インターネットセミナー

配列検索入門

JAICI
化学情報協会

本日の内容

1. STN の配列データベース
2. 配列検索の種類
3. 検索例 【DEMO】

1. STN の配列データベース

STN には，核酸・タンパク質配列を検索できるファイルが 4 つ搭載されています

- REGISTRY ファイル
- DGENE ファイル
- PCTGEN ファイル
- USGENE ファイル

各ファイルの特長を理解しましょう

JAICI 化学情報協会

REGISTRY ファイル

CAplus/CA ファイルで索引された配列を収録！

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
収録期間	1957 年 –
更新頻度	毎日
特長	– 雑誌論文に記載された配列情報も収録 – 収録源 (文献情報) を確認したいときは CAplus ファイルへクロスオーバー

JAICI 化学情報協会

DGENE ファイル

WPI ファイルで索引された配列を収録！

製作者	Thomson Reuters
収録源	WPI ファイルのベーシック特許
収録期間	1981 年 –
更新頻度	隔週
特長	– 配列に焦点を当てた独自抄録を収録 – 収録源（特許）情報も収録

JAICI 化学情報協会

PCTGEN ファイル

PCT 出願特許に記載された配列を収録！

製作者	WIPO, FIZ Karlsruhe
収録源	PCT 出願
収録期間	2001 年 –
更新頻度	毎週
特長	– 速報性に優れている （最短タイムラグは公報発行後 1 日）

JAICI 化学情報協会

USGENE ファイル

米国特許に記載された配列を収録！

製作者	SequenceBase Corporation
収録源	INSCD, NCBI/EMBL-EBI, USPTO PSOPS, 米国特許の Sequence Listing
収録期間	1982 年 –
更新頻度	毎週
特長	– 米国登録特許由来の配列情報も収録 – 速報性に優れている （タイムラグは公報発行後 3 日以内） – 著者抄録と全クレームを収録

JAICI 化学情報協会

2. STN の配列検索

STN の配列検索は 3 種類

① 完全配列検索

質問式と完全に一致する配列を検索

② 部分配列検索

質問式を一部に含む配列を検索

③ ホモロジー検索

質問式と類似した配列を検索

JAICI 化学情報協会

① 完全配列検索

質問式　　GCCCAAGCTGGCATCCGTCA

↓　*質問式と完全に一致する配列を検索*

回答例　　GCCCAAGCTGGCATCCGTCA

JAICI 化学情報協会

① 完全配列検索の検索方法

=> 検索コマンド　コード/検索タイプ

S (REGISTRY ファイル)
RUN GETSEQ (その他のファイル)

SQEN (核酸)
SQEP (タンパク質)

