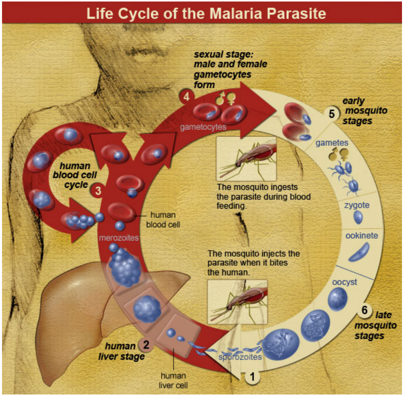
Transfer Learning of Genome Wide Transcription Dynamics during Malaria Infection

Venelin Mitov
ETH Zürich

February 5, 2014
Thesis presentation

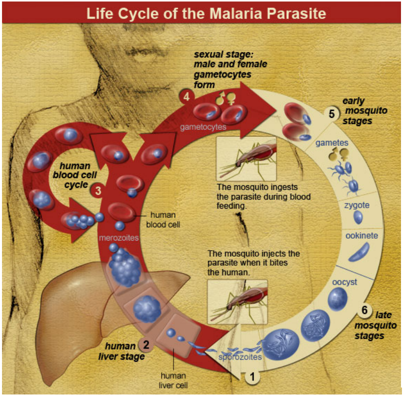
Outline

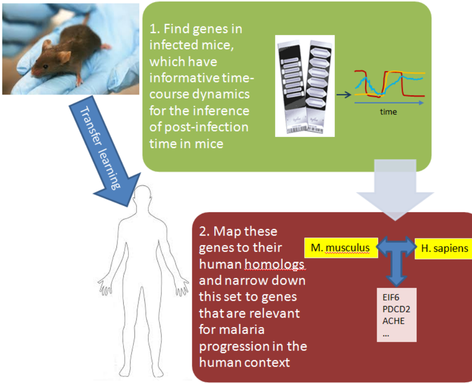
A Transfer Learning Approach



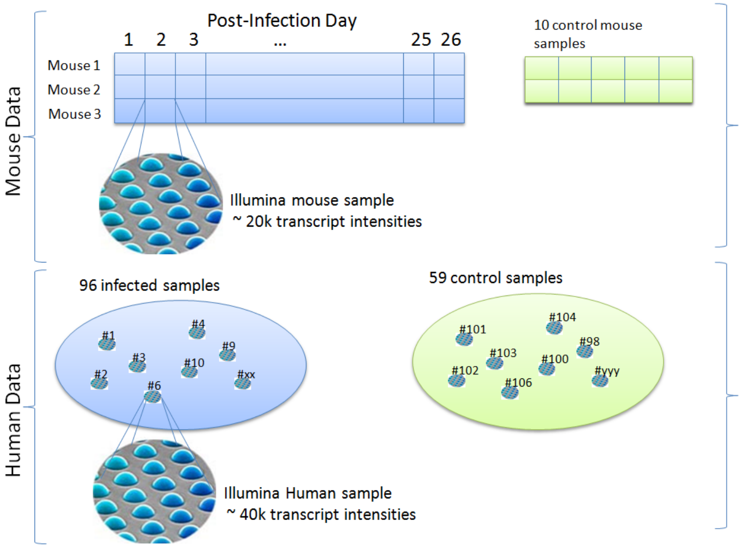
Courtesy: National Institute of Allergy and Infectious Diseases

A Transfer Learning Approach

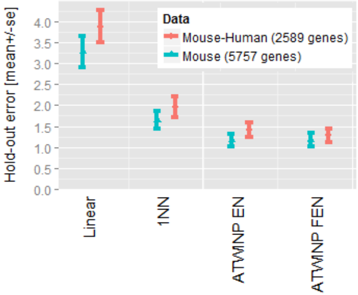




Courtesy: National Institute of Allergy and Infectious Diseases

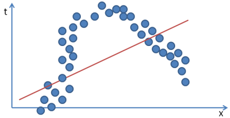

Mouse Data
Post-Infection Day
1 2 3 ... 25 26
Mouse 1
Mouse 2
Mouse 3
10 control mouse samples
Illumina mouse sample
~ 20k transcript intensities
Human Data
96 infected samples
#1
#2
#3
#4
#6
#9
#10
#xx
59 control samples
#101
#102
#103
#104
#106
#100
#98
#yyy
Illumina Human sample
~ 40k transcript intensities

