

「持続可能型社会への貢献遺伝子候補の探索」

— 持続可能型社会への貢献遺伝子データベースの構築 —

文部科学省が進めている「ライフサイエンス分野の統合データベース整備事業」の人材養成プログラムの活動として、ここに紹介するデータベースを具体例に、長浜バイオ大の学生実習においてバイオ分野のデータベースを作成しており、そのテキストの公開を計画しています。本テキストはその目的で作成しており、これを使えばあなたの家からでも、学校からでもインターネットを使って、「持続可能型社会へ貢献する可能性を持つ遺伝子」を発掘できます。着目する対象を医薬学の分野の課題へ変更すれば、「健康に貢献する遺伝子」の発掘も可能になります。

バイオサイエンス分野やバイオ産業分野で重要になる、多様なデータベースの統合的な利用やバイオデータベースの作成に関する基礎技術の習得が可能になります。

あなたも「持続可能型社会への貢献遺伝子データベース」の作成に参加しませんか。

長浜バイオ大学 バイオサイエンス学部
生命情報科学コース
生体分子情報学研究室 編集

1 はじめに

自然環境の保全や浄化に役立つバイオ技術の開発やその教育は、21 世紀にますます重要になる課題です。環境浄化や保全に役立つ、広い意味では「持続可能型社会の実現に貢献できる」可能性を持つ遺伝子を、国際塩基配列データベース (DDBJ/EMBL/GenBank) に登録されている塩基配列から発掘し、新規なデータベース「持続可能型社会への貢献遺伝子データベース」として世界へ発信しましょう。

現在、我々が知り得ている微生物は、環境中に生育する微生物の 0.1%より遥かに少ないと言われており、多様な自然環境で生育する微生物類については、培養が困難な例が大半を占めています。これら難培養性微生物については、通常の実験的な研究がなされておらず、膨大なゲノム資源が科学的にも産業的にも未開拓のままに残されています。これら難培養性微生物類のゲノムは新規な遺伝子類を豊富に保有すると考えられ、難培養性微生物類を多数含む環境中の試料から、培養を行わずにゲノム DNA の混合物を抽出し、ゲノム断片をクローン化し、その断片の配列決定を行い、産業上有用な遺伝子を探索する試みが世界各地で大規模に行われています。このような試みは、「メタゲノム解析」と呼ばれており、インターネットを使って <http://www.genomesonline.org/gold.cgi> の Metagenomes の項目を参照すると、最新の状況を知ることができます。

しかしながら、このメタゲノム解析で得られた環境由来 DNA 配列は、遺伝子に関する情報の品質、特に遺伝子機能についての記載が登録者によってまちまちであり、有用な遺伝子を含んでいる可能性が高いにもかかわらず、利用者にとって必要な情報が記載されていない場合が多数あります。そこで、本テキストでは、「自然環境の浄化や保全に役立つ遺伝子」の発掘を目標に、国際塩基配列データベース (DDBJ/EMBL/GenBank) に収録されている環境由来 DNA 配列から有用な遺伝子を探索する、具体的な方法を紹介します。

長浜バイオ大では3年生の全員を対象に、このテキストに沿った学生実習を行っており、環境由来 DNA 配列から多数の遺伝子を新規に発掘しています。それらの新規遺伝子を「持続可能型社会への貢献遺伝子データベース」として収録を開始しており、学生の氏名を記入して世界へ発信する予定です。本テキストに沿って、皆さんも「自然環境の浄化や保全に役立つ」広い意味では「持続可能型社会への貢献遺伝子」の候補を探索して下さい。長浜バイオ大の学生が見つけていない、新しい遺伝子を発掘された方で、希望される方は、長浜バイオ大の「持続可能型社会への貢献遺伝子データベース」へ登録が可能です。氏名を記入して世界へ発信するシステムを、「ライフサイエンス分野の統合データベース整備事業」の人材養成活動の一環として作成中です。

2 「持続可能型社会の実現に、貢献が期待できる遺伝子やタンパク質」の候補の検索

まず自分の興味のある「持続可能型社会の実現に、貢献が期待できる遺伝子やタンパク質」の候補を設定して下さい。例えば、環境浄化、環境ホルモン分解、重金属除去、有機水銀分解、PCB 分解、土壌浄化、バイオエタノール、農薬分解、硝酸性窒素除去、有機リン除去、CO₂ 固定、フェノール除去、石油分解、等に能力を発揮できる遺伝子類やタンパク質類は、「持続可能型社会の実現に貢献する」可能性を持つと考えられます。以下の候補検索例を参考に、皆さんで独自性の高い考えを出して下さい。

設定した項目（例えば、バイオエタノール）について、その項目に該当すると考えられる遺伝子やタンパク質を、Google 日本 (<http://www.google.co.jp/>) の「日本語のページを検索」を用いて探して下さい。

例えば、「バイオエタノール」を検索項目として、それに関係する遺伝子やタンパク質を検索する場合、「バイオエタノール 遺伝子」や、「バイオエタノール タンパク質」等を Key Word（キーワード：検索文字）として検索すると良いでしょう。ここで、バイオエタノールと遺伝

子の間に空白を入れる方が良い(「」は検索の欄へは入れない)でしょう。「バイオエタノール」の代わりに、「環境ホルモン分解」等、皆さんが設定した対象で良いですが、「遺伝子」や「タンパク質」を入れたほうが目的に近い情報が得られ易いです。「PCB 分解」のように対象が限定されている場合には、空白なしの「PCB 分解遺伝子」の方が、検索は能率的になりますが、検索の範囲が限定され過ぎるおそれがあります。検索に用いる用語について、空白ありとなしの両方について、検索を試してみると良いでしょう。「PCB 分解」の場合では、「PCB」と「分解」の間についても、空白ありとなしで試みるのも良いでしょう。

Google で得られた情報は、目的とする「遺伝子」や「タンパク質」そのものに関する説明ではないことが多いので、複数の検索結果の文章を読み進めて、目的とする「遺伝子」や「タンパク質」の名前を探し出します。この操作で複数の遺伝子やタンパク質を探し出しておくほうが、興味深い遺伝子発見につながります。当然のことですが、複数の異なった Key Word で検索する方が、目的とする遺伝子やタンパク質を探し出す確率が高まります。この際、検索で見つかった遺伝子名やタンパク質名に少しでも差があれば、レポート(様式1)に記載しておきましょう。個々の遺伝子や検索結果の文章中に、遺伝子やタンパク質の英語名や、学術的な省略名(遺伝子名等)があれば、それもレポート(様式1)に記載しておきましょう。

一個の検索項目(例えば、「ダイオキシン分解」)でも、複数の遺伝子やタンパク質が検索できる可能性があります。一方、遺伝子やタンパク質名が見つからなかった検索項目も、レポートに残しておきましょう。興味深い項目は、後で Google の「日本語のページを検索」以外の手段で検索を試みる際に有用になります。

2.1 得られた「遺伝子」や「タンパク質」名の機能や特徴と英語名の調査

以降の解析では、日本で運用されている国際塩基配列データベース DDBJ を利用します。データベースの使用法や内容の説明等は日本語で記載されていますが、データベース自体は国際的に共同して作成されており、英語が使用されています。遺伝子名もタンパク質名も英語で記載された学術用語になっています。従って、DDBJ を使用する前に、得られた「遺伝子」や「タンパク質」名の英語名を知る必要があります。遺伝子やタンパク質の英語名を知る方法としては、インターネットを使って、ライフサイエンス辞書(<http://wwwsoc.nii.ac.jp/lbsdproject/ja/service/webbsd/index.html>)を使うのが便利です。それでも英語名が見つからない場合は、自分で英語名にしてみましょう。例えば、「PCB 分解遺伝子」の場合、Google の「和英 分解」の検索をすれば、「分解」に関係する英単語(動詞や名詞)が探せるので、「PCB degrading gene」のように、英語名として適切と思える名前を探しだせます。

「遺伝子」や「タンパク質」の英語名を知ることで、DDBJ を使用して遺伝子やタンパク質の配列を取得する用意が整いました。長浜バイオ大の学生実習の経験からすると、ここで得られた英語名が、本当に自分の目標にする「遺伝子」や「タンパク質」なのかを確認しておくことが重要です。得られた「遺伝子」や「タンパク質」の英語名を Key Word として、Google の「日本語のページを検索」をもう一度用いて、それらの英語名の「遺伝子」や「タンパク質」に関する情報を検索し、それらの機能や特徴を再確認して下さい。英語名に加えて、日本語名を用いた Key Word の検索を繰り返すことで遺伝子やタンパク質の機能や特徴を正しく理解でき、目標とする有用遺伝子に関する理解が深まります。英語に自信のある方は、Google の「ウェブ全体から検索」で世界中の情報から検索することで、より正確で最新の情報を得ることができます。

これらの検索の過程で得られた知識をレポートに記載しておくとい良いでしょう。また、このような操作で、同一の遺伝子やタンパク質が複数の名前を持つ例に出合うと思いますが、それらの英語名についても以降の解析に使用して下さい。

2.2. 調べた遺伝子が国際塩基配列データベースに登録されているかどうかの調査

得られた遺伝子やタンパク質の英語名を Key Word として、DDBJ の検索システムの「ARSA」を用いて、そのタンパク質遺伝子について、既にどのようなタンパク質のアミノ酸配列が知られているのかを検索してみましょう。Google の「日本語のページを検索」で「DDBJ」を検索し、DDBJ Homepage を開き、左側の欄にある「検索・解析」のなかの ARSA の項目をクリックします。

The image shows two screenshots of the DDBJ (DNA Data Bank of Japan) website. The top screenshot is the main homepage, and the bottom screenshot is the ARSA (All-round Retrieval of Sequence and Annotation) search interface.

Top Screenshot (DDBJ Homepage):

- The browser address bar shows <http://www.ddbj.nig.ac.jp/index-j.html>.
- The DDBJ logo and navigation menu are visible.
- On the left sidebar, under "塩基配列の登録" (Sequence Registration), the "ARSA" link is highlighted with a red arrow and the text "ARSA をクリックする。" (Click ARSA).
- The main content area features a banner for "DDBJ : DNA Data Bank of Japan" and a "Hot Topics" section with recent news.

Bottom Screenshot (ARSA Search Interface):

- The header reads "ARSA All-round Retrieval of Sequence and Annotation".
- An "お知らせ" (Notice) box contains information about system maintenance and data updates.
- The "Quick Search" section has a red arrow pointing to the "All Databases" dropdown menu.
- A red arrow points to the search input box, which contains the text "key word を入力する" (Enter keyword).
- Another red arrow points to the "Search" button, with the text "クリックする" (Click) next to it.
- The search input box contains the example keyword "PCB degrading".
- Below the search box, there is a note: "検索条件を複数指定する場合は、語句を半角空白で区切ってください。AND条件で結合します。" (When specifying multiple search conditions, please separate the phrases with half-width spaces. They will be combined with AND conditions).
- The "DDBJ Advanced Search" section offers three options: "Simple Query", "Standard Query", and "Extended Query", each with a brief description of its capabilities.
- The "Search History" section at the bottom includes a text input field, a "参照..." (Reference) button, and an "Upload" button.

