Introduction to Bioinformatics

2. DNA Sequence Retrieval and comparison

Benjamin F. Matthews
United States Department of Agriculture
Soybean Genomics and Improvement Laboratory
Beltsville, MD 20708
matthewb@ba.ars.usda.gov

What we will cover today

- Retrieving a known DNA sequence
- Similarity searching with a DNA sequence
- BLAST
Retrieving a DNA sequence

You read a paper and..

★ Is full-length clone of gene available?
★ Is at least some of the DNA sequence available? (EST sequence)?
Finding Sequences in Databases

✿ The public DNA and protein sequence databases are huge.

✿ In order for these databases to be useful, the data must be readily accessible to researchers.

What Are You Looking For?

✿ A gene?
  ✿ DNA or protein sequence?
  ✿ DNA sequences are essentially all in GenBank
    ✿ Genomic, mRNA, cDNA, EST?

✿ Proteins are harder to pin down
  ✿ GenPept (GenBank Peptides) is huge and poorly annotated - lots of junk
  ✿ SwissProt is carefully annotated, but not fully comprehensive
  ✿ PIR is somewhere in between
Large Databases

- Once upon a time, GenBank sent out sequence updates on CD-ROM disks a few times per year.
- Now GenBank is over 95 Gigabytes (28 billion bases)
- Most biocomputing sites update their copy of GenBank every day over the internet.
- Scientists access GenBank directly over the Web

You can search DNA sequence database

- Retrieve known sequences by
  - Keyword search
  - Accession numbers
- If you know some DNA sequence
  - Compare your DNA sequence with those in database
  - Basic Local Alignment Search Tool (BLAST) searches
Retrieve a DNA sequence

- **ENTREZ**
- Click – Nucleotide
  - GenBank
  - OR
- Accession number
- Keyword search

**Entrez** is a Tool for Finding Sequences

- **GenBank** is managed by the **NCBI** (National Center for Biotechnology Information) which is a part of the US National Library of Medicine.

- NCBI has created a Web-based tool called **Entrez** for finding sequences in **GenBank**.
- Each sequence in **GenBank** has a unique “accession number”.
- **Entrez** can also search for keywords such as gene names, protein names, and the names of organisms or biological functions
Entrez is a Database

- The Entrez database contains all of the nucleotide and protein sequences in GenBank (updated daily) along with all of the literature in MEDLINE and the 3-D protein structures in PDB (Protein Data Base).

- Entrez is much more than a database, it is a both a powerful search engine and a pre-computed list of relationships among all of its data elements.

Entrez is Internally Cross-linked

- DNA and protein sequences are linked to other similar sequences.

- Medline citations are linked to other citations that contain similar keywords.

- 3-D structures are linked to similar structures.
Databases contain more than just DNA & protein sequences

MEDLINE

Nucleotide Sequences

Protein Sequences

Genomes

Structures
GenBank

- National Institute of Health, National Library of Medicine, National Center for Biotechnology Information

- Retrieve a sequence from GenBank
- Analyze raw sequence data
  - Base calling
  - Editing
  - Obtaining a consensus sequence
  - Translating
  - Restriction mapping
  - Similarity comparisons
  - Motif searches
Accession Numbers!!

- Databases are designed to be searched by accession numbers (and locus IDs)
- These are guaranteed to be non-redundant, accurate, and not to change.
- Searching by gene names and keywords is inexact and retrieves more than one record usually
Type in a Query term

Enter your search words in the query box and hit the “Go” button
GenBank Records

- Databases are composed of records
- Flat File Format
- Provides information
- Standard, consistent organization of data
Flat file format

- Organized in a structured manner
- One big file
- Large body of information assembled and distributed in consistent format
- Lack support for procession transactions (inserts and updates)
Some Fields of GenBank Record

- Locus Name
- Sequence length
- Molecule type
- Definition
- GenBank accession number
- Version
- Keywords
- Source
- Organism
- Reference

- Reference
- Authors
- Title
- Journal
- Medline
- Other references
- features
- Amino acid translation
- Nucleotide sequence
Locus Name

Unique

- Up to 10 characters
- 6 characters
  - Genus species
- 8 characters
  - Just accession number

