





1. SciFinder の配列情報

1-1. 配列情報の概要

データベース名	REGISTRY ファイル	
製作者	CAS (Chemical Abstracts Service)	
概 要	<ul style="list-style-type: none"> ・CA に収録された雑誌論文, 特許中の主要化学物質の辞書ファイル ・現在, 約 70 % のレコードがタンパク質, ペプチド, 核酸の配列情報 	
収録源	<ul style="list-style-type: none"> ・CAplus, CA, CAOLD に収録されている雑誌論文, 特許 (54 ヶ国 + 3 国際機関) ・GenBank からの情報 	
収録期間	1957 年 ~ 現在 (一部, 1957 年以前の文献由来の配列もある)	
レコード構成	物質単位 (一配列ずつ)	
収録情報	タンパク質配列	○
	核酸配列	○
	その他	<ul style="list-style-type: none"> ・配列に関連する情報(生物名, 配列長, 特徴表など) ・構造, 分子式 (短い配列のみ)
収録件数	核酸	約 5,520 万件
	タンパク質	約 710 万件
	合計	約 6,230 万件
更新頻度	毎 日	
CAS RN 付与率	100 %	
配列検索機能	実行方法	<ul style="list-style-type: none"> ・ 配列検索機能は Windows 版のみ ・ Explore ボタン  から Nucleotide or Protein Sequence  を選択 ・ KMP (Keep Me Posted) 機能を用いて, アラート登録も可能
	核酸	・ホモロジー検索 (BLASTn, tBLASTn, tBLASTx)
	タンパク質	・ホモロジー検索 (BLASTp, BLASTx)
タスク	1タスク (配列検索→文献検索→出力)	

▼ 核酸配列の収録基準

- ・ 塩基数が 9 以上の核酸. ただしプローブ, プライマーについては長さの制限はない.
- ・ CAplus ファイルの特許由来の配列
 - 1957 年～1999 年 9 月 : 請求範囲および実施例でクレームされているか, 新規性
に
関与している完全配列*.
 - 1999 年 10 月～ : 特許由来の配列: 特許に記載されたすべての配列*1.
新規性の有無や記載位置は関係ない.
- *1 2005 年～: 4,000 以上の配列を記載している特許の配列は収録しない.
- ・ CAplus ファイルの非特許由来の配列
 - 1957 年～ : 新規性に関与している完全配列*
- ・ GenBank ファイル由来の配列*2
 - 1957 年～2005 年 2 月 : 未発表, 不完全な残基を含む部分配列も収録
 - 2005 年 3 月～ : 参照文献情報をもたない配列, および GSS (Genome Survey
Sequence: ゲノム由来の断片配列)・EST (Expressed
Sequence Tags: 発現配列タグ)は収録対象外となった

*2 2007 年～: 1000 以上の GenBank アクセッション番号を記載している雑誌論文の配列は収録しない.

▼ 核酸配列の登録のルール

- ・ 一つでも核酸塩基が異なるものは別の核酸として登録.
- ・ 配列が同じであっても, 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ものは別の核酸として登録 (その情報は特徴表に表記される).
- ・ CAplus ファイル由来の配列と GenBank 由来の配列は, 同じ配列でも別レコードとして登録.
- ・ GenBank 由来の更新前と更新後の配列は, 同じ配列でも 別レコードとして登録.
- ・ CAplus ファイルの 2002 年以降の特許由来の配列については, 修飾基を含めて全く同一の配列でも特許番号ごとに別レコードとして登録.

▼ 収録される配列例

- 天然の核酸
- 融合遺伝子と核酸
- 遺伝子工学による遺伝子と合成遺伝子
- 化学修飾された遺伝子とオリゴヌクレオチド
- オリゴヌクレオチドプローブと PCR(複製連鎖反応)プライマー
- 一般的でないヌクレオチドを有する配列
- タンパク質または RNA 産生のためのコード情報をもつ遺伝子配列
- 調整領域
- ペプチド核酸(PNA)

* 完全配列とは

- 著者が完全だと明記した配列
- より大きな核酸分子に含まれ, タンパク質または RNA 産生の全てのコード情報を有する遺伝子領域
- 開始と終了を有する遺伝子領域(プロモータ, 調整領域など)

▼ 核酸配列のレコード例

Detail of Substance 1

File Edit Help

Registry Number: 593306-24-2

CA Index Name: DNA (mouse immunoglobulin κ -chain V region C-terminal fragment-specifying cDNA plus 3'-flank) (9C1)

Other Names: 75: PN: US20030166562 TABLE: 1 claimed DNA; GenBank AB017349

Class Identifier: Manual Registration

Sequence Length: 360

Nucleic Acid Count: 82 a 103 c 90 g 85 t

GenBank (R) Definitions and Features:
 ACCESSION NUMBER: AB017349
 VERSION NUMBER: AB017349.1 GI:4519566
 DEFINITION: Mus musculus mRNA for immunoglobulin light chain V region, partial cds.
 ORGANISM: Mus musculus
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus

FEATURE TABLE:

Feature Key	Location	Qualifier
source	1..360	/organism="Mus musculus" /db-xref="taxon:10090"
gene	1..360	/gene="12/B1 kappa"
CDS	<1..354	/gene="12/B1 kappa" /note="derived from anti-idiotypic antibody against anti-gibberellin A4 antibody" /codon-start=1 /product="immunoglobulin light chain V region" /protein-id="BAA75653.1" /db-xref="GI:4519567" /db-xref="IMGTLIGM:AB017349" /translation="AAAQIVLTQSPA\MSASPGEKVTMTCSASSV SFMVWYQQKPRSSPKPWYLTSLNLSGVPARF SGSGSGTSYSLTISSMEAEADAATYYCQQWSSN PLTFGAGTKLELKRVDAAPTV"

Annotations:

Source	Feature	Location	Description	Reference
Not Given			US20030166562 TABLE 1; claimed	

