1. ゲノム情報解析法のバイオ実課題への応用：健康への貢献遺伝子データベースの構築

1. はじめに

長浜バイオ大学は、ゲノム塩基配列の解読が盛んになった時期に開学しており、ゲノム情報を代表例とする大量情報のコンピュータ処理のバイオ分野での重要性を考え、コンピュータとバイオの両方の知識と技術を持つ人材の育成に努めてきた。この全国的にも特徴のある教育が評価され、文部科学省が行っている「ライフサイエンス分野の統合データベース整備事業」において、学部レベルでの人材養成プログラムの担当校に選ばれた。本情報解析実習は、その文部科学省のプログラムの一環でもある。

ヒトを含む広範な生物種のゲノム配列が決定され、人類全体の共有財産として世界に公開されている。ゲノムの大量塩基配列やそこから得られるタンパク質アミノ酸配列のコンピュータを用いた情報処理は、医薬学分野に於いて益々その重要性を増している。本学においても、医薬学に関連した産業分野での就職やその関連研究分野の大学院への進学を希望する学生も多い。本実習では、「健康に貢献する可能性のある遺伝子」を、各自がコンピュータを用いて国際 DNA データベースを対象に探索して、新しい発見が得られれば、その結果を「健康への貢献遺伝子データベース」として収録し、公開することを目指している。それでは「なぜ実習でありながら新規な発見が可能なのか」から説明を行う。

最近の DNA 解読技術の進歩は著しいものがあり、新規性の高い生物種に由来する遺伝子塩基配列も大量に解読され、遺伝子機能の推定を行う情報処理が追いつかない状況下にある。国際 DNA データベース (DDBJ/EMBL/GenBank) に収録された塩基配列について、機能が未知のままに残されている遺伝子の候補は既に 500 万件を超えている。そのなかには医薬学的に有用なものも多数含まれているはずであるが、機能に関する記載が一切なされておらず、利用価値が低いままに国際 DNA データベースに収録されている。特に最近になって注目を集めているのが、環境微生物由来の遺伝子候補群である。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る。これら難培養性微生物については、培養操作を含む通常の実験的な研究が行えず、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きた。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、注目に値する。最近、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、その断片の塩基配列決定を行い、科学的や産業的に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る。このような解析は、「メタゲノム解析」と呼ばれている。メタゲノム解析で大量なゲノム断片配列が決定され、データベースに収録されているが、これらの環境由来の DNA 配列については、現時点では遺伝子機能や由来生物系統に関するアノテーション情報の品質は登録者によってまちまちである。有用な機能を持つ遺伝子を多く含んでいる可能性が高いにもかかわらず、情報として記載されていない場合が大半である。

本実習では、これらの環境微生物に由来する DNA 配列群を対象にして、コンピュータ解析で、有用な「お宝遺伝子」を発掘して、データベース化して学生の氏名入りで公開を行う。実習でありながら、新規な知識発見が可能なことを体験してほしい。但し、公開を行うことには責任も伴う。間違いをしないように充分に注意することが重要であり、この姿勢の習得も実習の主要な目的である。氏名入りでの公開を希望しない人は、レポートにその旨を記載しておくこと。

昨年度の生命情報科学専門実習においては、「持続可能型社会に貢献する可能性のある遺伝子」を、各自が具体的な課題を設定して、環境微生物ゲノムに由来する約 500 万件の遺伝子候補群の中からコンピュータを用いて探し出し、「持続可能型社会への貢献遺伝子のデータベース」を構築した。「農薬等の環境汚染物質の分解」や「バイオエタノール生産」に貢献する可能性のある遺伝子などを既に約 7000 件発掘しており、学生の氏名入りで公開している (URL: http://dbcls.nagahama-i-bio.ac.jp/)。

本年度は、環境微生物ゲノムに由来する大量な遺伝子候補群から、医薬学的に有用な遺伝子類、例えば抗生物質・免疫抑制剤・ウイルス治療薬・花粉症治療薬・抗腫瘍薬・機能性健康食品素材・臨床検査試薬・サプリメント等の生産に利用が可能な遺伝子類を探し出す。具体的には、国際塩基配列データベース(DDBJ/EMBL/GenBank)に登録されている、環境微生物試料のメタゲノム解析で得られた配列を検索の対象として、各自が具体的な課題を設定して、有用な遺伝子を発掘して、新規なデータベース「**健康への貢献遺伝子データベース**」として国内外へ発信をする。「健康に役立つ」をキーワードに、環境微生物の DNA 配列から有用な遺伝子を探索する試みを通じて、ゲノム配列解析手法や機能情報を含むアノテーション情報の付加の基礎を学ぶ。

操作の全体としての流れ

レポート（テーマ検索レポート，既知遺伝子レポート）へ記載する事項に関する操作。

2. 1) テーマ設定。
2. 2) そのテーマに合った薬剤や物質を知る。
2. 3) その薬剤や物質の合成を行っている、微生物由来の酵素の遺伝子を知る。
2. 4) その薬剤や物質や酵素タンパク質遺伝子の英語名を知る。
2. 5) それらの英語名を用いて、着目の薬剤や物質の合成を行っている酵素タンパク質遺伝子に関して、既に知られているアミノ酸配列を国際データベースから取得する。

これに引き続いて、新規遺伝子候補レポートへ記載する操作へ進む。

3. 1)得られた既知のアミノ酸配列と相同性の高い配列を持つタンパク質の遺伝子を持つ DNA 断片を、環境試料のメタゲノム解析で得られている大量なゲノム断片配列から探し出す。
3. 2) 得られたゲノム断片配列に遺伝子に関する記載が無ければ、タンパク質遺伝子の開始と終止点の位置を確定する。
3. 3) 得られたタンパク質の全域のアミノ酸配列を用いて、このアミノ酸配列と相同性の高い配列をデータベースから取得し、系統樹を作成して、得られたゲノム断片配列の由来する生物系統を推定する。

2．「健康への貢献が期待できるタンパク質遺伝子」を、データベースに収録された機能が既知の遺伝子セットから検索する方法：レポート（テーマ検索レポート，既知遺伝子レポート）へ記載する事項

2. 1 テーマ設定。

微生物の代謝産物には人間の叡智を超えた多様な化合物が存在し、有用な医薬品素材等の探索源として多くの可能性を秘めている。微生物は、ペニシリンやストレプトマイシンなどの抗生物質をはじめ、多くの薬理作用を示す物質を産生している。カビの代謝産物のひとつサイクロスポリンが、臓器移植時の拒絶反応を抑える為に必要不可欠な免疫抑制剤として臨床応用され、移植手術の成功率が飛躍的に高まったことなども挙げられる。

微生物が生産する抗がん剤，抗 HIV ウイルス薬，抗加齢(アンチエイジング)、花粉症治療薬，アレルギー治療薬，アトピーに効果のある免疫抑制剤や抗生物質等はテーマとなる。さらには、整腸作用のある機能性食品素材やサプリメント並びに次世代の機能性食品素材（抗ストレス，抗アレルギー，抗疲労，睡眠促進等）の産生に能力を発揮できそうな微生物由来の遺伝子類やタンパク質類は、「健康に貢献する」可能性を持つと考えられる。多くの女性が関心を持っているアンチエイジングについても、天然系の微生物由来のヒアルロン酸やエラスチン等

があれば探索テーマとして挙げてよい。

各自で独自性の高いテーマを出してもらいたい。複数のテーマでも良く、途中でテーマを変更しても良い。

2.2 テーマに合った薬剤や物質名を知る。

次に、テーマ（例えば、抗がん剤）を設定後に、どのような手順で、そのテーマに沿った遺伝子（この場合は、抗がん剤の生産に役立つ可能性のある微生物の酵素遺伝子）をどのような手順で探索したら良いのかを解説する。

例えば、「抗がん剤」をテーマとする場合、まずどのような微生物由来の抗がん剤が知られているかについて、Google 日本（http://www.google.co.jp/）の「日本語のページを検索」を用いて探す。但し、「抗がん剤」だけを Key Word（検索文字：キーワード）にして検索すると、余りにも多数のヒットが見られる。当然のことながら、化学合成された抗がん剤や高等植物が生産する抗がん剤に関するレポートや文献もヒットする。目的が、微生物の生産する抗がん剤を知りたいので、「抗がん剤　微生物由来」と入れると不要なヒットが減少する。それ以外にも、「抗癌剤　微生物由来」や「抗腫瘍薬　微生物由来」や「がん治療薬　微生物由来」等を試みることで、検索の範囲が広がる。ここで、各単語の間に空白を入れる方が良い（「　」は検索の欄へは入れない）。

（画面 1）

検索された文献の内容を良く理解した上で、微生物が生産することに確信が持てた抗がん剤の薬剤名を、「レポート.xls」ファイルの“テーマ検索レポート”シートの中の“検索用に使用した薬剤・物質名”の欄に記載する（“テーマ検索レポート”シートのテーマの欄には、検索テーマに用いたテーマ名、この場合は「抗がん剤」を記載しておく）。同一のテーマで、複数の異なった薬剤が得られた場合には、薬剤ごとに別の行に情報を記入していく。但し、同一の薬剤が異なった名前を持っている場合は、同一の “検索用に使用した薬剤・物質名”の欄内に併記すれば良い。薬剤に関しては、各製薬メーカーが独自の商品名で販売をしているが通例である。このような商品名では以降の検索が困難になるので、商品名ではなく、研究分野で通用している学術的な名称を探し出すことが大切である。ある薬剤や物質が微生物由来かどうか不明の時は、「物質名生産菌」を検索して微生物が見つかれば、それは微生物由来と判断できる。

2.3 薬剤や物質を合成している、微生物由来のタンパク質遺伝子名を知る。

次に、「レポート.xls」ファイルの“テーマ検索レポート”シートに記載した物質・薬剤を微生物が合成する際には、どのような酵素遺伝子類が関わっているのかを調べる。「薬剤名　生合成遺伝子」や「薬剤名　生合成経路遺伝子」のような検索を行うと、“テーマ検索レポート”シートに記入した薬剤の合成を行っている酵素遺伝子を検索できる可能性が高い。「薬剤名遺伝子」のような検索を行うと、その薬剤が作用を及ぼしている、ヒト遺伝子側を検索する確

立が高くなるので、注意が必要である。「薬剤名　生合成遺伝子」としても、Google の検索では、薬剤名と生合成と遺伝子が別個の検索の対象になるので、着目の薬剤名がヒトの何かの生合成経路の遺伝子産物（酵素等）に影響を及ぼす文献を探し出す場合もある。薬剤の生産には関与しない目的外の遺伝子がヒットしている可能性が常に存在しているので、文献の内容を充分に吟味することが重要である。

対象とする薬剤・物質の種類によっては、その合成に関する酵素学的な研究の方が遺伝子の研究よりも盛んなことがあり、酵素やタンパク質名が記載された文献やレポートの方が多いこともある。「薬剤名　生合成遺伝子」とした検索では良い結果が得られない場合には、「薬剤名 生合成酵素　微生物由来」のような工夫も有効なことがある。そのような工夫をした上でも、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ヒットが混在していることもあるので、複数の検索文章を読み進めて、目的とする「遺伝子」や「タンパク質」の名前を探しだす。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、最終的には興味深い遺伝子の新規発見につながる確率が高まる。

文献の内容を吟味して、目的物質の生産に関与する微生物遺伝子である確信が持てた例について、その遺伝子名を“テーマ検索レポート”シートの“Google で得られた遺伝子やタンパク質名”の欄に記載する。なお、本実習で設定したテーマを、「レポート.xls」ファイルの テーマ検索レポート/既知遺伝子レポート/新規遺伝子候補レポート の各シートの“テーマ”の欄に記載しておく。この**「レポート.xls」ファイル（テーマ検索レポート/既知遺伝子レポート/新規遺伝子候補レポート の３シートからなる）をデータベース登録用のレポートとして提出してもらう。**

この際、得られた遺伝子やタンパク質の名称を、個々の遺伝子名やタンパク質名ごとに異なった行に記載する。一種類の薬剤、例えば一種類の抗生物質を合成するのにも、通常は複数段階の酵素反応が必要であり、従って複数種類の酵素が関与している。検索で見つかった遺伝子名やタンパク質名に差があれば、その複数種類の酵素である可能性が高い。同一の酵素やその遺伝子の別名であることが明らかな場合を除いて、レポートの別行に記載しておく。個々の遺伝子やタンパク質の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、それもテーマ検索レポート”シートの“Google で得られた遺伝子やタンパク質の英語名”の欄へ記載しておく。

各自が設定したテーマ（例えば、「抗生物質」）から見出された一つの薬剤・物質名（例えば、「ペニシリン」）について、アシルトランスフェラーゼ，イソペニシリンNシンセターゼ，イソペニシリンNエピメラーゼなどの複数の遺伝子やタンパク質が検索できる場合もあるであろう。同じ薬剤・物質名に対応するこのような複数の行については、その全てについて、レポートの“テーマ”の欄にその設定項目(この場合は、「抗生物質」)を、“検索用に使用した薬剤・物質名”の欄に薬剤・物質名（この場合は、「ペニシリン」）を、コピー＆ペーストしておく。コピー＆ペーストに慣れれば、文字入力よりも間違いが少ない。コピー＆ペーストは、コピーしたい領域をマウスで選択後に、キーボードの「Ctrl」キーを押しながら「c」キーを押し（これでコピー出来る）、次に貼り付けたいところにカーソルを移動させて、「Ctrl」キーを押しながら「v」キーを押すと、貼り付けが完了する。遺伝子やタンパク質名が見つからなかったテーマでも、レポートのテーマの欄に残しておこう。興味深いテーマは、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る。

この段階で注意して置くべき重要な点がある。自分が興味を持つ微生物由来の薬剤名は見つかったのに、その薬剤を合成する酵素が Google 検索では見つかっていない例も多いと思う。Google の「日本語のページの検索」を行っているので、日本の研究者が余り研究していない

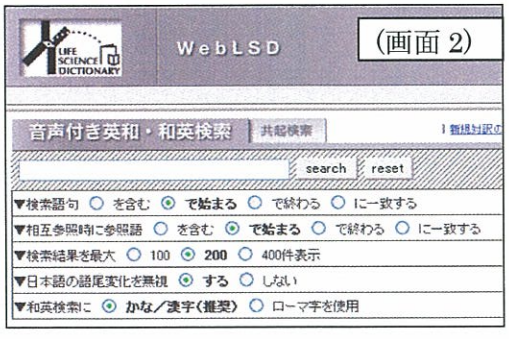
薬剤に関しては、日本語でのレポートが存在しない可能性が高い。そのような場合は、Google の「ウエブ全体からの検索」を行うのも一つの方法であるが、多数の英語レポートを読解する能力が要求される。本実習では、この方針はとらずに、後に説明する DDBJ の「ARSA」の機能を用いて酵素遺伝子の探索を行う。

2.4 得られた「薬剤や物質名」並びに「遺伝子」や「タンパク質」の英語名を知る方法

自分が目的とする薬剤・物質の合成を行っている酵素の遺伝子名が見つかったら、次はその酵素タンパク質アミノ酸配列が既に知られているのかを、DDBJ に収録されている国際配列データベースを対象にして検索を行う。国際配列データベースは英語で記載されているので、検索を行うためには着目遺伝子やタンパク質の英語名を知る必要がある。着目する遺伝子の情報を Google を用いて検索する過程で既に英語名が判明している場合には、その英語名をテーマ検索レポートの“Google で得られた遺伝子やタンパク質の英語名”の欄に記載する。その「遺伝子やタンパク質の機能や特徴」に関する興味深い説明文があれば、10～20 行程度に要約してテーマ検索レポートの「Google で得られた遺伝子やタンパク質の機能や特徴」の欄へ記載しておく。

遺伝子が Google 検索で分かっていなくても、薬剤や物質名の英語名が分かっていれば、DDBJ の ARSA を用いた検索機能で、着目の薬剤の合成酵素のアミノ酸配列を検索が可能な場合が多い。今回はこの方針で以降の検索を進める。薬剤や遺伝子やタンパク質の英語名を知る方法としては、ライフサイエンス辞書 Web LSD (http://lsd.pharm.kyoto-u.ac.jp/ja/service/weblsd/index.html) を用いる方法が便利で一般性が高い。従って、英語名が探せていない場合だけでなく、Google で遺伝子やタンパク質の英語名が判明している場合でも、ライフサイエンス辞書を用いて、英語名を確認しておく方が良い。ライフサイエンス辞書の URL は Google で「ライフサイエンス辞書」を Key Word として検索すれはすぐに見つかる。「お気に入り」に追加しておくと良い。ライフサイエンス分野の語彙が多く便利である。これを利用すれば、遺伝子名以外にも、バイオ分野の専門用語について英語名を探し出せるし、英語を和訳することも可能である。「シソーラス」として同義語も記載されているが、専門用語に関しては、同義語ではなく類似語も混在しているので、その点は注意が必要である。

得られた「薬剤や物質名の英語名」をテーマ検索レポートの“ライフサイエンス辞書で得られた薬剤や物質の英語名”の欄に記入する。「遺伝子やタンパク質の英語名」はテーマ検索レポートの“ライフサイエンス辞書で得られた遺伝子やタンパク質の英語名”の欄に記入する。余裕のあるグループは、この英語名を用いて Google 検索を行い、それら遺伝子やタンパク質の機能や特徴を英文として 10 行程度にまとめて、既知遺伝子レポートの“functions and characteristics”の欄へ追加することを試みよ。



※ライフサイエンス辞書の検索画面

また、ライフサイエンス分野専用ではないが、アルク英辞郎（URL：http://www.alc.co.jp/）は、登録語彙数が膨大で翻訳家も使用しているサイトである。こちらも試してみるとよい。

（画面 3）

2.5 調べた酵素遺伝子由来のタンパク質アミノ酸配列が国際配列データベースに登録されているかどうかの DDBJ を用いた調査

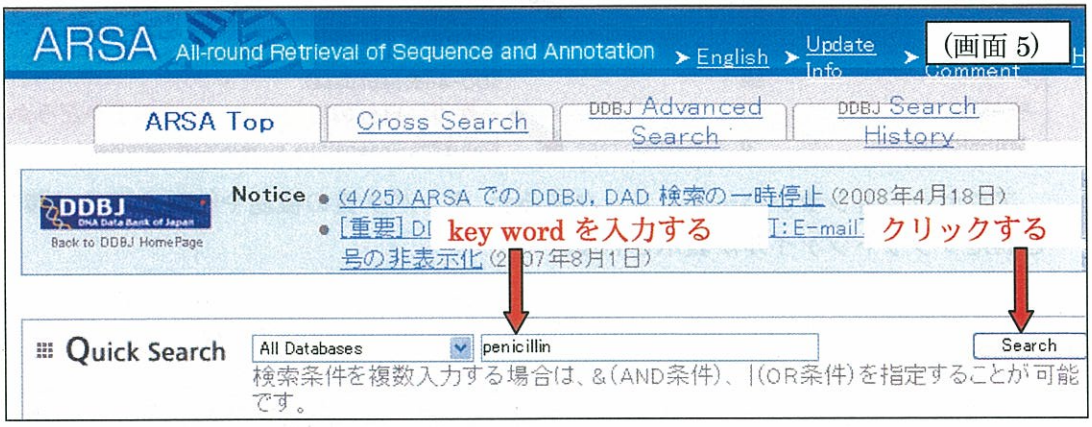
得られた薬剤や遺伝子やタンパク質の英語名を Key Word として、DDBJ(日本 DNA データバンク)の検索 tool の「ARSA」を用いて、着目の薬剤を合成する酵素タンパク質の遺伝子について、どの様なタンパク質アミノ酸配列が既に知られているかを検索する。「ARSA」は DDBJ が収録している国際的な 20 種類のデータベースを対象に、検索 Key Word が記載されている配列データを高速で取得するための検索 tool（検索エンジン）である。詳細な説明は、本テキストの巻末の付録部分を参照されたい。

ここで、なぜ遺伝子の塩基配列ではなく、タンパク質アミノ酸配列の方を検索するのかとの疑問が生じると思う。この疑問は、遺伝子進化の重要な問題と関係している。同じ機能を持つ酵素でも、生物種が異なればアミノ酸配列に差異が存在するのが通例である。近縁の生物種間ではその差異は小さく、同一のアミノ酸配列を持つこともあるが、系統的に遠くなるに従って、その差異が大きくなる。言い換えれば、系統的に遠い生物種間の比較になるに従い、進化的に同一の起源を持ち、同一の機能を持つ酵素についても、それら生物種の配列間での相同性を見出すのが段々困難になってくる。次に、この酵素遺伝子の塩基配列に着目した場合を考えてみる。近縁の生物種間で酵素のアミノ酸配列が同一の場合であっても、塩基配列はかなり異なっている例が多く見られる。同一のアミノ酸に対応している複数種類のコドン(同義コドン)間での差異（同義置換）が、近縁の生物種間でも頻度高く起こっていることが知られている。系統的に遠くなるに従い、アミノ酸を変化させる差異（非同義置換）も多数加わるので、塩基配列レベルの差異はさらに大きくなり、それらの配列間での相同性を見出すのが実質的には不可能になってしまう。今回の実習では、同じ機能を持つ酵素を、配列相同性検索を用いて広範囲の微生物ゲノム配列から探し出そうとしているので、塩基配列を比較することは適切ではない。

1) 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」のなかの ARSA の項目をクリックする。

2) Quick Search で All Databases を対象に、得られている「薬剤や遺伝子やタンパク質の英語名」、並びに以下に説明する「目的に適合する検索を行うための英語の用語」を半角文字で記入して、「Search」を開始する。検索条件を複数指定する場合は、語句の間を半角空白で区切っておく。

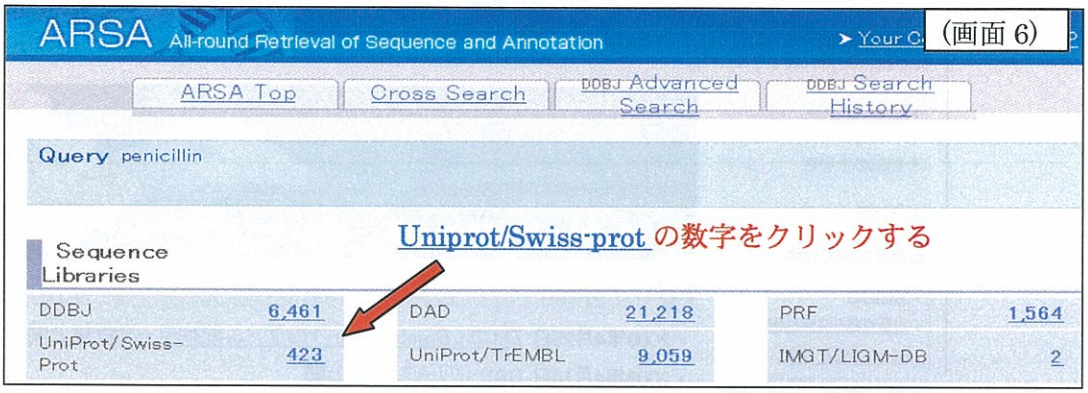
まず、“Google で得られた薬剤・物質の英語名”と“ライフサイエンス辞書で得られた薬剤・物質の英語名”の欄に着目し、両方の欄で同一の英語名が得られているものは信頼性が高いと推定して、この薬剤英語名を使用して、この薬剤を合成する酵素の既知のアミノ酸を配列を取得する（どちらかの欄にしかない場合は、ライフサイエンス辞書で得られた薬剤・物質の英語名の方から使用する）。テキストでは、抗生物質の「penicillin」をその薬剤の英語名の具体例として、このペニシリンを合成している遺伝子を探索する操作方法を紹介する。皆さんは、独自のテーマから探し出した薬剤・物質の英語名で探索を開始すればよい。試しに ARSA のキーワード検索で、“penicillin”と入力して、「Search」を開始してみると、



