



Improving the prediction of regulatory SNPs using functional information

Yiqiang Zhao¹ , Matt Mort² , David Cooper² & Sean D. Mooney¹

1. Center for Computational Biology and Bioinformatics,
Department of Medical and Molecular Genetics,
Indiana University School of Medicine,
410 W. 10th Street, Suite 5000, Indianapolis, IN
46202, USA
Email: mooney@compbio.iupui.edu

2. Institute of Medical Genetics, Cardiff University,
Heath Park, Cardiff, CF14 4XN, UK

Challenge: Can regulatory SNPs be predicted using sequence and bioinformatics? How important are relative genomic features and the annotations on nearby genes in making predictions? Using a training set of experimentally validated disease-causing regulatory SNPs from the Human Gene Mutation Database (HGMD) and background SNPs from dbSNP, we set out to answer these, and other questions.



Definition of cis-acting SNPs



We found 338 regulatory SNPs (rSNPs) and 183,500 background SNPs (bSNPs) in this region

Features Investigated

From SNP

- Distance to TSS
- SNP diversity
- Derived allele frequency
- Flanking GC content (21)
- Flanking conservation score (21)
- In CpG Island ?
- In Enhancer ?
- In RNA Polymerase II Enriched Region ?
- In Nuclease Hypersensitive Sites ?
- In Conserved Non-coding Sequences ?

From Gene

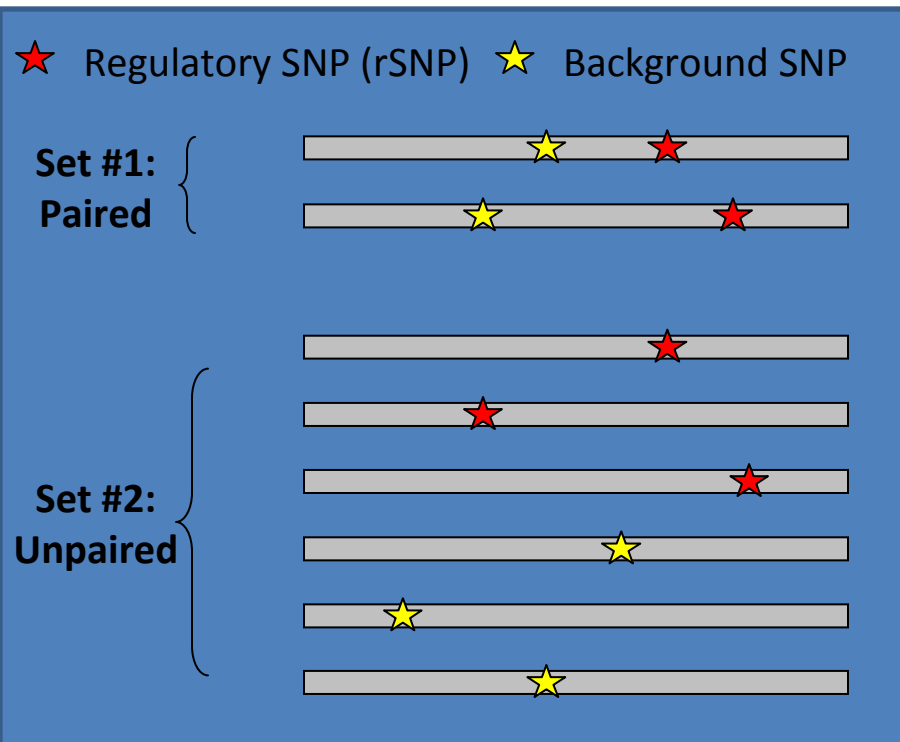
- Gene expression (Mean/Max/CV)
- Codon usage (Fop/ENC)
- Functional complexity (MF/BP)
- Protein-Protein interaction complexity



Training sets

Two training sets were defined, Paired and Unpaired. The paired set is designed to test how good SNP based features are for rSNP prediction.

