Found In Translation: A machine learning model for mouse to human inference

Rachelly Normand
Shen-Orr lab

Bioinfo4Women seminars program
Barcelona Supercomputing Center
October 31st 2019

Normand et al., Nature Methods, 2018
Mice are widely used in biomedical research
The cross-species gap is evident in health and disease

• Species differences in health.

• Mouse models that fail to mimic human disease.

• Clinical trial failure – TGN1412.

Liao & Zhang, PNAS 2013
McDonald et al., Physiological Reports 2015
Wikins et al., Nova Science Publishers 2011

Monoclonal antibody TGN1412 trial failure explained by species differences in CD28 expression on CD4+ effector memory T-cells
Eastwood et al., BJP 2010
The utility of mouse models is a much debated issue

Conservation and divergence in the transcriptional programs of the human and mouse immune systems

Shay et al., PNAS 2013

Genomic responses in mouse models poorly mimic human inflammatory diseases

Seok et al., PNAS 2013
The utility of mouse models is a much debated issue.
Experimental attempts to bridge the cross-species gap are long and costly

Myeloablation

12 weeks

>25% human CD45+ cells in peripheral blood

Human hematopoietic stem cells (CD34+)

Jackson Laboratory website
Prior computational solutions consist static assemblies of species differences
Found In Translation (FIT) model utilizes public data to improve mouse to human translation.
The training data FIT learns from

**Human**
- **Disease**: Microarrays, RNA-seq
- **Control**: Microarrays, RNA-seq
- **Effect size**
- **Cross-species pair (CSP)**

**Mouse**
- **Disease**: Microarrays, RNA-seq
- **Control**: Microarrays, RNA-seq
- **Effect size**
High quality comprehensive training compendium manually assembled from GEO

- 170 cross-species pairs from 28 different diseases.
- Only high quality datasets are included.
- Uniform pre-processing.
- Careful cross-species pairing.
Cross-species gap is quantified by comparing mouse and human differentially expressed genes.

170 CSPs X 117 threshold pairs ≈ 20,000 tests
Only 1 in 20 differentially expressed genes are shared between mouse and human.
FIT learns cross-species relationship per gene
The human prediction is the mean of the distribution created from bootstrapping.
FIT is evaluated in a leave-one-disease-out mode

170 comparisons
FIT is evaluated in a leave-one-disease-out mode

170 comparisons

Disease A

Disease C

Disease D

Disease B

Training

Test
FIT is evaluated in a leave-one-disease-out mode

170 comparisons

170 CSPs x 117 threshold pairs ≈ 20,000 tests
How FIT works

Training data of Sdf2
(leaving one disease out)

[Graph showing data distribution and comparison]

Prediction and comparison to human

[Graph showing comparison between different conditions]
FIT rescues human DEGs undetected by the mouse

Fold-change = 0.25
q-value = 1
FIT performance is variable but can be predicted with high accuracy.
An SVM classifier predicts FIT’s relevance correctly in 88% of cases.

| True identification | Good performance prediction | 23% |
| False identification | Poor performance prediction | 61% |

q-value=0.1
fold-change=0.15

23/26 = 88% True relevance

Correct in 84% of predictions
The classifier is accurate in 80% of the cases on average across thresholds.
FIT can triple the amount of true positive genes
FIT increases true positive fraction by 20-50%
Confidence intervals allow rescuing more human differentially expressed genes.
FIT does not repeatedly boost a specific group of genes

Percentage of comparisons in which the gene is a FIT DEG in

Number of genes
Novel discovery in IBD based on FIT prediction

- Human data as a silver standard
- FIT captures signals that may be missed by a single human experiment
FIT is able to boost the signal of disease-relevant pathways
How FIT should be used

Mouse experiment

Mouse relevant genes

SVM classifier

FIT model

Human predictions

Human relevant genes

Downstream analysis
FIT is available as an R package and a web tool

www.mouse2man.org

shenorrLab/FIT.mouse2man

A machine-learning model to improve mouse to human inference

What is FIT?

Cross-species differences form barriers to translational research that ultimately hinder the success of clinical trials. Yet systematic incorporation of the wealth of knowledge on species differences in the interpretation of animal model data is lacking. We have

How does FIT work?

Due to the complexity of translating mouse to human biology FIT's performance in terms of predicting values that resemble human data, is variable across datasets. Therefore, in order to make FIT more effective
Summary

• Increasing the relevance to humans with zero experimental cost.
• The human data as “silver-standard”.
• FIT is a novel paradigm of data re-use and can be applied to other mapping problems.
Acknowledgements

Technion
- Prof. Shai Shen-Orr
- Dr. Elina Starosvetzky
- Mayan Briller
- Dr. Renaud Gaujoux
- Amit Ziv-Kenet
- Dr. Gali Shalev-Malul

Stanford University
- Prof. Rob Tibshirani
- Wenfei Du
Model development process

- Finding predictive factors for expression conservation (per single genes)
  - TFs binding similarity
  - Sequence homology
  - Conservation at the gene/pathway level
- Protein-protein interactions network (per single genes)
- Hierarchical Bayesian model (training data dominated)
- Linear regression with weights for different diseases (noise even within disease, distances very big)
- Shrinking the slope to 0 (training data dominated)
Gene and dataset comparison metrics (before/after FIT)