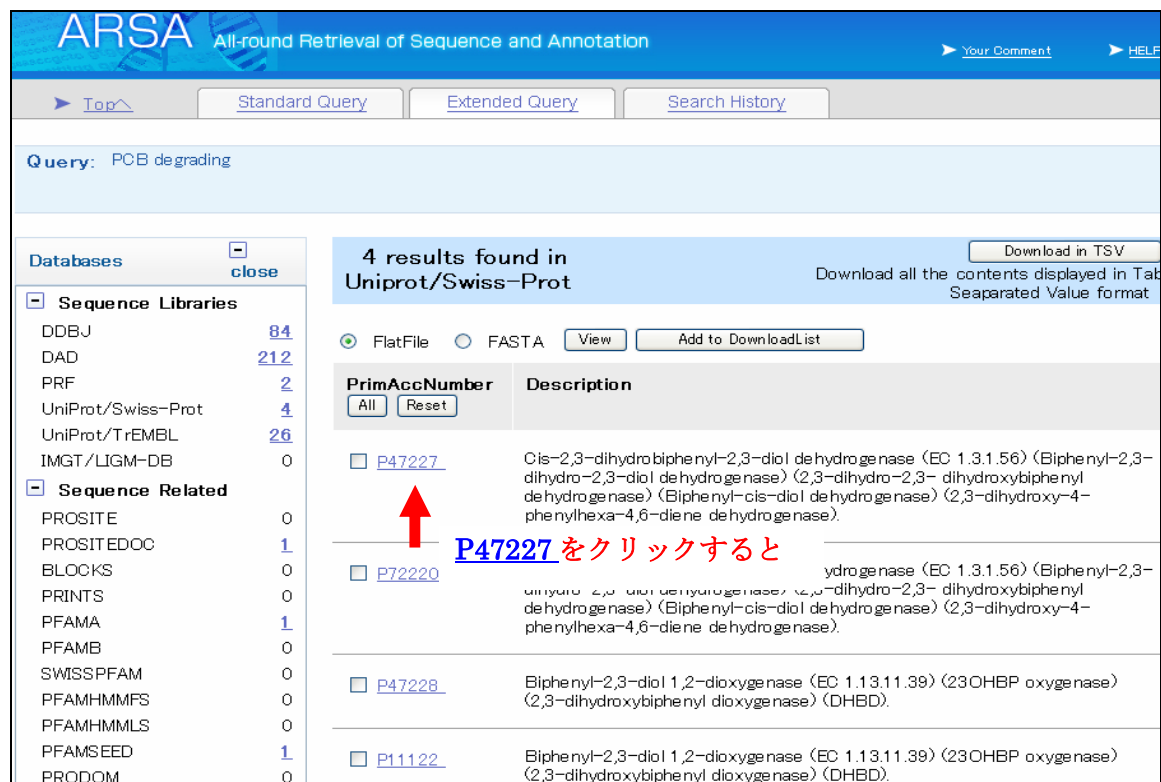
- 1) Quick SearchでAll Databases（赤矢印：DDBJ中に存在している23データベースの全体）を対象に、「検索条件を複数指定する場合は、語句を半角空白で区切ってください」との表示に従って、得られている「遺伝子やタンパク質の英語名」を半角文字でkey wordとして記入して、「Search」をクリックします。この際、ARSAは遺伝子やタンパク質を検索することを前提にしているので、検索の英語名において、「gene」や「protein」は入れない方がよいです。

waitの後に、空色の数字の0の部分が短時間の後に色々な数字に変化します。それがAll Databases（DDBJ中に存在している23データベースの全体）に収録されている大量なデータを対象にして、検索された結果の件数です。DDBJ（青矢印）ではDNA塩基配列の情報が得られますが、今回の解析では、タンパク質のアミノ酸配列を対象にします。Uniprot/Swiss-Prot（赤矢印）がそのタンパク質のアミノ酸配列を収録したデータベースを対象にした検索の結果です。この数字をクリックすると、ARSAの検索で見出された遺伝子のタンパク質のアミノ酸配列や機能や特徴が得られます。

ARSA All-round Retrieval of Sequence and Annotation			
Top Simple Query Standard Query Extended Query Search History			
Hit Counts List			
Query PCB degrading gene			
Sequence Libraries			
•DDBJ	83	•DAD	224
•PRF	2	•UniProt/Swiss-Prot	4
•UniProt/TrEMBL	12	•IMG/IMG-DB	0
Sequence Related			
•PROSITE	0	•PROSITEDOC	1
•BLOCKS	0	•PRINTS	1
•PFAMA	1	•PFAMB	0
•SWISSPFAM	0	•PFAMHMMFS	0
•PFAMHMMLS	0	•PFAMSEED	1
•PRODOM	0	•ENZYME	0
Protein 3D Structures			
•PDB	10	•HSSP	0
•FSSP	0		

次のような画面が表示されます。

なお、DDBJ 側の都合で、ARSA が使えない時があります。その際には、DDBJ の最初の画面の ARSA の上にある SRS を使用しても検索が可能です。その場合は、以下の画面等は少し違うものになります。



ARSA All-round Retrieval of Sequence and Annotation

Query: PCB degrading

Databases: Sequence Libraries

- DDBJ 84
- DAD 212
- PRF 2
- UniProt/Swiss-Prot 4
- UniProt/TrEMBL 26
- IMG/IMG-DB 0

Sequence Related

- PROSITE 0
- PROSITEDOC 1
- BLOCKS 0
- PRINTS 0
- PFAMA 1
- PFAMB 0
- SWISSPFAM 0
- PFAMHMMFS 0
- PFAMHMLS 0
- PFAMSEED 1
- PRODOM 0

4 results found in Uniprot/Swiss-Prot

Download in TSV

Download all the contents displayed in Table in Seperated Value format

FlatFile FASTA View Add to DownloadList

PrimAccNumber	Description
<input type="checkbox"/> P47227	Cis-2,3-dihydrobiphenyl-2,3-diol dehydrogenase (EC 1.3.1.56) (Biphenyl-2,3-dihydro-2,3-diol dehydrogenase) (2,3-dihydro-2,3-dihydroxybiphenyl dehydrogenase) (Biphenyl-cis-diol dehydrogenase) (2,3-dihydroxy-4-phenylhexa-4,6-diene dehydrogenase).
<input type="checkbox"/> P72220	hydrogenase (EC 1.3.1.56) (Biphenyl-2,3-dihydro-2,3-dihydroxybiphenyl dehydrogenase) (Biphenyl-cis-diol dehydrogenase) (2,3-dihydroxy-4-phenylhexa-4,6-diene dehydrogenase).
<input type="checkbox"/> P47228	Biphenyl-2,3-diol 1,2-dioxygenase (EC 1.13.11.39) (23OHBP oxygenase) (2,3-dihydroxybiphenyl dioxygenase) (DHBD).
<input type="checkbox"/> P11122	Biphenyl-2,3-diol 1,2-dioxygenase (EC 1.13.11.39) (23OHBP oxygenase) (2,3-dihydroxybiphenyl dioxygenase) (DHBD).

P47227 をクリックすると

別 Window で次の画面が表示されます。

http://getentry.ddbj.nig.ac.jp - DDBJ sequence search by accession number - Micro

Number = [P47227]

AC : アクセション番号

DE の行 : 遺伝子名

生物種名

GN の行 : 遺伝子の略号

アミノ酸配列

```

ID   BPHB_BURXI               Reviewed:      27
AC   P47227; Q13FT5; Q52435;
DT   01-FEB-1996, integrated into UniProtKB/Swiss-Prot.
DT   01-FEB-1996, sequence version 1.
DT   20-MAR-2007, entry version 54.
DE   Cis-2,3-dihydrobiphenyl-2,3-diol dehydrogenase (EC 1.3.1.56)
DE   (Biphenyl-2,3-dihydro-2,3-diol dehydrogenase) (2,3-dihydro-2,3-
DE   dihydroxybiphenyl dehydrogenase) (Biphenyl-cis-diol dehydrogenase)
DE   (2,3-dihydroxy-4-phenylhexa-4,6-diene dehydrogenase).
GN   Name=bphB; OrderedLocusNames=Bxeno_C1126; ORFNames=Bxe_C1192;
OS   Burkholderia xenovorans (strain LB400).
OC   Bacteria; Proteobacteria; Betaproteobacteria; Burkholderiales;
OC   Burkholderiaceae; Burkholderia.
OX   NCBI_TaxID=266265;

      (中 略)
FT   HELIX           238   241
FT   STRAND          248   254
FT   HELIX           255   257
FT   HELIX           269   273
SQ   SEQUENCE        277 AA;  28901 MW;  00194120BD4E12D1 CRC64;
      MKLKGEAVLI TGGASGLGRA LVDRFVAEGA KVAVLKSAE RLAELETDHG DNVLGIVGDV
      RSLEDQKQAA SRCVARFGKI DTLIPNAGIW DYSTALVDLP EESLDAFDE VFHINVKGVI
      HAVKACLPAL VASRGNVIFT ISNAGFYNG GGPLYTAAKH AIVGLVRELA FELAPYVRVN
      GVGSGGINS DLRGPSSLGMG SKAISTVPLA DMLKSVLPIG RMPEVEEYTG AYVFFATRGD
      AAPATGALLN YDGGLGVRGF FSGAGGNDLL EQLNIHP
//

```

「PCB degrading」で検索した結果から、Cis-2,3-dihydrobiphenyl-2,3-diol dehydrogenase (遺伝子の略号 *bphB*) という酵素が、国際塩基配列データベースに登録されていることが確認出来ました。キーワードによっては、目的の遺伝子類ではないものがヒットしている可能性もあります。検索で見出された配列情報(遺伝子名や機能に関する情報等)をよく確認してください。この酵素が PCB 分解に貢献しているという確信が得られたら、次にこのアミノ酸配列情報を出発点に、これらとの配列の類似性を指標に、多様な環境中で生息する生物類のゲノム解析で得られた膨大な遺伝子の候補セット(400 万件以上の遺伝子の候補が公的なデータベースに、機能未知として収録されている)の中から、類似性の高い配列を検索します。

ARSA の検索では、複数のタンパク質の配列が Uniprot/Swiss-Prot (タンパク質のアミノ酸配列を収録したデータベース) から得られるのが通例です(この例では 4 件)。複数のタンパク質の配列が得られる原因の一つは、着目する機能(この例では PCB 分解を意味する PCB degrading)に関与するタンパク質に関して、酵素活性を異にしており、アミノ酸配列も大きく(または全く)異なっているようなタンパク質の例が存在するからです。これらの異なったタンパク質類は、いずれも環境由来の遺伝子候補からの有用遺伝子の発掘で重要になります。

別の原因としては、同一機能のタンパク質でも、生物種によりタンパク質の配列が多

少しは異なっている場合です。多様な生物は、少しずつ遺伝子 DNA の配列を変化させながら進化してきました。従って同じ機能を持つタンパク質でも、異なった生物種のタンパク質は多少異なったアミノ酸配列を持つ場合が多く見られます。進化的に近い生物であればあるほど、その DNA 配列やアミノ酸配列は高い類似性を持っています。配列相同性検索 (homology search : 調べたい配列に対して類似性の高い配列を、データベース全体から検索する手法) により、着目配列に対して類似性の高い配列を、データベース全体から検索することが出来ます。但し、余り近い関係にあるアミノ酸配列を用いた場合、環境由来の DNA からの遺伝子発掘において、環境由来の同じ配列が見つかるだけです。効率的とは言えません。類似配列が多い場合には、その一部を選択するだけで充分です。ここでは、ARSA の結果を見て、酵素の種類を示す EC 番号の同じもの (この例では EC1.13.11.39) については、どれかを選ぶことで進んでください。ARSA の結果の記載中に fragmental の文字がある場合は、タンパク質の配列の一部しか Uniprot/Swiss-Prot に収録されていないので、その番号のタンパク質は以降の解析からは除外します。

詳細な作業手順は「次章 3. 環境由来 DNA 断片配列からの、持続可能型社会への貢献遺伝子の探索」でもう一度解説をしますので、ここまでは大まかな作業の流れを理解してもらえば結構です。

発見した遺伝子候補を、長浜バイオ大学のデータベースから皆さんの所属名・名前を記載して世界へ発信することも可能です。ご希望の場合は、レポート (様式 1 と様式 2 の 2 種類) に、必要事項を本テキストの解説に従ってご記入下さい。

- ・ 様式 1 は、「持続可能型社会の実現に、貢献が期待できる遺伝子やタンパク質」の候補選びと、それに関する既知遺伝子の情報に関するレポートです。
- ・ 様式 2 は、目的とする既知遺伝子と相同性が得られた環境由来 DNA 配列に関しての、遺伝子領域等に関するレポートです。誰もその遺伝子の存在に気付いていない場合には、環境由来 DNA 配列から新規に遺伝子を発掘したことになり、「持続可能型社会への貢献遺伝子」データベースへ収録し、公開する意味が生じます。

レポート (様式 1・様式 2) の必要な方は下記へお知らせ下さい。

送信先 : h_uehara@nagahama-i-bio.ac.jp

長浜バイオ大学 生体分子情報学研究室 遺伝子探索 係

※ 件名に (遺伝子探索) と明記願います。

3. 環境由来 DNA 断片配列からの、持続可能型社会への貢献遺伝子の探索の手順

3.1. 実習手順

今までは色々な事項を説明して来ましたので、ここで実習手順の概要をまとめておきます。

- ① 持続可能型社会の実現に、貢献が期待できる可能性を持つ遺伝子やタンパク質の候補を探します。
(使用サイト等：Google, ライフサイエンス辞書、NCBI PubMed, 専門誌や書籍)
- ② 得られた遺伝子やタンパク質の英語名を Key Word として、既知のアミノ酸配列を取得します。(使用サイト：DDBJ の ARSA)
※ ここまでで、レポートの様式 1 を作成するとよいでしょう。
※ 以降は新しく説明する操作手順です。
- ③ 取得したアミノ酸配列と環境由来 DNA 配列に存在するタンパク質候補との配列同一性検索を行う。環境由来 DNA 配列の中に目的の遺伝子やタンパク質と類似な配列が存在するかどうかを調べます。
(使用サイト：NCBI tblastn データベース設定 environmental.seq)
- ④ 同一性検索でヒット(発見)した環境由来 DNA 配列を取得します。
(使用サイト：NCBI)
- ⑤ 取得した DNA 配列について、遺伝子領域を確定しアミノ酸配列を取得します。
(使用サイト：NCBI ORF finder)
- ⑥ 確定した遺伝子領域のアミノ酸配列から、生物種由来を推定します。どの微生物種と近いかを調べ、既知微生物との類縁関係を見ます。取得した DNA 配列を持つ生物種そのものが推定できる訳ではないことに注意して下さい。
(使用サイト：NCBI protein blast (blastp))
- ⑦ レポート(所定様式有り。様式 1 と様式 2 の 2 種類。)を作成します。
(余力のある人は、以下についてもチャレンジし追加の情報を得るとよいでしょう。)
- ⑧ 取得したタンパク質の特徴や機能を調べます。(使用サイト：EBI InterProScan 等)
- ⑨ 代謝経路(Pathway)を調べます。(使用サイト：KEGG)

