Tools for processing next-generation sequencing data:

Statistical models connecting “Chip-seq Inputs” to “RNA-seq Outputs”

Mark Gerstein
Yale

Slides at Lectures.GersteinLab.org
(See Last Slide for References & More Info.)
THE SEQUENCE EXPLOSION

Cost per million base pairs of sequence (log scale)

Automated Sanger Sequencing:
Based on a decades-old method, at the peak of the technique, a single machine could produce hundreds of thousands of base pairs in a single run.

$10,000

$1,000

$100

$10

Third-generation Sequencing:
Companies such as Helicos Biosciences already read sequence from short, single DNA molecules. Others, such as Pacific Biosciences, Oxford Nanopore and Ion Torrent say they can read from longer molecules as they pass through a pore.

Sequencing by Ligation:
This technique, employed in SOLiD and Polonator instruments uses a different chemistry from previous technologies and samples every base twice, reducing the error rate.

International HapMap Project

23andMe

deCODEme

Navigenics

A Thousand Genomes

nature

Science

The human genome

Whole Genome Shotgun Sequence

Gene sequence stored in international public databases

2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010

The Trace archive, started in 2000, houses raw sequence data, and currently holds 1.8 trillion base pairs.
Personal genomes will become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.
Only ~1% of personal genome sequences are canonical protein coding genes regions (CDS).

The remaining 99% are non-coding.
Importance of Dark Matter of the Genome

- Non-coding regions contain the control elements for coding regions.
- Some non-coding regions are functional & are pervasively transcribed.
- “Molecular Fossils” in the non-coding genome represent a historical record of the genome.
- Most disease-associated mutations (e.g. GWAS hits) are in non-coding regions.

[Gravitational lensing by dark matter in Abell 1689 – HST (NASA, ESA)]
How might we annotate a human text?

If you want to be a thorough-going world traveler, you need to learn 6,912 ways to say “Where is the toilet, please?” That’s the number of languages known to be spoken by the peoples of planet Earth, according to Ethnologue.com.

If you want to be the complete polyglot, you also have quite a challenge ahead of you, learning all the ways to say:

```
printf("hello, world\n");
```

(This one is in C.) A catalog maintained by Bill Kinnersley of the University of Kansas lists about 2,500 programming languages. Another survey, compiled by Diamuid Piggott, puts the total even higher, at more than 8,500. And keep in mind that whereas human languages have had millennia to evolve and diversify, all the computer languages have sprung up in just 50 years. Even by the more-conservative standards of the Kinnersley count, that means we’ve been inventing one language a week, on average, ever since Fortran.

For ethnologists, linguistic diversity is a cultural resource to be nurtured and preserved, much like biodiversity. A good-enough notation—for expressing an algorithm or defining a data structure.

There are programmers of my acquaintance who will dispute that last statement. I expect to hear from them. They will argue—zealously, ardently, vehemently—that we have indeed found the right programming language, and for me to claim otherwise is willful ignorance. The one true language may not yet be perfect, they’ll concede, but it’s built on a sound foundation and solves the main problems, and now we should all work together to refine and improve it. The catch, of course, is that each of these friends will decide which end of a boiled egg to crack. This famous tempest in an egg cup was replayed 250 years later by designers of computer hardware and communications protocols. When a block of data is stored or transmitted, either the least-significant bit or the most-significant bit can go first. Which way is better? It hardly matters, although life would be easier if everyone made the same choice. But that’s not what has happened, and so quite a lot of hardware and software is needed just to swap ends at boundaries between systems.

This modern echo of Swift’s Endian wars was first pointed out by Danny Cohen of the University of Southern California in a brilliant 1980 memo, “On holy wars and a plea for peace.” The memo, subsequently published in Computer, was widely read and admired; the plea for peace was ignored.

