

Bioinformatics: Life Sciences on Your Computer

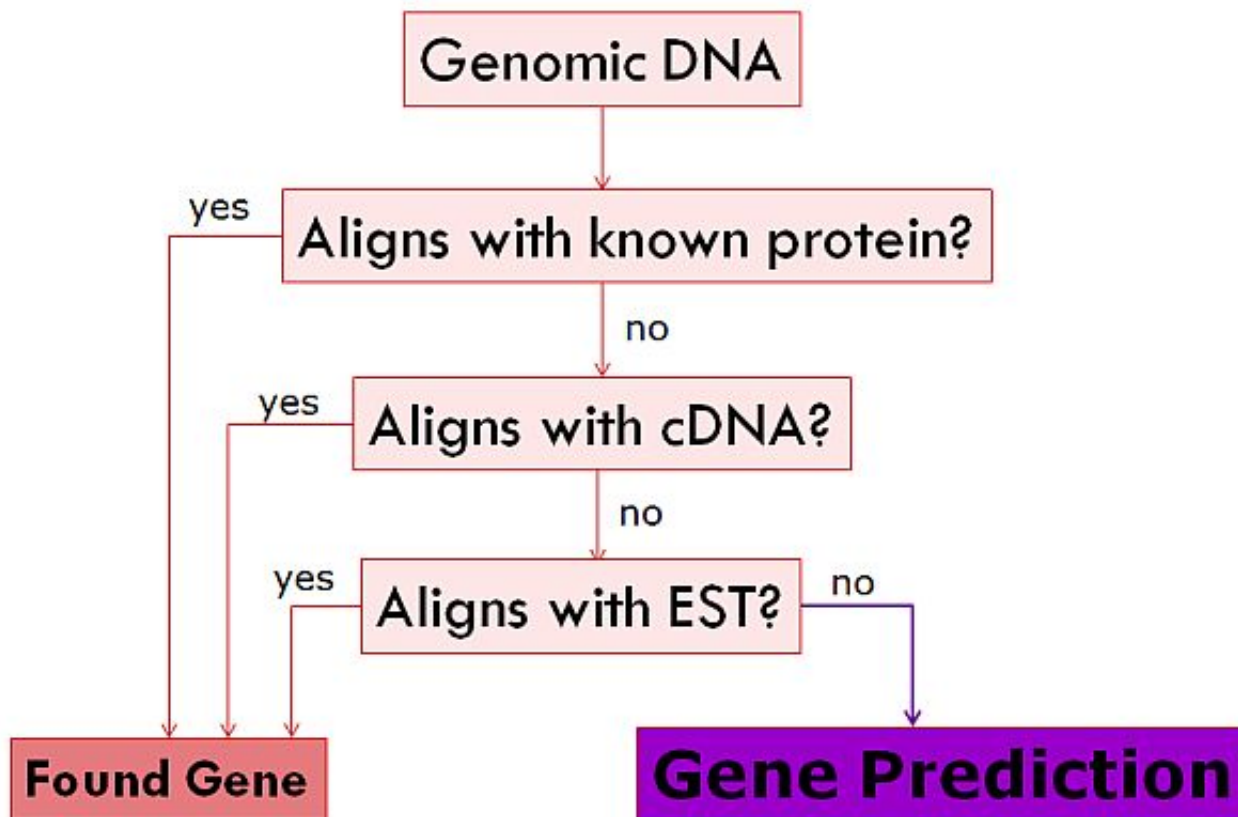
Prokaryotic Gene Prediction

This is supplemental reading for the video lecture on Bacterial Gene Finding.

Ab initio or de novo gene prediction

The focus of this section is *ab initio* or *de novo* gene prediction. The *de novo* gene prediction uses only genomic DNA as an input and then predicts the locations of genes. In bacteria, that usually means finding open reading frames (ORFs). Exons are not (usually) an issue with bacterial DNA. In eukaryotes, the challenge is finding exons and precisely determining the exon-intron boundaries.

As many genomes became well-characterized, gene prediction has taken a bit of a backseat in bioinformatics focus in recent years. However, with next generation sequencing technologies, many new genomes are being sequenced and there is a bit of an analysis gap. One of the first follow up steps to sequencing a new genome is to find the genes. The development of better automated programs will be of great importance in that effort.



