Functional Annotation

Background and Strategy

Bhanu Gandham
Maxine Harlemon
Ke Qi
Diana Williams
Xin Wu

Shashidhar Ravishankar
Subin Park
Menghan Li
Khalid Alhumimidi
Marzia Rizvi
• What is functional annotation?
• What are important features in our organisms of interest/background biology?
• What tools are available?
  • Tools for annotation of different features
  • Pipelines
• Make a new pipeline that includes tools missing from available pipelines
• Create a database of Variations
What is Functional Annotation?

• Attaching biological information to genomic information
• What does a particular area of the genome do? What is its biological consequence?
• Types of information gained:
  • biochemical function
  • biological function
  • involved regulation and interactions
  • expression

What is Functional Annotation?
Why is Functional Annotation Important?

- Identify the function of each area in the genome
- What is the role of than function in the cell?
- How is it the same or different in other organisms?
• Extrinsic annotation
  • Comparisons made against various databases like Ensembl, GENCODE, Gene Ontology Consortium, Uniprot, et al.
  • Quality of annotation dependent on quality of databases and similarity to other entries in the databases
• Intrinsic annotation
  • Based on characteristic of the gene or protein sequence
  • Examples: transmembrane proteins, signal peptides
• Closely linked with genome annotation and comparative genomics

Functional Annotation
• Serotypes/serogroups are groups within a single species of microorganisms, such as bacteria or viruses, which share distinctive surface structures.
• Serotyping/serogrouping is used for grouping sub species based on the chemical composition of the cell surface.
• It is used to distinguish between sub species which may look the same under the microscope but behave differently when interacting with the immune system and varies in the severity of infection.

Traditional Typing Methods - Serotype
• There are 12 Different serogroups for *N. meningitidis* based on the chemical composition of the capsular polysaccharide on the cell wall.

• The *N. meningitidis* serogroups which are the major cause of the disease nowadays are: A, B, C, Y and W135.
Slide Agglutination for Serogrouping (SASG)

Takes one day to grow in the appropriate media. Rating the intensity of the agglutination reaction at the final step: +2 (+50%) means serotype B.
Monoclonal Antibodies (Mabs)

Each dot represents a combination between the antibody and its antigen.

Dot-blot serotyping/serosubtyping
Haemophilus influenzae typing

- gram-negative, pleomorphic rods or coccobacilli
- large, colorless-to-grey, opaque colonies on a CAP
- 6 serotypes a-f and non-typable
• Both *H. influenzae* and *Haemophilus haemolyticus* require hemin (X factor) and NAD (V factor)
• *H. influenzae* does not display beta-hemolysis on horse blood agar
• *H. haemolyticus* may or may not display beta-hemolysis on horse blood agar, making it challenging to distinguish from *H. influenzae*

**H. influenzae serotypes**
Genetic Differences

*Neisseria meningitidis*

- 12 serogroups
- 6 of the serogroups cause meningococcal disease (A, B, C, W, Y)
- Capsule key virulence determinant
- cap locus
Genetic Differences

*Neisseria meningitidis*

Harrison et al. 2013
## Gene Targets for genotyping using real-time PCR assays

<table>
<thead>
<tr>
<th>Sero-group</th>
<th>Gene Target Name</th>
<th>Alternate Gene Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td><em>sacB</em></td>
<td><em>siaD</em></td>
</tr>
<tr>
<td>B</td>
<td><em>synD</em></td>
<td><em>siaD</em> of B</td>
</tr>
<tr>
<td>C</td>
<td><em>synE</em></td>
<td><em>siaD</em> of C</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>siaDC</em></td>
</tr>
<tr>
<td>W135</td>
<td><em>synG</em></td>
<td><em>siaD</em> of W135</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>siaDW</em></td>
</tr>
<tr>
<td>X</td>
<td><em>xcbB</em></td>
<td><em>siaD</em> of Y</td>
</tr>
<tr>
<td>Y</td>
<td><em>synF</em></td>
<td><em>siaDY</em></td>
</tr>
</tbody>
</table>

