Beyond Genomics:
Detecting Codes and Signals in the Cellular Transcriptome

Brendan J. Frey
University of Toronto
Purpose of my talk

To identify aspects of bioinformatics in which attendees of ISIT may be able to make significant contributions
Beyond Genomics:
Detecting Codes and Signals in the Cellular Transcriptome

Brendan J. Frey
University of Toronto
The Genome
Starting point: Discrete biological sequences

• Symbols are **Bases:** G, C, A, T

• Examples of biological sequences
  – Genes
  – DNA
  – Chromosomes
  – Proteins
  – Peptides
  – RNA
  – Viruses
  – HIV
Chromosomes: Inherited DNA sequence

DNA Sequence (GCATTTCATGC…)

Cell replication

Sexual cell reproduction

Nucleus
The genome

- **Genome**: Chromosomal DNA sequence from an organism or species

- Examples

<table>
<thead>
<tr>
<th>Genome</th>
<th>Length (bases)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>3,000 million (750MB)</td>
</tr>
<tr>
<td>Mouse</td>
<td>2,600 million</td>
</tr>
<tr>
<td>Fly</td>
<td>100 million</td>
</tr>
<tr>
<td>Yeast</td>
<td>13 million</td>
</tr>
</tbody>
</table>
Genes

- A gene is a subsequence of the genome that encodes a functioning bio-molecule

- The library of known genes
  - Comprises only 1% of genome sequence
  - Increases in diversity every year
  - Is probably far from complete
The Transcriptome
Genome: The digital backbone of molecular biology

**Transcripts**: Perform functions encoded in the genome
Traditional genes

DNA \rightarrow \text{Transcription} \rightarrow \text{Transcript (RNA)} \rightarrow \text{Translation} \rightarrow \text{Protein}

Input: DNA \rightarrow \text{Output: Transcript}

Input: Transcript \rightarrow \text{Output: Protein}
Transcription

Gene

Upstream region

Exon

Intron

DNA

Regulatory proteins

Transcript (RNA)

Transcription proteins
Transcription

- Codewords in the **upstream region** bind to corresponding **regulatory proteins**

DNA → CGTGGATAGTAGTGAT → Exon

- **Code**: Set of regulatory codewords
- **Signals**: Concentrations of regulatory proteins and the output transcript
Splicing of transcripts

Transcript (RNA)

Exon

Intron

Regulatory proteins
Splicing of transcripts

Transcript (RNA)

Exon

Intron

Regulatory proteins

Splicing proteins
Splicing of transcripts

- The intron is spliced out
- However, splicing may occur quite differently…
Splicing of transcripts

Transcript (RNA)

Exon

Intron

Regulatory proteins

Splicing proteins
Splicing of transcripts

Regulatory proteins

Splicing proteins
Splicing of transcripts

The middle exon is ‘skipped’, leading to a different transcript
Splicing of transcripts

- Codewords in the introns and exons bind to corresponding regulatory proteins
- Code: Set of regulatory codewords
- Signals: Concentrations of regulatory proteins and different spliced transcripts

Brendan Frey
The modern transcriptome

Cell nucleus

Genome

Non-functional transcripts

TRANSCRIPTION

Liver

Transcript (RNA)

SPLICING

Brain and Liver

Transcript (mRNA)

Non-traditional transcript

Protein A

Protein B

TRANSLATION

Protein

Brendan Frey
The modern **transcriptome**

... it turns out to be surprising in many ways

# genes, ½ trans, 60% AS, 18k AS, 20% dis, 10k ncRNA
The Resources
Your collaborators can do lab work…

- **Sequencing**: Snag an actual transcript and figure out its sequence
- **Microarrays**: Find out if your predicted transcript fragment is expressed in a tissue sample
- **Mass spectrometry**: Find out if a protein is present in a sample
Databases

• Genomes
• Genome annotations
• Libraries of observed transcript fragments
• Microarray datasets containing measured concentrations of transcripts
• ...
Measuring transcript concentrations using microarrays

