

新規データベース USGENE ファイル

2007.5



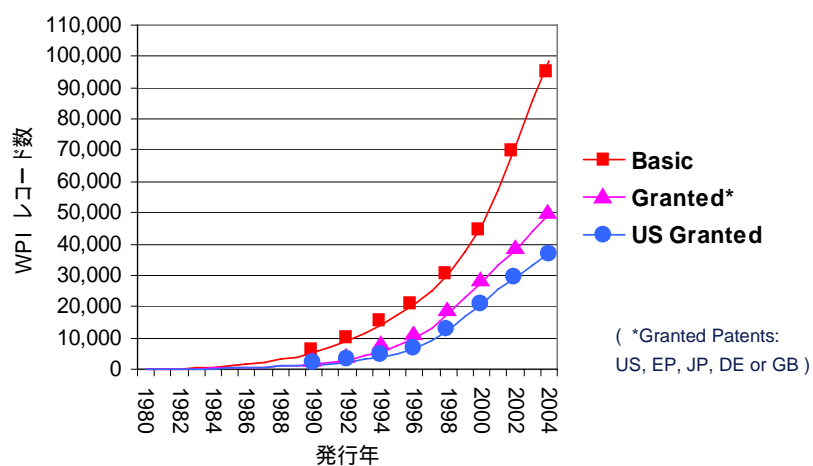
STN[®]

USGENE ファイル

Coming Soon!

STN ユーザーミーティング 2007

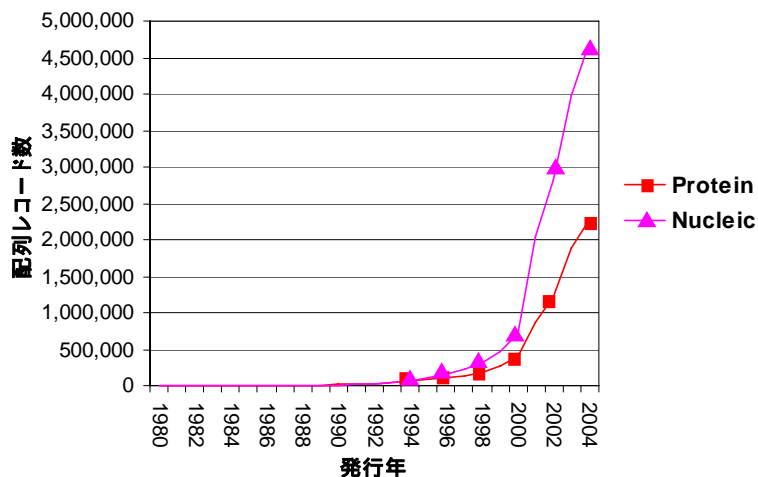
配列特許のデータベース収録状況 1



(*Granted Patents:
US, EP, JP, DE or GB)

(DGENE ファイルと WPI ファイルでの解析)

配列特許のデータベース収録状況 2




(DGENE ファイルでの解析)

STN ユーザーミーティング 2007-3

FIZ Karlsruhe

STN

配列情報を収録しているデータベース


- *USGENESM* 
 - *The USPTO Genetic Sequence Database*
- REGISTRY
 - Chemical Abstracts Service (CAS) Registry File
- DGENE
 - Thomson Scientific GENESEQTM
- PCTGEN
 - WIPO/PCT Patent Application Biosequences
- GenBank
 - National Center for Biotechnology Information (NCBI) Genetic Sequence Data Bank

FIZ Karlsruhe

STN ユーザーミーティング 2007-4

STN

配列検索可能なデータベース

- **USGENESM** 
– The USPTO Genetic Sequence Database
- **REGISTRY**
– Chemical Abstracts Service (CAS) Registry File
- **DGENE**
– Thomson Scientific GENESEQTM
- **PCTGEN**
– WIPO/PCT Patent Application Biosequences