Sequence:

```

1  gcggcgcgac aaattgttct caccagctct ccagcagtta tgtctgcac
51  tccaggggag aaggtcacca tgacctgcag tgccagctca agtgtaaagt
101 tcatgtactg gtaccagcag aagccaagat cctcccccac accctggatt
151 tatctcacat ccaacctggc ttctggagtc cctgctcgtc tcagtggcag
201 tgggtctggg acctcttact ctctcacaat cagcagcatg gaggctgaag
251 atgctgccac ttattactgc cagcagtgga gtagtaatcc gctcagttc
301 ggtgctggga ccaagctgga gctgaaacgg gttgatgctg caccaactgt
351 ataagagctc
  
```

-- Resources --

References: ~1

STN Files: CAPLUS, CA, GENBANK, TOXCENTER, USPATFULL

(Additional Information is available through STN International. Contact your information specialist, a local CAS representative, or the CAS Help Desk for Assistance)

Database: REGISTRY

Close

特許由来の配列は特許番号も

GenBank 由来の情報も

特許中の記載位置

▼ タンパク質・ペプチド配列の収録基準

- ・ アミノ酸残基数が 4 以上のタンパク質, ペプチド
- ・ CAplus ファイルの特許由来の配列
 - ～1987 年 : 請求範囲および実施例でクレームされているか, 新規性に関与している完全配列*.
 - 1988 年～ : 上記に加え, クレームされているか, 新規性に関与している部分配列 (ギャップを含まないもの).
 - 1999 年～ : 特許に記載されたすべての配列*¹.

*1 2005 年～:4,000 以上の配列を記載している特許の配列は収録しない。

- ・ CAplus ファイルの非特許由来の配列
 - ～1990 年 : 新規性に関与している完全配列*.
 - 1991 年～ : 上記に加え, 選択された雑誌由来の新規性に関与している部分配列 (20 以上のアミノ酸配列でギャップを含まないもの).
 - 1999 年～ : 新規性に関与しているすべての完全配列*, および部分配列.
- ・ GenBank 由来の翻訳配列
 - GenBank の核酸配列の仮コーディング領域から翻訳したタンパク質配列.
 - Other Names には, 翻訳された核酸の GenBank 番号が含まれる.
 - ～2005 年 2 月 : 未発表, 不完全な残基を含む部分配列も収録
 - 2005 年 3 月～ : 出典のある配列のみ収録

▼ タンパク質・ペプチド配列の登録のルール

- ・ 一つでもアミノ酸が異なるものは別のペプチド, タンパク質として登録.
- ・ 配列が同じであっても, 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ものは別のペプチド, タンパク質として登録.
- ・ CAplus ファイル由来の配列と GenBank 由来の配列は, 同じ配列でも別レコードとして登録.
- ・ GenBank 由来の配列は, 1 GenBank 番号につき 1 つのレコードとして登録.
- ・ CAplus ファイルの 2002 年以降の特許由来の配列については, 修飾基を含めて全く同一の配列でも特許番号ごとに別レコードとして登録.

▼ 収録される配列例

- 天然のペプチド, タンパク質
- 多鎖タンパク質
- 環状ペプチド
- 特異なアミノ酸を含む配列
- 融合タンパク質
- ペプチド金属錯体
- 化学修飾されたペプチド, タンパク質
- 遺伝子操作または合成で得られたタンパク質

* 完全な配列とは

- 著者が完全だと明記した配列
- 核酸配列 (AUG コドンで始まりストップコドンで終了する) から翻訳された配列

▼ タンパク質配列のレコード例

Detail of Substance 1

File Edit Help

Registry Number: 594889-71-1

CA Index Name: Transport protein, org. cation transporter (human alternative splicing isoform hOCT2-A) (9CI)

Other Names: 2: PN: JP2003250576 SEQID: 2 claimed protein 特許由来の配列は特許番号も

Class Identifier: Manual Registration

Sequence Length: 483

Annotations:

Source	Feature	Location	Description	Reference
Not Given				JP2003250576 SEQID 2; claimed 特許中の記載位置

Sequence:

```

1  MPTTVDDVLE  HGGEFHFFQK  QMFFLLALLS  ATFAPYVGI  VFLGFTPDHR
51  CRSPGVAELS  LRCGWSPAEE  LNYTVPGPGP  AGEASPRQCR  RYEVDWNQST
101 FDCVDPLASL  DTNRSRLPLG  PCRDRGWVYET  PGSSIVTEFN  LVCANSWMLD
151 LQSSVNVVGF  FIGSMSIGYI  ADRFGRKLCL  LTTVLINAAA  GVLMAISPTY
201 TWMLIFRLIQ  GLVSKAGWLI  GYLITFVVG  RRYRRTVGF  YQVAYTVGLL
251 VLAGVAYALP  HWRWLQFTVA  LPNFFFLYY  WCIPESPRWL  ISQNKNAEAM
301 RIIKHIAKKN  GKSLPASLQR  LRLEEETGKK  LNPSFLDLVR  TPQIRKHTMI
351 LMYNWFVTSSV  LYQGLIMHMG  LAGDNIYLDF  FYSALVEFPA  AFMIILTIDR
401 IGRRYPWAAS  NMVAGAACLA  SVFIPGGKFQ  VKLESYLQDP  GERECHGPLI
451 GKPCNLSSKS  IWKDKLECSI  WDPSEQIHMA  SLL

```

-- Resources --

References: ~1

STN Files: CAPLUS, CA

Database: REGISTRY

Close

2. SciFinder のホモロジー検索機能

2-1. 検索機能の概要

▼ ホモロジー検索の主なプログラム

プログラム	概要
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桁違いに高速処理できる。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。
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ン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。
Smith-Waterman	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スコアを算定する。このため、計算量は他のプログラムとは比較にならないほど膨大になる。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。

