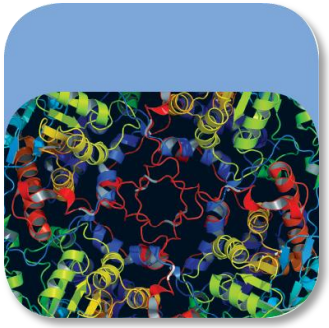


BECKER
Medical Library

NUCLEIC ACID SEQUENCE ANALYSIS

Kristi Holmes, PhD
holmeskr@wustl.edu
February 14, 2010

Washington University in St. Louis
School of Medicine

Information directories

***Nucleic Acids Research* Database Issue**

- The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. Cochrane GR, Galperin MY. Nucleic Acids Res. 2010 Jan;38(Database issue):D1-4. Epub 2009 Dec 3. PMID: 19965766 [PubMed - in process] Related articles Free article
- Complete table of contents for the NAR database issue *(Tip: to see the table of contents from the database issue for a previous year, just reduce the volume number in the URL (to the complete table of contents) by one.)*
- Searchable database of summary papers

***Nucleic Acids Research* Web Server Issue**

- 2009 Web Server complete table of contents
- Searchable database of web server summaries

Nucleic Acids Research* Methods *index

Bioinformatics Links Directory (described in an NAR article, July 2007 web server issue)

ExPASY Life Science Directory

- >1000 links on a single page, organized by category

BioMed Central Databases collection

Biocatalog by EBI

- database providing summary and access information for a wide range of molecular biology databases and software; browse category of interest or search complete db with EMBL SRS server

Online Bioinformatics Resources Collection (OBRC) from the Health Sciences Library System, University of Pittsburgh.

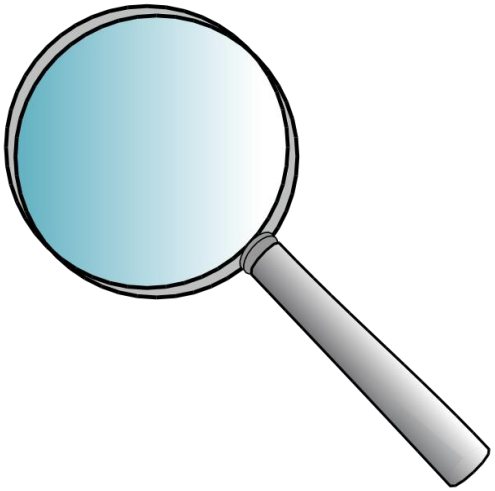
-- From NAWBIS Information Hubs for Molecular Biology Databases and Software, Renata Geer

What's next?

- Finding a sequence
- Sequence manipulation
- Restriction mapping
- Primer design
- Sequence alignments
- Vector screening

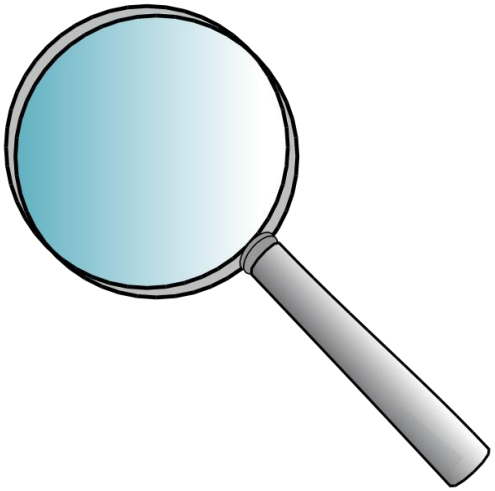
Finding a sequence

- Nucleotide database at NCBI
- Looking for a given gene? Go to Entrez Gene at NCBI
- NCBI Handbook:
  - Entrez Gene: A Directory of Genes
  - Entrez Gene Help
- Looking for a genomic region or for a specific gene plus upstream and downstream sequence? Try Map Viewer at NCBI
- NCBI Handbook:
  - Using Map Viewer to Explore Genomes
  - Exercises: Using Map Viewer