配列データベースの収録タイムラグ

- **USGENESM** 
– 毎週更新 公報発行後 **7 日以内** に収録
- **REGISTRY**
– 毎日更新 公報発行後 **27 日以内** に収録
- **DGENE**
– 隔週更新 公報発行後 **3 ヶ月以内** に収録
- **PCTGEN**
– 毎週更新 公報発行後 **24 時間以内** に収録

各データベースの配列収録範囲

- USGENE ファイルは, USPTOから発行された **公開公報** と **登録公報** の配列情報を収録
- REGISTRY ファイルは, CA ファイルのベーシック特許と雑誌論文の配列情報を収録
- DGENE ファイルは WPI ファイルのベーシック特許の配列情報を収録
- PCTGEN ファイルは 電子的に提出されたPCT 公開公報の配列情報を収録
- 同じ特許ファミリーであっても, 明細書に記載されている配列情報が異なる場合があるので, 複数のデータベースを利用するとよい

例: 各特許の配列収録状況

L1 ANSWER 1 OF 1 CAPLUS COPYRIGHT 2006 ACS on STN

AN 1999:549367 CAPLUS

DN 131:166204

TI A system for screening for complex between transcripti control of immune function

IN Mach, Bernard

PA Novimmune S.A., Switz.


FAN.CNT 1

PATENT NO.	KIND	DATE	APPLICATION NO.	DATE
WO 9942571	A1	19990826	WO 1999-FR376	19990219
W: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, , . . .				
FR 2775003	A1	19990820	FR 1998-2025	19980219
AU 9924312	A1	19990906	AU 1999-24312	19990219
EP 1068310	A1	20010117	EP 1999-903787	19990219
R: DE, FR, GB				
JP 2002504325	T2	20020212	JP 2000-20511	20000219
US 6379894	B1	20020430	US 2000-20511	20000219
PRAI FR 1998-2025	A	19980219		
WO 1999-FR376	W	19990219		

データベースに収録されている配列の数:

- 3 配列 WO 9942571 (REGISTRY ファイル)
- 4 配列 FR 2775003 (DGENE ファイル)
- 6 配列 US 6370894 (USGENE ファイル)

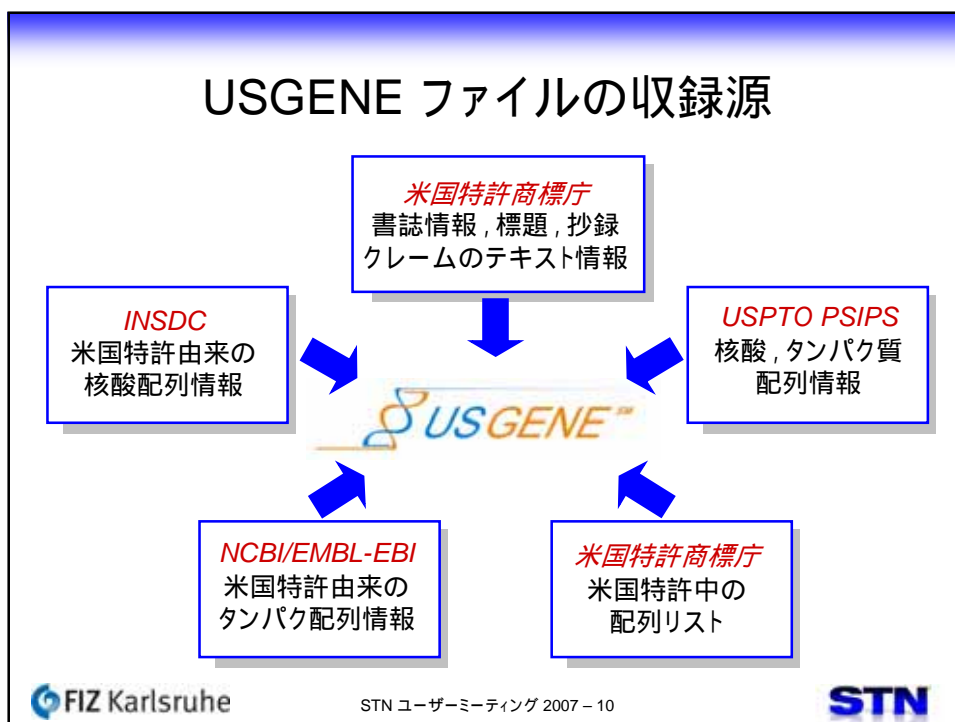
配列検索は, できる限り多くのデータベースを利用して検索することが非常に重要!