それぞれの操作で、いま何の作業をしているのか、目的や使用しているデータの形式等を、よく理解しながら進めるようにしましょう。参考までに、全ての操作手順の後に、次の解説等を付しておきます。

(付録 1)：DDBJ キーワード検索システム ARSA について

(付録 2)：国際塩基配列データベース(DDBJ/EMBL/GenBank)フラットファイルについて

(付録 3)：PCB 分解

(付録 4)：BLAST について

また、バイオ関連の用語についての検索は、次のサイトなどが便利です。

- jabion の用語辞書 <http://www.bioportal.jp/BioTerms/>
- ライフサイエンス辞書
<http://lsd.pharm.kyoto-u.ac.jp/ja/service/webisd/index.html>

3.2. キーワード検索から、既知のアミノ酸配列取得

それでは、実際の詳細な手順を、ARSA の使用法の復習を含めて、具体的な画面例を添えながら説明していきます。作業手順を説明するために例として、環境汚染物質である PCB（ポリ塩素化ビフェニル）の分解能を有する酵素遺伝子をキーワードにして、説明をしていくことにします。Google や文献の調査で、PCB を分解する酵素として *Rhodococcus* 属細菌に *bphA*（ビフェニル 2,3-ジオキシゲナーゼ）をはじめ、*bphB*, *bphC*, *bphD*, *bphE*, *bphF*, *bphG* の 7 種の異なった酵素が関係していることが、判明した場合を例に解説します（付録 3 参照）。皆さんの場合は、各自が設定した Key Words で、色々なタンパク質が見つかるはずです。

以下の例では、*bphF* をキーワードにして検索を開始した例を紹介します。

まずは、DDBJ のホームページを立ち上げましょう。Google 等の検索サイトで DDBJ を検索すればトップに出てきます。

URL を直接打ち込んでも構いません。<http://www.ddbj.nig.ac.jp/Welcome-j.html>

DDBJ のホームページが立ち上がったら、ARSA をクリックします。ARSA についての詳細は（付録 1）を参照するとよいでしょう。

アドレス http://www.ddbj.nig.ac.jp/index-j.html (画面 1)

Accession DNA Protein Taxonomy Site Search

http://www.ddbj.nig.ac.jp/Welcome-j.html

HOME 塩基配列の登録 検索・解析 FTP・WebAPI ドキュメント 統計情報 お問い合わせ English

DDBJ の紹介
利用の手引き
Q&A集

塩基配列の登録
SAKURA
大量データの登録(MSS)
データの修正・更新

検索
getentry
ARSA
SRS
TxSearch
BLAST
PSI-BLAST
FASTA
SSEARCH

ARSA をクリック

DDBJ : DNA Data Bank of Japan
DDBJは欧州の EBI/EMBL および
米国の NCBI/GenBank との密接な連携のもと
「DDBJ/EMBL/GenBank国際塩基配列データベース」を
構築している三大国際 DNA データバンクのひとつです。

Hot Topics 一覧へ

2007年10月12日 DAD リリース 41.0 公開
2007年10月3日 ARSA 機能改善(第2弾) 予定
2007年10月1日 DDBJ アノデータ募集
2007年10月5日 [今月の DDBJトピック] DDBJ HP デザイン一新
2007年8月1日 [重要]DDBJフラットファイルフォーマット改訂:E-mailアドレスと電話番号, FAX番号の非表示化

メンテナンス情報 一覧へ

2007年10月4日 国立遺伝学研究所ネットワークサービス・supernig の一時停止

次の画面で、① bphF と入力して、② Search アイコンをクリックします。

http://arsa.ddbj.nig.ac.jp/top-j.html (画面 2)

ARSA All-round Retrieval of Sequence and Annotation

DDBJ DNA Data Bank of Japan
Back to DDBJ
HomePage

お知らせ 2007年6月22日よりARSA機能エンハンス版の公開(6/20-22) ARSA サービスの休止 (2007年6月5日)

① bphF と入力

Quick Search

All Databases bphF Search

検索条件を複数指定する場合は、語句を半角空白で区切っていただきます。

DDBJ Advanced Search

Standard Query 検索する項目を指定でき、項目とキーワードで指定できます。

Extended Query Standard Search よりも項目が細かく分かれた条件を指定しての検索が可能です。

②クリック

ARSA の Quick Search で All Databases を選択していますので (赤の矢印)、ARSA 検索できる 23 種類の国際的な配列データベースの全体に対して、一括で検索を実施した結果画面が表示されます。以降の解析では、画面 3 で表示されている UniProt/Swiss-Prot (世界的に評価の高いタンパク質のアミノ酸配列のデータベース) のデータを使用することになります。

UniProt/Swiss-Prot の検索ヒット件数 [11](#) をクリックします。

ARSA All-round Retrieval of Sequence and Annotation (画面 3)

Top Standard Query Extended Query Search History

Hit Counts List

Query bphF

Sequence Libraries			
•DDBJ	20	•DAD	22
•PRF	20	•UniProt/Swiss-Prot	11
•UniProt/TrEMBL	10	•IMG/TrEMBL	0
Sequence Related			
•PROSITE	2	•PROSITEDOC	0
•BLOCKS	0	•PRINTS	0
•PFAMA	0	•PFAMB	0
•SWISSPFAM	4	•PFAMHMMFS	0

クリックする

ヒットした UniProt/Swiss-Prot のアクセッションナンバー (登録・整理番号) 等が表示されます。UniProt/Swiss-Prot データベースは更新が続いていますので、画面 3 とは異なった結果になり、11 以上のヒット件数が表示される可能性があります。ヒットしたリストを順次参照してゆきます。ここでは、[005151](#) の 4-hydroxy-2-oxovalerate aldolase を参照する場合での作業手順の説明を行うことにします。画面 4 の [005151](#) をクリックすると別 Window (画面 5) が開き、生物種やアミノ酸配列情報が得られます。アミノ酸配列情報をコピーしておきます。

005151 を例に説明した操作で、画面 4 の各ヒット項目については逐次クリックをして生物種やアミノ酸配列情報を得ておきましょう。その際に、画面 4 の記載事項として、(partial)の文字がある項目については、そのタンパク質の部分的なアミノ酸配列しか UniProt/Swiss-Prot データベースに登録されていないので、参照しなくても良いです。

なお、DDBJ 側の都合で、ARSA が使えない時があります。その際には、画面 1 の ARSA の上にある SRS を使用しても検索は可能です。その場合は、以下の画面等は少し違うものになります。

Query: bphF (画面 4)

Databases close

☒ Sequence Libraries

- DDBJ [20](#)
- DAD [22](#)
- PRF [20](#)
- UniProt/Swiss-Prot [11](#)
- UniProt/TrEMBL [10](#)
- IMG/CLIM-DB 0

☒ Sequence Related

- PROSITE [2](#)
- PROSITE-DOC 0
- BLOCKS 0
- PRINTS 0
- PFAMA 0
- PFAMB 0
- SWISS-PROT [4](#)
- PFAM-HMMFS 0
- PFAM-HMMLS 0
- PFAM-SEED 0
- PRODOM [6](#)
- ENZYME 0

☒ Protein 3D Structures

11 results found in UniProt/Swiss-Prot Download all the c

☒ FlatFile ☐ FASTA View Add to DownloadList

PrimAccNumber	Description
P37333	Biphenyl dioxygenase subunit alpha (EC 1.14.12.18) (Biphenyl 2,3- dioxygenase).
Q46372	Biphenyl dioxygenase subunit alpha (EC 1.14.12.18) (Biphenyl 2,3- dioxygenase).
Q52028	Biphenyl dioxygenase subunit alpha (EC 1.14.12.18) (Biphenyl 2,3- dioxygenase).
P37334	Biphenyl dioxygenase subunit beta (EC 1.14.12.18) (Biphenyl 2,3- dioxygenase).
Q46373	Biphenyl dioxygenase subunit beta (EC 1.14.12.18) (Biphenyl 2,3- dioxygenase).
P37332	Biphenyl dioxygenase system ferredoxin subunit.
P51014	4-hydroxy-2-oxovalerate aldolase (EC 4.1.3.-) (HOA).
005151	4-hydroxy-2-oxovalerate aldolase (EC 4.1.2.-).
O86013	4-hydroxy-2-oxovalerate aldolase (EC 4.1.2.-).

クリックすると、別 windows (画面 5) が表示され、その画面の下の方に配列が表示される。

別 Window で開かれた画面 (画面 5)。生物種等を確認し、アミノ酸配列情報をコピーします。
AC の行の番号、記号の全体をレポート (様式 1) に記入しておきます。

http://getentry.ddbj.nig.ac.jp - DDBJ sequence search by accession number - Microsoft Int... (画面 5)

Number = [005151]

ID BPHE_RHOSR Reviewed: 258 AA

AC 005151; Q75WN6; ← AC 行の アクセション番号をレポート (様式 1) に記入する。

DT 01-DEC-2000, integrated into UniProtKB/Swiss-Prot.

DT 01-JUL-1997, sequence version 1.

DT 23-JAN-2007, entry version 31.

DE 4-hydroxy-2-oxovalerate aldolase (EC 4.1.2.-) ← 遺伝子名 : レポート (様式 1) に記入する。

GN Name=bphF; Synonyms=etbF; OrderedLocusNames=RHA1_ro10138;

OS Rhodococcus sp. (strain RHA1). ← OS 行 : 生物種名が記載されている。

OG Plasmid pRHL2.

OC Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;

OC Corynebacterineae; Nocardaceae; Rhodococcus.

OX NCBI_TaxID=101510;

RN [1]

← 遺伝子略号: 複数行になることもある GN の行の内容を全て、
(中 略) レポート (様式 1) に記入する。

RL Submitted (JUL-2006) to the EMBL/GenBank/DDBJ databases.

CC -!- FUNCTION: Catalyzes the production of pyruvate and acetaldehyde
from 4-hydroxy-2-oxovalerate.

CC -!- PATHWAY: Xenobiotic degradation; biphenyl degradation.

CC -!- INDUCTION: By growth on ethylbenzene or biphenyl.

CC -!- SIMILARITY: Belongs to the hpcH/hpaI aldolase family.

CC -----

CC Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms

CC Distributed under the Creative Commons Attribution-NoDerivs License

CC -----

DR EMBL: D78322; BAA18937.1; -; Genomic_DNA.

DR EMBL: AB120955; BAC92717.1; -; Genomic_DNA.

DR EMBL: CP000433; ABH00331.1; -; Genomic_DNA.

DR PIR: JC6327; JC6327.

DR HSSP: P23522; 1DXE.

DR GenomeReviews; CP000433_GR; RHA1_ro10138.

DR KEGG: rha:RHA1_ro10138; -.

DR InterPro; IPR005000; HpcH_HpaI.

DR Pfam; PF03328; HpcH_HpaI; 1.

KW Aromatic hydrocarbons catabolism; Complete proteome; Lyase; Plasmid.

FT CHAIN 1 258 4-hydroxy-2-oxovalerate aldolase.

FT /FTId=PRO_0000207097.

SQ SEQUENCE 258 AA; 27159 MW; A539C46C4DA0190F CRC64;

MQSPINSFKK ALAEGRTQIG FWLALGDAYS AEVCAGAGFD WLLIDGEGHAP QDLRSVLAQL
QVIGAYRDCH AAVRVPSADT TVIKQYLDLG AQSLLVPMVD TADEAAAVVR ACRYPPGGIR
GVGGARASRW GRYPRLHEA DEQVCVVQA ETALALSNI AIAEVDGIDG VFIGTADLAA
SLGFPGNPAH PEVQDAILDA LQVRRAAGKA PGVLTPEDL AQKYLHGAV FVAVGIDTHL
LAKQTSALAA RFAQVAYS

//

↑ アミノ酸配列部分をコピーする。

(前の画面で FASTA 形式を選択し保存してもよい)

画面 5 のアミノ酸配列部分のコピー操作が分からない方もいると思います。Windows を使用の方は、赤色の線で囲まれたアミノ酸配列を含む部分をドラッグした後に、マウスの右クリックをして、コピーを選び、レポート (様式 1) の画面の所定に位置へカーソルを移してから、右クリックして貼り付けを選びます。右クリックをせずに、アミノ酸配列を含む部分をドラッグした後に、コントロールキー (Ctrl キー) を押しながら c を押して (コ

ピー操作に対応)、レポート(様式 1)の画面の所定に位置へカーソルを移してから、Ctrl キーを押しながら v を押すと貼り付けの操作が完了します。

ここまでで、自分が設定したテーマに関する遺伝子類について、すでに知られている遺伝子類のアミノ酸配列を得ることが出来ました。これらのタンパク質については、既に知られていたものであり、あなたの発見とはいえません。次に、得られたこの既知のタンパク質のアミノ酸配列を出発点に、これらとの配列の類似性を指標に、下記の 3.3 で紹介する方法で、多様な環境中で生息する環境由来の膨大な DNA 配列の中から、同一または類似な機能を持つ遺伝子を網羅的に探索していきます。その DNA 配列にタンパク質遺伝子の記載がされていなければ、それを「持続可能型社会への貢献遺伝子データベース」へ登録して、世界へ発信する意味が出ます。

3.3. 取得した既知のアミノ酸配列と環境由来 DNA 配列との相同性検索

UniProt/Swiss-Prot データベースから得られた、既知のアミノ酸配列を用いて、環境由来データベースを対象に配列相同性検索を行います。


NCBI の BLAST (tblastn) を用い、データベースの設定を environmental samples (環境由来) にすると、膨大な環境由来 DNA 配列からの網羅的な検索が可能となります (BLAST に関しては、付録 4 を参照して下さい)。environmental samples (環境由来サンプル) のデータベースには Sargasso 海由来の配列数が大量に収録されているため、検索結果も Marine (海洋) 由来のものが大半を占める傾向がありますが、いくつかは異なる環境由来のものも含まれているでしょう。どんな環境由来の配列から取得できるかにも注目しながら探索するとよいでしょう。

- 3.3.1. NCBI BLAST を立ち上げます。<http://www.ncbi.nlm.nih.gov/BLAST/> (NCBI のトップ画面のメニューバーにある「BLAST」コマンドをクリックしても構いません。) 次に tblastn をクリックします。

Basic BLAST

(画面 6)

Choose a BLAST program to run.