---

**Genetic Differences**

*Neisseria meningitidis*
• Capsule transport gene $ctrA$ and Cu, Zn superoxide dismutase gene $sodC$ targeted in species-specific assays
• $sodC$ not genetically linked to cap locus
  • useful for nongroupable $N. meningitidis$ that do not contain intact $ctrA$
• Two Main groups: typeable (encapsulated) and non-typeable (non-encapsulated)
• 6 capsule serotypes (a-f)
• Two phylogenetic divisions of *H. influenzae*:
  • I: most of a & b, all of c, d, and e; flanked by direct repeats of insertion element IS1016
  • II: f, some of a & b, IS1016 elsewhere but not flanking
• cap locus
Genetic Differences

*Haemophilus influenzae*

- Three functionally unique regions
- I and III common to all capsular types, homologs to *E. coli, N. meningitidis, A. pleuropneumoniae, P. multocida, M. haemolytica*
- II: serotype-specific biosynthesis genes

Satola et al. 2003
Genetic Differences

*Haemophilus influenzae*

CDC
- *hpd*
  - Protein D
  - highly conserved
  - in all encapsulated and non-encapsulated *H. influenzae*
Genetic Differences

*Haemophilus haemolyticus*

- Usually non-pathogenic
- Rarely causes subacute endocarditis
- Phylogenetic tree based on genes adk, pgi, recA, infB, and 16s rRNA
- iga gene encoding for enzyme immunoglobulin A1 protease and 7F3 epitope of P6 in *H. influenzae* but not *H. haemolyticus*
- Standard tests do not distinguish between it and Hi-X and V

Sandstedt et al. 2008
Possible Predicted Features

- Coding:
  - Domains/Motifs
  - Signaling Peptides
  - Transmembrane Regions
  - Operons
  - Pathways

- Non-Coding:
  - Ribosomes
  - tRNA
  - CRISPRs
  - small RNA
• **Sequence Motifs** are short, recurring patterns in **DNA** that are presumed to have a biological function. Often they indicate sequence-specific binding sites for proteins such as nucleases and transcription factors (TF).

• **A domain** is a discrete structural unit that is assumed to evolve, exist and fold independently of the rest of the protein and to have its own function.
Common motif in DNA binding proteins

Domains/Motifs
Importance of Motifs

- Indicates common structural protein domains
- Identifies similar function
- Other possible biological functions, e.g. Transcriptional factors, mRNA processing etc.
Importance of Protein Domains

- Bind to other molecules in the cell
- Signal transduction pathways
- Genetically engineering novel proteins
- Pharmaceutical importance
Algorithmic approaches for both DNA motifs and protein domain searches

- Enumeration
- Probabilistic Optimization
Enumeration

Word-enumeration techniques count the number of occurrences of a motif, defined as a string of letters \{a,c,g,t and sometime with degenerate letters, e.g., y = c,t and r = a,g\} of certain length (e.g., 6–20) in the sequence data. When no degenerate letters are used in the motif profile/model, a subsequence is considered an instance of the motif when the number of mismatches between the subsequence and the motif is less than a threshold. The motifs are then rank-ordered based on their overrepresentation, thus, these approaches guarantee the global optimum—e.g., producing motifs with the highest overrepresentation.

Domains/Motifs
Probabilistic Optimization

Most probabilistic methods involve iteratively scanning the input sequences to identify potential motifs and then updating the motifs to improve the likelihood of the model until convergence.

Domains/Motifs
Tools

• **Cd Search**: CD-Search is NCBI's interface to searching the Conserved Domain Database with protein or nucleotide query sequences. It uses RPS-BLAST, a variant of PSI-BLAST, to quickly scan a set of pre-calculated position-specific scoring matrices (PSSMs) with a protein query.

• **Interproscan**: InterProScan is the software package that allows sequences (protein and nucleic) to be scanned against InterPro's signatures. Signatures are predictive models, provided by several different databases (referred to as member databases), that make up the InterPro consortium.

Domains/Motifs
• A signal peptide is a short (5-30 amino acids long) peptide present at the N-terminus of the majority of newly synthesized proteins that are destined towards the secretory pathway.

• Signal peptides do not consist of a strict consensus sequence but have a three-region design consisting of a positively charged N-terminal region (n-region), a hydrophobic central region (h-region) and a neutral, polar C-terminal region (c-region) with small and neutral amino acids at position -1 and -3.

Signal Peptides
Signal Peptides
• SP prediction is fundamentally important as it impacts on other features such as transmembrane topology, subcellular localization, structure modeling and prediction, assignment of putative functions to novel proteins and identification of putative cleavage sites in database annotation.