＊コードの入力ルール
核酸は「5’末端 → 3’末端」, タンパク質は「N 末端 → C 末端」の順で入力

JAICI 化学情報協会

② 部分配列検索

質問式　　GCCCAAGCTGGCATCCGTCA

質問式を一部に含む配列を検索

回答例　UTCGCCCAAGCTGGCATCCGTCAGT

JAICI 化学情報協会

② 部分配列検索の検索方法

=> 検索コマンド　コード/検索タイプ

S（REGISTRY ファイル）

RUN GETSEQ（その他のファイル）

SQSN（核酸）

SQSP（タンパク質）

JAICI 化学情報協会

③ ホモロジー検索

質問式　　GCCCAAGCTGGCATCCGTCA

質問式と類似した配列を検索

回答例　　UTCGCCCAAGCTGGUTTCCGTCAGT

JAICI 化学情報協会

③ ホモロジー検索の検索方法

◆ REGISTRY ファイル

- 専用ソフトウェアを利用する

◆ DGENE, PCTGEN, USGENE ファイル

- コマンド検索 (RUN コマンド)
- Assistants 機能 (検索補助機能) を利用

JAICI 化学情報協会

3. 検索例 【DEMO】

◆ 検索例 1

DGENE ファイルで，核酸配列 ACCGGCCGGT を検索する

① 完全配列検索
② 部分配列検索

それぞれの回答の違いを確認する

JAICI 化学情報協会

JAICI 化学情報協会

◆ 検索例 1

① 完全配列検索

```
=> FILE DGENE

=> RUN GETSEQ ACCGGCCGGT/SQEN      ← 核酸の完全配列検索 (/SQEN)
L1           1 ACCGGCCGGT/SQEN

=> D TRI ALIGN                     ← TRIAL, ALIGN 表示形式で表示 (無料)

L1    ANSWER 1 OF 1  DGENE  COPYRIGHT 2012 THOMSON REUTERS on STN
AN    ACC69624  DNA       DGENE
TI    Tumor-specific promoter for producing e.g. transformant adenovirus to   ← 標題
      highly proliferate in ovarian cancer cells, applicable in gene therapy
      for treating ovarian cancer -
DESC  Human tumour-specific promoter related oligonucleotide #2.               ← 配列の説明
KW    Human; tumour-specific promoter; tumour; cytostatic; ovarian cancer;
      gene therapy; ss.                                                        ← キーワード
SQL   10                                                                       ← 配列長
SEQ                        ← 配列データ
          accggccggt       ← 質問式と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
HITS AT:  1-10             ← ヒット位置
```

② 部分配列検索

```
=> RUN GETSEQ ACCGGCCGGT/SQSN      ← 核酸の部分配列検索 (/SQSN)
L2           6891 ACCGGCCGGT/SQSN

=> S L2 NOT L1                     ← 完全配列検索で得られた回答を除く
L3           6890 L2 NOT L1

=> D TRI ALIGN                     ← TRIAL, ALIGN 表示形式で表示 (無料)

L3    ANSWER 1 OF 6890  DGENE  COPYRIGHT 2012 THOMSON REUTERS on STN
AN    BAC54469  cDNA       DGENE
TI    New isolated strain of Muscodor strobelii, where the strain is
      Agricultural Research Service Culture Collection accession number NRRL
      50288, is useful for killing a plant pathogen selected from e.g.
      Aspergillus fumigatus.
DESC  Muscodor strobelii expressed sequence tag (EST), SEQ ID 38964.
KW    EST; crop improvement; crop plant pathogen; disease resistance; expressed
      sequence tag; microorganism; palm oil; plant bacterial disease; plant
      fungal disease; ss.
SQL   6298       ← 配列長
SEQ
          accggccggt       ← 質問式と完全に一致した配列
          ==========
HITS AT:  3443-3452        ← 配列長 6298 の核酸配列のうち, 3443-3452 の部分が質問式と一致
```

=> D ALL　　← ALL 表示形式で表示（特許情報や抄録，全配列コード等を確認できる）

```
L3    ANSWER 1 OF 6890  DGENE  COPYRIGHT 2012 THOMSON REUTERS on STN
AN    BAC54469  cDNA       DGENE
TI    New isolated strain of Muscodor strobelii, where the strain is
      Agricultural Research Service Culture Collection accession number NRRL
      50288, is useful for killing a plant pathogen selected from e.g.
      Aspergillus fumigatus.
IN    Green W A; Herrgard M J; Kerovuo J S; Lomelin D; Mathur E J; Richarson T
      H; Schwartz A S; Strobel G A
PA    (SYNT-N)    SYNTHETIC GENOMICS INC.
PI    WO 2010115156   A2 20101007  72      ← WPI のベーシック特許のみ収録
AI    WO 2010-US29859      20100402
PRAI  US 2009-166681P      20090403
      US 2009-230648P      20090731
PSL   Claim 22; SEQ ID NO 38964           ← 明細書中の記載位置や配列番号
DED   06 DEC 2012  (first entry)
DT    Patent
LA    English
OS    2010-M76294 [68]                    ← WPI ファイルのレコード番号
DESC  Muscodor strobelii expressed sequence tag (EST), SEQ ID 38964.
KW    EST; crop improvement; crop plant pathogen; disease resistance; expressed
      sequence tag; microorganism; palm oil; plant bacterial disease; plant
      fungal disease; ss.
ORGN  Muscodor sp. WG-2009a.
AB    The present invention relates to a novel endophytic fungal strain,
      Muscodor strobelii NRRL 50288, which produces volatile organic compounds
      (I) e.g. isobutyric acid, which has biological activity against plant
      pathogens selected from Ganoderma boninense, Aspergillus fumigatus,
      Botrytis cinerea, Cerpospora betae, Curvularia sp., Geotrichum candidum,
      Mycosphaerella fijiensis, Phytophthora palmivora, Phytophthora ramorum,
      Pythium ultimum, Rhizoctonia solani, Rhizopus sp., Schizophyllum sp.,
      Sclerotinia sclerotiorum, Verticillium dahliae and Xanthomonas
      axonopodis. The strain is also useful for treating, inhibiting or
      preventing the development of a plant pathogenic disease; and for
      killing, inhibiting or preventing the development of an organism selected
      from fungus, bacterium, microorganism, nematode and insect. The present
      sequence is an Expressed Sequence Tag (EST) from the fungal strain of the
      invention.
NA    1590 A; 1516 C; 1511 G; 1681 T; 0 U; 0 Other
SQL   6298                                ← 配列長
SEQ
        1 aacataaacc tgaatgtcag ctatatagat catttaacca aaggcatgtc
       51 tcatcctacg cgctgcagtc attacgatgt tcaagcatca cagatgattt
      101 cgtactataa caccactgtc gatatctgct gcgtctcaat tgacgctcct
      151 gctcagctta caagccttca actaacattg aaagagcaaa gcgcaccatt
      201 catatataca tgttcccaaa tctccagcag tgcaaagtgg ctgcagatgt
      251 tgggcaagga ggctgtcgtg gcttgtggca cgtggattga tttgtacctc
      301 gaaatgtatt gtgtgagacg taaaatgtaa gtgaagtaag tccctttgcc
      351 tacgtcgggg tttcacccaa gatctcatca tcaaaacgag gccgcgcggt
            :
     3351 cgttcggatt ttcttgatat attatagtcc taccttagct atcaaagaat
     3401 gtaaacttcc aatcatttat tgcgtgttca tgatcggtaa gtaccggccg
                                                    ========
     3451 gtagcaatac gtcacacgag cagaatgaca aggcagcctc gttcccatac
          ==
              :
     6051 gggaagcaat gcgaacaggg aaagtctgtt tctatgagga taccaaaggg
     6101 acggtacttt tggaaagtgg aaaagatgtt agagttcaca cgagtcagtg
     6151 ggcatttctt ggatggaatt ggcgctccgg agtttgacaa ggaactttca
     6201 gtggattaat agataccgct tttgaagtta gaggattaca tgttatttgg
     6251 ttactgtctg gctatataca aatttgaata atcgcaatgg tatatctc
HITS AT:  3443-3452
```