<u>nucleotide blast</u>	Search a nucleotide database using a nucleotide query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
<u>protein blast</u>	Search protein database using a protein query <i>Algorithms: blastp, psi-blast, phi-blast</i>
<u>blastx</u>	Search protein database using a translated nucleotide query
 <u>tblastn</u>	Search translated nucleotide database using a protein query
<u>tblastx</u>	Search translated nucleotide database using a translated nucleotide query

クリックする

BLAST Basic Local Alignment Search Tool (画面 7)

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ tblastn: TBLASTN search translated nucleotide database using a protein query. more... Reset

Enter Query Sequence

Enter accession number, gi, or FASTA sequence [Clear](#) [Query subrange](#)

MQSPINSFKK ALAEGRTQIG FWLALGDAYS AEVCAGAGFD WLLIDGEHAP QDLRSVLAQL
QVIGAYRDCH AAVRVPSADT TVIKQYLDLG AQSLLVPMVD TADEAAVVR ACRYPPGGIR
GVGGARASRW GRYPRYLHEA DEQVCVVQA ETALALSNE AIAEVDGIDG VFIGTADLAA
SLGFFGNPAH PEVQDAILDA LQVRAAGKA PGVLTFVEDL AQKYLAHGAV FVAVGIDTHL
LAKQTSALAA RFAQVAYS

Or, upload file [参照](#)

Job Title

Enter a description

Choose Search Set

Database

Organism [Optional](#)

Entrez Query [Optional](#)

[クリック](#)

BLAST

[Algorithm parameters](#)

[Note: Parameter](#)

(画面 5)で取得したアミノ酸配列を貼り付ける

(env_nt)を選択


検索に時間のかかる(2~3分)場合もありますが、画面8のような検索結果が表示されます。画面8のQuery(クエリ)とは、データベースへの問い合わせに使用した文字列のことであり、ここでは大きな括弧内へ貼り付けた既知のアミノ酸配列に対応する。そのクエリー配列が258アミノ酸よりなることが、赤丸の中にLength=258として記載されている。0, 50, ..., 200, 250の数字の上の赤の太い横線が、258アミノ酸よりなるクエリー配列を示す。数字の下に表示された複数本の赤の横線(この場合は21本)は、環境由来DNA中に見つかった、クエリー配列と配列の類似性の高いタンパク質配列の長さを表示している。いずれの配列の長さもクエリー配列の全長に近いが、数本は明らかに先頭(クエリー配列の0番に近い領域)ないしは末端(クエリー配列の250番に近い領域)が短くなっている。このような部分では、配列の類似性が低くなっているか、環境由来DNA中に得られているタンパク質配列が不完全で、遺伝子の先頭や末端がクローン化された遺伝子で欠損している可能性が考えられる。もっと赤横棒が短いことも多い。

画面8の例では、[AACY023301218.1]を含む多くの配列で、クエリーに対してほぼ全長をカバーしているので、これらは *bphF* 遺伝子 を持っている可能性が高いと考えられます。

まずは、[AACY023301218.1]の Score 307 をクリックします。

作業手順を説明する為に、[AACY023301218.1]だけを例として、次に紹介していますが、皆さんは目的とする遺伝子類候補の可能性のあるもの(E-valueが、 $1e-10$ 以下のもの)について、確認を行って下さい。E-valueは、偶然な現象として二つの配列がたまたま類似性を持つ確率を表示している。 e^{-10} は自然対数表示で通常は e^{-10} と表示する。 $-$ 以降の数字が大きいくほど、数字としては小さい。 e^{-82} や e^{-78} は e^{-10} 以下の数であり、偶然に類似する確率

が極めて低く、生物学的に意味のある類似性であることを示します。

 **BLAST** *Basic Local Alignment Search Tool*

[Home](#) [Recent Results](#) [Saved Strategies](#) [Help](#)

(画面 8)

▶ [NCBI/ BLAST/ tblastn/ Formatting Results - 9SJP8ATN012](#)

[\[Reformat these Results\]](#) [\[Edit\]](#)

Job Title: Protein sequence(258 letters)

TBLASTN 2.2.17 (Jun-24-2007)

Reference:

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, 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.

RID: 9SJP8ATN012

Database: environmental samples
5,006,910 sequences; 4,930,790,886 total letters

If you have any problems or questions with the results of this search please refer to the [BLAST FAQs](#)

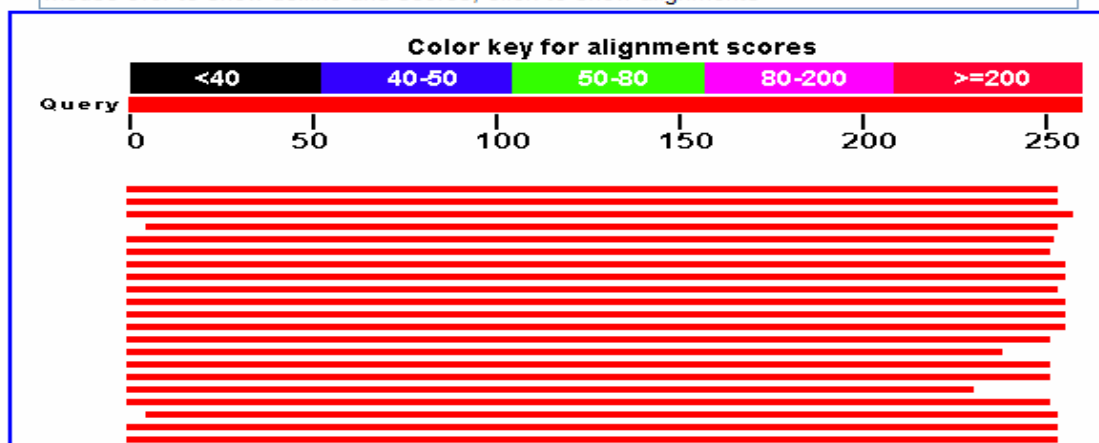
[Taxonomy reports](#)

Query=
Length=258

クエリー配列全長：カバー率を求める計算で使う。

Distribution of 100 Blast Hits on the Query Sequence

Mouse-over to show define and scores, click to show alignments



アクセッション番号 レポート様式2に記入

Sequences producing significant alignments:

クリックする（候補になりうるもの全てについて確認を行う）

			Score (Bits)	E Value
gb AACY023301218.1 	Marine metagenome	ctg_1101668108569, whol...	307	3e-82
gb AACY020561156.1 	Marine metagenome	1096626843456, whole genom	293	6e-78
gb AACY020347673.1 	Marine metagenome	1096626444319, whole genom	278	2e-73
gb AACY023307893.1 	Marine metagenome	ctg_1101668115244, whol...	272	1e-71
gb AATN01007600.1 	Metagenome sequence	ctg5152, whole genome sho	250	5e-65
gb AACY020561156.1 	Marine metagenome	1096626732855, whole genom	240	5e-62
gb AACY020288491.1 	Marine metagenome	1096626776326, whole genom	239	1e-61
gb AACY023894137.1 	Marine metagenome	ctg_1101668701488, whol...	238	2e-61
gb AACY020297008.1 	Marine metagenome	1096626373724, whole genom	238	3e-61
gb AACY020543920.1 	Marine metagenome	1096626800317, whole genom	237	4e-61
gb AACY023519705.1 	Marine metagenome	ctg_1101668327056, whol...	237	4e-61
gb AACY020058757.1 	Marine metagenome	1096626065840, whole genom	236	7e-61
gb AACY021030769.1 	Marine metagenome	1092351262426, whole genom	235	2e-60

次にヒットしている領域を確認します。次の画面で、**Frame=-2**、**DNA 配列位置 90-857** でヒットしていることがわかります。後の Orf Finder で領域を選択する際に重要な情報となるので、レポート様式 2 に控えておきます。この 90 と 857 の数字はアミノ酸に翻訳する前の DNA 配列における位置を示す数字です。

レポート(様式 2)に記入する。遺伝子候補になるか否かの判断に使う。

```

>gb|AACY023301218.1| Marine metagenome ctg_1101668108569, whole ge (画面 9)
Length=1326
Score = 307 bits (787), Expect = 3e-82
Identities = 153/256 (59%), Positives = 193/256 (75%), Gaps = 4/256 (1%)
Frame = -2

Query 1  MQSPINSFKKALAEGRQTQIGFWLALGDAYSAEVCAGAGFDWLLIDGEHAPQDLRSVLAQL 60
          M++P+N FKKA+ E QIG W+ L +YSAE+CA AGFDWLLIDGEHAP +L+++ QL
Sbjct 857 MKTPVNRFKKAITEKHAQIGLWMGLASSYSAEMCAMAGFDWLLIDGEHAPNNLQTIQQQL 678

Query 61  QVIGAYRDCHAAVRVPSADTTVIKQYLDLGAQSLVPMVDTADEAAAVRACRYPP---- 116
          Q I AY +A RVP DT +IKQYLDLGA++LL+PM+DT +A V+A RYP
Sbjct 677 QSIAAYFVNNALARVPVGD TALIKQYLDLGAETLLIPMIDTPQQAECQVCQAMRYPQNDGK 498

Query 117 GGIRGVGGARASRWGFYPRYLHEADEQVCVVVQAETALALSNLEAIAEVDGIDGVFIGTA 176
          GG+RG+GGARASRWG +E Y +EA+EQVC++VQAET AL NL+AIA G+DGVFIG A
Sbjct 497 GGVRGMGGARASRWGMFENYTNEANEQVCLLVQAETREALKNLDAIANTPGVDGVFIGPA 318

Query 177 DLAASLGFGPNPAHPEVQDAITDALQVRRAAGKAPGVLTPVEDLAQKYLAHGAVFVAVGI 236
          DL+ASLG G+P HPEVQ AI DA+ R+ AGKAPG+LT A+ YL+ GA+FVAVG+
Sbjct 317 DLSASLGHVGDPNHPEVQAAIEDAIAIRILKAGKAPGILTSDVTQAKHYLSLGALFVAVGL 138

Query 237 DTHLLAKQTSALAARF 252
          DT +L +QTSAL +F
Sbjct 137 DTQILVRQTSALVRQF 90
  
```

遺伝子候補とする条件を、今回は次のようにします。

- ① E-value: $1e-10$ 以下
 - ② Identities (アミノ酸配列の一致率) : 30%以上
 - ③ クエリー配列(この例では画面 8 に記載されている 258 塩基)の内で、環境由来 DNA のタンパク質と相同性のあった領域 (ヒットしている領域 : 画面 9 の Positives=の分母の数でこの例では 256) がカバーする割合(この例でのカバー率は $256/258$) : 30%以上
- 全てを満たすものについて、遺伝子候補と判断します。

(画面 9)に戻り、着目の遺伝子候補としての条件を満たしているか確認しておきましょう。

- ① E-value= $3e-82$ で条件を満たす。
 - ② Identities (一致率) =59% で条件を満たす。
 - ③ カバー率=ヒットしている領域のクエリー配列の長さ/クエリー配列の全長
で求められる。 この場合 $256/258 = 0.99$ (99%) で条件を満たす。
- 全ての条件を満たしているので、目的とする遺伝子の候補と判断しましょう。

これで、bphF をキーワードにして得た既知のアミノ酸配列と、相同性の高い環境由来の遺伝子候補が検索出来ました。遺伝子候補が得られた環境由来 DNA 配列 [AACY023301218.1]の全長に着目して、遺伝子領域の確定を、特に遺伝子の開始と終止点の特定をしましょう。