1. Fabricate microarray with probes
2. Extract transcripts from cell
3. Add florescent tag
4. Hybridize tagged sequence to microarray
5. Excite florescent tag with laser and measure intensity
Inkjet printer technology
Hughes et al, Nature Biotech 2001

Print nucleic acid sequences using inkjet printer
Then and now…

• First microarrays (late 1990s)
  – ‘Cancer chips’, ‘gene chips’, …
  – 5,000-10,000 probes per slide
  – Noisy

• Current microarrays
  – ‘Sub-gene resolution’
  – 200,000 probes per slide
  – Low noise
  – Multi-chip designs are cost effective
The Case Study:
Discovering protein-making transcripts using factor graphs

BJ Frey, ..., TR Hughes
Nature Genetics, September 2005
Controversy about the gene library

Despite Frey et al’s impressive computational reconstruction of gene structure, we argue that this does not prove the complexity of the transcriptome

- FANTOM/RIKEN Consortium
  Science, March 2006

How it all started…
Research on the transcriptome

- Analysis of genome
  - 2001-2005

- Detection of transcripts
  - 1960’s-2000
  - 2001-2006

Our project
- 2003-2005

Brendan Frey
Estimates of number of undiscovered genes

Genome: ~10,000
(IHGSC, Nature)

Genome: ~3000
(IHGSC, Nature)

Kapranov et al, Rinn et al, Shoemaker et al: ~300,000

Bertone et al: ~11,000
(Science)
Our microarrays

- Our genome analysis highlighted 1 million possible exons (~180,000 already known)
- We designed one 60-base probe for each possible exon
Our samples (37 tissues)

Twelve pools of mouse mRNA

<table>
<thead>
<tr>
<th>Pool</th>
<th>Composition (mRNA per array hybridization)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Heart (2 μg), Skeletal muscle (2 μg)</td>
</tr>
<tr>
<td>2</td>
<td>Liver (2 μg)</td>
</tr>
<tr>
<td>3</td>
<td>Whole brain (1.5 μg), Cerebellum (0.48 μg), Olfactory bulb (0.15 μg)</td>
</tr>
<tr>
<td>4</td>
<td>Colon (0.96 μg), Intestine (1.04 μg)</td>
</tr>
<tr>
<td>5</td>
<td>Testis (3 μg), Epididymis (0.4 μg)</td>
</tr>
<tr>
<td>6</td>
<td>Femur (0.9 μg), Knee (0.4 μg), Calvaria (0.06 μg), Teeth+mandible (1.3 μg), Teeth (0.4 μg)</td>
</tr>
<tr>
<td>7</td>
<td>15d Embryo (1.3 μg), 12.5d Embryo (12.5 μg), 9.5d Embryo (0.3 μg), 14.5d Embryo head (0.25 μg), ES cells (0.24 μg)</td>
</tr>
<tr>
<td>8</td>
<td>Digit (1.3 μg), Tongue (0.6 μg), Trachea (0.15 μg)</td>
</tr>
<tr>
<td>9</td>
<td>Pancreas (1 μg), Mammary gland (0.9 μg), Adrenal gland (0.25 μg), Prostate gland (0.25 μg)</td>
</tr>
<tr>
<td>10</td>
<td>Salivary gland (1.26 μg), Lymph node (0.74 μg)</td>
</tr>
<tr>
<td>11</td>
<td>12.5d Placenta (1.15 μg), 9.5d Placenta (0.5 μg), 15d Placenta (0.35 μg)</td>
</tr>
<tr>
<td>12</td>
<td>Lung (1 μg), Kidney (1 μg), Adipose (1 μg), Bladder (0.05 μg)</td>
</tr>
</tbody>
</table>
Signal: The data
(small part of the data from Chromosome 4)

Each column is an expression profile

Example of a transcript

Code:
A ‘vector repetition code with deletions’
The transcript model

Each transcript is modeled using

A prototype expression profile

# probes before prototype (e.g., 1)
# probes after prototype (e.g., 4)

Flag indicating whether each probe corresponds to an exon

Brendan Frey
The prototype for $x_i$ is $x_{i+r_i}$, $r_i \in \{-W, \ldots, W\}$. We use $W=100$.

