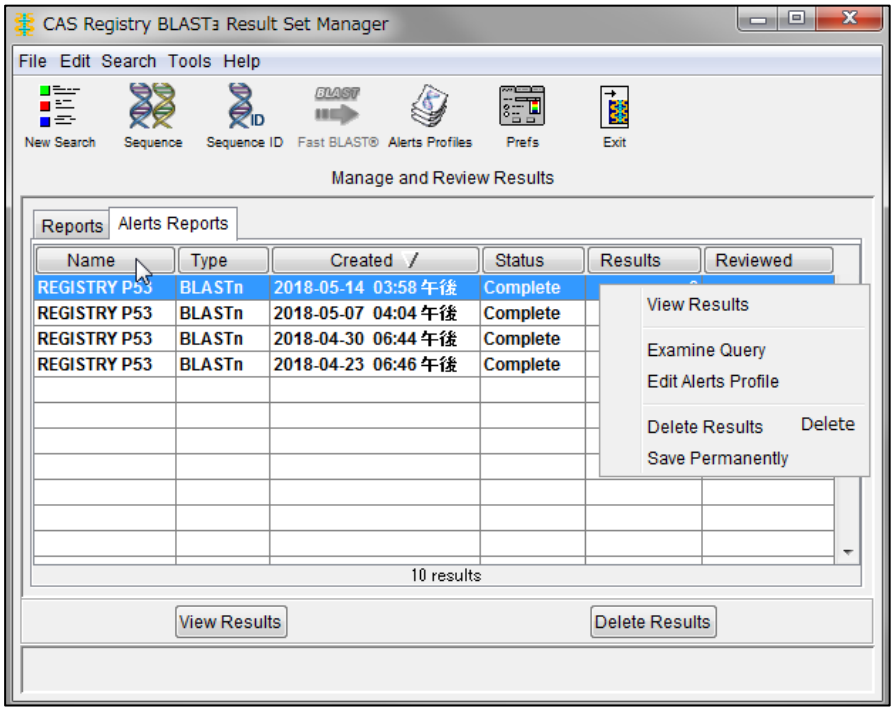


- REGISTRY BLAST のアラート送信先のメールアドレスを入力しておくと, アラート実行時に通知を受けとれる
- アラートごとに保存できる最大回答セット数を設定できる (最大 20 件)

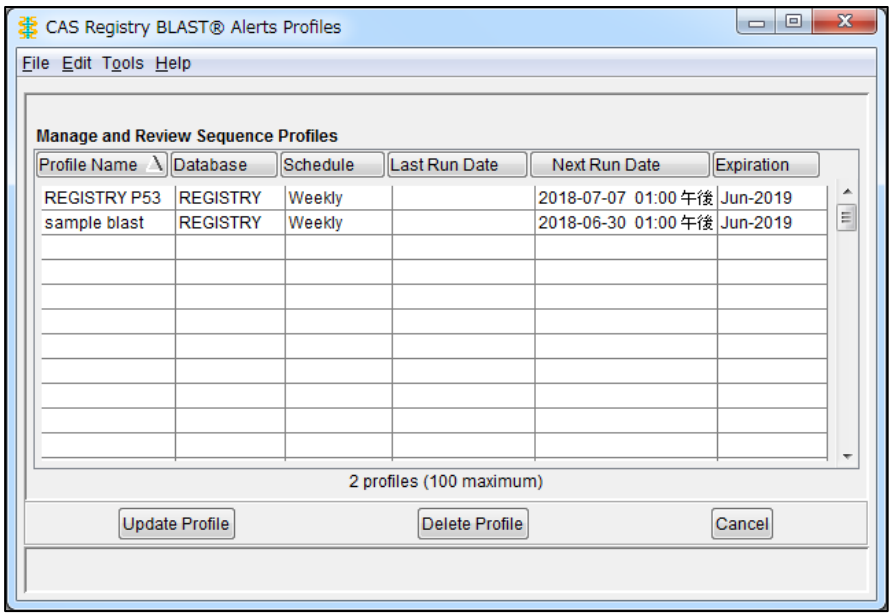
③ アラート検索の回答表示

REGISTRY BLAST を起動して,「Alerts Reports」タブを選択するとアラートの回答セットのリストが表示される. 表示したい回答セットを選択して, 「View Results」をクリックする.