Better to search for accession number than Locus Name
GenBank Accession Number

- Unique identifier for sequence record
- Usually a combination of letter(s) and numbers
- Do not change even if information changes
- Newer accession numbers to new submission using some of this data
Refine the Query

- Often a search finds too many (or too few) sequences, so you can go back and try again with more (or fewer) keywords in your query.
- The “History” feature allows you to combine any of your past queries.
- The “Limits” feature allows you to limit a query to specific organisms, sequences submitted during a specific period of time, etc.
- [Many other features are designed to search for literature in MEDLINE]
DNA similarity search

Find related sequences

* Find ESTs
  - If not full-length, may allow assembly from ESTs
* Find other family members
  - Organization and function
* Find similar genes from other cultivars
  - SNP discovery
* Find similar genes from other organisms
  - Phylogenetic relationships
ESTs (Expressed Sequence Tags)

- partial cDNA sequences
- dbEST at NCBI
  - a comprehensive set of all public EST data
- UniGene at NCBI
  - clusters of ESTs and known genes from key species
    - does NOT have consensus sequences
    - has far too many clusters to be representative of real genes (129 K human clusters)

Find related DNA sequences

- Similarity Search (BLAST)
- NCBI GenBank database
BLAST Searches

- Compare your sequence with database
- Nucleotide
- Protein
- Targeted to a genome

**BLAST**

- Basic Local Alignment Search Tool
- Local alignment
- Tutorial at:
BLAST

::* Discontiguous Mega BLAST
  * Comparison of diverged sequences especially from different organisms
  * Alignments with low degree of identity
  * Looks for hits in “non-consecutive positions”

::* Mega BLAST
  * Slight differences in similarity
  * Not effective at low degree of identity
  * Faster; handles longer sequences

::* BLAST
  * Local alignment tool
You have a sequence.
Does it have similarity to other known genes???
Copy DNA sequence from file

GGTTGTGCAAGTCCATTTTCTATTGACTTCTCCTCATTGATCCAAGATGAACAGCAC
ACATGCACTTGAATGGTACCATACCTGCTAAGCAGTGTCTAGTAGCTTCCTCATTTTTC
ATGCTTCAATCTCCTATTTAGCCTCAACTCTGGTTCAAAAATTGCTGGTCCACCAGGTGGTGC
AATCCAAAGGAAGGTAGGTGTAATCATCAATTTCTAAGGGGTGAAAGCTTCAACCAAGGC
CTTTCGATGTTTCGATGGATGAGTCCAGGAACATCTTTAGGAGATGAAAGTTGAGTCC
TACCTGGCAGATGTTGCCTCAGCAATAGCAGCTTCGCTGCGTAGCTGTGAGAGATGGTTTTCGATGCT
CCACACAAGCTCAAAAACAGGCTTCTCTGCTGACAGGGATCTGATCCAACAATGACAGC
GCTGACCATTCTCAAAAACAGGCTTCTCGCCACAAAGGCTGCTAAGGTGATGATGCAACTCGG
GACCAGAATGTGACTGCTGTGATCTCCGA
Your request has been successfully submitted and put into the Blast Queue.

**Query** = (1509 letters)

Your search was limited by an Entrez query: Glycine max

The request ID is 1098221208-15060-31262165038.BLASTQ4

The results are estimated to be ready in 37 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.
200401] was used to synthesize the cDNA. First-strand synthesis was performed with 5-mercaptopyrime, since the ligated cDNA was hemi-ethylated. A modification of Stratagene’s first-strand synthesis primer was used. An ‘anchor’ nucleotide (V4A, C, or G) was added in the 3’ end of the primer [5’CCACGAGACAGACAGTCTCTCTCT] to anchor the primer at the 5’ end of the polyA tract. After second-strand synthesis, the cDNA ends were filled in with clever Fru DNA, ligated to EcoRI adapters, and subsequently phosphorylated. The cDNA was then precipitated and redissolved in water. Hybridization was performed in a GeneQuest hybridization and washing system. The pGEM- vector was digested with XhoI and EcoRI, resulting in a 7-kb fragment that was directly subcloned into the vector. This library was constructed in the laboratory of Dr. Paul Rea and Dr. Virginia M. Coryell at Northern Arizona University.
FASTA/BLAST Statistics

