Gene expression bioinformatics:
Part 1. High-throughput quantitative genomics

September 17, 2013

Alvin T. Kho
Boston Children's Hospital

alvin_kho@hms.harvard.edu
Characterizing a biological system.

- Organizational scales and constituent features of a biological system
- **Central Dogma of molecular biology (CD):**
  - Centrality of genes in investigations of biological studies.
  - Reductionism.
- Identification of gene to DNA.
- The concept of a gene, gene expression, transcriptome.

2 central objectives in quantitative genomics

- 1. Identify genes
- 2. Quantify gene expression.

High-throughput gene expression quantification principles / technologies:

- Sequencing - SAGE
- Nucleotide complementarity – Microarrays
- Generalizations.

References.

Next lecture: Survey comp methods for modeling high-throughput transcriptome data, an application in cancer biology.
Characterizing a biological system

- Organizational scales of biological systems - multi-scalar
  - Molecular, chemical: Gene, protein, other bio-molecules.
  - A cell: autonomous or non-autonomous
  - A multi-cellular tissue/organ or organism
  - Ecosystem: community of organisms in an environment.
  - Add time as parameter: A biological system-state.

- Constituent features of a biological system-state
  - Objects/Attributes:
    - Micro scale: Genotype, genetic (gene, protein) epigenetic (non-DNA), etc.
    - Macro scale: Phenotype, histo/morphology, behavior, etc.

- Environment

- Characteristics of a biological system-state
  - Constituent features are multi-scalar in space/time: Granularities of objects/attributes.
  - Connections and interactions within/across organizational scales.
    - Combinatorial complexity.
Characterizing a biological system

- **Prototypical questions for a biological system-state**
  - Identify objects / attributes / environmental factors that characterize a biological system state.
  - Characterize the interaction of objects / attributes / environment factors that transform the state of a biological system.
  - Distinguish Causative (Generative / Mechanistic) from Correlative (Descriptive) interactions.

- **Disease = a gene environment interaction that moves the biological system out of reference “normal” state(s).**
Characterizing a biological system

- **Basic biology question**: Identify features and their connections/interactions across/within organizational scales of a biological system state.

**Microscopic** molecular scale objects / attributes

- Genotype
- Protein networks
- Transcriptome
- Epigenetics

**Macroscopic** large scale, phenomenological observations

- Phenotype
  - Cellular organization
  - Morphology
  - Behavior

**TASK**: Characterize object / attribute / environment interaction

Environment
Central dogma of molecular biology (CD)

- Original CD statement (Crick, Nature 1958)
  - ... detailed residue-by-residue transfer of sequential information ... such information cannot be transferred from protein to either protein or nucleic acids.

- Over-simplified CD: DNA -> RNA -> Protein
  - DNA: C, G, A, T double strand
  - RNA: C, G, A, U single strand

- Reductionist assumption
  - A biological event is completely characterized by this information transfer. Question: which genes, RNA, proteins exactly?

From Kohane et al, MITpress, 2003
Exceptions to over-simplified CD

- **Retroviruses**
  - RNA into DNA via reverse transcriptase: E.g., avian sarcoma/leukosis viruses, mouse leukemia viruses, human immunodeficiency virus (HIV)
  - RNA (virus) --> DNA (host) --> RNA (virus) --> Protein (virus)

- **Primitive RNA viruses**
  - Error-prone RNA replication. E.g., hepatitis B, rabies, Dengue, Ebola, flu
  - Genetic RNA --> Intermediate RNA --> Protein

- **Prions**
  - Self-replicating proteins. E.g., Creuzfeldt-Jakob, “mad cow”, kuru
  - Protein --> Protein

- **DNA-modifying proteins (not technically)**
  - DNA-repair proteins: MCM family

- **Retrotransposons (not technically)**
  - Mobile DNA (specifically *genetic*) segments in eukarya. Esp. plants, >90% wheat genome.
  - Retrotransposon DNA --> RNA --> DNA
DNA is a molecule
but what is a gene ?

1. “Gene” is an agent, abstraction.