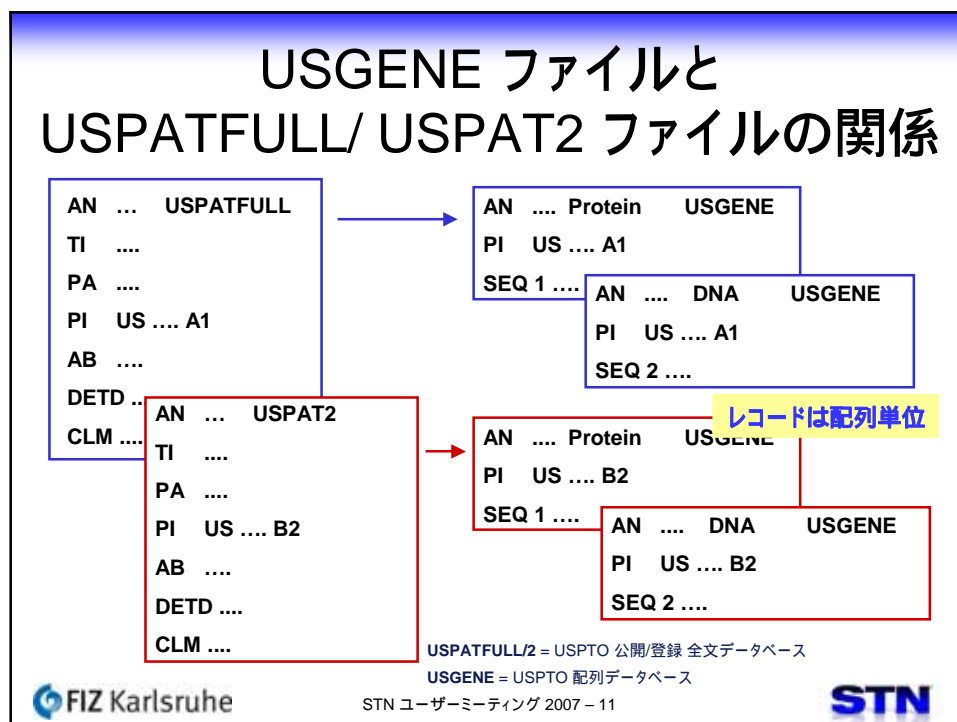


- データベース製作者:
the SequenceBase Corporation
- 収録範囲: 米国公開特許, 登録特許
- 収録件数: 約 600 万レコード
- 特許出願人から入手した情報
- 書誌情報, 標題, 抄録, クレーム (レコードは配列単位)
- 毎週更新
- 1982 年から現在

参考: <http://www.fiz-k.com/usgene>

FIZ Karlsruhe STN ユーザーミーティング 2007-9





USGENE ファイルの収録情報

- USPTO で発行された公開特許と登録特許のタンパク質と核酸配列情報を収録
- 生物名, 配列長, SEQ ID 番号, 分子タイプ
- 配列の特徴や修飾を収録する表
- 発明者 **標題, 抄録** と **クレーム**
- 特許出願人と発明者のフルネーム
- 特許情報, 出願情報, PCT 経由の特許の場合は, PCT 公開番号と日付

