Making Sense of Biological Sequences

Genome Annotation Overview

Nelson Ndegwa
Bioinformatics Department
ILRI/IITA
Genome Annotation

The process of taking the raw DNA sequence produced by the genome-sequencing projects and adding the layers of analysis and interpretation necessary to extract its biological significance and place it into the context of our understanding of biological processes.
Objectives

• Give an overview of the genome annotation processes at the:
  1. Nucleotide-level
  2. Protein-level
  3. Process-level

• Introduce an example annotation platform
Disclaimer

Not going to review all the annotation tools in this talk. Simply, to summarize the key concepts, software and how they are integrated with various data sources to provide annotations.
Why annotate?
Bioinfo - magic?
Why annotate?

- Annotation bridges the gap from the sequence to the biology of the organism
- To identify the key features of the genome – genes and their products.

A genome is only as good as its annotation.
1. Nucleotide level annotation

> Some FASTA formatted file

```
GTACCTTGATTTTCGTATTCTGAGGCTGCTGCTTTAGCGGTAGCCCTTTTGGTTTCGTTGGAACCGGA
GCGCGGGAATTACAGATAAAATTAAAAACTGCGACTGCGCGAGCGTGCAGCAGACTTCCTGAGAAGG
GACAGGCTGTGGGTATCTCAGATAACTGCGACTGCGCGGCGTGAGCTCGCTGAGACTTCCTGGACGG
GACAGGCTGTGGGTATTCTTAGGGGGTAGGGGCGGGAACCTGAGAGGCGTAAGGCGTG
GATTTCGAAGCTGACAGATGGGTATTCTTTTAGGGGGGTTAGGGGCGGGAACCTGAGAGGCGTAAGG
GCTGTGGGTATTCTTAGGGGGTAGGGGGCTGCAGCTCTATCAGCGAAGGCAGCTGCGACTGCG
AGTGTCCGTGGGGGAATCCTCTGAGGTACCGGAATATGCTTGGAGGGGACACTATGTGCTTTTA
CGTCCGGCTGTGGCTGTTATGAGGTACCGGAATATGCTTGGAGGGGACACTATGTGCTTTTA
AATACAAATGCGAGCCGGGCTGTGGCCCGGCTACATCAGAGGAGGGCAGGAGAATCGCTAGAGC
GCCGGGAGGCGGAGGGTGCCAGTGGAGCCGCGACTCCAGGGCGACAGACGAGA
AATACAAAAATTAGCCGGGCGTGGTGCCGCTGCTACTCAGGAGGCTGAGGCAGGAGAATCGCTAGA
AATACAAAAATTAGCCGGGCGTGGTGCCGCTGCTACTCAGGAGGCTGAGGCAGGAGAATCGCTAGA
```

[ILRI logo]

[IIITA logo]
Where are:

- Genes?
- Genetic markers?
- tRNAs?
- rRNAs?
- Repetitive elements?
- SNPs?
Gene finding

Gene finding refers to identifying stretches of sequences (genes) in genomic DNA that are biologically functional.

It is crucial in understanding the genome of a species.
Figure 1: A protein-coding gene. (This picture is from Sanja Rogic's lecture slides, "Computational Gene Finding")
Gene finding

- **Prokaryotes**: Small genomes e.g. ~ 5 Million in bacteria, gene finding is a matter of identifying long ORFs.
- Challenges arise when long ORFs overlap on opposite strands.
- E.g *Haemophilous influenzae* ~ 85% of its genome is in coding regions
- Calling genes involves using programs that carry out 6 frame translation and identifies all ORFs longer than a chosen threshold.
Gene finding

• **Eukaryotes:** For yeast, <25% of its genome is in coding regions. For human its <3%

• Splicing and alternative splicing

• Intron & exons present

• Large genomes e.g. human ~3Billion bases
Splicing leads to alternatively spliced mRNAs, which are then translated into proteins.
Approaches

• Computational methodology for finding genes in a genome has evolved significantly over the last 20 years.
• Many approaches have been proposed to find genes in both prokaryotes and eukaryotes.
• These approaches mainly fall into three categories: homology-based approaches, Ab Initio approaches and comparative genomics approaches.
Homology-based Approaches

These approaches are based on the similarity of sequences.

• Given a library of sequences of other organisms, we search target sequence in this library and identify library sequences (known genes) that resemble the target sequence.

• If the identified sequences are genes, the target sequence is probably (putatively) a gene.

• These approaches are able to find biologically relevant genes.
Homology-based approaches

Limitation:

• Identifies only genes that code for proteins present in the database.

Example:

• BLAST (Basic Local Alignment Search Tool)
• FASTA
Ab Initio Approaches

Rely on statistical qualities of exons rather than on homologies.

Examples: GenScan, fgenesh, HMMER

Limitations:
Prone to many false-positives.
Often unable to predict splice variants correctly.
Cannot predict 3’ and 5’ UTRs.
Ab Initio Approaches

Application:
Using prediction programs together with other data sources e.g. EST data can provide valuable sources of annotation.
3. Comparative Genomics Approaches

Capitalize on the continual availability of more genomes becoming publicly available from the various genome sequencing projects.