2. DNA is one of many classes of molecules in a cell.

3. Why DNA is the class of molecules that transmit biological information, phenotype ?
Concept of a gene

- **Chronology:**
  - 1854-65 “Unit factor” of inheritance. Gregor Mendel (Brno), *On the Origin of Species*, 1859
  - 1869 Nucleic acids isolated. Johann Miescher (Tübingen)
  - 1952 DNA (not protein) is genetic material. Hershey & Chase (NY)
  - 1953 DNA as double helix, structurally sensible. Watson, Crick & Franklin (Cambridge, UK),

  *This structure has 2 helical chains each coiled around the same axis ... it has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.*

- **Definition of a gene**
  - A fundamental physical and functional unit of heredity that is a DNA sequence located on a specific site on a chromosome which encodes a specific functional product (RNA, protein). (From NCBI)
  - A basic and complete unit of genetic information
  - Question: Let $X = \text{DNA sequence } k\text{-base pairs long } (k, \text{integer}).$ Is gene definition sufficient to decide if $X$ is/contains gene segment? No. Gene-finding algorithms ... large false positive rate
Concept of a gene

- Definition of genome / genotype
  - **Genome** - the total genetic material (DNA) in a living organism. **Genotype** – total genetic information in a living organism. (From NCBI)

- Content of genome (eukaryote)
  - Genes (~1.5% genome), gene-related DNA (~36% genome), intergenic DNA (~62.5% genome)
    - Exons (coding), introns (non-coding, eukarya). Coding = transmission into mRNA.
    - Pseudogene
    - Etc: microsatellites, genome-wide repeats. E.g., transposons, long/short interspersed nuclear elements

- Eukaryote vs. Prokaryote genomes structural differences

- Genome size not correlated with organism complexity: Corn (2x10^9 bp) vs human (3x10^9 bp). Arabidopsis (water-cress, plant) 1x10^8 bp vs Psilotum (whisk fern, primitive vascular plant – no leave, root, flower) 3x10^11 bp

Gene-related DNA = 1.15 Gbp
Intergenic DNA = 2.0 Gbp

Human genome = 3.2 Gbp
Concept of a gene

- Gene coding regions (exons) not contiguous along chromosomal stretch. Non-coding introns in between.

Figure 1.14. A segment of the human genome showing the location of genes, gene segments, pseudogenes, genome-wide repeats and microsatellites in a 50-kb segment of the human beta T-cell receptor locus on chromosome 7. Redrawn from Rowen et al. (1996). From T.A. Brown, Genomes 2

Intergenic DNA = Junk? Probably not. Muotri et al. (Nature 2005 16 June issue). L1 retrotransposon gene-hopping --> neuronal cell fate for rat neural stem cells. Small molecule RNA e.g. microRNA.
Concept of gene expression

- Definition of **gene expression** / transcription:
  - *The process by which information encoded in a gene is transcribed into RNA, and then typically into protein.* (From NCBI)

- Gene expression is a function of intrinsic & extrinsic factors:
  - Time and developmental stage
  - Space or location. E.g., brain vs. muscle
  - Response to autocrine, micro-, global environmental cues
  - Disease / cell state
Concept of gene expression

- **Definition of transcriptome**
  - Total mRNA present in a biological system (e.g., cell) at a particular time and state.
  - Environment dependent!!

**Alternative Splicing**

3 mRNA isoforms of one gene

Different isoforms --> different protein products, i.e., different function.
Key assumption and goal in genomics

- **Assumption** (reductionist): A biological system is completely characterized by its transcriptome state (i.e., total mRNA profile)
  - Depends on Central Dogma.
  - Recall “Gene expression is a function of ...”, 2 slides ago.