▼ 核酸ホモロジー検索の検索タイプ

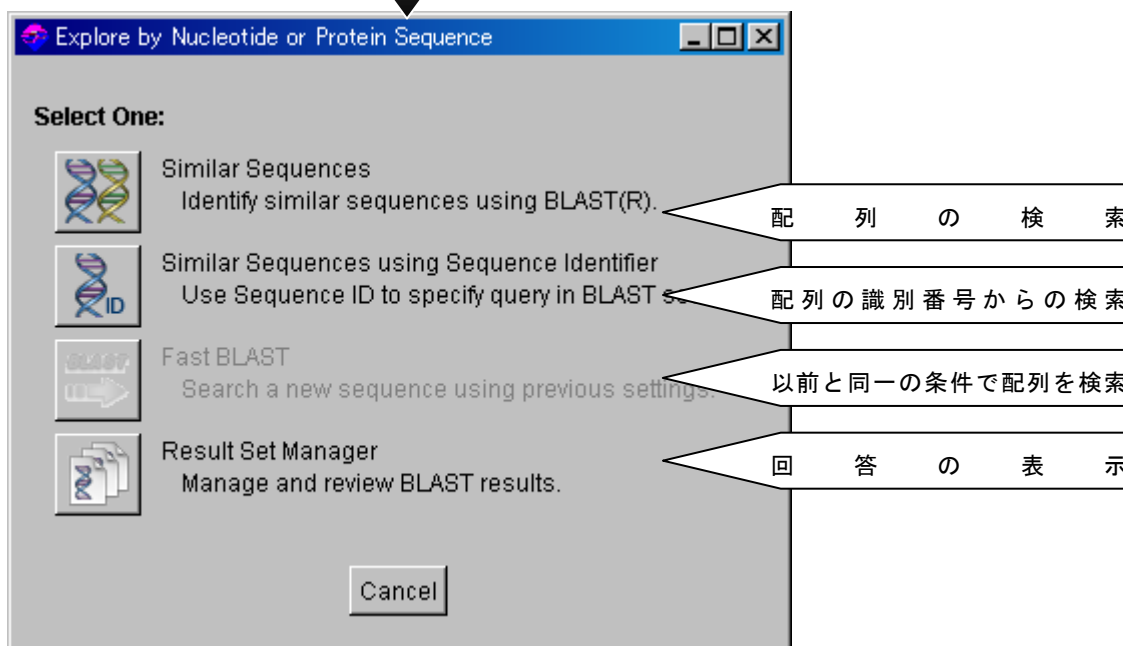
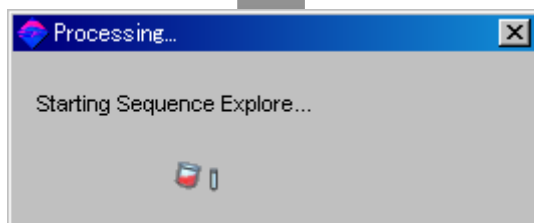
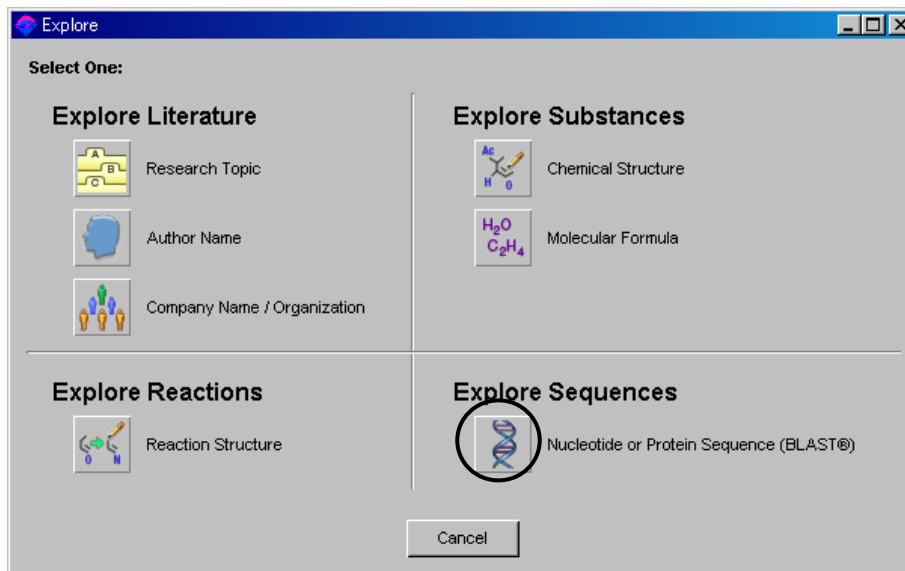
検索タイプ	検索機能	配列質問式
BLASTn	塩基配列の質問式に類似した塩基配列を検索	塩基配列
tBLASTn	アミノ酸配列の質問式を塩基配列に翻訳して、これに類似した塩基配列を検索	アミノ酸配列
tBLASTx	塩基配列の質問式をアミノ酸配列に翻訳して、これに類似したアミノ酸配列に翻訳された塩基配列を検索	塩基配列

▼ タンパク質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
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、これに類似したアミノ酸配列を検索	塩基配列