特許情報

抄録
（配列独自）

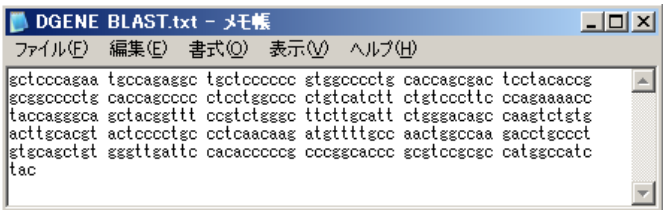
配列情報

3. 検索例 【DEMO】

◆ 検索例 2

DGENE ファイル，REGISTRY ファイルで
BLAST ホモロジー検索を行う

＊長い質問式の場合は，事前にテキストファイルを作成しておく



JAICI 化学情報協会

ホモロジー検索のプログラム

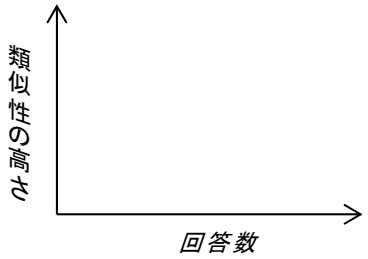
	BLAST	GETSIM
処理速度	速い	遅い（バッチ検索が望ましい）
類似性の高い配列	○	○
類似性の低い配列	△	○
比較方法	短い配列を比較 ギャップはあまり考慮されない	配列全体を比較 ギャップも考慮される
感受性	デフォールト設定では低い	高い
ファイル	REGISTY DGENE, PCTGEN, USGENE	DGENE, PCTGEN, USGENE