- **Goal**: Use transcriptome data to unravel the environment + genotype interactions that engender phenotype
  - Limitations? Reductionistic. Might proteins provide more accurate picture?
  - Gene-environment interaction → phenotypic transformation
Measuring gene expression

- mRNA level as proxy for gene expression

- 2 Steps:
  - 1. Identify molecule.
  - 2. Quantify molecule
    - (Identify) DNA libraries: Genomic, cDNA.
    - (Quantify) Low throughput: northern blot, RT-PCR.
    - (Quantify) High throughput: SAGE, microarrays. Our focus.
Step 1. DNA libraries (Identification)

- **DNA libraries** = collections of clones DNA fragments: *Genomic* & *cDNA* via sequencing. **Assembling fragments into complete sequence is bioinformatics challenge:** 1. coverage (redundancy), 2. minimal sequence length

- **Genomic libraries**
  - Overlapping genomic DNA fragments representing ~**entire genome** of an organism.

- **Complementary DNA libraries** (genomic material that is “expressed”, genes)
  - All mRNA species in a particular biological system/state. Contains only **coding regions** of genome (minus introns), proxy for expressed proteins.
  - mRNA $\rightarrow$ cDNA (chemically more stable) $\rightarrow$ **Expressed Sequence Tags**. Each EST gets a Genbank ID.
  - A gene may be “covered” by >1 EST's. Redundancy. Human genome $\rightarrow$ >4 million EST's. Estimate # of genes in human genome $\sim$30K.
  - EST's screened. All EST's associated with a particular gene forms an EST cluster for that gene. This cluster is assigned a Unigene ID.

```
Biological system at specific state
(organism, region, time)

Biological material with mRNA

mRNA $\rightarrow$ cDNA
```
Step 1. cDNA libraries (Identification)

- Example of a Unigene cluster of >1 ESTs. Human FoxP2 (Chr.7q31.1, 3 isoforms) gene has 51 EST sequences in it's Unigene cluster (Hs.656280).

<table>
<thead>
<tr>
<th>Genbank ID</th>
<th>Description</th>
<th>Tissue of Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>BF700673.1</td>
<td>Clone IMAGE:4285527, 5' read</td>
<td>brain</td>
</tr>
<tr>
<td>T97069.1</td>
<td>Clone IMAGE:121181, 5' read</td>
<td>mixed</td>
</tr>
<tr>
<td>T96957.1</td>
<td>Clone IMAGE:121181, 3' read</td>
<td>mixed</td>
</tr>
<tr>
<td>BU521502.1</td>
<td>Clone IMAGE:6527367, 5' read</td>
<td>uterus</td>
</tr>
<tr>
<td>BQ948273.1</td>
<td>Clone IMAGE:6473507, 5' read</td>
<td>uterus</td>
</tr>
<tr>
<td>AL711700.1</td>
<td>Clone DKFZp686E0284, 5' read</td>
<td>muscle</td>
</tr>
<tr>
<td>BM725479.1</td>
<td>Clone UI-E-Ej0-iae-p-18-0-UI, 5' read</td>
<td>other</td>
</tr>
<tr>
<td>BM701645.1</td>
<td>Clone UI-E-Ej0-ahh-24-0-UI, 5' read</td>
<td>other</td>
</tr>
<tr>
<td>BI752226.1</td>
<td>Clone IMAGE:5192788, 5' read</td>
<td>brain</td>
</tr>
<tr>
<td>N31133.1</td>
<td>Clone IMAGE:265380, 5' read</td>
<td>skin</td>
</tr>
<tr>
<td>N21118.1</td>
<td>Clone IMAGE:265380, 3' read</td>
<td>skin</td>
</tr>
<tr>
<td>DN990126.1</td>
<td>Clone TC100653, 5' read</td>
<td>Whole brain</td>
</tr>
<tr>
<td>AV658847.1</td>
<td>Clone GLCFQG08, 3' read</td>
<td>liver</td>
</tr>
<tr>
<td>AV658824.1</td>
<td>Clone GLCFQEO9, 3' read</td>
<td>liver</td>
</tr>
<tr>
<td>CV573230.1</td>
<td>Clone od33g10, 5' read</td>
<td>eye</td>
</tr>
<tr>
<td>CR738014.1</td>
<td>Clone IMAGp998F084735_:IMAGE:1929991, 5' read</td>
<td>lung</td>
</tr>
<tr>
<td>BP871788.1</td>
<td>Clone HKR01979</td>
<td>embryonal kidney</td>
</tr>
<tr>
<td>BE068078.1</td>
<td>Clone (no-name)</td>
<td>mammary gland</td>
</tr>
<tr>
<td>CD637513.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>CD637512.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>CD637511.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>CD637510.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>CD637509.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>CD637508.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>BX481950.1</td>
<td>Clone DKFZp686D03228, 5' read</td>
<td>muscle</td>
</tr>
<tr>
<td>CD001942.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>CB410738.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>CB410682.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>CB410681.1</td>
<td>Clone (no-name)</td>
<td>other</td>
</tr>
<tr>
<td>BX280996.1</td>
<td>Clone IMAGp998G13581_:IMAGE:265380</td>
<td>skin</td>
</tr>
<tr>
<td>CB111825.1</td>
<td>Clone B1T694954-5-A03, 5' read</td>
<td>brain</td>
</tr>
<tr>
<td>T06261.1</td>
<td>Clone HFBDR02</td>
<td>brain</td>
</tr>
<tr>
<td>AI459612.1</td>
<td>Clone IMAGE:2152081, 3' read</td>
<td>colon</td>
</tr>
<tr>
<td>AI624789.1</td>
<td>Clone IMAGE:2231455, 5' read</td>
<td>uterus</td>
</tr>
<tr>
<td>AI798932.1</td>
<td>Clone IMAGE:2348762, 3' read</td>
<td>mixed</td>
</tr>
<tr>
<td>CK430225.1</td>
<td>Clone qj46f12, 5' read</td>
<td>eye</td>
</tr>
<tr>
<td>CV569620.1</td>
<td>Clone od07e09, 5' read</td>
<td>eye</td>
</tr>
<tr>
<td>BF678535.1</td>
<td>Clone IMAGE:4250207, 5' read</td>
<td>prostate</td>
</tr>
<tr>
<td>BG722650.1</td>
<td>Clone IMAGE:4826916, 5' read</td>
<td>testis</td>
</tr>
<tr>
<td>BI495413.1</td>
<td>Clone IMAGE:2539657, 3' read</td>
<td>other</td>
</tr>
</tbody>
</table>