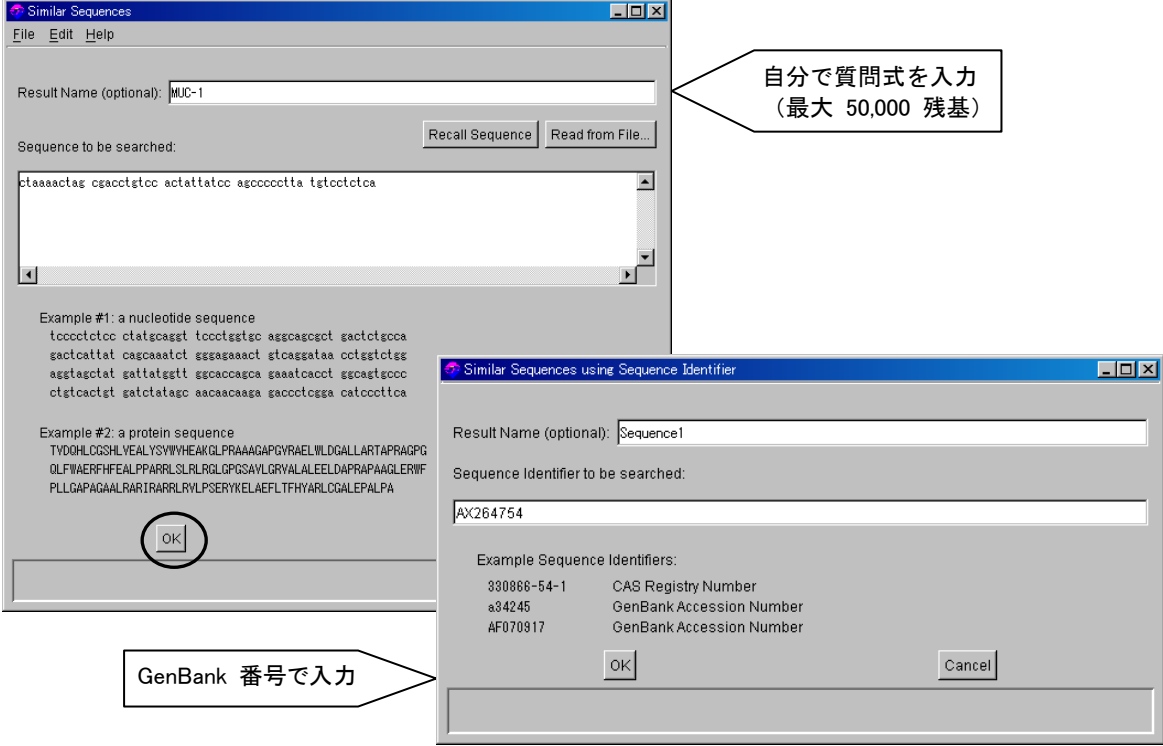
1. プログラムの起動

Explore ダイアログボックスから Nucleotide or Protein Sequence を選ぶと、しばらくして Explore by Nucleotide or Protein Sequence ダイアログボックスが表示される。

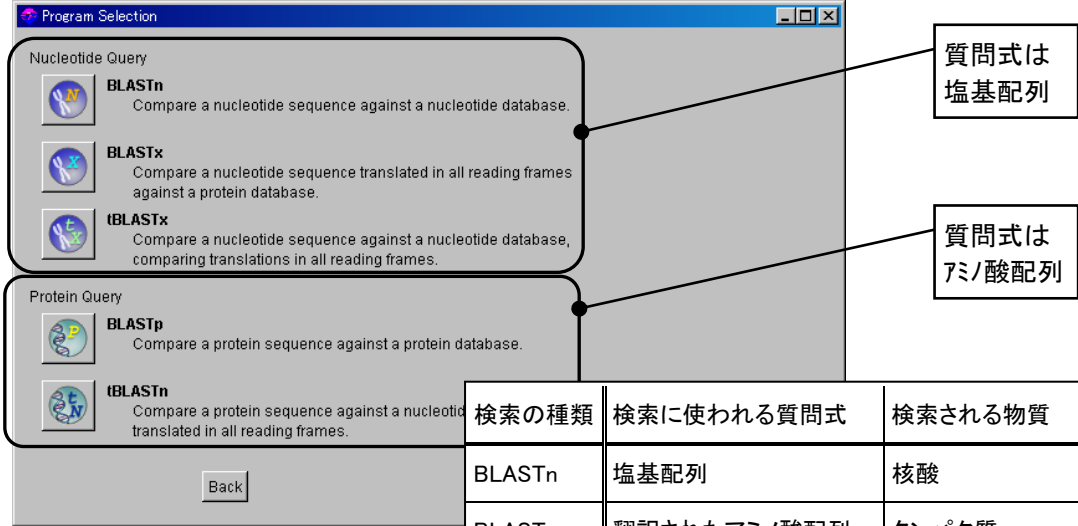


2. 配列質問式の入力

Similar Sequences 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. 質問式の名称および質問式を入力し, OK をクリックする. 質問式は, 別のファイルから読み込むこともできる. また, Similar Sequences using Sequence Identifier をクリックして, 質問式を CAS 登録番号または GenBank 番号で指定することもできる.



Program Selection ダイアログボックスが表示されるので, 利用したい検索の種類を選択する.

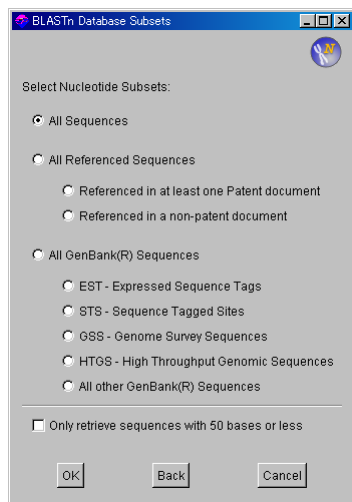


検索の種類	検索に使われる質問式	検索される物質
BLASTn	塩基配列	核酸
BLASTx	翻訳されたアミノ酸配列	タンパク質
tBLASTx	翻訳されたアミノ酸配列	アミノ酸配列に翻訳された核酸
BLASTp	アミノ酸配列	タンパク質
tBLASTn	翻訳された塩基配列	核酸

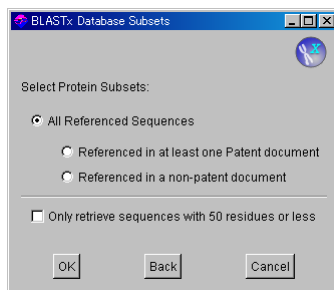
3. 検索条件の設定

核酸配列の検索を選択すると、Select Nucleotide Database Subsets ダイアログボックスが表示される。検索する範囲を選択する。CA より収録された配列, GenBank から収録された配列, 双方からの配列が選択できる。CA からの収録配列は特許由来の配列に限定することもできる。

