JOHNS HOPKINS
UNIVERSITY

Zanvyl Krieger School of Arts and Sciences
Advanced Academic Programs

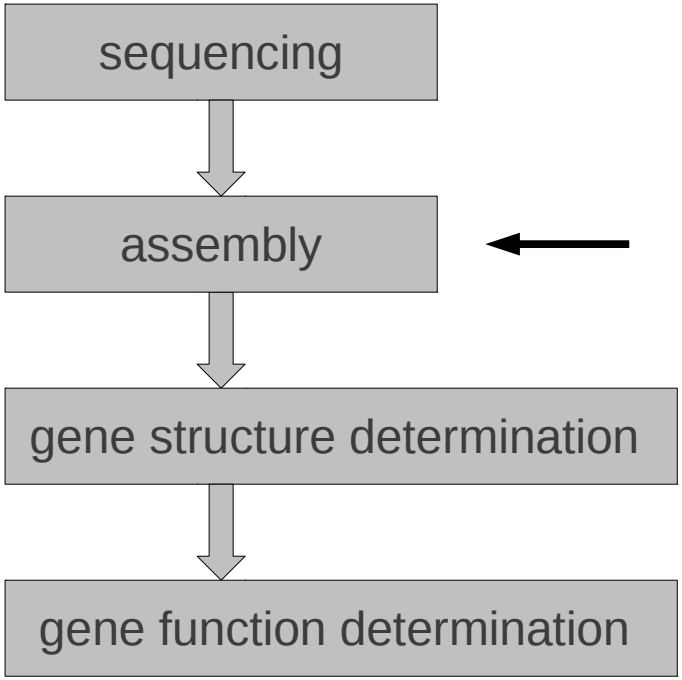
410.734.81 and 410.734.82
Practical Introduction to Metagenomics

Topic: Functional annotation

Instructor: Joshua Orvis

The annotation process

Annotation strategies differ between organism types (such as prokaryotes, eukaryotes, etc.) but the high-level steps of analysis into the annotation process are typically the same:



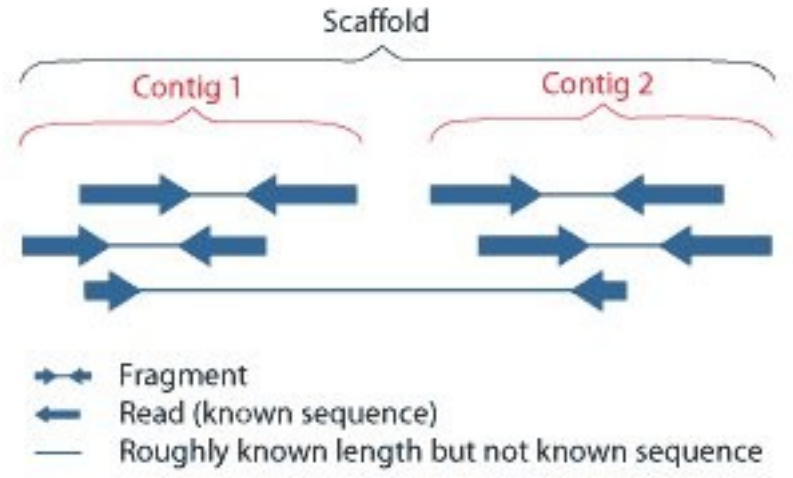
This is actually optional. For many applications, and depending on the sequence type, the raw read data may be sufficient for analysis.

First we define where the genes or fragments of them appear to be.

This week's lesson – what to the predicted genes appear to DO?

Because their genes typically lack introns, archaeal genomes are typically annotated using the same methods as prokaryotes. Viruses exhibit genes with structural properties of both which, along with their typically small genome size, typically require custom consideration.

Which molecules?



Let's go back a bit to the statement that assembly is arguably optional. Why is this so?

