A Machine Learning Pipeline for Phenotype Prediction from Genotype Data

Giorgio Guzzetta, Giuseppe Jurman, Cesare Furlanello

FBK-MPBA: Predictive Models for Biomedicine and Environment

http://mpba.fbk.eu

Funds from: Hiperdart FP7 and BICG AIRC
Fitting Quantitative Phenotypes from Genotype

- Quantitative phenotypes emerge everywhere in systems biology and biomedicine
- Complex common diseases with high individual variability and heterogeneous symptoms (no easy categorization among affected or affected/unaffected cases)
- Pre-condition for studying trajectories
- Combined approaches: estimate from molecular data the parameters needed to improve infectious disease modeling (e.g., susceptibility)
Example: Autism Spectrum Disorders are measured by semi-structured diagnostic assessments and sets of clinical tests. No easy categorization. Variable response to treatment. Evidence of different individual trajectories.

Also: molecular basis of response to pharmacological treatment in Major Depression
Reference study: A Monte Carlo Markov Chain (MCMC) model for predicting quantitative traits from genome-wide SNP data (Lee et al., 2008)

**GSCAN Public mice dataset** from the Wellcome Trust Center for Human Genetics (http://gscan.well.ox.ac.uk)

- Familiar, phenotype and genotype (biallelic)

**IN THIS STUDY**

- 2 Quantitative phenotypes:
  a. % of CD8+ cells (CD8)
  b. Mean Cell Haemoglobin (MCH)

- # samples:
  - 1,521 (CD8)
  - 1,591 (MCH)

- # features: 12,113

Reference Data and Study

Fitting physiological traits in mice

**Predicting Unobserved Phenotypes for Complex Traits from Whole-Genome SNP Data**

Sang Hong Lee, Julius H. van der Werf, Ben J. Hayes, Michael E. Goddard, Peter M. Visscher

**RESEARCH HIGHLIGHTS**

**Fitting phenotypes**

Analyzing the results of genome-wide association studies is a challenging task. This study implemented a novel approach called "Reference phenotypic prediction from genome-wide SNP data" (Lee et al., 2008). The approach combines genetic information with clinical data to predict phenotypes accurately and efficiently. This method uses a Monte Carlo Markov Chain (MCMC) model to estimate the effects of individual SNPs on a particular trait. The results show that the method is highly accurate, with a high explanatory power and low prediction error. This approach can be applied to predict new physiological traits in mice.
### Experimental populations (e.g. mice)

<table>
<thead>
<tr>
<th>R Package</th>
<th>Domain</th>
<th>Features</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>R/qtl</td>
<td>Experimental crosses</td>
<td>• QTL mapping</td>
<td>• HMM for dealing with missing data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Genotyping errors identification</td>
<td>• Interval mapping (EM algorithm)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Single-QTL genome scans</td>
<td>• Haley-Knott regression</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Two-QTL, two-dimensional genome scans</td>
<td>• Multiple imputation</td>
</tr>
<tr>
<td>HAPPY</td>
<td>Heterogeneous stocks</td>
<td>• Ancestral haplotype reconstruction</td>
<td>• Dynamic programming</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• QTL fine mapping</td>
<td>• Linear regression</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• ANOVA</td>
</tr>
<tr>
<td>bqtl</td>
<td>Inbred crosses, recombinant</td>
<td>• QTL mapping</td>
<td>• Likelihood-based</td>
</tr>
<tr>
<td></td>
<td>inbred lines</td>
<td></td>
<td>• Bayesian techniques</td>
</tr>
</tbody>
</table>
## Natural populations (e.g. humans)

<table>
<thead>
<tr>
<th>Software</th>
<th>Domain</th>
<th>Features</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENEHUNTER</td>
<td>Sib-pair</td>
<td>• Linkage analysis</td>
<td>• Haseman-Elston regression (traditional and EM)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Maximum likelihood</td>
</tr>
<tr>
<td>MERLIN</td>
<td>Any</td>
<td>• Linkage analysis</td>
<td>• Gene flow trees</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Linkage disequilibrium adjustment</td>
<td>• Sham regression</td>
</tr>
<tr>
<td>LOKI</td>
<td>Any</td>
<td>• Segregation analysis</td>
<td>• Monte Carlo Markov Chain</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Linkage analysis</td>
<td></td>
</tr>
<tr>
<td>PLINK</td>
<td>Any</td>
<td>• Association</td>
<td>• Standard linear regression</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Epistasis tests</td>
<td></td>
</tr>
</tbody>
</table>

**Hoggart et al 2008:** "Testing one SNP at a time does not fully realise the potential of genome-wide association studies to identify multiple causal variants, which is a plausible scenario for many complex diseases."

