

Author's response to reviews

Title: Spatially Uniform ReliefF (SURF) for Computationally-Efficient Filtering of Gene-Gene Interactions

Authors:

Casey S Greene (casey.s.greene@dartmouth.edu)

Nadia M Penrod (nadia.m.penrod@dartmouth.edu)

Jeff Kiralis (jeff.kiralis@dartmouth.edu)

Jason H Moore (jason.h.moore@dartmouth.edu)

Version: 2 Date: 26 June 2009

Author's response to reviews:

Thank you very much for the thoughtful reviews. We have addressed each criticism below point by point.

Reviewer: Brett McKinney

Summary:

ReliefF is a datamining algorithm capable of identifying attribute interactions, which is a significant challenge in genetic associations studies. ReliefF is a metric-based method that scores the relevance of an attribute to the phenotype by calculating how well the attribute separates nearest-neighbor instances with different phenotype status. The number of nearest neighbors is a user-chosen parameter, which is typically set to 10 with little further discussion. Instead of considering a fixed number of nearest neighbors, the authors consider a variable number of nearest neighbors within a fixed radius of a given instance, and they provide a statistical framework for guiding the choice of this radius.

Minor Essential Revisions:

The authors use "the mean distance as calculated from the data" as the the distance threshold. Could the authors clarify what "mean" they refer to? The mean distance across all pairs of instances? The mean of the distance between the given instance of interest and its nearest neighbors? Are hit and miss instances treated separately?

We have specified that we refer to the mean distance across all pairs of instances and that hits and misses are not treated separately.

In Relief-F, the distance between instances is typically calculated in the space of

all attributes. However, in the development of SURF it appears that only two SNPs (attributes) are considered. Why?

We use all attributes to calculate distance and now clarify this.

The equations for the M and H cardinalities (after Eq.2 in the appendix) include a factor of $N/2$. I assume this is due to an assumption of a balanced data set; an equal number of cases and controls? Will this hinder analysis of unbalanced data sets?

This is due to the balanced data assumption. We hope to provide a thorough examination of these methods on unbalanced data in the future.

Is this assumption also the source of $1/2$ in Eq. 4 of the main body?

The $1/2$ is here since each individual in TM1 and TH1 changes the score of a relevant SNP by $1/2$ on the average so this does not arise because of the balanced data assumption. We now make this clear.

Reviewer: Marylyn Ritchie

Abstract:

It says that ReliefF has low power when the interaction effect is small. How small? Many people in the field expect the interaction effects to be small (while others think it will be large) and so they may be put off by that statement right in the abstract and may not read any further. They will think this approach cannot work.

The issue is noise more than size, so we now specify that the issue is noisy data. We also now note that TuRF largely ameliorates this specific concern.

What is the user-selectable parameter? This makes no sense here.

We remove a nuisance parameter (that is a user-selectable parameter that is difficult to properly set without already knowing the optimum for a dataset dependent parameter value) and now say this.

What is meant by "amining"? I cannot find a definition for this word.

This was an error and has been corrected.

The phrase "discover detecting" is awkward and does not make sense.

This was an error and has been corrected.

Background:

The are many more updated references than what is listed in 1,2. Those are ok but there should be newer ones as well.

We now include some more recent and updated references.

In the third paragraph where it says "small data sets" it would be good to explain whether you mean small in terms of the number of samples or SNPs.

We mean small in terms of SNPs and now specify this.

Page 3, second paragraph. The first sentence is awkwardly worded. It also says that Relief algorithms detect interacting pairs of attributes. Later in the paper is says that it does not detect interacting pairs (page 8). It would also be good to define relief algorithms in this paragraph. The general reader of this will not know.

This was poor wording because these algorithms detect SNPs associated with disease (pair-wise or otherwise) but do not provide a model or specify which pairs interact (beyond what amounts to "X and Y are both associated in some way"). We now properly discuss these algorithms finding disease associated SNPs. We also now provide a brief definition of Relief algorithms here.

The statement that stochastic approaches fail without additional information is highly dependent on the types of models being evaluated. Models with main effects or main effects with interactions can be found without Relief algorithms.

We now specify that this is in the case of purely epistatic effects.

Also, purely epistatic models in data with linkage disequilibrium between noise loci and the functional loci can be detected in stochastic algorithms.

Here we avoid the discussion of functional and non-functional loci and focus solely on relevant and irrelevant loci. Loci which are non-functional but correlated with functional loci would, in this context, be considered relevant.

What is meant by "a set of assumptions regarding variance"?