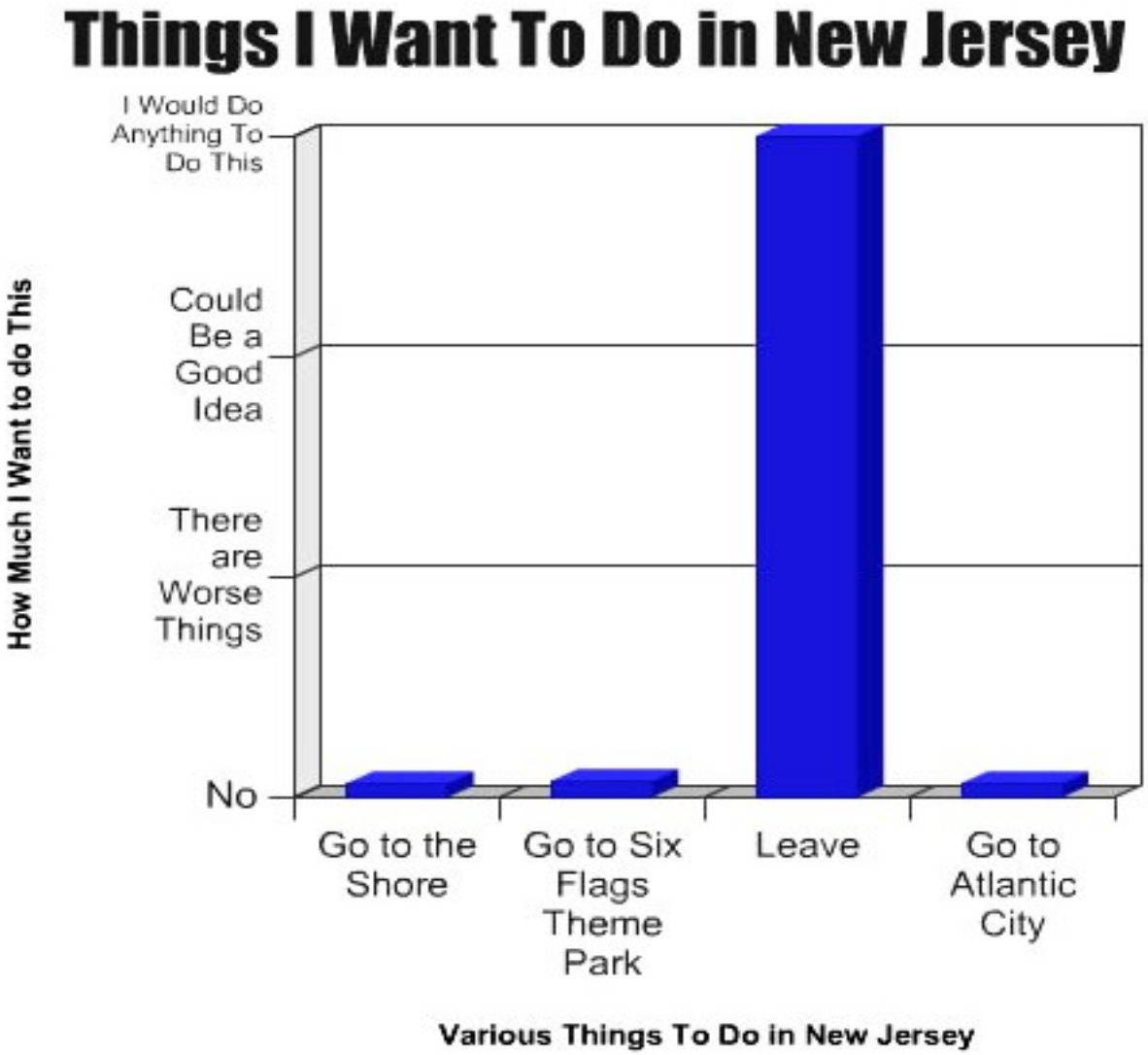
Traditionally, a large collection of sequencing reads undergo assembly into *contigs*. These contigs can even be ordered, oriented and a stitched together into *scaffolds*. (Often the scaffolds have controlled sequences inserted between contigs to delineate contig boundaries.)

