Investigating questions in biology using computational approaches

Balaji Santhanam
MRC Laboratory of Molecular Biology, Cambridge

At what level is this talk pitched at?

Data

- Development of methods
- Algorithms, programs, etc
- Uncovering general principles
- Discovery using computational approaches
- Prioritising experiments
- Interpreting experimental results

Focused
Genome-scale

Qualitative
Quantitative

- Sequences
- Protein interactions
- qPCR
- Transcription data

"Biologically" inclined
"Computationally" inclined
Comprehensive understanding of a gene

- **Identity information**
  - Sequence
  - Structure

- **Conditional information**
  - Transcript abundance
  - Protein abundance
  - Half-lives

- **Regulatory information**
  - DNA-protein interaction
  - DNA-DNA interaction
  - Post-translational control

- **Interaction information**
  - Genetic interactions
  - Protein interactions

Transcriptomics + mass cytometry: Time course and populations
Gene

No clue or want to more

Conducted high-throughput Experiments

Aim to make sense of the data

And/or

Identify most relevant set of genes
Outline

- Introduction to resources and tools (20-25 minutes)
- Some case studies (15 mins)
- High-throughput data

How can one extract information?

<table>
<thead>
<tr>
<th>Low information content</th>
<th>High information content</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Query</strong></td>
<td><strong>Network analyses</strong></td>
</tr>
<tr>
<td>Search tool</td>
<td><a href="http://www.cytoscape.org">www.cytoscape.org</a></td>
</tr>
<tr>
<td>Database</td>
<td>BIOGRID database</td>
</tr>
<tr>
<td>Data</td>
<td>ArgoExpress database</td>
</tr>
<tr>
<td>Ranking, statistics</td>
<td>Next-gen sequencing data analyses</td>
</tr>
<tr>
<td><strong>Output</strong></td>
<td></td>
</tr>
<tr>
<td>Referencing</td>
<td></td>
</tr>
<tr>
<td>Literature</td>
<td></td>
</tr>
</tbody>
</table>

Information retrieval (think of Google)
What is the list of databases that is currently available?

Over 1700 databases covering various aspects of molecular and cell biology

I. Nucleic acid sequence, structure, and regulation
II. Protein sequence and structure, motifs and domains
III. Metabolic and signalling pathways, enzymes
IV. Viruses, bacteria, protozoa, and fungi
V. Human genome, model organisms, comparative genomics
VI. Genomic variation, diseases and drugs
VII. Plant databases
VIII. Metagenomics

Prominent recent databases
- Cell Model Passports: Human Cancer Cell Models
- Cancer SEA: Cancer Single-Cell Atlas
- Editiome Disease Knowledgebase: Curated collection of RNA editing events
- ViBrismDB: Tomographic transcriptome

ViBrism Database: Expression meets Tomography

https://vibrism.neuroinf.jp/
Question #2: Tools

What is the list of tools, web-servers and programs that are currently available?

https://www.ebi.ac.uk/services

https://genome.ucsc.edu/
Welcome to DAVID 6.8

2003 - 2017

The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.8 compiles a full Knowledgebase updated to the latest version of our original web-accessible program. DAVID now provides a comprehensive set of functional annotation tools for investigations to understand biological meaning behind large lists of genes. For any given gene list, DAVID tools are able to:

- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related gene groups
- Cluster redundant annotation terms
- Visualize genes on BioCarta & KEGG pathway maps
- Display related gene-to-term or 2-D views
- Search for other functionally related genes in the list
- List intersecting proteins
- Explore gene names in batch
- Identify gene-gene associations

What’s important in DAVID?

- DAVID 6.8
- 100% of DAVID tools and gene arrays supported
- Novel Classification Algorithms
- Pre-built Allele-integrated and lineage backgrounds
- Gene-customized gene background
- Enhanced calculating speed

Statistics of DAVID

DAVID Biinformatics Resources Citations

https://david.ncifcrf.gov/
Question #3: Literature

What is the list of databases, web-servers and programs that are currently available to explore the literature?


http://apps.isiknowledge.com

http://www.scopus.com/home.url
http://www.ihop-net.org/UniPub/iHOP/
### Explosion of information about living systems

