Applications in Computational Biology

Overfitting and Generalization: Deleteriousness Prediction

Deleteriousness Prediction

Assessing the impact of missense variants

- Given the availability of more and more sequencing data on individual patients, one fundamental question to ask is: Is a variant at a particular position in the genome deleterious?

Deleteriousness Prediction

Assessing the impact of missense variants

- Given the availability of more and more sequencing data on individual patients, one fundamental question to ask is: Is a variant at a particular position in the genome deleterious?
- Even when restricting ourselves to missense variants that cause an amino acid change, one is usually left with tens of thousands of these variants.

Deleteriousness Prediction

Assessing the impact of missense variants

- Given the availability of more and more sequencing data on individual patients, one fundamental question to ask is: Is a variant at a particular position in the genome deleterious?
- Even when restricting ourselves to missense variants that cause an amino acid change, one is usually left with tens of thousands of these variants.
- This motivated the development of a large number of computational tools to predict the deleteriousness of missense variants.

Deleteriousness Prediction

Assessing the impact of missense variants

- A multitude of definitions of deleteriousness, datasets and algorithms for deleteriousness prediction (SIFT, POLYPHEN, MUTATIONTASTER, LRT, GERP, FatHMM) exists.

Deleteriousness Prediction

Assessing the impact of missense variants

- A multitude of definitions of deleteriousness, datasets and algorithms for deleteriousness prediction (SIFT, POLYPHEN, MUTATIONTASTER, LRT, GERP, FatHMM) exists.
- For the practitioner, it is extremely hard to choose among this plethora of approaches.

Deleteriousness Prediction

Assessing the impact of missense variants

- A multitude of definitions of deleteriousness, datasets and algorithms for deleteriousness prediction (SIFT, POLYPHEN, MUTATIONTASTER, LRT, GERP, FatHMM) exists.
- For the practitioner, it is extremely hard to choose among this plethora of approaches.
- Our goal: To provide the cleanest and most comprehensive comparative evaluation of different deleteriousness predictors on a wide variety of datasets (Grimm et al., Human Mutation 2015).

Deleteriousness Prediction

Assessing the impact of missense variants

- A multitude of definitions of deleteriousness, datasets and algorithms for deleteriousness prediction (SIFT, POLYPHEN, MUTATIONTASTER, LRT, GERP, FatHMM) exists.
- For the practitioner, it is extremely hard to choose among this plethora of approaches.
- Our goal: To provide the cleanest and most comprehensive comparative evaluation of different deleteriousness predictors on a wide variety of datasets (Grimm et al., Human Mutation 2015).
- We compared 10 methods on 5 widely used datasets.

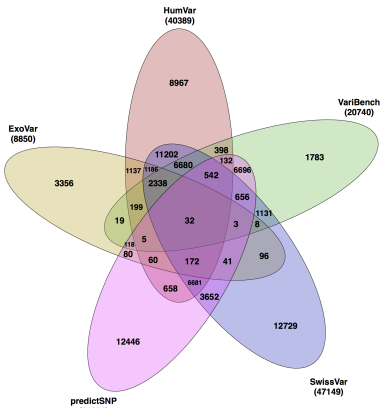
Deleteriousness Prediction

Two Major Types of Circularities

- **Type 1 Circularity**: The common benchmark datasets used for training and testing tools overlap to a large degree.
- **Type 2 Circularity**: Most proteins contain only deleterious or only neutral variants. A naive majority class vote within a protein gives (artificially) excellent results.

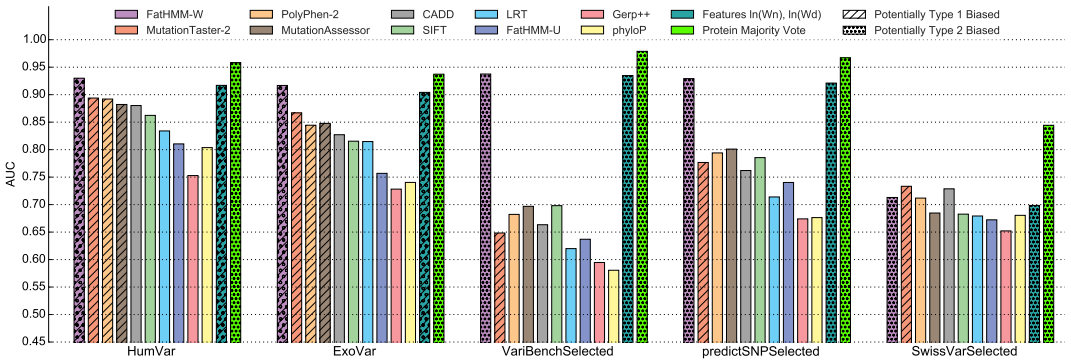
Deleteriousness Prediction

Type 1 Circularity



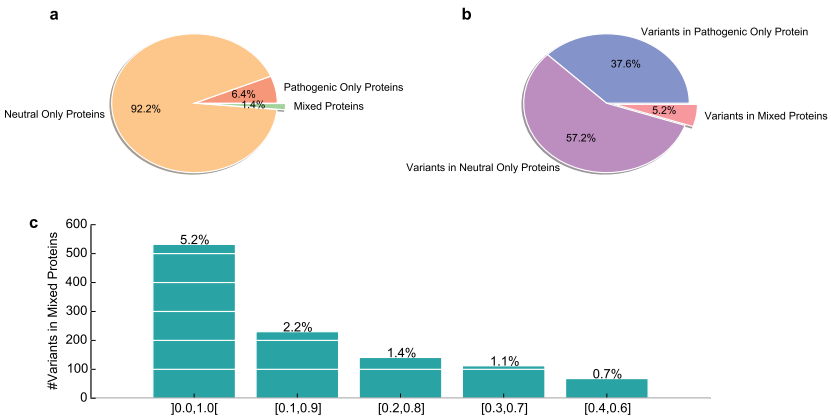
Deleteriousness Prediction

Comparative Evaluation



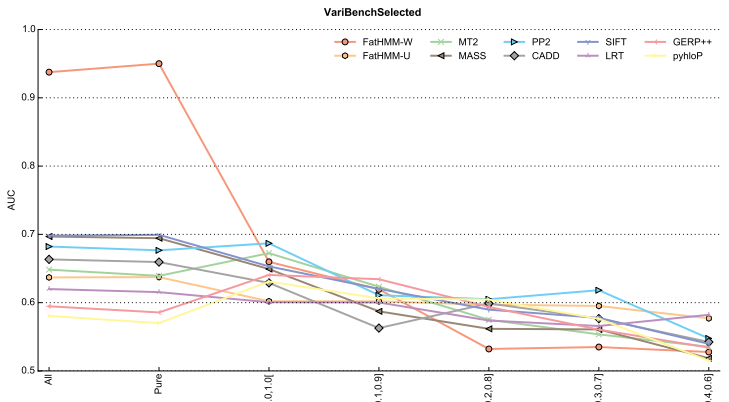
Deleteriousness Prediction

Fraction of 'mixed' proteins in VariBenchSelected



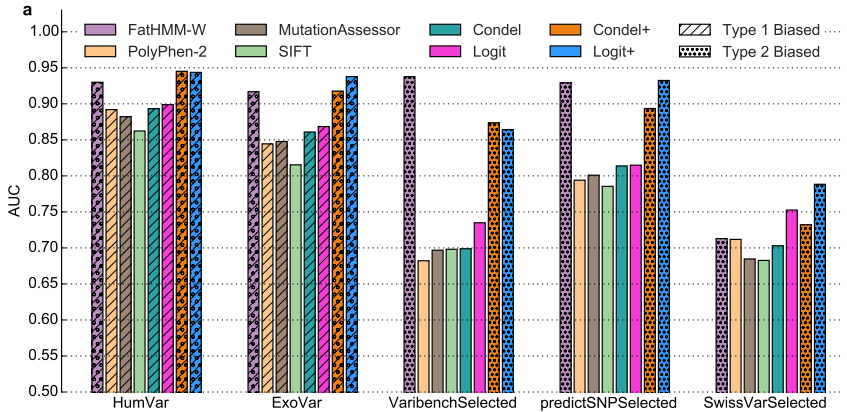
Deleteriousness Prediction

Type 2 Circularity: Predictive performance versus neutral/deleterious ratio



Deleteriousness Prediction

Impact of Circularity on Combining Predictors



Deleteriousness Prediction

Conclusions

- A comparative evaluation of deleteriousness prediction tools is complicated by two types of circularity:

Deleteriousness Prediction

Conclusions

- A comparative evaluation of deleteriousness prediction tools is complicated by two types of circularity:
- Type 1 circularity can only be avoided by cleanly separating training and test dataset.

Deleteriousness Prediction

Conclusions

- A comparative evaluation of deleteriousness prediction tools is complicated by two types of circularity:
- Type 1 circularity can only be avoided by cleanly separating training and test dataset.
- Type 2 circularity can only be avoided by stratifiying training and test dataset with respect to protein membership.

Deleteriousness Prediction

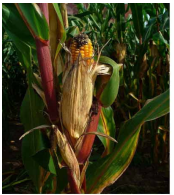
Conclusions

- A comparative evaluation of deleteriousness prediction tools is complicated by two types of circularity:
- Type 1 circularity can only be avoided by cleanly separating training and test dataset.
- Type 2 circularity can only be avoided by stratifiying training and test dataset with respect to protein membership.
- A severe complication in practice is that many authors only publish their prediction tool, but not the features used to train the predictors. Retraining the models to ensure a clean, circularity-free prediction is practically impossible.

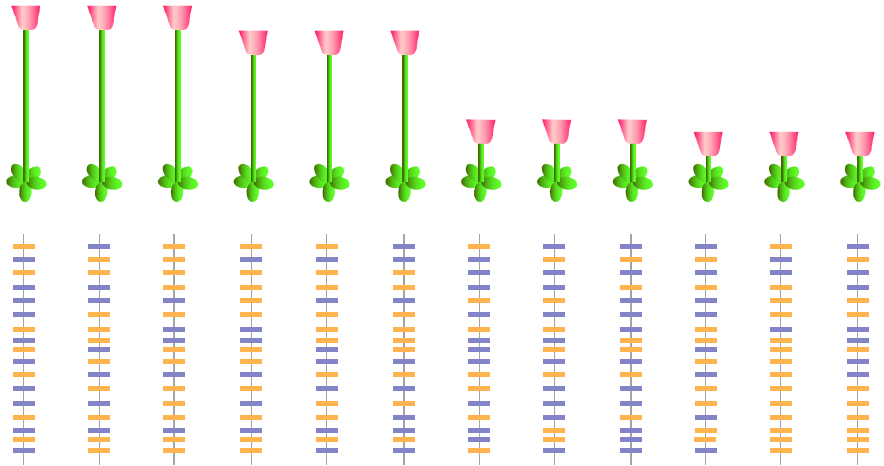
Phenotype Prediction and Epistasis

Classic and New Questions

- Genetics
  - How does genotypic variation lead to phenotypic variation?
  - Can we predict phenotypes based on the genotype of an individual?
- Recent progress
  - Genotypes can be determined at an unprecedented level of detail
  - Phenotypes can be recorded in an automated manner



Genome-wide Association Mapping



by courtesy of D. Weigel

Prediction of Complex Phenotypes

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive

News & Comment > News > 2015 > October > Article

NATURE | NEWS

Epigenetic 'tags' linked to homosexuality in men

Twin study reveals five DNA markers that are associated with sexual orientation.

Sara Reardon

08 October 2015 | Updated: 12 October 2015

Rights & Permissions

Prediction of Complex Phenotypes

Claim that sexual orientation can be predicted (Source: Science— DOI: 10.1126/science.aad4686)

- At ASHG 2015, the Vilain lab from UCLA claimed that certain methylation patterns in the human genome are predictive for sexual orientation.
- Tuck Ngun from this lab considered methylation patterns at 140,000 regions in the DNA of 37 pairs of male identical twins who were discordant and 10 pairs who were both homosexual.
- They reported to have identified five regions in the genome where the methylation pattern appears very closely linked to sexual orientation.
- The team reached 70% prediction accuracy when splitting the discordant twin pairs into 2 groups, one for training, one for testing.

Prediction of Complex Phenotypes

Criticisms (by Ed Yong, The Atlantic)

- Sample size is very small.

Prediction of Complex Phenotypes

Criticisms (by Ed Yong, The Atlantic)

- Sample size is very small.
  - Ngun: "Yes, we were underpowered."

Prediction of Complex Phenotypes

Criticisms (by Ed Yong, The Atlantic)

- Sample size is very small.
  - Ngun: "Yes, we were underpowered."
- Overfitting on the test set.

Prediction of Complex Phenotypes

Criticisms (by Ed Yong, The Atlantic)

- Sample size is very small.
  - Ngun: "Yes, we were underpowered."
- Overfitting on the test set.
  - Ngun: "All models (from the very first to the final one) were built using JUST the training data.... If performance was unsatisfactory, we remade the model by selecting a different set of predictors/features/data based on information from the TRAINING set and then reevaluating on the test set."

Prediction of Complex Phenotypes

Criticisms (by Ed Yong, The Atlantic)

- No correction for multiple testing

Prediction of Complex Phenotypes

Criticisms (by Ed Yong, The Atlantic)

- No correction for multiple testing
  - Ngun: "We are not testing whether each of the 6000 marks/loci are significantly associated with sexual orientation. If we had done that, multiple testing correction would have certainly been warranted. But we didn't. The single test we did was to ask whether the final model we had built was performing better than random guessing."

Prediction of Complex Phenotypes

Lessons we should learn

1. Predicting complex traits from high-dimensional molecular data is (becoming) a reality. Low sample size is still an important obstacle.

2. It is important to build predictors that generalize to unseen data and to avoid overfitting.

3. When searching high-dimensional spaces for higher-order associations, multiple testing correction is an enormous problem.

Prediction of Complex Phenotypes

Lessons we should learn

1. Predicting complex traits from high-dimensional molecular data is (becoming) a reality. Low sample size is still an important obstacle.
   - Roqueiro, Witteveen et al. Bioinformatics/ISMB, 2015; Sugiyama and Borgwardt, NIPS 2015
2. It is important to build predictors that generalize to unseen data and to avoid overfitting.
   - Grimm et al., Human Mutation 2015
3. When searching high-dimensional spaces for higher-order associations, multiple testing correction is an enormous problem.
   - Sugiyama et al., SDM 2015; Llinares-Lopez et al., Bioinformatics/ISMB 2015, KDD 2015

Phenotype Prediction

Arabidopsis phenotypes (99-199 plants, 250k SNPs, Atwell et al., 2010)

Phenotype	$\mathrm{AUC}_{\mathrm{SVM}}$
Chlorosis at 22°C	0.629 ± 0.003
Anthocyanin at 16°C	0.569 ± 0.003
Anthocyanin at 22°C	0.609 ± 0.003
Leaf Roll at 10°C	0.696 ± 0.002
Leaf Roll at 22°C	0.587 ± 0.004

Phenotype Prediction

Why is there room for improvement?

- We assume additive effects of SNPs, ignore gene-gene interactions and gene-environment interactions.
- We ignore population structure, that is systematic ancestry differences of cases and controls.

Epistasis - what it means I (Cordell, 2002)

Bateson's masking effect model

- Bateson defines epistasis as a masking effect, whereby a variant or allele at one locus prevents the variant at another locus from manifesting its effect.

Genotype at locus B/G	gg	gG	GG
bb	White	Grey	Grey
bB	Black	Grey	Grey
BB	Black	Grey	Grey

Table: Example of phenotypes (e.g. hair colour) obtained from different genotypes at two loci interacting epistatically under Bateson's (1909) definition of epistasis.

Epistasis - what it means II (Cordell, 2002)

Epistasis in a general sense

Genotype at locus A/B	bb	bB	BB
aa	0	0	0
aA	0	1	1
AA	0	1	1

Table: Example of penetrance table for two loci interacting epistatically in a general sense

Epistasis - what it means III (Cordell, 2002)

Genetic heterogeneity model

Genotype at locus A/B	bb	bB	BB
aa	0	0	1
aA	0	0	1
AA	1	1	1

Table: Example of penetrance table for two loci acting together in a heterogeneity model

Epistasis - what it means IV

Regression model

- Most popular statistical definition:

$$\mathbf{y} = \theta_i \mathbf{x}_i + \theta_j \mathbf{x}_j + \theta_{(i,j)} \mathbf{x}_i \odot \mathbf{x}_j + \epsilon \tag{1}$$

- Test whether $\theta_{(i,j)}$ is significantly different from zero; rank pairs by the resulting p-value.
- Other common measures of association include e.g. the F-statistics and Pearson's correlation coefficient.

Epistasis - what it means V (Marchini et al., 2005)

Model 1: Multiplicative interaction within and between loci

Locus A/B	bb	bB	BB
aa	α	$\alpha(1+\theta_2)$	$\alpha(1+\theta_2)^2$
aA	$\alpha(1+\theta_1)$	$\alpha(1+\theta_1)(1+\theta_2)$	$\alpha(1+\theta_1)(1+\theta_2)^2$
AA	$\alpha(1+\theta_1)^2$	$\alpha(1+\theta_1)^2(1+\theta_2)$	$\alpha(1+\theta_1)^2(1+\theta_2)^2$

Table: The odds increase multiplicatively with genotype both within and between loci.

Epistasis - what it means VI (Marchini et al., 2005)

Model 2: Two-locus interaction with multiplicative effects

Locus A/B	bb	bB	BB
aa	α	α	α
aA	α	$\alpha(1+\theta)$	$\alpha(1+\theta)^2$
AA	α	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^4$

Table: In this model, the odds have a baseline value (α) unless both loci have at least one disease-associated allele. After that, the odds increase multiplicatively within and between genotypes.

Epistasis - what it means VII (Marchini et al., 2005)

Model 3: Two-locus interaction with threshold effects

Locus A/B	bb	bB	BB
aa	α	α	α
aA	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$
AA	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$

Table: In this model, the odds have a baseline value (α) unless both loci have at least one disease-associated allele. In this case, the odds-ratio is $\alpha(1+\theta)$.

Epistasis - what it means VIII (Marchini et al., 2005)

a

Multiplicative within and between loci

	bb	Bb	BB
aa	α	$\alpha(1+\theta_2)$	$\alpha(1-\theta_2)^2$
Aa	$\alpha(1+\theta_1)$	$\alpha(1+\theta_1)(1+\theta_2)$	$\alpha(1+\theta_1)(1+\theta_2)^2$
AA	$\alpha(1+\theta_1)^2$	$\alpha(1+\theta_1)^2(1+\theta_2)$	$\alpha(1+\theta_1)^2(1+\theta_2)^2$

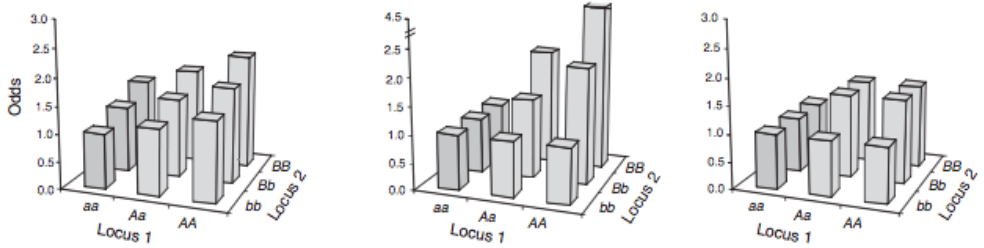
Two-locus interaction multiplicative effects

	bb	Bb	BB
aa	α	α	α
Aa	α	$\alpha(1+\theta)$	$\alpha(1+\theta)^2$
AA	α	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^4$

Two-locus interaction threshold effects

	bb	Bb	BB
aa	α	α	α
Aa	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$
AA	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$

b



Impact of Epistasis

Examples of Epistasis

- Epistasis is conjectured to be one source of missing heritability (Manolio et al., 2009)
- Genetic interactions are one indicator that epistasis is a major factor in the genotype-phenotype relationship (e.g. Boone et al., 2007)
- Pairs of genes have been reported to affect complex diseases such as breast cancer (Ashworth et al., 2011):
  - Loss of either BRCA1 or BRCA2 tumor suppressor gene function in cells triggers a cell-cycle arrest at the G2/M checkpoint that can be suppressed by the inactivation of P53 (Connor et al., 1997 and Liu et al., 2007).
  - Loss of VHL (Von Hippel-Lindau tumor suppressor) function normally causes cellular senescence, but inactivation of a second tumor suppressor, RB (Retinoblastoma), can suppress this process (Young et al., 2008).

Bottlenecks in two-locus mapping

Scale of the problem

- Typical datasets include order $10^5 - 10^7$ SNPs.
- Hence we have to consider order $10^{10} - 10^{14}$ SNP pairs.
- Enormous multiple hypothesis testing problem.
- Enormous computational runtime problem.

Common approaches in the literature

Exhaustive enumeration

- Only with special hardware such as GPU implementations: EPIBLASTER (Kam-Thong et al., EJHG 2010)

Filtering approaches

- Statistical criterion, e.g. SNPs with large main effect (Zhang et al., 2007)
- Biological criterion, e.g. underlying PPI (Emily et al., 2009)

Index structure approaches

- fastANOVA, branch-and-bound on SNPs (Zhang et al., 2008)
- TEAM, efficient updates of contingency tables (Zhang et al., 2010)