Human Genome Analysis:

Understanding & Annotating Promoters Using Statistical Models

Slides downloadable from Lectures.GersteinLab.org
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:

```c
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 Diarmuid 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.

Every programmer knows there is one true programming language. A new one every week

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 have 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

[B Hayes, Am. Sci. (Jul.- Aug. ’06)]
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
Comparative "Rollout"

Subprojects:
- Transcriptome
- Chromatin
- TFs
Comparative ENCODE Functional Genomics Resource
(EncodeProject.org/modENCODE.org)

- Broad sampling of conditions across transcriptomes & regulomes for human, worm & fly
  - embryo & ES cells
  - developmental time course (worm-fly)
- In total: >100B reads (>1000 datasets)
  - RNAseq: ~65B reads, ~585 sets
Previous studies have compared RNA transcription between closely related organisms (e.g., RNA-seq within mammals, Brawand et al. '11) ...
Previous studies have compared RNA transcription between closely related organisms (e.g. RNA-seq within mammals, Brawand et al. '11) …

… or integrated diverse –omic information within each of several species (eg modencode '10)
Comparative ENCODE

Previous studies have compared RNA transcription between closely related organisms (e.g. RNA-seq within mammals, Brawand et al. '11) ...

A first effort to comprehensively integrate diverse data across distantly related species

... or integrated diverse –omic information within each of several species (e.g. modencode '10)
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.
Focus on Promoters

• Key Questions
  – How do we define the active regions of promoter?
  – For an active promoter, how do we relate it bound TFs, its epigenetic marks & its chromatin state to the level of transcription?
  – Are these definitions & relationships conserved between very different species?

Focus on Promoters

- Significance of the "Answers"
  - Unified understand of “ancient” principles of gene expression in animals
  - This is useful in better annotating the human genome
  - Also, useful in understanding how epigenetic changes (eg in cancer) drive expression

Human Genome Analysis:

Understanding & Annotating Promoters Using Statistical Models

- Genome Annotation Intro
  - ENCODE
  - Lots of Matched Data
    - On upstream activity and gene expression

- HM models of gene expression
  - Works for ncRNAs as well as genes
  - Developing a universal cross-species model

- TF models
  - Variable importance of regions around genes for chromatin & TFs
  - TF & HM signals are redundant for ‘predicting’ expression
  - Surprisingly, a few TFs are quite predictive
The Problem: Relating Genomic Inputs to Outputs
Inputs v Outputs:
Upstream Binding/Modification v Expression

Spearman's correlation:
Pol II, 0.64;
H3K4me3, 0.58
Histone Modification (HM) model

Chromatin features: Histone modifications

Predictors

RNA-Seq data

Prediction target: Gene expression level

~10000 refseq genes

Bin 1, Bin 2, ..., Bin160

Gene k
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
Human Results

Pearson's $r = 0.9$ (p-value $< 2.2 \times 10^{-16}$)
RMSE = 1.9
Classification: AUC = 0.95
Regression: $r = 0.77$ (RMSE = 2.3)
Application of chromatin model in 5 species: Consistent Performance

>50% of variation of expression levels can be explained by HMs
Comparison of Models for Gene Expression, Building a Universal Model

Human, Worm & Fly

Universal Model is Built Simultaneously on Data from all 3 Organisms & Predicts on all 3 with a Single Set of Parameters
Performance of Universal, cross-organism Model

- works almost as well as species specific models
- works for both mRNAs and ncRNAs

### Prediction Accuracy for Protein-coding Genes

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Worm</th>
<th>Fly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Trained in Human</td>
<td>.82</td>
<td>.66</td>
<td>.69</td>
</tr>
<tr>
<td>Model Trained in Worm</td>
<td>.66</td>
<td>.74</td>
<td>.70</td>
</tr>
<tr>
<td>Model Trained in Fly</td>
<td>.69</td>
<td>.68</td>
<td>.84</td>
</tr>
</tbody>
</table>

### Prediction Accuracy of Universal Model

<table>
<thead>
<tr>
<th></th>
<th>Protein coding</th>
<th>ncRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein coding</td>
<td>.80</td>
<td>.69</td>
</tr>
<tr>
<td>ncRNA</td>
<td>.73</td>
<td>.51</td>
</tr>
</tbody>
</table>

[ENCODER-modencode Transcriptome paper, submitted]
Human Genome Analysis:
Understanding & Annotating Promoters Using Statistical Models

- **Genome Annotation Intro**
  - ENCODE
  - Lots of Matched Data
    - On upstream activity and gene expression

- **HM models of gene expression**
  - Works for ncRNAs as well as genes
  - Developing a universal cross-species model

- **TF models**
  - Variable importance of regions around genes for chromatin & TFs
  - TF & HM signals are redundant for ‘predicting’ expression
  - Surprisingly, a few TFs are quite predictive
Doing a Model with TFs: Positive and negative regulators from correlating TF signal at TSS with gene expression.