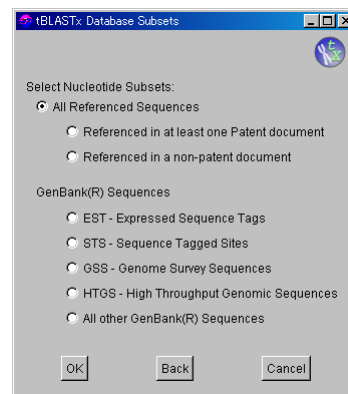
BLASTn



BLASTx

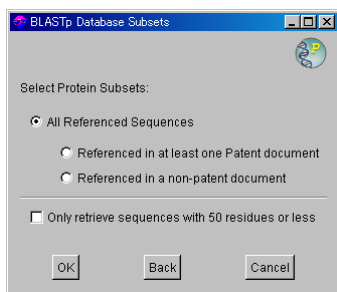


tBLASTx

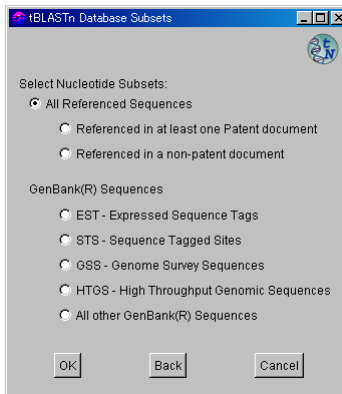


タンパク質配列の検索を選択すると、Select Protein Database Subsets ダイアログボックスが表示される。検索する範囲を選択する。

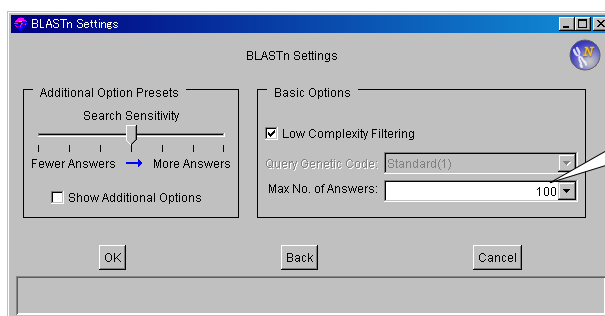
BLASTp



tBLASTn



OK をクリックすると、<BLAST> Settings ダイアログボックスが表示され、その他のオプションが設定できる。OK をクリックすると検索が実行される。

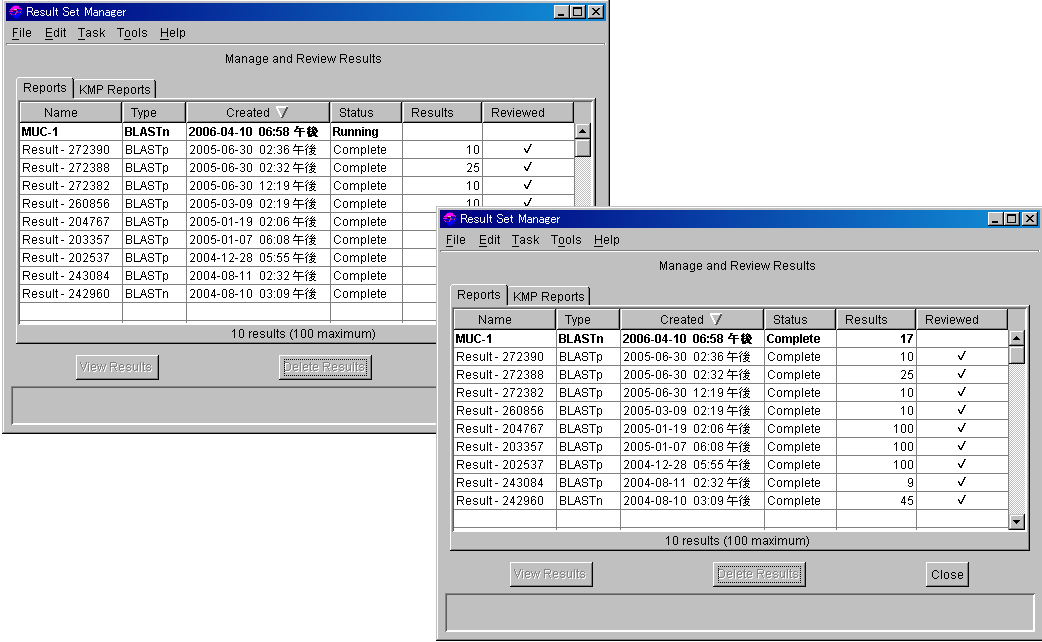


回答件数の上限
(最大 1,000)

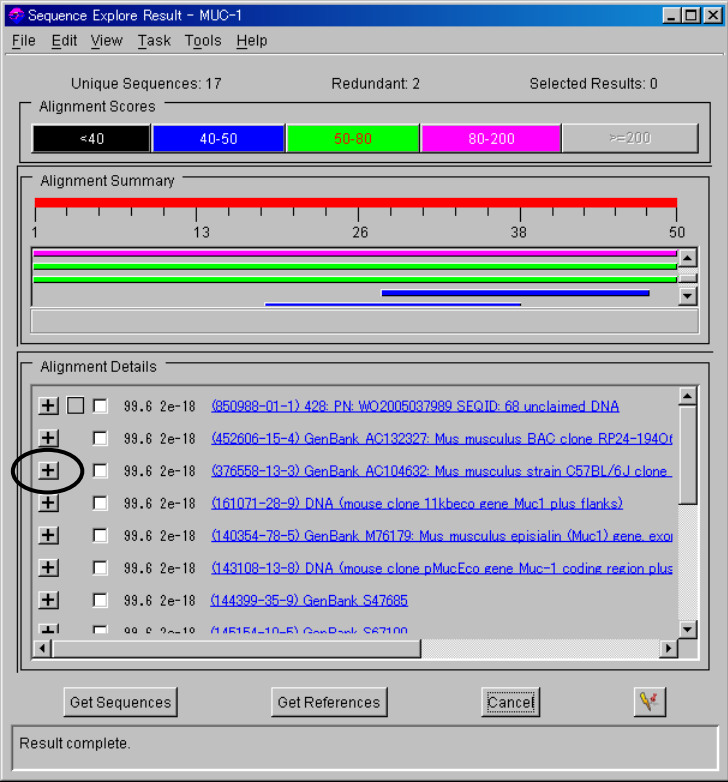
配列検索を実行している間には、別の SciFinder の検索を行うことができる。