- $E()$ value is equivalent to standard P value
- Significant if $E() < 0.05$ (smaller numbers are more significant)
  - The $E$-value represents the likelihood that the observed alignment is due to chance alone. A value of 1 indicates that an alignment this good would happen by chance with any random sequence searched against this database.
- The histogram should follow expectations (asterisks) except for hits

Interpretation of output

- very low $E()$ values ($e^{-100}$) are homologs or identical genes
- moderate $E()$ values are related genes
- long list of gradually declining of $E()$ values indicates a large gene family
- long regions of moderate similarity are more significant than short regions of high identity
What this does for you

- You identified what gene is encoded by your clone’s sequence
- Perhaps you may have found the function of your gene
- You have more cDNA sequences to add together to build a consensus and perhaps a full-length cDNA

Biological Relevance

- It is up to you, the biologist to scrutinize these alignments and determine if they are significant.
- Were you looking for a short region of nearly identical sequence or a larger region of general similarity?
- Are the mismatches conservative ones?
- Are the matching regions important structural components of the genes or just introns and flanking regions?
Borderline similarity

What to do with matches with E() values in the 0.5 -1.0 range?

this is the “Twilight Zone”

retest these sequences and look for related hits (not just your original query sequence)

similarity is transitive:

if A~B and B~C, then A~C

Advanced Similarity Techniques

Automated ways of using the results of one search to initiate multiple searches

**INCA (Iterative Neighborhood Cluster Analysis)**

[http://itsa.ucsf.edu/~gram/home/inca/](http://itsa.ucsf.edu/~gram/home/inca/)

- Takes results of one BLAST search, does new searches with each one, then combines all results into a single list
- JAVA applet, compatibility problems on some computers

**PSI BLAST**


- Creates a “position specific scoring matrix” from the results of one BLAST search
- Uses this matrix to do another search
- builds a family of related sequences
- can’t trust the resulting e-values
FASTA format

- One of three formats used for sequences
- Begins with single-line description followed by sequence data
- Description line starts with “>”
- Example:

```plaintext
>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFAALLKCNADAYQGKTNCSNVSVHCTNLMNNTTVTGGILLSNYSENRT
QIQQKHTSNDSAILNKHNLNYLTVTCKRPGNKTALPVTIMAGLTVFHSQKYNLRRQAWC
HPFNSWKGAWKVEECIVNLPRKERYRGTDPKRIFFQRQWQGQPETANLWFNCHEFFYCYK
MDWFNLNYNLTVDADHYNECKNNTSGTKSGNKRAGPGPCVQRTYVACHRSVIMLETISKK
TYAPPREGHLECTSTVTGMTVLNYPNKRTNTVLSGOESIAAEDRYKLVETIPGF
APTEVRRYTGGHERQKRPVFVXXXXXXXXXXXXXXXXXXXXXXXXVQSLLAGILQQKNL
LAAVEAQQMLKTLIWGVK
```

The nucleic acid codes supported are:

- A -> adenosine
- C -> cytidine
- G -> guanine
- T -> thymidine
- U -> uridine
- A -> G A (purine)
- C -> G T (pyrimidine)
- G -> G T (keto)
- Y -> T C (pyrimidine)
- K -> G T (keto)
- M -> A C (amino)
- S -> G C (strong)
- W -> A T (weak)
- R -> G A (purine)
- H -> A C T
- V -> G C A
- N -> A G C T (any)

- gap of indeterminate length
accepted amino acid codes are:

A alanine           P proline
B aspartate or asparagine Q glutamine
C cystine           R arginine
D aspartate         S serine
E glutamate         T threonine
F phenylalanine     U selenocysteine
G glycine           V valine
H histidine         W tryptophan
I isoleucine        Y tyrosine
K lysine            Z glutamate or glutamine
L leucine           X any
M methionine        * translation stop
N asparagine

- gap of indeterminate length

Types of data integrated in genome browsers

- Genomic sequence
- RefSeq mRNAs (non-redundant)
- GenBank mRNAs (redundant)
- ESTs
- Gene predictions
- SNPs
- Homologous sequences from other organisms
- STSs
Other Sequence Search Tools

SRS (Sequence Retrieval Service) was created by Dr. Thure Etzold: [CABIOS 9(1); 49-57 (1993)]

It is a meta search engine for all types of biological data in hundreds of databases as well as about 20 sequence analysis programs

SRS can be accessed over the WWW from many servers (mostly in Europe):

http://srs.ebi.ac.uk/
http://www.infobiogen.fr/srs6bin/cgi-bin/wgetz?-page+top
http://www.sanger.ac.uk/srs6bin/cgi-bin/wgetz?-page+top
http://iubio.bio.indiana.edu/srs6bin/cgi-bin/wgetz?-page+top
Why So Many Databases?

