特集　大きなサイトの知らない部分

NCBI

田中　俊典

Ⅰ．はじめに

NCBIとは米国の国立衛生研究所（NIH; National Institutes of Health）の中にある国立医学図書館（NLM ; National Library of Medicine）の一部門で、国立生物工学情報センター（National Center for Biotechnology Information）のことです。1988年の設立以後、生命科学、とくに分子生物学や生物情報科学（bioinformatics）の研究に資するために、種々のデータベースの構築やそれらを利用するためのソフトウェアの開発をしています。これらのリソースの使用は基本的に無償で、全世界の研究者が利用しており、今やNCBIが無ければ全く研究は進まないと言っても過言ではないでしょう。

NCBIにあるたくさんの機能の中でも文献データベースであるPubMedは医学生物学系のデータベースとして最も有名です。PubMedの元のデータはほとんどがMEDLINE（NLMのデータベース）由来なのですが、MEDLINEがカバーしていないデータ、例えばfull text文献へのリンクや古い文献などが含まれています。したがってPubMedはMEDLINEのスーパーセットと言うことができるでしょう[1]。

拙稿ではたくさんあるNCBIの活動の中から、PubMed以外のいくつかを、あまり細かい所には立ち入れませんが、筆者の能力の及ぶ範囲で紹介していきたいと思います。実はNCBIにはチュートリアルやヘルプといったリソースも充実しており[2]、この辺の話はそれらを見ていただいた方が間違いがないので、もし興味がおありの方はぜひ参照していただければと思います。

たなか　としふみ：藍野大学医療保健学部臨床工学科教授
t-tanaka@me-u.aino.ac.jp

Ⅱ．概観

表1はThe NCBI Handbook[3]の目次の項目部分だけを抽出したものです。Part.1のデータベースを見ていただくと、11の項目が並んでいます。

例えばPart.1の最初にGenBankがありますが、これは核酸配列、すなわちDNAやRNAのデータベース（アミノ酸の配列情報も含まれます）で、ヨーロッパ（EMBL; the European Molecular Biology Laboratory）や日本（DDBJ; the DNA Data Bank of Japan）との協力でできています。実際は、この11項目以外にも多くのデータベースがあります。

Part.3にはこれらのデータベースを用いるために有用なツールなどが書かれています。この中でEntrezというのは複数のデータベースに渡って検索できる非常に強力なサーチエンジンです。図1はNCBIの最初のページにある検索ボックスにEntrezでprionという語を入れ、「all databases」を選択して、検索した画面です（prionというのは蛋白質の一種で、ウシの狂牛病やヒトのクロイツフェルト・ヤコブ病に関係することで知られています）。

たくさんのアイコンがありますが、それらがすべてデータベースを表しており、これらのすべてのデータベースからprionという語を検索した結果が表示されています。左の一番上にはPubMedがありますね。アイコンの左にある13484という数字はPubMedを検索して引っか

表1　The NCBI Handbookの目次抜粋

Part 1. The Databases
1. GenBank : The Nucleotide Sequence Database
2. PubMed : The Bibliographic Database
3. Macromolecular Structure Databases
4. The Taxonomy Project
5. The Single Nucleotide Polymorphism Database（dbSNP）of Nucleotide Sequence Variation
6. The Gene Expression Omnibus（GEO）: A Gene Expression and Hybridization Repository
7. Online Mendelian Inheritance in Man（OMIM）: A Directory of Human Genes and Genetic Disorders
8. The NCBI BookShelf : Searchable Biomedical Books
9. PubMed Central（PMC）: An Archive for Literature from Life Sciences Journals
10. The SKY/CGH Database for Spectral Karyotyping and Comparative Genomic Hybridization Data
11. The Major Histocompatibility Complex Database, dbMHC

Part 2. Data Flow and Processing
12. Sequin : A Sequence Submission and Editing Tool
13. The Processing of Biological Sequence Data at NCBI
14. Genome Assembly and Annotation Process

Part 3. Querying and Linking the Data
15. The Entrez Search and Retrieval System
16. The BLAST Sequence Analysis Tool
17. LinkOut : Linking to External Resources from Entrez Databases
18. The Reference Sequence（RefSeq）Database
19. Gene : A Directory of Genes
20. Using the Map Viewer to Explore Genomes
21. UniGene : A Unified View of the Transcriptome
22. The Clusters of Orthologous Groups（COGs）Database : Phylogenetic Classification of Proteins from Complete Genomes

