

## **Bioinformatics Databases and Tools: The Basics**

In this session, you will see how published researchers find and use data from GenBank and other databases. We will explore the basics of applying those techniques to examine sequence, 3-D protein structures and genomic map data from animal or plant online resources and introduce further training opportunities.

Goals include comfort in documenting bioinformatics searching or analysis strategy as part of your experiment recorded in methods sections in lab notebooks, proposals, articles and your thesis and enhanced communication with advisors about your exploration of bioinformatics data, methods and literature.

Contact: Kristine Alpi, Director, Veterinary Medicine Library 513 – 6219 <u>Kristine\_Alpi@ncsu.edu</u> Mohan Ramaswamy, Assoc. Head Research & Grad Services, 513-3157 <u>mohan\_ramaswamy@ncsu.edu</u>

## **Understanding Bioinformatics in Published Research**

S NCBI Resources 🛛 How To 🖂		Sign in to NCBI					
Publiced.gov PubMed v re-dox active antibiotics divergent bacteria	Search						
US National Library of Medicine National Institutes of Health Save search Advanced		Help					
Display Settings: ♥ Abstract Send to: ♥ See 1 citation found using an alternative search:	Science MAAS FREE WITH REGISTRATION	PMC full-text archive					
Science, 2008 Aug 29;321(5893):1203-6. doi: 10.1126/science.1160619.							
Redox-active antibiotics control gene expression and community behavior in divergent bacteria. Dietrich LE, Teal TK, Price-Whelan A, Newman DK. Denatment of Biology Massachusetts hatting of Technology 77 Massachusetts Avenue, Cambridge MA 01239, USA.	Save items 📼						
Abstract It is thought that bacteria excrete redox-active pigments as antibiotics to inhibit competitors. In Pseudomonas aeruginosa, the endogenous antibiotic pyocyanin activates SoxR, a transcription factor conserved in Proteo- and Actinobacteria. In Escherichia coli, SoxR regulates the superoxide stress response. Bioinformatic analysis coupled with gene expression studies in P, aeruginosa and Streptomyces coelicolor revealed that the majority of	Related citations in PubMed						
SoxR regulons in bacteria lack the genes required for stress responses, despite the fact that many of these organisms still produce redox-active small molecules, which indicates that redox-active pigments play a role independent of oxidative stress. These compounds had profound effects on the structural organization of colony biofilms in both P. aeruginosa and S. coelicolor, which shows that "secondary metabolites" play important	Cited by 41 PubMed Central articles 💿						
conserved roles in gene expression and development. PMID: 18755976 [PubMed - indexed for MEDLINE] PMCID: PMC2745639 Free PMC Article	Related information Related Citations	Weighted = pre- computed similarity searches					
Images from this publication. See all images (4) Free text	Gene						
	Nucleotide (Weighted)						
	Protein (RefSeq)						
	Protein (Weighted)						
	Protein Clusters						
	References for this PMC Article Substance (MoSH Kowword)	<b>DofSog</b> is					
2 m 1	Tavonomy via GenBank	Reiseq is					
Poli	GEO Profiles	curated by					
🛨 Publication Types, MeSH Terms, Substances, Grant Support	Free in PMC	scientists					
+ LinkOut - more resources	Cited in PMC						

#### http://www.sciencemag.org/cgi/content/full/321/5893/1203

(from the Methods) In this study, we investigated the distribution of the E. coli-type oxidative stress response by performing a BLAST search for SoxR and SoxS in the bacterial domain (15). SoxR was found in sequences from 176 strains in the phyla Proteobacteria and Actinobacteria (Fig. 1A), 123 of which come from completed genomes. The occurrence of SoxS was restricted to the family Enterobacteriaceae. To identify alternative SoxR targets in non-enterics, we **searched all available complete bacterial genomes** (616) for the presence of soxRboxes (i.e., SoxR-binding sites in the promoter regions of target genes) using a position weight matrix (PWM) derived from the soxRbox sequences of 12 diverse SoxRcontaining bacteria (fig. S1B). This PWM permits statistically robust predictions of SoxR binding to a