• The systematic functional annotation of biological sequences using Gene Ontology requires a precise knowledge of the subcellular localization, where SP prediction has a fundamental input.

Signal Peptides
• Detection of Signal peptides: algorithms and models such as artificial neural network, hidden Markov model, dynamic Bayesian network, position weight matrix
• Popular SP prediction tools:
  1. SignalP: ANN, HMM
  2. Phobius: HMM
  3. PrediSi: PWM
  4. SPOCTOPUS: NN, HMM
• Transmembrane proteins are helix-loop structures which are found embedded in the cell membrane.

• They serve as a gateway for the cell, enabling transport substances in and out of the cell.

• Transmembrane proteins have some general characteristics which can be used to identify them.
  • Hydrophobicity of helix.
  • Abundance of positive charge on the cytoplasmic side.
  • Pattern of cytoplasmic and non-cytoplasmic loop structures.

• The most widely used \textit{ab-initio} methods to detect membrane proteins use HMMs to determine the structure and topology.

Transmembrane Proteins
HMMs

Eddy 2004
TMHMM

Krogh et al. 2001
**HMMTOP**

Tusnady and Simon 1998
Gene Ontology (GO) project provides consistent descriptions of gene products.

- Developed since 1998
  - FlyBase
  - SGD
  - MGD

- Includes three terms
  - Biological Process
  - Molecular Function
  - Cellular Component

- Database updates EVERYDAY

Gene Ontology
• Relationship Structure:
  • DAG (directed acyclic graph)
• Edge logic:
  • is_a
  • part_of
• Application:
  • Gene annotation
  • Expression analysis (GO enrichment)
• Website Tools for GO annotation
  • AmiGO 2
    http://amigo2.berkeleybop.org/amigo
  • DAVID
    http://david.abcc.ncifcrf.gov/
  • And hundreds more!

• Command line Tool for GO annotation
  • b2g4pipe

Gene Ontology
• The **ribosome** is a large and complex molecular machine, found within all living cells, that serves as the site of biological protein synthesis (translation).

• Ribosomes link amino acids together in the order specified by messenger RNA (mRNA) molecules.

• Ribosomes consist of two major components — the **small** ribosomal subunit which **reads** the RNA, and the **large** subunit which **joins** amino acids to form a polypeptide chain.

• Each subunit is composed of one or more ribosomal RNA (rRNA) molecules and a variety of proteins.

**Ribosomes**
• **Ribosomal ribonucleic acid** (rRNA) is the RNA component of the ribosome, which is approximately 60% rRNA and 40% protein by weight.

• In prokaryotes a small 30S ribosomal subunit contains the 16S ribosomal RNA.

• The large 50S ribosomal subunit contains two rRNA species (the 5S and 23S ribosomal RNAs).

**Ribosomes**
• The ribosomal RNAs form two subunits and the mRNA is sandwiched between the two, and the ribosome catalyzes the formation of a peptide bond between the two amino acids that are contained in the rRNA.
• rRNA is one of only a few gene products present in all cells.
• rRNA sequences are widely used for working out taxonomic groups and evolutionary relationships among organisms, since they are of ancient origin and are found in all known forms of life.
• **FACT**: Study of ribosomal RNA led to the definition of three separate “Domains” of life; Eukaryotes, Bacteria, and Archaea.
• The Ribosomal Database Project (RDP) and SILVA are high quality ribosomal RNA databases of aligned and annotated rRNA sequences from the major sequence repositories (GenBank, EMBL, and DDBJ).

• RNAmmer: The program uses hidden Markov models trained on data from the 5S ribosomal RNA database and the European ribosomal RNA database project.
Probabilistic parameters of a hidden Markov model (example) $X$ — states $y$ — possible observations $a$— state transition probabilities $b$ — output probabilities
• A transfer RNA is a small RNA molecules involved in the translation of the nucleic acid message into the amino acids
• Typical Structure
  • 5’P
  • D arm
  • anticodon arm
  • variable arm
  • TyC arm
  • 3' amino acid acceptor stem

Transfer RNAs (tRNAs)
• Methods based on
  • Search for linear sequence
  • Covariance models
  • Rule based systems

Prediction of tRNAs
Prediction of tRNAs (tRNAscan)

Fichant and Burks 1991
• Clustered Regularly Interspaced Short Palindromic Repeats

• Repeat sequences, 21-47 nucleotides in length, with interspaced non repeating *spacer* sequences of similar size.