JAICI 化学情報協会

◆ 検索例 2

① DGENE ファイル

```
=> FILE DGENE

=> UPL R BLAST              ← 質問式をアップロード
L4   GENERATED

=> D LQUE                   ← 質問式を確認

L1    ANSWER 1  DGENE  COPYRIGHT 2012 THOMSON REUTERS on STN
LQUE  gctcccagaatgccagaggctgctccccccgtggcccctgcaccagcgactcctacaccggcggcccctgcac
      cagccccctcctggcccctgtcatcttctgtcccttcccagaaaacctaccagggcagctacggtttccgtct
      gggcttcttgcattctgggacagccaagtctgtgacttgcacgtactcccctgccctcaacaagatgttttgc
      caactggccaagacctgccctgtgcagctgtgggttgattccacacccccgcccggcacccgcgtccgcgcca
      tggccatctac

=> RUN BLAST L4/SQN -F F  ← BLAST ホモロジー検索
                            * -F F は低複雑度領域フィルタを外すオプション設定
                            * デフォールトで，入力した配列コードとその相補鎖が検索される

1528 ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0

      QUERY SELF SCORE VALUE IS   601     ← 回答が質問式と完全に一致した場合のスコア値
      BEST ANSWER SCORE VALUE IS  601     ← 今回得られた回答の最高スコア

Similarity
Score
      601 |
          ||||||||||
          ||||||||||
          ||||||||||||
          ||||||||||||
          ||||||||||||
          ||||||||||||
          ||||||||||||
          ||||||||||||
          ||||||||||||||||
      301 ||||||||||||||||
          ||||||||||||||||||
          ||||||||||||||||||||
          ||||||||||||||||||||
          ||||||||||||||||||||||
          ||||||||||||||||||||||
          ||||||||||||||||||||||
          ||||||||||||||||||||||
          ||||||||||||||||||||||||||
          ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Answer Count     310       620       930       1240      1550

ENTER EITHER THE NUMBER OF ANSWERS YOU WISH TO KEEP
 OR ENTER MINIMUM PERCENT OF SELF SCORE FOLLOWED BY %
 (BEST ANSWER PERCENTAGE OF SELF SCORE IS 100%)
ENTER (ALL) OR ? :ALL       ← ALL と入力すると，回答セット（L5）には全件の回答が含まれる
L5    RUN STATEMENT CREATED
L5        1528 GCTCCCAGAATGCCAGAGGCTGCTCCCCCCGTGGCCCCTGCACCAGCGAC
               TCCTACACCGGCGGCCCCTGCACCAGCCCCCTCCTGGCCCCTGTCATCTT
               CTGTCCCTTCCCAGAAAACCTACCAGGGCAGCTACGGTTTCCGTCTGGGC
               TTCTTGCATTCTGGGACAGCCAAGTCTGTGACTTGCACGTACTCCCCTGC
               CCTCAACAAGATGTTTTGCCAACTGGCCAAGACCTGCCCTGTGCAGCTGT
               GGGTTGATTCCACACCCCCGCCCGGCACCCGCGTCCGCGCCATGGCCATC
               TAC/SQN.-F F
```



=> D TRI ALIGN　　← 1 番目の回答を表示

```
L5    ANSWER 1 OF 1528  DGENE  COPYRIGHT 2012 THOMSON REUTERS on STN
AN    BAE35297  DNA       DGENE
TI    New isolated clonal LIVP strain other than clonal strain whose genome
      comprises sequence of nucleotides, useful for e.g. treating proliferative
      disorder in subject, and detecting tumor or metastasis in subject.
DESC  LIVP clonal strain related heterologous gene, SEQ 268.
KW    breast tumor; cancer; colon tumor; cytostatic; diagnostic test; ds;
      genetically engineered microorganism; immune stimulation; lung tumor;
      metastasis; microorganism identification; neoplasm; ovary tumor; p53
      gene; p53 tumor suppressor protein; pancreas tumor; prostate tumor;
      therapeutic; vaccine live-attenuated; vaccine, anticancer; vaccine,
      antiviral; viral infection; viral replication; virucide.
SQL   2586
BLASTALIGN
   Query  = 303 letters                          ← 質問式は 303 コード
   Length = 2586                                 ← 配列長
   Score  =  585 bits (295), Expect = e-171      ← スコア値*
   Identities = 301/303 (99%)                    ← 同一性パーセント*
   Strand = Plus / Plus
```

