Algorithms for finding sequence motifs in HT-Selex data

Introduction

Ron Shamir

Lab instructor: Yaron Orenstein

October 2013
Outline

0. Workshop goals and schedule
1. Basic biology (+ some stories)
2. Gene regulation and motif finding
3. PBM and HT-SELEX
0. Workshop goals and schedule
Motivation

• Biological processes are regulated by genes (and other molecules)
• Understanding how this regulation works is a holy grail of biomedical research
• Many experimental and technological developments aim to achieve this understanding
• Our goal: analyze the data produced by the two newest technologies (HT-SELEX and PBM).
The workshop in a nutshell

1. Input 1: HTS experiment data

2. Develop a method to build a model from the data

3. Input 2: Test data: PBM experiment

4. Output: prediction. Use the model to predict sequence ranks

5. Evaluate the prediction quality vs. the hidden signal

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>CATGTAAGAGTTGACTCTGGTCTGTTCTAAT</td>
<td>28926</td>
</tr>
<tr>
<td>TTGCTCATCAGAGTCGCGTAACAGGCTTTC</td>
<td>1457</td>
</tr>
<tr>
<td>TCCAGTTTAGGGCCGCCCAGCCTTAA</td>
<td>12972</td>
</tr>
<tr>
<td></td>
<td>(hidden)</td>
</tr>
<tr>
<td>CATGTAGCCCTTAACTGTGACTAAAGCC</td>
<td>33755</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>CATGTAAGAGTTGACTCTGGTCTGTTCTAAT</td>
<td>11</td>
</tr>
<tr>
<td>TTGCTCATCAGAGTCGCGTAACAGGCTTTC</td>
<td>5</td>
</tr>
<tr>
<td>TCCAGTTTAGGGCCGCCCAGCCTTAA</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>(hidden)</td>
</tr>
<tr>
<td>CATGTAGCCCTTAACTGTGACTAAAGCC</td>
<td>4</td>
</tr>
</tbody>
</table>
Administrata

• Project should be written in Java/Linux
• Project can be done in pairs
• Pairs/singletons – please inform Yaron by next Monday who is on your team.
• Meetings – introductory lectures on week 1,2. Then individual meetings with groups (on the workshop time slot)
• Questions? Contact Yaron or me
• Website:
  www.cs.tau.ac.il/~rshamir/workshop/13
Grading

- 15% for the design
- 25% for the implementation (10% for modularity, clarity, documentation, \(f(r,k) \times 15\%\) for efficiency)
- 20% for the final report and presentation
- \(f(r,k) \times 50\%\) for the accuracy of the test results
  - \(f(r,k) \times 15\%\) for test 1
  - \(f(r,k) \times 20\%\) for test 2
  - \(f(r,k) \times 15\%\) for test 3
- Where
  - \(r = \text{group's rank in test out of } k \text{ groups (top rank } r=k\)}
  - \(f(r,k) = 0.5 + 0.5 \times r/k\)
- So a uniformly top ranking group can get 110, and uniformly least ranking can still get 82.
- Ties will be scored...
Schedule

1. 19/11 Individual meetings and first progress report
2. 10/12 Submission of Test 1 results
3. 24/12 submission of Design document
4. 14/1 submission of Test 2 + executable
5. 18/2 Class meeting and final presentation

You are always welcome to meet us. Contact us by email.
1. Basic Biology

Slides with Adi Akavia
Nucleotides Chain

Double helix, 2-stranded helix

Gregor Mendel
laws of inheritance, “gene”
1866

Watson and Crick
DNA Discovery
1953

Nucleotides/ Bases:
Adenine (A),
Guanine (G),
Cytosine (C),
Thymine (T).

phosphate
sugar
Backbone

Weak hydrogen bonds between base pairs
DNA (Deoxy-Ribonucleic acid)