But this process is time-consuming, resource intensive and error prone, especially in metagenomic sequences. If your specific analysis type doesn't benefit from long input sequences or more-complete coding genes/coding regions then it's reasonable to skip it.

150bp reads, for example, are sufficient for many taxonomy studies. But for those which need full gene sequences some assembly must be done, since only small protein domains could be recovered from reads of that length.

In this course we've performed analyses at both the read-level (FR plots) and on assemblies (gene prediction.) It's important to keep in mind that assembly is a process you should only do if you can justify a need for it for your analysis.

With apologies to some students



Annotation attributes

Before we get into the algorithms and methods of gathering functional evidence for a gene, we should first consider the end-goal. What does it mean to have a functionally annotated gene? Is it enough to just have a *gene product name*? A *gene symbol*? What else?

I call any of the data you would attach to a gene (actually, to a protein) its **annotation attributes**. What these should be is actually quite debated, so let's look at NCBI, the standard repository in the US for all published annotation, and see what some of these attributes are in a truncated *E. coli* entry:

```
CDS             3734..5020
                /gene="thrC"
                /locus_tag="b0004"
                /EC_number="4.2.3.1"
                /product="threonine synthase"
                /protein_id="NP_414545.1"
                /GO_component="GO:0005737"
                /GO_process="GO:0009088"
```

gene: AKA "gene symbol" or "gene name", these are the common abbreviations molecular biologists have passed around for the last few decades.

locus_tag: Identifier only unique within that entry.

EC_number: Finally, something systematic - an ID within a hierarchical database of enzyme nomenclature.

product: AKA "gene product name", this is a free-text description of the protein's functionality. Most loved by biologists for its utility, the complete lack of standards in naming here makes this less ideal for computational comparisons. Differences in naming conventions, regional spellings, etc. mean that the same gene in even two different strains may not have the same value here.

protein_id: Issued by NCBI, this is a 'globally' unique, versioned identifier for the protein.

GO_component and **GO_process**: Computational gold here. (see next slide)

Gene Ontology

“The Gene Ontology project provides an *ontology* of defined terms representing gene product properties. The ontology covers three domains: **cellular component**, the parts of a cell or its extracellular environment; **molecular function**, the elemental activities of a gene product at the molecular level, such as binding or catalysis; and **biological process**, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