<table>
<thead>
<tr>
<th>Category</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence</td>
<td>&gt; 45,000,000 sequences from &gt; 160,000 organisms (Genbank, NCBI, UniProt)</td>
</tr>
<tr>
<td>Structure</td>
<td>150,861 structures from &gt; 1500 organisms (PDB, MSD)</td>
</tr>
<tr>
<td>Expression</td>
<td>&gt;100,000 different conditions &gt; 200 organisms (SMD, GEO, ArrayExpress)</td>
</tr>
<tr>
<td>Interaction</td>
<td>&gt;800,000 interactions 60 organisms (Bind, DIP, BIOGRID, publications)</td>
</tr>
<tr>
<td>Literature</td>
<td>Over 50 million abstracts and papers Numerous organisms (PubMed, ISI, Scopus)</td>
</tr>
</tbody>
</table>

**Major challenge – How to exploit this information?**

---

**Sequence**
- **Query sequence**
  - BLAST, PSI-BLAST, etc
  - Sequence database (NCBI, ENSEMBL)
  - E-value, score, etc
  - Sequence Alignment

**Structure**
- **Query structure**
  - DALI, TopSearch, VAST, etc
  - Structure database (PDB)
  - p-value, score, etc
  - Structure Alignments
Expression database (GEO, BioGPS, Broad Institute)

Relevant conditions

Correlation coefficient, hierarchical clustering

Other transcripts with similar expression profile

Query gene

Gene list

Sub-graph extraction

Genetic interaction networks, Protein interaction network, pathways, etc (BIOGRID, INTact, KEGG)

P-value, enrichments

Network of functional association

Functional enrichment

How to integrate in a context-specific manner?

Outline

• Introduction to resources and tools (20-25 minutes)

• A case studies (15 mins)

• High-throughput data
Previous comparative genomic analysis of eukaryotes suggested lack of detectable transcription factors in Plasmodium.

**Large number of genes**
- 5,300 genes with over 700 metabolic enzymes
- Extensive complement of chromosomal regulatory proteins
- Extensive complement signaling proteins (GTPases, kinases)

**Complex life cycle**

The Problem!
*How does this pathogen regulate gene expression?*

Possible explanations for the paradoxical observation

**Alternative regulatory mechanisms**
- Chromatin-level regulation
- Post-translational modification
- RNA based regulation

**Undetected transcription factors**
- Distantly related or unrelated to known DNA binding domains

Proteome of Plasmodium

Profiles & HMMs of known DBDs

The suspect!

AT-Hook

PF14_0633

SEG

Uncharacterized Globular domain ~60 aa

Profiles & HMMs of known DBDs

- bZIP
- Homeo
- MADs
- Forkhead
- HMG
- AT-hook
- ARID

**The suspect!**
Characterization of the globular domain – sequence analysis I

Non-redundant database

Profiles + HMM

Non-redundant database

Lineage specific expansion in Apicomplexa

Floral Homeotic protein Q

(Triticum)

49L, an endonuclease

(X. oryzae phage Xp10)

Globular region maps to AP2 DNA-binding domain

Characterization of the globular domain – structural analysis I

Predicted SS of ApiAP2

SS of ATERF1

A. thaliana ethylene response factor

(ATERF1 - 1gcc – NMR structure)

Binds GC rich sequences

12 residues show a strong pattern of conservation and these are involved in key stabilizing hydrophobic interactions that determine the path of the backbone in the three strands and helix of the AP2 domain

Core fold of the ApiAp2 domain will be similar to the plant AP2 DNA-binding domain

Balaji et al, 2005
Characterization of the globular domain – expression analysis I

**Complex life cycle**

Liver

Mosquito

Human

RBC infection & merozoite burst

**Intra-erythrocyte developmental cycle**

DeRisi Lab

Characterization of the globular domain – expression analysis II

**22 Transcription factors**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Ring stage</th>
<th>Trophozoite stage</th>
<th>Early Schizont stage</th>
<th>Schizont stage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Co-expressed genes**

<table>
<thead>
<tr>
<th>Genes</th>
<th>Time points</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

**Striking expression pattern in specific developmental stages suggests that they could mediate transcriptional regulation of stage specific genes**
Integration of different types of experimental data allowed us to discover potential transcription factors in the *Plasmodium* genome.

Integration of data can generate experimentally testable hypotheses.