Part 4. User Support
23. User Services : Helping You Find Your Way
24. Exercises : Using Map Viewer

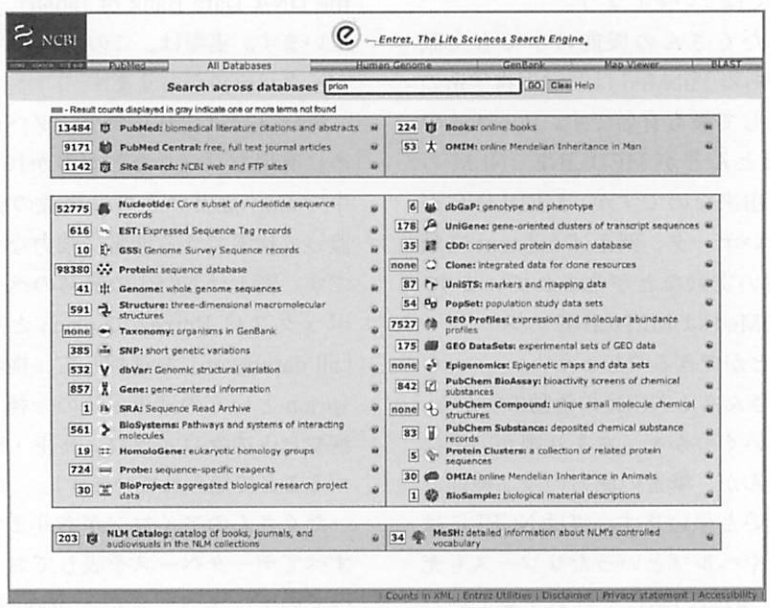


図1　Entrezでprionを検索した画面

かってきた件数を示しています（2012年9月20日現在の件数です）。ここをクリックすればその文献すべてにアクセスすることができるという訳です。このようにCross-database searchができるというのもNCBIデータベースの大きな特徴の一つと言えるかもしれません。また、表1にはありませんが、比較的新しいサービスとして、MyNCBIといって検索利用時に利用者が自分で使用方法に合わせてカスタマイズできるツールなどもあり、どんどん便利に進化しているようです。

Ⅲ．分子生物学の予備知識

データベースの解説をする前に、DNAやRNAといった用語について確認をしておきましょう。少しややこしいですがおつきあいください。

地球上のすべての生命はその遺伝情報をDNAあるいはRNAという高分子の形にして保持しています。DNAすなわちデオキシリボ核酸（RNAはリボ核酸）はヌクレオチドという分子が鎖状につながった長い紐状の分子です。ヌクレオチドには5種類あり、記号でA,G,C,T,Uと表されます。DNAはこのうちA,G,C,Tの4種類、RNAはA,G,C,Uの4種類のヌクレオチドから出来上がっています。このA,G,C,T（あるいはA,G,C,U）の並び方が、つまるところ遺伝情報という訳です。たとえば

・・・AAGCTTTATAGGGCGA・・・

といった具合に延々と繋がっているのがDNA（RNA）なのです。この並びを配列（シーケンス）と呼びます。

さてわれわれヒトのような真核細胞生物は、通常はこの長いDNAをいくつかの部分に切り分けて、くるくると巻き取って細胞の核の中にしまい込んでいます。この一塊を染色体と言います。ヒトの場合は46本の染色体を持っていますが、実はこのうち半分は母親から、残り半分は父親から受け継いだもので、ほとんど同じものなのです（全く同じという訳ではもちろんありません）。つまりほとんど同じDNAを2セット持っているということです。この1セットを「ゲノム」と呼びます。

ところで生物はDNA（言い換えると遺伝情報）を使って何をしているのでしょうか。実はその情報から蛋白質を作っているのです。ヒトの細胞を例にとると、まずゲノムのDNAのシーケンスを、RNAに写し取ります（この操作を「転写：transscript」と言い、できたRNAをメッセンジャーRNAと呼びます）。次にメッセンジャーRNAのシーケンスを元にして、今度はアミノ酸を並べてつないでいきます（これを「翻訳：translation」と言います）。詳細は省略しますが、3つのヌクレオチドに対して1つのアミノ酸が対応するようになっています。アミノ酸がつながった分子ということは、すなわち蛋白質、という訳です。比喩的に言うと、DNAの上に蛋白質の作り方が書いてある、ということになります。注意しておきたいのは、長いゲノムのDNAのうちで蛋白質の作り方が書かれているのは、ほんのごく一部の場所であり、ほとんどの場所は何のためにあるのかよくわかっていません。この蛋白質の作り方が書かれているゲノムDNAの部分を「遺伝子：gene」と言うのです。さらにややこしいことに、遺伝子は飛び飛びの状態で書かれていることが多いのです。そのためRNAに転写された後に、飛び飛び部分をくっつけて一つにする作業が必要です（これをスプライシングと言います）。