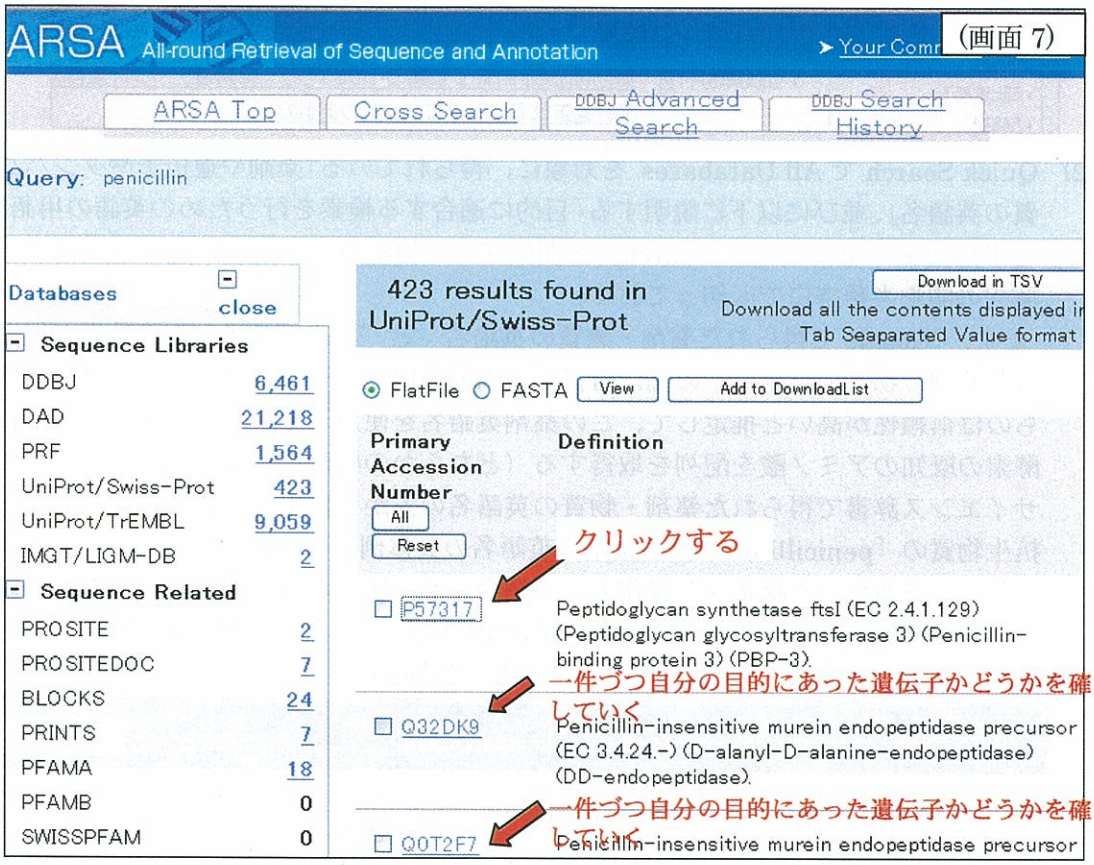
短時間の後に、wait の部分が数字に変化するが、それが All Databases に収録されている大量なデータを対象にして、検索された結果の件数である。

Uniprot/Swiss-prot はタンパク質アミノ酸配列を収録したデータベースを対象にし

た検索の結果である。この数字をクリックすると、ARSA の検索で見出された既知遺伝子のタンパク質の機能や特徴やアミノ酸配列が得られる。



検索された、Uniprot/Swiss-prot からのデータの一覧が表示される。この例では 423 件ものデータが検索された。



アクセッションナンバーをクリックすると、タンパク質の機能や特徴やアミノ酸配列が記載されたフラットファイルが表示される。

(画面 8)

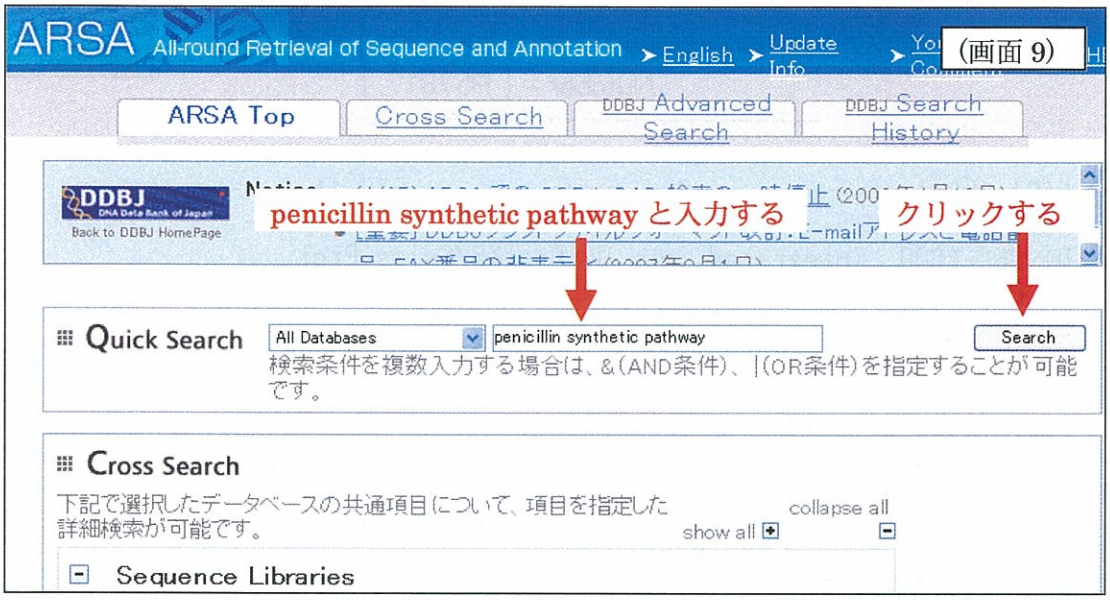
Number = [P57317]

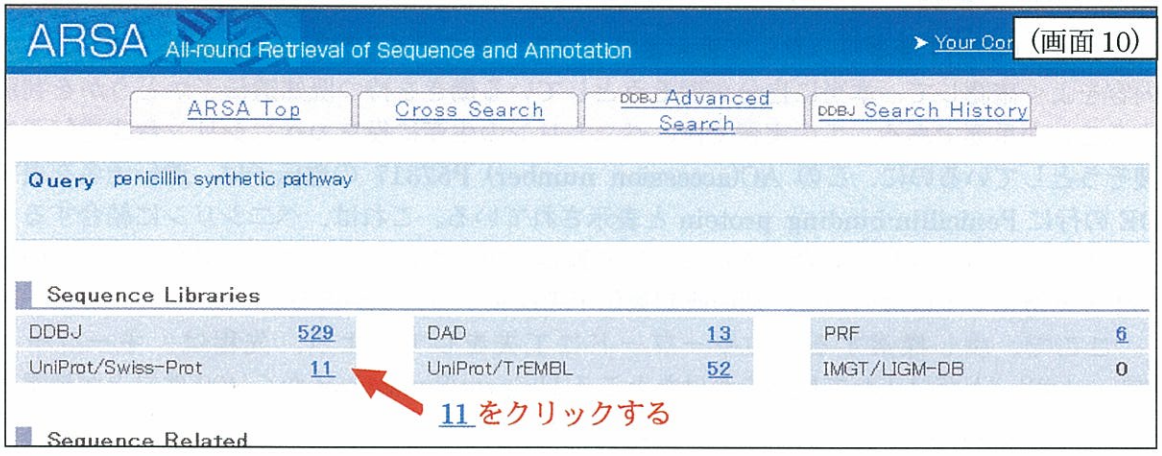
```
ID   FTSI_BUCAI              Reviewed;         579 AA.
AC   P57317;
DT   01-DEC-2000, integrated into UniProtKB/Swiss-Prot.
DT   01-DEC-2000, sequence version 1.
DT   29-APR-2008, entry version 48.
DE   Peptidoglycan synthetase ftsI (EC 2.4.1.129) (Peptidoglycan
DE   glycosyltransferase 3) (Penicillin-binding protein 3) (PBP-3).
GN   Name=ftsI; OrderedLocusNames=BU222;
OS   Buchnera aphidicola subsp. Acyrthosiphon pisum (Acyrthosiphon pisum
OS   symbiotic bacterium).
OC   Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
OC   Enterobacteriaceae; Buchnera.
OX   NCBI_TaxID=118099;
```

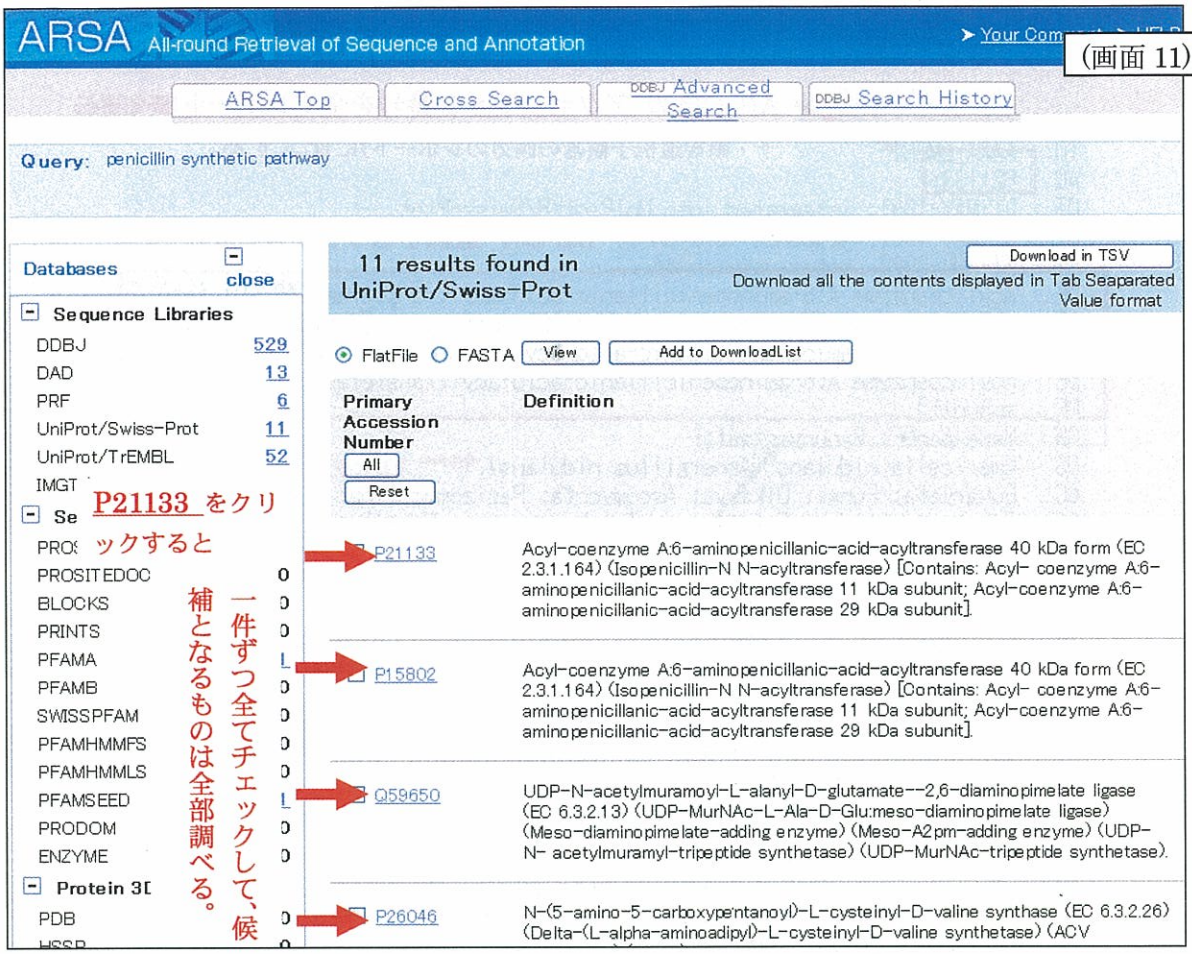
これは、P57317 をクリックして表示されるフラットファイルの上の方の画面である。この内容をよく確認して、**本当に自分が探そうとしている働きを持つ既知遺伝子かどうかを判別することが重要である。この実習では、ペニシリンの生産に役立つペニシリン合成遺伝子を捜そうとしているのに、この AC(accession number) P57317 の遺伝子は、遺伝子名を表す DE の行に Penicillin-binding protein と表示されている。これは、ペニシリンに結合するタンパク質で、ペニシリンの合成を行っている遺伝子ではない。このような場合は、自分の目的とする遺伝子ではないので、別の既知遺伝子を探索する必要がある。**

ここで、少し探索する際のキーワードに工夫をしてみよう。先程は、キーワードに"penicillin"とだけ入れて検索をかけたところ Uniprot/Swiss-prot から 400 件以上が検索された。その中にはどうも目的とは異なるとみられる遺伝子（例えばペニシリンが作用を及ぼす酵素の遺伝子）も多く含まれていた。**そこで、今度は "合成経路: synthetic pathway" というキーワードを加えて検索してみよう。**（他には" 生合成経路：biosynthetic pathway "や" 生合成：biosynthesis"というキーワードを加えてみてもよいだろう。）

これにより、ペニシリンの合成経路に関する、即ち中間代謝物の産生に役立つ遺伝子の候補が探索される可能性が高くなる。実際に行ってみると、Uniprot/Swiss-prot からは 11 件に絞られた（画面 9,10）。この 11 件について全て、1 件ずつ自分の目的とする遺伝子であるかどうかを確認していく。


ARSA All-round Retrieval of Sequence and Annotation
English
Update Info
(画面 9)
ARSA Top
Cross Search
DDBJ Advanced Search
DDBJ Search History
DDBJ DNA Data Bank of Japan
Back to DDBJ HomePage
penicillin synthetic pathway と入力する
クリックする
Quick Search
All Databases
penicillin synthetic pathway
Search
検索条件を複数入力する場合は、&(AND条件)、|(OR条件)を指定することが可能です。
Cross Search
下記で選択したデータベースの共通項目について、項目を指定した詳細検索が可能です。
show all
collapse all
Sequence Libraries


ARSA All-round Retrieval of Sequence and Annotation
(画面 10)
ARSA Top
Cross Search
DDBJ Advanced Search
DDBJ Search History
Query penicillin synthetic pathway
Sequence Libraries
DDBJ 529
DAD 13
PRF 6
UniProt/Swiss-Prot 11
UniProt/TrEMBL 52
IMGT/LIGM-DB 0
11 をクリックする
Sequence Related



(画面 11)

➢ 一つ目の P21133 をクリックすると、別 Window で次の画面が表示される。
自分の目的とする遺伝子かどうかを確認する。

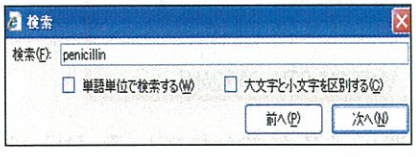
Number = [P21133]

(画面 12)

```
ID   AAAA_EMENI              Reviewed;         357 AA.
AC   P21133;
DT   01-MAY-1991, integrated into UniProtKB/Swiss-Prot.
DT   01-MAY-1991, sequence version 1.
DT   24-JUL-2007, entry version 44.
DE   Acyl-coenzyme A:6-aminopenicillanic-acid-acyltransferase 40 kDa form
DE   (EC 2.3.1.164) (Isopenicillin-N N-acyltransferase) [Contains: Acyl-
DE   coenzyme A:6-aminopenicillanic-acid-acyltransferase 11 kDa subunit;
DE   Acyl-coenzyme A:6-aminopenicillanic-acid-acyltransferase 29 kDa
DE   subunit].
GN   Name=penDE; Synonyms=aat;
OS   Emericella nidulans (Aspergillus nidulans).
OC   Eukaryota; Fungi; Dikarya; Ascomycota; Pezizomycotina; Eurotiomycetes;

RA   Schofield C.J., Sutherland J.D., Willis A.C.;
RT   "Acyl coenzyme A: 6-aminopenicillanic acid acyltransferase from
RT   Penicillium chrysogenum and Aspergillus nidulans.";
RL   FEBS Lett. 262:342-344(1990).
CC   -!- FUNCTION: Last enzyme in penicillin biosynthetic pathway, which
CC       converts isopenicillin N (IPN) to penicillin G, using phenyl-
CC       acetyl-CoA or phenoxyacetyl-CoA as acyl donors.
CC   -!- CATALYTIC ACTIVITY: Phenylacetyl-CoA + isopenicillin N + H(2)O =
CC       CoA + penicillin G + L-2-aminohexanedioate.
CC   -!- PATHWAY: Antibiotic biosynthesis; penicillin G biosynthesis;
CC       penicillin G from L-alpha-aminoadipate and L-cysteine and L-
CC       valine: step 3/3.
```

※「Ctrl」キーと「f」キーを同時に押すと、次のような小さな検索画面が表示される。ここに探したいキーワード (ここでは penicillin) を入れて探索すると便利である。



キーワード等を入力する際には、コピー&ペーストを利用して入力すると間違いがない。コピーしたい文字等をマウスで選択後に、キーボードの「Ctrl」キーを押しながら「c」キーを押し（これでコピー出来る）、次に貼り付けたいところにカーソルを移動させて、「Ctrl」キーを押しながら「v」キーを押すと、貼り付けが完了する。

Number = [P21133]　(画面 13)

AC の行の アクセッション番号を全てレポート(既知遺伝子・新規遺伝子候補の両方のレポート)に記入する。

DE の行 遺伝子名：全て既知遺伝子レポートに記入する。

OS の行　生物種

遺伝子の略号：GN の行を全て、既知遺伝子レポートに記入する。

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）

```
ID   AAAA_EMENI
AC   P21133;
DT   01-MAY-1991, integrated into UniProtKB/Swiss-Prot.
DT   01-MAY-1991, sequence version 1.
DT   24-JUL-2007, entry version 44.
DE   Acyl-coenzyme A:6-aminopenicillanic-acid-acyltransferase 40 kDa form
DE   (EC 2.3.1.164) (Isopenicillin-N N-acyltransferase) [Contains: Acyl-
DE   coenzyme A:6-aminopenicillanic-acid-acyltransferase 11 kDa subunit;
DE   Acyl-coenzyme A:6-aminopenicillanic-acid-acyltransferase 29 kDa
DE   subunit].
GN   Name=penDE; Synonyms=aat;
OS   Emericella nidulans (Aspergillus nidulans).
OC   Eukaryota; Fungi; Dikarya; Ascomycota; Pezizomycotina; Eurotiomycetes;
OC   Eurotiomycetidae; Eurotiales; Trichocomaceae; Emericella.
OX   NCBI_TaxID=162425;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RX   MEDLINE=90340281; PubMed=2166227; DOI=10.1007/BF00259395;
RA   Montenegro E., Barredo J.L., Gutierrez S., Diez B., Alvarez E.,
RA   Martin J.F.;
RT   "Cloning, characterization of the acyl-CoA:6-amino penicillanic acid
RT   acyltransferase gene of Aspergillus nidulans and linkage to the
RT   isopenicillin N synthase gene.";
RL   Mol. Gen. Genet. 221:322-330(1990).
RN   [2]
RA   Schofield C.J., Sutherland J.D., Willis A.C.;
RT   "Acyl coenzyme A: 6-aminopenicillanic acid acyltransferase from
RT   Penicillium chrysogenum and Aspergillus nidulans.";
RL   FEBS Lett. 262:342-344(1990).
CC   -!- FUNCTION: Last enzyme in penicillin biosynthetic pathway, which
CC       converts isopenicillin N (IPN) to penicillin G, using phenyl-
CC       acetyl-CoA or phenoxyacetyl-CoA as acyl donors.
CC   -!- CATALYTIC ACTIVITY: Phenylacetyl-CoA + isopenicillin N + H(2)O =
CC       CoA + penicillin G + L-2-aminohexanedioate.
CC   -!- PATHWAY: Antibiotic biosynthesis; penicillin G biosynthesis;
CC       penicillin G from L-alpha-aminoadipate and L-cysteine and L-
CC       valine: step 3/3.
CC   -!- PTM: The pre-AAT protein is probably synthesized as 40 kDa
CC       precursor which is then processed into an 11 kDa (protein A) and a
CC       29 kDa (protein B). The B protein carries AAT activity.
CC   -!- SIMILARITY: Belongs to the peptidase C45 family.
PE   1: Evidence at protein level;
KW   Acyltransferase; Antibiotic biosynthesis; Direct protein sequencing;
KW   Transferase; Zymogen.
FT   CHAIN         1    357       Acyl-coenzyme A:6-aminopenicillanic-acid-
FT                                acyltransferase 40 kDa form.
FT                                /FTId=PRO_0000020592.
FT   CHAIN         1    102       Acyl-coenzyme A:6-aminopenic
FT                                acyltransferase 11 kDa subun
FT                                /FTId=PRO_0000020593.
FT   CHAIN       103    357       Acyl-coenzyme A:6-aminopenic
FT                                acyltransferase 29 kDa subun...
FT                                /FTId=PRO_0000020594.
SQ   SEQUENCE   357 AA;  39236 MW;  89129F003CA0A00C CRC64;
     MLHVTCQGTP SEIGYHHGSA AKGEIAKAID FATGLIHGKT KKTQAELEQL LRELEQVMKQ
     RWPRYYEEIC GIAKGAEREV SEIVMLNTRT EFAYGLVEAR DGCTTVYCKT PNGALQGQNW
     DFFTATKENL IQLTICQPGL PTIKMITEAG IIGKVGFNSA GVAVNYNALH LHGLRPTGLP
     SHLALRMALE STSPSEAYEK IVSQGGMAAS AFIMVGNAHE AYGLEFSPIS LCKQVADTNG
     RIVHTNHCLL NHGPSAQELN PLPDSWSRHG RMEHLLSGFD GTKEAFAKLW EDEDNYPLSI
     CRAYKEGKSR GSTLFNIVFD HVGRKATVRL GRPNNPDETF VMTFSNLDTK SAIQANI
//
```

➢　一つ目の P21133 の情報を詳細に見ると、遺伝子名が書かれている DE の行に、

Isopenicillin-N N-acyltransferase（ペニシリンNと同種 N-アシル基転移酵素）と記載されている。これは、テーマ検索レポートで調べた“Google で得られた遺伝子やタンパク質名”の中の“アシルトランスフェラーゼ”、並びに“ライフサイエンス辞書で得られた遺伝子やタンパク質の英語名”の“acyltransferase”に合致していると考えられる。また CC 行の FUNCTION（機能）には Last enzyme in penicillin biosynthetic pathway･･･（ペニシリン生合成経路における最後の酵素･･･）と書かれている。これらのことから、この遺伝子はペニシリンの生合成経路において最後のほうで機能している酵素遺伝子であると判断出来る（英語の専門用語の意味はライフサイエンス辞書で調べると良い）。よってこのアミノ酸配列を問合わせ配列(クエリー配列)として選択し、次にメタゲノム配列の中から相同性の高い配列を探索していくことになる。2.3 の項でも指摘したように、一種類の抗生物質を合成するのにも、通常は複数種類の酵素が関与している、それらのいずれもが、今回の検索対象となるので、既知遺伝子レポートへ記載しておく。