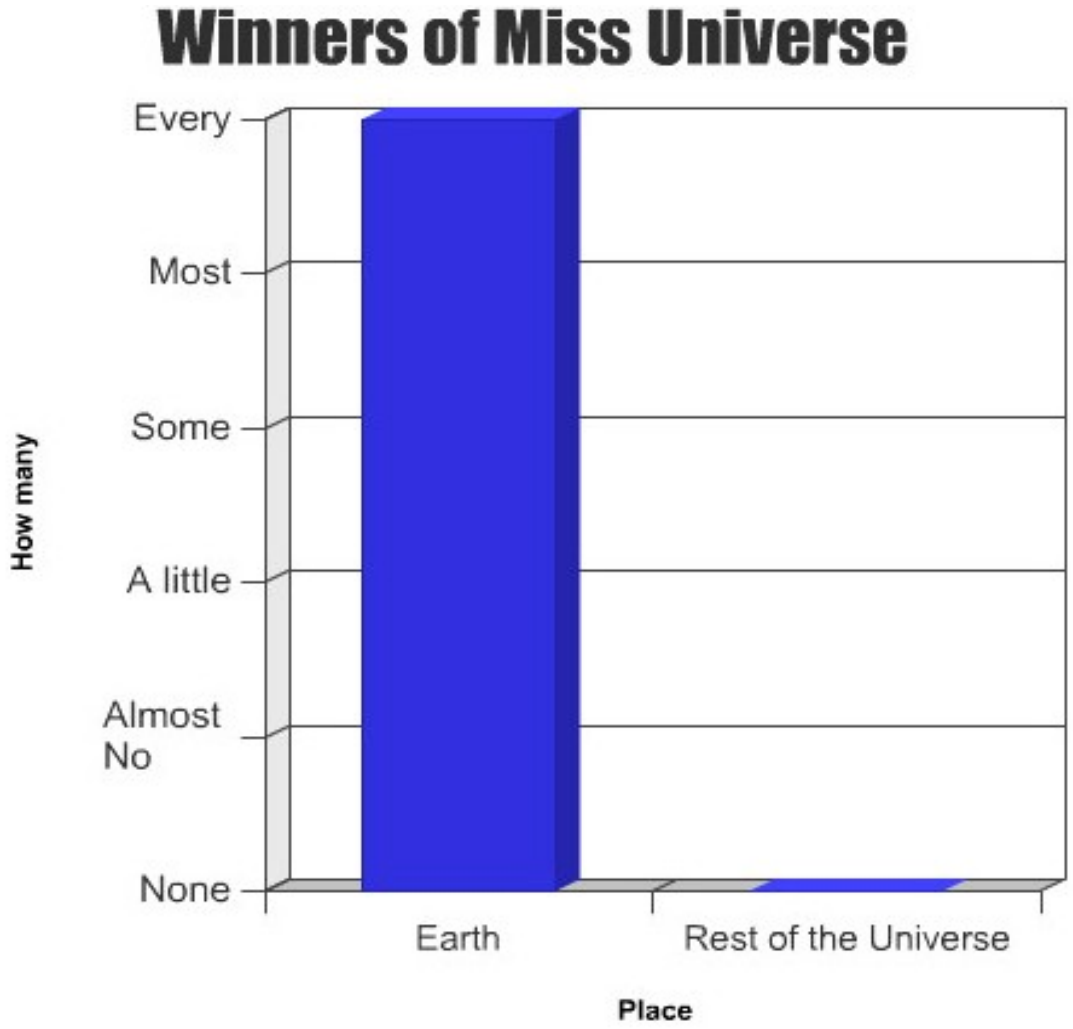
For example, the gene product cytochrome c can be described by the molecular function term oxidoreductase activity, the biological process terms oxidative phosphorylation and induction of cell death, and the cellular component terms mitochondrial matrix and mitochondrial inner membrane.

The GO ontology is structured as a directed acyclic graph, and each term has defined relationships to one or more other terms in the same domain, and sometimes to other domains. The GO vocabulary is designed to be species-neutral, and includes terms applicable to prokaryotes and eukaryotes, single and multicellular organisms.”

By assigning as many GO terms as possible to each protein, you can query or summarize any large collection of gene products at any level of functionality, or compare any collections of products, whether they are individual genomes or metagenomic annotation sets.

For more: http://www.geneontology.org/

Same goes for the “World Series”



Ontologies

An ontology here is any controlled collection of terms (identifiers and descriptions) which usually have relationships defined between them. GO is just one of these.

Ontologies are typically stored in *OBO format*, which is a simple plain-text format. Here's the complete entry for the thrC gene given in the E. coli example a few slides back:

```
[Term]
id: GO:0009088
name: threonine biosynthetic process
namespace: biological_process
def: "The chemical reactions and pathways resulting in the formation
of threonine (2-amino-3-hydroxybutyric acid), a polar, uncharged,
essential amino acid found in peptide linkage in
proteins." [GOC:jl, ISBN:0198506732]
subset: gosubset_prok
synonym: "threonine anabolism" EXACT []
synonym: "threonine biosynthesis" EXACT []
synonym: "threonine formation" EXACT []
synonym: "threonine synthesis" EXACT []
xref: MetaCyc:HOMOSER-THRESYN-PWY
xref: MetaCyc:THRESYN-PWY
is_a: GO:0006566 ! threonine metabolic process
is_a: GO:0009067 ! aspartate family amino acid biosynthetic process
```

The annotation need only contain the GO id (GO:0009088) and by keeping a copy of the entire ontology locally you can access this wealth of functional data.

You can download the entire GO ontology to view in plain text, use a web-based browser such as AmiGO, or download a local viewer/editor.