Step 1. cDNA libraries (Identification)

- Unigene cluster sizes
  - Not every Unigene cluster is mapped to a known gene. Different similarity levels, sequence matching.
  - Estimated # of genes in human genome ~30K
  - Human Unigene stats

<table>
<thead>
<tr>
<th>Cluster Size</th>
<th># of Unigene Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>~10,400</td>
</tr>
<tr>
<td>2</td>
<td>7,100</td>
</tr>
<tr>
<td>3-4</td>
<td>6,800</td>
</tr>
<tr>
<td>5-8</td>
<td>5,300</td>
</tr>
<tr>
<td>9-16</td>
<td>3,800</td>
</tr>
<tr>
<td>17-32</td>
<td>3,100</td>
</tr>
<tr>
<td>~500-1,000</td>
<td>1,500</td>
</tr>
<tr>
<td>~2,000-4,000</td>
<td>130</td>
</tr>
<tr>
<td>~8,000-16,000</td>
<td>12</td>
</tr>
<tr>
<td>~16,000-30,000</td>
<td>3</td>
</tr>
</tbody>
</table>

From J. Pevsner, Bioinformatics & Functional Genomics, 2005
Step 2. Comparing between cDNA libraries

- **Question:** Is gene X expressed in a biological system?

- **Binary** (present / absent) comparison between cDNA libraries derived from various tissue systems

  - Digital Differential Display (DDD). Statistical significance (p value) is assessed via Fisher exact test (non parametric). Contingency table. Null hypothesis: # of sequences for a given gene X is the same in the two cDNA libraries.
  
  - Not quantitative

  - Limitations: (1) Sequencing error (2) Depth of sequencing (3) tissue of origin contamination (4) library construction bias.