画面 12・13 で OS の行を確認しておくことも重要である。OS の行は、このクエリー配列がどの生物種に由来しているかが記載されている。同一の機能を持つ酵素であっても、生物種によってそのアミノ酸配列は異なるのが一般的であり、それらはタンパク質アミノ酸配列のデータベースでは異なったアクセション番号で生物種別に収録されている。一つの生物種が同じ機能を持つ複数の酵素を生産している場合もあるが、アミノ酸配列に差があれば、それらも異なったアクセッション番号で収録されている。この OS の行に書かれた生物種がヒトや高等動植物である場合は、その薬剤が作用する側のタンパク質である可能性が高い（英語の生物種名もライフサイエンス辞書で調べると良い）。これらは目的外の遺伝子なので、レポートに記載してはいけない。

ここで、画面 13 の一番左側の主な略号について説明を付しておく。

AC：データのアクセッション番号，　DE：遺伝子名　，GN：遺伝子の略号，
OS：生物種名，　RN：出典となった論文等，　RA：著者名，　RT：論文名，
RL：雑誌名，　RX：論文の MEDLINE#，　DR：他のデータベースでのアクセッション番号，　KW：検索で参照されるキーワード，　FT：機能，SQ：配列情報

➢ 次に（画面 11）の２つ目の、P15802　をクリックして同じように目的とする有用な遺伝子かどうかを確認する。

Number = [P15802]　(画面 14)

```
ID   AAAA_PENCH              Reviewed;         357 AA.
AC   P15802;
DT   01-APR-1990, integrated into UniProtKB/Swiss-Prot.
DT   01-APR-1990, sequence version 1.
DT   08-APR-2008; entry version 50.
DE   Acyl-coenzyme A:6-aminopenicillanic-acid-acyltransferase 40 kDa form
DE   (EC 2.3.1.164) (Isopenicillin-N N-acyltransferase) [Contains: Acyl-
DE   coenzyme A:6-aminopenicillanic-acid-acyltransferase 11 kDa subunit;
DE   Acyl-coenzyme A:6-aminopenicillanic-acid-acyltransferase 29 kDa
DE   subunit].
GN   Name=penDE; Synonyms=aat;
OS   Penicillium chrysogenum (Penicillium notatum).
OC   Eukaryota; Fungi; Dikarya; Ascomycota; Pezizomycotina; Eurotiomycetes;
```

```
RA   Aplin R.T., Baldwin J.E., Roach P.L., Robinson C.V., Schofield C.J.;
RT   "Investigations into the post-translational modification and mechanism
RT   of isopenicillin N:acyl-CoA acyltransferase using electrospray mass
RT   spectrometry.";
RL   Biochem. J. 294:357-363(1993).
CC   -!- FUNCTION: Last enzyme in penicillin biosynthetic pathway, which
CC       converts isopenicillin N (IPN) to penicillin G, using phenyl-
CC       acetyl-CoA or phenoxyacetyl-CoA as acyl donors.
CC   -!- CATALYTIC ACTIVITY: Phenylacetyl-CoA + isopenicillin N + H(2)O =
CC       CoA + penicillin G + L-2-aminohexanedioate.
CC   -!- PATHWAY: Antibiotic biosynthesis; penicillin G biosynthesis;
CC       penicillin G from L-alpha-aminoadipate and L-cysteine and L-
CC       valine: step 3/3.
```

この遺伝子もペニシリンの生合成経路において最後のほうで機能する酵素遺伝子であると判断出来る。よってこのアミノ酸配列を問合わせ配列(クエリー配列)として、次にメタゲノム配列の中から相同性の高い配列を探索していくことになる。

➢ 次に（画面 11）の 3 つ目の、Q59650　をクリックして同じように目的とする有用な遺伝子かどうかを確認する。

（画面 15）

```
Number = [ Q59650 ]

ID   MURE_PSEAE              Reviewed;         487 AA.
AC   Q59650;
DT   01-NOV-1997, integrated into UniProtKB/Swiss-Prot.
DT   08-DEC-2000, sequence version 2.
DT   29-APR-2008, entry version 64.
DE   UDP-N-acetylmuramoyl-L-alanyl-D-glutamate--2,6-diaminopimelate ligase
DE   (EC 6.3.2.13) (UDP-MurNAc-L-Ala-D-Glu:meso-diaminopimelate ligase)
DE   (Meso-diaminopimelate-adding enzyme) (Meso-A2pm-adding enzyme) (UDP-N-
DE   acetylmuramyl-tripeptide synthetase) (UDP-MurNAc-tripeptide
DE   synthetase).
GN   Name=murE; OrderedLocusNames=PA4417;
OS   Pseudomonas aeruginosa.
OC   Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales;
OC   Pseudomonadaceae; Pseudomonas.
```

```
RX   MEDLINE=96100768; PubMed=7486937;
RA   Liao X., Hancock R.E.W.;
RT   "Cloning and characterization of the Pseudomonas aeruginosa pbpB gene
RT   encoding penicillin-binding protein 3.";
RL   Antimicrob. Agents Chemother. 39:1871-1874(1995).
CC   -!- FUNCTION: Catalyzes the addition of meso-diaminopimelic acid to
CC       the nucleotide precursor UDP-N-acetylmuramoyl-L-alanyl-D-glutamate
CC       (UMAG) in the biosynthesis of bacterial cell-wall peptidoglycan
CC       (By similarity).
CC   -!- CATALYTIC ACTIVITY: ATP + UDP-N-acetylmuramoyl-L-alanyl-D-
CC       glutamate + meso-2,6-diaminoheptanedioate = ADP + phosphate + UDP-
CC       N-acetylmuramoyl-L-alanyl-D-gamma-glutamyl-meso-2,6-diamino-
CC       heptanedioate.
CC   -!- PATHWAY: Cell wall biosynthesis; peptidoglycan biosynthesis.
```

この Q59650 番号を持つ遺伝子は、論文名を表す RT の行に encoding penicillin-binding protein と表示されている。これは、ペニシリンに結合するタンパク質の合成に関与している遺伝子で、ペニシリンの合成を行っている遺伝子ではない。このような場合は、自分の目的とする遺伝子ではないので、このアミノ酸配列を問合わせ配列(クエリー配列)としてはいけない。（画面 9）で **“penicillin synthetic pathway”** とキーワードを入れているにもかかわらず、このようにペニシリンの合成を行っている遺伝子ではないものが検索されることもあるので、手間ではあるが全件チェックを行うことが必須である。

このようにして、（画面 11）で検索された 11 件の遺伝子について一件ずつ全てチェックし、自分の目的に合う遺伝子のみを選別し、そのアミノ酸配列を問合わせ配列(クエリー配列)として使用すること。ここで、目的にあわない遺伝子を選択してしまうと以後の作業は全く意味の無い検索となるので、慎重にチェックを行うこと。

➢ （画面 11）で検索された 11 件の遺伝子について一件ずつ全てチェックを行うと、最終的には、11 件のうち 7 件（P21133, P15802, P26046, P25464, P27742, P27743,

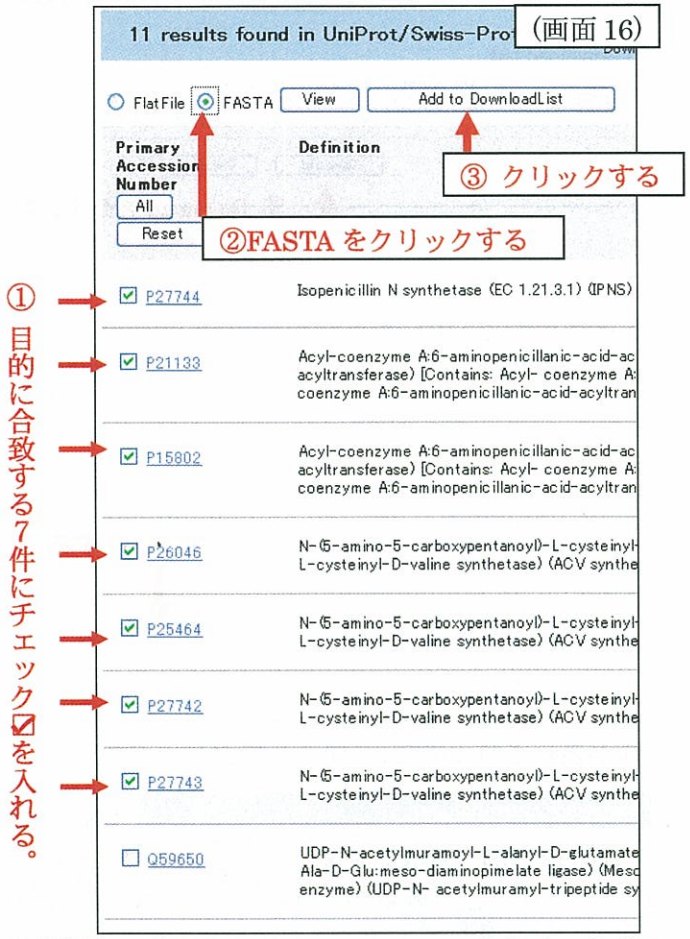
P27744）が目的に合致する既知遺伝子であることがわかる。

今回は、７件に絞れたので全てのアミノ酸配列を取得して、そのアミノ酸配列を問合わせ配列(クエリー配列)として、次に進めばよい。しかし、物質名によっては絞り込んでも 100 件を超えたりする場合もあるだろう。その中には非常に近縁で、配列も数箇所違うだけで、残りは全部同じといったものもあるに違いない。アミノ酸配列が非常に似ているものを、それぞれ別個に問合わせ配列(クエリー配列)として使用しても、相同性検索でヒットしてくる配列は同じものである可能性が高い。そこで、系統樹を作成して非常に近縁の配列については、その中の１つだけを代表で選ぶことにより、問合わせ配列(クエリー配列)を絞り込むほうがよい。

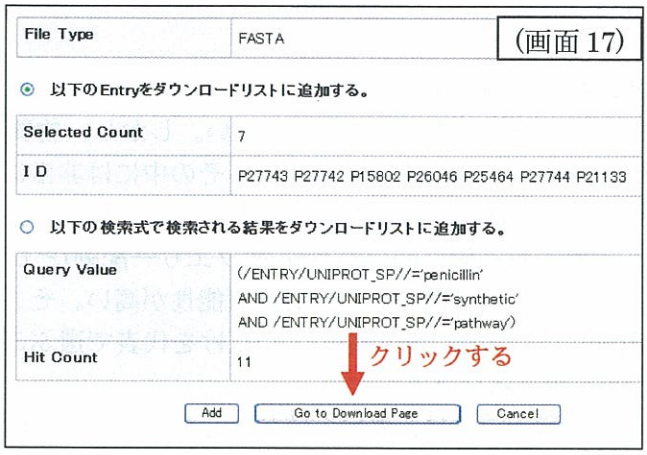
ここでは操作の説明の為に、先程の絞り込んだ７件を、系統樹を作成して非常に近縁なものについては代表のみとして、更に絞込みをかける方法を紹介する。

2.6 CLUSTALW による非常に近縁な配列の見つけ方

（画面 11）でペニシリンの合成を行っている遺伝子と判断できた遺伝子についてのみ☑を入れる。ここでは７件が該当した。

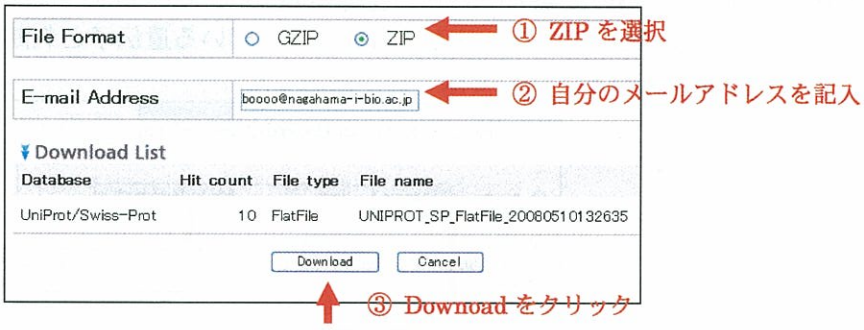


FASTA を選択し、「Add to Download List」をクリックする。
次の画面が表示されるので、「Go to Download Page」をクリックする。

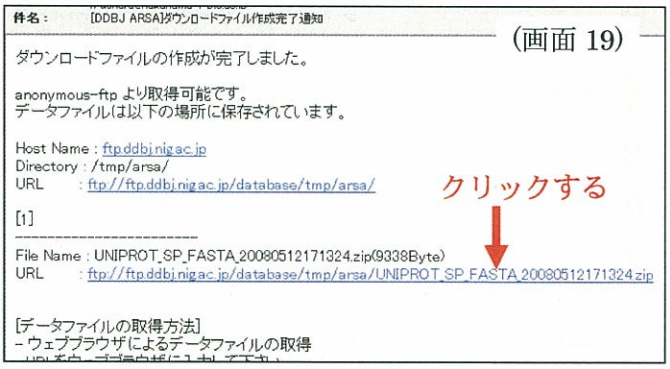


次の画面で、windows パソコンの場合は「ZIP」を選択し、自分のメールアドレスを記入し、最後に「Download」ボタンをクリックする。

(画面 18)



しばらくすると、次のメールが送られてくる。
ZIP ファイルのある URL をクリックする。



次の画面が表示されるので、「開く」をクリックする。

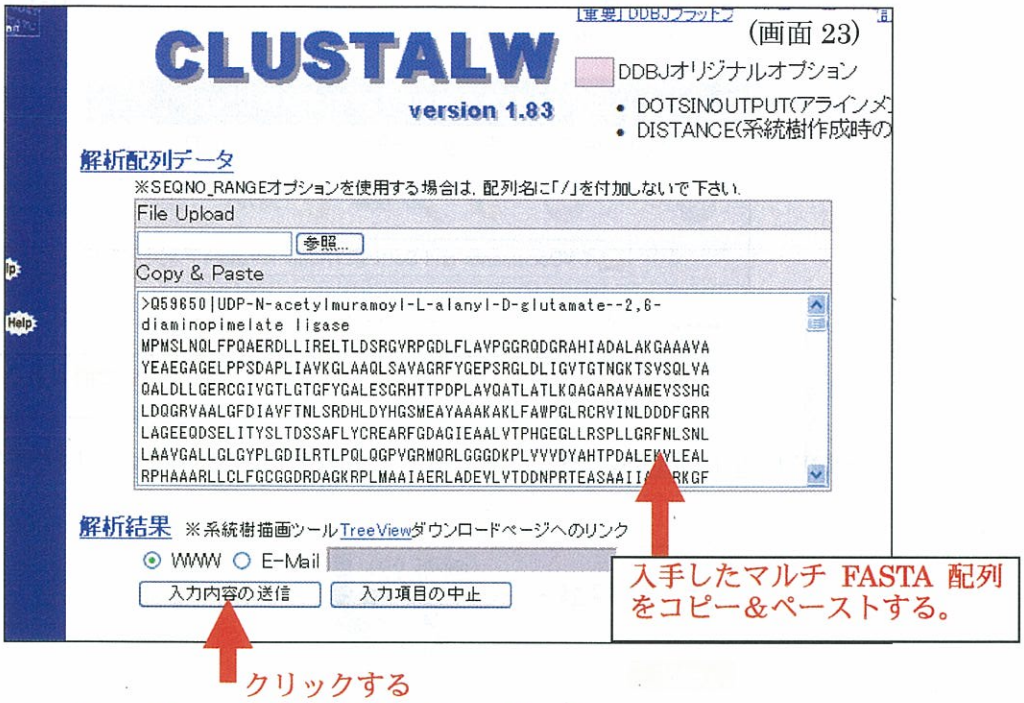
デスクトップに、マルチ FASTA ファイルが出来る。
（fasta 形式は”＞”で始まる行に配列の名前を書き、2 行目からアミノ配列/DNA 配列を記す形式で、マルチ FASTA 形式は、１つのファイルの中に FASTA 形式のファイルが複数連結されている形式である。）

このファイルを開く際は、ワードパット，ノートパッド，SAKURA 等のテキストファイルで開くとよい。

(画面 21)

次に、DDBJ（http://www.ddbj.nig.ac.jp/）の系統解析メニューの CLUSTALW を開ける。

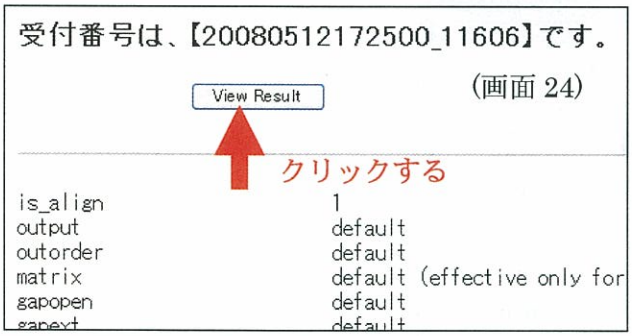
(画面 22)



(画面 23)

URL: http://clustalw.ddbj.nig.ac.jp/top-j.html

次の画面の「View Result」をクリックする。

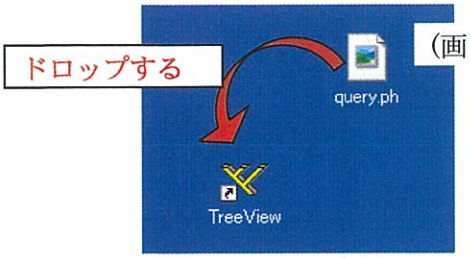


(画面 24)

表示される画面の下のほうの query.ph をクリックし、ファイルをデスクトップに保存する。

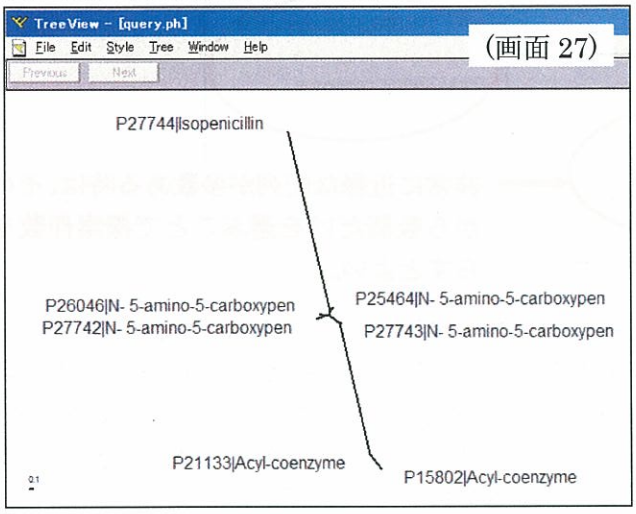
(画面 25)

デスクトップ上で、保存した query.ph ファイルを※「Tree View」にドロップする。



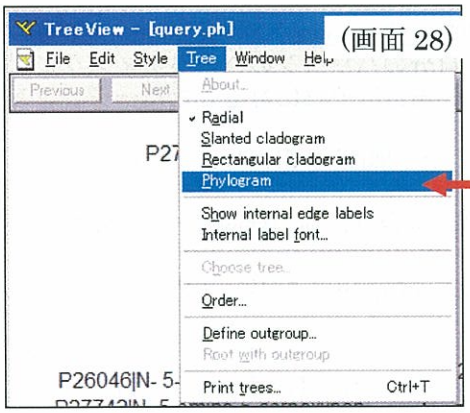
(画面 26)

次のような画面が表示される。



(画面 27)

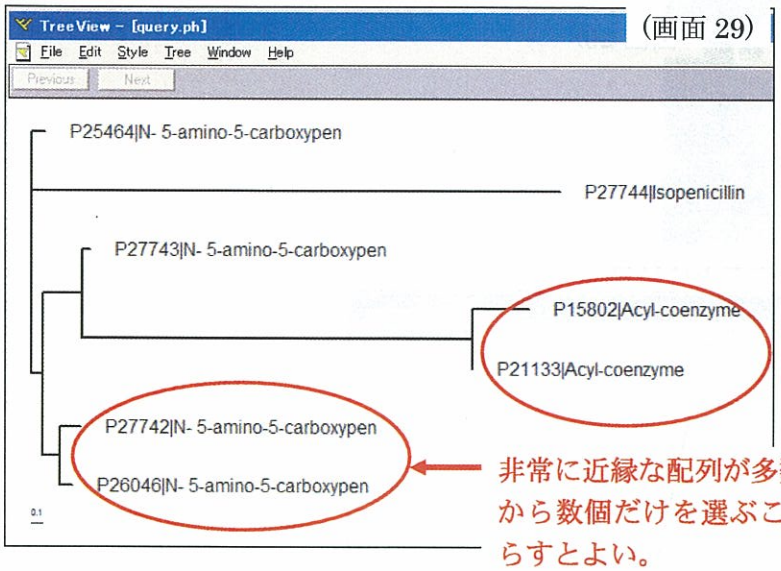
Tree の中の「Phylogram（系統樹）」を選択する。



(画面 28)

次のような系統樹が表示される。

この例では、数が少ないのであまり意味はないが、多数の遺伝子を絞り込む際には、非常に近縁な配列間では、そのアミノ酸配列も大半が同じであると考えられるので、近縁の配列の中からは代表で数個を選択するとよい。



（画面 29）

非常に近縁な配列が多数ある時は、その中から数個だけを選ぶことで探索件数を減らすとよい。

※ 長浜バイオ大学の情報実習室のパソコンには既に Tree View（ClustalW で解析して得られた系統樹の数値情報を図示するためのフリーソフト）がインストールされており、デスクトップに表示されているので、それを使用すればよい。長浜バイオ大学外のパソコンで Tree View がインストールされていない場合は、次の要領でダウンロードすればよい。

1．"TreeView"のホームページにアクセスする。
http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

2．ダウンロードする。
ホームページ下部の Win32 (Windows 95 or Windows NT) (version 1.6.6)をクリックして、"treev32.zip"をダウンロードする（自分のパソコンのハードディスクに保存する）。
（注）WindowsXP 以外（マックなど）の場合はそれに対応したバージョンを選択する。

3．展開（解凍）
保存した"treev32.zip"を右クリックして″全て展開″を選択し、「展開ウィザード」を起動してその指示に従って展開する。
（注１）zip ファイルが展開できない環境の場合は、別の展開ツールを用いる。

