How to Search for a *C. elegans* Homologue of Your Gene

線虫の遺伝子ホモログの検索

Kondo, Kazunori

Department of Bioengineering, Faculty of Engineering, Soka University,
Tangi-cho, Hachioji, Tokyo 192-8577, Japan
FAX: 81-426-91-9312, E-mail: kondo@scc1.t.soka.ac.jp

Key Words: *C. elegans*, EST, genome sequence, homologue

Abstract

A nematode *Caenorhabditis elegans* has been used as a model organism for the study of animal development and neurons. Recently, essentially complete DNA sequence of the genome was determined and published [Science (1998), **282**, 2011–2045]. *C. elegans* has become of interest in studying the genes whose biological functions are unknown to biologists who are not studying *C. elegans*, because not only classic genetics but also reverse genetics such as gene knockout can be used in *C. elegans*. In this manuscript I will briefly explain the methods of searching for the *C. elegans* homologue of your interested genes using the Internet. How to use DDBJ has already been described [TIGG (1999), **11**, 119–127]. Here I write about the homepages of Washington University Genome Sequencing Center, The Sanger Center and ACeDB.

要約

線虫 *C. elegans* は、発生や神経系を研究するためのモデル生物としてよく用いられていますが、最近、ゲノムの全塩基配列がほぼ決定され、Science (1998), **282**, 2011–2045 に発表されました。*C. elegans* は、遺伝子をノックアウトする逆遺伝学的手法や古典遺伝学的手法も使えることから、生物学的機能の明らかでない遺伝子を研究するためのモデル生物として、線虫研究者以外からも注目されてきています。

ここでは、自分の研究している(興味のある)遺伝子のホモログが、線虫に存在しているかどうか、インターネット上で検索するにはどうしたらよいのかについて簡単に説明します。DDBJ を用いる方法についてはすでに紹介されています[TIGG (1999), **11**, 119–127]ので、ワシントン大学及びサンガーセンターのサーチサービスを用いる方法について説明します。

A. BLAST Search of Washington University Genome Sequencing Center

Start up your browser application such as Netscape and directly access the BLAST search of Washington University Genome Sequencing Center (<http://genome.wustl.edu/gsc/cgi-bin/blast/client.pl>) or click the “BLAST” on the home page of Washington University Genome Sequencing Center (<http://genome.wustl.edu/gsc/index.shtml>). It is not necessary to register or obtain a password. Sequences of the genes you have determined or obtained through papers or the Internet can be analyzed for the *C. elegans* counterparts by homology search with the BLAST search program.

In the case where your sequence is DNA, select “BLASTN (DNA query sequence)” at BLAST type (Fig.1.①). For amino acids sequence, select “TBLASTN (protein query sequence)”. Then select *C. elegans* genomic database “*C. elegans* (all; both St. Louis and Sanger)” at Fig. 1.②. You may choose “*C. briggsae* (St. Louis project only)” “*Arabidopsis thaliana* (St. Louis project only)” or “Human (St. Louis project only)” instead. Select options: Default position of DUST/SEG filter that eliminates false high score due to polyA etc. is on. If this is not desired, select “Do not filter query sequence with DUST/SEG”

A. ワシントン大学のBLASTを用いる

ネットスケープ等のブラウザを用いて、ワシントン大学 Genome Sequencing Center のホームページ (<http://genome.wustl.edu/gsc/index.shtml>) にアクセスし、Search service の BLAST をクリックするか、あるいは直接 BLAST Server のページ (<http://genome.wustl.edu/gsc/cgi-bin/blast/client.pl>) にアクセスします(図1)。パスワードの入手や登録の必要はありません。自分が塩基配列を決定した遺伝子、あるいは論文やインターネットを通じて取得した遺伝子などと相同性を持つ線虫 *C. elegans* の遺伝子を、相同性検索プログラム BLAST によって解析できます。

まず、検索したい配列が DNA 塩基配列の場合、BLAST type(図1 ①)のところで BLASTN (DNA query sequence) を選択します。アミノ酸配列の場合は、TBLASTN (protein query sequence) を選びます。次に database(図1 ②)を選びます。ここでは *C. elegans* のゲノム塩基配列データベース、*C. elegans* (all; both St. Louis and Sanger) を選びますが、他にも *C. elegans* と近縁種の *C. briggsae* (St. Louis project only)、*Arabidopsis thaliana* (St. Louis project only)、Human (St. Louis project only) も選択可能です。次に検索オプションを指定します。Filter(図1 ③)は、デフォルトでは DUST/SEG プログラムでフィルターをかけ、ポリ A などによる相同性を除く設定になっていますが、フィルターをかけたくない場合には Do not filter query sequence with DUST/SEG を選