3.4. 遺伝子領域の確定

環境由来ゲノムから検索した、相同性の高い遺伝子候補について、自分が新たな発見者であることの確認と、更に詳しく遺伝子領域の確定をしていきましょう。

3.4.1. 環境由来等の確認

先程の（画面 9）で、「[gb|AACY023301218.1](#)」をクリックすると、塩基配列のほか、塩基配列を決めた登録者、関連文献、生物種や機能情報など必要な情報が全て記載されているデータファイルが表示されます。配列の取得情報も記載されているので、どんな環境由来かを確認し、レポート様式 2 の環境由来の欄に記入しておきましょう。

環境由来 DNA 配列は、国際 DNA データベースでは、大半が機能未知な遺伝子候補の DNA 配列データとして残されています。皆さんが見つけた環境由来メタゲノムからの遺伝子候補は、大半はこれまで機能未知だったものである可能性が高いですが、中には既に機能の注釈付けが行われている場合も考えられます。自分がその DNA 配列の機能に関しての発見者（国際 DNA データベースで判断する限りは）であることも、この画面の中で確認しておきましょう。具体的には、データファイルの COMMENT 欄や、FEATURES の CDS の/product 欄（遺伝子の機能についての記載欄）に、自分の求めている機能が既に記載されていないかを確認しておきましょう。

本候補では、データファイルの COMMENT 行に、*bphF* に関連する記載は無く、また FEATURES の” CDS の/product 欄” 自体がありませんので、少なくとも公的データベースを参照する限りでは、自分が発見者であると言えるでしょう。「持続可能型社会への貢献遺伝子データベース」へ登録する意味が出てきました。

```
> gb|AACY023301218.1 Marine metagenome ctg_1101668108569, whole (画面 9)
Length=1326

Score = 307 bits (787) Expect = 3e-82
Identities = 151 / 193 (78%) Mismatches = 42 / 193 (22%) Gaps = 4 / 256 (1%)
Frame = -2

Query 1      MQSPINSFKKALAEGRQTQIGFWLALGDAYSAEVCAGAGFDWLLIDGEHAPQDLRSVLAQL 60
             M++P+N FKKA+ E   QIG W+ L  +YSAE+CA AGFDWLLIDGEHAP +L+++ QL
Sbjct 857     MKTPVNRFRFKKAITEKHAQIGLWMGLASSYSAEMCAMAGFDWLLIDGEHAPNNLQTIQQQL 678

Query 61      QVIGAYRDCHAAVRVPSADTTVIKQYLDLGAQSLVPMVDTADEAAAVVRACRYPP---- 116
             Q I AY  +A  RVP  DT +IKQYLDLGA++LL+PM+DT  +A   V+A RYP
Sbjct 677     QSIAAYPVNNAIARVPVGDALIKQYLDLGAETLLIPMIDTPQQAEQCVQAMRYPQNDGK 498

Query 117     GGIRGVGGARASRWGRYPYVLEHAEQVGVVVOAETALALSNLEATAEVDGIDGVFEGTA 176
```

NCBI Nucleotide

PubMed Nucleotide Protein Genome

Search Nucleotide for [] Go Clear

Limits Preview/Index History

Display FASTA Show 5 Send to Hide: ☐ sequence ☐ all but gene, CDS and

Range: from begin to end ☐ Reverse complemented strand

(画面 10)

1: [AACY023301218](#). Reports Marine metagenome...[gi:130393036]

[Comment](#) [Features](#) [Sequence](#)

LOCUS AACY023301218 1326 bp DNA linear UNA 13-MAR-2007
 DEFINITION Marine metagenome ctg_1101668108569, whole genome shotgun sequence.
 ACCESSION AACY023301218 [AACY020000000](#)
 VERSION AACY023301218.1 GI:130393036
 PROJECT GenomeProject:[13694](#)
 KEYWORDS WGS.
 SOURCE marine metagenome
 ORGANISM [marine metagenome](#)
 unclassified sequences; metagenomes; ecological metagenomes.
 REFERENCE 1 (bases 1 to 1326)
 AUTHORS Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K.E., Nelson,W., Fouts,D.E., Levy,S., Knap,A.H., Lomas,M.W., Nealson,K., White,O., Peterson,J., Hoffman,J., Parsons,R., E and Smith,H.O.
 TITLE Environmental genome shot
 JOURNAL Science 304 (5667), 66-74
 PUBMED [15001713](#)

既に機能の注釈がされていないかの確認：
 COMENT 行や、FEATURES の CDS の/product
 欄(遺伝子の機能についての記載)に、既に機能が
 記載されていないかを、確認しておく。

JOURNAL Submitted (02-MAR-2007)
 Center Drive, Rockville, MD 20850
 COMMENT For complete environmental metadata relating to this record,
 background on the Global Ocean Sampling expedition, as well as
 additional analysis results, please visit the CAMERA website
 (<http://camera.calit2.net>).

FEATURES Location/Qualifiers
 source 1..1326

/organism="marine metagenome"
 /mol_type="genomic DNA"
 /isolation_source="isolated as part of a large dataset
 composed predominantly from surface water marine samples
 collected along a voyage from Eastern North American coast
 to the Eastern Pacific Ocean, including locations in the
 Sargasso Sea, Panama Canal, and the Galapagos Islands"
 /db_xref="taxon:408172"
 /environmental_sample
 /note="metagenomic"

レポート様式 2 の

環境由来の欄に記入
 する。(コピー&ペース
 ト)

ORIGIN
 1 ttgcagcgaa cgcgcaacag gcagttgatg tgtcagcatg a
 61 gttggcaacg caatgccagc atctggcttg aattgacgca c
 121 accaaaattt gagtgtccag cccaccgcc acaaacaatg t
 181 ttggcttggt tcacatctga cgtaaaatg cctggtgctt taccagcttt caaatgcgt

次に、タンパク質配列を指定する領域を知る目的で、ORF Finder にかけるための DNA 配列を取得します。先程の画面 10 の Display を、下記の画面 10 のように FASTA に変更すると、画面 11 のように計算機処理しやすい FASTA 形式の DNA 配列のみが表示されます。

NCBI (画面 10)

PubMed Nucleotide Protein

Search for

Limits Preview

Display **FASTA** (selected)
 GenBank
 GenBank(Full)
 INSDSeq XML
 FASTA
 TinySeq XML
 ASN.1
 XML
 Graph
 GI List
 Brief
 Summary

Range to

☐ 1: [AACY023301218](#) Reports Marine metagenome...[gi:130393036]

Comr atures Sequence

LOCUS AACY023301218 1326 bp
 DEFINITION Marine metagenome ctg_1101668108569
 ACCESSION AACY023301218 [AACY02000000](#)
 VERSION AACY023301218.1 GI:130393036
 PROJECT GenomeProject:13694
 KEYWORDS WGS.

Display **FASTA** Show Send to

Range: from to ☐ Reverse complemented

☐ 1: [AACY023301218](#) Reports Marine metagenome...[gi:130393036]

```
>gi|130393036|gb|AACY023301218.1| Marine metagenome ctg_1101668108569, v
TTGCAGCGAACGCGCAACAGGCAGTTGATGTGTCAGCATGATCAATACACCTGTCTGAAGTTGGCAACG
CAATGCCAGCATCTGGCTTGAATTGACGCACCAATGCCGAAGTTTGACGCACCAAAATTTGAGTGTCCAG
CCCCACCGCCACAAACAATGCGCCGAGCGACAAATAGTGTGTTGGCTTGGGTCACATCTGACGTCAAATG
CCTGGTGCTTTACCAGCTTTCAAATGCGTGCAATAGCGTCTTCAATAGCGGCTTGCACCTCAGGGTGGT
TGGGGTCGCCCACATGACCCAATGAAGCCGACAAATCTGCAGGGCCAATGAACACACCATCCACGCCAGG
TGTGTTGGCAATAGCGTCTAAGTTTTTCAAGGCTTCACGCGTTTCGGCTTGACCAACAAGCAAACCTTGT
TCATTGGCTTCGTTGGTGTAATTGGGAAACATGCCCCAGCGCGAAGCGCGCACCGCCCATACCGCGCA
CACCGCCCTTACCATCGTTTTGGGGGTAGCGCATGGCTTGACACATTGCTCAGCTTGCTGGGGTGTGTC
AATCATCGGAATGAGCAAGGTCTCGGCGCCCAATCCAAATACTGCTTGATGAGCGCAGTGTGCGCCACG
GGAACGCGCGCAATTGCGTTGTTGACCGGGTAGGCCGCAATGCTTTGAAGTTGTTGCTGAATAGTTTGCA
AGTTATTGGGCGCATGCTCGCCATCAATTAACAACCAAGTCGAAGCCAGCCATGGCGCACATCTCTGCGCT
GTAAGTGTGCGGAGGCCCATCCACAAACCAATTTGGGCGTGTGTTTTCAGTGATGGCTTTTTTAAAGCGG
TTCACGGGTGTTTTATCATGTCTCTTTCTTAAATGAATTGAAATTGCAAACAGCCAAGTGGGCCGTAGT
CGGCATCGAACACATCGTCCACTTTGGCTGGAAGTGGTTTAGTAAAGGAGCCCGCCAAACACAATTTGTCC
GGCTTTCAAATGCTCGCCCCATGGGGCCAACTTGTGGCAAGCCATGCAATGCCACCGCAGGATGGCCT
TGAACACCTGCCGCCAAACCGTCTCTTCGACCACGCCATTCAATTTCAAATAGCGCCGCACCATGGCA
AATTGGTTGTGATTGGATTGACTTTTTCTGCGCCACCAAGAATGCCTGCGTTGGCTGCGTTGTGCTGAT
GGTGTCAAACACTTTGCGCATCACTTTGGTGTGGCGGTCAAATTGCTCAATGCGCGAATCGATGATTTCA
ATCGCAGGCGTAACATAGTCAGTAGCTGCCAACACTTGATCGACCGTGACGTCTGGCCCTTGCAGG
```

この DNA 配列をコピー&ペーストして、次の NCBI ORF finder にて遺伝子領域の確定を行います。

3.4.2. 遺伝子領域の確定

遺伝子予測の基本的な手法として、ORF の抽出作業があります。ゲノム配列中で、mRNA に転写される部分が遺伝子となりますが、その内でタンパク質配列を指定する領域は長い ORF (Open Reading Frame ; 終止コドンが出てこないコドンの読み枠) として特定できます。ORF Finder を使えば、先程皆さんが取得した DNA 配列から簡単に、ORF を抽出してくれます。実際に行ってみましょう。

NCBI の ORF Finder を立ち上げます。<http://www.ncbi.nlm.nih.gov/projects/gorf/> (NCBI トップページ上の右の [Hot Spots] の下のほうにある [ORF Finder] をクリックしても構いません。)

先程 (画面 11) の DNA 配列のコピーをテキストボックスに貼り付け、[OrfFind] アイコンをクリックします。

アドレス http://www.ncbi.nlm.nih.gov/gorf/gorf.html (画面 12)

NCBI ORF Finder (Open Reading Frame Finder)

PubMed Entrez BLAST

NCBI Tools for data mining

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool already in the database. This tool identifies all open reading frames using the standard or alternative sequence database using the WWW. ②クリックする ORF Finder Sequin sequence submission software.