4．インストール
展開されてできたフォルダ″treev32″内の″setup.exe″を実行（ダブルクリック）して、インストールを開始する。

5．インストール完了
正常にインストールされれば、スタートメニューに″TreeView″が現れる。

2.7　レポート様式への情報の記入

皆さんは、自分が選択した「薬剤・物質名」について、上記の手順でその薬剤を合成する酵素のアミノ酸配列を取得する。確信を持てた配列が取得できたら、既知遺伝子レポートへ記入をする。注意すべき点は、アクセション番号ごとに（言い換えれば画面 13 の画面ごとに）、エクセルファイルの新しい一行を使用する。まず、エクセルファイルの新しい一行に、遺伝子検索に用いる薬剤・物質名をコピー＆ペーストで、「遺伝子検索用に使用する薬剤・物質名」欄に記載し、英語名も「遺伝子検索用に使用する薬剤・物質の英語名」欄に際する。次に複数のアクセション番号が、得られている場合は、アクセション番号ごとに、Flat File から得られた情報や配列情報を、レポートの対応する欄へコピー＆ペーストの機能で記載をする。繰り返すが 1 画面情報（画面 13）につき 1 行を使う点は重要であり、各行ごとに遺伝子検索に使用する薬剤・物質名を記載しておく。この画面 13 からのコピー＆ペーストは、コピーしたい領域をマウスで選択後に、キーボードの「Ctrl」キーを押しながら「c」キーを押し（これでコピー出来る）、次に既知遺伝子レポートのエクセルファイルの貼り付けたいセルをマウスで選択して、「Ctrl」キーを押しながら「v」キーを押すと、貼り付けが完了する。

提出するレポートは、3 種類（テーマ検索レポート，既知遺伝子レポート，新規遺伝子候補レポート）ある。

- **テーマ検索レポートは、「健康に貢献が期待できる遺伝子やタンパク質」の候補選びに関するレポート。**
- **既知遺伝子レポートは、「健康への貢献が期待できる遺伝子やタンパク質」の候補についての既知遺伝子に関するレポート。**
- **新規遺伝子候補レポートは、目的とする既知遺伝子と高い相同性が得られた、新規探索遺伝子候補に関するレポート。**

「薬剤・物質の英語名」での検索を完了したら、遺伝子名やタンパク質名が書かれている DE の行の内容と、「Google で得られた遺伝子やタンパク質の英語名」とを見比べる。DE の行を見ても分かるように、同一の遺伝子やタンパク質でも色々な名前で呼ばれている。その内のどれかが、Google で得られた英語名と一致していると安心できる。一致した Google 側の英語名があれば、そのアクセション番号の記載されているエクセルファイルの行の「Google で得られた遺伝子やタンパク質の英語名」の欄に、その一致した英語名をコピー＆ペーストの機能で記載しておく。(画面からのコピー＆ペーストは、コピーしたい領域をマウスで選択後に、キーボードの「Ctrl」キーを押しながら「c」キーを押し（これでコピー出来る）、次に既知遺伝子レポート-2 のエクセルファイルの貼り付けたいセルをマウスで選択して、「Ctrl」キーを押しながら「v」キーを押すと、貼り付けが完了する。）それにしても、同一の酵素が色々な名称を持つのは余りにも不便である。大半の酵素には、EC 番号が付されており、DE の行に記載されている。この番号が一致すれば、確実性はさらに高まる。

「Google で得られた遺伝子やタンパク質の英語名」に自信があるのに、ARSA の「遺伝子検索用に使用する薬剤・物質の英語名」での検索では見つからない場合も存在する。その場合には、ARSA の Key Word に「Google で得られた遺伝子やタンパク質の英語名」を用いると良い。“遺伝子やタンパク質の英語名”に加えて、“biosynthesis”や“synthetic pathway”のような用語も追加すると安全性が高まる。但し、着目の“遺伝子やタンパク質”を微生物だけでなく、高等動植物も持っている場合には事情が複雑になる。OS の行に書かれた生物種がヒトや高等動植物である場合は、高等動植物が持つ酵素で、微生物の酵素とはアミノ酸配列が大きくかけ離れている可能性が高いので、レポートには記載しない。生物種の英語名もライフサイエンス辞書“Web LSD”（http://lsd.pharm.kyoto-u.ac.jp/ja/service/weblsd/index.html）で和名を調べるとよい。

3．環境由来 DNA 断片配列からの、健康への貢献遺伝子の新規探索

3.1 取得したアミノ酸配列と環境由来ＤＮＡ配列との相同性検索

得られた既知アミノ酸配列情報を用いて、環境由来データベースを対象に相同性検索を行う。

まず、BLAST を立ち上げる。ただし今回は、長浜バイオ大学のサーバー内に準備したものを使う。http://hpc31.nagahama-i-bio.ac.jp/blast/blast.html

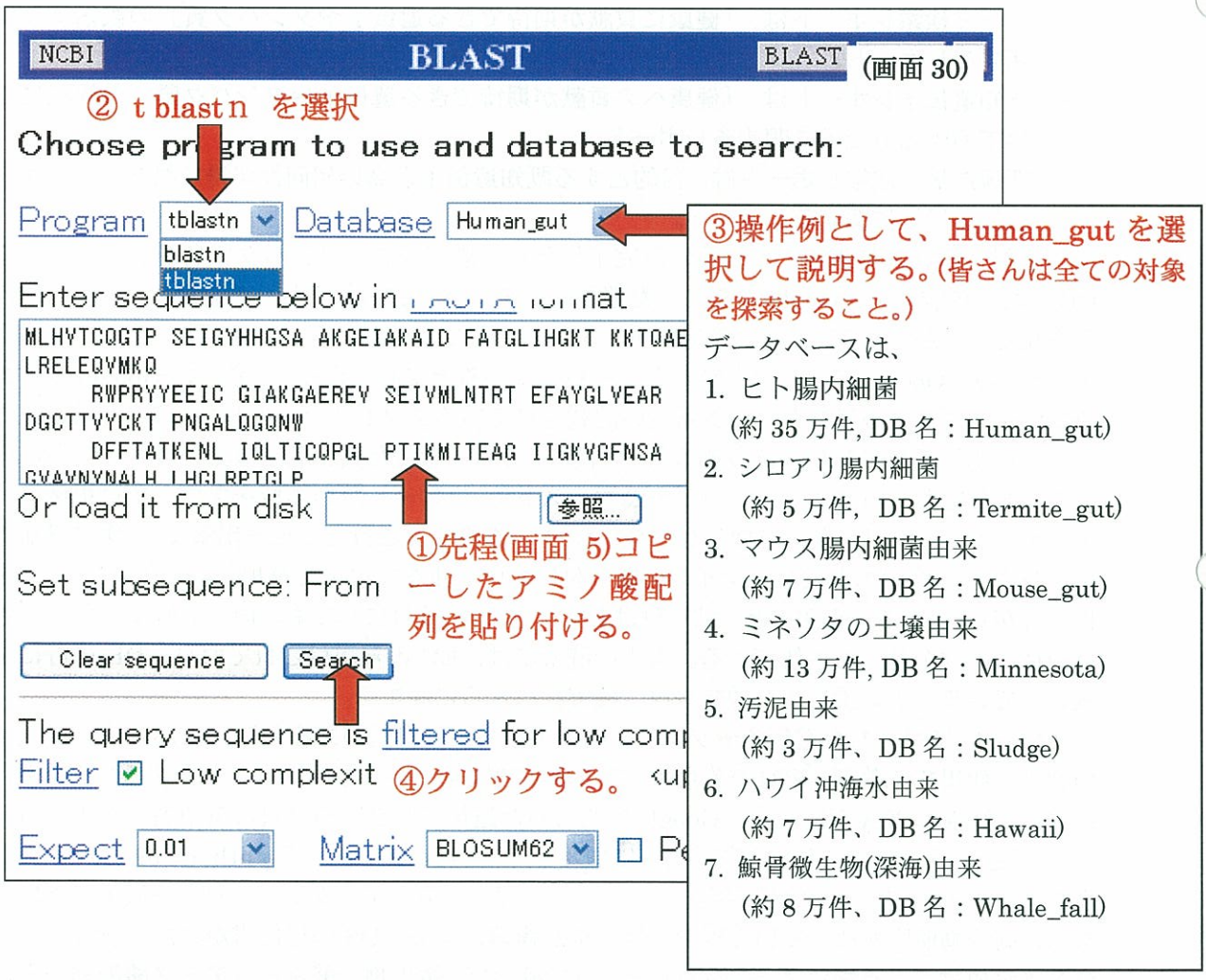
長浜バイオ大学のキャンパス外で作業を行う場合には、後の 3.4.（P36，画面 51）に記載した NCBI の BLAST を利用するとよい。

① 先程 (画面 13) コピーしたアミノ酸配列を貼り付ける。

② Program は「 **tblastn** 」を選択する。(BLAST の検索プログラムについては、副教材「基礎と実習バイオインフォマティクス」P43 の表 2. 6 を参考にするとよい。)

③ データベースは、ここでは Human_gut.fasta (ヒト腸内細菌) を選択したもので説明をする。(皆さんは全てを対象に探索を行い、多くの新規探索遺伝子候補を探すこと。)

④ Search をクリックする。



次のような画面が表示される。click here to format the BLAST-Output...をクリックする。
実行結果が表示される。

(画面 31)

Your query was send into the batch system.

Your query: blast7291859604
click here to format the BLAST-Output. ← クリックする。

(画面 32)

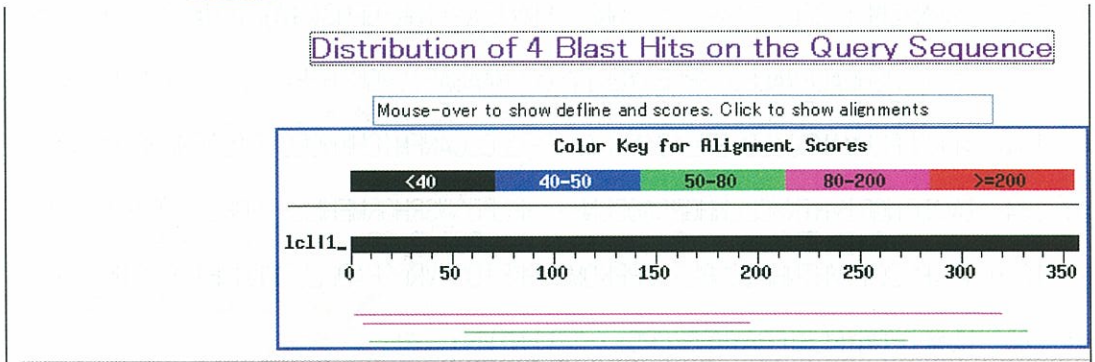
NCBI **BLAST Search Results** BLAST Entrez

TBLASTN 2.2.11 [Jun-05-2005]

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.

Database: Humangut.fasta
353,805 sequences; 462,989,954 total letters

Query= (357 letters) クエリー配列全長：カバー率を求める計算で使う。



新規遺伝子候補レポートの“ACCESSION#”に記入する

Sequences producing significant alignments:	Score (bits)	E Value
BABF01001330-3069	140	1e-32
BAAZ01030430-1027	108	6e-23
BAAU01012000-1118	71	1e-11
BABD01016542-996	65	8e-10

クリックする

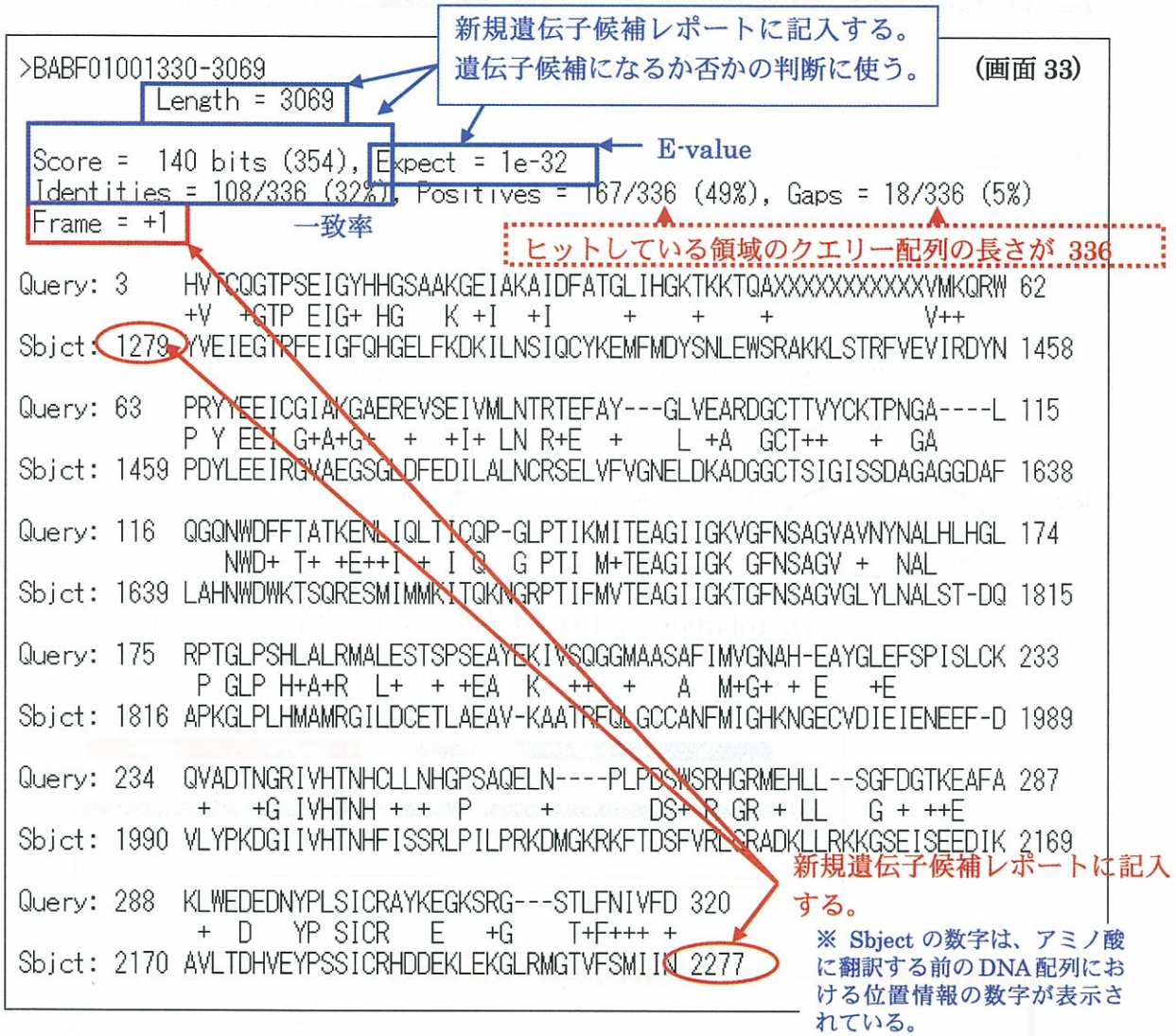
Accession 番号　　配列長 (bp)

・ 配列 ID は、[Accession 番号―配列長 (bp)] で記載されている。
・この例では、４件が配列相同性検索の結果ヒットした。これらはペニシリン生合成経路において機能している酵素遺伝子と同様な機能を持っている可能性がある。皆さんはこの４件全てについて、目的とする遺伝子候補としての条件を満たすか否かを次に確認していく。

ここでは、トップヒット[BABF01001330]の場合を例に説明を行う。Score 140 をクリックする。次の画面が表示される。

この画面から、ヒットしている領域を確認する。画面中の、Frame＝＋１，塩基位置1279..2277（後の Orf Finder で領域を選択する際に重要な情報となるので、新規遺伝子候補レポートに控えておくこと）でヒットしていることがわかる。

クエリーに対して、ほぼ全長をカバーしているので、[BABF01001330]はペニシリン生合成経路において機能している酵素遺伝子の可能性が高いとみられる。



遺伝子候補とする条件を、次のようにする。

① E-value: 1e-10 以下

② Identities（一致率）：３０％以上

③ カバー率(ヒットしている領域の)：５０％以上

全てを満たすものについて、遺伝子候補と判断しよう。

(画面 33) に戻り、遺伝子候補としての条件を満たしているか確認しておこう。

- Evalue＝1e-32　で条件を満たす。
- Identities（一致率）＝ 32% で条件を満たす。
- **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　336／357 ＝ 0.94 (94%) で条件を満たす。

全ての条件を満たしているので、目的遺伝子候補とする。

これで、環境由来(今回はヒト腸内細菌)ゲノムを対象にして、ペニシリンをテーマにして、ペニシリンの生合成経路において機能する酵素 Acyl-coenzyme の既知アミノ酸配列と、相同性の高い環境由来遺伝子候補が検索出来た。

続いて、遺伝子領域の確定をしていこう。

3.2 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、更に詳しく遺伝子領域を確定してみよう。

ＮＣＢＩを立ち上げる。　http://www.ncbi.nlm.nih.gov/

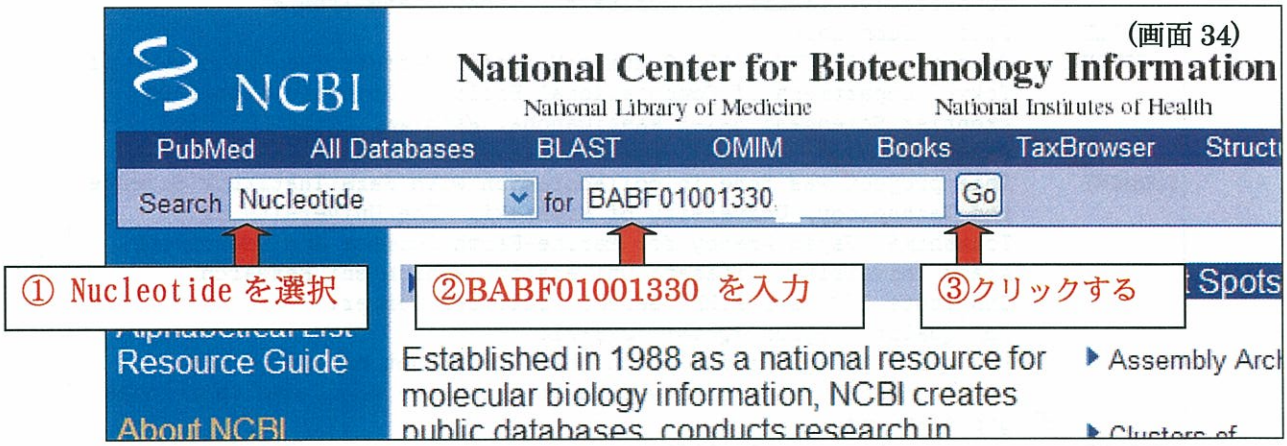
先程のアクセッション番号 BABF01001330 のアミノ酸 fasta データを得る。

① Search で、Nucleotide を選択する。今までのアミノ酸配列の解析とは違い、DNA 配列であることに注意すること。

② For には、BABF01001330　と入力する。

注）長浜バイオ大学ローカル BLAST で表示された、BABF01001330-3069　の -3069 の部分は配列長(bp)を表しているので、アクセッション番号としては、BABF01001330 のみを入力する。

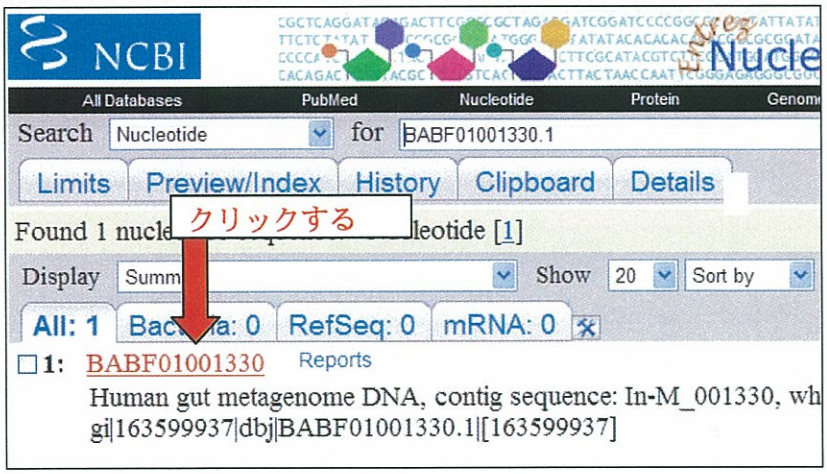
(画面 34)



次の検索結果が表示される。

表示される Flat File　の中には、配列の取得情報も記載されている。どんな環境由来かを確認し、所定様式のレポートにも記入しておこう。Flat File についての説明は（付録２）を参照するとよい。次に FASTA ファイル(DNA 配列)を入手しよう。

(画面 35)



(画面 36)

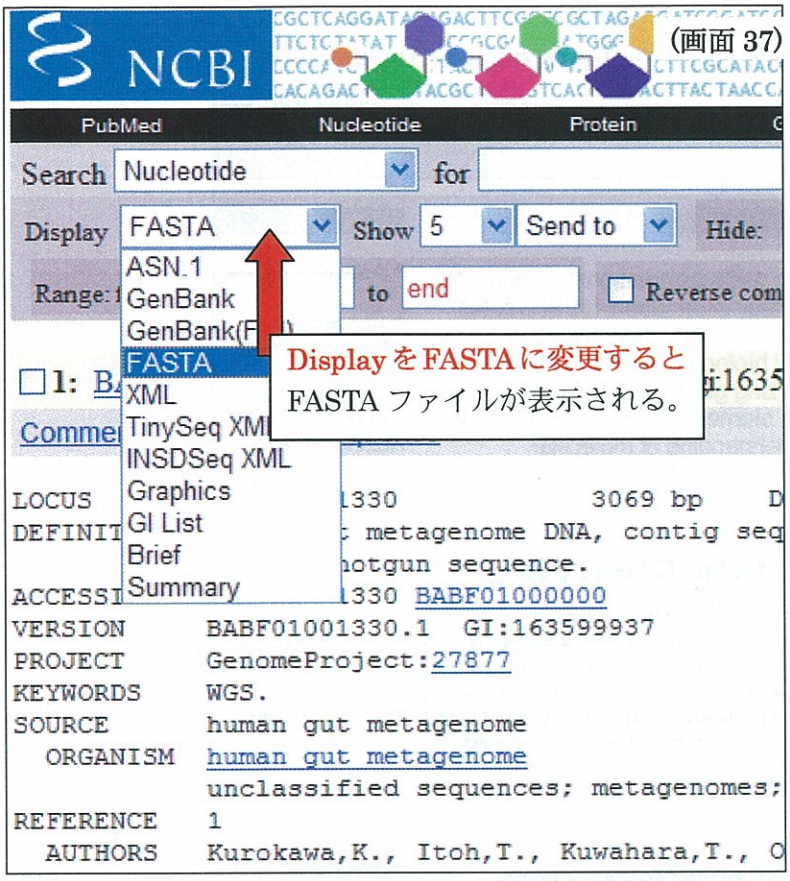
1: BABF01001330. Reports Human gut metagen...[gi:163599937]