＊ スコア値（Score）とは，質問式と回答が一致した局所領域において共通性を算出した値

＊ 同一性パーセント（Identities）とは，回答の質問式類似領域中で，質問式と一致する割合

↓質問式と一致しないコードが存在

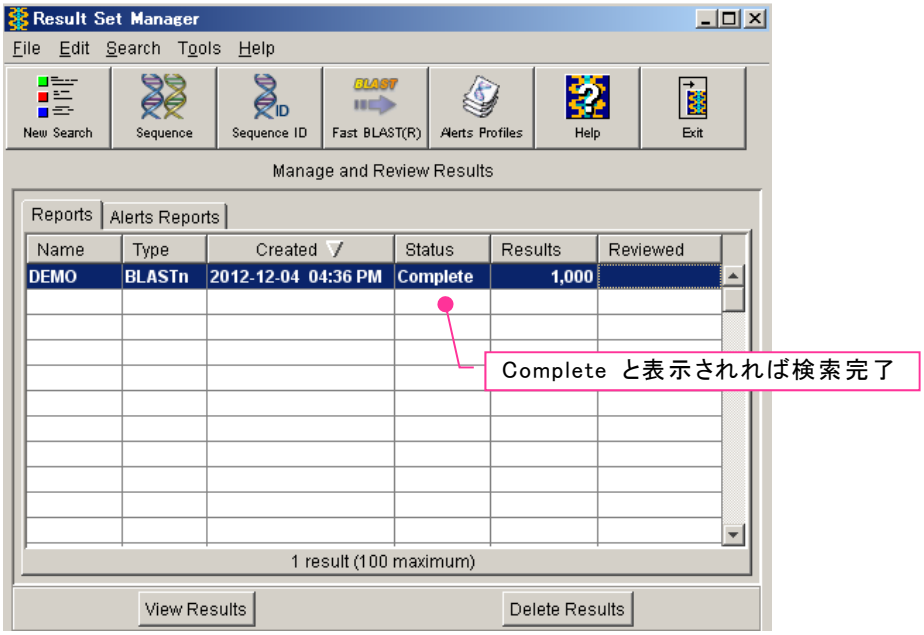
```
  Query: 1   gctcccagaatgccagaggctgctccccccgtggcccctgcaccagcgactcctacaccg
             ||||||||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||
  Sbjct: 384 gctcccagaatgccagaggctgctccccccgtggcccctgcaccagcagctcctacaccg

  Query: 61  gcggcccctgcaccagcccccctcctggcccctgtcatcttctgtcccttcccagaaaacc
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  Sbjct: 444 gcggcccctgcaccagcccccctcctggcccctgtcatcttctgtcccttcccagaaaacc

  Query: 121 taccagggcagctacggtttccgtctgggcttcttgcattctgggacagccaagtctgtg
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  Sbjct: 504 taccagggcagctacggtttccgtctgggcttcttgcattctgggacagccaagtctgtg

  Query: 181 acttgcacgtactcccctgccctcaacaagatgttttgccaactggccaagacctgccct
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  Sbjct: 564 acttgcacgtactcccctgccctcaacaagatgttttgccaactggccaagacctgccct

  Query: 241 gtgcagctgtgggttgattccacacccccgcccggcacccgcgtccgcgccatggccatc
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  Sbjct: 624 gtgcagctgtgggttgattccacacccccgcccggcacccgcgtccgcgccatggccatc

  Query: 301 tac 303
             |||
  Sbjct: 684 tac 686
```