Gathering evidence - BLAST

Many of the tools used to gather evidence you will have used already in prerequisite courses. Though they differ between different pipelines, some very common searches include:

BLAST: You can't have come this far in the program without having used this ubiquitous tool. When choosing a database to compare to you must consider the scale of your input and computational resources.

NR - NCBI's entire non-redundant database might be fine for even an entire single genome, but the database is full of poorly annotated or unannotated sequence and searches against it won't scale to metagenomic data sets.

RefSeq - A selection of NR but with annotation updated in a relatively consistent manner by NCBI. Aims to be a “stable” and “well-annotated” reference.

UniProtKB/SwissProt - A manually annotated and reviewed non-redundant protein database, “which brings together experimental results, computed features and scientific conclusions.”

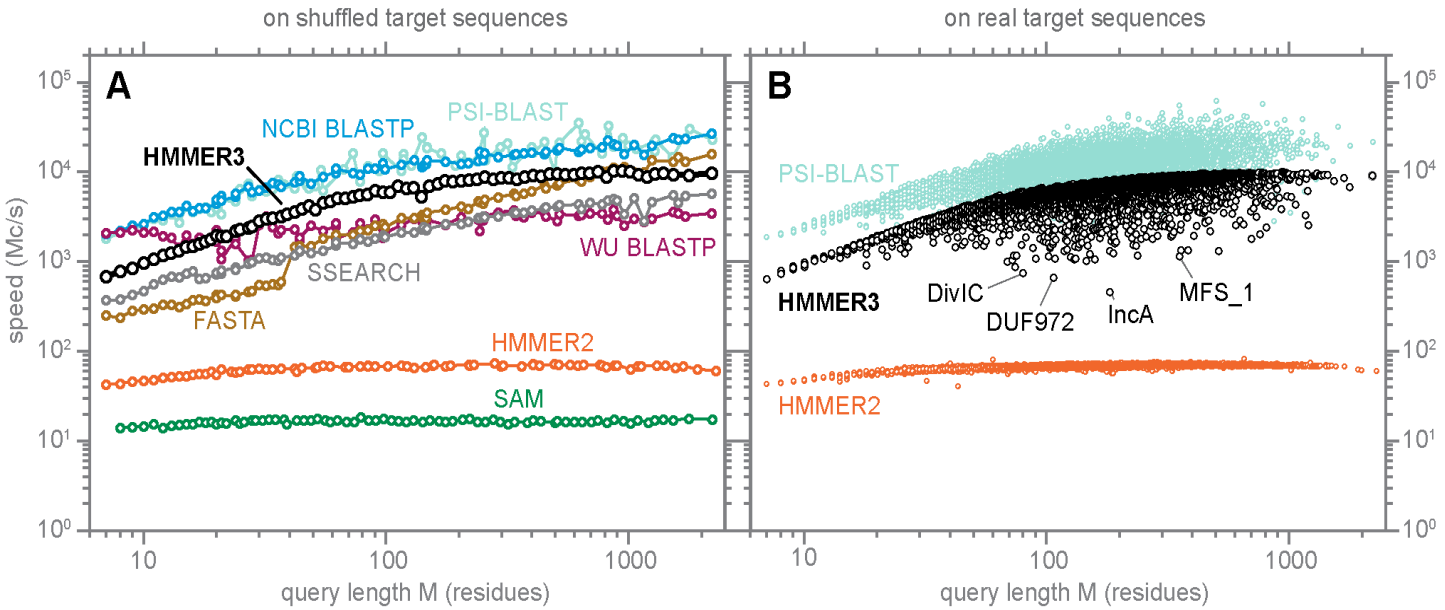
Many of the matches in these more-curated databases contain multiple annotation attributes, including GO terms, which you can choose to transitively annotate your own genes.

Gathering evidence - HMMer

HMMer uses probabilistic hidden markov models to search databases for homologs of protein sequences. In my experience, using this to search the Pfam database of protein families has provided the most accurate and descriptive annotation attributes in my projects, but this was at the cost of significant computational resources.