Fig. 1. Washington University Genome Sequencing Center BLAST Server page.

at Fig. 1.③.; Default of Repeat masker which masks repetitive sequences such as Line/L1 is also on. If this is not desired, select "Do not mask query sequence with RepeatMasker" at Fig. 1.④. The search results will be displayed in the browser. In case it takes a lot of time because of too many users, you may select Email for the results. Select "Email me the BLAST output instead" at Fig. 1.⑤ and type your Email address in the box (Fig. 1.⑥). Input your sequence in "Please enter your sequence here" (Fig. 1.⑦) in FASTA format with copy/paste and Click "Submit" (Fig. 1.⑧).

B. BLAST Search of The Sanger Center

BLAST Search of The Sanger Center has cDNA (EST) databases in addition to genome databases. The genome database of The Sanger Center and that of Washington University contain basically the same data except for a few latest ones, as those centers exchange data with each other. To use this browser, access the homepage of The Sanger Center (<http://www.sanger.ac.uk/>) and click "C. elegans" at Database Search

択します。RepeatMasker(図1 ④)は、Line/L1 等のあらかじめ登録された繰り返し配列をマスクするプログラムで、デフォルトではオンになっています。もしこれが望ましくない場合は Do not mask query sequence with RepeatMasker を選びます。さらに、デフォルトでは検索結果はブラウザに表示されますが、Email を指定すると検索結果が指定したアドレスに電子メールで送られてきます。使用者が多い、あるいは回線が込み合っていて時間がかかるような場合にはオプション(図1 ⑤)で Email me the BLAST output instead を指定し、図1 の⑥の Please enter your email address のところに自分の Email アドレスを入力するとよいでしょう。最後に、Please enter your sequence here(図1 ⑦)の枠内に自分の配列を FASTA フォーマットでタイプするか copy-paste 機能を使って入力し、Submit(図1 ⑧)をクリックします。

B. サンガーセンターのBLASTを用いる

サンガーセンターの BLAST Search では、線虫ゲノム配列だけでなく cDNA (EST)データベース検索もできます。ゲノムの塩基配列に関しては、ワシントン大とサンガーセンターは互いにデータを交換しあっているため、わずかな時間のずれはありますが基本的には同じデータベースです。まず、サンガーセ

Fig. 2. The Sanger Center BLAST Server page.

Service, or directly access The Sanger Center *C. elegans* BLAST server (http://www.sanger.ac.uk/Projects/C_elegans/blast_server.shtml) (Fig.2). The operation of this browser is similar to that of Washington University. First, select database (Fig.2. ①): select "*C. elegans* genomic" when you search *C. elegans* genomic DNA sequences; select "*C. elegans* EST" for cDNA database; you may select the peptide database "Wormpep" which is predicted from DNA sequences or transposon Tc1 sequence-tagged sites "Tc1 STS". Next, click one of the circles (Fig. 2.②) to select a suitable program for your query sequence and database selected. For example, when you have a DNA sequence for query and select "*C. elegans* genomic" for the database, you can select either "BLASTN" or "TBLASTX". Input maximum number of alignments you want to be reported (Fig.2.③), whose default number is 100. If you do not need the filter, click the box (Fig.2.④). Then copy & paste your query sequence with FASTA format in the box (Fig.2.⑤) or specify the file name of your sequence located on the accessible disk with Fig. 2.⑥. Results would be sent to the browser on default. You may choose

