**Introduction**

Single Nucleotide Polymorphisms are the most common type of genetic variation in humans. Recent studies show that SNPs can be associated to the risk of development of complex diseases such as cancer. But the individual contribution of each SNP seems to be very small compared to the contribution of SNP combinations.

Among various computational approaches to identify interacting SNPs in a context of genotypic individual-level data, this web-tool gives the user the possibility to analyze three of them: Polymorphism Interaction Analysis v.2, Multifactor Dimensionality Reduction, Entropy-Based SNP-SNP Interaction Method and Multi-Approach SNP-SNP Interaction Analysis (novel). The main goal of these methods is to search and rank the combinations of features (SNPs or environmental factors) most likely associates with the disease. In addition, they calculate the analysis time, allowing the user to make a performance analysis.

The site offers:

ANALYSIS: Automatic execution of analysis using different approaches.

METHODS: A brief yet robust explanation of the methods, including their input data format.

REFERENCES: Pointers to papers that present the methods in detail.

**Polymorphism Interaction Analysis – Method Overview**

The Polymorphism Interaction Analysis (see session “References” [1,2]) is a method designed exclusively to identify possible interactions among features like SNPs or environmental factors. It is based on 7 scoring functions calculated for each possible factor combination and an overall scoring function that is calculated summing the normalized scoring functions for each combination of SNPs and/or other factors.

For the first 5 scores it creates a n-dimensional table (where n is the order of the interaction) that accounts for the total of cases and controls for each possible attribute values (see the image 1). Then a cross-validation is performed dividing the data in equal parts and accounting for the number of TP, FP, TN and FN using the test (and maybe the train too) data. That cross validation is repeated a certain number of times and it results in a contingency table (or confusion matrix) similar to the one in image 2. With that table it is possible to calculate the following scores:

(imagem dos 5 primeiros scores)

The other 2 scores are calculated based on two split-measures used in other methods like CART (Classification and Regression Trees). They are the Gini Index and the Absolute Probability Difference (APD). Differently than the first five scores they use the entire data instead of dividing it in training and testing sets. Suppose that we have the genotype-phenotype table as shown on image 3, the following equations gives us the Gini Index and the APD.

(imagens do gini e do apd)

Finally, the overall score is obtained by normalizing each scoring function so that the highest value is set to 50, then summing all the score values for each combination. Note that the maximum score that can be obtained this way is 350 as there are 7 scores. If the numbers of cases and controls are the same or the fractional occupation is used, the %correct score gives the same value as the sensitivity+specificity, and then only one of them is used, making the overall maximum score of 300.

**Parameters details**

* Base Path: Defines the path and the name of the base in the user’s PC.
* Use Pathway: Check if the base contains information about pathway of each feature, for more details see session “Data Input Format”.
* Order: Interaction order. Ex: if the value passed is 2, then the algorithm will search for combination between 2 features.
* No. Top Results: The number of top combinations that will be shown in the result screen. If it is larger than the total number of combinations, the result will truncate to exhibit only the maximum number of combinations. The pathway analysis is restricted to that number of combinations.
* Shuffle: It represents the number of times the data will be shuffled at the start of each cross-validation step.
* Fract: Represents how much of the data will be used as the test data in the cross-validation analysis. It must be a float between 0 and 1. Ex: If 0.1 is passed, then 10% of the data will be used as test at each cross validation interval testing.
* NTime: Tells how much cross-validations repetitions will be performed. The TP, FP, TN, FN numbers are summed for all the cross-validation repetitions in each feature combination.
* Ratsl: It’s a floating value that cut off some attribute combinations that does not satisfy the following criteria: Max(cases,controls) / Min(cases,controls) <= Ratsl. It is used to cut the combinations extremely affected by missing data.
* Ifract: Must be 0 or 1. Where 1 means that we are using fractional occupation to populate the genotype-phenotype table and 0 means that we use the actual values.
* ITrain: Must be 0 or 1. If 1 it means that we are using the training set to calculate the contingency table (along with the testing data), if 0 it means that we are using only the testing data.
* Lootr: If the ITrain parameter is 1, it tells how the training set values are used to populate the contingency table. For detailed information on each possibility see reference [2] on session “References”.

**Results Shown on this tool**

The first two light and dark blue tables are respectively the parameters used in the testing and the analysis time divided by each algorithm part and the total time.

Then it's shown the top combinations of features returned by each scoring function. If pathway analysis was requested then another table will be shown with the results for that analysis.

**Multifactor Dimensionality Reduction – Method Overview**

The Multifactor Dimensionality Reduction (see session “References” [3-5]) is based on the idea of reduce the dimensionality of each attribute, so that this new attribute is made easier and faster to be predicted. It performs a cross-validation strategy to select, based on the balanced accuracy predictor, the best model that explains the splitting between cases and controls.