We remove the "set of assumptions regarding variance" statement and now provide the paper citation as the paper has been published. The specific assumptions are available in the reference.

When you say a single nearest neighbor, is that one person, one SNP, one person at all of their SNPs? Based on how many SNPs?

The nearest neighbor is the nearest single individual based on all SNPs. We now specify this.

Does deleting SNPs with lowest Relief weights make it difficult to detect SNPs

with purely interactive effects?

This actually makes it easier to detect SNPs with purely epistatic disease association because noisy SNPs are most often removed. This means that the re-estimation can more accurately evaluate the relevance of the remaining SNPs. We now clarify this.

In general, this first Background section is very difficult to follow. A figure would be helpful to explain these different Relief algorithms in this section and the next.

We have added a new figure, Figure 1, which shows how neighbors are selected for each of the Relief algorithms.

Results:

All of the references to figures are listed as "figure ??"

This occurred when figures were removed for submission. These labels have been added back in.

In the third paragraph you discuss the powers when SURF is used. This is not power. It is really sensitivity or detection.

This is an excellent point. We now discuss these results in terms of success rate which is a much more appropriate term. We note that the success rate we are discussing is the ability to get both relevant SNPs above a certain rank.

On page 8 it says "direct replacement of these methods". What methods?

We now specify that we are talking about direct replacement of ReliefF is what may improve these frameworks.

Methods:

The use of 99th, 95th, 75th percentiles in the first paragraph is not clear. Percentiles of what? I can figure it out once I read further, but it would be good to describe it here.

We now specify that these are percentiles of SNPs.

On page 9, it says that we test each method with fixed parameters. What methods? What are the parameters?

We change the wording to indicate that the parameters are to follow.

How long does it take to run 1000 SNPs? Is it feasible to run 1M? Does the distance metric work in that space?

Running 1000 SNPs with SURF is fast (12 seconds for 1000 SNPs with 800 individuals). The method scales linearly with respect to the number of SNPs, so running 1M is feasible. The distance metric would work in that space as well,

although the influence of the relevant SNPs on the signal would be reduced which could affect success rate.

Tables/Figures

Where are the other 30 penetrance models? They should be provided.

We now provide these models as supplementary material.

Did you evaluate everything from 99th - 75th percentile?

We did not evaluate the 99th to 75th percentiles. We only evaluated those percentiles that we thought would be commonly used in practice. These percentiles correspond to a researcher reducing a 1000 SNP dataset to 10, 50, or 250 SNPs for a more thorough epistasis analysis.

Figure 2 legend says figure ??

Are these plots an average across the 30 models with 5 shown in each plot?

We now specify that this is the case.

Figure 1 is not clear. The bar plots need to be better defined. This does not get the point across.

We have added examples to indicate how these plots are to be read.

Reviewer: David V Conti

The authors present an extension to existing algorithms designed to detect interacting variants. Specifically, they introduce a Spatially Uniform ReliefF (SURF) algorithm that weights neighbors of an individual. Simply put, the original ReliefF algorithm uses a fixed number of neighbors, while SURF uses all neighbors within a fixed distance. The authors perform a simulation study to demonstrate increased power across a variety of heritabilities and sample sizes. In general, I found this manuscript very hard to follow. Most of the relevant information is included in the Appendix and only those that are already familiar with the approaches will follow the Methods section without constant reference to the Appendix. Since this is the case, the paper would benefit including more of the material within the Appendix in the main part of the paper to help guide the reader.

We have moved some material from the appendix to the paper to improve the clarity.

My biggest concern is the presentation of power. Power in this paper is defined

as: "The percentage of time that a method scores both relevant SNPs above a given threshold..." This does not give us the complete story and thus the results do not allow for complete comparison of methods. Most investigations are also interested in:

Under the null when there are no simulated epistatic SNPs, how many SNPs do I score above a given threshold? This is typically called the Type I error and it is necessary for two methods to have the same Type I error in order to interpret power. It is unclear if these methods have the same Type I error. Related to this is the False Discovery Rate. Of those that I declare significant, how many are false? This would also be of interest. Does the use of spatially related weighting help reduce the False Discovery Rate, since I am utilizing information in a more efficient manner? Currently, the results are very misleading. I view showing the complete context of these results as necessary for publication.

We were using power very loosely. We now report results in terms of success rate. Because these methods are used for filtering, we consider success rate the number of times that both relevant SNPs are scored above a given threshold. We set this standard because further analysis steps can not succeed if both relevant SNPs are not discovered.