Finding a sequence

- EMBL Nucleotide Sequence Database (also known as EMBL-Bank)
- The EMBL Nucleotide Sequence Database is the European member of the tripartite International Nucleotide Sequence Database Collaboration DDBJ/EMBL/GenBank. Direct access to hundreds of completed genome sequences plus according protein translations is available via EBI's Genome Server. Automatic genome annotation, graphical views and web-searchable datasets are available from the Ensembl project.
- EBI Nucleotide databases
- Mine Ensembl with BioMart and export sequences or tables in text, html, or Excel format



Sequence manipulation

- Sequence Manipulation Suite
  - The Sequence Manipulation Suite is a collection of JavaScript programs for generating, formatting, and analyzing short DNA and protein sequences. It is commonly used by molecular biologists, for teaching, and for program and algorithm testing.
  - See the about the Sequence Manipulation Suite page for more information about individual Sequence Manipulation Suite programs.
  - You can easily mirror the Sequence Manipulation Suite on your own web site, or you can use it off-line.
- ReadSeq – biosequence conversion tool
  - Converts input DNA/AA sequence to specified format (Input format is determined automatically).
  - Information on READSEQ is maintained at the IUbio Archive site at University of Indiana.

- BioMart is a query-oriented data management system developed jointly by the Ontario Institute for Cancer Research (OICR) and the European Bioinformatics Institute (EBI).

- GMOD wiki entry for BioMart
  - Documentation
  - BioMart Tutorial
  - Mailing Lists
  - Download & Install
  - BioMart @ Ensembl

http://www.biomart.org/

Restriction Mapping

What does this mean?

- Determine the number of restriction sites for each enzyme in the database for your sequence.
- Determine the nucleotide position of the cut for each restriction enzyme in your sequence.
- List the enzymes that do not cut your sequence.
- List separately the enzymes that cut only once in your sequence.
- Show a graphical representation of the restriction sites in your sequence.
- Show a textual representation of the restriction sites aligned to your sequence.

Where can I look for help? The web!

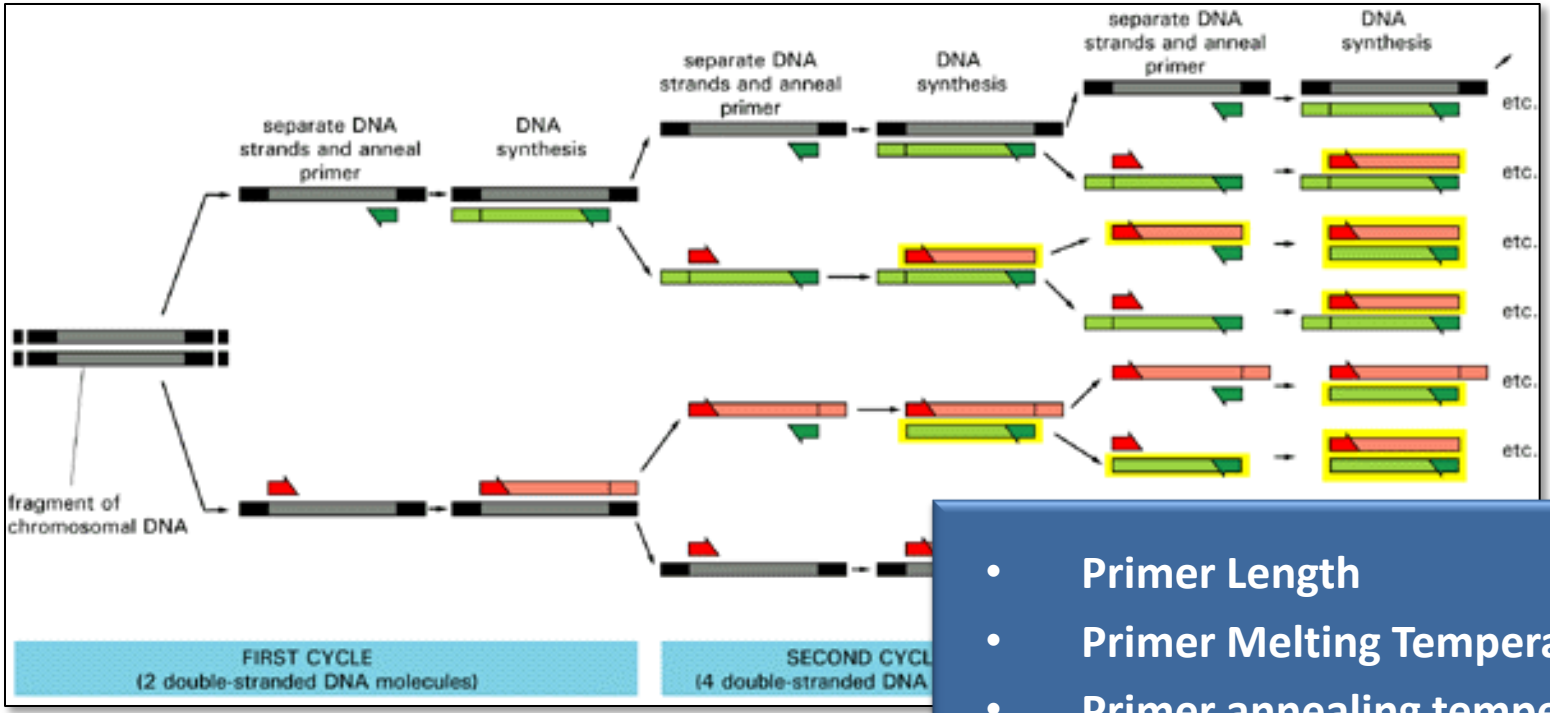
- There are a number of online restriction mapping tools…