それでは細かい話はこの辺にして、Entrezを使って実際にいろいろなデータベースにアクセスしてみましょう。

Ⅳ．核酸配列の検索

上でも触れましたが、NCBIにおける核酸の情報はGenBankというデータベースにまとめられています。GenBankには10万を越える生物種のデータが登録されています。そこにはゲノムのシーケンスもあれば、RNAのデータ、さらにアミノ酸の配列まで含まれています。

さきほどのEntrezの画面（図1）を見てください。ここからprion遺伝子（遺伝子名はPRNPです）を含む、ヒトのゲノムのシーケンスを探してみましょう。左の上から4つ目にnucleotideというデータベースがありますが、これを利用します。しかしこのままではヒット数は52,777件と多いので、「ゲノムのDNA」、「ヒトのデータ」条件を絞ってみます。そうすると41件が抽出されました。その中から説明文を手がかりにして、目的のデータを選び出した結果が図2です（ごく一部です）。

図のa, t, g, cの4つのアルファベットがずらっと並んでいる部分がDNAの配列、すなわちゲノムの遺伝情報を表現しています。chromosome＝"20"と表示されているのは、20番染色体に属するゲノムDNAだという意味です。これでPRNPが、20番染色体のゲノムの中にあることはわかりましたが、実際にこのゲノムの配列のどの部分がPRNPの配列なのかはわかりづらいですね。そこで、ゲノムのデータではなく、遺伝子の配列そのもの（上で述べたように飛び飛びになっている情報も含めて）を知りたいときは、図1の左下にある、geneというデータベースを使えば簡単です。結果は省略しますが、これによってPRNPのDNA配列を知ることができます。

V. 蛋白質の検索

IV. では核酸の配列を調べましたが、NCBIには蛋白質の配列（この場合はアミノ酸配列ということになります）のデータベースもあります。ついでにこれも見ておきましょう。図1のproteinというデータベースからヒトのprion蛋白の配列データを選び出した結果が図3です

```
FEATURES             Location/Qualifiers
     source          1..11079
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /chromosome="20"
ORIGIN
        1 aatgacagtt tgggtgaagc agtatgtcca tgtaaaacca cttttccaaa ccttgcaaat
       61 ccacagtggt gaaatgtgat gcagttatgg tcaatgatat taagcataat atttttggtg
      121 gggcatctgg taagaacctt aaaagagagg aaaaaagtag ttaagaggca ttatccatta
      181 gcctttgcct ttctgctttc ttctgcatga agcatgggca aaatgttgga gtcaatgcag
      241 ccatcttgca aacattcagc aaaagcacag tctaggaatg tattttaaaa agactagaaa
      301 ttgcctcata cttgatatca ttgtgaagct attatagcat tccttgattt tatgttatat
      361 aagaataata aacccctaat ttataaaagg caacattttt catgttttat gtttcagcct
      421 attgtacttg tgaatagatt atgtattata gatcaaatgt gatttagaaa tatagccctt
      481 tctccaatct atcattgatg ggcatttagg ttgattctgt cttttatatt gggaatagtg
      541 ctgcaatgaa catacgcgtg catagaggat caggaaagat aaacaatggg taccaggctt
          ....
```

図2　nucleotideデータベースの検索結果（一部）

```
     CDS             1..253
                     /gene="PRNP"
                     /coded_by="AB300823.1:11..772"
                     /note="PrP 129M"
ORIGIN
        1 manlgcwmlv lfvatwsdlg lckkrpkpgg wntggsrypg qgspggnryp pqggggwgqp
       61 hgggwgqphg ggwgqphggg wgqphgggwg qgggthsqwn kpskpktnmk hmagaaaaga
      121 vvgglggyml gsamsrpiih fgsdyedryy renmhrypnq vyyrpmdeys nqnnfvhdcv
      181 nitikqhtvt tttkgenfte tdvkmmervv eqmcitqyer esqayyqrgs smvlfssppv
      241 illisflifl ivg
//
```