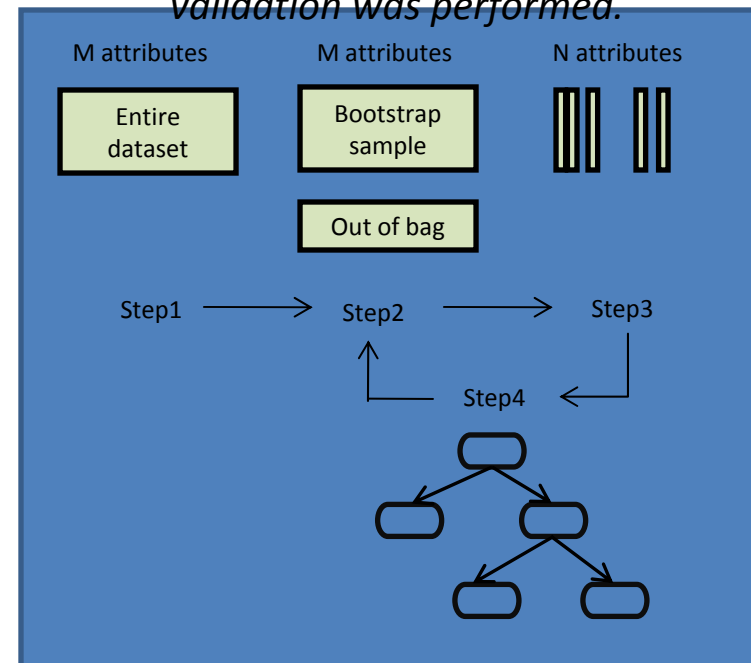
The unpaired set is designed to test how good gene based features and SNP based features are for rSNP prediction.



Evaluation Approach

We tried SVM / Random Forest / Bayes Network.

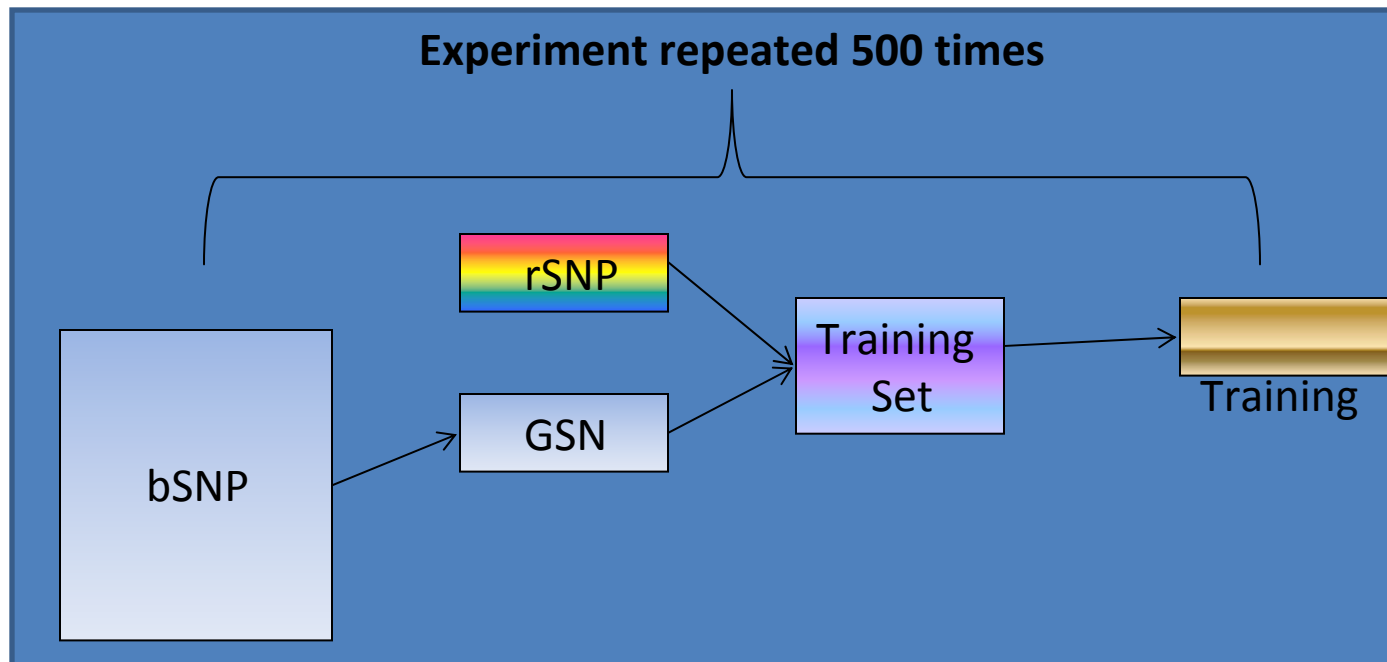
The RF algorithm, implemented in WEKA (Waikato Environment for Knowledge Analysis), was chosen with the generation of 100 trees because of better performance. 10-fold cross-validation was performed.





Statistical Analysis

*The training set is highly imbalanced (**bSNPs** \gg **rSNPs**). To collect statistics of our approach, we collected 500 samples of a gold standard negative (GSN) randomly selected from the extensive bSNP set. Classification were performed against 500 training sets obtained using the GSN set and the rSNPs.*

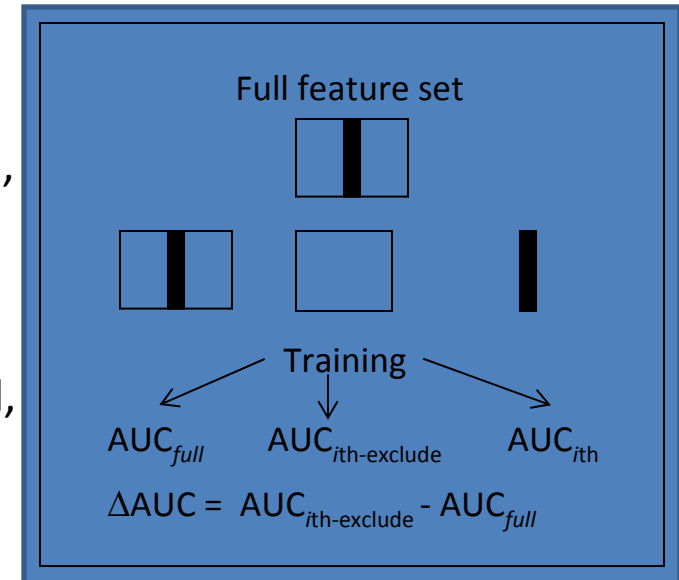


For the overall performance of the model, we reported the average value for AUC, specificity/sensitivity or other measurements.



Four approaches for evaluation of the relative importance of features

1. information gain (univariate filter),
2. correlation-based feature selection (multivariate filter) ,
3. individual Area under the Curve (AUC) of a Receiver Operator Characteristics (ROC) Plot for each feature under the same classification method as the whole model,
4. AUC index



For the i^{th} feature, we calculate the decrease of the AUC from the model with the full feature set to the model with the feature set of which the i^{th} feature is excluded. The value indicates the importance of the i^{th} feature.

Taking those methods together, we can find feature dependence, and indentify truly important features.



Results and Discussion

R1, For the unpaired dataset, we obtain an **AUC (0.889+-0.011)**, sensitivity(0.843+-0.015) and specificity(0.755+-0.019) for overall performance.

R2, For the unpaired dataset, we found the best features are: **Functional complexity, maximum value of gene expression, protein-protein interaction complexity and distance to TSS**. It is interesting that most of the best features come from associated gene.

R3, For the paired dataset, that all the **gene features were not included**, the overall performances are only **AUC(0.735+-0.017)**, sensitivity(0.670+-0.024) and specificity(0.667+-0.020). Few SNP based features are important except distance to TSS.

P1, The randomly selected background dataset in this study is problematic and likely contains regulatory SNPs.

P2, Functional annotation is important, but it is perhaps not surprising as regulatory SNPs likely come from functionally well characterized genes.