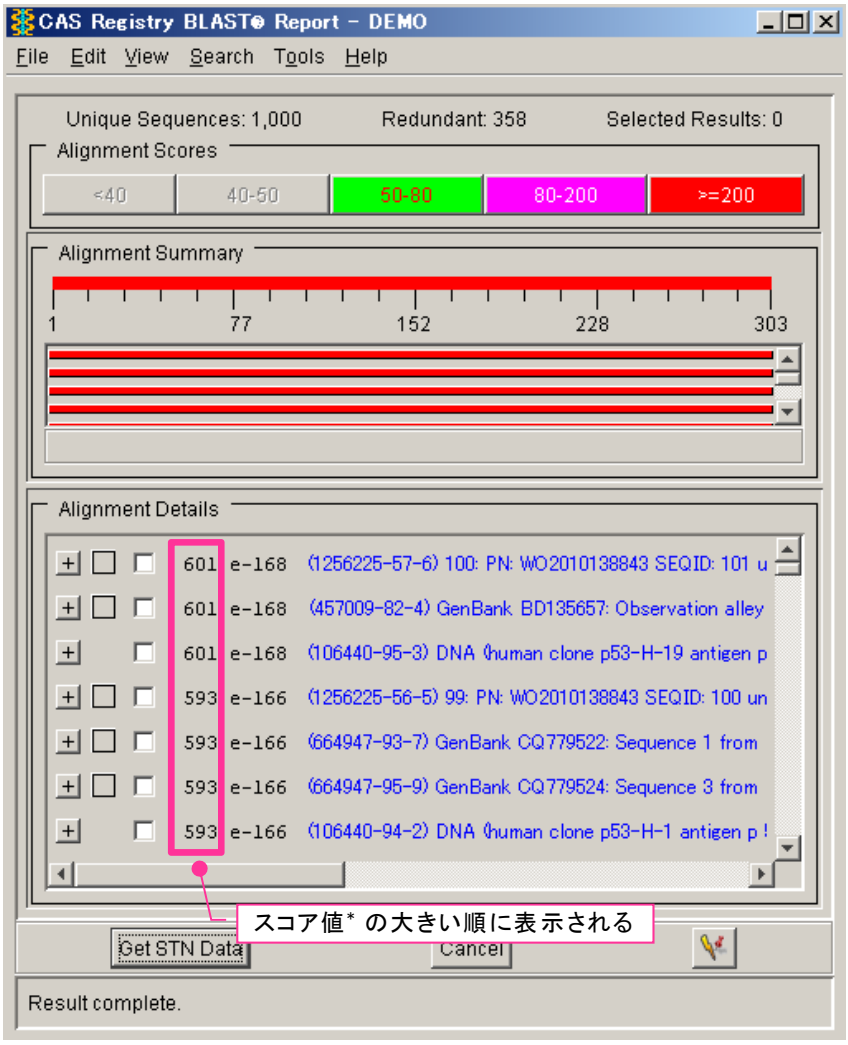
```
=> SORT L5 1- SCORE D IDENT D      ← スコア値（SCORE）と同一性（IDENT）の高い順に並び替え
L6          1528 SORT L5 1- SCORE D IDENT D

=> D 1 1528 BIB SCORE ALIGN        ← スコア値の最も高い回答と最も低い回答を表示

L6    ANSWER 1 OF 1528  DGENE  COPYRIGHT 2012 THOMSON REUTERS on STN
AN    AYM36275  cDNA       DGENE
TI    Evaluating a patient with acute lymphoblastic leukemia (ALL) that is
      characterized by the presence of Philadelphia chromosome comprises
      generating an expression profile of ALL biomarkers from a test biological
      sample.
IN    Zuo Z; Luthra R
PA    (TEXA)      UNIV TEXAS SYSTEM.
PI    WO 2010138843   A2 20101202               43
AI    WO 2010-US36623      20100528
PRAI  US 2009-182228P      20090529
PSL   Disclosure; SEQ ID NO 101
DT    Patent
LA    English
OS    2010-P75161 [82]
CR    DDBJ: M14695
      PC-NCBI: gi339815
      PC_ENCPRO-NCBI: gi339816
DESC  Acute lymphoblastic leukemia prognosis determining DNA marker, SEQ 101.
SCORE 601      100% of query self score 601
BLASTALIGN
   Query  = 303 letters                          ← 質問式は 303 コード
   Length = 1303                                ← 配列長
   Score  =  601 bits (303), Expect = e-176     ← スコア値
   Identities = 303/303 (100%)                  ← 同一性パーセント
   Strand = Plus / Plus

  Query: 1   gctcccagaatgccagaggctgctccccccgtggcccctgcaccagcgactcctacaccg
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  Sbjct: 308 gctcccagaatgccagaggctgctccccccgtggcccctgcaccagcgactcctacaccg

  Query: 61  gcggcccctgcaccagccccctcctggcccctgtcatcttctgtcccttcccagaaaacc
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  Sbjct: 368 gcggcccctgcaccagccccctcctggcccctgtcatcttctgtcccttcccagaaaacc

  Query: 121 taccagggcagctacggtttccgtctgggcttcttgcattctgggacagccaagtctgtg
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  Sbjct: 428 taccagggcagctacggtttccgtctgggcttcttgcattctgggacagccaagtctgtg

  Query: 181 acttgcacgtactcccctgccctcaacaagatgttttgccaactggccaagacctgccct
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  Sbjct: 488 acttgcacgtactcccctgccctcaacaagatgttttgccaactggccaagacctgccct

  Query: 241 gtgcagctgtgggttgattccacacccccgcccggcacccgcgtccgcgccatggccatc
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  Sbjct: 548 gtgcagctgtgggttgattccacacccccgcccggcacccgcgtccgcgccatggccatc

  Query: 301 tac 303
             |||
  Sbjct: 608 tac 610
```