センターのホームページ(<http://www.sanger.ac.uk/>)にアクセスし、Database Search Service の *C. elegans* をクリックするか、あるいは直接 The Sanger Center *C. elegans* BLAST server のページ(http://www.sanger.ac.uk/Projects/C_elegans/blast_server.shtml)にアクセスします(図2)。使い方はワシントン大のものと同様で、最初に database(図2 ①)を選びます。ゲノム DNA を検索する場合は、*C. elegans* のゲノム塩基配列データベース、*C. elegans* genomic を選び、cDNA を検索したい場合は、*C. elegans* ESTs を選びますが、他にゲノムの DNA 配列から予想されたペプチドのデータベース(Wormpep)、及びトランスポゾン Tc1 sequence-tagged sites (Tc1 STS)も選択可能です。次に、図2の②のところのサークルをクリックして、手持ちの配列と選んだデータベースに適したプログラムを選びます。例えば、手持ちの配列が DNA でデータベースに *C. elegans* genomic を選んだ場合には、BLASTN または TBLASTX が選択可能です。図2の③は、表示されるアラインメント数の上限を指定します。デフォルトは 100 になっています。フィルターをかけたくない場合は、図2④をクリックします。そして、自分の検索したい配列を FASTA フォーマットで、図2の⑤の枠内に copy-paste 機能を使って入力するか、または図2の⑥をクリックして、アクセス可能なディスク上にある目

BLAST Response Follows: Job ID 19714

Here are your results as generated by the BLAST software.

TBLASTN 2.0MP-WashU [10-Aug-1999] [linux-alpha 22:33:14 11-Aug-1999]

Copyright (C) 1996-1999 Washington University, Saint Louis, Missouri USA.
All Rights Reserved.

Reference: Gish, W. (1996-1999) <http://blast.wustl.edu>

Notice: statistical significance is estimated under the assumption that the equivalent of one entire reading frame of the database codes for protein and that significant alignments will involve only coding reading frames.

Query= NO_NAME_SPECIFIED
(559 letters)

Database: /usr/databases/alleleaid

4212 sequences; 109,902,956 total letters.

Searching... 10...20...30...40...50...60...70...80...90...100% done

Sequences producing High-scoring Segment Pairs:				Smallest Sum	
	Reading Frame	High Score	Probability P(N)	N	
ZK688	-3	1024	1.8e-131	4	
H38K22/980630	-3	538	2.1e-95	6	
V39E4B/981221	+2	388	2.2e-91	6	
V39E4	Contig ID=07041; Length=340806; Order=U...	+3	388	9.5e-90	6
V46H3A	-1	486	4.4e-87	3	
V45F10D/971013	+3	620	6.0e-78	5	
V47D3A/990115	-1	272	2.4e-68	6	
V47D3	Contig ID=05154; Length=341258; Order=U...	-1	272	2.9e-67	6
V79H2	Contig ID=04089; Length=249626; Order=U...	-3	337	3.4e-55	5
V75B8A/981022	-1	337	5.7e-55	5	
V116F11	Contig ID=03144; Length=6097; Order=Unk...	+3	378	4.2e-37	2
V116F11	Contig ID=04062; Length=2255; Order=Unk...	-1	376	3.3e-31	1
V66A7A/980825	+1	187	1.9e-30	4	
V116F11	Contig ID=03707; Length=4539; Order=Unk...	-2	147	1.3e-06	1
C33R11/960315	-2	83	0.99994	1	

ZK688
Length = 36,977

Ninus Strand HSPs:

Score = 1024 (365.5 bits), Expect = 1.8e-131, Sum P(4) = 1.8e-131
Identities = 185/333 (55%), Positives = 244/333 (73%), Frame = -3

Query: 175 KUPVHUHANEGRSGLIARALKGARUSKQQUITFLDANCECTUGMLEPLLRKHORATUV 234
++P+H++ +E RSGLIARAL G+ ++KG+++ FLDAH E T GMLEPL++R+ DR+ UU
Sbjct: 34650 QIPHLVHLENRSGLIARALTGSEHAKGKILLFLDAHVEUTDGLLEPLVSRAVEDRKRUV 34471

Query: 235 CPIIDUISDQTFEVHAGSDHTYGGFHWKLNFRVYVUPQREHARKGORTLPURTPTNAGG 294
PIIDUISDQTFEV+ S+ T+GGFHW LNFRVYV UP+RE++R+ DR++P++TPT+AGG
Sbjct: 34470 APIIDUISDQTFEVATSETTGGFHWKLNFRVYVUPKRELNAGSDRSNP1QTPTIAGG 34291

Query: 295 LFSIDRDYFQEIQTVDAGMDIHGGENLEISFRHWCGGTLEIVTCSHUGHUFKATPVTF 354
LF+ID+ +F +IG+VD GN +HGGENLEISFRHW CGG+LEI CS UGHUFK TPVTF
Sbjct: 34290 LFRIDKOFFYDIGSYDEGHQVHGGENLEISFRHWCGGSLEIHPCSRUGHUFKQTPVTF 34111

Fig. 3. Example of search result with WU BLAST Server.