- DNA is located in the cell nucleus
- **Bases:**
  - Adenine (A)
  - Guanine (G)
  - Cytosine (C)
  - Thymine (T)
- **Bonds:**
  - G - C
  - A - T
- Length of human DNA $\sim 3 \times 10^9$ bp (= base pairs)
DNA and Chromosomes

• DNA: 4 bases
  molecule: ACGT

• Complementary strands: A-T; C-G

• Allows duplication

• Chromosome: contiguous stretch of DNA

• Genome: totality of DNA material
Genes

- **Gene**: a segment of DNA that specifies the sequence of a protein.
- Contains one or more regulatory sequences that either increase or decrease the rate of its transcription
- **Genes are 2-3% of human DNA**
- the rest - non-coding “junk DNA”

**Red**: a region that encodes a protein sequence
**Black**: a non-coding region (a single gene usually contains more than one)
**Green**: a regulatory sequence
DNA → RNA → protein

The hard disk

One program

transcription

Its output

translation
From Gene to Protein

Translating DNA into Protein

When genes are expressed, the genetic information (base sequence) on DNA is first transcribed (copied) to a molecule of messenger RNA in a process similar to DNA replication. The mRNA molecules then leave the cell nucleus and enter the cytoplasm, where triplets of bases (codons) forming the genetic code specify the particular amino acids that make up an individual protein. This process, called translation, is accomplished by ribosomes (cellular components composed of proteins and another class of RNA) that read the genetic code from the mRNA, and transfer RNAs (tRNAs) that transport amino acids to the ribosomes for attachment to the growing protein.
Replication

- A short oligonucleotide that starts a new strand.

**Figure 2:** Transcription of DNA into RNA

The complementary strand has the same sequence as the transcript RNA (except U for T).

The RNA polymerase melts the DNA and catalyses the synthesis of the RNA.
The Genetic Code

- **Codon** - a triplet of bases, codes a specific amino acid (except the stop codons)
- **Stop codons** - signal termination of the protein synthesis process
- Different codons may code the same amino acid

The genetic code, written by convention in the form in which the codons appear in mRNA. The three terminator codons, UAA, UAG, and UGA, are boxed in red; the AUG initiator codon is shown in green.

http://ntri.tamuk.edu/cell/ribosomes.html
The Human Genome: numbers

- 23 pairs of chromosomes
- \(\sim 3,200,000,000\) bases
- \(\sim 25,000\) genes
- Gene length: 1000-3000 bases, spanning 30-40,000 bases
- \(\sim 1,000,000\) protein variants
Hybridization

- DNA double strands form by “gluing” of complementary single strands
- Complementarity rule: A-T, G-C

Use probe to identify if target contains a particular sequence
The Human Genome Project

- Project planned for 15 years, initiated 1990
- US budget: 3 billion Dollars
- Main players: US, Europe, Japan
- Over 50 participating laboratories
2. Introduction to Promoter Analysis

Slides with Chaim Linhart
Regulation of Expression

• Each cell contains a copy of the whole genome - but utilizes only a subset of the genes
• Most genes are highly regulated - their expression is limited to specific tissues, developmental stages, physiological condition

How is the expression of genes regulated?

One way is through transcriptional regulation
Regulation of Transcription

- A gene’s transcription regulation is mainly encoded in the DNA in a region called the **promoter**.
- Each promoter contains several short DNA subsequences, called **binding sites (BSs)** that are bound by specific proteins called **transcription factors (TFs)**.
Regulation of Transcription (II)

TFs bound to their BSs

Transcription machinery

Gene start
Regulation of Transcription (III)

• By binding to a gene’s promoter, TFs promote or repress the recruitment of the transcription machinery

• A gene’s transcription is determined by the specific combination of BSs in its promoter
WH-questions

• √ Why are we looking for common BSs?

• *What* exactly are we trying to find?

• *Where* should we look for it?