The release of HMMer3 in 2011, though, saw vast improvements in the speed of the tool. Indeed, the author now advertises that it is essentially as fast as BLAST.

Speed benchmarks of HMMer3 [source]



IGS annotation SOP paper

The IGS Standard Operating Procedure for Automated Prokaryotic Annotation
Galens K, Giglio MG, et al.
Standards in Genomic Sciences, 2011.
PMID: 21677861

This pipeline has evolved over more than 17 years from the developers at TIGR who sequenced and annotated the first free-living organism.

It has been expanded considerably and is now offered as a free service for individual genomes, with each taking 8-12 hours to run.

This paper details each tool/algorithm used in this pipeline, including gathering evidence and applying a hierarchical rule-based system to determine annotated attributes.

[full discloser – I'm an author here]

Standards in Genomic Sciences (2011) 4:244-251 DOI:10.4056/sigs.1223234

The IGS Standard Operating Procedure for Automated Prokaryotic Annotation

Kevin Galens*, Joshua Orvis, Sean Daugherty, Heather H. Creasy, Sam Angiuoli, Owen White, Jennifer Wortman, Anup Mahurkar, Michelle Gwinn Giglio

Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA

* Corresponding Author: **kgalens@som.umaryland.edu**

Keywords: Institute for Genome Sciences, functional annotation, structural annotation, microbial genomics, prokaryotic genomics, annotation pipeline, pFunc, Glimmer, HMM, BER, Ergatis, Manatee, IGS Annotation Engine

The Institute for Genome Sciences (IGS) has developed a prokaryotic annotation pipeline that is used for coding gene/RNA prediction and functional annotation of *Bacteria* and *Archaea*. The fully automated pipeline accepts one or many genomic sequences as input and produces output in a variety of standard formats. Functional annotation is primarily based on similarity searches and motif finding combined with a hierarchical rule based annotation system. The output annotations can also be loaded into a relational database and accessed through visualization tools.

Introduction

The IGS prokaryotic annotation pipeline can be used for the annotation of *Bacteria* and *Archaea*. This pipeline forms the core of the IGS Annotation Engine [1], a free annotation service for prokaryotic sequences. It is also used as the annotation system for prokaryotes sequenced under the IGS Genome Sequencing Center for Infectious Disease [2]. The IGS prokaryotic annotation pipeline can be applied to both draft and finished sequences and has been successfully used in the annotation of hundreds of genomes. The pipeline includes gene finding, protein searches, and the pFunc evidence hierarchy that produces automated functional annotation. The output of this pipeline can be stored in a Chado [3] relational database and can be accessed with Manatee [4] for annotation visualization and curation (Figure 1). Output of the pipeline is also available in a variety of flat file formats. The pipeline is managed using the Ergatis [5] framework and is available on Sourceforge.

Requirements

The IGS prokaryotic annotation pipeline accepts a multi-sequence nucleotide fasta file as input. Annotation can also be performed on an existing set of gene predictions, which simply skips the structural prediction steps of coding sequences. In addition, the name and locus tag prefix (if applicable) of the organism are also required. Structural prediction is performed on the input sequences, followed by similarity searches against public datasets. The final steps of the pipeline include running polypeptide analysis tools as well as automated functional annotation. The output is then converted to various output formats as required. The pipeline uses open source or free software whenever possible. All unique tools written specifically for the pipeline are written in PERL and distributed under the GNU public license on the Ergatis Sourceforge website.

Procedure

Structural Annotation

The pipeline starts by splitting the multi-sequence nucleotide fasta file into individual files. Non-coding RNA and protein coding genes are predicted first, in parallel on each input sequence.

Non-coding RNA Structural Annotation

Non-coding RNA genes are predicted using RNAmmer [6] and tRNA-scanSE [7]. RNAmmer predicts rRNA genes (5s, 16s, and 23s) using the

RAST paper

The RAST Server: Rapid Annotations using Subsystems Technology
Aziz RK, Zagnitko O, et al.
BMC Genomics, Feb. 8 2008
PMID: 18261238

Another freely-available public tool and pipeline, RAST is probably the most different from the IGS SOP.