USGENE ファイル レコード例

```

L1 ANSWER 1 OF 1 USGENE COPYRIGHT 2006 SEQUENCEBASE CORP on STN
AN 6639063.3883 (1) DNA (2) USGENE
TI EST's and encoded human proteins (Patent) (3)
IN Edwards Jean-Baptiste Dumas Milne (Paris, FR)
   Jobert Severin (Paris, FR)
   Giordano Jean-Yves (Paris, FR)
PA Genset S A (FR)
PI US 6639063 B1 20031028 (4)
AI US 2000-621976 20000721
ORGN Homo Sapiens (5)
AB The sequences of 5' ESTs and consensus contigated 5' ESTs derived from
   mRNAs encoding secreted proteins are disclosed. The 5' ESTs and consensus
   contigated 5' ESTs may be used to obtain cDNAs and genomic DNAs
   corresponding to the 5' ESTs and consensus contigated 5' ESTs. The 5'
   (6) ESTs and consensus contigated 5' ESTs may also be used in diagnostic,
   forensic, gene therapy, and chromosome mapping procedures. Upstream
   regulatory sequences may also be obtained using the 5' ESTs and consensus
   contigated ESTs. The 5' ESTs and consensus contigated 5' ESTs may also be
   used to design expression vectors and secretion vectors.
    
```

USGENE のレコードには
書誌情報や、発明者標題、
抄録が収録されています

USGENE ファイル レコード例 (続き)

- 1) USGENE レコード番号 (AN), 配列同定番号を含む (SEQ ID 番号)
- 2) 分子タイプ (MTY)
- 3) 発明者標題 と 特許情報のフラグ
 - "PublishedApplication" (公開特許)
 - "Patent" (登録特許)
- 4) 書誌情報 - 特許情報, 出願情報, 特許出願人, 発明者
- 5) 生物種 (記載がある場合) - 配列の起源
- 6) 発明者抄録

USGENE ファイル レコード例 (続き)

CLM US6639063 B1: What is claimed is:

USGENE ファイルでは、クレームを検索することができます

1. An isolated, purified, or recombinant polynucleotide encoding a signal peptide, wherein: a) the polynucleotide encodes a signal peptide (7) consisting of residues -102 to -1 of SEQ ID NO: 3986; and b) wherein said signal peptide directs the secretion of a polypeptide when located at the amino terminus of a polypeptide.
2. The isolated, purified, or recombinant polynucleotide of claim 1, wherein: said polynucleotide consists of nucleotides 144 to 449 of SEQ ID NO: 126.
3. An isolated, purified, or recombinant polynucleotide encoding a signal peptide, wherein: a) said polynucleotide, consists of a nucleic acid sequence having at least 90% homology to the polynucleotide of claim 2; and b) said polynucleotide, encodes a si

SSO NUCLEIC; PSIPS; GRANTED (8)

USGENE ファイルは、複数の情報源からの情報を収集しています

USGENE ファイル レコード例 (続き)

SQL 231 (9)

SEQ

```

1 tactactagt ctccttgaag tatatggtgt cgccacatt ttgctgcagt
51 tcacttttaa ttctaagaa ggttgtttc acttggtgtt tttttaatct (10)
101 ctaagaatg aatagtagga atattagtag caacacctta aactcatgtc
151 acattttaa attcacagaa catctacaca cacattatgt tattaggtaa
201 acagtggtg acagcctgca ttagtttaa g
    
```

USGENE ファイルのレコードは配列単位です

FEATURE TABLE:

Key	Location
CDS	24..197

(11)

USGENE ファイル レコード例 (続き)

- 7) クレーム全文
- 8) 配列の情報源 (SSO)
 - nucleic (核酸) あるいは protein (タンパク質)
 - PSIPS/USPTO, NCBI 等
 - 公開特許 または 登録特許
- 9) 配列長 (SQL)
- 10) 配列 (SEQ)
 - USGENE のレコードは配列単位
- 11) 特徴表
 - 配列の修飾やそのほかの特徴の情報 (記載がある場合)

USGENE ファイルは, DGENE ファイル と同じように検索できます!