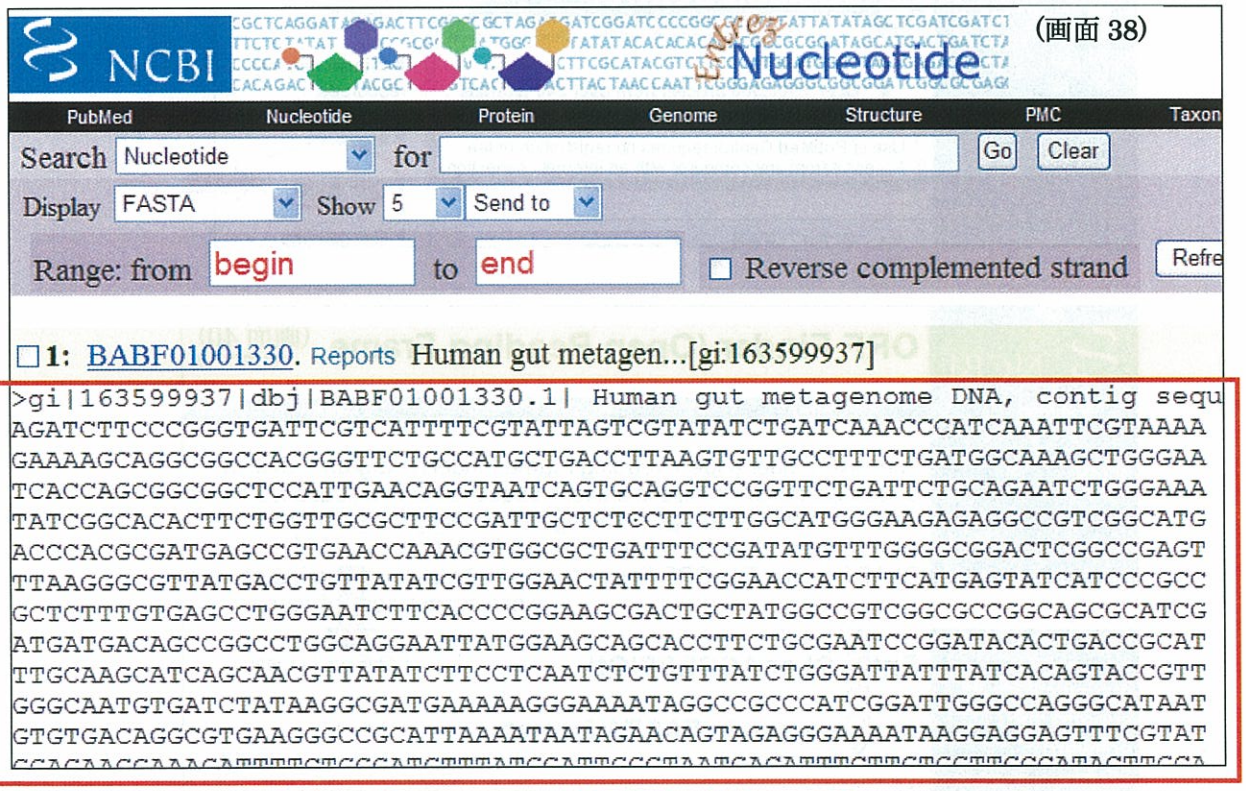
Comment Features Sequence

```
LOCUS       BABF01001330            3069 bp    DNA     linear   ENV 05-DEC-2007
DEFINITION  Human gut metagenome DNA, contig sequence: In-M_001330, whole
            genome shotgun sequence.
ACCESSION   BABF01001330 BABF01000000
VERSION     BABF01001330.1  GI:163599937
PROJECT     GenomeProject:27877
KEYWORDS    WGS.
SOURCE      human gut metagenome
  ORGANISM  human gut metagenome
            unclassified sequences; metagenomes; organismal metagenomes.
REFERENCE   1
  AUTHORS   Kurokawa,K., Itoh,T., Kuwahara,T., Oshima,K., Toh,H., Toyoda,A.,
            Takami,H., Morita,H., Sharma,V.K., Srivastava,T.P., Taylor,T.D.,
            Noguchi,H., Mori,H., Ogura,
            Sakaki,Y., Hayashi,T. and H
  TITLE     Comparative metagenomics re
            human gut microbiomes
  JOURNAL   DNA Res. 14 (4), 169-181 (2
   PUBMED   17916580
REFERENCE   2  (bases 1 to 3069)
  AUTHORS   Hattori,M., Oshima,K., Toyoda,A., Kurokawa,K., Hayashi,T.,
            Kuwahara,T., Takami,H., Morita,H., Itoh,T. and Itoh,K.
  TITLE     Direct Submission
  JOURNAL   Submitted (01-JUN-2007) Contact:Masahira Hattori The University of
            Tokyo, Department of Computational Biology, Graduate School of
            Frontier Sciences; 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561,
            Japan
COMMENT     This project was done in collaboration with Nara Institute Science
            and Technology, University of Miyazaki,  The University of
            Tokushima, Japan Agency for Marine-Earth Science and Technology,
            Azabu University, Kitasato University, Mitsubishi Research
            Institute INC. and RIKEN Genomic Sciences Center.
            low quality region (Q<15): 209-211
            Low quality regions for which bases could not be called are
            represented by runs of n within the sequence data, and associated
            gap features.
FEATURES             Location/Qualifiers
     source          1..3069
                     /organism="human gut metagenome"
                     /mol_type="genomic DNA"
                     /isolation_source="feces from healthy human sample In-M,
                     Infant, Female"
                     /db_xref="taxon:408170"
                     /environmental_sample
                     /note="synonym: human gut microbiota
                     contig sequence: In-M_001330"
```

既に機能に関するアノテーション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、既に記載されていないかを、確認しておく。

新規遺伝子候補レポートの環境由来の欄に記入する。(コピー&ペースト)



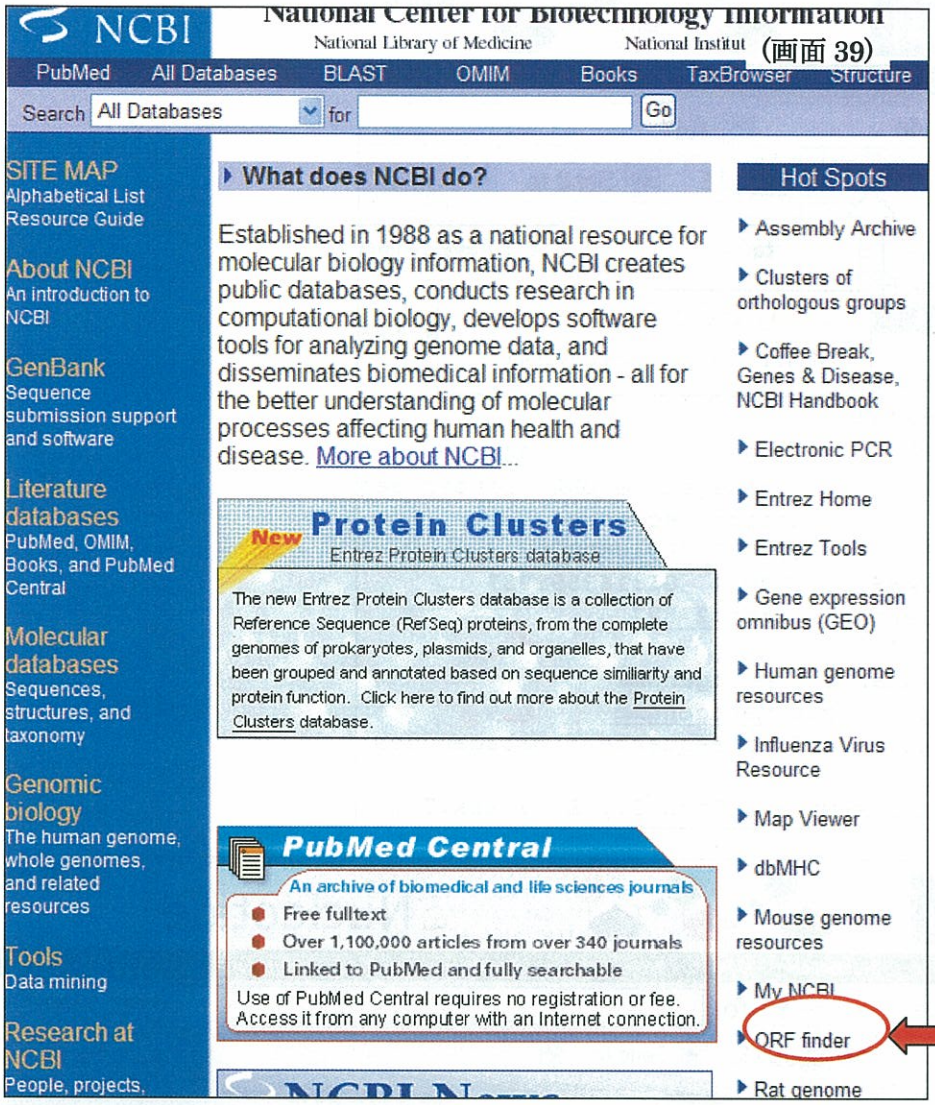


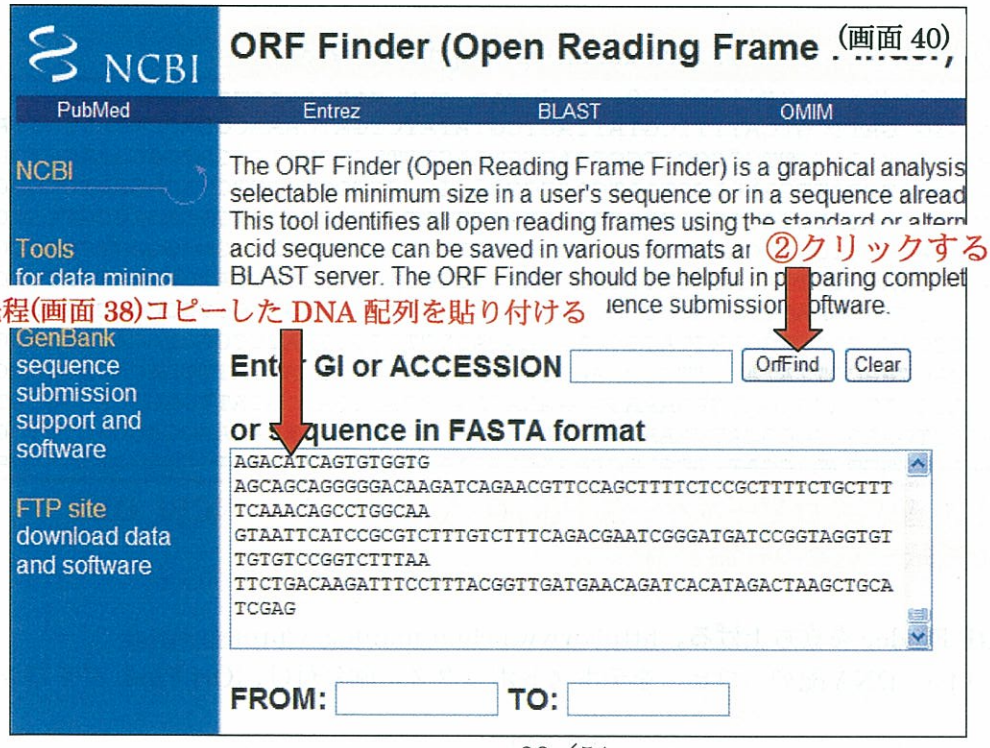
この DNA 配列をコピー＆ペースト(画面 40)して、次の NCBI ORF finder で遺伝子領域の有無を調べる。

NCBI の ORF Finder を立ち上げる。 http://www.ncbi.nlm.nih.gov/projects/gorf/

先程 (画面 14) の DNA 配列のコピーをテキストボックスに貼り付け、[Orf Find] アイコンを

クリックする。

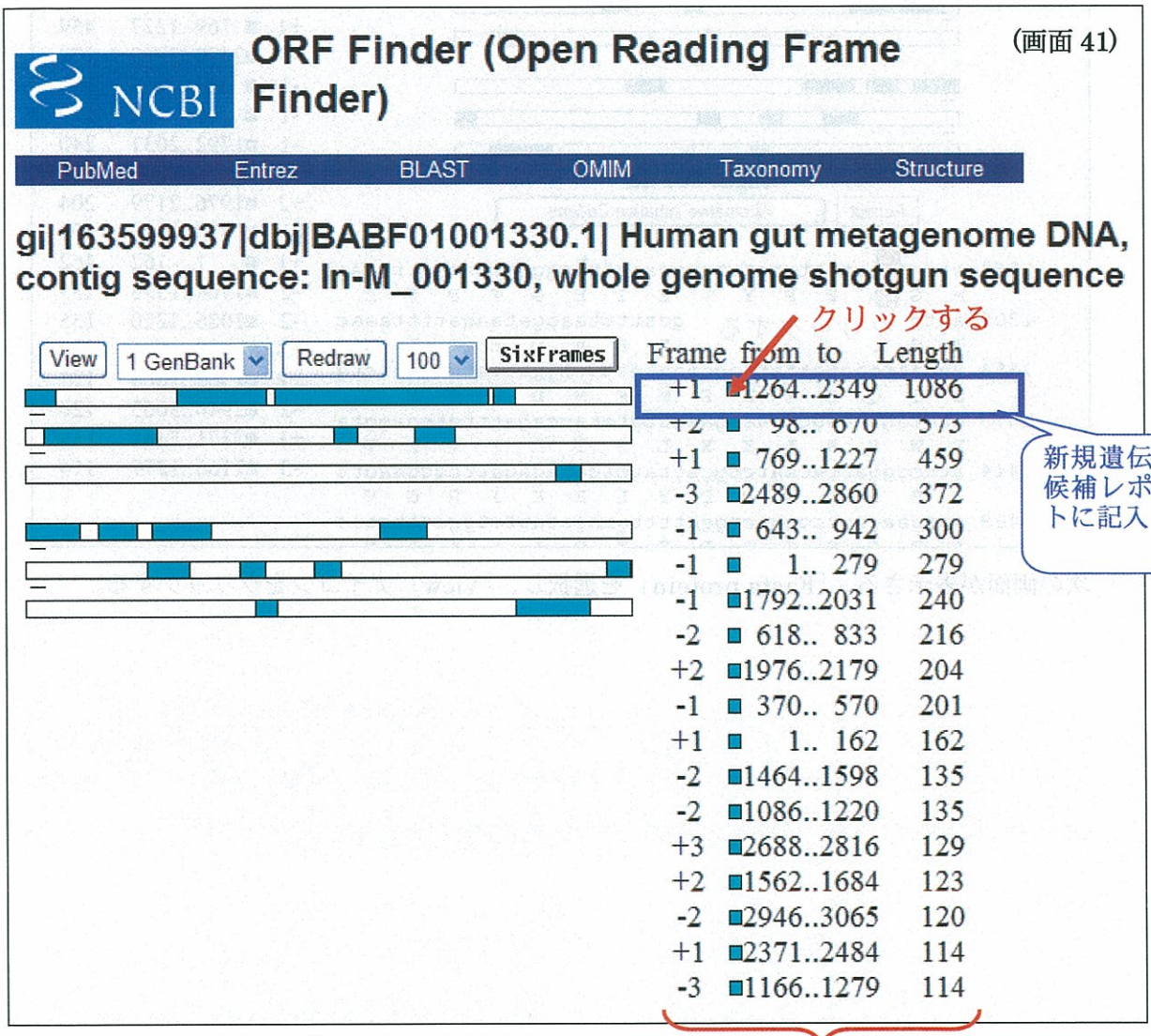

NCBI
National Center for Biotechnology Information
National Library of Medicine
National Institut
(画面 39)
PubMed
All Databases
BLAST
OMIM
Books
TaxBrowser
Structure
Search All Databases for Go
SITE MAP
Alphabetical List
Resource Guide
About NCBI
An introduction to NCBI
GenBank
Sequence submission support and software
Literature databases
PubMed, OMIM, Books, and PubMed Central
Molecular databases
Sequences, structures, and taxonomy
Genomic biology
The human genome, whole genomes, and related resources
Tools
Data mining
Research at NCBI
People, projects,
What does NCBI do?
Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More about NCBI...
New
Protein Clusters
Entrez Protein Clusters database
The new Entrez Protein Clusters database is a collection of Reference Sequence (RefSeq) proteins, from the complete genomes of prokaryotes, plasmids, and organelles, that have been grouped and annotated based on sequence similiarity and protein function. Click here to find out more about the Protein Clusters database.
PubMed Central
An archive of biomedical and life sciences journals
Free fulltext
Over 1,100,000 articles from over 340 journals
Linked to PubMed and fully searchable
Use of PubMed Central requires no registration or fee. Access it from any computer with an Internet connection.
Hot Spots
Assembly Archive
Clusters of orthologous groups
Coffee Break, Genes & Disease, NCBI Handbook
Electronic PCR
Entrez Home
Entrez Tools
Gene expression omnibus (GEO)
Human genome resources
Influenza Virus Resource
Map Viewer
dbMHC
Mouse genome resources
My NCBI
ORF finder
Rat genome
クリックする


NCBI
ORF Finder (Open Reading Frame
(画面 40)
PubMed
Entrez
BLAST
OMIM
NCBI
Tools
for data mining
GenBank
sequence submission support and software
FTP site
download data and software
The ORF Finder (Open Reading Frame Finder) is a graphical analysis
selectable minimum size in a user's sequence or in a sequence alread
This tool identifies all open reading frames using the standard or alter
acid sequence can be saved in various formats ar
BLAST server. The ORF Finder should be helpful in p paring complet
ience submissio oftware.
②クリックする
① 先程(画面 38)コピーした DNA 配列を貼り付ける
Ent GI or ACCESSION
OrfFind
Clear
or sequence in FASTA format
AGACATCAGTGTGGTG
AGCAGCAGGGGGACAAGATCAGAACGTTCCAGCTTTTCTCCGCTTTTCTGCTTT
TCAAACAGCCTGGCAA
GTAATTCATCCGCGTCTTTGTCTTTCAGACGAATCGGGATGATCCGGTAGGTGT
TGTGTCCGGTCTTTAA
TTCTGACAAGATTTCCTTTACGGTTGATGAACAGATCACATAGACTAAGCTGCA
TCGAG
FROM:
TO:

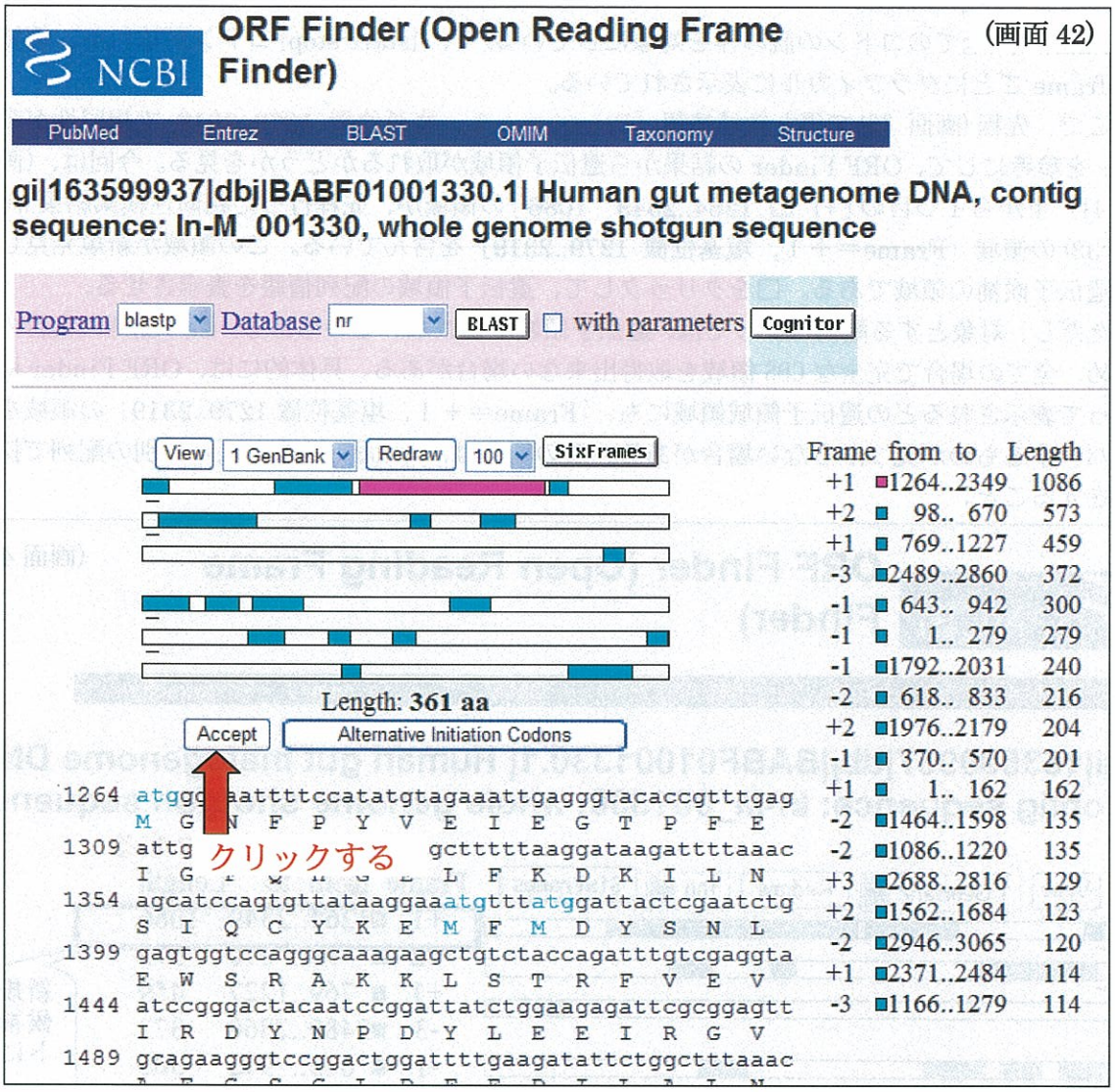
次の画面が表示される。
6 frame 分(全てのコドンの読み枠を対象にしている)で、[start-stop]コドンが取れる。領域が frame ごとにグラフィカルに表示されている。
ここで、先程(画面 33)で得た領域情報「**Frame＝＋１，塩基位置 1279..2319** で相同性が高い」を参考にして、ORF Finder の結果から遺伝子領域が取れるかどうかを見る。今回は、(画面 41) 上から１つ目の[+1 □ **1264..2349　1086**]の領域が、先程行った相同性検索結果(画面 33)の領域「**Frame＝＋１，塩基位置 1279..2319**」を含んでいる。この領域が新規発見した遺伝子候補の領域である。□をクリックして、遺伝子領域の配列情報を表示させる。

ただし、対象とする配列によっては、遺伝子領域がとれないものもある。断片配列の集合のため、全ての場合で完全な CDS 領域を取得出来ない場合がある。具体的には、ORF Finder によって表示されるどの遺伝子領域領域にも、「Frame＝＋１，塩基位置 1279..2319」の領域をカバーするものが見つからない場合がある。この場合は、(画面 32)まで戻り、別の配列で探索をすること。