Bacterial gene structure

First, some terminology. The phrase, "upstream from," is equivalent to "toward the 5' side of." So most bacterial genes have the following structure:

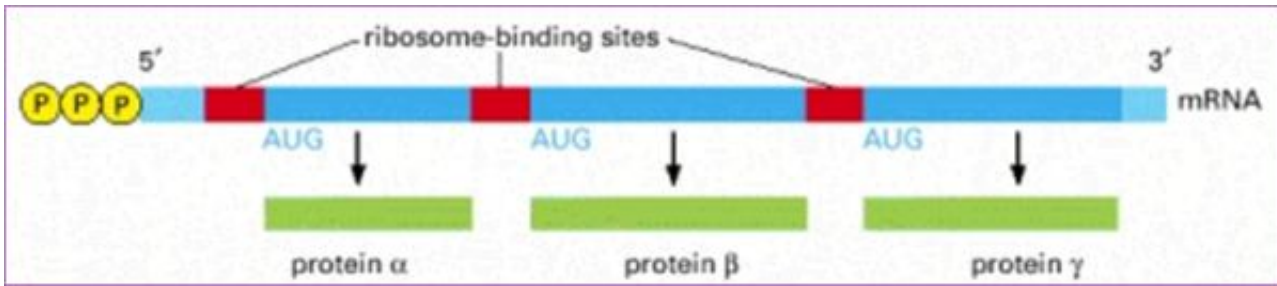
Promoters with consensus sequences at -35 and -10 (some genes)

The transcription start site (defined as +1)

A ribosome-binding site, called the Shine-Dalgarno sequence

One or more open reading frames (start codon to stop codon)

Transcription termination site



--image from Alberts et al. (2002). *Molecular Biology of the Cell*. Garland Science

Most importantly, not that the +1 position is defined as the first transcribed nucleotide, or the transcription start site. The -35 sequence is 35 bases upstream from the *transcription* start site. **The -35 sequence is NOT 35 bases upstream from a start codon.** That would be the *translation* start site.

Gene prediction strategy

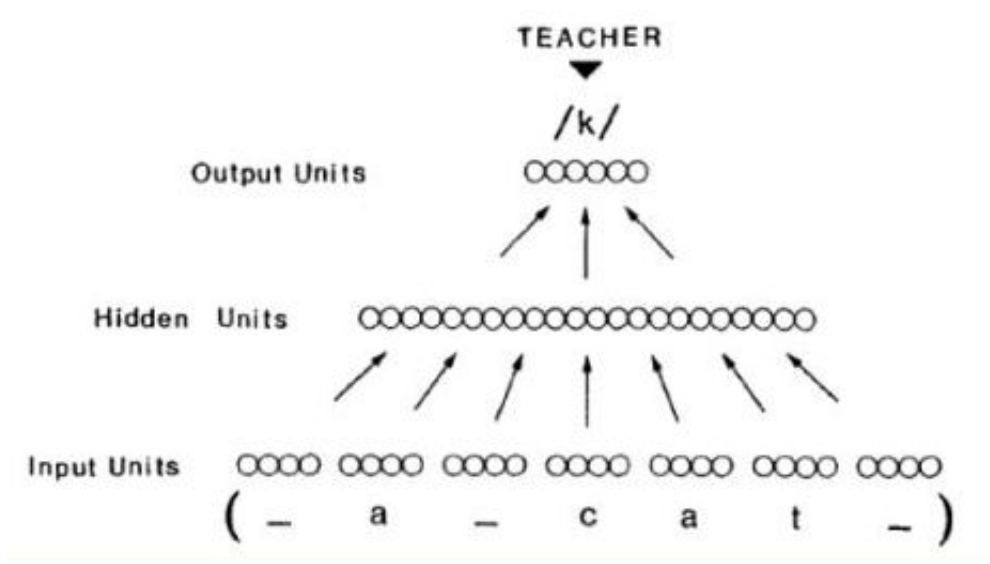
Most prokaryotic genomes are relatively small: 0.5-10 Mb. Transcription termination sites are stem-loop structures that help find the ends of genes. Most programs rely heavily on finding open reading frames. An open reading frame is a start codon and a stop codon "in the same frame," meaning that if you read the DNA triplets, the stop codon is relevant. Usually the longer open reading frames represent genes. The stop codon is easier to define. Sometimes there are multiple possibilities (all in frame) for the start codon. Finding the "real" one can be a little tricky.

The neural network approach works well with bacteria, but it requires a training set. More on neural networks below.

Verifying each gene prediction is usually not too difficult. Running BLAST searches can determine if a similar protein exists in the database. Because so many bacteria have been very well-characterized, it is rare to find a novel bacterial gene with no homologs in the database.

Neural networks

Let's start by reading the work of Sejnowski and Rosenberg from 1987. The initial experiment involve reading simple text and connecting sound with text. Below is a figure from that paper.