Restriction Mapping

- WebCutter 2.0
- NEBcutter
- WatCut

Try one of these tools with this sequence or with one of your own.

GAPDH[gene] AND homo sapiens[organism]



Primer Design Guidelines from Premier Biosoft

PCR amplification - *Molecular Biology of the Cell, 3rd ed.*

denaturation, annealing and extension

- **Primer Length**
- **Primer Melting Temperature**
- **Primer annealing temperature**
- **GC Content**
- **GC Clamp**
- **Primer Secondary Structures**
- **Repeats**
- **Runs**
- **3' End Stability**
- **Avoid Template secondary structure**
- **Avoid Cross homology**

Primer Design – tools

Primer3

Highlights:

- Select optimal primer pairs for PCR reactions using user-specifiable parameters such as %GC content, melting temperature (Tm), and many more constraints.
- Determine primer-dimer possibilities.
- Select "internal oligo" intended to be used as hybridization probe to detect PCR product after amplification.
- Uses DNA sequence in FASTA format.

- Primer3 Wiki

Primer Design Assistant (PDA)

Highlights:

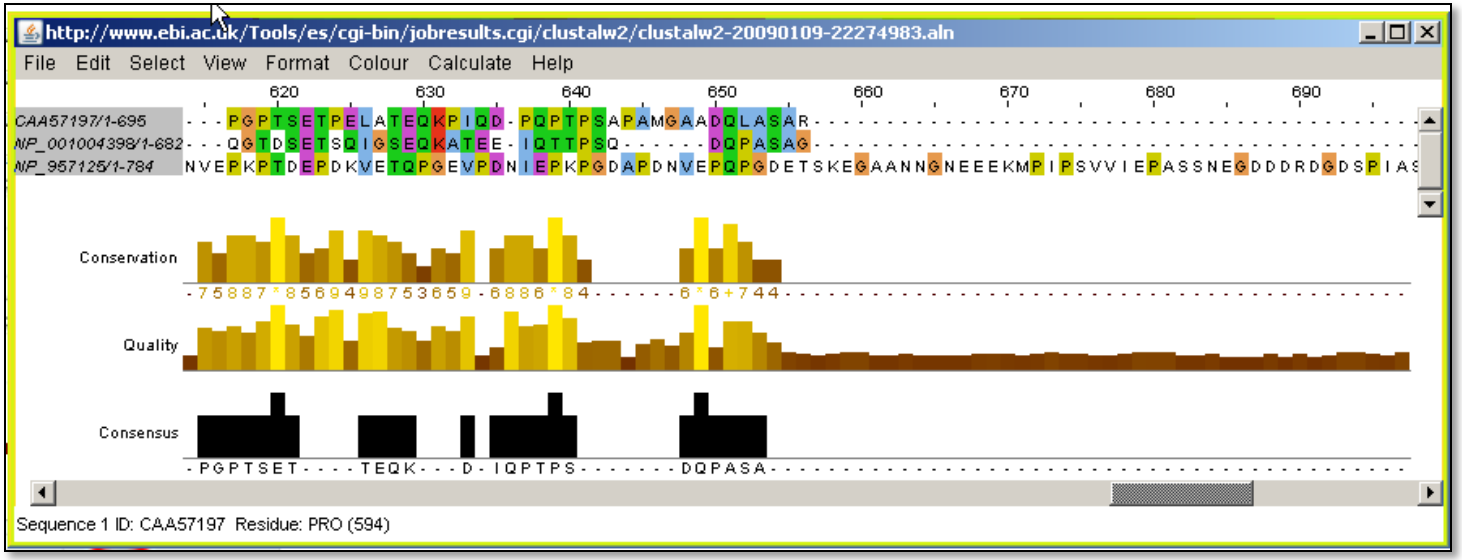
- Primer Design Assistant (PDA) is a web interface primer design service combined with thermodynamic theory to evaluate the fitness of primers.
- Advanced options on 5' GC content, 3' GC content, dimer check and hairpin check are available.
- The option of covered region constrains the PCR product to cover a user-defined segment.
- PDA accepts single sequence query or multiple ones in FASTA format.
- It produces optimal and homogeneous primer pairs that meet the need in experimental design with large-scaled PCR amplifications.
- Considering the system loading, the size of a submitted sequence is limited to 10 kb and the total sequence number in a query is limited to 20.