• *How* can we find it?
Models for Binding Sites
(I) Exact string(s)

Example:
BS = TACACC, TACGGGC

CAATGCAGGATACACCAGATCGGTA
GGAGTAGCAGGCAAGTCCCCATGTGA
AGGCTTGAGACCAGACTCTCATACACCCTA
(II) String with mismatches

Example:

BS = TACACC + 1 mismatch

CAATGCAGGATTACACC GATCGGTA

GGAGTACAGCAAGTCCCCATGTGA

AGGCTGGACCAGACTCTACCTA
(III) Degenerate string

a.k.a consensus

Example:

BS = TASDAC (S={C,G} D={A,G,T})

CAATGCGAGGATACACGATCGGTAA
GGAGTAGTACCAAGTCCCCCATGTGTA
AGGCTGGGACCAGACTCTACTACGACTA
(IV) Position Weight Matrix (PWM)

a.k.a Position Specific Scoring Matrix (PSSM)

**Example:**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.1</td>
<td>0.8</td>
<td>0</td>
<td>0.7</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0.1</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>T</td>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Need to set score threshold

ATGCAGGATACACCGATCGGTA \[0.0605\]
GGAGTAGAGCAAGTCCCGTGA \[0.0605\]
AAGACTCTACAATTATGGCGT \[0.0151\]

ATGCAGGATACACCGATCGGTA
GGAGTAGAGCAAGTCCCGTGA
AAGACTCTACAATTATGGCGT
Motif Logo representations
3. PBM and HT-SELEX
DNA Microarrays

- A DNA microarray allows scientists to perform an experiment on thousands of genes at the same time.
- Each spot on a microarray contains multiple identical strands of DNA.
- The DNA sequence on each spot is unique.
- Each spot represents one gene.
- Thousands of spots are arrayed in orderly rows and columns on a solid surface (usually glass).
- The precise location and sequence of each spot is recorded in a computer database.
- Microarrays can be the size of a microscope slide, or even smaller.

http://learn.genetics.utah.edu/content/labs/microarray/analysis/
Protein Binding Microarrays

Berger et al, Nat. Biotech 2006

- Generate an array of double-stranded DNA with all possible 10-mers
PBM (2)
PBM - implementation

- ~41K probes, each 35nt long
- Probes contain all possible 10-mers
- Experiment gives binding intensity of the TF to each probe

- For each 8-mer, can combine signals from all probes that contain it (or differ in 1nt) to obtain signal
SELEX
Systematic Evolution of Ligands by EXponential enrichment

Start with a random pool of double-stranded DNA sequence of a fixed length (20-50nt)

Each cycle contains higher fraction of bound sequences

After each cycle – sequence a sample from the pool

Pool of DNA sequences

Amplify

Select bound sequences

Bind TF to the sequences
High-Throughput SELEX (HT-SELEX)

- Same - but using high throughput technologies (in particular ultra cheap and fast deep sequencing techniques)

Jolma et al, Genome Res 2010
HT-SELEX (2)

- Output: for each of cycles 0,1,..4/5 , a sample of the sequences present.
- Sequence length: 14-50
- Sample size: thousands to millions per cycle
The workshop in a nutshell (again)

1. Input 1: HTS experiment data

2. Develop a method to build a model from the data

3. Input 2: Test data: PBM experiment

4. Output: prediction. Use the model to predict sequence ranks

5. Evaluate the prediction quality vs. the hidden signal

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>CATGTAAGAGTTGACTCTGGTCTGT</td>
<td>28926</td>
</tr>
<tr>
<td>TGGCTCATCAGAGTCCCGTGAAAC</td>
<td>1457</td>
</tr>
<tr>
<td>TCCAGTTTAGGTGGCGCCCGGAACC</td>
<td>12972</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>CATGTAAGAGTTGACTCTGGTCTGT</td>
<td>11</td>
</tr>
<tr>
<td>TGGCTCATCAGAGTCCCGTGAAAC</td>
<td>5</td>
</tr>
<tr>
<td>TCCAGTTTAGGTGGCGCCCGGAACC</td>
<td>200</td>
</tr>
</tbody>
</table>

(hidden)
HT-SELEX:


PBM:

Fin