4. 配列検索結果の表示

配列検索結果を表示するには、Explore by Nucleotide or Protein Sequence ダイアログボックスで Result Set Manager を選択する(p7 の下図参照)。以前、自分のログイン ID を用いて実行された配列検索の状況も確認できる。実行中の検索については、Status が Running になっている。検索が終了すると、Status は Complete になる。



表示する回答をハイライトし、View Results をクリックすると、Sequence Explore Result-[質問式名] ダイアログボックスが表示される。回答の左の+記号をクリックすると、各配列の詳細が表示される。なお、すべての配列の詳細は、File メニューから Print from Browser を選択するとブラウザ上に表示される。



件数情報
 類似度分布
 アライメントの概略
 アライメントの詳細
 (類似度の高い順)

(配列の詳細画面例)

Alignment Details

99.6 2e-18 [\(376558-13-3\) GenBank AC104632: Mus musculus strain C57BL/6J clone](#)

Length = 210499
Score = 99.6 Expect = 2e-18
Identities = 50/50 (100%)
Strand = Plus / Minus

Query: 1 ctaaaactagcgacctgtccactattatccagcccccttatgtcctctca 50
|||||
Subject: 180443 ctaaaactagcgacctgtccactattatccagcccccttatgtcctctca 180394

99.6 2e-18 [\(161071-28-9\) DNA \(mouse clone 11kbeco gene Muc1 plus flanks\)](#)

99.6 2e-18 [\(140354-78-5\) GenBank M76179: Mus musculus episialin \(Muc1\) gene_exo](#)

(すべての配列の詳細レポート例)

MUC-1 - Microsoft Internet Explorer

Sequence Explore Results

Query Input

Result Name: MUC-1
Program: BLASTn
Subsets: Patents
Non-patents
EST
STS
GSS
HTGS
Other GenBank(R)

Return Sequence Length: All
Date Range: Up to 2006-04-10
Low Complexity Filtering: On
Max No. of Answers: 100
Expectation Value: 10
Word Size: 11
Open Gap Cost: 5
Extend Gap Cost: 2
Penalty for Mismatch: -3
Reward for Match: 1

Sequence:
ctaaaactag cgacctgtcc actattatoc agccccctta tgcctctca

Result Summary

Result Name: MUC-1
Program: BLASTn
Creation Date/Time: 06/04/10 18:58
Unique Sequences: 17
Total Sequences: 19

Alignment Details

(850908-01-1) 428; PN: W02005037989 SEQID: 68 unclaimed DNA
Length = 3576 Score = 99.6 Expect = 2e-18

(548574-13-6) 68; PN: US20030118592 SEQID: 68 unclaimed DNA
(140321-64-8) GenBank M64928: Mus musculus mucin (Muc-1) gene, exons 1-6.

Score = 99.6 Expect = 2e-18
Identities = 50/50 (100%)
Strand = Plus / Plus

Query: 1 ctaaaactagcgacctgtccactattatccagcccccttatgtcctctca 50
|||||
Subject: 1 ctaaaactagcgacctgtccactattatccagcccccttatgtcctctca 50

(452606-15-4) GenBank AC132327: Mus musculus BAC clone RP24-19406 from 3, complete sequence.
Length = 174267 Score = 99.6 Expect = 2e-18