http://www.hsls.pitt.edu/guides/genetics/obrc/

Sequence alignments

ClustalW2

- a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms.
- ClustalW@ FAQ includes information about supported sequence formats
- Download Clustal to run locally
- Help documentation
- Multiple sequence alignment with the Clustal series of programs. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. Nucleic Acids Res. 2003 Jul 1;31(13):3497-500.

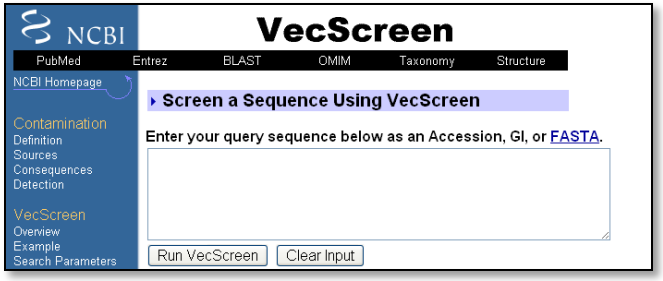


Sequence alignments

Other similar applications for sequence alignments

- Align - This tool is used to compare 2 sequences. When you want an alignment that covers the whole length of both sequences, use needle. When you are trying to find the best region of similarity between two sequences, use water.
- Kalign - A fast and accurate multiple sequence alignment algorithm.
- MAFFT - MAFFT (**M**ultiple **A**lignment using **F**ast **F**ourier **T**ransform) is a high speed multiple sequence alignment program.
- MUSCLE - MUSCLE stands for **MU**ltiple **S**equence **C**omparison by **L**og-**E**xpectation. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options.
- T-Coffee - will allow you to combine results obtained with several alignment methods. For instance if you have an alignment coming fromClustalW2, an other alignment coming from Dialign, and a structural alignment of some of your sequences, T-Coffee will combine all that information and produce a new multiple sequence having the best agreement whith all these methods. By default, T-Coffee will compare all you sequences two by two, producing a global alignment and a series of local alignments (using lalign). The program will then combine all these alignments into a multiple alignment.