Email instead for the results to be sent; click “x” (Fig.2.⑦) and put your Email address in the box (Fig.2.⑧). After all parameters are set, click “Submit Query” (Fig.2.⑨).

To search for cDNA database, select “*C. elegans* EST” in the Sanger Center BLAST Search or use DDBJ BLAST Search.

C. Results of The Search and How to Obtain Sequences

An example of search results is shown in Fig. 3, in which TBLASTN program on the Washington University BLAST Server was used, the query sequence was human “Polypeptide N-acetylgalactosaminyltransferase (GalNAC-T1) (Accession Number: Q10472)”, and database was “*C. elegans* (all; both St. Louis and Sanger)”. Fifteen sequences were obtained with their alignments (not all are shown). The *C. elegans* genome database consists of sequences of independent cosmid clones. Each sequence of these clones has been sent to databases such as EMBL or GenBank, too. To obtain those sequences, access Entrez Nucleotide Sequence Search (<http://>

的の配列のファイル名を指定します。さらに、デフォルトでは検索結果はブラウザに表示されますが、図2の⑦をクリックして、⑧の枠内に自分の Email アドレスを指定すると、検索結果が指定したアドレスに電子メールで送られてきます。パラメーターをすべて入力し終わったら Submit Query(図2 ⑨)をクリックします。

このサンガーセンターの BLAST Search でデータベースに *C. elegans* EST を選ぶか、あるいは DDBJ を用いて *C. elegans* EST を選べば、cDNA 配列データベースを検索することができます。

C. 検索結果と配列の入手

検索結果の例を図3に示します。ここでは、ワシントン大学の BLAST Server で、ヒトのN-アセチルガラクトサミン転位酵素 Polypeptide N-acetylgalactosaminyltransferase (GalNAC-T1) (ACCESSIONナンバー：Q10472)のアミノ酸配列を用いて、プログラムは TBLASTN を選び、線虫の対応する遺伝子を検索しました。15 個の配列がひっかかってきました。図3ではカットしてありますが、それぞれの配列の、アミノ酸配列でのアラインメントが示されています。これら BLAST Server 中の線虫のゲ

Fig. 4. Example of search result with Sanger Center BLAST Server.

www.ncbi.nlm.nih.gov/Entrez/nucleotide.html), for example, and input the clone name "ZK688" as a keyword for the search to download the sequence. In case the clone is not submitted to EMBL or GenBank yet, use ACeDB (see below).

Figure 4 shows the result by The Sanger Center *C. elegans* BLAST Server using the same query sequence and parameters as used in Fig. 3. Sixteen clones were found, which is more than that picked up in Washington University BLAST search. In this page (Fig.4), you can just click the names of the clones with underlines to access the DNA sequence. To obtain the sequences of the clones not underlined, use ACeDB (see below). Comparing these two results (Fig.3. & Fig.4), you will notice that the BLAST result of The Sanger Center has two more clones, 9th Y32G9.Contig365 and 15th Y32G9.Contig361, which are not in the BLAST result of Washington University. Instead, the C33A11 is found only in the BLAST result of Washington University. The reason why the C33A11 was picked up only by the BLAST result of Washington University is probably because

ノム配列データベースは、塩基配列決定時に用いられた個々のコスミドクローンごとに独立に入力されています。個々のクロンの塩基配列は、完成版ができ次第 EMBL,GenBank 等のデータベースにも送られています。それぞれの配列を入手するためには、例えば、Entrez Nucleotide Sequence Search (http://www.ncbi.nlm.nih.gov/Entrez/nucleotide.html)等のページにアクセスして、クロンの名前 ZK688 をキーワードとして入力し、目的の塩基配列をダウンロードします。EMBL,GenBank 等のデータベースにはまだ登録されていないクロンの場合は ACeDB (後述)を使用します。

図4は、同じ検索をサンガーセンターの *C. elegans* BLAST serverで行ったもので、ワシントン大学の BLAST Server の時より1つ多い16個がひっかかっています。こちらのページでは、アンダーラインのあるクロンには、そこをクリックするだけで塩基配列にアクセスすることができます。アンダーラインのないクロンは ACeDB を用います。これら2つの検索結果をよく見比べてみると、サンガーセンターのサーチでは、9番目の Y32G9.Contig365 と 15番目の Y32G9.Contig361(図4)が余分で、かわりにワシントン大の 15番目の、C33A11/960315(図3)