Enter GI or ACCESSION OrfFind Clear

① 先程(画面 11)の DNA 配列を貼り付ける

submission support and software FTP site download data and software

or sequence in FASTA format

```
>gi|130393036|gb|AACY023301218.1| Marine metagenome  
ctg_1101668108569, whole genome shotgun sequence  
TTGCAGCGAACGCGCAACAGGCAGTTGATGTGTCAGCATGATCAATACACCTGT  
CCTGAAGTTGGCAACG  
CAATGCCAGCATCTGGCTTGAATTGACGCACCAATGCCGAAGTTTGACGCACCA  
AAATTTGAGTGTCAG  
CCCCACCGCCACAAACAATGCGCCGAGCGACAAATAGTGTGTTGGCTTGGGTCAC  
ATCTGACGTCAAAATG
```

FROM: TO:

次の画面 13 が表示されます。

ゲノム DNA は 2 本鎖よりなっていますが、その一方だけがデータベースに収録されています。着目する環境由来のタンパク質の配列を指定する DNA 配列 (コーディング配列) が、データベースに収録されていない側 (マイナスストランド側: - strand 側) の場合は、マイナス frame の ORF と呼びます。全てのコドンの読み枠を対象にした 6 frame 分について、開始 (start) と終止 (stop) コドンが取れる領域が frame ごとにグラフィカルに表示されています。ここで紹介する環境配列の例は、マイナス frame です。画面 9 の sbjct (相同性検索で見つかった環境配列) では、857 が開始側で 90 が終止側です。クエリーの既知タンパク質の遺伝子配列と逆方向の番号付けなのは、マイナス frame の ORFであることを示しています。

ここで、画面 9 で得た領域情報「Frame=-2, 塩基位置 90...857 で相同性が高い」が、ORF Finder の抽出した遺伝子領域の結果の中に含まれているかどうかを確認します。今回は、上から 1 つ目の [-2 □ 42..860 819] の領域が、画面 9 で行った相同性検索結果の領域情報「塩基位置 90...857」を含んでいます。但し、この画面 9 の相同領域「塩基位置 90...857」と、画面 13 で得られた環境由来のタンパク質遺伝子の ORF「塩基位置 42..860」とで、開始と終止の位置に差があります。これは既知タンパク質と環境タンパク質のアミノ酸配列の類似性が、タンパク質の末端部分でやや低いことを意味します。この程度の差異は、一般的に見られる現象です。

これで、目的としていた *bphF* 遺伝子候補が、海洋由来のメタゲノム [AACY023301218.1] の塩基配列の中に見つかりました。

☐ 部をクリックして、遺伝子領域の配列情報を表示させます。

画面 13 で表示されたどの遺伝子の候補領域にも、自分が画面 9 で得た領域情報がほとんど合致しない場合があります。環境 DNA がタンパク質遺伝子の一部の配列しか持っていない、例えば遺伝子の開始コドンを含む部分を欠損している可能性が考えられます。このような場合はレポート様式 2 のコメント欄に” fragment (断片)” と記入しておいて下さい。

再び画面 8 に戻り、別の候補 gb|AACY020561156.1|, gb|AACY020347673.1|, gb|AACY023307893.1| 等で、順次 ORF Finder の結果からの遺伝子領域に含まれる候補を探し続けて下さい。

NCBI ORF Finder (Open Reading Frame Finder) (画面 13)

PubMed Entrez BLAST OMIM Taxonomy Struc

gi|130393036|gb|AACY023301218.1| Marine metagenome

View 1 GenBank Redraw 100 SixFrames

クリックする

レポート様式 2 へ記入

読み枠

開始塩基位置

終了塩基位置

Frame	from	to	Length
-2	42..	860	819
+3	804..	1325	523
-2	870..	1211	342
-1	949..	1116	168
-3	563..	724	162
-3	1..	154	154
+2	104..	247	144
+1	523..	663	141
-3	377..	505	129
+1	1162..	1278	117
+1	328..	441	114
-3	230..	340	111
+2	1001..	1105	105
-3	986..	1090	105

全ての候補領域が表示される。

領域がピンク色で強調表示され、遺伝子領域の配列情報が表示されます。
[Accept] コマンドをクリックします。

(画面 14)

View 1 GenBank Redraw 100 SixFrames

Length: 272 aa

Accept Alternative Initiation Codons

クリックする

860 atgatgaacacccgtgaaccgctttaaaaaagccatcactgaa
M M P T P V N R F K K A I T E
815 ggttgtggatggcctcgccagcagttac
R R A Q L G L W M G L A S S Y
770 agcgcagagatgtgcccattgctggttcgactggttgtaatt
S A E M C A M A G F D W L L I
725 gatggcgagcatgcgccaataacttgcaaactattcagcaacaa
D G E H A P N N L Q T I Q Q Q
680 cttcaaacgattgcgacgtaccggtcagcaacgcaattgcgacg

次の画面が表示されます。Fasta protein を選択し、[View] アイコンをクリックします。

(画面 15)

ORF Finder (Open Reading Frame Finder)

NCBI

PubMed Entrez BLAST OMIM Taxonomy Structure

gi|130393036|gb|AACY023301218.1| Marine metagenome ctg_110

View 3 Fasta protein Redraw 100 SixFrames

1 GenBank
2 Fasta nucleotide
3 Fasta protein
4 ASN.1

① 「3 Fasta protein」を選択し
② [View]コマンドをクリックする。

Frame	from	to	Length
-2	42..	860	819
+3	804..	1325	523
-2	870..	1211	342
-1	949..	1116	168
-3	563..	724	162
-3	1..	154	154
+2	104..	247	144
+1	523..	663	141
-3	377..	505	129
+1	1162..	1278	117
+1	328..	441	114
-3	230..	340	111
+2	1001..	1105	105
-3	986..	1090	105

レポート(様式 2)には、「2 Fasta nucleotide」を選択して表示される DNA 配列情報も記入しておく。

次のようなファスタ形式と呼ぶ計算機処理の容易な形式でのアミノ酸配列が表示されます。

http://www.ncbi.nlm.nih.gov/gorf/orfig.cgi - Microsoft Internet Explorer (画面 16)

ファイル(E) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

戻る 検索 お気に入り

アドレス http://www.ncbi.nlm.nih.gov/gorf/orfig.cgi

```
>|c|Sequence 1 ORF:42..860 Frame -2
MMKTPVNRFFKKAITEKHAQIGLWMGLASSYSAEMCAMAGFDWLLIDGEHAPNNLQTIQQQLQSI AAYPVN
NAIARVPVGDALIKQYLDLGAETLLIPMIDTPQQAEQCVQAMRYPQNDGKGGVVRGMGGARASRWGMFPN
YTNEANEQVCLLVQAETREALKNLDAIANTPGVGVFIPADLSASLGHVGPDPNHPEVQAAIEDAIARIL
KAGKAPGILTSDVTQAKHYLSLGA LFVAVGLDTQILVRQTSALVRQFKPDAGIALPTSGQVY*
```

アミノ酸配列をコピーしておく。レポート(様式 2)に記入する。

アミノ酸配列をコピーしておきます。

3.5. 環境 DNA より見つかった遺伝子が由来するゲノムの推定

これまでに、キーワードに基づく既知のアミノ酸配列を取得し、環境由来の混合ゲノム解析で得られた膨大な遺伝子の候補から、最初に設定したタンパク質遺伝子とアミノ酸配列の類似したタンパク質の遺伝子を検索し、更にその中の遺伝子領域まで調べることが出来ました。最後に、環境から得られたタンパク質遺伝子が、どのような生物種のゲノムに由来しているのかを推定しておきましょう。難培養性の微生物の場合、今まで研究の進んでいなかった新種である可能性が考えられます。そのような場合でも、どの生物種と近いかを調べ、既知の生物種との類縁関係を推定することは可能です。但し、取得した DNA 配列を持つ生物種そのものが推定できる訳ではないことに十分に注意して下さい。あくまでも、現在判明しているタンパク質配列の類似性を利用した推定であり、新規性の高いタンパク質については、類似性を比較する対象が少数に限定されるので、著しく不正確な生物種の推定になる恐れがあります。長浜バイオ大の我々の研究グループはこの不十分さを解消するための研究開発を行っています。興味のある方は、下記の参考文献を参照して下さい。

阿部貴志、池村淑道、木ノ内誠、中村由紀子、前野聖、金谷重彦 ” SOM の医学・生物学からバイオ産業への応用まで ” 「自己組織化マップとその応用」、徳高ら編、シュプリンガー・ジャパン、p87-98 (2007)

- 3.5.1. NCBI の BLAST を立ち上げます。 <http://www.ncbi.nlm.nih.gov/BLAST/> (NCBI のトップ画面のメニューバーにある「BLAST」コマンドをクリックしても構いません。) **protein blast** を選択します。

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help (画面 17)

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

[Learn more](#) about how to use the new BLAST design

BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

<input type="checkbox"/> Human	<input type="checkbox"/> Oryza sativa	<input type="checkbox"/> Gallus gallus
<input type="checkbox"/> Mouse	<input type="checkbox"/> Bos taurus	<input type="checkbox"/> Pan troglodytes
<input type="checkbox"/> Rat	<input type="checkbox"/> Danio rerio	<input type="checkbox"/> Microbes
<input type="checkbox"/> Arabidopsis thaliana	<input type="checkbox"/> Drosophila melanogaster	<input type="checkbox"/> Apis mellifera

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query Algorithms: blastn, megablast, discontinuous megablast
protein blast	Search protein database using a protein query Algorithms: blastp, psi-blast, phi-blast
blastx	Search protein database using a translated nucleotide query

クリック →

3.4.2(画面 16)で得たアミノ酸配列を貼り付け、blastp を実行します。
データベースの設定は、Non-redundant protein sequences (nr)でよい。

BLAST Basic Local Alignment Search Tool (画面 18)

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ blastp suite: BLASTP programs search protein databases using a protein query. [more..](#)

Enter Query Sequence

Enter accession number, gi, or FASTA sequence [Clear](#)

```
>lc1|Sequence 1 ORF:42..860 Frame -2
MMKTPVNRFKKAITEKHAQIGLWMGLASSYSAEMCAMAGFDWLLIDGEHAPNNLQTIQQQLQSIAAYP
VN
NAIARVPVGD TALIKQYLDLGAETLLIPMIDTPQQAEEQCVQAMRYPQNDGKGGVRGMGGARASRWGMF
PN
```

Query subrange

From

To

① (画面 16)で得たアミノ酸配列を貼り付ける。

Or, upload file [参照...](#)

Job Title

Enter a descriptive title for your BLAST search

Choose Search Set

Database

Organism

Optional

Entrez Query

Program Selection

Algorithm

☒ blastp (protein-protein BLAST) [Blastp](#)

☐ PSI-BLAST (Position-Specific Iterated BLAST)

☐ PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

②クリック

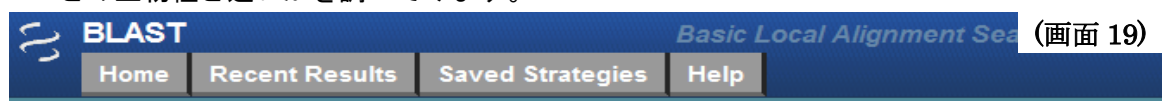
BLAST

Search database **nr** using **Blastp (protein-protein BLAST)**

☐ Show results in a new window

次のように、blastの結果が表示されます。この画面で、

- ④ [Taxonomy reports](#) をクリックすると、ヒットした配列の生物情報が表示されます。
 - ⑤ [Distance tree of results](#) をクリックすると、相同配列間での系統樹を見ることが出来ます。
- どの生物種と近いかを調べてみます。



▶ [NCBI/ BLAST/ blastp/ Formatting Results - 9SNA9MCC012](#)

[\[Reformat these Results\]](#) [\[E\]](#)

Job Title: lcl|Sequence 1 ORF:42..860 Frame -2

BLASTP 2.2.17 (Jun-24-2007)

Reference:

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, 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.

Reference:

Schäffer, Alejandro A., L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and Stephen F. Altschul (2001), "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements", Nucleic Acids Res. 29:2994-3005.

RID: 9SNA9MCC012

Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
5,304,967 sequences; 1,837,733,703 total letters

If you have any problems or questions with the results of this search please refer to the [BLAST FAQs](#)

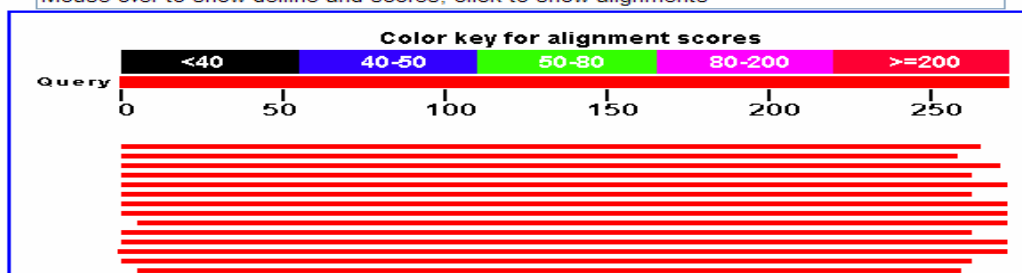
[Taxonomy reports](#)

← クリックすると、ヒットした配列の生物情報が表示される

Query= lcl|Sequence 1 ORF:42..860 Frame -2
Length=272

Distribution of 100 Blast Hits on the Query Sequence

Mouse-over to show define and scores, click to show alignments



[Distance tree of results](#) NEW

← クリックすると、相同配列間での系統樹が表示される

Sequences producing significant alignments:		Score (Bits)	E Value
ref ZP_01579792.1 	2,4-dihydroxyhept-2-ene-1,7-dioic acid ald...	315	1e-84
ref YP_708489.1 	4-hydroxy-2-oxovalerate aldolase [Rhodococcu...	300	6e-80 G
ref YP_995594.1 	2-dehydro-3-deoxyglucarate aldolase [Vermine...	298	3e-79 G
ref YP_522218.1 	2-dehydro-3-deoxyglucarate aldolase [Rhodofe...	294	3e-78 G
ref ZP_01716263.1 	2,4-dihydroxyhept-2-ene-1,7-dioic acid ald...	294	3e-78
ref ZP_01367563.1 	hypothetical protein PaerPA_01004715 [Pseu...	291	3e-77
ref YP_608657.1 	2,4-dihydroxyhept-2-ene-1,7-dioic acid aldol...	290	6e-77 G
ref NP_252817.1 	hypothetical protein PA4128 [Pseudomonas aer...	289	1e-76 G
ref NP_246474.1 	HpaI [Pasteurella multocida subsp. multocida...	286	9e-76 G
ref YP_788994.1 	putative 2,4-dihydroxyhept-2-ene-1,7-dioic a	285	1e-75 G