Vector Screening - VecScreen



- A **contaminated** sequence is one that does not faithfully represent the genetic information from the biological source organism/organelle because it contains one or more sequence segments of foreign origin.
- The primary consequences of contamination are:
  - Time and effort wasted on meaningless analyses
  - Erroneous conclusions drawn about the biological significance of the sequence
  - Misassembly of sequence contigs and false clustering of Expressed Sequence Tags (ESTs)
  - Delay in the release of the sequence in a public database
  - Pollution of public databases

1. VecScreen is a system for quickly identifying segments of a nucleic acid sequence that may be of vector origin. VecScreen detects contamination by running a BLAST sequence similarity search against the UniVec vector sequence database. VecScreen then categorizes the matches, eliminates redundant hits, and shows the location of contaminating and suspect segments on a simple graphical display.
2. Screens for vector contamination may also be conducted by running a sequence similarity search, such as BLAST, against other sequence databases, for example NCBI's vector database, or the EMVEC vector database from the European Bioinformatics Institute (EBI).
3. Another method used to detect vector contamination is to search the sequence for restriction sites. (Software for restriction site analysis is widely available. Sequences can also be analyzed via the Internet using Webcutter.) Clusters of restriction sites often indicate sequence derived from the multiple cloning site (MCS) of a vector.

http://www.ncbi.nlm.nih.gov/VecScreen/contam.html

TRY IT OUT

A few more things:

Translating DNA into protein

- ExPASy Translation Tool
- EMBOSS Transeq from EBI.
- DNA to Protein Translation

Finding the promoter:

- Promoter Scan from the Bioinformatics and Molecular Analysis section of NIH.
- TFSearch from the Computational Biology Research Center of Japan.

NEW!!

Serial Cloner

- Serial Cloner provides tools with an intuitive interface that assists you in DNA cloning, sequence analysis and visualization
- Macintosh and Windows compatible
- Powerful graphical display tools and simple interfaces help the analysis and construction steps in a very intuitive way.
- Allows local alignment with multi-frame translation in addition to remote NCBI Blast2Seq.
- Finds simultaneously all occurrences of a sub-sequence, a restriction site or any ORF. Homologous recombination and Gateway cloning window.
- Version 1.3 includes virtual cutter Window with simulated gel migration, small Web browser with instant parsing of NCBI/EMBL entries, silent restriction map window, and consensus extraction after local alignment.

Sequences

GeneCards Hot Genes

Escherichia coli UTI89, complete genome.

- Chen, SL, et al. Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: A comparative genomics approach. Proc. Natl. Acad. Sci. U.S.A. 103 (15), 5977-5982 (2006)

TUTORIALS

- 2Can Support Portal The bioinformatics educational resource.
- OpenHelix Tutorials
  - All tutorials
  - Free tutorials
- Bioinformation FAQs on a TON of topics written by Yannick Pouliot at Lane Medical Library (Stanford)
- NCBI Tutorials
  - Genome Workbench Tutorials
  - Cn3D 4.1 Tutorial
  - BLAST information guide
  - Entrez tutorial
  - PubMed Tutorial
  - Entrez GEO Profiles and Entrez GEO DataSets query tutorial
  - PubMed Central
- **Bioinformatics Tutorials – Bioinformatics@Becker Blog (list)**

References

- Baxevanis, A.D. and Ouellette, B.F.F., eds., *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, third edition. Wiley, 2005. ISBN 0-471-47878-4
- Geer, R.C., Messersmith, D.J, Alpi, K., Bhagwat, M., Chattopadhyay, A., Gaedeke, N., Lyon, J., Minie, M.E., Morris, R.C., Ohles, J.A., Osterbur, D.L. & Tennant, M.R. 2002. NCBI Advanced Workshop for Bioinformatics Information Specialists. [Online] *Additional Analytical Tools: What Else Is Out There?* http://www.ncbi.nlm.nih.gov/Class/NAWBIS/. [*date revised July 23, 2006; date cited February 13, 2010*]
- Chen, YB, Chattopadhyay A., Bergen P., Gadd C and Tannery N. 2007. The online Bioinformatics resources collection at the University of Pittsburgh Health Sciences Library System - A one-stop gateway to online Bioinformatics databases and software tools. *Nucleic Acids Research 2007 Database Issue,* 35:D780-D785 http://www.hsls.pitt.edu/guides/genetics/obrc [*date cited February 13, 2010*]
- Primer Design Guidelines from Premier Biosoft. http://www.premierbiosoft.com/tech_notes/PCR_Primer_Design.html [*date cited February 13, 2010*]