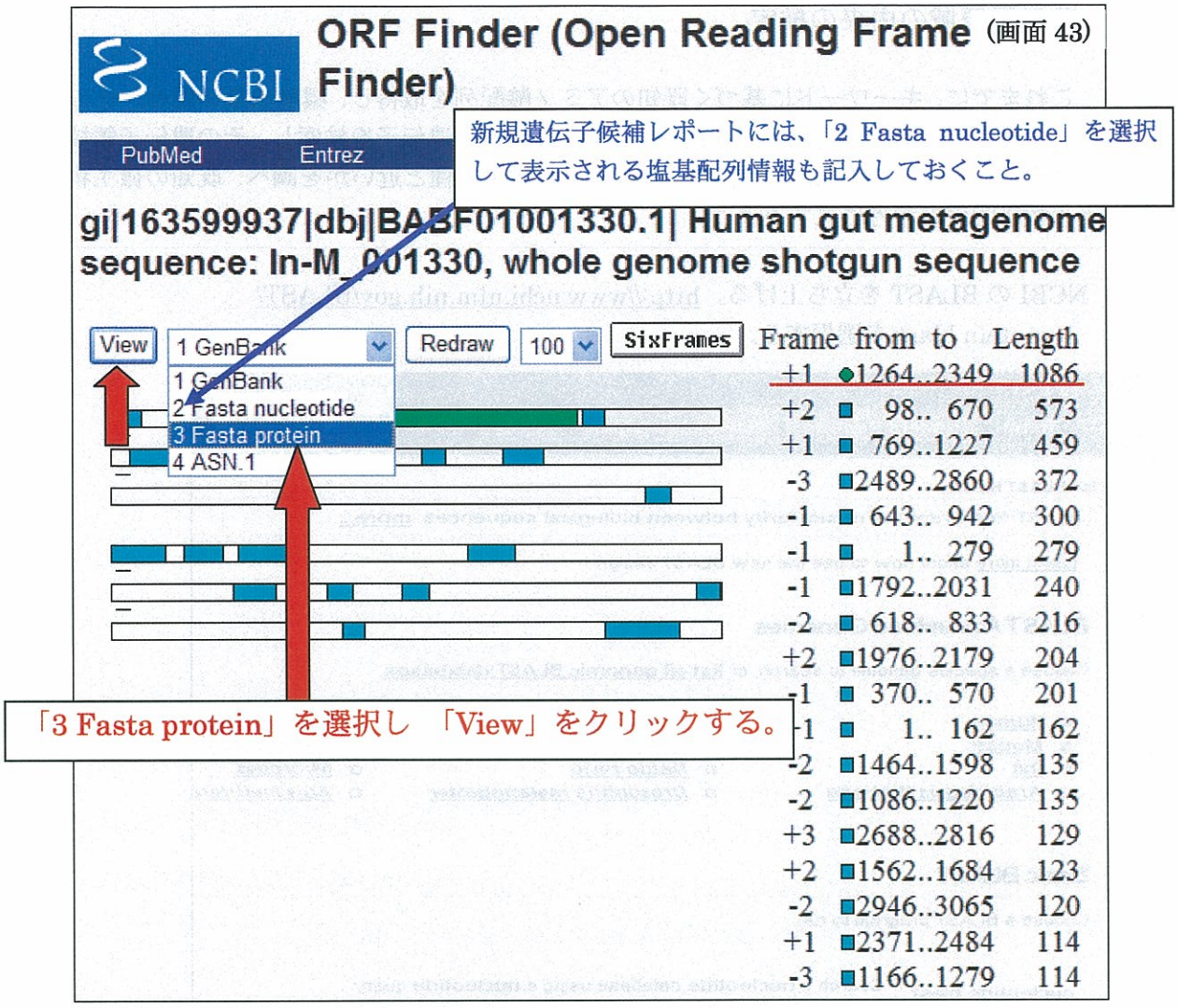
(画面 41)



次の画面のように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る。
Accept をクリックする。



次の画面が表示さる。「Fasta protein」を選択し、「View」アイコンをクリックする。



次のようなアミノ酸配列が表示されるので、コピーしておく。

(画面 44)

```
>lcl|Sequence 1 ORF:1264..2349 Frame +1
MGNFPYVEIEGTPFEIGFQHGELFKDKILNSIQCYKEMFMDYSNLEWSRAKKLSTRFVEVIRDYNPDYLE
EIRGVAEGSGLDFEDILALNCRSELVFVGNELDKADGGCTSIGISSDAGAGGDAFLAHNWDWKTSQRESM
IMMKITQKNGRPTIFMVTEAGIIGKTGFNSAGVGLYLNALSTDQAPKGLPLHMAMRGILDCETLAEAVKA
ATRFQLGCCANFMIGHKNGECVDIEIENEEFDVLYPKDGIIVHTNHFISSRLPILPRKDMGKRKFTDSFV
RLGRADKLLRKKGSEISEEDIKAVLTDHVEYPSSICRHDDEKLEKGLRMGTVFSMIINLTKGEILFCKGN
PCELEYEKYRI*
```

アミノ酸配列をコピーしておく。新規遺伝子候補レポートに記入する。

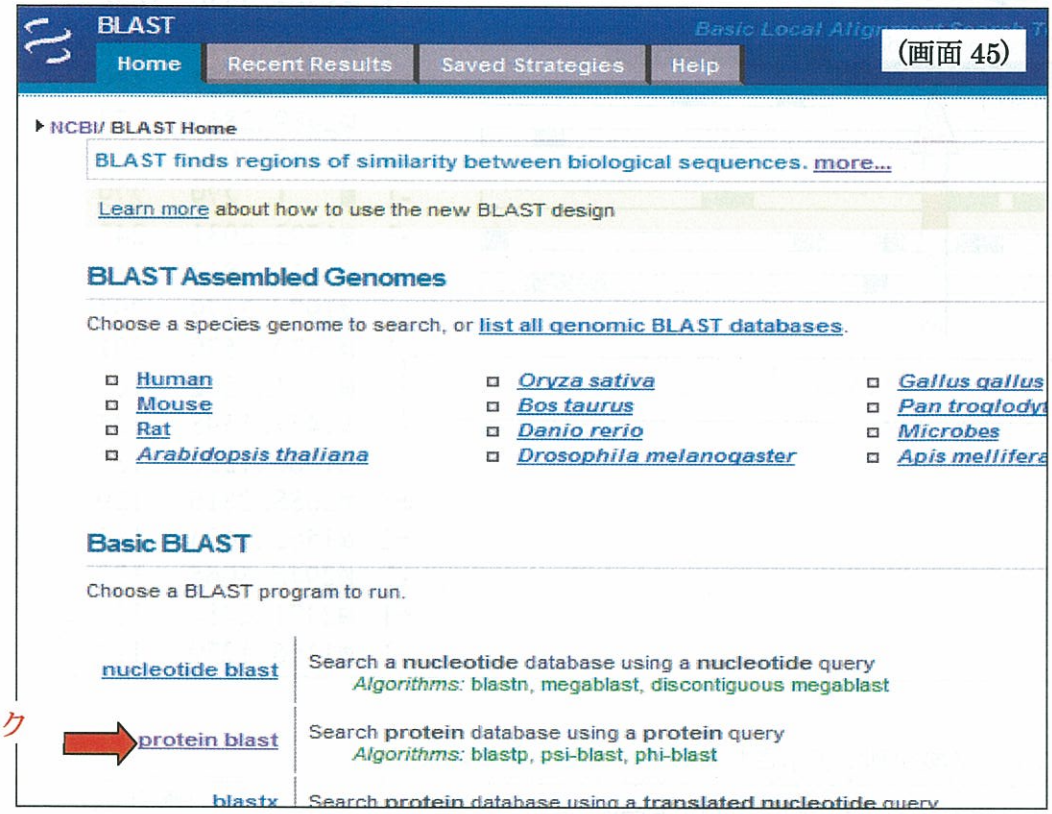
アミノ酸配列をコピーし、既知の微生物ならば何の微生物に近いかを調べるために、NCBI Blast(blastp)を実行し、相同性の高い生物種を確認していく。

ここでは、ORF Finder による遺伝子領域の確定方法を説明したが、東京大学らのグループが環境由来 DNA 断片配列に特化した遺伝子予測プログラム“MetaGene”を開発している。本テキスト最後のほうの（付録 3）で紹介をするので、余力のある人は ORF Finder による結果と MetaGene による結果とを見比べてみるとよい。

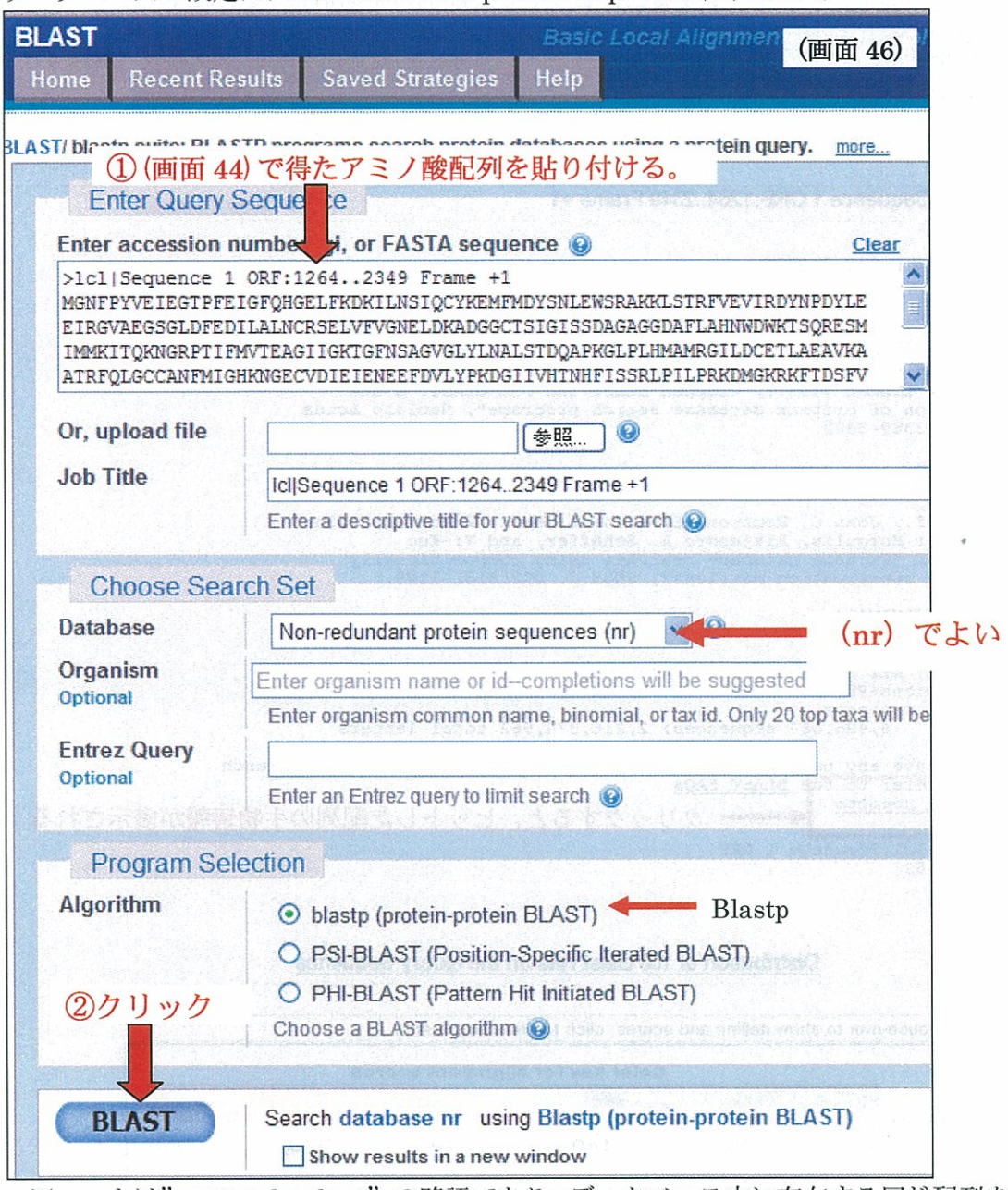
3.3　遺伝子領域の由来の推定

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、相同性のある遺伝子を検索し、その遺伝子領域までを調べることが出来た。最後に、既知のどの微生物種と近いかを調べ、既知の微生物との類縁関係を参考に見てみよう。

NCBI の BLAST を立ち上げる。http://www.ncbi.nlm.nih.gov/BLAST/
protein blast を選択する。



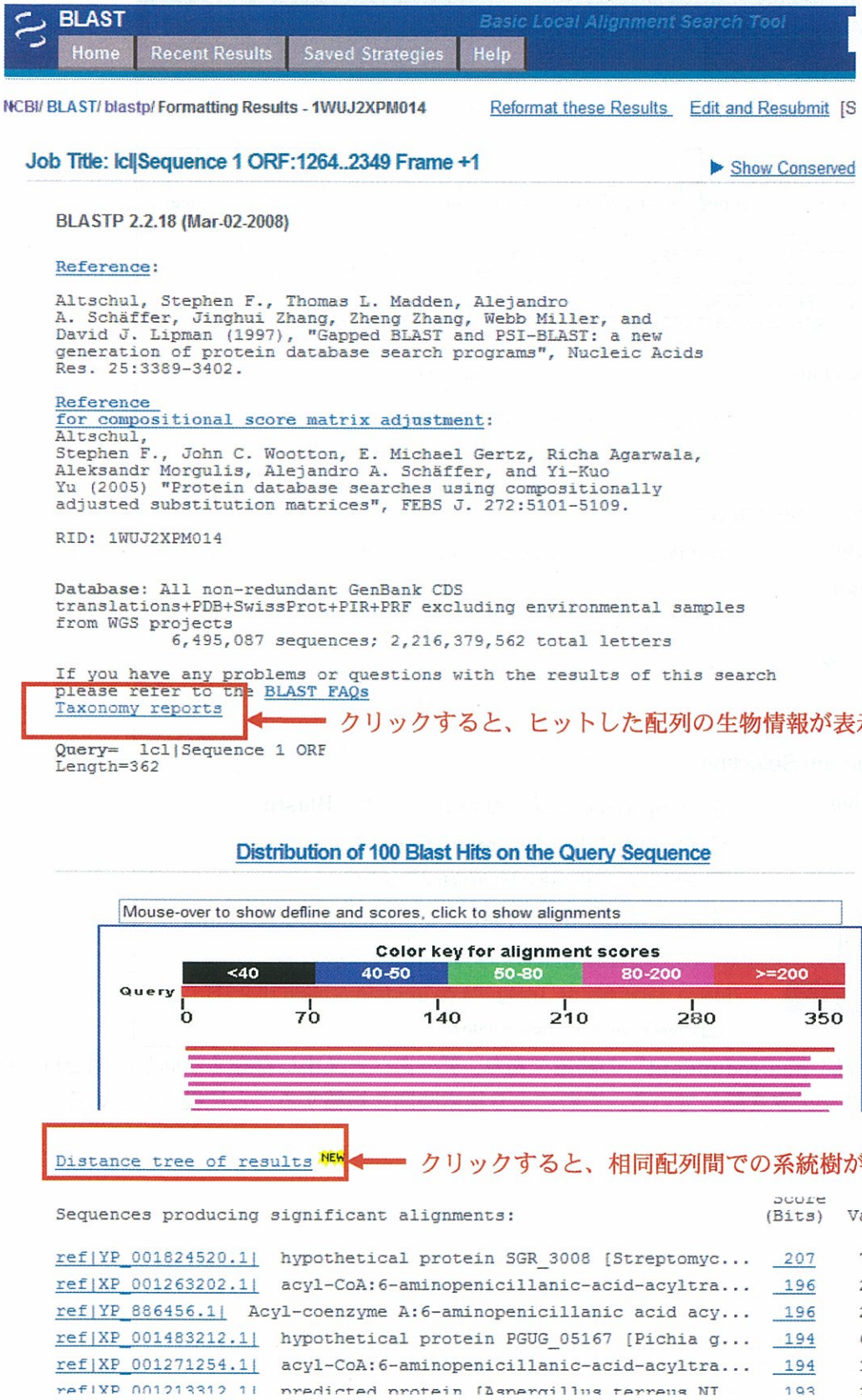
(画面 44) で得たアミノ酸配列を貼り付け、blastp を実行する。
データベースの設定は、Non-redundant protein sequences(nr)でよい。



※ nr とは”non-redundant”の略語であり、データベース中に存在する同じ配列を持つ重複したデータを除いている。これにより検索時間を短縮することに役立っている。

次のように、blast の結果が表示される。この画面で、

➢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る。



(画面 47)

Taxonomy reports をクリックした画面

Lineage Report

(画面 48)

クリックする

cellular organisms
. Bacteria [bacteria]
. . Actinomycetales [high GC Gram+]

. . . Streptomyces griseus subsp. griseus NBRC 13350	207	2 hits	[high GC Gram+]	hypothetical protein SGR_3008 [Streptomyces griseus subsp
. . . Mycobacterium smegmatis str. MC2 155	196	2 hits	[high GC Gram+]	Acyl-coenzyme A:6-aminopenicillanic acid acyl-transferase
. . . Saccharopolyspora erythraea NRRL 2338	122	2 hits	[high GC Gram+]	peptidase C45, acyl-coenzyme A/6-aminopenicillanic acid a
. . . Rhodococcus sp. RHA1	107	4 hits	[high GC Gram+]	probable isopenicillin-N N-acyltransferase [Rhodococcus s
. . . Arthrobacter aurescens TC1	103	2 hits	[high GC Gram+]	hypothetical protein AAur_1233 [Arthrobacter aurescens TC
. . . Agromyces sp. KY5R	83	1 hit	[high GC Gram+]	Acyl-coenzyme A/6-aminopenicillanic acid acyl-transferase
. . . Brevibacterium linens BL2	65	1 hit	[high GC Gram+]	hypothetical protein BlinB01001127 [Brevibacterium linens
. . . Frankia sp. CcI3	58	2 hits	[high GC Gram+]	peptidase C45, acyl-coenzyme A:6-aminopenicillanic acid a
. . Polaromonas sp. JS666	179	2 hits	[b-proteobacteria]	peptidase C45, acyl-coenzyme A:6-aminopenicillanic acid a
. . Planctomyces maris DSM 8797	179	2 hits	[planctomycetes]	hypothetical protein PM8797T_07442 [Planctomyces maris DS
. . Verminephrobacter eiseniae EF01-2	177	2 hits	[b-proteobacteria]	peptidase C45, acyl-coenzyme A [Verminephrobacter eisenia
. . Bradyrhizobium sp. ORS278	176	2 hits	[a-proteobacteria]	hypothetical protein BRADO2616 [Bradyrhizobium sp. ORS278

NCBI Taxonomy Browser (画面 49)

Entrez PubMed Nucleotide Protein Genome Structure PMC

Search for as complete name lock Go Clear

Display 3 levels using filter: none

Streptomyces griseus subsp. griseus NBRC 13350

Taxonomy ID: 455632
Rank: no rank
Genetic code: Translation table 11 (Bacterial and Plant Plastid)
Other names:
synonym: **Streptomyces griseus subsp. griseus IFO 13350**
equivalent name: **Streptomyces griseus subsp. griseus strain NBRC 13350**
equivalent name: **Streptomyces griseus subsp. griseus str. NBRC 13350**

Lineage(full)
cellular organisms; Bacteria; Actinobacteria; Actinobacteria (class); Actinobacteridae; Actinomycetales; Streptomycineae; Streptomycetaceae; Streptomyces; Streptomyces griseus; Streptomyces griseus subsp. griseus

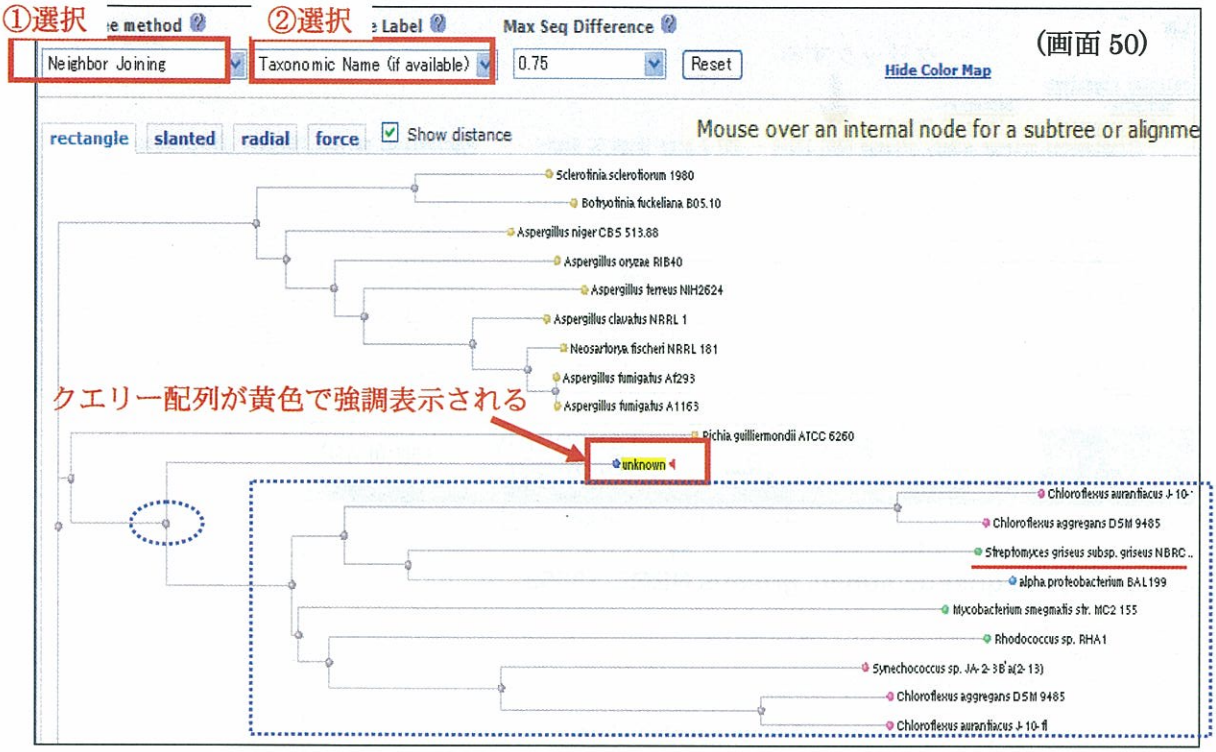
新規遺伝子候補レポートにコピー＆ペーストで記入する。

Taxonomy reports(画面 48)の画面より、トップヒットの“Streptomyces griseus subsp”が既知の微生物の中では最も近いと考えられる。

次に、この“Streptomyces griseus subsp”がどのような微生物かを、文献やネット等で調べてみる。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 等を調べる。

➢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る。
参考に見ておこう。

① Tree method で Neighbor Joining を選択し、
② Sequence Label で Taxonomic Name を選択 する。
次のような系統樹が表示される。



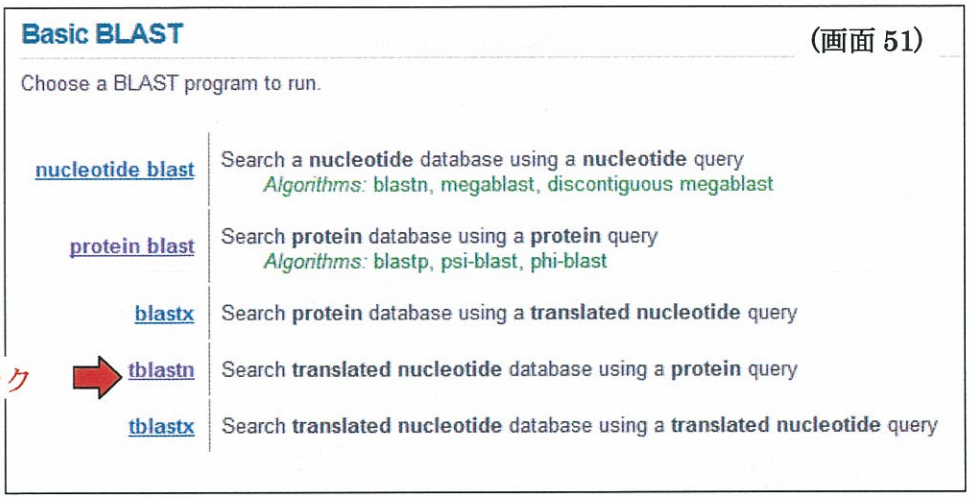
この系統樹を見ると、Taxonomy reports（画面 48）で最もスコアの高かった Streptomyces griseus subsp. griseus NBRC 13350 が、問合わせ（クエリー）配列に最も近縁とは言い難い。点線で囲った 9 つの微生物種のうち、どれが最も近縁かを判断することは難しい。これは、新規探索した遺伝子候補が、新規性の高い生物種の遺伝子を見つけてきた可能性が高いとも考えることが出来る。