Fig. 5. AceDB simple search.

these two searches have different cut-off values of the score to be shown on the browsers. Analysis using ACeDB revealed that the clone Y32G9 which overlaps with the Y46H3 (5th clone) and thus this is another clone corresponding to the same region of the genome. These two redundant clones appear to be already erased from the Washington University database, but not from The Sanger Center database yet. The genome sequencing of *C. elegans* is reported to be nearly completed (>99%). Analyses of these data, however, are still in the process. Multiple clones corresponding to the same regions of genome might be in the database. These redundancies will soon be cleared, but you should be careful of these redundancies when you analyze the search.

D. Analyses of the *C. elegans* Genes

In order to judge whether the sequences picked up by the BLAST searches are the *C. elegans* homologues of your genes or not, it would be convenient to use ACeDB. ACeDB was originally developed for the *C. elegans* genome project, from which its name was derived (A *C. elegans* DataBase). However, the tools in it have been generalized to be much more flexible and the same software is now used for many different genomic databases from bacteria to fungi to plants to man. It is also planned to update, upgrade and expand it into a comprehensive information resource. As ACeDB has multi-functions, here I explain

が欠けています。C33A11 がワシントン大のサーチでだけ引かかる理由は、カットするスコアの最小値の違いだと思われます。サンガーセンターの9番目と15番目のY32G9は、ACeDBで調べてみますと、5番目のY46H3Aとオーバーラップするクローンで、結局、ゲノム上の同じ領域に対応する別クローンであることがわかります。ワシントン大のデータベースではすでに整理されて1つにまとめられていたクローンが、サンガーセンターでは整理が少し遅れているようです。*C. elegans* ゲノムの全塩基配列はほぼ(>99%)決定されたと発表されましたが、BLAST Server上のデータは完全には整理されていないようで、ゲノム上のまったく同じ領域のDNA配列が異なった名前で複数はいっていることがあるようです。徐々に改善されていくと思いますが、その点は自分で確認する必要があります。

D. 線虫の遺伝子の解析

BLASTサーチで得られた検索結果を解析して、ひっかかった配列が目的の遺伝子に対応する線虫の遺伝子ホモログかどうかを判断するわけですが、ACeDBを用いると便利です。ACeDBは、最初はCエレガンスのゲノムプロジェクトのために作られたプログラムで、A *C. elegans* DataBaseの略です。現在では、線虫だけでなくヒトやシロイヌナズナのゲノム解析等、広い分野に応用されています。線虫のACeDBも、塩基配列だけでなく、文献や突然変異体の検索等、様々な機能を持ち、さらに拡張される計画もスタートしたようです。非常に多機能なので、ここでは最も簡単なSimple search機能だけを説明します。ま

[Data release policy](#) and [Guidelines and conditions on use of data](#)

The Sanger Centre

| [WormHome](#) | [Info](#) | [HGP](#) | [Projects](#) | [Database Searches](#) | [Software](#) | [Teams](#) | [Search](#) |

C. elegans Genome Database Please take a moment to complete the *C. elegans* database user survey

[Simple Search](#) [AltaVista Search](#) [Class Browser](#) [Text Search](#) [Ace Query](#)

C. elegans Simple Search

Select the type of object you are looking for and optionally type in a name or a wildcard pattern (? for any one character, * for zero or more characters). If no name is entered, the search displays all objects of the selected type. *Anything* searches for the entered text across the entire database.

☒ Anything ☐ Author ☐ Strain ☐ Name:
☐ Gene ☐ Genetic Map ☐ Clone
☐ Sequence ☐ Sequence Map