- NCBI BLAST ホモロジー検索
 - RUN BLAST
- FASTA ホモロジー検索
 - RUN GETSIM
- 完全配列/部分配列検索
 - RUN GETSEQ
- BATCH 検索と アラート検索

BLAST 検索の基本 7 ステップ

- 1) 質問式の保存, アップロード, 確認 (L1)
- 2) BLAST 検索
- 3) 保存回答数の決定 (L2)
- 4) 相同性の高い順に回答を並べ替え
(=> SORT SCORE) (L3)
- 5) 配列情報を含む無料の表示形式で回答を確認
例 => D L3 TRI ALIGN 1-
- 6) 表示する回答を選択して, 書誌情報等を表示
例 => D L3 BIB ALIGN 1,3,10
- 7) 検索経過が記録されていることを確認後, セッションを
切断する

1) 質問式の保存, アップロードと確認

```
=> FILE PCTGEN
```

```
=> UPL R BLAST
```

これらのコマンドは STN Express のウィザードを利用した場合に, 自動的に実行されます

```
UPLOAD SUCCESSFULLY COMPLETED
```

```
L1 GENERATED
```

```
=> D LQUE L1
```

```
L1 ANSWER 1 PCTGEN COPYRIGHT 2006 WIPO on STN
VQTVPLSRLFDHAMLEAHRAHELAIPTYQEFEEETYIPKDKYSFLHDSQT
SFCFSDSIPTPSNMEETQQKSNLELLRISLLIESWLEPVRFLRSMFANN
LVYDTSDDYHLLKDLLEGIQTLMGRLEDGSRRTGQILKQTYSKFDTNS
HNHDALLKNYGLLYCFRKDMKQVETFLRMVQCRSVEGSCGF
```

```
=>
```

この配列の L 番号は, USGENE, DGENE, PCTGEN ファイルで, そのまま利用することができます. REGISTRY ファイルの BLAST 検索では, 別の方法を用います

2) BLAST 検索

```
=> FILE USGENE
```

```
=> RUN BLAST L1 /SQP -F F
```

```
BLAST Version 2.2
```

```
The BLAST software is used herein with permission of the
National Center for Biotechnology Information (NCBI) of
the National Library of Medicine (NLM). See also, Altschul,
Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui
Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein
database search programs." Nucleic Acids Res. 25:3389-3402
```

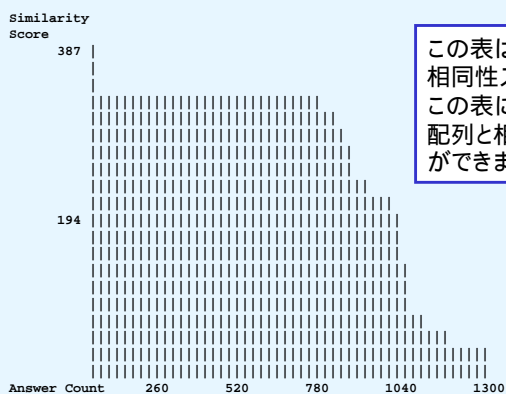
```
BLAST SEARCHING . . . .
```

USGENEは、簡単に配列特許を
検索できるデータベースです

注) この検索は USGENE テストファイルでの結果です
実際の検索結果と異なる場合があります

3) 保存回答数の決定

1297 ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0



この表は 横軸はヒットした配列数、縦軸は
相同性スコアを示しています。
この表により、回答数の中の相同性の高い
配列と相同性の低い配列の割合を知ること
ができます

すべて (ALL) の回答を指定

HOW MANY ANSWERS WOULD YOU LIKE TO KEEP ? (ALL) OR ? : ALL

4) 相同性の高い順に回答を並べ替え

```
L2      RUN STATEMENT CREATED
L2      1297 VQTVPLSRLFDHAMLEAHRAHELAIDTYQEFEETYIPKDQKYSFLHDSQT
          SFCFSDSIPTPSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANN
          LVYDTSDDYHLLKDLEEGIQTLMGRLEDGSRRTGQILKQTYSKFDINS
          HNHDALLKNYGLLYCFRKMDDKVETFLRMVQCRSVEGSCGF/SQP.-F F
```