Application:
An attractive use of this is to use the presence of conserved sequences to deduce all functional regions in a genome. E.g. protein coding genes are likely to be conserved in closely related species. E.g. human & mouse.
Comparative Genomics Approaches

Software:
Artemis comparison tool (ACT)
E.Coli genome comparison in ACT
2. Protein-level annotation

Concerned with the compilation of a definitive catalogue of proteins of the organisms, name them and assign putative function.

Humans have between 20-25,000 genes, Haemophilus influenzae ~1,709 genes.
2. Protein-level annotation

Most genome sequencing projects, generally continue to have numerous proteins of unknown functions.

Protein family classification uses similarities to better characterize proteins.
Protein comparison

• Cf of proteins **between species** is a rich source of functional annotation based on high similarity.

• E.g. if a well characterized protein in yeast is known to be involved in the initiation of DNA replication, then it's likely that a predicted protein from human sequence similar to it performs the same function.
Caveat

- Orthologue vs. Paralogue

Orthologues: Genes directly descended from a common ancestor after speciation likely to perform same function.

Paralogues: Genes duplicated in the same organism might not have the same functional role.
Typical protein annotation pipeline

Search for protein similarity:
Will search for similarity using BLASTP or PSI-BLAST tools against several different databases of protein sequences.

Databases: SwissProt & SwissProt-TrEMBL
SwissProt

• Human curated collection of confirmed protein sequences. (nr)
• Extensively cross-referenced with other sequence & structure dbs.
• Have descriptions of function and biological role.
• Highly reliable
• Few cf to all sequenced genomes.
TrEMBL

• Automatically filtered.
• Non-redundant.
• Domain searches performed but not verified.
• Next in line to be added to Swissprot
Typical protein annotation cont’d

Domain searches: alternative to Swissprot/TrEMBL

• Protein domains are also essential in determining the function of predicted genes.
• Different DBs can be used.
• Each has strengths and weaknesses.
Typical protein annotation cont’d

Databases of functional domains:

• **PFAM**: Collection of HMM profiles & alignments for common protein families.
• **PRINTS**: DB of short protein motifs, folds and domains.
• **PROSITE**: Longer protein signatures of known profiles.
• **ProDom**: collection of protein domains derived from PSI-BLAST
• **BLOCKS**: conserved protein regions and their multiple alignments.
• **SMART**: curated collection of protein domains
Which protein database to use?

**InterPro:**

- It integrates all the protein signatures and domains in one resource.
- Provides least-redundant and extensive annotation currently available.
Nucleotide + Protein Level Annotation

Source: Yandell et. al 2012 (Eukaryotic Genome Annotation)
3. Process-level annotation

How do the building blocks of genes and proteins relate to, say:

• Cell cycle,
• Apoptosis,
• Metabolism or
• Health & Disease
• Resistance?
Relating the genome to biological process
Gene Ontology (GO)

A standard vocabulary for describing function of Eukaryotic genes.
Gene Ontology (GO)

Consists of 3 subparts:

• **Molecular function** – task carried out by individual genes e.g. enzymatic activity

• **Biological Process** – Used for broader biological roles e.g. meiosis

• **Cellular component** – describes genes in terms of subcellular structure e.g. localized to nucleus.
GO

Organized as a hierarchy of terms (DAG) – that allows a term to appear in several places in the hierarchy.

This flexibility allows genes to be annotated to whatever level of specificity the current biological understanding allows.
HT- Techniques support GO

Process-level annotation extends beyond purely computational work. HT-techs provide vital clues to the roles that genes and proteins have in biological processes and provide a rich layer of annotation.
Example HT-techs

- Microarray expression analysis
- Direct assay of Protein expression (MS)
- RNA interference
- GFP assays
- Y2H assays
### Experimental Evidence Codes

- **EXP**: Inferred from Experiment
- **IDA**: Inferred from Direct Assay
- **IPI**: Inferred from Physical Interaction
- **IMP**: Inferred from Mutant Phenotype
- **IGI**: Inferred from Genetic Interaction
- **IEP**: Inferred from Expression Pattern

### Computational Analysis Evidence Codes

- **ISS**: Inferred from Sequence or Structural Similarity
- **ISO**: Inferred from Sequence Orthology
- **ISA**: Inferred from Sequence Alignment
- **ISM**: Inferred from Sequence Model
- **IGC**: Inferred from Genomic Context
- **RCA**: Inferred from Reviewed Computational Analysis

### Author Statement Evidence Codes

- **TAS**: Traceable Author Statement
- **NAS**: Non-traceable Author Statement

### Curator Statement Evidence Codes

- **IC**: Inferred by Curator
- **ND**: No biological Data available

### Automatically-assigned Evidence Codes

- **IEA**: Inferred from Electronic Annotation

**Evidence codes**

- **IDA**: enzyme assay
- **IPI**: e.g. Y2H

**Subcategories of ISS**

- BLASTs, orthology comparison, HMMs

**Review papers**
How scientists use the GO

• Access gene product functional information
• Analyse high-throughput genomic or proteomic datasets
• Validation of experimental techniques
• Get a broad overview of a proteome
• Obtain functional information for novel gene products
Bregje Wertheim at the Centre for Evolutionary Genomics, Department of Biology, UCL and Eugene Schuster Group, EBI.
Implication

Conventional bench research and genome annotation begin to merge.

Each experiment adds an item of information to our knowledge of biology and this in turn enhances our understanding of the genome through the genes and proteins it touches.