Taxonomy reports をクリックした画面

(画面 20)

Lineage Report

```

Bacteria [subbacteria]
├── Proteobacteria [proteobacteria]
│   ├── Betaproteobacteria [b-proteobacteria]
│   │   ├── Burkholderiales [b-proteobacteria]
│   │   │   ├── Comamonadaceae [b-proteobacteria]
│   │   │   │   ├── Delftia acidovorans SPH-1 ----- 315 4 hits [b-proteobacteria] 2,4-dihydroxy-2-dehydro-3-oxoheptanoate
│   │   │   │   ├── Therminephrobacter eiseniae FWH-2 ----- 298 2 hits [b-proteobacteria] 2-dehydro-3-oxoheptanoate
│   │   │   │   ├── Rhodococcus fascians T118 ----- 294 2 hits [b-proteobacteria] 2-dehydro-3-oxoheptanoate
│   │   │   │   ├── Polaromonas sp. JS666 ----- 249 6 hits [b-proteobacteria] 2,4-dihydroxy-2-dehydro-3-oxoheptanoate
│   │   │   │   ├── Acidovorax avenae subsp. citrulli AAC00-1 ----- 243 2 hits [b-proteobacteria] 2,4-dihydroxy-2-dehydro-3-oxoheptanoate
│   │   │   │   ├── Comamonas testosteroni KF-1 ----- 238 2 hits [b-proteobacteria] 2,4-dihydroxy-2-dehydro-3-oxoheptanoate
│   │   │   │   ├── Bordetella bronchiseptica RB50 ----- 283 2 hits [b-proteobacteria] 2,4-dihydroxy-2-dehydro-3-oxoheptanoate
│   │   │   │   ├── Ralstonia pickettii 12J ----- 265 2 hits [b-proteobacteria] 2-dehydro-3-oxoheptanoate
│   │   │   │   ├── Burkholderia sp. 383 ----- 265 2 hits [b-proteobacteria] 2-dehydro-3-oxoheptanoate
│   │   │   │   └── Burkholderia cenocepacia MC0-3 ----- 263 2 hits [b-proteobacteria] 2,4-dihydroxy-2-dehydro-3-oxoheptanoate

```

トップスコア (315) の [Delftia acidovorans SPH-1](#) をクリックすると、生物種の分類情報（界，門，綱，目，科，属，種）等が表示されます。

NCBI Entrez PubMed Nucleotide Protein Genome Structure PMC

Search for as complete name ☐ lock

Display 3 levels using filter: none

Delftia acidovorans SPH-1

Taxonomy ID: 398578
Rank: no rank
Genetic code: [Translation table 11 \(Bacterial and Plant Plastid\)](#)
Other names:
equivalent name: **Delftia acidovorans strain SPH-1**
equivalent name: **Delftia acidovorans str. SPH-1**

レポート様式2の、相同性検索から予測される種名 欄に記入する

↓

Lineage(full)
[cellular organisms](#); [Bacteria](#); [Proteobacteria](#); [Betaproteobacteria](#); [Burkholderiales](#); [Comamonadaceae](#)
[Delftia](#); [Delftia acidovorans](#)

Comments and References:

Distance tree of results をクリックし、

①Tree methodで「Neighbor Joining」を選択し、

②Sequence Label で[Taxonomic Name]を選択した時の画面が、次のように表示されます。

(画面 22) の系統樹の画面が小さくて見にくい場合は、系統樹の画面の上にカーソルを持っていき、右クリックして「画面を表示する」を選択すると拡大表示されます。

Tree view for rid: 9VADV8NP (画面 22)

This tree was produced using BLAST

①選択 Tree method ? Neighbor Joining

②選択 Sequence Label ? Taxonomic Name (if available)

Max Seq Difference ? 0.75 Reset

rectangle slanted radial force ☒ Show

クエリ配列が黄色で強調表示される

Rhodococcus sp. RHA1

Verminephrobacter eiseniae EF01-2

Rhodofera

unknown

Delftia acidovorans SPH-1

Bordetella bronchiseptica R50

Bordetella adum 197N

Ralstonia pickettii 12J

Ralstonia pickettii 12D

Dechloromonas aromatica RCB

Taxonomy reports (画面 20) 並びに Distance tree of results (画面 22) の画面より、*Delftia acidovorans* SPH-1 (別名 *Comamonas acidovorans* SPH-1 コマモナス・アシドボランス) が一番近く、この微生物が由来の候補として考えられます。

次に、この *Delftia acidovorans* SPH-1 がどのような微生物かを、文献やネット等で調べてみましょう。どのような環境下にいるのか、機能についてコメントがないか、他の応用例がないか 等を調べてみるとよいでしょう。調べを進めると、*Delftia acidovorans* SPH-1 は、プラスチック分解菌で、2,4-dihydroxyhept-2-ene-1,7-dioic acid aldolase という酵素を持っていることが既に知られています。このことから今回探索した遺伝子候補が、探索目的としていたアルドラーゼの酵素遺伝子を持つ可能性が、より確からしくなったといえます。

皆さんがされた解析では Taxonomy reports (画面 20) と Distance tree of results (画面 22) の画面で得られる近縁の生物種が異なる場合が多いと思います。環境から得られたタンパク質遺伝子の由来する生物種、あるいはその遺伝子について、余り研究が進んでいない場合にはそのような事が起きます。判断に困る場合は、Taxonomy reports (画面 20) のトップスコアの生物種を優先させておいてください。今回の場合 (画面 20) のトップスコア 315 の *Delftia acidovorans* SPH-1 になります。

以上で、作業は終了です。皆さんが興味を持たれた、自然環境の浄化や保全に役立つ、広い意味では「持続可能型社会の実現に貢献できる」可能性を持つ遺伝子を、国際塩基配列データベースに機能未知のままに残されている塩基配列から探索し、その遺伝子の由来する生物種が既知のどんな生物種と近縁であるかの情報まで得ることが出来ました。

長浜バイオ大学では、学部生が探索した自然環境の浄化や保全に役立つ遺伝子類候補を「持続可能型社会への貢献遺伝子データベース」として公開し発信していく予定です。希望される方については、皆さんが発見された持続可能型社会への貢献遺伝子の候補を、このデータベースに所属・氏名を記載して登録することも可能です。レポート (様式 1・様式 2) の電子ファイルの必要な方はお知らせ下さい。データベースへの登録を希望される場合は、レポート (様式 1・様式 2) を添付頂き、電子メールにてお送り下さい。

送信先 : h_uehara@nagahama-i-bio.ac.jp

長浜バイオ大学 生体分子情報学研究室 遺伝子探索 係

※ 件名に (遺伝子探索) と明記願います。

以 上

(付録 1) キーワード検索システム ARSA (<http://arsa.ddbj.nig.ac.jp/>) について

キーワード検索とは、私たちが日常使っている単語や技術用語を、検索用の手がかり (検索語) として、データベースを検索し、検索語を含むエントリを取り出すことを指す。

データベース内のエントリにはデータ項目が予め登録されている (予約語)。国際塩基配列データベース (DDBJ/EMBL/GenBank) では、フラットファイルに記載されている DEFINITION, Author 名, 登録日, ORGANISMS などが、予約語として、使用できる。

ここで、キーワード検索の仕組みについて、簡単に説明する。一般に以下の二つの検索方法があります。

フィールド検索 : 予約語と検索語を組み合わせで使用する方法

全文検索 : 検索語だけを与えて、エントリ全体を検索対象とする方法

また、検索したいキーワードの組み合わせを以下の仕組み (論理演算子) を使用して、キーワードを組み合わせで検索を行う事ができる。

---使用できる論理演算子について

and, or, not → 2 つ以上の検索語を指定して検索するときに使用。

and : A かつ B

or : A もしくは B

not : A 以外のもの

例) 「ヒトとマウス以外の生物の神経組織で発現している遺伝子」 ⇔ not “ヒト” and not “マウス” and “神経組織” と表すことができる。

さて、ここでは、国際塩基配列データベース (DDBJ/EMBL/GenBank) 中に登録されている塩基配列データを検索するためのシステム ARSA (All-round Retrieval of Sequence and Annotation) の使用方法について、簡単に紹介する。

ARSA とは、最新の国際塩基配列データのみならず、アミノ酸配列データ (UniProt/Swiss-Prot, UniProt/TrEMBL)、モチーフ配列データ等 23 のデータベースについて、データベース間を横断的に検索することが可能なキーワード検索システムである。特に、検索スピードが速く、例えば、” human” といった検索キーワードを入れた場合 (検索結果は、100 万件以上) には、10 秒前後で検索結果を得ることができるのが特徴である。

以下に各ページの簡単な説明と使い方を示す。詳しくは、ARSA の HELP を参照すると良い。

ARSA

All-round Retrieval of Sequence and Annotation

[English](#)
[Your Comment](#)

お知らせ

- ◆ [\(5/15 12:00-14:00\) 国立遺伝学研究所ネットワークサービス停止 \(2007年5月10日\)](#)
- ◆ [2007年2月のシステム移行に伴うお知らせ \(重要\) \(2007年3月23日\)](#)
- ◆ [データベースの説明](#)
- ◆ [データ更新情報](#)

Quick Search

All Databases

human

Search

検索条件を複数指定する場合は、語句を半角空白で区切ってください。AND条件で結合します。

DBDJ Advanced Search

Simple Query

DBDJデータを対象にしたキーワードでの検索が可能です。

Standard Query

検索する項目を指定でき、項目とキーワードを5セットまで指定できます。

Extended Query

Standard Search よりも項目が細かく分かれており、詳細な条件を指定しての検索が可能です。

Search History

参照...

Upload

検索履歴を表示します。<1件以上ヒットした検索条件のみ表示>

とりあえず、ここに検索したいキーワードを入力し、クリック。

ARSAトップページ

ARSA

All-round Retrieval of Sequence and Annotation

[Your Comment](#)

[Top](#)
[Standard Query](#)
[Extended Query](#)
[Search History](#)

Hit Counts List

Query

human

Sequence Libraries

• DDBJ	2,725,818	• DAD	765,484
• PRF	95,316	• UniProt/Swiss-Prot	31,901
• UniProt/TrEMBL	451,227	• IMG/LIGM-DB	63,947

Sequence Related

• PROSITE	1,473	• PROSITE-DOC	403
• BLOCKS	8	• PRINTS	1,388
• PFAMA	711	• PFAM-B	0
• SWISS-PROT	53,841	• PFAM-HMM-FS	5
• PFAM-HMM-LS	5	• PFAM-SEED	711
• PRODOM	42,338	• ENZYME	1,010

Protein 3D Structures

• PDB	12,237	• HSSP	724
• FSSP	655		

Metabolic Pathways

• LENZ-1	316	• L-COMPOUND	4
----------	---------------------	--------------	-------------------

ここをクリックすると、各データベースの詳細が表示

Hit Counts ページ (各データベースの検索結果)

ARSA All-round Retrieval of Sequence and Annotation

Top Standard Query Extended Query Search History

Query: human

ヒット件数 データベース名

検索結果をタブ区切りテキストで取得

2,725,818 results found in DDBJ

Download in TSV

Download all the contents displayed in Tab Separated Value format

FlatFile XML FASTA View Add to DownloadList

Primary Accession Number	Definition	Sequence Length
AB080903	Bifidobacterium sp. Cj59 gene for 16S rRNA, partial sequence	506
AB080904	Bifidobacterium sp. Cj59 gene for 16S rRNA, partial sequence	502
AB080905	Propionibacterium sp. Cj82 gene for 16S rRNA, partial sequence	502
AB091384	Abiotrophia defectiva dltC and dltB genes for D-alanyl carrier protein and peptidoglycan biosynthesis protein, complete and partial cds	633
AB128799	Coxiella burnetii icd gene for isocitrate dehydrogenase, partial cds, isolate:604 1w	370
AB128801	Coxiella burnetii icd gene for isocitrate dehydrogenase, partial cds, isolate:SS4 2nd 11.7 (No.72)	370

クリックするとフラットファイルが表示される

Sequence Libraries

Database	Count
DDBJ	2,725,818
DAD	765,484
PRF	95,316
UniProt/Swiss-Prot	31,901
UniProt/TrEMBL	451,227
IMG/IMG-DB	63,947

Sequence Related

Database	Count
PROSITE	1,473
PROSITE/DOC	403
BLOCKS	8
PRINTS	1,388
PFAMA	711
PFAMB	0
SWISSPFAM	53,841
PFAMHMMFS	5
PFAMHMMLS	5
PFAMSEED	711
PRODOM	42,338
ENZYME	1,010

Protein 3D Structures

Database	Count
PDB	12,237
HSSP	724
ESSP	655

検索結果詳細表示

その他、検索機能について

- Standard Query
 - DB の項目をプールダウンメニューで選択後、項目ごとに検索条件を指定して検索を行える(最大5項目まで)。
- Extended Query
 - 各 DB の項目ごとに検索条件の指定が可能のため、Standard Search より詳細に検索条件を指定して検索を行える。
- Search History
 - 各検索(Simple Search, Standard Search, Extended Search)画面から検索した結果の履歴を表示。各履歴から絞り込み検索が可能。また、履歴をローカルマシンに保存することも可能。Upload File ローカルマシンに保存した検索履歴をアップロードする。アップロードしたい検索履歴ファイルを指定して、Submit ボタンをクリックする。