Another feud—largely forgotten, I think, but never settled by truce or treaty—focused on the semicolon. In Algol and Pascal, program statements had to be separated by semicolons. For example, `in x:=0; y:=x+1; z:=2` the semicolons tell the compiler where one statement ends and the next begins. C
Sources of Annotation: Comparative & Functional

Signal processing of raw experimental data:
- Removing artefacts
- Normalization
- Window smoothing

Segmentation of processed data into active regions:
- Binding sites
- Transcriptionally active regions

Group active regions into larger annotation blocks
Concepts:

Raw Tracks to Networks & Relating Genomic Inputs to Outputs

Signal processing of raw experimental data:
- Removing artefacts
- Normalization
- Window smoothing

Segmentation of processed data into active regions:
- Binding sites
- Transcriptionally active regions

Group active regions into larger annotation blocks

Further analysis: Building regulatory networks

Inputs (mostly Chip-seq of TFs & Chromatin)

Outputs (mostly RNA-seq)


[Science 330:6012]
Outline

• Using His. Mods. to predict gene expression
• Comparing this with TF binding
• Using His. Mods to predict TF binding
Modeling Transcription: Connecting Inputs & Outputs

• Models connecting various types of high-level data
  – HMs+TFs => gene expression
  – HMs => TFs

ENCODE Production Project

Consortium Comprises ~50 Labs

Subprojects:
- Transcriptome
- Chromatin
- TFs
Histone Modification (HM) model

Chromatin features: Histone modifications

Predictors

RNA-Seq data

Prediction target: Gene expression level

Gene k

Bin 40-1 (TSS-4kb to TSS)  Bin 41-80 (TSS to TSS+4kb)  Bin 120-81 (TTS-4kb to TTS)  Bin 121-160 (TTS to TTS+4kb)

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Features:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histone modifications</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bins</th>
<th>Predictors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bin 1</td>
<td>HM1, 2, 3, .....</td>
</tr>
<tr>
<td>Bin 2</td>
<td></td>
</tr>
<tr>
<td>......</td>
<td></td>
</tr>
<tr>
<td>Bin160</td>
<td></td>
</tr>
</tbody>
</table>

RNA-Seq data

Prediction target: Gene expression level

[Cheng et al. (11), Genome Biol. 12: R15]
His. mods around TSS & TTS are clearly related to level of gene expression, in a position-dependent fashion.
Integrate all histone modifications to predict gene expression levels

Classify H/L genes (SVM)

Predict expression values

Magnitude of Prediction from a “bin” around the TSS

Cheng et al. (11) Genome Biol. 12: R15

[Image: Graphs showing AUC and observed expression values vs. predicted expression values from 13HM and miRNA expression models]
HM models are tissue specific

HM model—Best prediction is achieved by using histone modification and expression data from the same developmental stage

[Cheng et al. ('11) Genome Biol. 12: R15]
Scale up to Human
Application of chromatin model in 5 species: Consistent Performance

>50% of variation of expression levels can be explained by HMs
TFs => Expr
Doing a Model with TFs:
Positive and negative regulators from correlating
TF signal at TSS with gene expression

[Cheng et al. (’11) PLOS CB]
Predictor v2: 2-levels, now with TFs

[Cheng et al. NAR ('11)]
Mouse ESC Models Illuminates Different Regions of Influence for TFs vs HMs

- Datasets
  - ChIP-Seq for 12 TFs (Chen et al. 2008)
  - ChIP-Seq for 7 HMs (Meissner et al.’08; Mikkelsen et al. ’07)
  - RNA-Seq (Cloonan et al. 2008)

A TF+HM model that combine TF and HM features does NOT improve accuracy!