soxRbox. Of the 123 soxR containing genomes, 121 contain soxRboxes. SoxRboxes were also found in 27 genomes (19 were Firmicutes) that do not contain a soxR homolog. **The results of our analysis (table S1 and http://soxRbox.mit.edu**) were consistent with gene expression studies made in the Gram-negative bacteria E. coli, S. enterica (10), P. aeruginosa (8, 13, 14), and Agrobacterium tumefaciens (16), which validates our search algorithm. Supporting Online Material for this article: See Table S1. www.sciencemag.org/cgi/content/full/321/5893/1203/DC1

*Exercise:* Search PubMed for articles by a faculty member in your department that contain data links to NCBI resources or discuss genetic techniques. Record any interesting PMID numbers. Example: yoder ja AND (molecular sequence data OR genetic techniques)

**Core Resources** – *How many of you work with European collaborators?* **European Bioinformatics Institute (EBI)** <u>http://www.ebi.ac.uk/</u>

**National Center for Biotechnology Information (NCBI)** <u>http://www.ncbi.nlm.nih.gov/</u> *Exercise:* Take a look at the NCBI <u>How-To's</u>: Learn how to accomplish specific tasks at NCBI. Identify one that might be relevant to the work done in your program or lab.

## **Topic Searching at NCBI**

The Cross-Database Search is best for a quick view of the content of most NCBI databases by subject. To get all the search options and limits, it is more effective to use individual database interfaces.

N	NCBI	Entrez, The Life Sciences Search Engine						
HOME S	EARCH SITE MAP	PubMed All Databases	Hu	iman Genome	GenBank Map Viewer	BLAS		
Search across databases purple sweet potato								
	- Result cou	nts displayed in gray indicate one or more terms not found						
	101 🚺	PubMed: biomedical literature citations and abstracts	0	1 📋	Books: online books	0		
	230 🥤	PubMed Central: free, full text journal articles	0	none 렀	OMIM: online Mendelian Inheritance in Man	0		
	none 👿	Site Search: NCBI web and FTP sites	0					
	45 🌏	Nucleotide: Core subset of nucleotide sequence records	Ø	16 鶲 d	IbGaP: genotype and phenotype	0		
	1816 😁	EST: Expressed Sequence Tag records	Ø	none 🤗 u	IniGene: gene-oriented clusters of transcript sequences	0		
	none 関	GSS: Genome Survey Sequence records	0	none 🛃 c	DD: conserved protein domain database	0		
	32 👯	Protein: sequence database	0	none 🔘 c	Clone: integrated data for clone resources	0		
	none 🕕	Genome: whole genome sequences	0	none 🏠 u	JniSTS: markers and mapping data	0		
	1 🤤	Structure: three-dimensional macromolecular structures	0	1 🔂 P	opSet: population study data sets	0		
	1	Taxonomy: organisms in GenBank	0	none 🊺 G	EO Profiles: expression and molecular abundance profiles	0		
	none 🥡	SNP: short genetic variations	Ø	none 🥮 G	EO DataSets: experimental sets of GEO data	0		
	none 📢	dbVar: Genomic structural variation	Ø	none 🎒 E	pigenomics: Epigenetic maps and data sets	0		
	none 【	Gene: gene-centered information	Ø	none 📝 💡	DubChem BioAssay: bioactivity screens of chemical ubstances	0		
	1 🕕	SRA: Sequence Read Archive	0	none 🛞 g	PubChem Compound: unique small molecule chemical tructures	0		
	none 🄇	BioSystems: Pathways and systems of interacting molecule	s 🕜	none 🕕 P	PubChem Substance: deposited chemical substance records	0		
	none 🜐	HomoloGene: eukaryotic homology groups	0	none 🎯 P	Protein Clusters: a collection of related protein sequences	0		
	none 🧰	Probe: sequence-specific reagents	0	none 👩 a	MIA: online Mendelian Inheritance in Animals	0		
	1	BioProject: aggregated biological research project data	0	2 💞 в	tioSample: biological material descriptions	0		

NCSU has a sweet potato breeding program. http://potatoes.ncsu.edu/SPRe leases.html

What can we learn from NCBI about the purple sweet potato?