アラート検索の回答セットは古い順に自動的に削除されるため, 残しておきたい回答セットは右クリックし,「Save Permanently」をクリックすると,「Reports」リストに移る.



④ Alerts Profiles をクリックすると登録したアラートの一覧を表示できる. アラートの設定を変更したい場合は Update Profile をクリックする.



APPENDIX

■ REGISTRY BLAST 検索のパラメータ設定

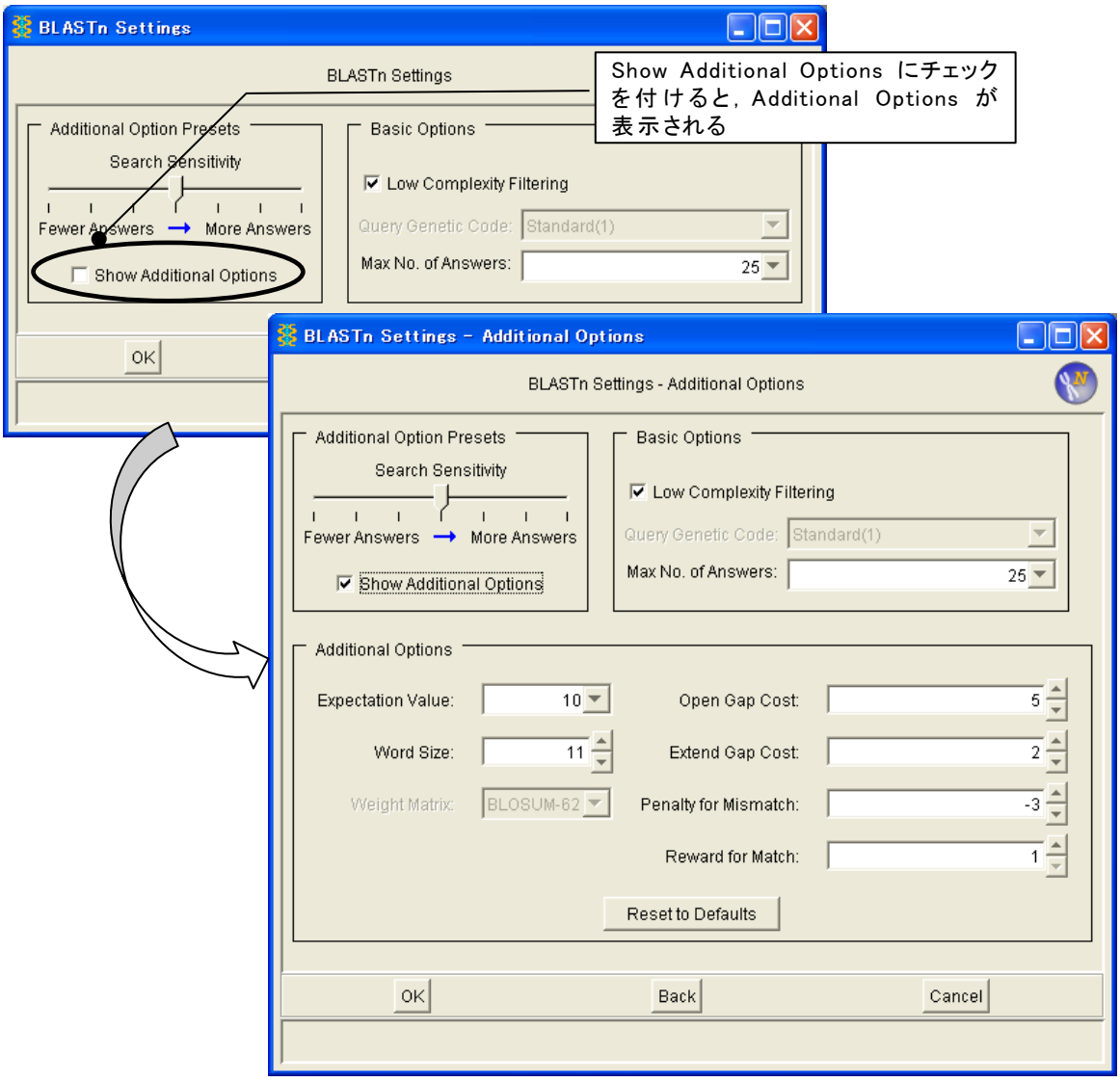
REGISTRY BLAST 検索のパラメータ設定

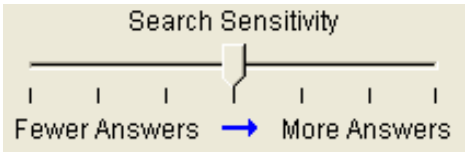
■ REGISTRY ファイルの BLAST ホモロジー検索におけるパラメータ設定