<table>
<thead>
<tr>
<th></th>
<th>Gene X</th>
<th>Genes other than Gene X</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>cDNA library A</strong></td>
<td># EST's mapped to Gene X</td>
<td># EST's not mapped to Gene X</td>
</tr>
<tr>
<td><strong>cDNA library B</strong></td>
<td># EST's mapped to Gene X</td>
<td># EST's not mapped to Gene X</td>
</tr>
</tbody>
</table>

Step 2. SAGE (Quantification)

- **Serial Analysis of Gene Expression**
  - Need SAGE library: a 1-1 map between SAGE tags and genes & EST's
  - Get mRNA to construct cDNA.
  - From each cDNA transcript, cut a short sequence tag (SAGE tag) of 10-14 bps from a **specific** position (3'-end typically) that uniquely identifies the transcript.
  - Tags have uniform length.
  - Concatenate all SAGE tags into one concatamer -> clone and sequence.
  - # of times a specific tag observed = expression level of specific gene/mRNA

Details:

From www.sagenet.com, Note cartoon hides a lot of vital details!
Step 2. SAGE (Quantification)

Example of SAGE result: 3 types of transcripts relative to SAGE library

<table>
<thead>
<tr>
<th>Tag_Sequence</th>
<th>Count</th>
<th>Gene Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATATTGTCAA</td>
<td>5</td>
<td>translation elongation factor 1 gamma</td>
</tr>
<tr>
<td>AAATCGGAAT</td>
<td>2</td>
<td>T-complex protein 1, z-subunit</td>
</tr>
<tr>
<td>ACGCCTTCG</td>
<td>1</td>
<td>no match</td>
</tr>
<tr>
<td>GCCCTTGGTA</td>
<td>81</td>
<td>rpa1 mRNA fragment for r ribosomal protein</td>
</tr>
<tr>
<td>GTTAACCATC</td>
<td>45</td>
<td>ubiquitin 52-AA extension protein</td>
</tr>
<tr>
<td>CCGCGGTGGG</td>
<td>9</td>
<td>SF1 protein (SF1 gene)</td>
</tr>
<tr>
<td>TTTTTGGTTAA</td>
<td>99</td>
<td>NADH dehydrogenase 3 (ND3) gene</td>
</tr>
<tr>
<td>GCAAACCGGG</td>
<td>63</td>
<td>rpL21</td>
</tr>
<tr>
<td>GGAGCCCGCC</td>
<td>45</td>
<td>ribosomal protein L18a</td>
</tr>
<tr>
<td>GCCGCAACA</td>
<td>34</td>
<td>ribosomal protein S31</td>
</tr>
<tr>
<td>GCCGAAGTTG</td>
<td>50</td>
<td>ribosomal protein S5 homolog (M(1)15D)</td>
</tr>
<tr>
<td>GTGGCTTACA</td>
<td>4</td>
<td>EST BB205679</td>
</tr>
</tbody>
</table>

From www.embl-heidelberg.de/info/sage
Step 2. SAGE (Quantification)

- SAGE limitations:
  - Tag specificity. Short SAGE tag size may lead to identification problems. 1 tag mapping to >1 genes is a problem.
  - Restriction enzyme action variability. Each SAGE tag must have constant length, otherwise problems arise in sequencing concatamer. Restriction enzyme may not yield tags of uniform length. Not all mRNA species have the same enzyme recognition sequence. Temperature dependency of enzyme.
  - What is appropriate control/reference system for comparison? Really a more general problem that we will revisit as we cover microarrays and other high-throughput assaying technologies.
Step 2. DNA microarrays (Quantification)

- What is a DNA microarray (chip)?
  - A collection of single-stranded DNA (known sequences of genes / EST's) anchored at one end onto a substrate, typically in the form of a gridded array. Different DNA species placed on separate grids.
  - ssDNA fragments (called probes), not entire gene sequence is placed. Why?
  - Inspired by PCR & southern blots for DNA. Exploits parallelism.
  - Mechanistic principle: Nucleic acid complementarity – i.e., hybridization of complement partners, A ↔ T, G ↔ C.
  - ssDNA on chip hybridizes to complementary strand in solution (derived from biological system). Complementary strand is fluorescent labeled.
  - Assumption: Fluorescence is (linearly) proportional to gene expression level.