特許情報

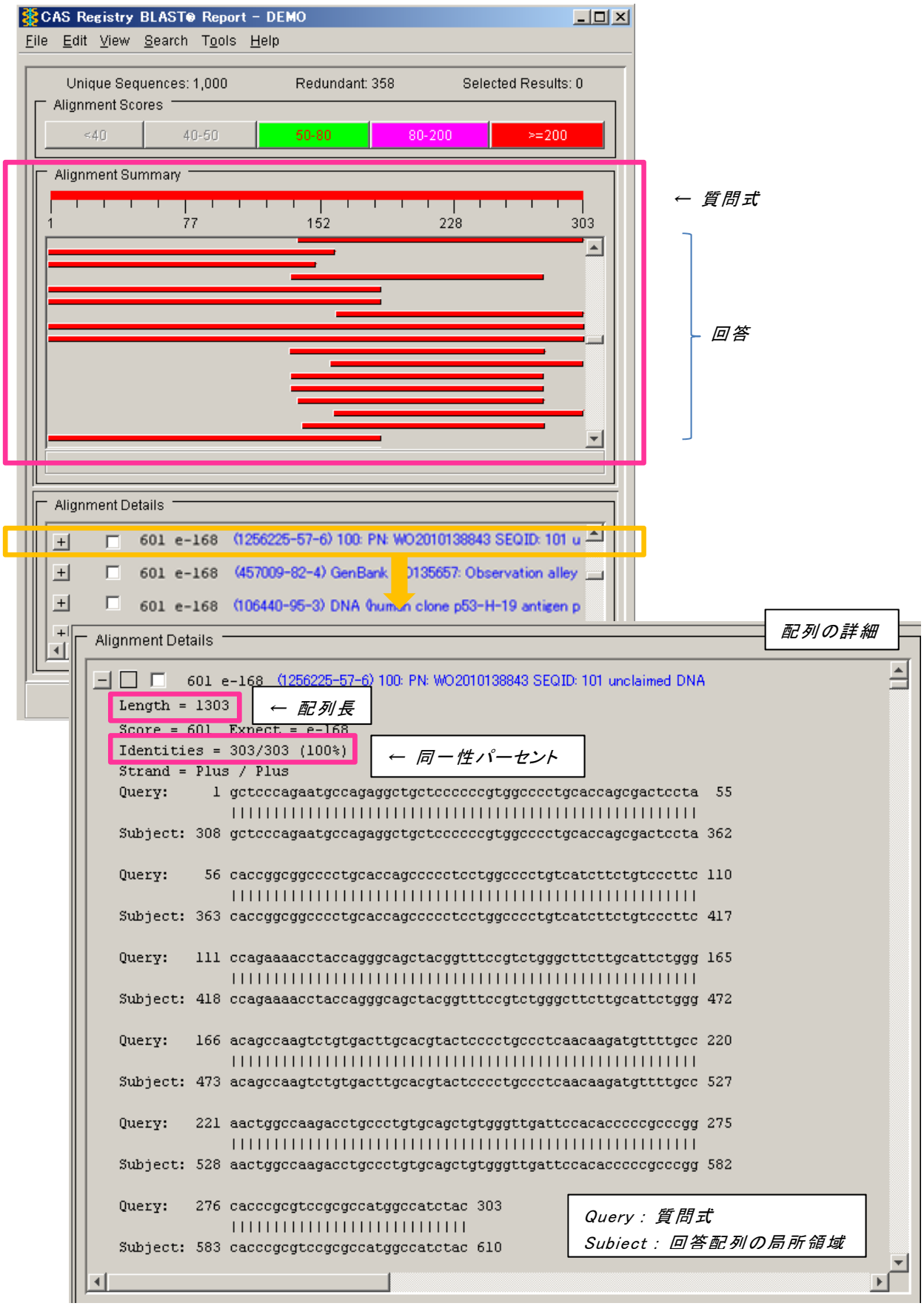
```
L6    ANSWER 1528 OF 1528  DGENE  COPYRIGHT 2012 THOMSON REUTERS on STN
AN    ADZ15573  DNA       DGENE
TI    Constructing a mutant p53 gene library by performing first PCR using
      oligonucleotide specifying mutation induction as primer, performing
      second PCR using product of first PCR, as megaprimer and PCR cloning a
      PCR product in gap repair vector.
PA    (TOHO-N)    TOHOKU TECHNOARCH KK.
PI    JP 2003265187   A  20030924               664
AI    JP 2002-76990        20020319
PRAI  JP 2002-76990        20020319
PSL   Claim 8; SEQ ID NO 385
DT    Patent
LA    Japanese
OS    2004-183645 [18]
DESC  Mutagenic PCR primer used to amplify human p53 cDNA - SEQ ID 383.
SCORE 40      6% of query self score 601
BLASTALIGN
   Query  = 303 letters                              ← 質問式は 303 コード
   Length = 26                                       ← 配列長
   Score  = 40.1 bits (20), Expect = 5e-09           ← スコア値
   Identities = 23/24 (95%)                          ← 同一性パーセント
   Strand = Plus / Minus

  Query: 1  gctcccagaatgccagaggctgct 24
            |||||||||| ||||||||||||
  Sbjct: 24 gctcccagaaggccagaggctgct 1
```