• Found in approximately 40% of bacterial genomes investigated till now.

• Known to provide immunity against viruses and plasmids.
• *Ab initio* detection of CRISPR sequences involves identification of repeat regions with interspaced sequences of similar length.

• There are two different approach which we are evaluating:
  • Alignment based approaches.
  • K-mer based approaches.
CRISPR detection using alignment based method (piler-cr)

Edgar 2011
CRISPR detection using k-mer based method (CRT)

- Uses a k-mer matching approach to detect repeat regions within a given range.
- Once exact matching repeating k-mers are found, the k-mers are extended to check if they satisfy the conditions of a CRIPR array.

Bland et al. 2007
Small RNA (sRNA) is small (30-500 nt) non-coding RNA produced by bacteria.

- Highly structured and contain stem-loops.
- Functions - gene regulation
  - Regulation of RNA polymerase e.g. 6S RNA.
  - Regulation of signal recognition particle e.g. 4.5S RNA.
  - Regulation of stress responser e.g. dsrA, rprA and oxyS for rpoS gene.
  - Regulation of outer membrane proteins e.g. RNAIII of S. aureus.
- Database for sRNA
- BSRD

http://www.bac-srna.org/BSRD

Small RNA
• Genes coding for functionally related proteins are clustered along the DNA.
• Allows coordinated protein synthesis in response to the needs of the cell.
• Typical Structure:
  • Regulatory genes
  • Promoter
  • Operator
  • Structural genes
  • Terminator
Example: Lac Operon

Lactose binds to the inhibitor and prevents it binding to the operator, thereby allowing transcription.

Binds to the operator to inhibit transcription.

Enzymes to import lactose and break it down.

Ralston 2008
Prediction of Operons

- Methods based on
  - finding signals
    - Rely on well studied operons
  - finding gene clusters
    - Rely on genome sequences
1. Search all genes from available genomes against one another using BLASTP.
2. A conserved gene pair is defined as two adjacent genes (A,B) for which a homologous gene pair (A’,B’).
3. Genes in conserved S pairs are candidates for membership in the same operon (SO pairs).
4. Calculate probability of the genes of being in the same operon.

\[
P \geq \left(1 + k \times \frac{2N\text{(directions)} + N\text{(S pairs)} - N\text{(genes)}}{N\text{(S pairs)}} \right) - \left(\frac{0.1G}{N\text{(conserved S)}}\right)^h
\]

where
- P - probability
- k - constant
- N(S Pairs) - Pair of genes on same strands
- N(genes) - Number of genes
- 2N(directions) - Number of operons
- G - number of genomes in the database
- h - number of unrelated genomes
• High-throughput technologies often identify hundreds of genes related to biological or pathological processes. From these genes one wants to identify biological pathways that may be involved and diseases that may be implicated.

• Most Common Pathways:
  • Metabolic Pathways
  • Gene Regulatory Pathways
  • Signal Transduction Pathways
Metabolic Pathway

Pathways
Gene on DNA → Primary transcript → mRNA → Protein

Gene Regulatory Pathway

Pathways
Signal Transduction Pathway
ab initio prediction of pathways and diseases is challenging. One feasible approach is to use existing databases of known metabolic and signaling pathways and databases of known disease-associated genes as the starting point for annotation of a new set of genes.

Databases we are looking into:
- KEGG PATHWAY,
- KEGG DISEASE,
- PID,
- BioCyc,
- Reactome
- Panther
KOBASES 2.0 – Pathway Annotation Tool

Pathways
• General Requirements:
  • command line-based, not web-based, better for scaling
  • are the tools being used in available pipelines the best available?
  • What tools do we need that are not in available pipelines but would be helpful for our task?

Tools for Functional Annotation
- Test pipelines and compare results
- Test individual tools and compare to each other and pipeline results
- Test tools not contained within the pipelines
- Create new pipeline
  - Optimized individual tools, may or may not contain elements of existing pipelines
- Create database of variations

Strategy and Deliverables
• Laslett and Canback. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 32:11-16.