It bases its attempts to achieve accuracy, consistency, and completeness on the use of a growing library of subsystems that are manually curated, and on protein families largely derived from the subsystems (FIGfams)

Notably, the RAST system automatically includes metabolic reconstruction capability (next week's topic)

BMC Genomics

BioMed Central

Database

Open Access

The RAST Server: Rapid Annotations using Subsystems Technology

Ramy K Aziz[8,9], Daniela Bartels[3], Aaron A Best[7], Matthew DeJongh[7], Terrence Disz[2,3], Robert A Edwards[1,2], Kevin Formsma[7], Svetlana Gerdes[1], Elizabeth M Glass[2], Michael Kubal[3], Folker Meyer[2,3], Gary J Olsen[4,2], Robert Olson[2,3], Andrei L Osterman[1,5], Ross A Overbeek*[1], Leslie K McNeil[6], Daniel Paarmann[3], Tobias Paczian[3], Bruce Parrello[1], Gordon D Pusch[1,3], Claudia Reich[6], Rick Stevens[2,3], Olga Vassieva[1], Veronika Vonstein[1], Andreas Wilke[3] and Olga Zagnitko[1]

Address: [1]Fellowship for Interpretation of Genomes, Burr Ridge, IL 60527, USA, [2]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA, [3]Computation Institute, University of Chicago, Chicago, IL 60637, USA, [4]Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA, [5]The Burnham Institute, San Diego, CA 92037, USA, [6]National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA, [7]Hope College, Holland, MI 49423, USA, [8]University of Tennessee, Health Science Center, Memphis, TN 38136, USA and [9]Department of Microbiology and Immunology, Cairo University, Cairo, Egypt

Email: Ramy K Aziz - ramy.aziz@gmail.com; Daniela Bartels - bartels@mcs.anl.gov; Aaron A Best - Best@hope.edu; Matthew DeJongh - dejongh@hope.edu; Terrence Disz - disz@mcs.anl.gov; Robert A Edwards - RobE@theFIG.info; Kevin Formsma - kevin.formsma@hope.edu; Svetlana Gerdes - Sveta@theFIG.info; Elizabeth M Glass - marland@mcs.anl.gov; Michael Kubal - mkubal@mcs.anl.gov; Folker Meyer - folker@mcs.anl.gov; Gary J Olsen - gary@life.uiuc.edu; Robert Olson - olson@mcs.anl.gov; Andrei L Osterman - osterman@burnham.org; Ross A Overbeek* - Ross@theFIG.info; Leslie K McNeil - lkmcneil@ncsa.uiuc.edu; Daniel Paarmann - paarmann@mcs.anl.gov; Tobias Paczian - paczian@mcs.anl.gov; Bruce Parrello - drake@mkrules.net; Gordon D Pusch - gdpusch@xnet.com; Claudia Reich - creich@ncsa.uiuc.edu; Rick Stevens - stevens@anl.gov; Olga Vassieva - OlgaV@theFIG.info; Veronika Vonstein - Veronika@theFIG.info; Andreas Wilke - wilke@mcs.anl.gov; Olga Zagnitko - OlgaZ@theFIG.info

* Corresponding author

Published: 8 February 2008
BMC Genomics 2008, 9:75 doi:10.1186/1471-2164-9-75

Received: 12 September 2007
Accepted: 8 February 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/75