Answer set arranged by similarity score, enter

=> S L2 AND SOMATOMAMMOTROPIN/CLM AND AY<1996 AND GRANTED/SSO

```
L3      39 L2 AND SOMATOMAMMOTROPIN/CLM AND AY<1996 AND GRANTED/SSO
```

=> **SOR L3 SCORE D**

```
PROCESSING COMPLETED FOR L3
L4      39 SOR L3 SCORE D
```

ここでは、配列検索の結果に対して、さらにクレーム中の言葉 (SOMATOMAMMOTROPIN) や、出願年 (1996 年より前)、登録特許 (GRANTED) に限定しています

配列検索後は、SORT SCORE D を実行して、相同性の高い順に並び替えます
キーワードや日付で検索結果を絞りこんだ場合には、最後に SORT SCORE D を指示します

5) 配列情報を含む無料の表示形式で回答を確認

```
=> D L4 TRI ALIGN 1-39; FILE STNGUIDE
```

```
L4      ANSWER 1 OF 39 USGENE COPYRIGHT 2006 SEQUENCEBASE CORP on STN
TI      Recombinant DNA transfer vectors (Patent)
MTY     Protein
SQL     191
ORGN    Unknown
BLASTALIGN
        Query = 191 letters
        Length = 191
        Score = 387 bits (995), Expect = e-113
        Identities = 189/191 (98%), Positives = 191/191 (100%)
Query: 1 VQTVPLSRLFDHAMLEAHRAHLAIDTYQEFEETYIPKDQKYSFLHDSQTSFCFSDSIPT
          VQTVPLSRLFDHAMLEAHRAHLAIDTYQEFEETYIPKDQKYSFLHDSQTSFCFSDSIPT
Sbjct: 1 VQTVPLSRLFDHAMLEAHRAHLAIDTYQEFEETYIPKDQKYSFLHDSQTSFCFSDSIPT
Query: 61 PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDDYHLLKDLEEG
          PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDDYHLLKDLEEG
Sbjct: 61 PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDDYHLLKDLEEG
          . . . . .
```

標題や、配列長、ホモロジー検索時のアライメント情報を無料で出力することができます
また、標題を見ることにより、その配列の収録されている特許種類が簡単にわかります

6) 表示する回答を選択して 書誌情報等を表示

=> D BIB SSO AB CLM ALIGN 1 3 10

```
L4 ANSWER 1 OF 39 USGENE COPYRIGHT 2006 SEQUENCEBASE CORP on STN
AN 4363877.1 Protein USGENE
TI Recombinant DNA transfer vectors (Patent)
IN Goodman Howard M. (San Francisco, CA)
  Shine John (San Francisco, CA)
  Seeburg Peter H. (San Francisco, CA)
PA The Regents of the University of Cal
PI US 4363877 A 19821214
AI US 1978-897710 19780419
ORGN Unknown
SSO PROTEIN; EMBL; GRANTED
AB Recombinant DNA transfer vectors con
  somatomammotropin and for human growth hormone.
```

このヒットした配列情報は、登録特許 US4363877 由来の配列で、出願年は 1978 年、さらに、クレームに somatomammotropin のキーワードを含んでいます

```
CLM US4363877 A: What is claimed is:
  1. A recombinant DNA transfer vector comprising codons for human
  chorionic somatomammotropin comprising the nucleotide . . . .
```

BLASTALIGN

まとめ

- USGENE ファイルは、DGENE ファイルや REGISTRY ファイルに通常収録されていない**米国の登録特許の配列情報**が検索できます
- USGENE ファイルは、EMBL-EBI が収録していない公開特許の配列データも収録しています
- DGENE ファイルは、特許配列検索に使われる代表的なデータベースです。
- REGISTRY ファイルは、DGENE ファイルよりも速報性に優れ、詳細な配列情報が収録されています。
- 米国の配列情報を検索する場合には、USGENE, REGISTRY, DGENE ファイルを併用すると幅広く検索することができます