② REGISGRY ファイル（検索タイプ BLASTn）



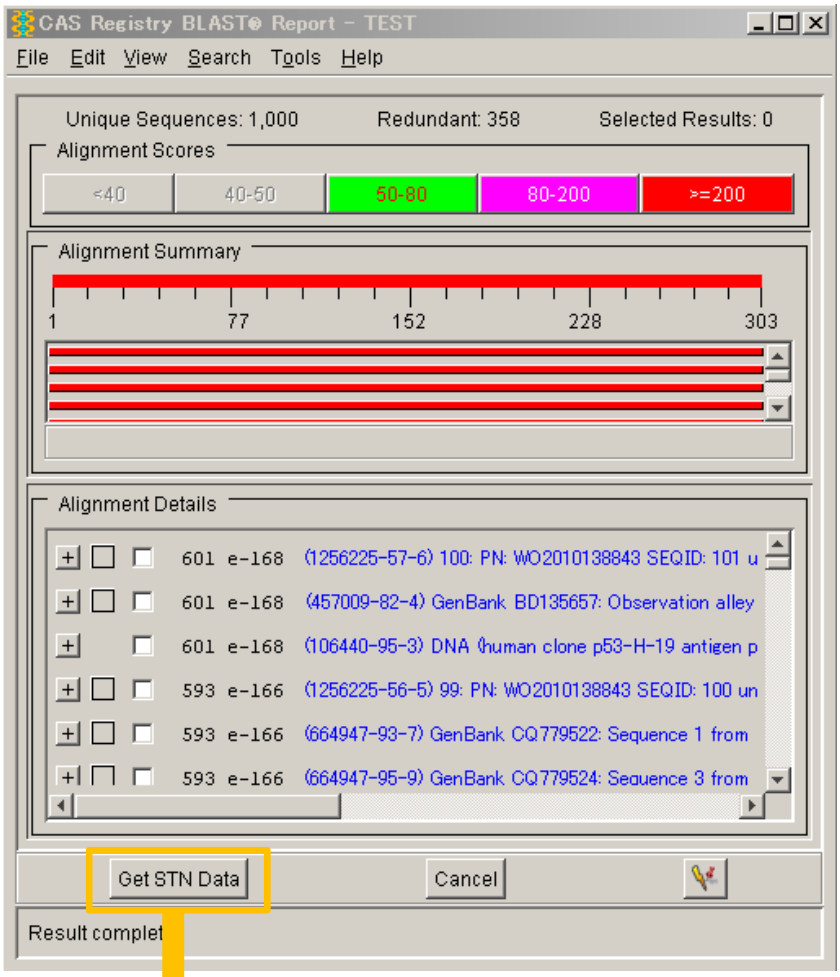


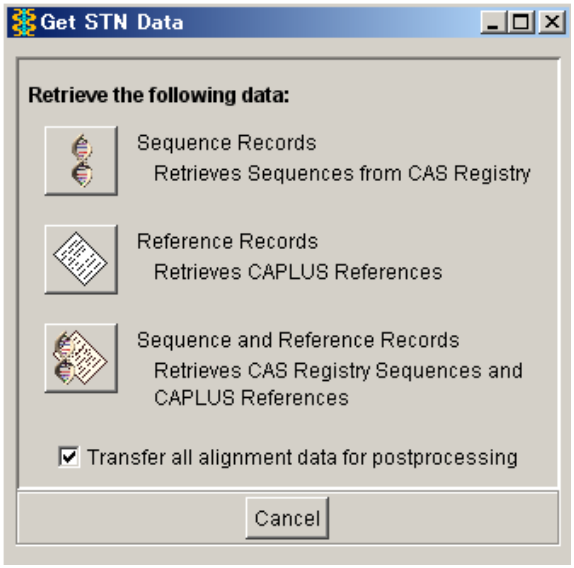
＊ スコア値（Score）とは，質問式と回答が一致した局所領域において共通性を算出した値



＊ 同一性パーセント（Identities）とは，回答の質問式類似領域中で，質問式と一致する割合

【参考】STN へデータを取り込むには，下記の方法で CAS 登録番号を抽出する





← *REGISTRY ファイルで CAS 登録番号を検索*

← *REGISTRY ファイルで CAS 登録番号を検索し，CAplus ファイルへクロスオーバー*

← *REGISTRY ファイルで CAS 登録番号を検索・回答表示し，CAplus ファイルへクロスオーバー*

← *チェックを入れると，ダウンロードデータに BLAST ホモロジー検索の結果が組み込まれる（STN Express のみ）*