Predictor v2:
2-levels, now with TFs
Human Results

Pearson’s $r = 0.81$; RMSE = 2.57
Classification: AUC = 0.89
Regression: $r = 0.62$; RMSE = 3.06
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!
TF model accuracy only needs a small number of TFs for high accuracy (>90%)
Optimally arrange TFs into 3 levels by simulated annealing, maximizing downward-pointing edges.

Hierarchy height distribution approximated by 3 levels.

Probing direction framed as an optimization problem.
Avg. correlation between binding signal of TF & gene expression of its target

Integration of TF hierarchy with other ‘omic information:
more influential & connected TFs on the top
Human Genome Analysis:

Understanding & Annotating Promoters Using Statistical Models

• Genome Annotation Intro
  • ENCODE
  • Lots of Matched Data
    • On upstream activity and gene expression

• HM models of gene expression
  • Works for ncRNAs as well as genes
  • Developing a universal cross-species model

• TF models
  • Variable importance of regions around genes for chromatin & TFs
  • TF & HM signals are redundant for ‘predicting’ expression
  • Surprisingly, a few TFs are quite predictive
modENCODE/ENCODE Transcriptome Group

Joel Rozowsky, Koon-Kiu Yan, Daifeng Wang,

Chao Cheng, James B. Brown, Carrie A. Davis, LaDeana Hillier, Cristina Sisu,

Jingyi Jessica Li, Baikang Pei, Arif O. Harmanci, Michael O. Duff, Sarah Djebali, Roger
P. Alexander, Burak H. Alver, Raymond K. Auerbach, Kimberly Bell, Peter J. Bickel, Max E.
Boeck, Nathan P. Boley, Benjamin W. Booth, Lucy Cherbas, Peter Cherbas, Chao Di, Alex
Dobin, Jorg Drenkow, Brent Ewing, Gang Fang, Megan Fastuca, Elise A. Feingold, Adam
Frankish, Guanjun Gao, Peter J. Good, Phil Green, Roderic Guigo, Ann Hammonds, Jen
Harlow, Roger A. Hoskins, Cedric Howald, Long Hu, Haiyan Huang, Tim J. P. Hubbard, Chau
Huynh, Sonali Jha, Dionna Kasper, Masaomi Kato, Thomas C. Kaufman, Rob Kitchen, Erik
Ladewig, Julien Lagarde, Eric Lai, Jing Leng, Zhi Lu, Michael MacCoss, Gemma May,
Rebecca McWhirter, Gennifer Merrihew, David M. Miller, Ali Mortazavi, Rabi Murad, Brian Oliver,
Sara Olson, Peter Park, Michael J. Pazin, Norbert Perrimon, Dmitri Pervouchine, Valerie
Reinke, Alexandre Reymond, Garrett Robinson, Anastasia Samsonova, Gary I. Saunders, Felix
Schlesinger, Anurag Sethi, Frank J. Slack, William C. Spencer, Marcus H. Stoiber, Pnina
Strasbourger, Andrea Tanzer, Owen A. Thompson, Kenneth H. Wan, Guilin Wang, Huaien Wang,
Kathie L. Watkins, Jiayu Wen, Kejia Wen, Chenghai Xue, Li Yang, Kevin Yip, Chris Zaleski, Yan
Zhang, Henry Zheng, Steven E. Brenner, Brenton R. Graveley, Susan E.
Celniker, Thomas R Gingeras, Robert Waterston
TF-v-expr:

worm-HM:

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
Default Theme

• Default Outline Level 1
  – Level 2
SUBJECT: Networks

DESCRIPTION:

NOTES:
This PPT should work on mac & PC. Paper references in the talk were mostly from Papers.GersteinLab.org.

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).

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.