Are there gene function studies in other types of sweet potato that we might considering running in the purple potato?



## **Entrez Gene**

The best starting point to see what is known about the sequence, structure, function, clinical implications and availability of any gene you are studying. Look at the human coagulation factor XI record in full. Note aliases that could be used in comprehensive literature searching across databases.

*Exercise:* Find sequences from at least two other organisms known to be within the 20 most similar to the sequence of human coagulation factor XI. Give the accession number of the sequences and the name of the organisms (both popular and scientific). NB: There are many ways to do this: HomoloGene, Protein similarity searches (pre-computed or with BLAST, Blink).

1. \_\_\_\_\_ 2.

Sequence Searching and Alignment

Search tools used to find statistically significant matches, based on similarity, to a protein or nucleotide sequence of interest. Matches may provide information on inferred function of a gene or protein or help to determine whether an implied homology between two sequences is justified.

- Find conserved domains common to many sequences in your sequence of interest.
- Compare known sequences for similarity.
- Search for sequence motifs or patterns that are similar to a sequence of interest in a particular region.
- Search for a protein sequence of interest using a nucleotide sequence as the query and vice versa.
- Clean suspected cloning vector sequences from your sequence.
- Create primers using Primer-BLAST

IDENTITY - The extent to which two sequences are invariant.

SIMILARITY - The extent to which sequences are related. Similarity makes no statement about descent from a common ancestor.

#### **Global vs Local Alignments**

Needleman/Wunsch - Finds the best Global alignment between any two sequences.

• CPU and time intensive. Often misses domain and/or motif alignments in sequences.

<u>Smith/Waterman</u> – An extension of Needleman – Wunsch that compares segments of all possible lengths (Local) between two sequences to maximize alignment.

• Very sensitive search. CPU and time intensive.

<u>FASTA</u> – Local alignment. Uses a lookup table to increase speed. Sensitivity and speed are determined by the size of the "word" used for the initial lookup table.

• Sensitive search. Fast.

BLAST (Basic Local Alignment Search Tool) blast.ncbi.nlm.nih.gov/

Fairly sensitive search. Very fast. BLAST is a set of similarity search programs.

Gapped BLAST algorithm allows gaps (deletions and insertions) to be introduced into the alignments that are returned. Allowing gaps means that similar regions are not broken into several segments. Scoring of gapped alignments tends to reflect biological relationships more closely. Ungapped searches are possible. If possible, use the protein sequence for BLAST searches. The BLAST site does allow a global Needleman/Wunsch comparison of two nucleotide sequences. To use other global or local alignments, go to EBI's Sequence Similarity & Analysis page.



Under **Help**, the **Getting Started** category offers the **BLAST Program Selection Guide** with database descriptions. If preparing your own sequence for searching, use FASTA format.

Expect value (E) is a parameter that describes the number of hits one can "expect" to see just by chance when searching a database of a particular size. It decreases exponentially with the Score (S) that is assigned to a match between two sequences. Essentially, the E value describes the random background noise that exists for matches between sequences.

#### **BLAST for Comparing Protein Sequences**

**Position-Specific Iterated BLAST (PSI-BLAST)** provides an automated version of a "profile" search, which is a sensitive way to look for sequence homologues. The program uses the information from any significant alignments returned to construct a position-specific score matrix, which replaces the query sequence for the next round of searching. PSI-BLAST may be iterated until no new significant alignments are found.