図3　proteinデータベースの検索結果（一部）

(一部です)。

下半分にアルファベットの暗号のような文字列がありますが、この文字一つ一つがアミノ酸を表しています。先ほどの核酸のように4種類ではなく、アミノ酸は20種類あるので、このような文字列になっているのです。

蛋白質は生物の体の中でいろいろな機能を持つのですが、そこで重要なのは立体構造です。立体構造は基本的にはアミノ酸の配列で決まります。つまり、例えば2つの蛋白質の一部分に共通のアミノ酸配列があると、少なくともその部分はよく似た立体構造となり、機能もよく似たものになる(だろう)と考えられるのです。このような共通部分をたくさん集めてデータベース化したものがCDD;conserved protein domain databaseです。図1の右の真ん中より少し上あたりにあります。このデータベースを利用すれば、機能はよくわからないけれどもアミノ酸配列はわかっている蛋白質の立体構造を推定することができ、さらには機能まで推測することができます。逆にある種のアミノ酸配列をもった蛋白質を人工的に作れば、ある特定の機能をもたせることができる(かもしれない)という利用の仕方も可能です。これは薬を設計したりするときに非常に有用です。

Ⅵ. BLAST

今まで見てきた検索というのは、prionという物質の遺伝子の核酸配列やアミノ酸配列を知るための方法でした。しかし逆に、例えば手元に何かわからない核酸の配列データがあるのだけれど、これが一体何の配列なのかわからないといった場合はどうすればよいのでしょうか。あるいは、ある詳細不明な蛋白質を分析して、アミノ酸配列は知ることができたけれども、この蛋白質はどういった蛋白質なのか、それとも全く新しい蛋白質なのかを知りたい。そういうときのためにNCBIではBLASTというツールを用意しています。BLASTとはBasic Local Alignment Search Toolのことです。

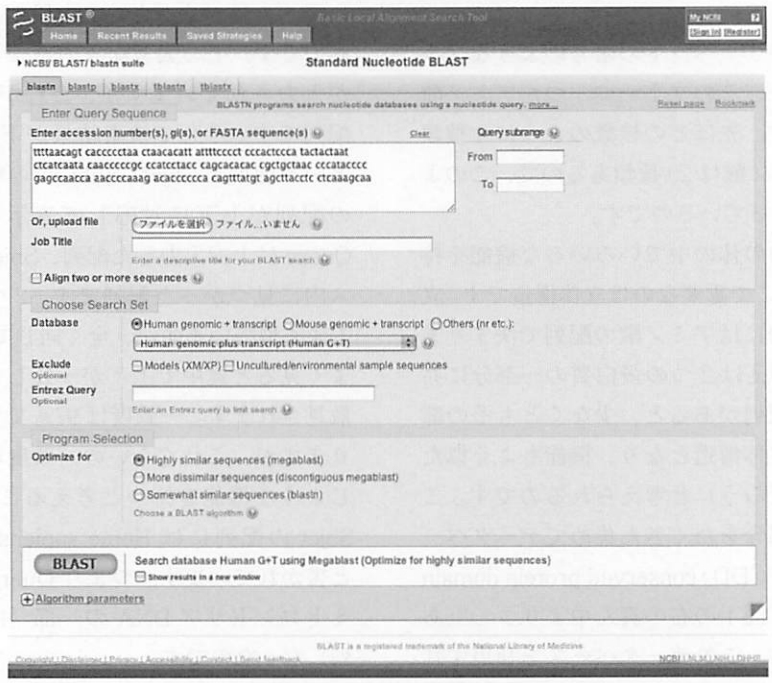
図4を見てください。上段がBLASTの検索画面です。上の大きな入力ボックスに核酸配列が入力されていますが、これが調べたいDNA配列です。そして検索結果が下段です。Queryという配列とSbjct(Subjectの略)という2つの配列が上下に対応して表示されています。Queryは上で入力した配列、Sbjctがデータベース内に見つかった配列です。2つの配列はとてもよく似ていますが、全く同じではありません。よく見ると途中で上下が一致していない場所が散見されます。一致度は97%です。ものにもよりますが、これぐらいの一致度ならば、ほぼ同じような機能を持つと考えることもできます。Sbjctの配列にはHomo sapiens mitochondrionと書かれています。つまりQuery配列はヒトのミトコンドリアDNAの一部(にとても近い配列)だったのです。