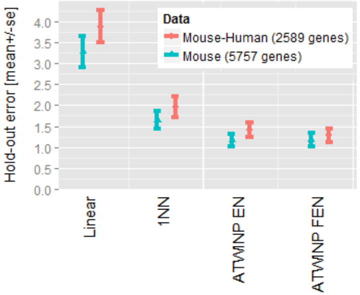
Results from Leave-One-Mouse-Out Cross Validation



Results from Leave-One-Mouse-Out Cross Validation

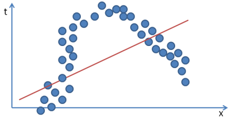
Linear regression

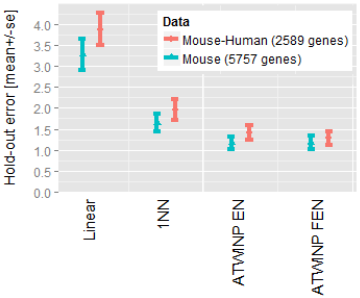




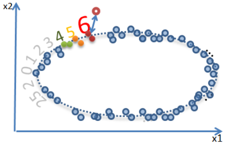
Results from Leave-One-Mouse-Out Cross Validation

Linear regression



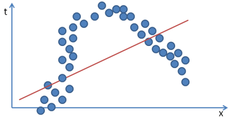


Classification (1^{st} NN)

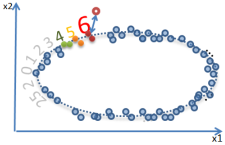


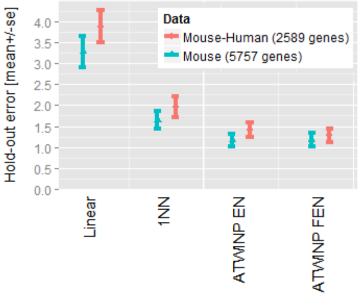
Results from Leave-One-Mouse-Out Cross Validation

Linear regression

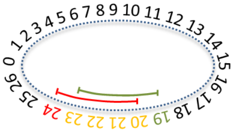


Classification (1^{st} NN)



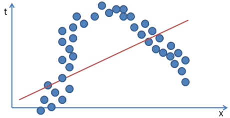


Aggregated Time Windows

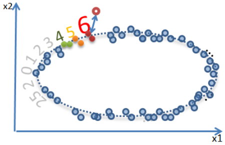


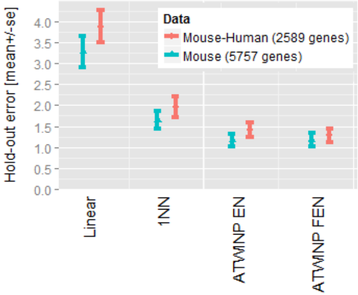
Results from Leave-One-Mouse-Out Cross Validation

Linear regression

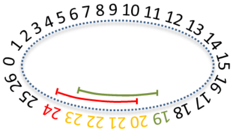


Classification (1^{st} NN)

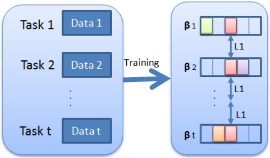




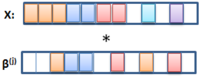
Aggregated Time Windows



+ fused logistic regression



One-Against-All Linear Logistic Regression



Model the logit function, $logit(\pi) := log(\pi/(1-\pi))$, as a linear function of $\mathbf{x}$:

$$logit(\pi^{(j)}(\mathbf{x})) \approx \mathbf{x}^T \boldsymbol{\beta}^{(j)}.$$

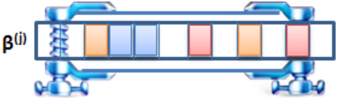
The negative log-likelihood is defined as:

$$-\ell^{(j)}(\boldsymbol{\beta}^{(j)}; [X|\mathbf{y}_j]) = \sum \log\left(\mathbf{1} + \exp(-\mathbf{y}_j \odot X\boldsymbol{\beta}^{(j)})\right), \quad j = 1, ..., t.$$