35 objects found

0	10	20	30
1	Clone: ZK688	Peptide: WP:ZK688.4	Sequence: ZK688.7 Sequence: ZK688.5
2	Clone: YK281	Peptide: WP:ZK688.5	Sequence: ZK688.6 Sequence: C07A9.8
3	DNA: ZK688	Protein: WP:ZK688.3	Sequence: C02C2.4 Sequence: ZK688.6
4	Locus: slv-3	Protein: WP:ZK688.4	Sequence: ZK688.3 Strain: BC4544
5	Peptide: WP:ZK688.6	Protein: WP:ZK688.5	Sequence: ZK688 Uni: CELZK688
6	Peptide: WP:ZK688.1	Protein: WP:ZK688.6	Sequence: ZK688.1
7	Peptide: WP:ZK688.7	Protein: WP:ZK688.7	Sequence: ZK688.2
8	Peptide: WP:ZK688.2	Protein: WP:ZK688.8	Sequence: ZK688.3
9	Peptide: WP:ZK688.8	Protein: WP:ZK688.2	Sequence: ZK688.4
10	Peptide: WP:ZK688.3	Protein: WP:ZK688.1	Sequence: C02C2

[For technical information about these pages see:](#)
[AceDB Home Page](#)
[AcePerl Home Page](#)

[Privacy Statement](#)
[www.broadinstitute.org/ce/ace">www.broadinstitute.org/ce/ace](#)

Fig. 6. Example of AceDB Simple search result.

only how to use "Simple search". First, access "Simple search" (<http://stein.cshl.org/perl/ace/simple/elegans> (Fig.5) or <http://stein.cshl.org/perl/ace/simple/elegans>). If you need only DNA sequences of the clone, select "Sequence", but usually you may need other useful information. So select "anything" (default parameter) and input a clone name in the box. The search result using "ZK688" as a keyword, which showed the highest score in the BLAST searches (Fig.3 & Fig.4) is shown in Fig. 6. The complete DNA sequences of ZK688 will be obtained when you click "Sequence: ZK688" (Fig.6.①) and click "Genomic DNA (36977 bp)" in the coming page (not shown). Further analysis with ACeDB (details of the methods are omitted here), revealed that ZK688 is located in the middle part of chromosome III: a peptide corresponding to the UDP-GALNAC homologous region of ZK688 is predicted as Protein: WP:ZK688.8 (Fig.6.②) and three cDNA clones, cm17a3, yk139e11.3, and yk139e11.5 derived from this region. Mutants corresponding to this region would be described if they were identified, but no mutant corresponding to Protein: WP:ZK688.8 is reported yet. Considering all these, this Protein: WP:ZK688.8 gene is probably a *C. elegans* counterpart (homologue) of human UDP-GALNAC, because of its high homology with mammalian UDP-GALNAC and existence of cDNAs. However, you cannot say it is really a UDP-GALNAC of *C. elegans*, since no data is available for its enzyme activity or mutants. Among 15 candidates obtained (Fig.3),

ず、Simple search のページ(<http://stein.cshl.org/perl/ace/simple/elegans>(図5)または<http://stein.cshl.org/perl/ace/simple/elegans>)にアクセスします。塩基配列だけがほしい場合は、Sequence を選べばよいのですが、他にも役に立つ情報が得られるので、デフォルトのanythingを選んで、枠内にクローンの名前を入力しリターンキーを押します。例えば上のサーチの結果で、スコアの最も高かったクローン ZK688 を入力すると、図6の結果が返ってきます。このクローンの全塩基配列は、アンダーラインのある Sequence: ZK688(図6 ①)をクリックし、さらに次の画面(示していません)の、アンダーラインのある Genomic DNA (36977 bp)をクリックすれば得られます。さらに、詳しくは述べませんが、simple Search で得られた他の項目をたどって見てみると、ホモロジーサーチでひっかかった部分は、3 番染色体の中央部に位置し、Protein: WP:ZK688.8(図6 ②)という名で同定(予想)されており、ヒト、ウシなどのUDP-GALNAC とホモロジーがあり、cDNAクローンcm17a3, yk139e11.3, yk139e11.5がこれとマッチすることなどが解ります。変異体は、同定されていれば記載があると思われますが、このProtein: WP:ZK688.8に対応するものはまだ記載されていません。結局このProtein: WP:ZK688.8遺伝子に関しては、哺乳類のUDP-GALNAC とのホモロジーが高く、対応する mRNA も発現しているので線虫の UDP-GALNAC ホモログである可能性は高いと思われますが、酵素活性や変異体のデータがないので決め手には欠けるといえるところです。図3で得られた他の候補者のうち、2 番目の H38K22、10 番目の

three clones, H38K22 (2nd), Y75B8A (10th) and Y66A7A (13th) have comments as "similar to Glycosyl transferases". Other clones have no comments for their identity. You should judge yourself whether they are real *C. elegans* homologues of your gene or not. Also it should be noticed that the identities described there are often putative.