系統樹は、画面 50 の①のところで選択するプログラムの種類により変わることがあるので、違うプログラムで系統樹を作成することを試みてもよい。

3. 4　他の環境由来サンプルからの探索 （長浜バイオ大学以外から探索する場合）

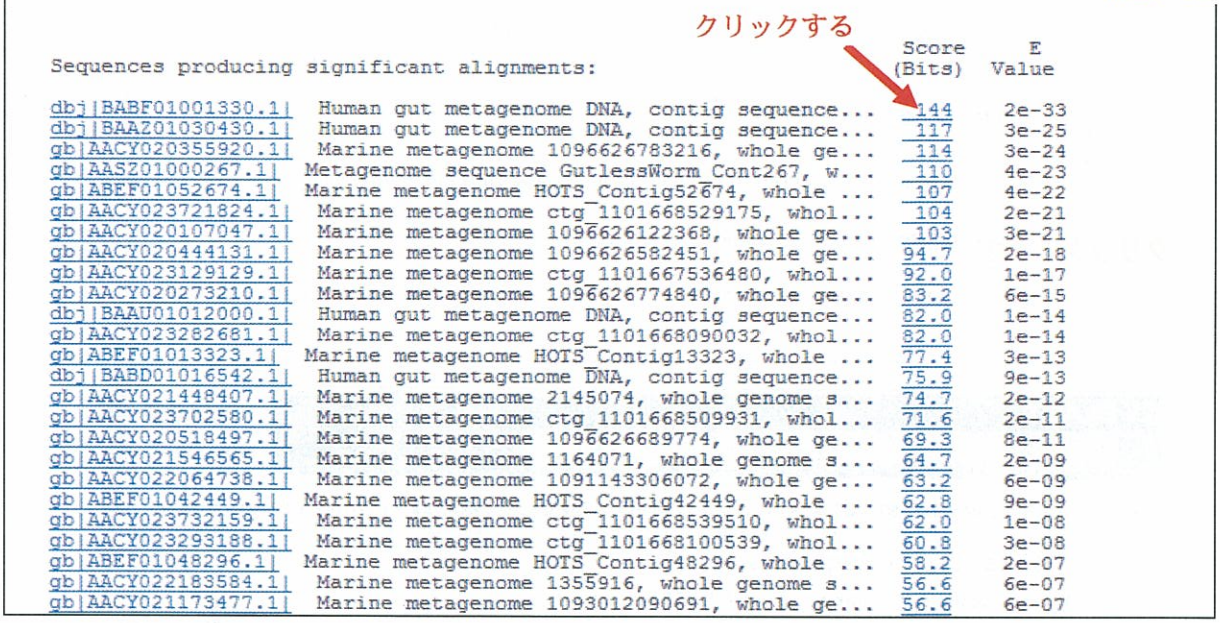
3.1 (画面 30) の手順で指示した、長浜バイオ大学ローカル環境下の BLAST のデータベースは、ヒトの腸内細菌由来(約 35 万件, DB 名：Human_gut.fasta)の他、計 7 種類の環境由来データベースのみであった。長浜バイオ大学ローカル環境下で良い結果が見出せていない場合や、**長浜バイオ大学の学生以外の方がご自宅等で探索される場合は、NCBI の BLAST（tblastn）を用い、データベースの設定を“environmental.seq”にすると、**上記のデータ以外も含む**膨大な環境由来ＤＮＡ配列からの検索が可能となる。**検索する環境由来メタゲノムの対象を広げて、更に別の新たな遺伝子領域候補を探してみよう。

NCBI　BLAST を立ち上げる。http://www.ncbi.nlm.nih.gov/BLAST/
次に **tblastn** をクリックする。



検索結果が表示される。(これは、画面下の方)

(画面 53)
画面 32 に
相当する。



ヒト腸内細菌由来のものに加えて、海由来 (Sargasso sea) 由来のものも検索されている。どんな環境由来の配列から取得できそうか、遺伝子領域はあるのかどうか、あればその機能や応用例がないか、**3 章（P22～）の作業を繰り返してみる。**

今回の目的である健康への貢献遺伝子の候補を、膨大な環境由来 DNA 配列データベースの中から発掘し、長浜バイオ大学から世界へ発信していこう。

※ここで、これまでの実習手順の骨子をまとめておく。

① 健康に、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す。
（使用サイト等：Google，NCBI PubMed，専門誌，図書館の文献　等）

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する。　（使用サイト：DDBJ の ARSA）
※ここまでで、レポートの テーマ検索レポート/既知遺伝子レポートを作成するとよい。

③ 取得したアミノ酸配列と環境由来 DNA 配列との相同性検索を行う。環境由来 DNA 配列の中に目的の遺伝子やタンパク質が存在するかどうかを調べる。
（使用サイト：長浜ローカル環境の、NCBI の **tblastn** ）
URL：http://hpc31.nagahama-i-bio.ac.jp/blast/blast.html

④ 相同性検索でヒット (発見) した、環境由来微生物等の DNA 配列を取得する。
（使用サイト：NCBI トップページ）

⑤ 遺伝子領域を確定しアミノ酸配列を取得する。
（使用サイト：NCBI ORF finder，EMBOSS transeq ）

⑥ 確定した遺伝子領域のアミノ酸配列から、生物種由来を推定する。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る。
（使用サイト：NCBI protein blast(blastp) ）

⑦ 様々な環境由来データベースから、上記③～⑥を試みる。
（使用サイト：NCBI tblastn　データベース設定 environmental.seq）

⑧ レポート（所定様式有り。テーマ検索レポート/既知遺伝子レポート/新規遺伝子候

補レポートの 3 種類。）を作成する。

（余力のあるグループは、以下についてもチャレンジし追加のレポートを作成する。）

⑨ 取得した配列の機能を調べる。（使用サイト：EBI InterProScan 等）

⑩ Pathway を調べる。（使用サイト：KEGG）

自分達が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よう。参考までに、次の解説等を付しておく。

（付録 1）：DDBJ キーワード検索システム ARSA (http://arsa.ddbj.nig.ac.jp/)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）:探索した環境由来 DNA 断片配列中にコードされている遺伝子の探索**[上級者編]**

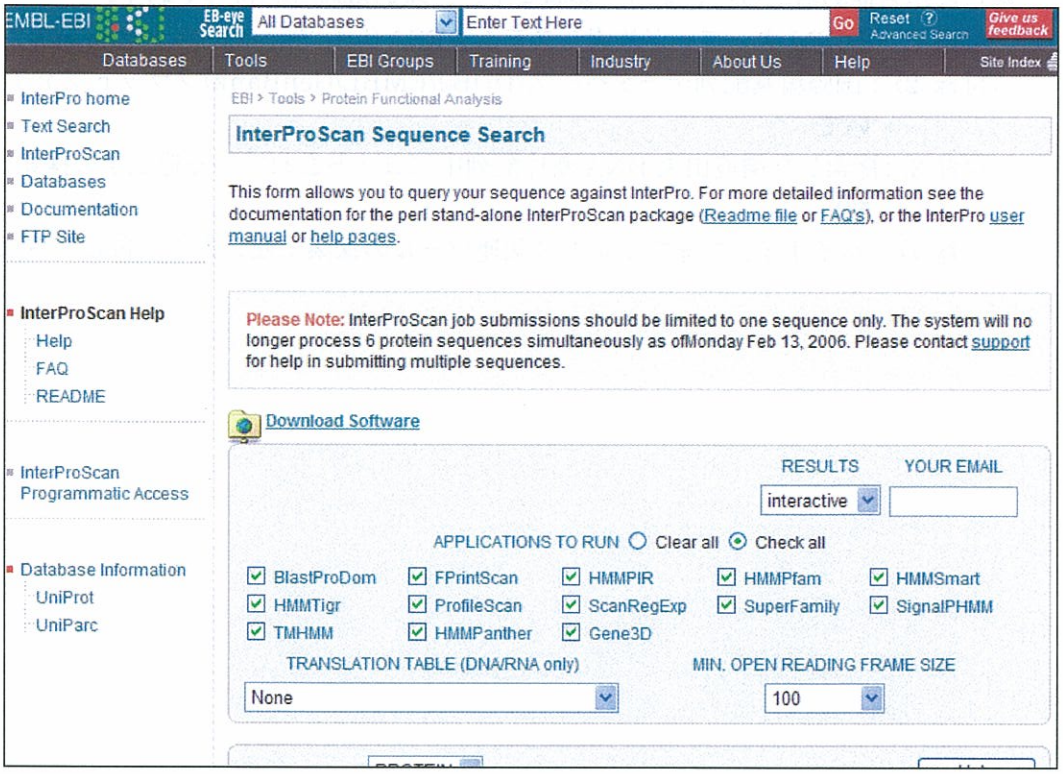
（付録 4）：バイオインフォマティクス関連ツールの探索や使い方を自習してみよう。

3.5（余力のあるグループは）取得した配列の機能を確認する

これまでの作業の中で、環境由来から取得した配列について、既にある程度の機能情報等が得られているかもしれない。余力のあるグループは、更にモチーフ検索を行い、モチーフの有無を調べ、それらの情報からも遺伝子の機能推定を行ってみてもらいたい。

ここでは、ツールの紹介だけをしておく。

EBI InterProScan： http://www.ebi.ac.uk/InterProScan/index.html

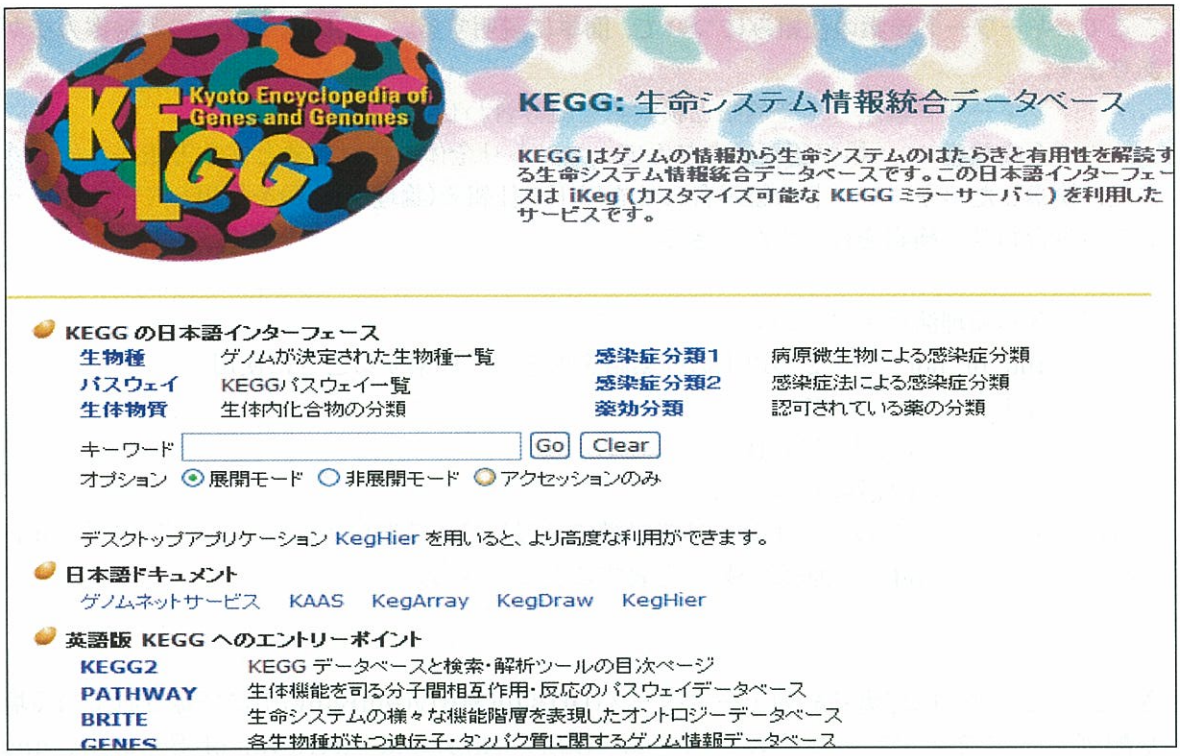


副教材「基礎と実習バイオインフォマティクス」　P60～62 に操作方法が記載されているので参照するとよい。

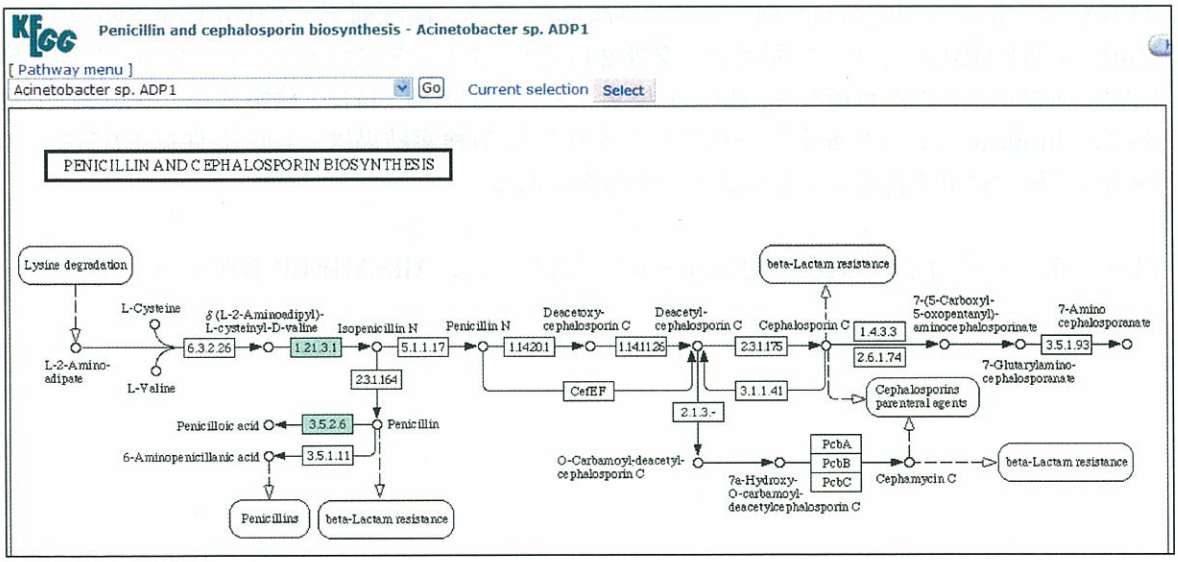
3.6（余力のあるグループは）Pathway を確認してみる

KEGG は、細胞レベルでの生命システムの機能に関する知識を、分子間相互作用ネットワーク（代謝、シグナル伝達、遺伝情報等）の二項関係に基づいた情報としてデータベース化（PATHWAY）されている。余力のあるグループはこちらも利用して、より広い視点から遺伝子の機能推定等を試みてもらいたい。

KEGG: 生命システム情報統合データベース： http://www.kegg.jp/



（画面例）



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て使用する方法
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て検索を行う事ができる。

---使用できる論理演算子について

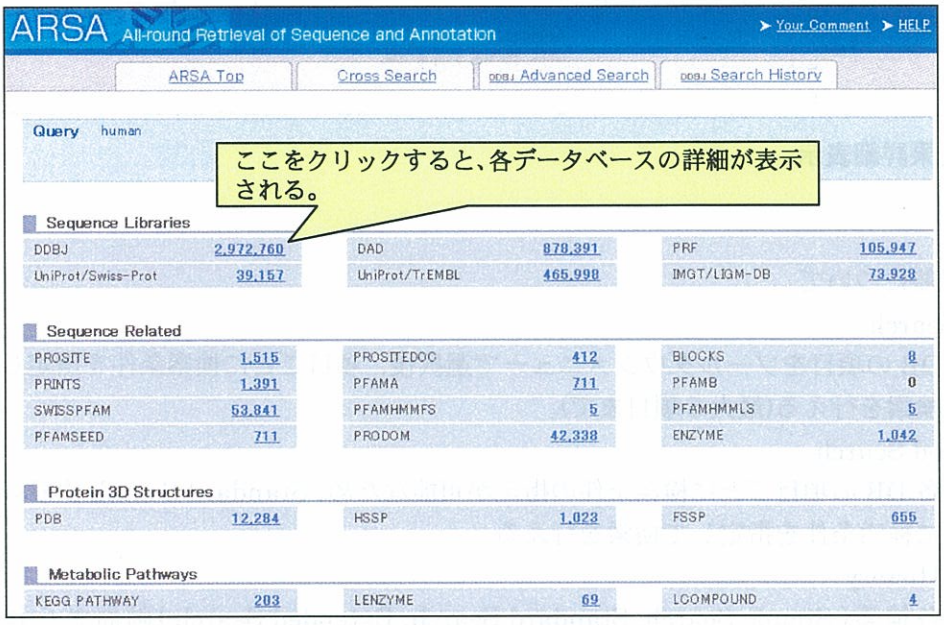
and, or, not　→　2 つ以上の検索語を指定して検索するときに使用。
and　：A かつ B
or　：A もしくは B
not　：A 以外のもの

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。

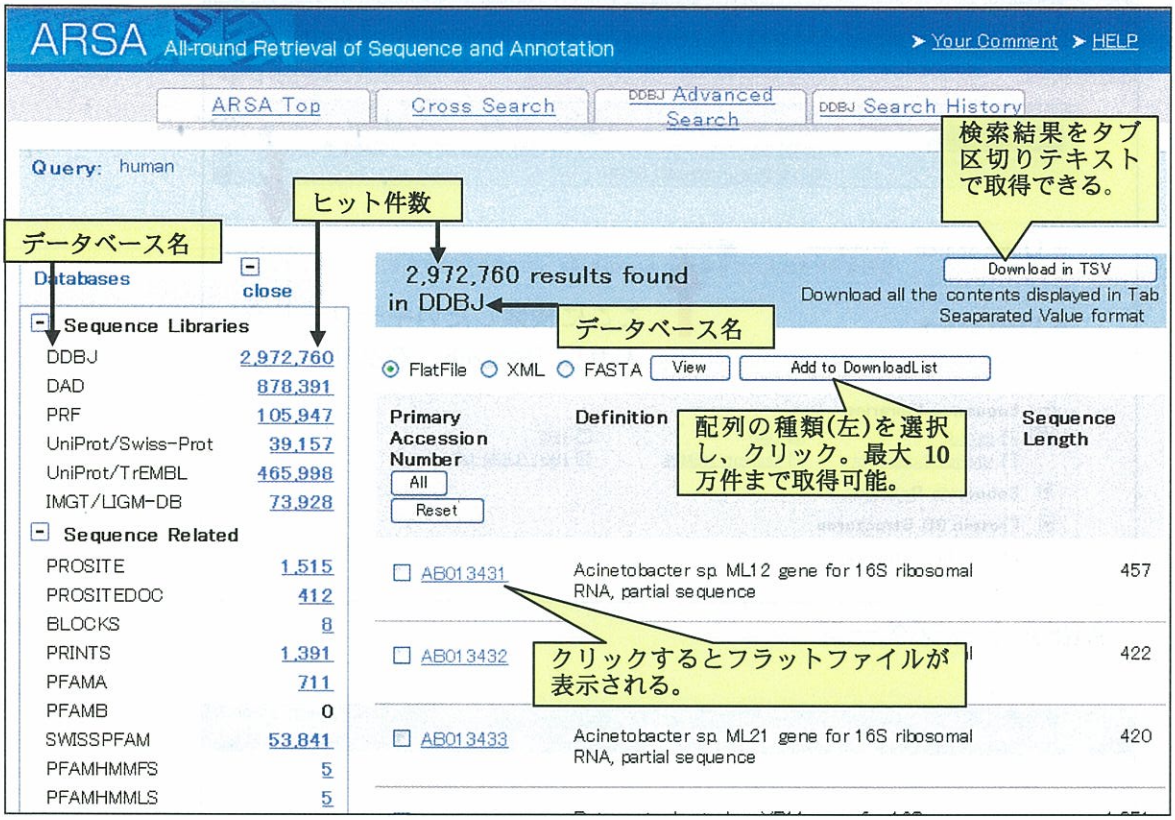
さて、ここでは、国際塩基配列データベース(DDBJ/EMBL/GenBank)中に登録されている塩基配列データを検索するためのシステム ARSA (All-round Retrieval of Sequence and Annotation)の使用方法について、簡単に紹介する。
ARSA とは、最新の国際塩基配列データのみならず、アミノ酸配列データ(UniProt/Swiss-Prot, UniProt/TrEMBL)、モチーフ配列データ等 23 のデータベースについて、データベース間を横断的に検索することが可能なキーワード検索システムである。特に、検索スピードが速く、例えば、”human”と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、ARSAHELP 参照すると良い。

ARSA トップページ



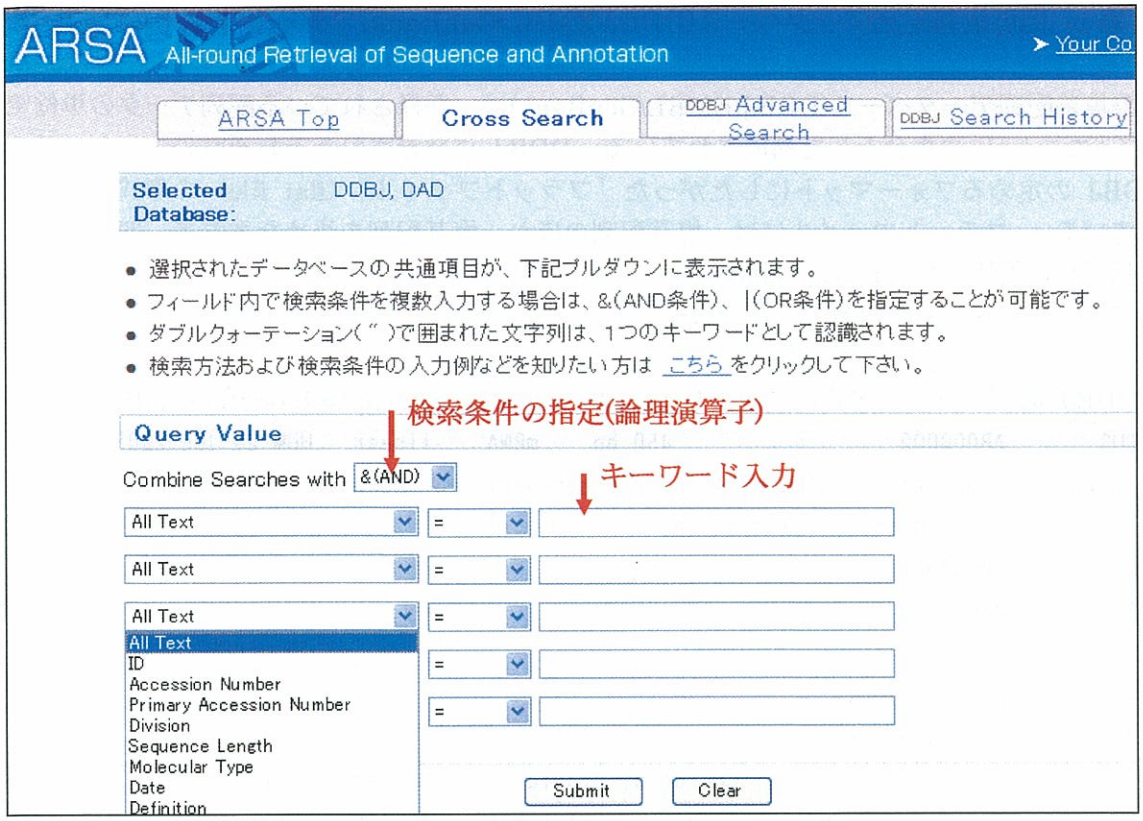
Hit Counts ページ（各 DB の検索結果件数）



検索結果詳細表示

その他、検索機能について

- Cross Search
  - 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)。
- Advanced Search
  - 各 DB の項目ごとに検索条件の指定が可能なため、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。