Maximum likelihood fit for $\boldsymbol{\beta}^{(j)}$:

$$\boldsymbol{\beta}^{(j)*} := \arg \min_{\boldsymbol{\beta}^{(j)} \in \mathbb{R}^{(1+d)}} \left\{-\ell^{(j)}(\boldsymbol{\beta}^{(j)}; [X|\mathbf{y}_j])\right\}$$

Regularization and Automatic Variable Selection

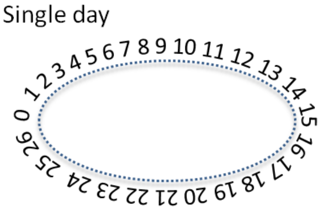


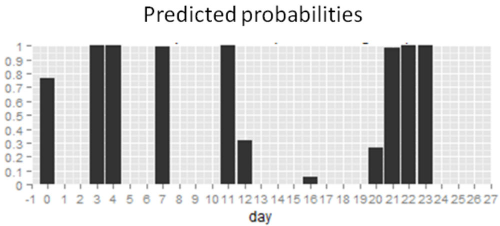
- L2-penalty (Ridge): $\frac{1}{2}\lambda_2||\boldsymbol{\beta}^{(j)}||_2^2 = \frac{1}{2}\lambda_2\sum_{k=1}^{d}\beta_k^{(j)^2}$
- L1-penalty (Lasso): $\lambda_1||\boldsymbol{\beta}^{(j)}||_1 = \lambda_1\sum_{k=1}^{d}|\beta_k|$
- Elastic Net penalty (Lasso+Ridge):$\lambda_1||\boldsymbol{\beta}^{(j)}||_1 + \frac{1}{2}\lambda_2||\boldsymbol{\beta}^{(j)}||_2^2$

Maximum A-Posteriori fit for $\boldsymbol{\beta}^{(j)}$:

$$\boldsymbol{\beta}^{(j)*} := \arg\min_{\boldsymbol{\beta}^{(j)}\in\mathbb{R}^{(1+d)}}\left\{-\ell^{(j)}(\boldsymbol{\beta}^{(j)};[X|\mathbf{y}_j]) + \lambda_1||\beta||_1 + \frac{1}{2}\lambda_2||\boldsymbol{\beta}^{(j)}||_2^2\right\}$$

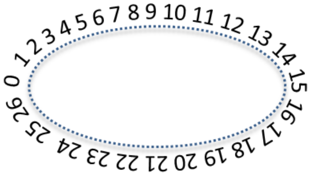
Single Day versus Time Window Prediction