Score = 99.6 Expect = 2e-18
Identities = 50/50 (100%)
Strand = Plus / Plus

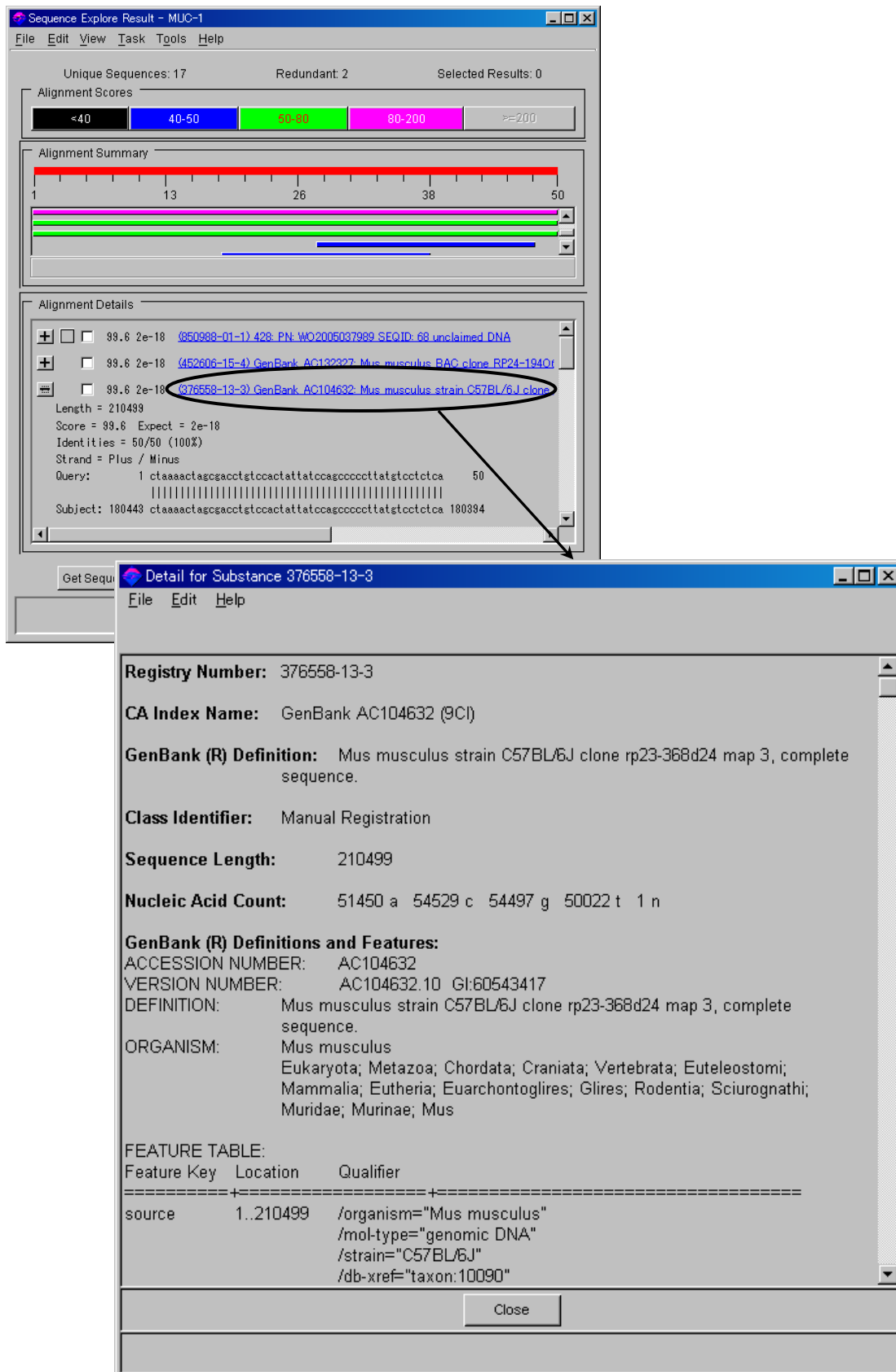
Query: 1 ctaaaactagcgacctgtccactattatccagcccccttatgtcctctca 50
|||||
Subject: 85992 ctaaaactagcgacctgtccactattatccagcccccttatgtcctctca 86041

(376558-13-3) GenBank AC104632: Mus musculus strain C57BL/6J clone rp23-368d24 map 3, complete sequence.
Length = 210499 Score = 99.6 Expect = 2e-18

Score = 99.6 Expect = 2e-18
Identities = 50/50 (100%)
Strand = Plus / Minus

Query: 1 ctaaaactagcgacctgtccactattatccagcccccttatgtcctctca 50
|||||
Subject: 180443 ctaaaactagcgacctgtccactattatccagcccccttatgtcctctca 180394

名称をクリックすると、各配列の詳しい情報が表示される。



5. 配列集合・文献集合の作成

Sequence Explore Result - [質問式名] ダイアログボックスで、興味のある配列にチェックを付けた後に、Get Sequences をクリックすると配列の集合が作成される。Get References をクリックすると、該当する配列を収録する文献が表示される。

The diagram illustrates the workflow for creating sequence and reference collections in SciFinder. It begins with the 'Sequence Explore Result' dialog box, where users can filter alignment scores and view alignment details. The 'Get Sequences' button is used to create a collection of sequences, and the 'Get References' button is used to retrieve literature references for the selected sequences. The resulting SciFinder windows show the sequence collection and the corresponding reference list.

SciFinder の配列検索は検索内容の機密を守る為に SSL プロトコルを採用している。社内の firewall の設定によってはそのままでは利用できない場合がある。

・ 生化学

1. 薬理学
2. ホルモン薬理学
3. 生化学的遺伝学
4. 毒物学
5. 農芸化学的 생물調節劑
6. 生化学一般
7. 酵素
8. 放射線化学
9. 生化学の方法
10. 微生物生化学
11. 植物生化学
12. 非ほ乳類生化学
13. ほ乳類生化学
14. ほ乳類病理生化学
15. 免疫化学
16. 発酵, 工業生物化学
17. 食品, 飼料化学
18. 動物栄養
19. 肥料, 土壌, 植物栄養
20. 歴史, 教育, ドキュメンテーション

・ 有機化学

21. 有機化学一般
22. 物理有機化学
23. 脂肪族化合物
24. 脂環式化合物
25. ベンゼン, ベンゼン誘導体, 縮合ベンゼノイド化合物
26. 生体分子, 合成類似体
27. 複素環式化合物(ヘテロ原子1個)
28. 複素環式化合物(ヘテロ原子2個)
29. 有機金属, 有機メタロイド化合物
30. テルペン, テルペノイド化合物
31. アルカロイド
32. ステロイド
33. 炭水化物
34. アミノ酸, ペプチド, タンパク質

・ 高分子化学

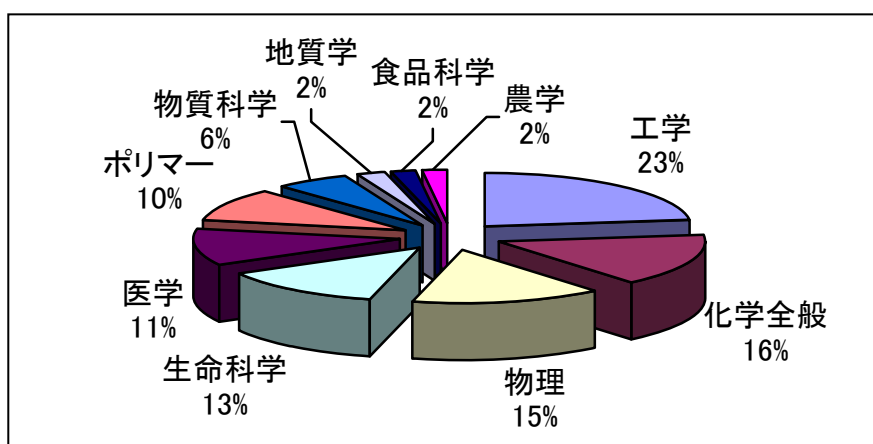
35. 合成高分子の化学
36. 合成高分子の物理的性質
37. プラスチックの製造, 加工
38. プラスチックの成型, 用途
39. 合成エラストマー, 天然ゴム
40. 織物
41. 染料, 蛍光増白剤, 写真増感剤
42. 塗料, インク, 関連製品
43. セルロース, リグニン, 紙, その他の木材製品
44. 工業炭水化物
45. 工業有機化学製品, 皮革, 脂肪, ロウ
46. 界面活性剤, 洗淨剤装置, 工場設備