Cross Search ページ

(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="MAKIKIGINGFGRIGRLVARVALQSDDVELVAVNDPFITTDYMT
                     YMFKYDTVHGQWKHHEVKVKDSKTLLFGEKEVTVFGCRNPKEIPWGETSAEFVVEYTG
                     VFTDKDKAVAQLKGGAKKV"
BASE COUNT          102 a          119 c          131 g          98 t
ORIGIN
        1 cccacgcgtc cggtcgcatc 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。
/tarans_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。

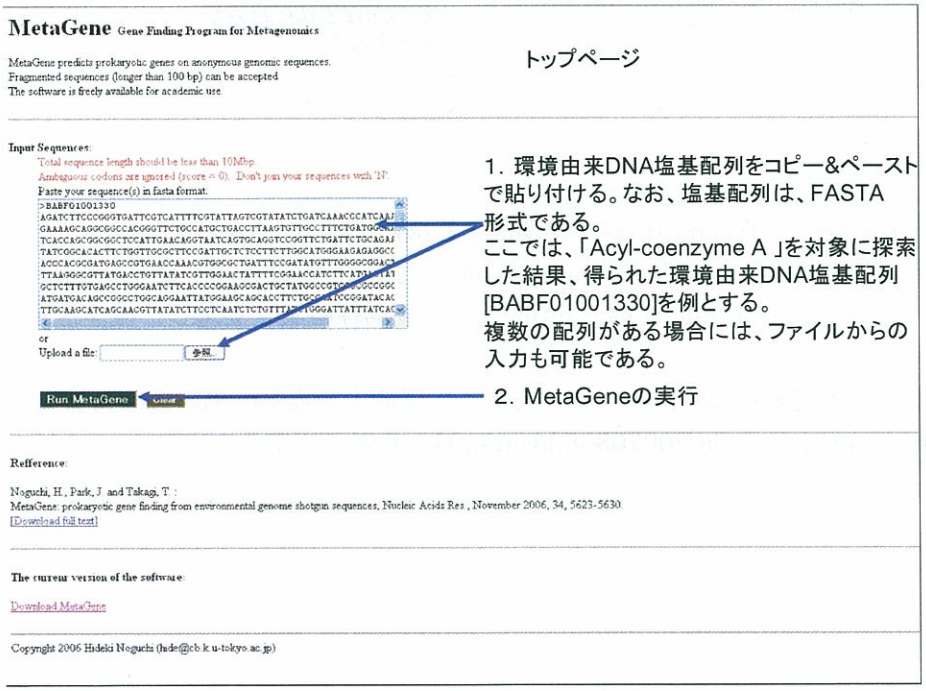
・ BASE COUNT 塩基配列組成

・ ORIGIN 塩基配列情報

（付録 3）：探索した環境由来 DNA 断片配列中にコードされている遺伝子の探索【上級者編】

一般に、ゲノム配列中にコードされている遺伝子領域を同定する方法として、本実習で紹介している既知遺伝子のアミノ酸配列情報を基に相同性検索を行い、ゲノム配列上に相同性領域が含まれるか否かを調べる方法のほかに、遺伝子領域に含まれるコドン組成や塩基配列の GC 含量などの遺伝子領域の特徴や遺伝子領域の制御領域の塩基配列組成の特徴などを基にした遺伝子領域予測プログラムが開発されており、この遺伝子領域予測プログラムを活用した遺伝子領域探索方法がある(生物ごとにコドンの使用パターンには生物固有な特徴があり、その特徴を活用したものである。コドン使用パターンとその多様性については、ここでは省略するが、詳しくは国立遺伝学研究所遺伝学電子博物館にて紹介されているので、そちらを参照されたい。http://www.nig.ac.jp/museum/evolution/04.html)。前者では、自分が目的とする有用遺伝子がコードされている環境由来 DNA 断片配列を探索するのに有用であり、後者では遺伝子領域の網羅的な予測や有用遺伝子がコードされているかもしれない環境由来DNA断片配列の他の領域に遺伝子領域がコードされているかどうかを調べるために有用である。
他の領域の遺伝子領域候補を探索することは、この遺伝子がオペロン構造を形成しているかどうかや環境由来 DNA 断片配列の系統を推定する上でも非常に重要である。
遺伝子領域予測プログラムは、DNA 塩基配列解読技術の進歩により、多くのゲノム配列が同定されおり、得られたゲノム配列中にコードされている遺伝子領域を同定するための方法として開発が進められており、原核生物では、同定プログラムの精度は 95%を超える場合もあり、ゲノムアノテーション(ゲノム配列中にどのような遺伝子があるかを注釈付けしてゆくこと)において、活用されている。
ここでは、遺伝子領域予測プログラムを活用した遺伝子探索方法について紹介する。ゲノム環境中より取得された環境由来 DNA 断片配列に特化した遺伝子領域予測プログラム MetaGene(Noguchi et al *Nucleic Acids Res.*, 34, 5623-5630, 2006.)が東京大学らのグループによって開発されており、Web 上で公開されている(http://metagene.cb.k.u-tokyo.ac.jp/metagene/)。Metagene を用いた遺伝子探索例を以下に紹介する。なお、MetaGene の詳細については上記の原著論文を参照されたい。

1．Metagene WEB ページでの環境由来 DNA 配列を投入し、実行する。



2．出力結果の取得

```
MetaGene Prediction Results

# BABF01001330
# gc = 0.480613
# bacteria
98   670  + 0 30.6338 complete
769  1227 + 0 25.8365 complete
1264 2349 + 0 25.6298 complete
2489 3069 - 2 31.8919 partial (lack 5'-end)
```

ここで、出力結果の書式は以下の通りである。
#配列ID(FASTAファイルのコメント行が表示)
#配列のGC含量
#生物種推定結果(bacteria or Archaea)
[開始位置] [終了位置] [ストランドの向き] [コドンの読み枠] [スコア] [完全長/部分配列]

3．予測された遺伝子領域の塩基配列の対象領域の抽出

4．予測された遺伝子領域遺伝子領域の相同性検索による機能同定

予測された遺伝子領域の DNA 配列を query 配列とし、NCBI の BLAST（blastn）を用いて、対象データを Nucleotide collection(nt)として相同性検索を実施し、機能推定を行う。

次の（付録 4）の中の図 5「NCBI BLAST を使って機能未知塩基配列の機能を推定する」（URL: http://togotv.dbcls.jp/20070808.html#p01）を参考にするとよい。

(付録 4) バイオインフォマティクス関連ツールの探索や使い方を自習してみよう。

バイオインフォマティクスの分野では、バイオ分野のデータベースと同様に、世界中の研究者によって、ゲノム配列やアミノ酸配列解析、遺伝子発現解析など様々な解析ツールが開発され、インターネットを通じて、公開されている。(The Bioinformatics Links Directory : http://bioinformatics.ca/links_directory/ によると 2,360 の解析ツールが公開され、体系的に整理・登録されている。) これらの解析ツールのうち、一般的に使用される BLAST などの配列相同性解析ツールや配列アラインメントツールなどについては、本学の他の実習を通じて紹介しているが、紹介している解析ツール以外にも有用な解析ツールが公開されている。しかしながら、実際にどのような有用なバイオインフォマティクス解析ツールが公開されているかを調べるには、バイオインフォマティクスに精通していないと調査・利用するのがなかなか難しいのが現状である。そのため、本学も参加している「ライフサイエンス分野の統合データベースプロジェクト」では、学部学生やバイオインフォマティクス初心者の方にも容易に解析ツールが検索できるように、解析ツールを体系化し、目的に沿った検索が可能な「WEB リソースポータルサイト for バイオインフォマティクス」
(http://tools.lifesciencedb.jp/cgi-bin/WebResourcePortal/WebResourcePortal.cgi)や、解析ツールの使用方法を実際の解析ツールの画面とともに説明つきの動画として配信する「統合 TV」(http://togotv.dbcls.jp/)が構築・公開されており、日本語の環境でバイオインフォマティクス関連ツールの探索や利用方法の自習が可能となっている。
ここでは、皆さんに本実習で利用する解析ツール以外についても自分で調べて、遣ってみたいという方のために「WEB リソースポータルサイト for バイオインフォマティクス」と「統合 TV」について紹介する。

1.「WEB リソースポータルサイト for バイオインフォマティクス」
(http://tools.lifesciencedb.jp/cgi-bin/WebResourcePortal/WebResourcePortal.cgi)

現在、バイオインフォマティクス関連解析ツールとして WEB 上で公開されている解析ツールとしては、BLAST をはじめ、2,300 を超える解析ツールが開発され、公開されている。解析ツールの種類も多岐にわたっており、自分の目的に合った解析ツールを探索するのが困難な状況になりつつある。そのため、自分の目的に合った解析ツールを容易に探索するために、解析ツールを日本語で解析の目的に沿って体系的に整理し、まとめて公開を試みているのが、ここで、紹介する「WEB リソースポータルサイト for バイオインフォマティクス」である。現在、解析ツールとして 456 件、解析ワークフロー(手順を記したもの)として、29 件が公開されている。
利用方法については、以下の画面イメージにて紹介を行う。

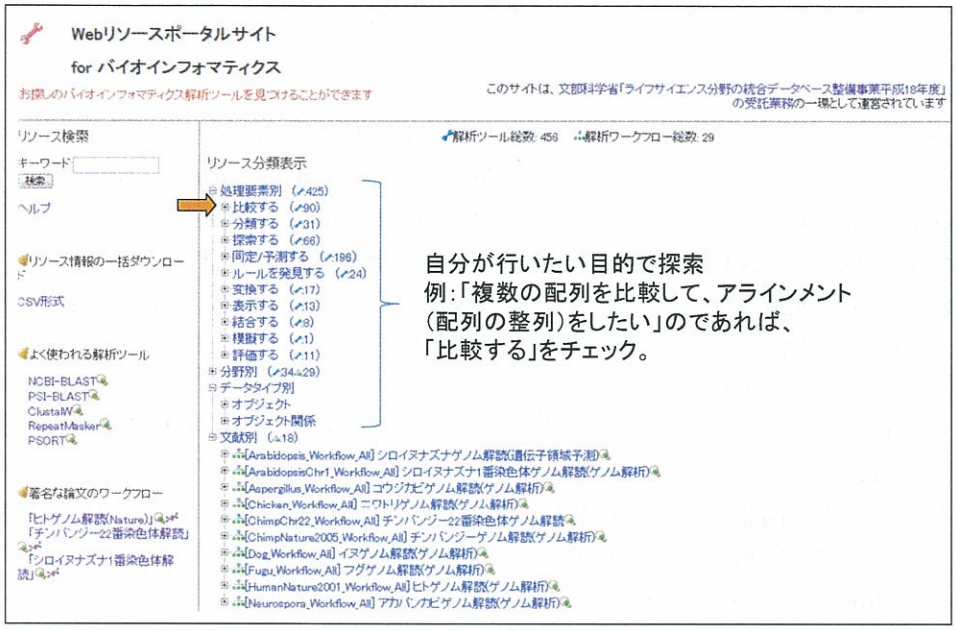


図 1. Web リソースポータルサイトトップページ。(矢印はクリックを示す。)

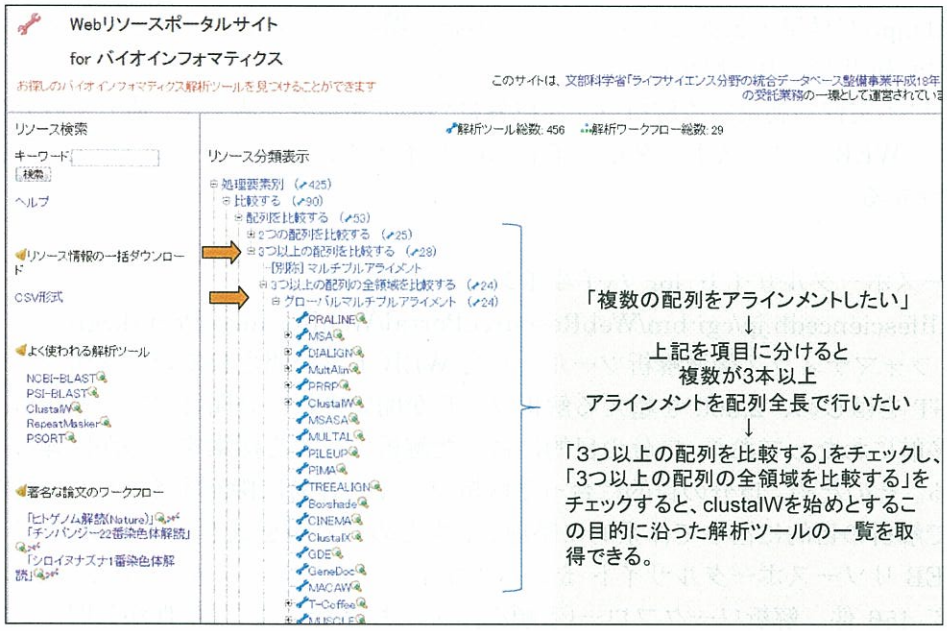


図 2 Web リソースポータル探索例。(矢印はクリックを示す。)

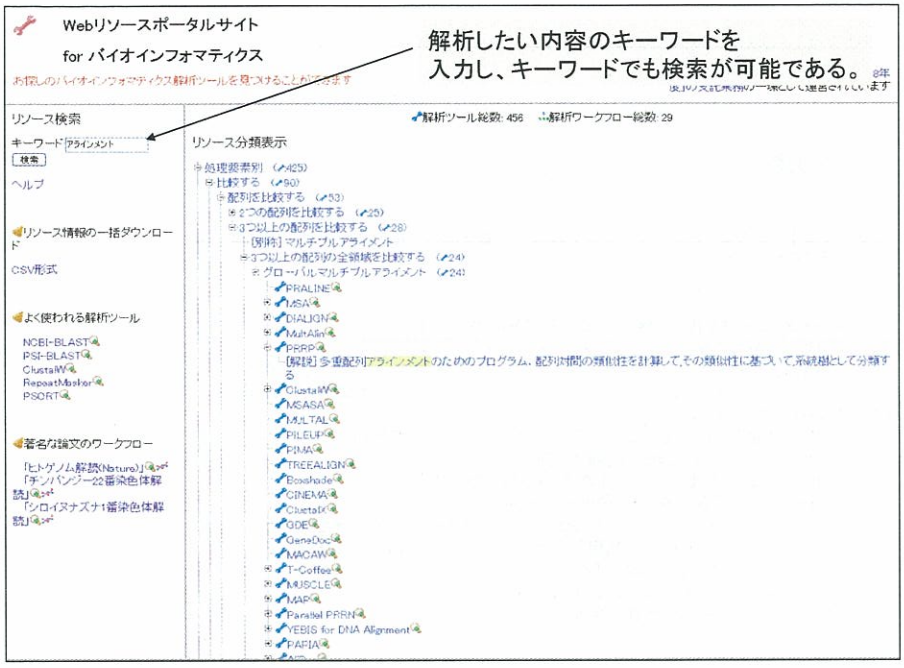


図 3　キーワード検索例。

2．「統合 TV(とうごうてれび)」(http://togotv.dbcls.jp/)

学部学生やバイオインフォマティクス初心者の方にも容易にバイオデータベースやバイオインフォマティクス解析ツールの利用方法を実際の操作手順の動画とともに日本語で紹介を行っている。本サイトに公開されている動画を利用して、バイオデータベースやバイオインフォマティクス解析ツールの使い方について自習が可能である。因みに、Podcast にも登録されており、iPod でも閲覧可能である。利用方法については、以下の画面イメージにて紹介を行う。

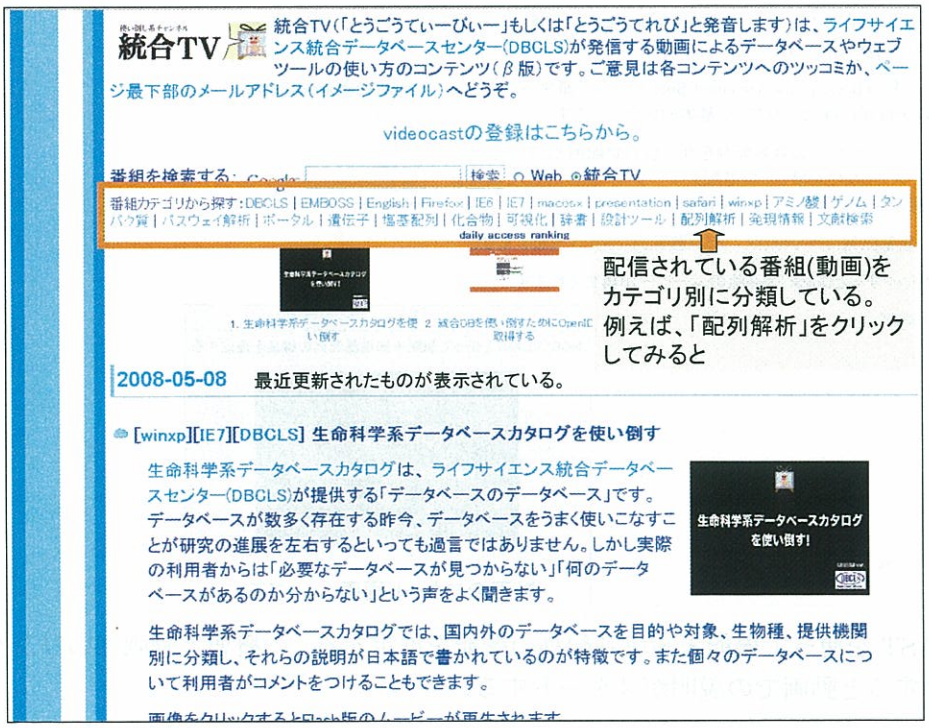


図 4 統合 TV (togo tv)トップ画面

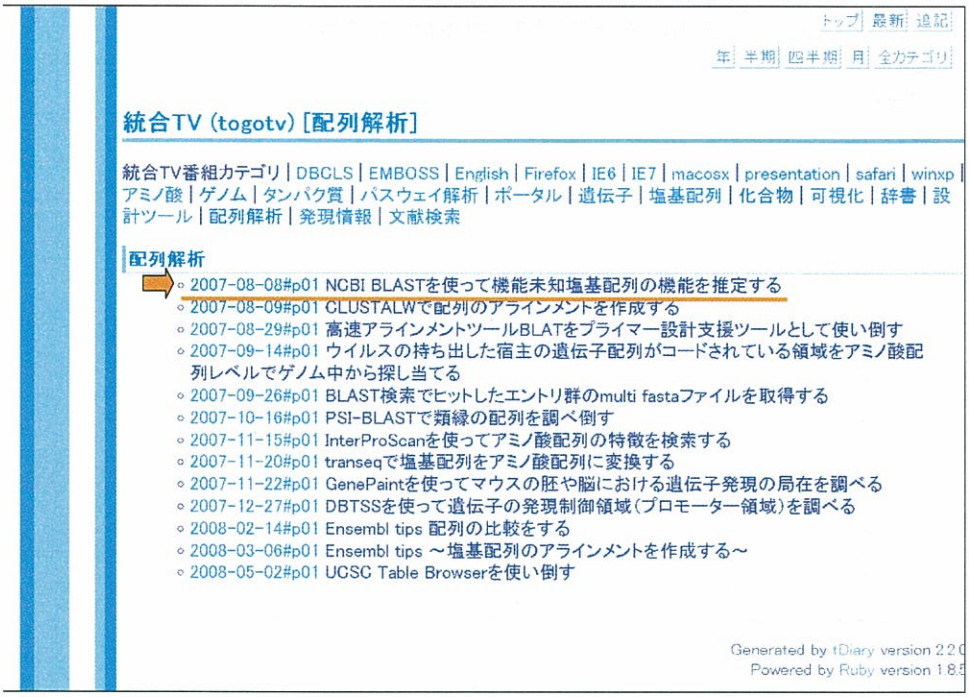


図 5 番組カテゴリ「配列解析」に属する番組一覧。ここで、「NCBI BLAST を使って機能未知塩基配列の機能を推定する」をクリックしてみる。

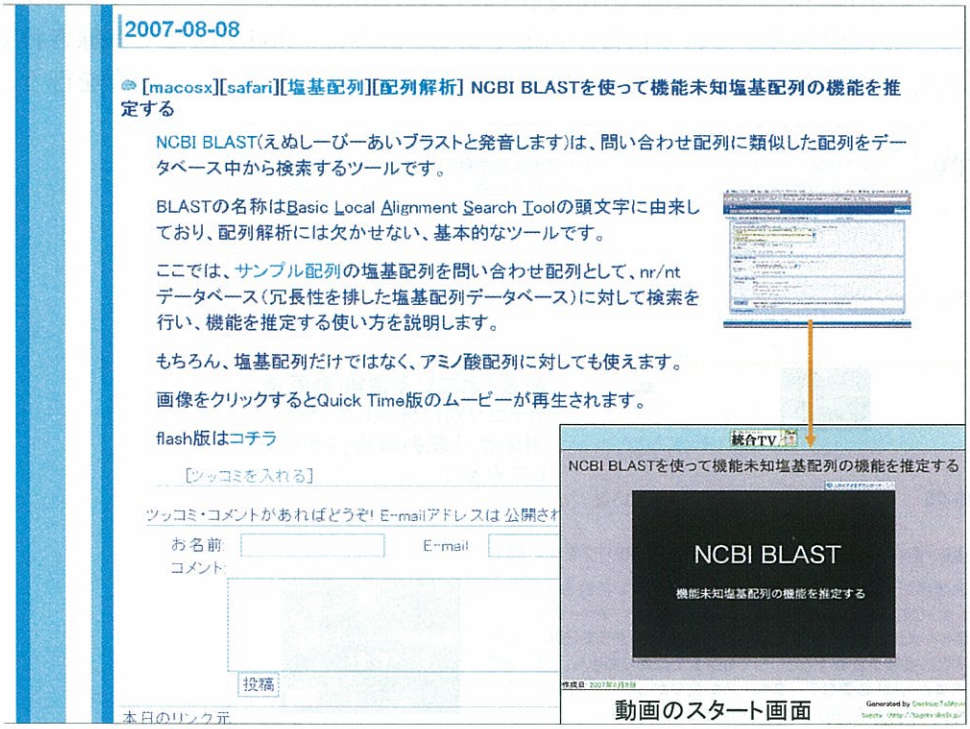


図 7 「NCBI BLAST を使って機能未知塩基配列の機能を推定する」の概要と動画へのリンクがあり、クリックすると動画での説明がスタートする。

以　上