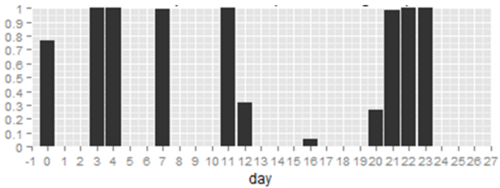


Single Day versus Time Window Prediction

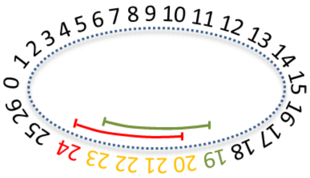
Single day

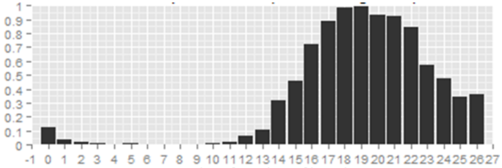


Predicted probabilities

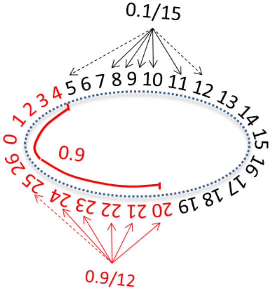


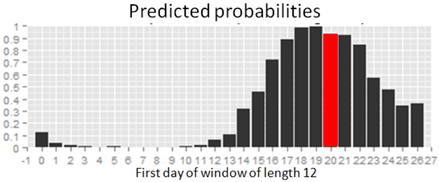
Time Window



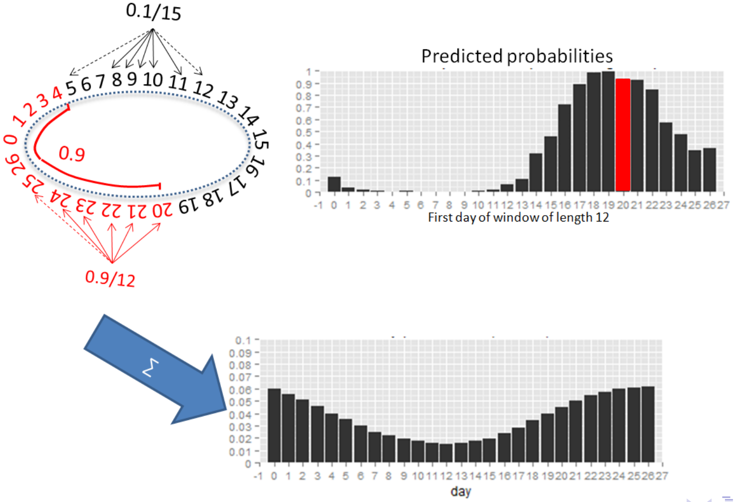


Aggregated Time Window Predictor (ATWINP)

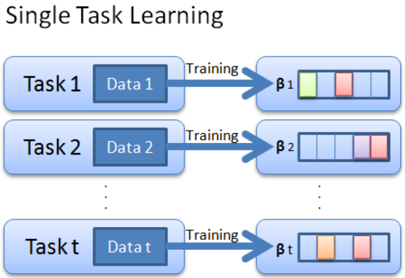




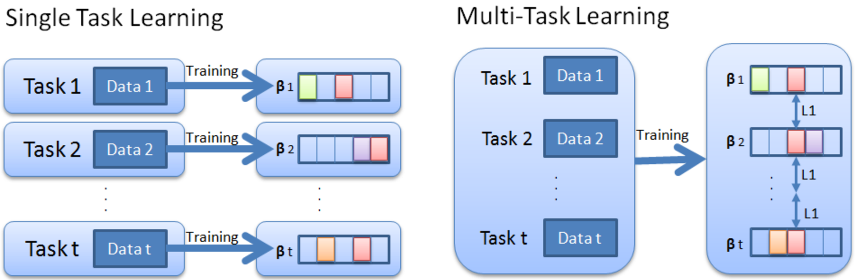
Aggregated Time Window Predictor (ATWINP)



The Idea of Multi-Task Learning



The Idea of Multi-Task Learning



Fused Elastic Net Logistic Regression (FLR)

Let $B := [\beta^{(1)}, ..., \beta^{(t)}] \in \mathbb{R}^{(1+d)\times t}$ be the coefficient matrix for all tasks and let $R \in \mathbb{R}^{t\times t}$ be a matrix defined in the following way:

$$R_{ij} := \begin{cases} 1 & \text{if } j = i-1 \text{ or } (i,j) = (1,t) \\ 0 & \text{otherwise} \end{cases}, \quad i,j = 1, ..., t.$$

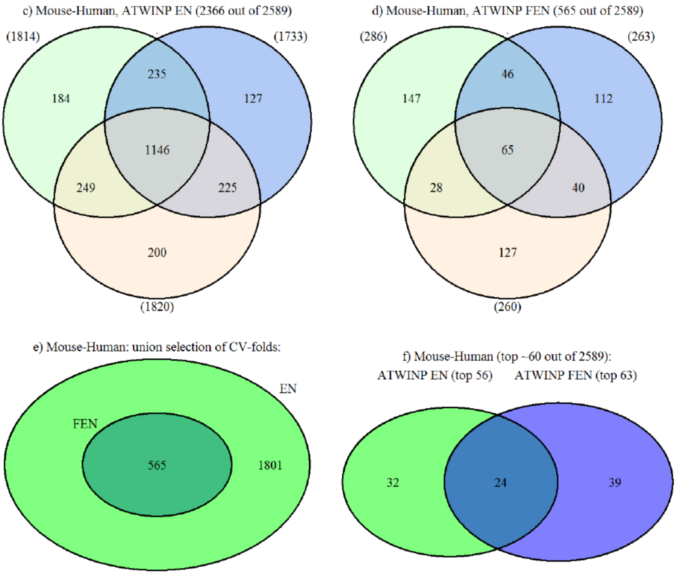
The multi-task fused elastic net negative log-likelihood is defined as:

$$\begin{aligned} -\ell^{MT}(B; [X|Y]) \quad := \quad & \sum \log\Big([1] + \exp(-Y \odot XB)\Big) \\ & + ||[\boldsymbol{\lambda}_1] \odot B||_1 + \frac{1}{2}||[\boldsymbol{\lambda}_2] \odot B||_2^2 \\ & + ||[\nu] \odot B(I - R)||_1 \end{aligned}$$

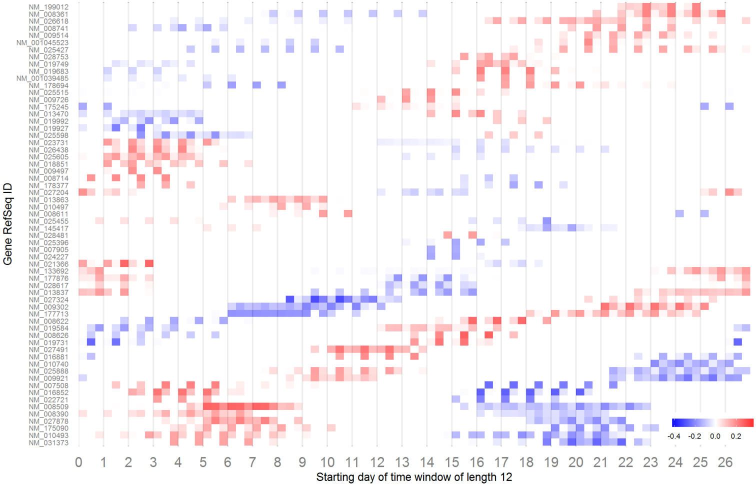
The Fused Elastic Neto Logistic Regression (FENLR) fit for B is obtained by solving

$$B^* = \arg \min_{B \in \mathbb{R}^{(1+d)\times t}} -\ell^{MT}(B; [X|Y]).$$

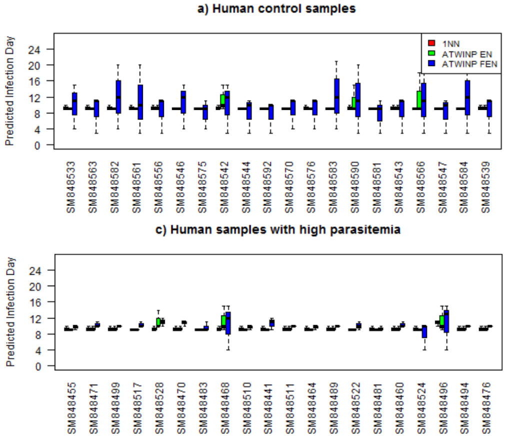
Selected Genes



Top 60 Genes, ATWINP FEN



Post-Infection Time Prediction in Human Patients



Discussion

- Our model can predict the post-infection time of an unlabeled infected mouse-sample with expected deviation of 1.28 days from the true post-infection time.
- The gene-expression profile of an infected host-organism preserves information with respect to the beginning of the infection, and can be used to characterize the disease progression on a fine time-scale.
- We were able to identify a set of genes that are informative for the disease progression in mice and we could quantify the effect of each selected gene at all points in the time-course of the infection.
- At the current time knowledge transfer from mouse to human patients cannot provide a valuable estimation of the post-infection time in humans.

Acknowledgements

- prof. Manfred Claassen - Advisor of the master thesis project
- Stefan, Eirini, Anita, Ana - colleagues in the Claassen's Group
- David and Brenda - Stanford Microbiology and Immunology Lab