Standard Queryについて

ARSA

All-round Retrieval of Sequence and Annotation

出力件数の選択

[Top](#)
[Simple Query](#)
[Standard Query](#)
[Extended Query](#)
[Search History](#)

Combine Searches with ANC
Number of Entries 30
☒ Display Xpath Expression

検索条件を複数指定する場合は、語句を半角空白で区切ってください。ダブルクォーテーション()で囲まれた文字列は、1つのキーワードとして認識されます。

Query Value

All text

=

ANC

All text

=

ANC

All text

=

ANC

All text

=

ANC

All text

=

ANC

All text

=

ANC

All text

Accession number

Primary Accession Number

Version

Division

Sequence Length

Molecular Type

Molecular Form

Date

Definition

Comment

Keywords

Organism

Taxon

Authors

Title

Journal

PubmedID

キーワード入力

検索項目については、プルダウンで項目を選択することが可能

Search

Clear

Results View Image

1st

Primary Accession Number

2nd

Definition

3rd

Sequence Length

4th

5th

6th

出力情報の選択

(付録 2) 国際塩基配列データベース (DDBJ/EMBL/GenBank) フラットファイルについて

国際塩基配列データベース (DDBJ/EMBL/GenBank) は、登録されている配列データの単位である「エントリ」の集合として構成されている。DDBJ に登録されたそれぞれのエントリは、DDBJ の定めるフォーマットにしたがった「フラットファイル」(flat file) の形式で公開されている。フラットファイルには、塩基配列のほか、塩基配列を決めた登録者、関連文献、生物種や機能情報など必要な情報が全て表示されている。そのため、フラットファイルに書かれていることを読み解けることが大事になる。

— DDBJ のデータ公開形式の説明より (<http://www.ddbj.nig.ac.jp/sub/ref10-j.html>)

LOCUS AB000000 450 bp mRNA linear HUM 08-JUL-2002
DEFINITION Homo sapiens GAPD mRNA for glyceraldehyde-3-phosphate dehydrogenase, partial cds.
ACCESSION AB000000
VERSION AB000000.1
KEYWORDS .
SOURCE Homo sapiens
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 450)
AUTHORS Mishima, H. and Shizuoka, T.
TITLE Direct Submission
JOURNAL Submitted (30-NOV-2000) to the DDBJ/EMBL/GenBank databases.
Hanako Mishima, National Institute of Genetics, DNA Data Bank of Japan; Yata 1111, Mishima, Shizuoka 411-8540, Japan (E-mail:mishima@supernig.nig.ac.jp, Tel:81-55-981-6853, Fax:81-55-981-6849)
REFERENCE 2 (sites)
AUTHORS Mishima, H., Shizuoka, T. and Fuji, I.
TITLE Glyceraldehyde-3-phosphate dehydrogenase expressed in human liver
JOURNAL Unpublished (2002)
COMMENT Human cDNA sequencing project.
FEATURES Location/Qualifiers
source 1..450
/chromosome="12"
/clone="GT200015"
/clone_lib="lambda gt11 human liver cDNA (GeneTech. No. 20)"
/map="12p13"
/mol_type="mRNA"
/organism="Homo sapiens"
/tissue_type="liver"
CDS 86..450
/codon_start=1

```

/gene="GAPD"
/product="glyceraldehyde-3-phosphate dehydrogenase"
/protein_id="BAA12345.1"
/transl_table=1
/translation="MAKIKIGINGFGRIGRLVARVALQSDDVELVAVNDPFIITDYM
YMFKYDTVHGQWKHHEVKVKDSKTLFGEKEVTVFGCRNPKEIPWGETSAEFVVEYTG
VFTDKDKAVAQLKGGAKKV"

```

BASE COUNT

102 a 119 c 131 g 98 t

ORIGIN

```

1  cccacgcgtc cggtogcatc gcacttgtag ctctcgaccc ccgcatctca tccctcctct
61 cgcttagttc agatcgaaat cgcaaatggc gaagattaag atcgggatca atgggttcgg
121 gaggatcggg aggctcgtgg ccagggtggc cctgcagagc gacgacgtcg agctcgtcgc
181 cgtcaacgac cccttcatca ccaccgacta catgacatac atgttcaagt atgacactgt
241 gcacggccag tggaagcatc atgaggttaa ggtgaaggac tccaagaccc ttctcttcgg
301 tgagaaggag gtcaccgtgt tcggctgcag gaaccctaag gagatcccat ggggtgagac
361 tagcgctgag tttgttgtgg agtacactgg tgttttcact gacaaggaca aggccgttgc
421 tcaacttaag ggtggtgcta agaaggtctg

```

//

各項目について、以下に簡単に説明する。

・ LOCUS

配列の全長、分子タイプ、最終更新日など

・ DEFINITION

エントリの要約。DDBJが自動構築する場合と登録者が直接記載する場合がある。一般的には、遺伝子を抽出した生物名、遺伝子名、タンパク質名、遺伝子全長の有無、保存株名などで構成。

・ ACCESSION

データバンクが発行する登録番号（アクセッション番号）。

・ VERSION

➤ 「アクセッション番号. バージョン番号」で記載。

・ KEYWORD

➤ キーワード。現在は、基本的にデータバンク側で必要がある場合に限り入力。

・ SOURCE

この塩基配列を決定した生物種に関する情報。学名だけではなく分子種やクローン名も記載。また、付随の ORGANISM には、学名が記載されるが、これは、GenBank の Taxonomy デ

データベースに登録されているものが記載。

- ・ **REFERENCE**

登録者の情報と参考文献情報。

- ・ **COMMENT**

➤ 登録者が付加するコメント。ex.) 実験情報

- ・ **FEATURES**

DDBJ/EMBL/GenBank の Feature Table Definition に沿って、配列付加情報を記載。

(http://www.ddbj.nig.ac.jp/FT/full_index.html)

ここで、代表的な Feature Key について、簡単に説明する。

source

配列の生物学的な由来やサンプル取得状況が記載されている。

Source の補足情報 (Qualifier key) として、

/isolation : サンプル取得地点情報

/organism : 生物種名

/mol_type : 分子種

/isolation_source : サンプル取得地点情報

/db_xref : 他の DB の ID 情報 (あれば)

/country : 取得国に関する情報

CDS

タンパク質のアミノ酸をコードする領域を記載。(ex: 86..450. 86 番目から 450 番目までに CDS 領域があることを指す。相補鎖の場合には、complement(xx..yy) と記載される。)

CDS の補足情報として、

/product : 遺伝子の機能についての記載。

/note : 著者らによりコメントを記載。

/trans_table : 対応するコドンテーブルの番号を記載。

tRNA

tRNA についての情報を記載。

rRNA

rRNA についての情報を記載。

- ・ **BASE COUNT** 塩基配列組成

- ・ **ORIGIN** 塩基配列情報

(付録 3) PCB 分解

PCB (ポリ塩素化ビフェニル) はビフェニルの水素原子を塩素原子で一部置換した一連の化合物の総称で、不燃性、絶縁性、溶媒性にすぐれ、ヒートポンプの熱媒体、コンデンサーの絶縁油、インキ溶剤などに 1960 年代まで広く用いられていた。1960 年代後半になり野生動物を脅かす環境汚染の元凶とされ、食用油への PCB 混入により深刻な被害者を出したカネミ油症事件などが発生したため、1970 年代初頭には各国で製造・使用・廃棄が厳しく制限され、市場から姿を消した環境汚染物質である。市場に出た PCB の 1/3 ほどが環境中に漏出したと推定されているが、化学的安定性に優れるため長期にわたり環境中に残留し、現在も環境汚染は無くなっていない。すでに環境中に広がってしまった PCB 汚染の処理とともに、現在大量に保管されているものを効率的に無害化するための対策も、重要な環境問題の一つである。

PCB の多くは油状物質であり (PCB には塩素の置換数、置換位置によって多くの類似体がある)、高温で分解することが出来る。しかし、700℃以下の低温で燃焼させた場合、毒性が桁違いに高いジベンゾフランやダイオキシンが生成するため、約 1,000℃以上の高温で処理する必要がある。さらに、このような処理施設建設には、パブリックアクセプタンス (社会的受け入れ; PA) を得にくいという問題もある。

そこで、エネルギー消費が少なく副産物も出来にくい、微生物を用いた生物処理 (バイオレメディエーション) が有効な手段として期待されている。

PCB は塩素置換のないビフェニルの共代謝によって分解されることが明らかになっている。ビフェニルを唯一の炭素源として生育できる土壌細菌は環境中に広く分布しており、比較的容易に分離することが出来る。ただし、どの程度の範囲の PCB 類似体が分解できるかは微生物によって異なっており、様々なタイプの PCB 分解菌が採取され、遺伝子レベルで解析されている。

長岡技術科学大学の福田教授のグループでは、世界的にトップクラスの PCB 分解能を有する、PCB 分解菌 *Rhodococcus* 属 RHA1 株の酵素と遺伝子の解析と強化を進めている。RHA1 株では、*bphA* (ビフェニル 2, 3-ジオキシゲナーゼ) をはじめ、*bphB*, *bphC*, *bphD*, *bphE*, *bphF*, *bphG* 他の PCB 分解に関連する各酵素遺伝子が存在する。それらがビフェニルで誘導されて協調的に PCB 分解に関わることを解明している。

参考文献等：福田雅夫, *Rhodococcus* 属細菌の PCB 分解システム, 蛋白質核酸酵素 Vol. 50 No. 12 (2005)

日本農芸化学会編, 人に役立つ微生物のはなし, 学会出版センター, 2002, 41-45
長岡技術科学大学 21 世紀 COE プログラム HP

<http://pelican.nagaokaut.ac.jp/GER/study/m21.html>

簡易的な実習例：例えば *bphA*, *bphB*, *bphC*, *bphD* で検索してみる。

- ① DDBJ の ARSA で、*bphA Rhodococcus* で検索をかけてみると、UniProt/Swiss-Prot で [005151](#) 1 件のみヒットする。画面 2 と異なり、生物種名 *Rhodococcus* を入れて検索範囲を限定しているので、1 件のみのヒットとなった。
- ② 次に、DDBJ の ARSA で、*bphB Rhodococcus* で検索をかけてみると、UniProt/Swiss-Prot で [P47230](#) 1 件のみヒットする。
- ③ 次に、DDBJ の ARSA で、*bphC Rhodococcus* で検索をかけてみると、UniProt/Swiss-Prot で、[P47231](#) (DHBD I) [P47232](#) (DHBD II) [P47233](#) (DHBD III) [069349](#) の 4 件がヒットする。
- ④ 次に、DDBJ の ARSA で、*bphD Rhodococcus* で検索をかけてみると、UniProt/Swiss-Prot で、[005151](#) 1 件のみヒットする。

(付録 4) BLAST について

BLAST (Basic Local Alignment Search Tool) は、DNA 配列（塩基配列）やアミノ酸配列の類似性を検索するための方法として最もよく使われているプログラムである。

比較したい配列の種類の違いにより、複数の検索プログラムがある。主に使われている BLAST の検索プログラムについて以下に紹介をする。

blastp

アミノ酸配列をクエリ配列として、ユーザが指定したアミノ酸配列データベースから類似するアミノ酸配列を検索する。

blastn

DNA 配列をクエリ配列として、ユーザが指定した DNA 配列データベースから類似する DNA 配列を検索する。

blastx

DNA 配列をクエリ配列として、全てのコドンの読み枠を対象にした 6 frame 分で翻訳しながら、ユーザが指定したアミノ酸配列データベースから類似するアミノ酸配列を検索する。

tblastn

アミノ酸配列をクエリ配列として、ユーザが指定した DNA 配列データベースを全てのコドンの読み枠を対象にした 6 frame 分で翻訳しながら、類似する DNA 配列を検索する。

tblastx

DNA 配列をクエリ配列として、全てのコドンの読み枠を対象にした 6 frame 分で翻訳し、ユーザが指定した DNA 配列データベースを同様に翻訳しながら、類似する DNA 配列を検索する。

以 上

持続可能型社会への貢献遺伝子候補の探索 —持続可能型社会への貢献遺伝子データベースの構築—

2007 年 8 月 5 日 初版

編集： 長浜バイオ大学 バイオサイエンス学部 生体分子情報学研究室
〒526-0829
滋賀県長浜市田村町 1266 番地

※ 本テキストの内容に関するお問い合わせ、遺伝子候補のデータベース登録依頼 等につきましては、本研究室 上原 啓史（E-mail: h_uehara@nagahama-i-bio.ac.jp）宛にお願いします。

本テキストの作成には、生体分子情報学研究室の他のメンバーとして、阿部貴志、棚橋佳世、池村淑道が参加している。