Some programs such as GeneFinder can predict the exon-intron structure of *C. elegans* genome sequences with considerable fidelity. Most of the polypeptide (start and stop codons, and exon-intron structure) written at "FEATURES" in GenBank reports have been predicted by GeneFinder and they are not determined by comparison with cDNA sequences (they have comments as "putative"). Some of the predicted genes were compared with cDNA sequences and confirmed, but not all the predictions were 100% correct. You should analyze yourself especially when you find genes totally unrelated to your genes are identified in the same region where your query sequence has homology. In ACeDB, correspondences of genomic DNA sequences and EST are not completely described, because it takes time to analyze the homologies. So it is recommended that you carry out searches for both genome and EST. When you first obtain an EST, do the searches for genomic DNA described above by using the EST as a probe, since cDNA sequences are usually not complete. When you have a genomic DNA sequence first, search EST as above. In the example case above, ACeDB contained many useful data for the clones picked up by the BLAST search. This is not always the case for many genes you might want to be analyzed. For a lot of clones in the database of ACeDB including a few of the clones shown in Fig.3 and 4, only DNA sequences are shown with no peptide prediction. Also cDNAs of genes with low expression might not be found yet. In such cases, you could analyze possible polypeptide sequences yourself with ORF-predicting programs such as BCM GeneFinder (<http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html>). Then you could judge the possibility that the gene might be a homologue of your gene, comparing the predicted sequence with your gene. It may be superfluous to say that genes with low homology with your gene cannot to be identified as a *C. elegans* homologue of your gene by these computer analyses alone.

Y75B8A と 13 番目の Y66A7A は、similar to Glycosyl transferases との記載がありますが、その他のクローンには何も記載されていません。これらが UDP-GALNAC 遺伝子のホモログかどうかに関しては、記載のあるものが本物かどうかも含めて、自分で判断することになります。

現在、線虫ゲノムのエキソン・イントロン構造は、GeneFinder 等のプログラムを用いて、かなりの精度で予測可能で、GenBank などの DNA データの FEATURES 欄に記載されているペプチド配列(開始・終止コドン、エキソン・イントロン構造)は、現段階ではほとんどのものは cDNA と比較したものではなく、ソフトウェアによって予測されたものです(putative と表記されている)。いくつかのものは cDNA 配列と比較されて、正しいことが確認されていますが、100% 正しいわけではありませんから、特に、検索に用いた遺伝子とまったく異なる遺伝子がある領域に記載されていたような場合は、どちらが本当か検討してみる必要があります。また、ACeDB で、ゲノムと EST のデータの対応関係は、解析に時間的な遅れがあるので完全には記載されていません。その部分についての両方のデータを入手することをおすすめします。先に EST をサーチした場合、得られた cDNA 配列は全長ではなく、部分的であることが多いので、この配列をプローブに、さらに上記の方法で検索して対応する部分の *C. elegans* ゲノム配列を入手します。先にゲノム配列を得た場合には、その部分の塩基配列をプローブに、*C. elegans* EST をサーチします。上の解析例では、ACeDB によってかなり解析が進んでいる場合を採り上げましたが、必ずしもこういうクローンにヒットするとは限りません。ペプチドの予測はまったく記載されていない、DNA 配列のみのクローンも多数あります(図3、図4 のうちのいくつかもこれに該当します)。EST も発現量の少ないものは見いだされてないことがあります。これらの場合にも、自分である程度解析を進めることができます。まず、例えば、BCM GeneFinder (<http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html>)などのプログラムを用いて、可能なポリペプチド配列(ORF)を予測し、それと自分の配列を比較して、遺伝子ホモログの可能性を判断します。しかし、相同性の弱いものは、これらのコンピューターによる解析からだけでは、目的の遺伝子の線虫ホモログかどうかを判断することは困難であることは言うまでもありません。

Profile of the Author

Kazunori Kondo Graduated from Kyoto University (in 1980), and obtained Ph.D. in Science at the same university (in 1985). He is a lecturer of Soka University(Department of Bioengineering, 1991-).

(Main themes)

He is studying fucosyltransferases in *C. elegans*: cDNA cloning, expression in culture cells and measuring their enzyme activities; Gene knockout is planned, too.

(Hobbies)

Listening to music with CD: World music including Asia, Europe and Africa; Also classic music.