Now I'm going to oversimplify this significantly. Note that it is connecting input with output--in this case--text and sound. The idea is that if you train the network to properly connect text and sound, it those patterns can be applied to new inputs. The new inputs was a second page of text. Notably, the trained network encountered the word "two" and pronounced it, "TWOH," including the W sound. So it did pick up patterns. It's a way to approximate learning.

The following year, Qian and Sejnowski tried a similar approach to connecting protein primary structure to protein secondary structure, with a significant level of success (that proved to be difficult to improve on using this method).

So for bacteria, use well-annotated genomes as a training set input and the gene locations from those genomes as an output. The NN should then be able to take an unknown input (an unknown prokaryotic genome) and predict the output--the predicted locations of genes in that new uncharacterized genome.

Discussion topic: What do you know about neural networks?

ORF Finder

The [NCBI ORF Finder](#) searches all six reading frames of a bacterial sequence for potential ORFs. The ORFs should be at least 30 triplets long to even consider as a potential coding region. Some would argue that 50-60 triplets are necessary. You can enter a GI number here, so let's put in the *E. coli* lac operon, with GI number 146575. Look for long ORFs (rectangles) that line up nicely. There is usually very little intergenic space between genes in prokaryotes.

If you click any ORF, you should get the translated sequence. You can then do blastp searches off this page using the predicted protein. A homolog is a strong confirmation of that prediction.

Discussion topic: Where are the genes in the lac operon? How does it display alternative start codons? Any trouble doing Blast off that page?

GLIMMER

Glimmer stands for Gene Locator and Interpolated Markov Model Modeller. It has always been a downloadable program that runs on a unix platform but a [web server](#) has been available for about two years now. It uses a neural network coupled with an interpolated Markov model (IMM).

It might be the most popular program for annotating newly sequenced bacterial genomes, since many who work with sequence assemblers tend to use unix.

The concept is to use the complete genome to find long open reading frames. Because long open reading frames are almost always actual genes in bacteria, the long orfs are used to build a training set for which to analyze the rest of the genome. If you are analyzing unknown DNA and do not have the entire genome, then a related genome will do for the training.

GLIMMER v3.02 is available at NCBI on a [web server](#) and as a downloadable program. The web server is fast but there are fewer options for training sets.

FGENESB

Available at <http://www.softberry.com>, [FGENESB](#) is the bacterial program for gene prediction among many programs available there. The "Choose closest organism" window is important--that selects the training set. Pick the genome that you think most resembles the genome that you are analyzing. Note that the menu is alphabetical, but archaea come first--to get to the bacteria, scroll below the archaea.

FGENESB uses a 5th-order HMM and linear discriminant analysis. It is the leading web-based program for prokaryotes and may be the best overall program.

Discussion topic: What is a 5th-order HMM? Why would that be useful in gene prediction? Hint: certain organisms have what is called "codon bias." Some have a further bias toward pairs of codons.

Step-by-Step Description of FGENESB annotation.

Please note that web version of FGENESB includes only gene prediction and greatly simplified operon prediction portion of the program, represented by steps 3 and 4 below. Complete FGENESB package is available only for local installation.

STEP 1. Finds all potential ribosomal RNA genes using BLAST against bacterial and/or archaeal rRNA databases, and masks detected rRNA genes.

STEP 2. Predicts tRNA genes using [tRNAscan-SE](#) program (Washington University) and masks detected tRNA genes.

STEP 3. Initial predictions of long ORFs that are used as a starting point for calculating parameters for gene prediction. Iterates until stabilizes. Generates parameters such as 5th-order in-frame Markov chains for coding regions, 2nd-order Markov models for region around start codon and upstream RBS site, Stop codon and probability distributions of ORF lengths.

STEP 4. Predicts operons based only on distances between predicted genes.

STEP 5. Runs BLASTP for predicted proteins against COG database, cog.pro.

STEP 6. Uses information about conservation of neighboring gene pairs in known genomes to improve operon prediction.

STEP 7. Runs BLASTP against NR for proteins having no COGs hits.

STEP 8. predicts potential promoters ([BPROM](#) program) or terminators (BTERM) in upstream and downstream regions, correspondingly, of predicted genes. BTERM is the program predicting bacterial -independent terminators with energy scoring based on discriminant function of hairpin elements.

STEP 9. Refines operon predictions using predicted promoters and terminators as additional evidences.

-<http://www.softberry.com/berry.phtml?topic=fgenesb&group=help&subgroup=gfindb>

Other gene-finding programs

GeneMark: <http://www.ncbi.nlm.nih.gov/genomes/MICROBES/genemark.cgi>

EasyGene 1.2b: <http://www.cbs.dtu.dk/services/EasyGene/>

Prodigal: <http://prodigal.ornl.gov/server.html>