・ 検索タイプ

検索タイプ	検索機能	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	
tBLASTn	データベース中の塩基配列をアミノ酸配列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	質問式とデータベース中の塩基配列をアミノ酸配列に翻訳し，類似した配列を検索	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	
BLASTx	塩基配列の質問式をアミノ酸配列に翻訳して，データベース中の類似したアミノ酸配列を検索	塩基配列	アミノ酸配列

・ パラメータ



パラメータ	概要，指定できる数値・項目
Search Sensitivity	期待値（Expectation Value）を指定することで検索の精度を変更する. スライダを右に移動すれば回答件数が多くなり，左に移動すれば少なくなる. 高い類似性を持つ配列だけを検索する場合は左に移動する. デフォルトは 10  期待値　1e −4　0.01　1　10　50　100　1000
Low Complexity Filtering	デフォルトはチェックされており，配列質問式に対して低複雑度領域のマスクフィルタリングを行なう. マスクを行うことで，統計的に有意であっても生物学的には無意味なアラインメント（例：共通の酸性，塩基性アミノ酸のリピートやプロリン過剰な領域など）を取り除くことができる. チェックをつけておくと，具体的には配列質問式中複雑でない配列部分（低複雑度配列，反復配列）がマスクされ検索に使用されなくなる. この際 BLASTn では DUST プログラム，その他の検索タイプでは SEG プログラムを用いて低複雑度配列が決定されている. 配列質問式中マスクされるコードは以下のコードに置き換えられ検索される. ・ 塩基配列の場合： N ・ アミノ酸配列の場合： X 以下の場合はフィルタリングを行わない方がよいためチェックをはずす. ・ 特許性調査の場合 ・ 短い配列質問式で検索する場合 ・ 低複雑度領域を多く持つ配列質問式で検索する場合
Query Genetic Code	BLASTx または tBLASTx でのみ指定する. Mold Mitochondrial には Mold, Protozoan, Coelenterate Mitochondrial and Mycoplasma/Spiroplasma が含まれている. ＜項目＞ ・ Standard（デフォルト） ・ Yeast Mitochondrial ・ Invertebrate Mitochondrial ・ Echinoderm Mitochondrial ・ Bacterial ・ Ascidian Mitochondrial ・ Blespharisma Macronuclear ・ Vertebrate Mitochondrial ・ Mold Mitochondrial ・ Ciliate Macronuclear ・ Euplotid Nuclear ・ Alternative Yeast Nuclear ・ Flatworm Mitochondrial

パラメータ（続き）

<table>
<tr><th>パラメータ</th><th>概要，指定できる数値・項目</th></tr>
<tr><td>Max No. of Answers</td><td>回答件数の上限値．デフォルトは 25 件で 1 から 1,000 までの数値を指定することができる．特許調査の場合は 1,000 に設定する．</td></tr>
<tr><td>Show Additional Options</td><td>チェックを入れると，さらに細かなパラメータを設定するための画面が表示される．
デフォルトはチェックなし．</td></tr>
<tr><td>Expectation Value</td><td>期待値（Expectation Value）とはデータベース中の配列に対してマッチする際の統計的有意性の閾値のこと．デフォルトは 10 で，同程度の大きさのデータベースの検索を行った場合，10 回のマッチが偶然でも起こりうることを示している．得られた統計的有意性が与えられた閾値よりも低い場合はホモロジー検索の回答に含まれるが，高い場合は回答に含まれない．