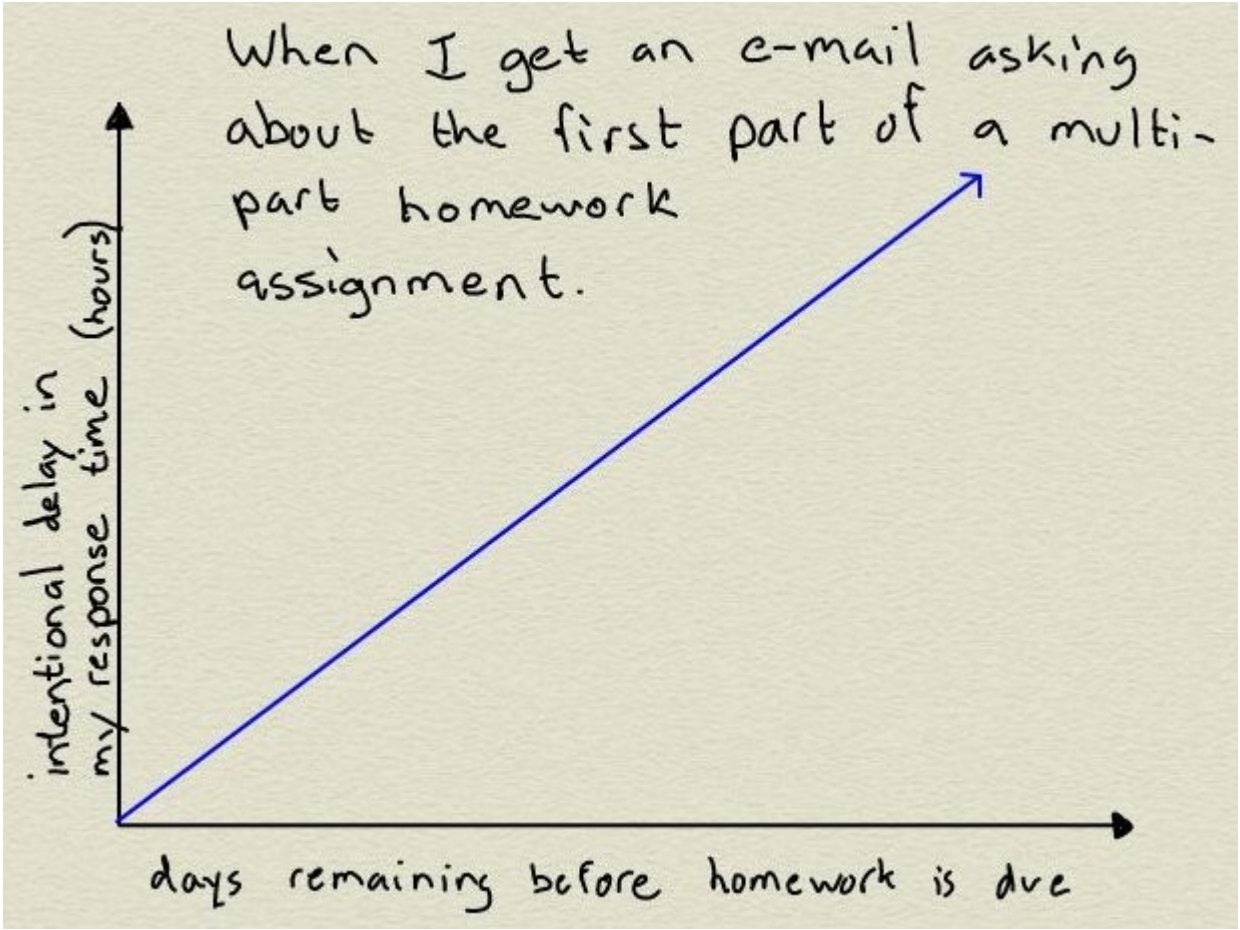
Abstract

Background: The number of prokaryotic genome sequences becoming available is growing steadily and is growing faster than our ability to accurately annotate them.

Description: We describe a fully automated service for annotating bacterial and archaeal genomes. The service identifies protein-encoding, rRNA and tRNA genes, assigns functions to the genes, predicts which subsystems are represented in the genome, uses this information to reconstruct the metabolic network and makes the output easily downloadable for the user. In addition, the annotated genome can be browsed in an environment that supports comparative analysis with the annotated genomes maintained in the SEED environment.

The service normally makes the annotated genome available within 12–24 hours of submission, but ultimately the quality of such a service will be judged in terms of accuracy, consistency, and

On procrastination



Further reading

Gene ontology (GO consortium)

http://www.geneontology.org/

Accelerated Profile HMM Searches

http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002195