• Percent agreement (per single genes)

• Spearman correlation (too global)

• Sensitivity + specificity (too global)

• Defining FIT differentially expressed genes fold change (+/- confidence Intervals)
Different training data

• 647 CSPs in 27 different diseases – pairing all possible pairs

• Taking only high quality pairs

• Disease type specific training: cancer/ infectious diseases

• Homogenous training: same number of datasets from every disease
Training compendium structure
Training compendium platforms

Human

- GPL6947 5%
- GPL97 5%
- GPL571 11%
- GPL81 8%
- GPL11154 15%
- GPL96 17%
- GPL570 18%

Mouse

- GPL339 6%
- GPL6887 8%
- GPL81 8%
- GPL13112 12%
- GPL96 17%
- GPL1261 39%
Training compendium tissues

Human

- Peripheral Blood Mononuclear Cell: 4%
- Renal Tubule: 4%
- Caudate Nucleus: 5%
- White Blood Cells: 5%
- Lung: 5%
- Muscle: 11%
- Cerebral Cortex: 15%
- Whole Blood: 25%

Mouse

- Heart: 4%
- Kidney: 4%
- Calf: 5%
- Spleen: 5%
- White Blood Cells: 5%
- Lung: 6%
- Cerebral Cortex: 6%
- Liver: 8%
- Corpus Striatum: 15%
- Whole Blood: 15%
The new mouse data dictates the prediction and the training data adjusts it

\[ Z_p^H(g) = \alpha_g + \beta_g \cdot Z_p^M(g) \]

- Minimizing the sum of squares
- Shrinking slope to 1 and the intercept to 0

\[
\min_{\alpha_g, \beta_g} \sum_d (Z_p^H(g) - \alpha_g - \beta_g Z_p^M(g))^2 + \lambda (|\alpha_g| + |1 - \beta_g|)
\]

\( g \) – gene
\( p \) – cross-species pair
\( d \) – disease
Automatic pipeline of gene expression analysis

Microarray datasets

1. Check that the dataset is:
   • log-transformed (if not – log2 transformed)
   • normalize (if not – quantile normalized)

2. Probe fold-change:
\[
FC_{\text{probe}} = \text{mean}_{\text{sample} \in \text{Disease}} (\log_2 (\text{Intensity}_{\text{sample}})) - \text{mean}_{\text{sample} \in \text{Control}} (\log_2 (\text{Intensity}_{\text{sample}}))
\]

3. Gene fold-change:
\[
FC_{\text{gene}} = \max (FC_{\text{probe}} | \text{probe} \in \text{gene})
\]

4. Gene Z-score:
\[
Z_{\text{test}} (g) = \frac{\text{mean}(Z_{\text{samp} \in \text{Disease}} (g)) - \text{mean}(Z_{\text{samp} \in \text{Control}} (g))}{\sqrt{\frac{\text{SD}_{\text{Control}}}{N_{\text{Control}}} + \frac{\text{SD}_{\text{Disease}}}{N_{\text{Disease}}}}}
\]
Automatic pipeline of gene expression analysis

RNAseq datasets

1. Raw reads downloaded with R package SRAdb.
2. Pseudo-mapped and quantified by Kallisto.
3. Effect size computed by Sleuth.
FIT is evaluated by accounting a full confusion matrix
Sensitivity and specificity are computed per dataset

$$\text{TP ratio} = \frac{\text{TP}_{\text{FIT}}}{\text{TP}_{\text{Mouse}}}$$

$$\text{TN ratio} = \frac{\text{TN}_{\text{FIT}}}{\text{TN}_{\text{Mouse}}}$$

$$\text{FP ratio} = \frac{\text{FP}_{\text{FIT}}}{\text{FP}_{\text{Mouse}}}$$

$$\text{FN ratio} = \frac{\text{FN}_{\text{FIT}}}{\text{FN}_{\text{Mouse}}}$$
FIT performance split into five classes based on the confusion matrix values

<table>
<thead>
<tr>
<th>Major signal gain</th>
<th>Minor signal gain</th>
<th>Minor signal loss</th>
<th>Major signal loss</th>
<th>Equal performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP ratio</td>
<td>↑</td>
<td>↑</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>TN ratio</td>
<td>↑</td>
<td>↑</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>FP ratio</td>
<td>↓</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>FN ratio</td>
<td></td>
<td></td>
<td>↑</td>
<td>↑</td>
</tr>
</tbody>
</table>

The rest
FIT increases the specificity in signal gain datasets
Mouse data has higher sensitivity in signal loss datasets
The training data dictates the prediction more than the new mouse data

**Burns vs. Sepsis**

**Burns vs. Adenocarcinoma**

**Observed in:**
- Bayesian inference
- Linear regression when slope is shrunk to 0
The distances between diseases are very small
Correlation-based distances between comparisons

The disease weights are very small for most of the disease pairs. => Over fitting.