新規データベース USGENE ファイル

まとめ

USGENE, DGENE, PCTFULL, REGISTRY, GenBank ファイルの収録状況

	USGENE	DGENE
収録源	米国公開特許, 米国登録特許 配列情報収録源: INSDC (核酸配列) NCBI/EMBL-EBI (タンパク質) USPTO PSIPS の配列情報, 米国特許中の配列リスト	WPI ファイルに収録されている ベーシック特許
収録期間	1982 年から現在	1981 年から現在
収録件数	600 万件	892 万件
レコード構成	配列単位	配列単位
収録情報	<ul style="list-style-type: none"> 配列関連情報 (配列, 生物名, 配列長, 特徴表) 特許の書誌情報 (標題, 特許出願人, 特許情報, 出願情報, 優先権出願情報など), 抄録, 全クレーム 	<ul style="list-style-type: none"> 配列関連情報 (配列, 生物名, 配列長, 特徴表) 特許の書誌情報 (標題, 特許出願人, 特許情報, 出願情報, 優先権出願情報など) 配列収録源の特許番号と配列の記載位置, 配列番号 WPI の対応特許情報 INPADOC の法的状況データ 英文抄録, 索引
実行方法	RUN コマンドを用いたパッケージプログラム	
検索機能	<ul style="list-style-type: none"> 完全配列検索 部分配列検索 ホモロジー検索 (BLAST, GETSIM) 完全配列ファミリー検索, 部分配列ファミリー検索 (タンパク質のみ) 	
更新頻度	毎週	隔週
タイムラグ	7 日以内	3 ヶ月

新規データベース USGENE ファイル

まとめ

(2007 年 5 月現在)

PCTFULL	REGISTRY	GenBank
特許出願人により、電子的に提出されたデータ	CAplus 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 7 月以降は出典のあるもののみ)	雑誌論文 研究者から直接受領したデータ 未発表, 不完全な残基を含む部分配列も収録 一部の特許 (米国, PCT 出願, 他) EMBL, DDBJ 由来のデータ
2001 年 8 月から現在	1957 年から現在	1982 年から現在
125 万件	5,985 万件	7,517 万件
配列単位	配列単位	配列単位
<ul style="list-style-type: none"> 配列関連情報 (配列, 生物名, 配列長, 特徴表) 出願の書誌情報 (標題, 特許出願人, 特許情報, 出願情報, 優先権出願情報など) 	<ul style="list-style-type: none"> 配列関連情報 (配列, 生物名, 配列長, 特徴表, CAS 登録番号) 配列収録源の特許番号と配列の記載位置, 配列番号 構造, 分子式 REGISTRY ファイルから CAplus/CA ファイルのクロスオーバー検索を利用すれば, 配列が収録されている特許や雑誌の書誌情報, 対応特許, 英文抄録, 索引を見ることができる 	<ul style="list-style-type: none"> 配列関連情報 (配列, 生物名, 配列長, 特徴表, CAS 登録番号, GenBank 番号) 出典の情報 (標題, 収録源, 著者名など)
RUN コマンドを用いたパッケージプログラム	<ul style="list-style-type: none"> ホモロジー検索: 専用ソフト ホモロジー検索以外: SEARCH コマンド 	<ul style="list-style-type: none"> 配列検索は不可 REGISTRY ファイルからのクロスオーバー検索が可能 (CAS 登録番号付与率は 58 %)
<ul style="list-style-type: none"> 完全配列検索 部分配列検索 ホモロジー検索 (BLAST, GETSIM) 完全配列ファミリー検索, 部分配列ファミリー検索 (タンパク質のみ) 	<ul style="list-style-type: none"> 完全配列検索 部分配列検索 ホモロジー検索 (BLAST) 完全配列ファミリー検索, 部分配列ファミリー検索 (タンパク質のみ) 	
毎週	毎日	毎日
7 日以内	27 日以内 (主要国)	