If GenBank has all sequence data and Entrez is such a good query tool, then why are there so many other sequence databases?

- Specialized data (single species, immunoglobulins, etc.)
- Better annotation (i.e. SwissProt)
- Sequences linked to other data (ACEDB)
- Subbornness and local pride - EMBL, DDBJ

Well designed databases are interlinked with others for supplemental data

It is very hard to get all relevant information across all databases for any gene
Other Genetic Databases

- Genome Sequence - where does a gene fall on the genome
  - integrate multiple layers of information
    - Sequence contigs, mRNAs, predicted exons, etc.
  - Single species?
- ESTs: dbEST @ NCBI
- SNPs: dbSNP @ NCBI, http://snp.cshl.org (SNP Consortium)
- Metabolism/Pathways
- Gene Function (Genome Ontology)
- Protein motifs/domains and protein families
Genome Databases

- New area - in desperate need of development
  - Chromosomes::Sequence::Contigs::Clones::
  - STS Markers::Genetic Markers::Genes::
  - Features::Expression data::Phenotype
- No single database can hold it all
- UCSC is probably the best right now
  - genome.ucsc.edu
- Need a data exchange and linkage infrastructure

European Bioinformatics Institute

- Products and services
- Databases
  - Literature
  - Microarray
  - nucleotide
- Toolbox with software
  - Similarity searches
  - Protein function
  - Sequence analysis
  - Structure
- http://www.ebi.ac.uk/
IDENTIFIERS

dbEST Id: 101883
GenBank Acc: T48601
GDB Id: 490761

CLONE INFO

Clone Id: IMAGE:69864 (3')
DNA type: cDNA

SEQUENCE
GGCGGCTCAGTAGCAGGTGCCGTCCACCTCCGCCATGACAACAGACACATTGACATGGGT
GGGTTTACCACCAAGCGTCCGATGGTCTTCTGTGTGAAGGCCAGCCAGGCGCCTCCATGG
CACCATGCAGGAGAAGGNCTCCCCCTTCTTCCAGTCCTCGGCTGCCACGCGCAGTATGCT
GGTCACACGAAGGTCGTGGTGCCCTGGCTGGNTCCTNCANGGATGCCCAAGTCAGGTACT
TNTCGCGGGCAGCTCCTGTGACCCCTGCAGCCAGCGAACCAGCACGTCCTTGGGGCTTN
AAGCNGCGCTACCAGGCACTTCAACCGTTCNCCAGCTTCGTTCAGGGCCANCTTC

Quality: High quality sequence stops at base: 277

LIBRARY

Lib Name: Stratagene placenta (#937225)
Organism: Homo sapiens
Sex: male
Organ: placenta
Lab host: DOGA cells (kanamycin resistant)
Vector: pBluescript SK-

COMMENTS

This clone is available royalty-free through LLNL; contact the IMAGE Consortium (info@image.llnl.gov) for further information.

PUTATIVE ID

Assign by submitter
similar to gb:S71043_rna1 IG ALPHA-2 CHAIN C REGION (HUMAN)

ENTRY CREATED

Feb 6 1995
Last Updated: Feb 6 1995

SOURCE

IMAGE Consortium, LLNL

This clone is available royalty-free through LLNL; contact the IMAGE Consortium (info@image.llnl.gov) for further information.

High quality sequence stops: 277
Database Search Strategies

- General search principles - not limited to sequence (or to biology)
- Use accession numbers whenever possible
- Start with broad keywords and narrow the search using more specific terms
- Try variants of spelling, numbers, etc.
- Search all relevant databases

- Be persistent!!

What we covered today

- Retrieving a known DNA sequence
- Similarity searching with a DNA sequence
- BLAST