- Microarray oligo probe design technicalities
  - GC content: Hybridization (binding) energy for GC > AT. Introduces non-linearity in hybridization rate for cDNA species with different %CG content. General problem.
  - Distance of ssDNA design sequence from 3' end. General problem.
Step 2. DNA microarrays (Quantification)

- Primarily 2 types of DNA microarrays
  - **Spotted**: (Pat Brown, Stanford). Robot attaches down previously prepared ssDNA probes of order $10^{2-3}$ bp long on substrate. Customizable --> heterogeneous (noisy)
  - **Oligonucleotide**: (e.g., Affymetrix). Photolithography. Typically standardized manufacturing and shorter (relative to spotted microarrays) length oligos

From Kohane et al. MITPress, 2003
From Southern et al. Nature Genetics 1999
Step 2. DNA microarrays (Quantification)

- Stages of a typical microarray experiment
  - Experimental design involving biological system under investigation. Replicates – biological and measurement / technical
  - RNA and (target) probe preparation: Extract mRNA. Convert to ss cDNA typically. Label with fluorescence.
  - Probe hybridization.
  - Fluorescence image analysis
  - Microarray data analysis (post image)

From J. Pevsner, Pg 181 Bioinformatics & Functional Genomics, 2005
Step 2. DNA microarrays (Quantification)

- 1 channel vs. 2 channel microarray usage
  - 2 channel
  - Internal control for fluorescence

Diagram:
- Control
  1. 2 channel
  2. Internal control for fluorescence
- Test
  3. 2 channel
  4. Competitive, 2-channel
Step 2. DNA microarrays (Quantification)

- **1 channel** vs. 2 channel microarray usage
  - 1 channel
  - Internal control?

1-channel

1. Biotin Labeled cRNA
2. Oligonucleotide microarray
3. Hybridized Array

SAPE
Streptavidin-phycoerythrin
Step 2. DNA microarrays (Quantification)

- Canonical usage and experimental designs
  - Comparing two groups. E.g., tumor/cancer vs. “normal” tissue
  - Time course (dosage-level) profiling
  - Suitable reference state is a general problem
  - Typical “statistically-sound” experiment design principles apply. Sample pooling mRNA? For 2-channel experiments: Dye swap.

- Microarray experiment (biological) assumptions
  - Central Dogma holds. Specifically that mRNA transcription is proportional to its associated protein translation
  - All mRNA have ~identical lifespans. ~Uniform degradation rate.
  - Principal cellular activities are ~entirely characterized by the transcriptome.
Step 2. DNA microarrays (Quantification)

- Generalizing chip parallelism/complementarity principle to measuring other principal biological components/characteristics
  - Protein microarrays. Identify protein targets, e.g., other proteins (protein-protein interaction), mRNA, other bio-active small molecules.
  - Tissue microarrays. Paraffin blocks of distinct biological tissue cores. Simultaneous histologic analysis, immunohistochemical (protein) & in situ (mRNA) analyses.
  - Reverse transfection microarrays. cDNA probes on grid with a cell culture on top. Cells assimilate probes.

- Limitations:
  - Probe specificity. Cross (RNA) species hybridization, promiscuous probes.
  - Non-uniform RNA degradation.
  - “Noise”. Next time.

- Future: Next generation sequencing. High efficiency, low cost.
Characterizing a biological system.

- Organizational scales and constituents of a biological system
- **Central Dogma of molecular biology (CD):**
  - Centrality of genes in investigations of biological mechanisms
- From DNA to gene.
- The concept of a gene and its expression.

2 central objectives in quantitative genomics

1. Identify genes
2. Quantify gene expression.

High-throughput gene expression quantification principles / technologies:

- Sequencing - SAGE
- Nucleotide complementarity – Microarrays
- Generalizations.

References.

Next lecture: Survey comp methods for modeling high-throughput transcriptome data, an application in cancer biology.
References

- **Reader-Friendly References**


  - NCBI has a many user-friendly primers on genomic biology: www.ncbi.nlm.nih.gov/books/NBK2345/. Coffee Break section, and library of searchable free books / tutorials