・ 応用化学・化学工学

47. 装置, 工場設備
48. 単位操作, プロセス
49. 工業無機化学製品
50. 推進薬, 爆薬
51. 化石燃料, 誘導製品, 関連製品
52. 電気化学的, 放射及び熱エネルギー工学
53. 鉱物, 地質化学
54. 抽出冶金学
55. 鉄, 鉄合金
56. 非鉄金属, 合金
57. セラミックス
58. セメント, コンクリート, 関連建設材料
59. 大気汚染, 産業衛生
60. 廃棄物処理, 処分
61. 水
62. 精油, 化粧品
63. 薬剤
64. 薬剤分析

・ 物理化学・無機化学・分析化学

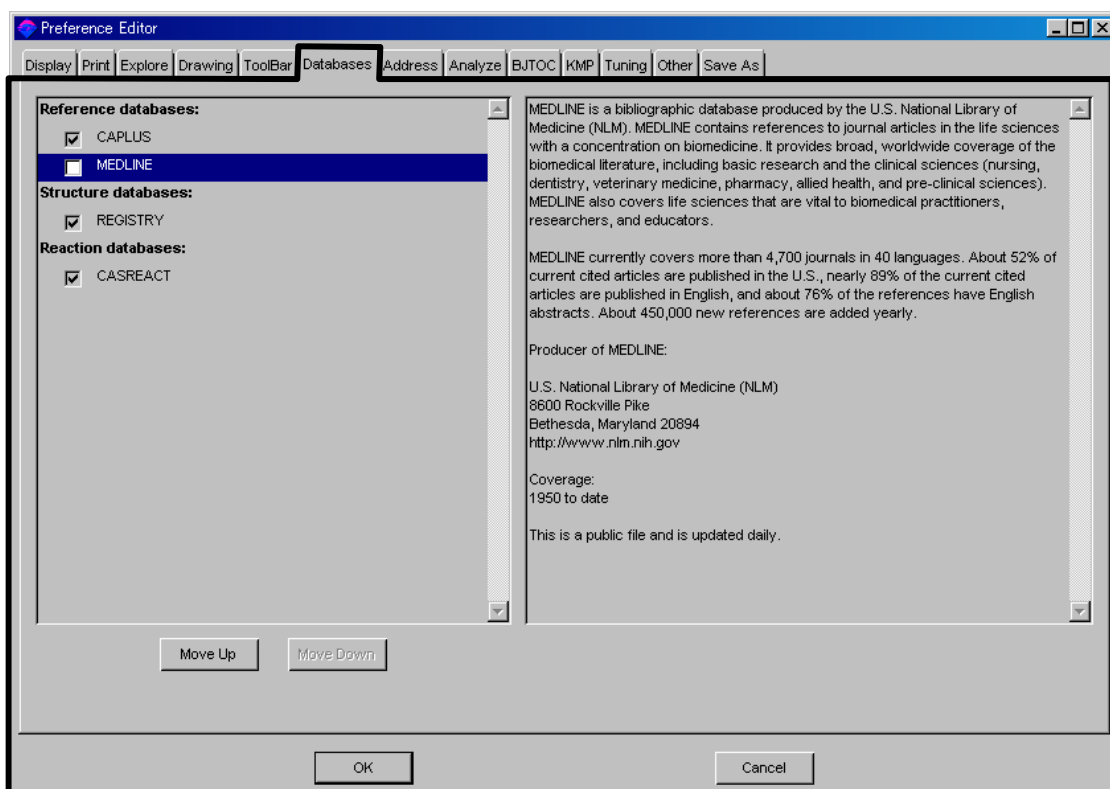
65. 物理化学一般
66. 界面化学, コロイド
67. 触媒化学, 反応動力学, 無機反応機構
68. 相平衡, 化学平衡, 溶液
69. 熱力学, 熱化学, 熱的性質
70. 原子核現象
71. 原子核工学
72. 電気化学
73. 光, 電子, 質量分光学, その他の関連する性質
74. 放射線化学, 光化学, 写真, その他の複写プロセス
75. 結晶学, 液晶
76. 電気的性質
77. 磁氣的現象
78. 無機化学薬品, 反応
79. 無機分析化学
80. 有機分析化学



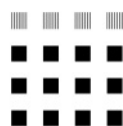


SciFinder では、各種初期設定を変更することができます。

タスクパッケージ契約の場合、デフォルトでは文献検索対象ファイルが、CAplus ファイルのみになっている。必要に応じて MEDLINE ファイルも検索対象にすることができる。



- ・ 配列検索を利用した後に文献情報を表示する場合は、MEDLINE に対する追加タスクはかからない。また、化学物質からの Explore (構造, 名称, 分子式, CAS 登録番号) を利用した後に、文献情報を表示した場合も同様に MEDLINE に対する追加タスクはかからないが、その他の検索では追加タスクがかかるので注意が必要である。



JAICI 社団法人 化学情報協会

情報事業部 ヘルプデスク

〒113-0021 東京都文京区本駒込6-25-4 中居ビル

TEL: 0120-003-462 FAX: 03-5978-3600

URL: www.jaici.or.jp

E-mail: support@jaici.or.jp