**Marchini et al 2009:** "The inconclusive findings identified with this study reflect the status of the field of autism genetics and suggest that classical approaches such as linkage SNP association and CNV analysis and association analyses alone may not be sufficient to deal with the genetic and phenotypic heterogeneity seen in autism."
Multivariate and ML approaches

1. Regularization
   • **This study (method)**: l1/l2, n=12000 x p=1600 (mice), various encodings, Intra/interfamily effects fully managed by the experimental plan (DAP)
   • CMU SailingLab (W84): heterogeneous multitask.

2. Kernel-based
   • **This study (baseline)**: epsilon-SVR : R/Libsvm implementation, with linear loss function, Kernels: Gaussian, Linear, poly, custom kernels.
   • Gonzalez-Recio et al., 2008 & 2009: kernel regression with ‘ad hoc’ kernel n=3500 x p=400 samples (broilers)

3. Bayesian
   • **This study (reference)**: Lee et al., 2008, MCMC (Reversible Jump), Encoding: (-1,0,1); Dominance, Intra/interfamily effects
   • de los Campos et al., 2009: Bayesian regression coupled with LASSO, largest study n=11000 x p=1900 (mice), Strong pedigree effect
   • Gonzalez-Recio et al., 2009: Bayesian regression, Dominance effect and interaction terms included, Hypotheses on priors
I1-I2 Regularization

1. Basis

RLS regression model with embedded feature selection (De Mol et al 09)

• A variant of the elastic net model (Zou&Hastie 05)

• Optimization problem: \( \beta_{l1/2} = \arg \min_{\beta} \frac{1}{n} \| Y - X\beta \|_2^2 + \tau \| \beta \|_1 + \mu \| \beta \|_2^2 \)

\( \beta \): regression weights; \( Y \): observed output;
\( X \): input data matrix; \( \mu, \tau \): regularization parameters)

\( \mu \) and \( \tau \) modulate the selection of the features:
- \( \mu \) preserves correlation among selected features
- \( \tau \) enforces the sparsity of the solution

3. Algorithm

• Feature selection step: by Iterative Soft Thresholding

• RLS Step: At each step correction of the weight bias by a RLS regression on selected features (RLS parameter: \( \lambda \))

Note: Additive-only genetic model. But (Hill et al. 08) show that most variance is of an additive type even when dominance effects are present.

3. A New Implementation
   • l1-l2 with double optimization implemented in Python/NumPy, now a component of the mlpym package (https://mlpy.fbk.eu), using its functions for data import, handling and cross-validation. Parallelized for HPC. Extensively tested on 550K SNPs input features.

5. Previous Results
   • DeMol et al., 2009: classification from real predictions on gene expression data of leukaemia (n=72, p=7,000), lung cancer (n=181, p=12,000) and prostate cancer (n=102, p=12,000)
   • Fardin et al., 2009: a signature for hypoxia in gene expression data, neuroblastoma cell lines (n=18, p=50,000), classification from real predictions.

Fardin, P. et al. The l1-l2 regularization framework unmask the hypoxia signature hidden in the transcriptome of a set of heterogeneous neuroblastoma cell lines. BMC Genomics (2009).
Data Analysis Protocol: overview

PREPROCESSING

SVR PIPELINE
- Dataset
- Encoding
- Imputing
- Landscaping
- Parameter search
- Bootstrap validation

L1L2 PIPELINE
- Dev/val split
- K-fold CV
- Parameter search
- Learn on whole dev sets

C. Furlanello – Phenotype prediction from genotype data – NIPS09/MLCB09 - 10
Preprocessing

Encoding
- Linear regression approaches need a numerical representation for ternary SNP data (dominant homozygous – AA, heterozygous – Aa, or recessive homozygous – aa)
- Options tested:
  - 0, 1, 2 (biological interpretation: # of alleles deviating from dominant)
  - -1, 0, 1
  - relative frequency of each allelic class at each locus over the sample population
- No significant variation in predictive power between 0, 1, 2 and -1, 0, 1
- Predictions worsen with the frequency-based encoding
- Final choice: 0, 1, 2 representation (classical in literature)

Imputing
- Few missing data in the GSCAN dataset (4.7%): random imputation with probability equal to the relative frequency of each allele at that locus in the population.
L1L2 workflow

Protocol inspired by the MAQC-II project guidelines (Shi et al., 2008)

- Split dataset in **15 development / validation sets** (50% bootstrap **interfamily** re-sampling)
- For each experiment, run a **10-fold CV** on the development set
- Select optimal parameter set \((\mu^*, \tau^*, \lambda^*)\) according to average prediction performances (accuracy, instability) across the 10 test sets
  - **Accuracy**: MSE or \(R^2\)
  - **Instability**: **Canberra distance** (Jurman et al., 2008) of lists selected in the 10 test sets ranked by abs weights
- Learn the 15 development sets using \((\mu^*, \tau^*, \lambda^*)\)
- Assess results on the 15 validation sets

Shi, L. et al. Reproducible and reliable microarray results through quality control: Good laboratory proficiency and appropriate data analysis practices are essential. Curr Opin Biotechnol (2008)

Model selection

![Graph of model selection showing instability vs mean squared error for different experiments.

Legend:
- Solid square: mu = 0.0004
- Solid triangle: mu = 0.0010
- Solid circle: mu = 0.0025
- Red circle: tau = 0.4
- Green circle: tau = 1.0
- Blue circle: tau = 2.5
- Black square: lam = 1

Experiments:
- EXP-1 to EXP-15]
Prediction accuracy

- Regression R for SVR, L1L2 and the reference MCMC model for each phenotype
- Standard deviation in parentheses

<table>
<thead>
<tr>
<th>Method</th>
<th>CD8+</th>
<th>MCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVR</td>
<td>0.551 (0.05)</td>
<td>0.379 (0.06)</td>
</tr>
<tr>
<td>L1L2</td>
<td>0.559 (0.06)</td>
<td>0.323 (0.045)</td>
</tr>
<tr>
<td>MCMC - additive</td>
<td>0.51 (0.05)</td>
<td>0.33 (0.06)</td>
</tr>
<tr>
<td>MCMC – additive+dominance</td>
<td>0.56 (0.06)</td>
<td>0.33 (0.09)</td>
</tr>
</tbody>
</table>

**NB: Interfamily experiments.** Members of a family were assigned all either to the training or to the test set, thus avoiding information leakage due to very high genetic similarity between individuals in the same family.
Top-k Analysis

- We define **top-ranked SNPs** those in the top 10-percentile of the distribution of the absolute weights in at least 14/15 exps.
- It is well-known that correlated, functionally important variables may be discarded or poorly ranked in feature selection methods.
- We define **top-correlated SNPs** those having an absolute Pearson correlation coefficient with at least one top-ranked SNP higher than 0.8.
- In our analyses, top-correlated SNPs are:

  (a) Clustered around the reference top-ranked SNPs: the median distance smaller for higher correlation levels
  (b) Highly ranked on average: the median absolute regression weight larger for higher correlation levels

For CD8: from 41 top-ranked to 182 top-correlated SNPs
Pooling top-ranked and top-correlated SNPs, both methods select many SNPs in candidate loci previously identified by GWAS (Valdar et al., 2006)

Conclusions

Fitting quantitative phenotypes from high-throughput data

- **L1/L2 with double optimization** can be used for building regression models and extract biomarkers in large scale GWAS studies.
- Apply carefully within an experimental Data Analysis protocol (DAP) to avoid bias. Still one module within a pipeline: each component a source of variability – see MAQC-II studies
- Use stability and biological hypothesis to detect potential markers

**Applications:**
- Effective in complex common diseases with high individual variability (e.g. neurogenomics): trajectories.
- Parameter estimation for infectious disease modeling
- Gene expression as a quantitative trait: identify traits predictive of gene networks and gene expression regulation (eQTL)
- Survival analysis
- May be used for classification purposes where classifiers fail
- GSCAN dataset
• GSCAN dataset
  most numerous families
SVR – Model selection
Protocol inspired from (Lee et al., 2008)

**Landscaping**
- For each parameter set \((C, \varepsilon, \sigma)\), train the SVM on 10 training sets (50% bootstrap re-sampling)
- **Interfamily sampling**: members of a family were assigned all either to the training or to the test set
  - avoid information leakage due to very high genetic similarity among individuals in the same family.
- Parameter space explored with a grid search
- Select the optimal parameter set \((C^*, \varepsilon^*, \sigma^*)\) as the one having the best average \(R^2\) on the test sets

**Bootstrap validation**
- Train the SVM with the optimal parameter set \((C^*, \varepsilon^*, \sigma^*)\) on 15 training sets 50% bootstrap re-sampling
- Evaluate predictions on the test sets