実はこの例で用いたQuery配列もGenBankからとってきた配列でした。何の配列かと言いますと、Homo sapiens neanderthalensis つまりネアンデルタール人のミトコンドリアゲノムだったのです。そのため現代人の配列との間には微妙な違いがあったわけです。それにしても化石から採取されたデータまであるなんて、とても面白いですね。ちなみに、こういったデータの蓄積によって、ネアンデルタール人はわれわれ人類の直接祖先ではなく、同時代に生息していた別の種だったと考えられるようになったということです。

Ⅶ. 病気との関連を調べる

最後に遺伝疾患に関するデータベースを紹介します。OMIM: Online Mendelian Inheritance in Manというもので、図1の右上の方にあります。ここをクリックするとprionに関連した疾患についての詳細な情報を得ることができます(図5)。

図5の下の方にはCreutzfeldt-Jakob diseaseやGerstmann-Sträussler diseaseといったprion関連の疾患が見えます。病気そのものの臨床病



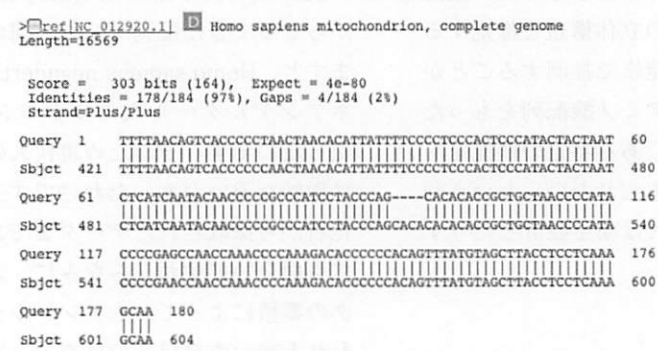


図4　BLASTによる検索

像など詳細が知りたければ、ここから簡単に参照することができますし、解説の文書中の引用文献はPubMedと、配列情報はGenBankとリンクしていますので効率よく知識を得ることができます。

Ⅷ. おわりに

以上、たくさんあるNCBIの機能の中から、いくつか紹介させていただきましたが、これはほんの入り口に過ぎません。それぞれのデータベースを本当に使いこなすためには専門的知識と専門的ニーズを持っていないと難しいと思います。とはいえ、データベースは誰にでも無料で利用できますし、仕事ではなくぶらりとNCBIを訪れて、気の向くままに遊んでみるというのもなかなか楽しいものです。そんな楽しい世界の一端を拙稿で感じていただければ幸いです。

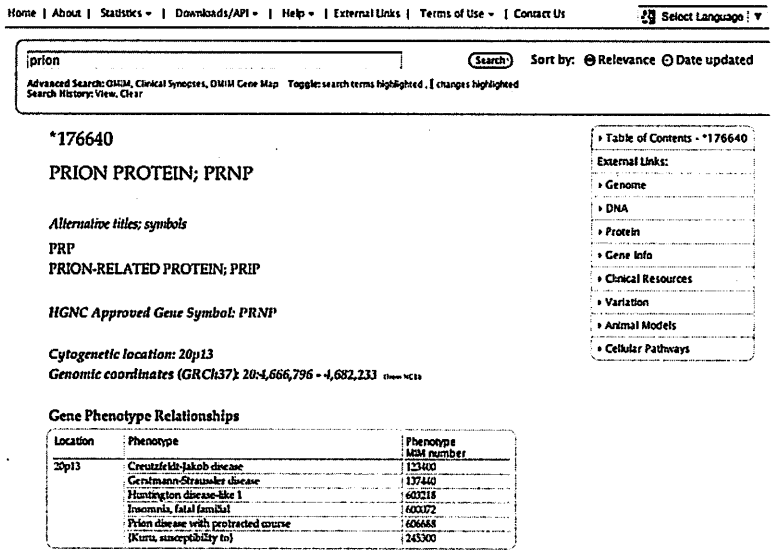


図5　OMIM（一部）

参考文献

1）PubMed と MEDLINE の違いについて. [引用 2012-09-20]
http : //www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html

2）NCBI Mini Courses. [引用　2012-09-20]
http : //www.ncbi.nlm.nih.gov/Class/minicourses/
日本語訳を科学技術振興機構が公開している；
http : //www-bird.jst.go.jp/minicourses/

3）The NCBI Handbook. [引用　2012-09-13]
http : //www.ncbi.nlm.nih.gov/books/NBK21101/