**Pattern Hit Initiated BLAST (PHI-BLAST)** combines matching of regular expressions with local alignments surrounding the match. Given a protein sequence S and a regular expression pattern P occurring in S, PHI-BLAST helps answer the question: *What other protein sequences both contain an occurrence of P and are homologous to S in the vicinity of the pattern occurrences?* 

Align two sequences using BLAST (bl2seq) <u>http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi</u> Pairwise BLAST compares two sequences using the BLAST algorithm. "Sequence 1" is the Query sequence and "Sequence 2" is the Subject sequence. You can upload each sequence or reference those already in GenBank.

*Exercise:* How similar are Influenza A virus (A/Hong Kong/1073/99(H9N2)) gi/32140171 and the Influenza A virus (A/Goose/Guangdong/1/96(H5N1)) gi/73852950 looking at segment 3?

#### Multiple Sequence Analysis

BLAST only offers multiple sequence alignment for 1) population data sets in the NCBI PopSet database, and 2) for proteins using the Constraint Based Protein Multiple Assignment Tool.

**CLUSTAL Omega** from the European Bioinformatics Institute can be used for multiple alignments, global or local. <u>http://www.ebi.ac.uk/Tools/msa/clustalo/</u>. The input for Clustal Omega is limited to a maximum of 2000 sequences or to a 2 MB file (whichever is smaller).

### Structures (use Cn3D free viewer to manipulate structure)

http://www.ncbi.nlm.nih.gov/Structure

View 1DMO, a calmodulin protein with 30 NMR models. Review the right sidebar filters. Only one of the models opens on the main screen. Click on the structure title link and for Data Set select "All 3D Structures." Review the display and labeling options in the Structure and Sequence viewing windows.

Why use Structure? Visualize possible points of interactions. Have a sequence but no structure? Find a sequence that is highly similar that has a known structure. You can then align your sequence to that sequence



in the viewing window using the Import function and highlight where things match up. To use a structure file in a different viewing program, change **File Format** from Cn3D to PDB.

## **Genomic Maps**

**Map Viewer** <u>http://www.ncbi.nlm.nih.gov/projects/mapview/</u> lets you browse available organisms and see the locations of known genes on any chromosome, as well as upstream and downstream data. You can also get to maps from Entrez Gene and other NCBI databases. From EBI: Ensembl.

To see the status of approved sequencing targets funded in the US which can help you think about what research opportunities might be coming up, go to <u>http://www.genome.gov/10002154</u>

### **Keeping Track and Keeping Up**

Print web search history and results pages to PDF and store PDFs as documents in your RefWorks account or other storage areas. Date printed on PDF output. Save search strategies (sequences or keywords) in a free MyNCBI account. <u>http://www.ncbi.nlm.nih.gov/sites/myncbi/</u>

### **Further Resources**

#### 2013 Database Summary Paper Category List - Nucleic Acids Research

http://www.oxfordjournals.org/nar/database/c/ The Nucleic Acids Research online Database Collection lists 1300+ selected molecular biology databases.

Exercise: Go to the NAR list and choose a category relevant to your area of study OR use the Search Summary Papers to search by keyword. List a database or two available that might be useful to you:

How recently is one of those databases updated?

### Learning More -

NCBI's web and video tutorials are online at http://www.ncbi.nlm.nih.gov/education/tutorials/

### **Resources at NCSU**

**Bioinformatics Consulting and Service Core** (BCSC) provides software, hardware, and analytical support related to bioinformatics, functional genomics, and life sciences research: <u>http://brc.ncsu.edu/consulting/</u>

#### Bioinformatics Research Center: Bioinformatics and Statistical Genetics <u>http://brc.ncsu.edu/</u>

Example: <u>Alternative Splicing Gallery</u> (ASG) is a web-based splicing graph database that integrates transcript information from Ensembl, RefSeq, STACK, TIGR gene index, and UniGene, in order to explore and visualize gene structure and alternative splicing and to provide an exhaustive transcript catalog.