**Confirmations**
Coevolution to identify protein functions


Thomas A. Hopf, Lucy J. Colwell, Robert Sheridan, Burkhard Rost, Chris Sander, Debora S. Marks Cell, 2012

Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science, 1999

Evfold.org
WW domain: Evolutionary perspective

Intrinsically unstructured or disordered proteins/regions

http://elm.eu.org/

Welcome to the Eukaryotic Linear Motif (ELM) resource

This computational biology resource mainly focuses on annotation and detection of eukaryotic linear motifs (ELMs) by providing both a repository of annotated motif data and an exploratory tool for motif prediction. ELMs, or short linear motifs (ELMs), are simple protein interaction sites composed of short sequences of adjacent amino acids. They are enriched in intrinsically disordered regions of the proteome and predict a wide range of functional and physical interactions (such as protein-protein, protein-DNA, or protein-RNA). They play crucial roles in signal regulation and are also of clinical importance, as altered ELM functions have been associated with several diseases. ELM families are often used by pathologists to analyze their data and include machinery (e.g., the Y ring).

ELM Prediction

The ELM prediction tool uses user-submitted peptide sequences for matching to the regular expression-defined in ELM. Statistics are made between motifs that are experimentally validated and those unresolved entries contained in the ELM database and matches that correspond to peptide motifs based on the sequence. Given ELMs are short and degenerate, overprediction is likely and many potentially ELMs will be false positives. However, prediction power is improved by using additional filters based on contextual information, including taxonomic cellular compartments, evolutionary conservation, and structural features.

Protein sequence

Enter protein identifier or accession number: (痴恨痴恨)
1.g. EPM_HUMAN, FSTAL_HUMAN, [DAK]
• Introduction to resources and tools (20-25 minutes)

• A case studies (15 mins)

• High-throughput data

Next-gen sequencing data analysis
Web server integrated platform:

- Galaxy server: https://galaxyproject.org

Software packages for differential gene expression


- **RSEM (RNA-Seq by Expectation-Maximization)** [http://deweylab.github.io/RSEM/](http://deweylab.github.io/RSEM/) integrated to EBSeg

Reference genome mapping and exon-exon junction identification

- **TopHat**: [http://ccb.jhu.edu/software/tophat/index.shtml](http://ccb.jhu.edu/software/tophat/index.shtml) uses Bowtie to map and identify splice junctions between exons
Global characteristics of the mapping tools

<table>
<thead>
<tr>
<th>Tool</th>
<th>Format</th>
<th>Algorithm</th>
<th>Threads</th>
<th>Gaps</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>BWA</td>
<td>SAM</td>
<td>BWT</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Novoalign</td>
<td>SAM</td>
<td>hash the ref.</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Bowtie</td>
<td>SAM</td>
<td>BWT</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>SOAP2</td>
<td>perso</td>
<td>BWT</td>
<td>yes</td>
<td>no</td>
<td>at most 2</td>
</tr>
<tr>
<td>BFAST</td>
<td>SAM</td>
<td>hash the ref.</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>SSHAIA2</td>
<td>SAM</td>
<td>hash the ref.</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>MScope</td>
<td>perso</td>
<td>suffix tree</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>GASSST</td>
<td>SAM</td>
<td>hash the ref.</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>PerM</td>
<td>SAM</td>
<td>hash the ref.</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>
RNA-Seq data analyses for differential gene expression

<table>
<thead>
<tr>
<th>Non-parametric methods</th>
<th>Parametric methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOISeq</td>
<td>DESeq</td>
</tr>
<tr>
<td>EBSeq</td>
<td>EdgeR</td>
</tr>
</tbody>
</table>

mRNA samples from mouse oocytes paired-end sequencing


Proteomics data

http://www.humanproteomemap.org/

http://www.proteinatlas.org/
Protein interactions databases

https://thebiogrid.org/

http://www.ebi.ac.uk/intact/

Huttlin et al, Cell 2015
Protein-protein interaction data set – BioPlex network

Huttlin et al, Nature 2017

www.cytoscape.org

Cytoscape
Network data integration, analysis and visualization in a box

Introduction
Download 3.3.0
Machine Learning

- Supervised learning
- Unsupervised learning

Supervised learning: Classification of input(s)
Unsupervised learning: Grouping of input(s)

Machine Learning 101: IRIS flower classification


The caret Package in R
General approach to investigate biological questions using a computational approach

WHY
1. Formulate the big question and have valid reasons

WHAT
1. Come up with several specific questions
2. Prioritise questions and prepare a checklist

HOW
1. Identify the database
2. Identify the tools
3. Be aware of the basic statistics
4. Retrieve and integrate the information

FRAME MORE WHY
1. Formulate hypothesis and READ A LOT!
2. Design experiments
3. Publish work & be happy ever after 😊