期待値が 0 に近づくほど類似性が高くなり回答件数は少なくなる．0 より大きい数値を入力する．

短い配列質問式で検索する場合は，期待値を大きくする．</td></tr>
<tr><td>Word Size</td><td>通常はデフォルト値を変更する必要はない．ただし，短い配列を検索する場合は値を小さくする．また，非常に長い配列質問式を用いたため検索が完了しない場合は値を大きくする．

＜数値＞
・ BLASTn では 7-23（デフォルトは 11）
・ BLASTn 以外の検索タイプでは 2-3（デフォルトは 3）</td></tr>
<tr><td>Weight Matrix*1</td><td>BLASTn 以外の検索タイプでスコア値を計算する置換行列を指定する．置換行列には PAM（Point Accepted Mutation）と BLOSUM（BLOcks Substitutuin Matrix）の二つのタイプがある．

＜項目＞
・ PAM30　・ PAM70
・ BLOSUM80　・ BLOSUM62（デフォルト）　・ BLOSUM45

配列質問式の配列長による置換行列と Gap Cost の組み合わせ例

<table>
<tr><th>配列長</th><th>置換行列</th><th>Gap Cost</th></tr>
<tr><td><35</td><td>PAM-30</td><td>(9, 1)</td></tr>
<tr><td>35 - 50</td><td>PAM-70</td><td>(10, 1)</td></tr>
<tr><td>50 - 85</td><td>BLOSUM-80</td><td>(10, 1)</td></tr>
<tr><td>>85</td><td>BLOSUM-62</td><td>(11, 1)</td></tr>
</table>
</td></tr>
<tr><td>Gap Cost*1</td><td>tBLASTn, BLASTx または BLASTp で指定する Open Gap Cost と Extend Gap Cost の数値の組み合わせ．</td></tr>
</table>

＊1 については，P.125 参照

パラメータ（続き）

パラメータ	概要，指定できる数値・項目
Open Gap Cost Extend Gap Cost	BLASTn でのみ指定する． Open Gap Cost と Extend Gap Cost の数値は組み合わせが決まっておらず，自由に指定できる．0 より大きい数値を指定する．デフォルトは 5, 2
Penalty for Mismatch	BLASTn でのみ指定する．0 より小さい数値を指定する．
Reward for Match	BLASTn でのみ指定する．スコア値を決めるため Weight Matrix の代わりに指定する．デフォルトは 1 で，1 から 8 までの数値を指定することができる．

*1 指定した置換行列の種類によって，Open Gap Cost と Extend Gap Cost の数値の組み合わせが決まっている．下線部分がデフォルト

置換行列	Open Gap Cost と Extend Gap Cost の組み合わせ
BLOSUM62	(11,2) (10,2) (9,2) (8,2) (7,2) (6,2) (13,1) (12,1) (**11,1**) (10, 1) (9,1)
BLOSUM80	(25,2) (13,2) (9,2) (8,2) (7,2) (6,2) (11,1) (**10,1**) (9,1)
BLOSUM45	(13,3) (12,3) (11,3) (10,3) (16,2) (**15,2**) (14,2) (13,2) (12,2) (19,1) (18,1) (17,1) (16,1)
PAM30	(7,2) (6,2) (5,2) (10,1) (**9,1**) (8,1)
PAM70	(8,2) (7 2) (6 2) (11,1) (**10 1**) (9,1)

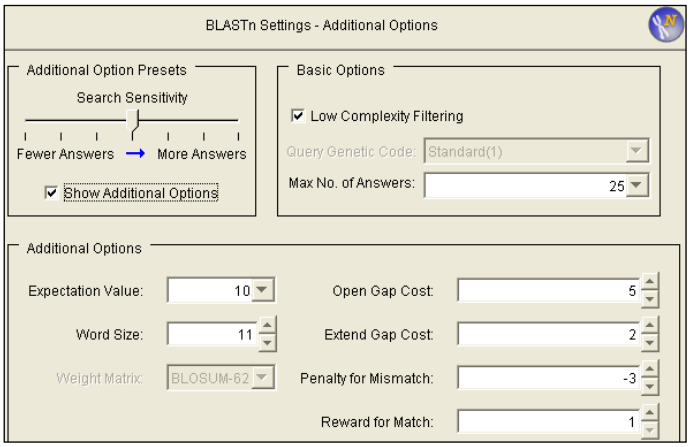
■ 配列長 30 以下の配列質問式を BLAST ホモロジー検索する場合のパラメータ設定