A Model with only a Few of the Thousands of Mouse TFs is able to Predict Well

[Cheng et al. NAR ('11, in press)]
Scale up to Human: TFs

Pearson’s $r=0.81$; RMSE=2.57
Classification: AUC = 0.89
Regression: $r = 0.62$; RMSE = 3.06

Different types of TFs have different correlation with gene expression

Prediction of Differential Expression

TF model– differential TF binding signals are predictive of differential expression levels between two human cell lines

TFs active in a particular cell type are the strongest predictors in that cell

HMs => TFs
Chromatin model: link histone modification patterns to TF binding

Predictors (HM)

<table>
<thead>
<tr>
<th>Bin 1 (100bp)</th>
<th>Bin 2</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>HM1</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>HM2</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>HM3</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

TF binding site?

| Bin1 | Bin2 | Bin3 | Bin4 | Bin5 | Bin6 | Bin7 | Bin8 | Bin9 | Bin10 | ...
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ChIP-seq Data

Machine Learning Method (SVM et al.)

ROC curve

AUC

TPR (sensitivity)

FPR (1-specificity)
Data for Predicting TF targets from HMs

• Worm (same as for gene expr.)
  – Chip-seq : 22 TFs + Pol2 in a variety of stages
  – Chromatin Chip-chip : >12 HMs mostly in EE & L3
• Yeast
  – TF binding data:
    • ChIP-chip for 203 yeast TFs
  – Chromatin modification profile data
    • ChIP-chip, Pokholok et al. 2005
    • 14 profiles: H3(YPD), H4(YPD), H3(H2O2), H3K9AC/H3(YPD), H3K14AC/H3(YPD), H3K14AC/WCE(YPD), H3K14AC/H3(H2O2), H4AC/H3(YPD), H4AC/H3(H2O2), H3K4ME1/H3(YPD), H3K4ME2/H3(YPD), H3K4ME3/H3(YPD), H3K36ME3/H3(YPD), H3K79ME3/H3(YPD)
  – Positional weighted matrices for yeast TFs

Relating the Chromatin Model to TF Binding Sites in worm

- Model Predicts Sites Fairly Accuracy
- Accuracy improved when coupled with PWM

[Gerstein et al., Science ‘10]
HIS+PWM leads to higher accuracy for predicting **SOME** yeast TF target genes – the His-sensitive ones

[Cheng et al. Genome Biol. ('11) 12:R111]
Histone-sensitive vs. -insensitive TFs

[Cheng et al. Genome Biol. ('11) 12:R111]
Models of Transcription Relating Various Genome Tracks of Information – His. Mods, TF Binding & Gene Expression

• Predictive models of gene expression
  – Applicable in many contexts
  – Work for miRNAs as well as genes
  – Show variable importance of regions around genes for chromatin & TFs
  – Show TF & HM signals are redundant for ‘predicting’ gene expression
  – Surprisingly, a few TFs, particularly TFSSs, are quite predictive

• Predictive models of TF Binding
  – Open chromatin model + PWM does better than features individually
  – TFs in yeast divide into 2 groups: His-sens & His-insens
  – His-sens TFs easier to predict & higher expression + more PPIs
Acknowledgements


- TF-model: Cheng C, Shou C, Yip KY, Gerstein MB.


- encode: Chao Cheng, Roger Alexander, Renqiang Min, Kevin Y. Yip, Jing Leng, Joel Rozowsky, Koon-kiu Yan, Xianjun Dong, Sarah Djebali, Yijun Ruan, Carrie A Davis, Piero Carninci, Timo Lassman, Thomas R. Gingeras, Roderic Guigó Serra, Ewan Birney, Zhiping Weng, Michael Snyder, Mark Gerstein + ENCODE consortia.
Default Theme

- Default Outline Level 1
  - Level 2
Info about content in this slide pack

• Some of the material in this presentation is unpublished, marked by in press or submitted citations on various slides. Please treat it as confidential material and do not redistribute.

• PERMISSIONS: This Presentation is copyright Mark Gerstein, Yale University, 2010. Please read permissions statement at http://www.gersteinlab.org/misc/permissions.html. Feel free to use images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).

• Paper references in the talk were mostly from Papers.GersteinLab.org.

• PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info. In particular, many of the images have particular EXIF tags, such as kwpotppt, that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt