eXtasy simplified - towards opening the black box

Dusan Popovic*, Alejandro Sifrim*, Yves Moreau and Bart De Moor
Department of Electrical Engineering (ESAT), STADIUS - iMinds Future Health Department
KU Leuven
Leuven, Belgium
* These authors contributed equally to this work

***Abstract*—Exome sequencing remarkably simplifies the search for mutations causing rare monogenic disorders. Still, due to a big number of potential candidate variants, computational methods are needed to facilitate this process. Recently, an algorithm based on genomic data fusion has been proposed in this context (eXtasy), which exhibits highly competitive performances among the state of the art methods. Nonetheless, being based on a Random Forest classifier, its core model is characterized by a prohibitive size, slow execution speed and difficulties associated with gaining insights in the decision-making process. Here we propose a simplification of the original eXtasy algorithm that retains superior ranking capability of former without suffering from the both high complexity and low interpretability.**

Keywords—eXtasy, variant prioritization, genomic data fusion, rare genetic disorders, interpretable model, random forest, decision trees, hybrid sequential system

I. INTRODUCTION

The discovery of mutations involved in the etiology of genetic disorders has been significantly accelerated in the recent years due to advances in massively parallel sequencing technologies. Amongst other approaches, sequencing of the exome (i.e. protein-coding region of a gene) in particular has been shown to be an efficient strategy for identification of causes of rare monogenic disorders [1]. However, a single exome typically harbors approximately 8000 non-synonimus variants, while DNA samples from affected individuals are often scarce.

Even after aggressive filtering of the exome against nSNVs (nonsynonymous single nucleotide variants) and loss-of-function mutations that are present in healthy individuals, roughly 200 candidate mutations still remain [2,3]. To overcome this problem, several computational methods for variant prioritization have been developed so far [4-10]. These methods mostly rely on evolutionary, biochemical and structural properties of the mutation in question. Recently, a new approach based on genomic data fusion has been proposed (eXtasy [11]) that displays vastly superior performance compared to the rest of the current state-of-the art methods.

eXtasy combines phenotypic-specific information, haploinsufficiency prediction and deleteriousness prediction scores of mutations to deliver variants ranking. Firstly, given the list of phenotypes associated with a disease of interest, the mutated genes are scored based on their similarity with known disease genes [12]. Secondly, resulting vectors of scores obtained per each phenotype/variant combination are augmented with several additional features : haploinsufficiency prediction [13], state-of-the art deleteriousness prediction scores [4-8] and conservation scores across several species [14,15]. Finally, this information is used to classify a mutation as disease-causing and to assign a score to it. As several values are obtained for each mutation (in context of associated phenotypes), only the maximum score is retained.

The core of the eXtasy algorithm is the Random Forest [16] classifier trained using the data corresponding to known disease-causing mutations (positives) and to rare mutations found in healthy individuals (negatives). The Random Forest is an ensemble method, so it combines prediction of many base models to reach a final decision. By doing so, it stabilizes the variance of low-bias/high-variance base learners (i.e. decision trees), improving overall performance. Randomness is injected into the algorithm by building each tree on a different bootstrap from data and by choosing each decision tree split from a random subset of all variables. This procedure increases the diversity of an ensemble, which is crucial for triggering the previously discussed effect. In addition, it has been shown that ensemble approaches have a smaller number of generalized degrees of freedom than some supposedly simple algorithms [17], making their predictions quite robust against small perturbations in the data.

However, these advantages come with the price of increased model complexity and reduced interpretability. The complexity translates to limitations associated with memory constraints and execution speed in practical applications. For example, the eXtasy stand-alone application comes with the model that is 100 Mb large (for 500 trees) and takes approximately 2 minutes to score all mutations from a single exome using a standard PC. Furthermore, being based on a Random Forest algorithm, the decision process of eXtasy takes place in a “black box”, which prevents users from gaining insight into it.

An attempt has been made to address this issue in the original study by using the Random Forest feature importance measures [16]. Yet, these are extensively criticized for displaying strange statistical properties [18]; and for being highly biased in presence of correlation between variables [19,20] or when data of mixed types is used [21]. Therefore, in this work we propose a straightforward simplification of the eXtasy model, aiming to address previously risen concerns. In addition, we demonstrate that the described model achieves similar performance as the original formulation, while being more interpretable, having a smaller memory footprint and faster prediction speed.

II. MATERIALS AND METHODS

A. The method

The basic idea behind our approach is usage of a simple classifier to model the decision boundary that is implicit in a more complex one (i.e. Random Forest). In this setup, two algorithms are organized sequentially, where the first learns from the data and then acts as a "teacher" to the second. Provided that an inductive bias of a simple method allows the construction of a frontier of the same type as that generated by the complex method, this procedure ideally results in preserving advantageous performances of the later in a more compact and comprehensible form. The described paradigm was already successfully employed in several different settings; amongst which to extract symbolic rules from neural network [22,23] and decision trees from bagged ensemble [24]. Our approach is somewhat similar to that in [24], with the exception of a different synthetic data generation scheme.

The general outlook of the algorithm is provided on Fig. 1. First, a Random Forest classifier is trained using class-balanced data, identical to the way the eXtasy model was built [11]. Then, artificial data is generated by sampling from the empirical distributions of the features. In particular, values of predictors are shuffled across examples. This randomization is repeated ten times to create a large data set with distributional properties similar to that of the real data. Empirical sampling has been chosen over uniform and model-based sampling due to a high skew of predictor values and a lack of fit to any standard distribution. Subsequently, simulated data is subjected to classification by the previously trained Random Forest, assigning labels to each generated example. Finally, the new data set is used to build a fully-grown decision tree. This model is deliberately not pruned, as the goal is to "overfit" the already generalized decision boundary of the Random Forest.

A single decision tree is used as it can represent function of arbitrary complexity over the instance space, analogously to the Random Forest. Also, the interpretation of a single tree decision process is quite straightforward compared to that of an ensemble system. However, it might not be immediately obvious how the splits of such a derived tree differ from that inferred directly from real data, and consequently, how positions of these in a tree indicate importance of corresponding variables. A graphical illustration of this difference on a toy example is displayed on Fig. 2.

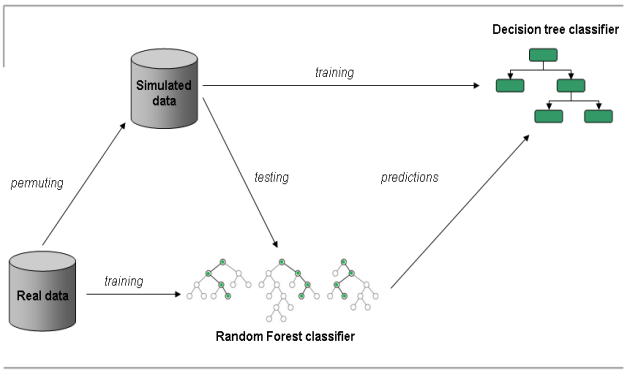

Fig. 1. Sequential algorithm for training the simplified eXtasy. The real data is used as an input for training the Random Forest classifier (as in the classical eXtasy) and to generate artificiall data that follows empirical distribution of the former. The output of the first classifier together with corresponding simulated examples is passed to the decision tree for the second round of training.

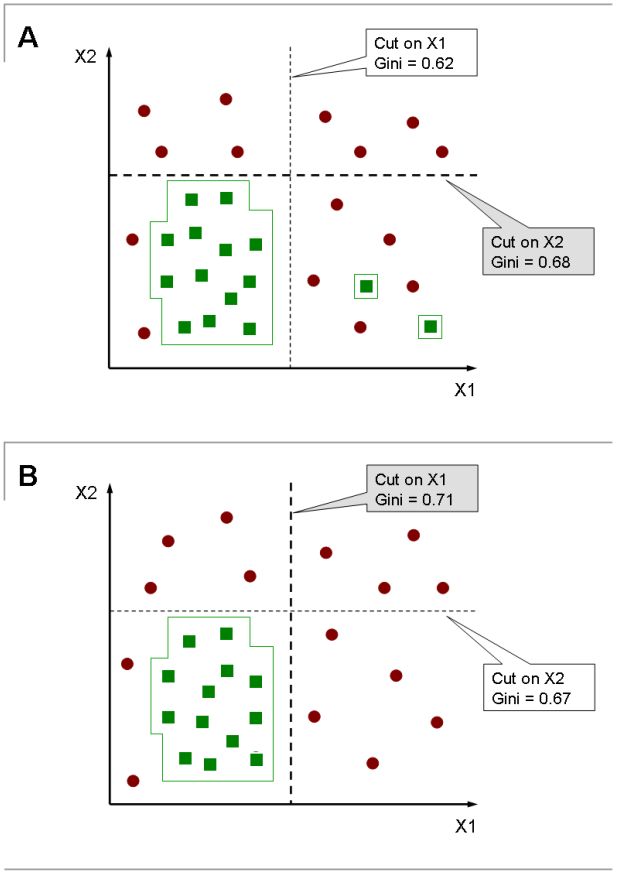

Fig. 2. An example of two different initial decision tree splits for the same classification problem. Squares and dots represent two classes; solid lines enclose implicit general concept while two empty squares indicate outliers. A) Variable X2 is chosen for the first split due to a higher Gini index. B) Data from the same distribution, but without outliers. Now the X1 is chosen for the split. In both cases the decision boundary can be learned by a single decision tree, but the ordering of variables changes as well as their relative importance in describing the underlying concept.

Briefly, decision tree induction algorithms are based on greedy, top-down recursive partitioning, thus their first (or any other) split might not be the most optimal with respect to the boundary constructed by the Random Forest. For example, a few outliers can render one variable more informative than others at a moment of choosing the optimal (current) split; even subsequent nodes generated using these instances would be pruned off later in a process. This effectively disregards them as non-informative during the validation, yet they bias the choice of variables during the tree induction phase. Conversely, outliers are routinely averaged out by bagging when the Random Forest decision boundary is build; thus they can not influence order in which variables are selected when a decision tree is used to model this already learned underlying concept.

The proposed algorithm is tested against the original eXtasy formulation and a classical decision tree classifier in the following manner – the whole dataset has been divided into training and testing partitions, with two-thirds of the total number of genes in the first, and one-third in the second part. Note that one gene contains many variants, while each variant has several corresponding data records (i.e. phenotypes). The validation scheme relies on stratification of the data on the highest level of granularity to prevent algorithms from overfitting gene-level information (i.e. to learn to recognize particular genes). After splitting, the class distribution of a training part is balanced, followed by training of all three methods on the same examples. In addition, the identical Random Forest classifier has been used both to access performance of the standard eXtasy and to generate synthetic data for the simplified version. To stabilize performance metrics, this process has been repeated one hundred times, randomizing the data division each time.

B. The data

The data set consists of two classes of mutations: Mendelian disease-causing variants and rare mutations present in healthy individuals as controls. For the disease-causing variants we obtained 24,454 nSNVs from the Human Gene Mutation Database (HGMD) associated in 1142 different Human Phenotype Ontology (HPO) terms. The HGMD is a database of expertly curated disease mutations published in scientific literature. Control variants were obtained from two different sources: the 1000 Genomes Project [25] and inhouse sequenced exomes (n=68) of healthy individuals. We selected nonsynonymous variants from the 1000 Genomes Project with a minor allele frequency lower than 1% (n=257556).

For the inhouse variants we selected variants not present in any publicly available variation repository (1000Genomes, dbSNP, NHLBI Exome Variant Server) and imposed a quality criterium of being sequenced at a depth of at least 20 (n=25429). For each of the phenotypes in the disease-variant set we sampled 500 variants from the pool of control variants and assigned them that given phenotype. Although the rare variants in our control sets could potentially have a functional impact (and thus have a lower frequency due to evolutionary negative selective pressure), it is safe to assume that it is unlikely that they would contribute to the randomly selected phenotype.

For each of the phenotypes in our data set we performed an Endeavour gene prioritization [12]. Training of Endeavour has been preformed using known disease-associated genes from the Online Mendelian Inheritance in Man database. We appended the predicted Endeavour scores to each variant-phenotype combination based on their respective gene and phenotype. Haploinsufficiency [13], conservation, deleteriousness prediction scores (SIFT, Polyphen2, MutationTaster, LRT obtained from dbNSFP v1.3 [26] and CAROL [8] score were also appended.

III. RESULTS AND DISCUSSION

Table 1. lists values of various performance metrics, together with standard deviations, for three methods under the study; as recommended in [27] for this type of benchmarks. Fig. 3 displays corresponding Receiver operating characteristic (ROC) curves. It is immediately apparent that the simplified version of eXtasy reaches almost the same performances as the original one across all indicators of interest; while the simple

TABLE I.

Metrics	**Method**		
	eXtasy	***eXtasy simplified***	***decision tree***
Accuracy	0.9404 (0.0045)	0.9348 (0.0052)	0.8743 (0.0076)
Sensitivity	0.8758 (0.0268)	0.8617 (0.0329)	0.8688 (0.0328)
Specificity	0.9496 (0.0057)	0.9450 (0.0067)	0.8748 (0.0104)
PPV	0.7158 (0.0298)	0.6946 (0.0328)	0.5019 (0.0348)
NPV	0.9816 (0.0031)	0.9795 (0.0035)	0.9789 (0.0048)
MCC	0.7586 (0.0245)	0.7373 (0.0293)	0.5977 (0.0299)

Average values of six performance measures obtained by testining the three methods : standard eXtasy (Random forest), simplified eXtasy and the decision tree built on the eXtasy data. PPV stands for the positive predictive value (precision), NPV for the negative predictive value and MCC for the Matthews correlation coefficient. The corresponding values of standard deviations are enclosed between brackets.

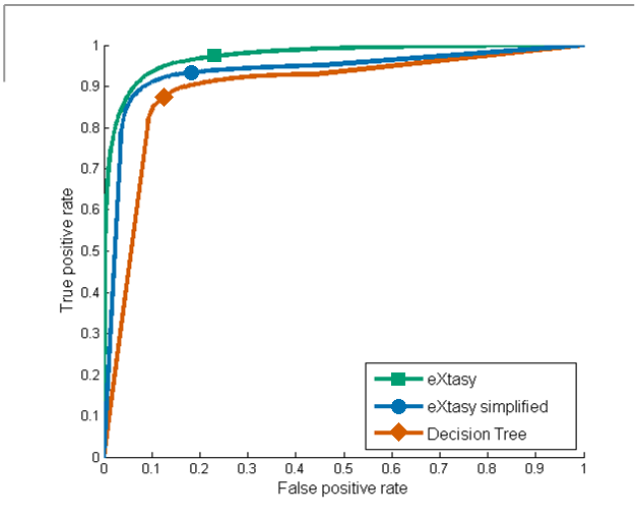

Fig. 3. ROC curves obtained by applying the three classifiers : standard eXtasy (Random Forest), simplified eXtasy and the decision tree built on the eXtasy data

decision tree does not - analogously to what was observed during the initial benchmark [11]. In ROC space though, the simplified version exhibits somewhat reduced discriminatory capability in certain regions.

However, in its operating point (i.e. a decision tree threshold) it behaves almost the same as the original eXtasy model. This distinction might be an artifact of a method used to estimate posterior probabilities of the decision tree outputs. Indeed, the ROC curve of the simplified eXtasy is rather straight in the region of low false positive rate, indicating that the corresponding scores might be too rough compared to that of the Random Forest. Thus, a proper calibration might help in overcoming this effect. Alternatively, if only the importance of variables for prediction is of interest, different trees can be constructed using the simulated data corresponding to various thresholds of the Random Forest. In this way, contribution of predictors to certain performance metrics (ex. precision) can be analyzed along the operating range of the classifier.

In terms of size, the simplified eXtasy considerably exceeds the decision tree constructed using the real data, as it contains approximately nine times more nodes than the latter. This is not surprising, as no pruning took place in this case, in contrast to a usual decision tree induction. Additionally, the simulated sample is ten times larger than the real one. Yet, the Random Forest classifier consists of unpruned trees too, thus the size of the new model is reduced compared to the old one by the total number of trees in the ensemble (originally 500). That is, in general, the size complexity of an unpruned decision tree is N times smaller than the complexity of a Random Forest that is built using the same data, where N stands for the number of trees in the ensemble. Consequently, as the execution speed scales linearly with the depth of a tree, the gain achieved in that aspect is also quite remarkable.

Fig. 4 depicts first few levels of the decision tree induced from the simulated data. The Random Forest classifier has been previously trained using the whole dataset. Sequence similarity (as determined by BLAST) and gene function annotation (as annotated in the Swissprot database) compared to the known phenotype-associated genes are among the most discriminatory features. Also the global Endeavour scores seems to play an important role in the decision-making process be it in the form of the rank, the p-value or the normalized p-value of the gene prioritization.

Interestingly, not only phenotype-specific information but also prediction of the deleterious impact of the mutation (here shown by the MutationTaster node) appears informative. This seems logical as not any mutation in a potentially phenotypically-associated gene might be disease-causing if it does not perturb protein function significantly. This demonstrates that the decision boundary of the Random Forest clearly takes into account both aspects of the mutation: the damaging functional impact and the phenotypical relevance of the gene in which it lies.

When the structure of this tree is compared with Random Forest feature importance measures from the initial study [11], two major observations can be made. Firstly, importance of sequence similarity and gene function annotation for discriminating between two classes seems to be consistent,

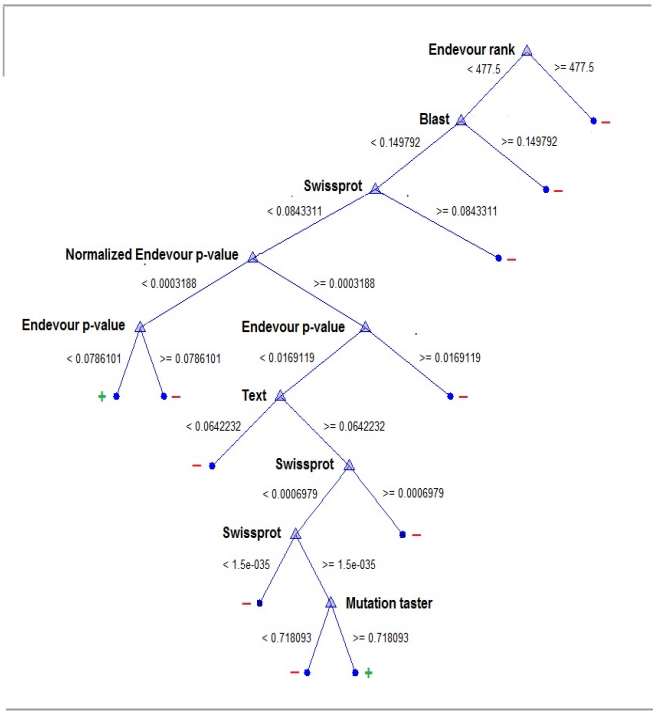

Fig. 4. Decision tree constructed on the data generated by Random Forest. For clarity of the display tree has been pruned such that only ten the most discriminative nodes remain. Note that, due to pruning, the labels (signs) associated with "terminal" nodes only indicate which class preveals in a node and should not be confused with the final decision provided by the model.

regardless to a method used to estimate it. This is the strong indication of the biological relevance of these features. Secondly, the Endevour scores apparently characterize this classification problem much more than previously implied. For being highly correlated, their importance was heavily underestimated by RF-FI; while here they constitute the most discriminative nodes in the tree corresponding to the simplified eXtasy. This suggests that importance of a variable for a particular classification problem can be indirectly assessed by modeling decision boundary of a complex classifier by a more interpretable one.

IV. CONCLUDIONS

We presented a simple method for extracting a comprehensible model from a Random Forest classifier and applied it to reduce the complexity of the eXtasy variant prioritization algorithm. The achieved performances are in line with that previously established in the original study, while the size (and consequently execution speed) of the model has been reduced considerably. In addition, we provide insights into the working mechanism behind eXtasy using an improved explanatory capacity of the new formulation. In the near future, we plan to further simplify the model through feature selection. Also, we plan to experimentally determine an appropriate method for calibrating posterior probabilities provided by a decision tree induction algorithm implementation(s), such that they can be used for ranking variants without losing performance along certain ranges of the operating space.

ACKNOWLEDGMENT

Bart De Moor and Yves Moreau are full professors at the Katholieke Universiteit Leuven, Belgium.

Research supported by:

- Research Council KU Leuven: GOA/10/09 MaNet, KUL PFV/10/016 SymBioSys, and several PhD/postdoc & fellow grants
- Industrial Research fund (IOF): IOF/HB/13/027 Logic Insulin, IOF: HB/12/022 Endometriosis
- Flemish Government:
  - FWO: PhD/postdoc grants, projects: G.0871.12N (Neural circuits), research community MLDM
  - IWT: PhD Grants; TBM-Logic Insulin, TBM Haplotyping, TBM Rectal Cancer, TBM IETA
  - Hercules Stichting: Hercules III PacBio RS
  - iMinds: SBO 2013; Art&D Instance
  - IMEC: phd grant
  - VLK van der Schueren: rectal cancer
  - VSC Tier 1: exome sequencing
- Federal Government: FOD: Cancer Plan 2012-2015 KPC-29-023 (prostate)
- COST: Action BM1104: Mass Spectrometry Imaging, Action BM1006: NGS Data analysis network

The scientific responsibility is assumed by its authors.

REFERENCES

[1] Ng SB et al. "Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome," Nat Genet., 42(9):790-793, September 2010.

[2] Lupski, J. R. et al. "Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy," The New England journal of medicine, 362(13), 1181–91, April 2010.

[3] Vissers, LE et al. "A de novo paradigm for mental retardation," Nature genetics, 42(12), 1109–12, December, 2010.

[4] Chun, S. & Fay, J. C. "Identification of deleterious mutations within three human genomes," Genome research, 19(9):1553–61, September 2009

[5] Ng, P. & Henikoff, S. "SIFT: Predicting amino acid changes that affect protein function," Nucleic acids research, 31, 3812, July 2003.

[6] Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. "MutationTaster evaluates disease-causing potential of sequence alterations," Nature methods 7, 575–6, August 2010.

[7] Adzhubei, IA et al. "A method and server for predicting damaging missense mutations," Nature methods 7, 248–9, April 2010.

[8] Lopes, MC et al. A Combined Functional Annotation Score for Non-Synonymous Variants. Human Heredity, 73(1), 47–51. March 2012.

[9] Kumar, S., Sanderford, M., Gray, V. E., Ye, J. & Liu, L. "Evolutionary diagnosis method for variants in personal exomes", Nature methods 9, 855–6, September 2012.

[10] O'Fallon, B., Wooderchak-Donahue, W., Bayrak-Toydemir, P., & Crockett, D. "VarRanker: rapid prioritization of sequence variations associated with human disease," BMC Bioinformatics, 14(Suppl 13), October 2013.

[11] Sifrim A & Popovic D et al. "eXtasy: variant prioritization by genomic data fusion," Nature Methods 10, 1083–1084, November 2013.

[12] Aerts, S. et al. "Gene prioritization through genomic data fusion," Nature biotechnology 24, 537–44, Jun 2006.

[13] Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E," Characterising and predicting haploinsufficiency in the human genome," PLoS genetics 6, e1001154, October 2010.

[14] Pertea, M., Pertea, G. M. & Salzberg, S. L. "Detection of lineage-specific evolutionary changes among primate species," BMC bioinformatics 12, 274, July 2011.

[15] Siepel, A. et al. "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," Genome research 15, 1034–50, August 2005.

[16] Breiman, L. Random Forests, Machine Learning , 45(1): 5–32, October 2001.

[17] J.F. Elder "The generalization paradox of ensembles," J. Comput. Graphical Stat., 12(4) , 853–864, December 2003

[18] Strobl C, Zeileis A, Danger: High Power! – Exploring the Statistical Properties of a Test for Random Forest Variable Importance. Proceedings of the 18th International Conference on Computational Statistics, Porto, Portugal 2008.

[19] Archer, K. J. and R. V. Kimes "Empirical characterization of random forest variable importance measures," Computational Statistics & Data Analysis 52(4), 2249–2260, January 2008.

[20] Tolo̧si L & Lengauer T, "Classification with correlated features: unreliability of feature ranking and solutions," Bioinformatics, 27(14): 1986-1994, July 2011.

[21] Strobl C, Boulesteix AL, Zeileis A, Hothorn T, Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics, 8:25, January 2007.

[22] Z. Zhou, Y. Jiang and S. Chen, "Extracting symbolic rules from trained neural network ensembles," AI Communications, 16(1):3-15, January 2003.

[23] L.U Fu, "Rule generation from neural networks," IEEE Transactions on Systems, Man and Cybernetics, vol. 28, no.8, pp. 1114-1124, August 1994.

[24] P. Domingos, "Knowledge discovery via multiple models," Intelligent Data Analysis, 2:187-202, 1998.

[25] Durbin, RM Et al. "A map of human genome variation from population-scale sequencing," Nature, 467(7319), 1061–1073. October 2010.

[26] Liu, X., Jian, X., & Boerwinkle, E. "dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions," Human mutation, 32(8), 894–9. August 2011

[27] Vihinen, M. "How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis," BMC genomics 13 Suppl 4, S2, June 2012.