検索配列	設定値
核酸	・ フィルタリングを行わない ・ 期待値 : 1000 ・ word size : 7
タンパク質	・ フィルタリングを行わない ・ 期待値 : 20000 ・ word size : 2 ・ 置換行列 : PAM30

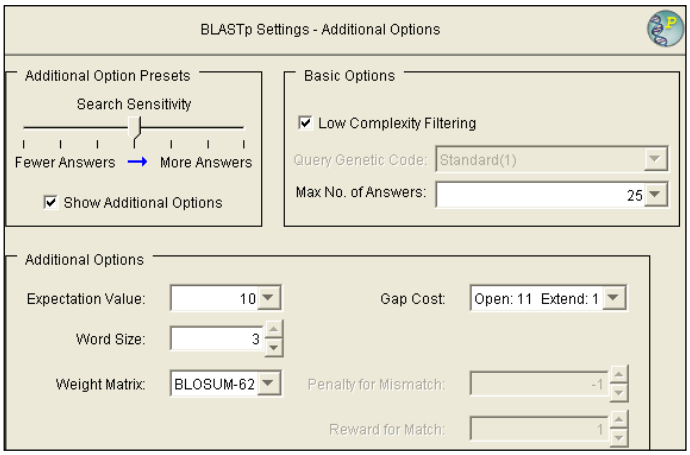
■ 最適なパラメータは検索のタイプによって異なり，デフォルト値で設定されている．特に Advanced Options 画面のパラメータは BLAST プログラムを熟知していないと，指定が難しい．

■ デフォルトの設定

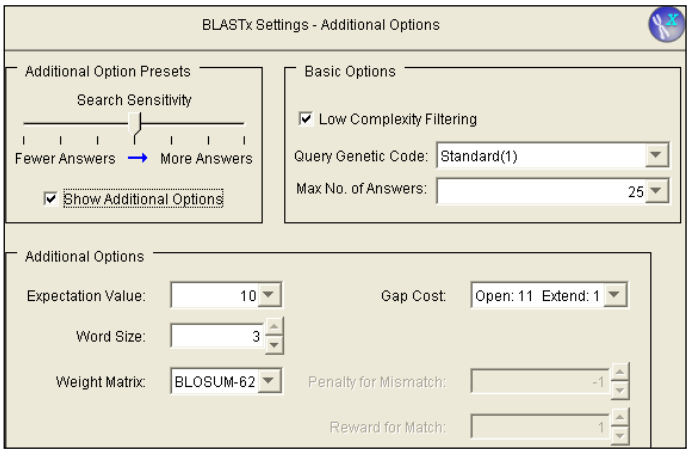
〈BLASTn〉



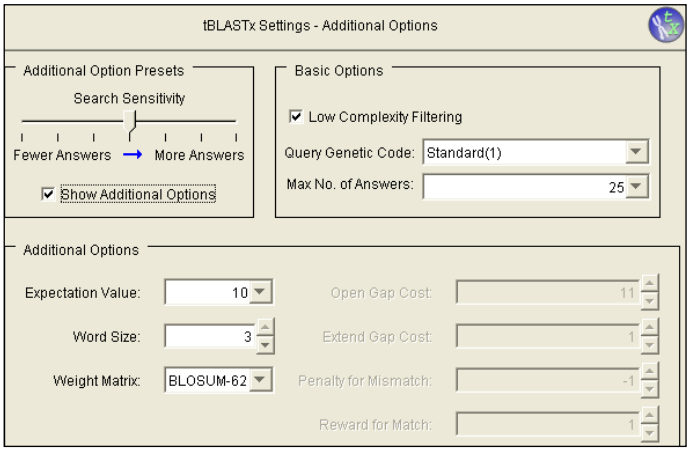
〈BLASTp〉〈tBLASTn〉



〈BLASTx〉



〈tBLASTx〉



JAICI
化学情報協会

情報事業部
〒113-0021 東京都文京区本駒込6-25-4 中居ビル
TEL: 0120-003-462 FAX: 03-5978-4090
URL: www.jaici.or.jp
E-mail: support@jaici.or.jp

2022.10