**ONLY 1 FREE PARAMETER:**

$\kappa$, probability of starting a transcript.

Brendan Frey
After expression data \((x)\) is observed, the factor graph becomes a tree.

Transcription start/stop indicator

Relative index of prototype

Exon versus non-exon indicator

Probe sensitivity & noise

Expression profile (genomic order)
After expression data \((x)\) is observed, the factor graph becomes a tree

Computation: The max-product algorithm performs \textit{exact inference and learning}.
Summary of results *

• 10 X more sensitive than other transcript-based methods
• Detected 155,839 exons
• Predicted ~30,000 new exons
• Reconciled discrepancies in thousands of known transcripts

* Exon false positive rate: 2.7%
Revisiting Estimates of number of undiscovered genes

- **2000**: Genome: ~10,000 (IHGSC, Nature)
- **2001**: Kapranov et al, Rinn et al, Shoemaker et al: ~300,000
- **2002**: Genome: ~3000 (IHGSC, Nature)
- **2003**: Bertone et al: ~11,000 (Science)
- **2004**: Frey et al: ~0 (Nature Genetics)
- **2005**: SURPRISE!

Brendan Frey
Contentious results

Bertone et al: ~11,000
(Science)

Frey et al: ~0
(Nature Genetics)

FANTOM3: 5,154
(FANTOM Consortium, Science)

Brendan Frey
... [We discovered] new mouse protein-coding transcripts, including 5,154 encoding previously-unidentified proteins …

- FANTOM/RIKEN Consortium
  Science, Sep 2005

We wondered: Are these really new genes?
… we found that **2917 of the FANTOM proteins** are in fact splice isoforms of **known transcripts** …

- Frey et al
Science, March 2006

… the **number of new protein-coding genes** found by us has been revised from **5154 to 2222** …

- FANTOM/RIKEN Consortium
Science, March 2006
Last word…

…the number of completely new protein-coding genes discovered by the FANTOM consortium is at most in the hundreds…

- Frey et al
  Science, March 2006
The Closing Remarks
Open problems

• Producing genome-wide libraries of functioning transcripts, including
  – Alternatively-spliced transcripts
  – Transcripts that don’t make proteins
• Understanding functions of transcripts
• Developing models of how transcription and alternative splicing are regulated
• Developing models of gene interactions
  – ‘Genetic networks’
Should you work in computational biology?

Pluses
• A major scientific frontier
• Potential for high impact on society

Minuses
• Mostly a collection of facts
• Mechanisms are complex and beyond our control
• Lacking a mathematical framework

Brendan Frey
Remember, communication theory also once lacked a mathematical framework…

“Ok, Zorg, lets try using a prefix code”
Should **you** work in computational biology?

**Pluses**
- A major scientific frontier
- Potential for high impact on society
- Lacking a mathematical framework

**Minuses**
- Mostly a collection of facts
- Mechanisms are complex and beyond our control
How do you enter this field?

- Hire a tutor (ie, student or postdoc)
- Hire a programmer
- Get involved in several ‘winner’ projects
- Be prepared to drop ‘loser’ projects
- Build mutually-beneficial collaborations
- How long will it take?
For more information…

- As of Friday July 14, 2006:
  
  http://www.psi.toronto.edu/isit2006.html
  
  – These slides
  
  – Pointers to helpful papers, databases, etc
Acknowledgements

**Frey Group**
- Quaid D Morris (postdoc)
- Leo Lee (postdoc)
- Yoseph Barash (postdoc)
- Ofer Shai (PhD)
- Inmar Givoni (PhD)
- Jim Huang (PhD)
- Marc Robinson (programmer)

**Genomics Collaborators**
- Hughes’ Lab
- Blencowe’s Lab
- Emili’s Lab
- Boone’s Lab

**Medical Collaborators:**
E Sat, J Rossant, BG
Bruneau, JE Aubin