As shown on the image 1, the algorithm starts by dividing the data into equal parts (cross-validation intervals). For each interval we choose the N factors (where N is the order of the interaction) next to be examined (it examines all the possible combinations) and put the data in a genotype-phenotype table (step 3). In the Step 4 each cell is categorized as “low-risk” if the case/control ratio is smaller than a certain threshold or “high-risk” if it is higher than that threshold, or “empty” if the cell contains no information. Then the classification error is calculated for each of the possible factor combinations and the lowest error is chosen to be the model that best explains the data in that cross-validation turn. In Step 6, the testing data that was separated is used to calculate the prediction error of that cross validation turn for the best model chosen by its classification error. This procedure is done a couple of times and the best classification combination is ranked. In addiction is shown the best prediction error of the best model for each cross-validation turn.

**Parameter Details**

* Base Path: Defines the path and the name of the base in the user’s PC.
* Use Pathway: Check if the base contains information about pathway of each feature, for more details see session “Data Input Format”.
* No. Top Results: The number of top combinations that will be shown in the result screen. If it is larger than the total number of combinations, the result will truncate to exhibit only the maximum number of combinations. The pathway analysis is restricted to that number of combinations.
* Shuffle: It represents the number of times the data will be shuffled at the start of each cross-validation step.
* Fract: Represents how much of the data will be used as the test data in the cross-validation analysis. It must be a float between 0 and 1. Ex: If 0.1 is passed, then 10% of the data will be used as test at each cross validation interval testing.
* NTime: Tells how much cross-validations repetitions will be performed. The TP, FP, TN, FN numbers are summed for all the cross-validation repetitions in each feature combination.
* Threshold: A floating value representing the threshold used on the categorization of each cell of the genotype-phenotype table as “low-risk” and “high-risk”.

**Results shown on this tool**

The first two light and dark blue tables are respectively the parameters used in the testing and the analysis time divided by each algorithm part and the total time.

It is shown two values bellow these tables that represent the mean balanced accuracy for all of the combinations calculated for the classification and prediction steps.

Then it's shown the top combinations of features returned by each scoring function. If pathway analysis was requested then another table will be shown with the results for that analysis.

**Random Forest – Method Overview**

The Random Forest (see session “References” [6,7]) is a well-known data mining method used at the classification task in many applications. There is a lot of works published that show the consistency of this method and it is used at the interaction analysis area as shown by [7].

The Random Forest consists of generating a set of Decision Trees that are calculated based on the ability of each feature to split the data between case and control. In the Random Forest algorithm, each node of the decision tree is selected among a subset of size M of the set of all attributes (where M << total of attributes). The population used to build each tree is sampled using a bootstrapping approach, on which N samples are selected (allowing repetitions) to be the training set (where N is the size of the training set) and the samples that are not selected are called Out-Of-Bag (OOB) samples.

After generating all trees (without pruning) we then calculate the margin of each individual with the trees that were generated leaving that individual OOB. The margin is the number of (correct – incorrect) / total of the predictions.

After that, for each feature combination, we permute the values of each attribute involved in the combination being tested and calculate the new margin based on this new data set with the attribute permuted. We then calculate the final score based on the ratio between the summation of the (margin – permuted margin) and the total of individuals. Then each feature combination is ranked based on that score.

**Parameter Details**

* Base Path: Defines the path and the name of the base in the user’s PC.
* Use Pathway: Check if the base contains information about pathway of each feature, for more details see session “Data Input Format”.
* Order: Interaction order. Ex: if the value passed is 2, then the algorithm will search for combination between 2 features.
* No. Top Results: The number of top combinations that will be shown in the result screen. If it is larger than the total number of combinations, the result will truncate to exhibit only the maximum number of combinations. The pathway analysis is restricted to that number of combinations.
* Initial Shuffle: It represents the number of times the data will be shuffled at the start of the algorithm.
* Scramble: Represents the number of times the feature vector will be shuffled.
* Permute: Represents the number of times the data set is permuted to calculate the new permuted margin.
* M Parameter: The size of the set containing the maximum number of attributes that the algorithm will use to generate each node of the decision tree.
* No. Trees: The number of trees in the forest.

**Results shown on this tool**

The first two light and dark blue tables are respectively the parameters used in the testing and the analysis time divided by each algorithm part and the total time.

Then it's shown the top combinations of features returned by each scoring function. If pathway analysis was requested then another table will be shown with the results for that analysis.

**ESNP2 – Method Overview**

There are two versions of ESNP2 (see session “References” [6]). The ESNP2-standard (ESNP2-S) aims to detect the interactions the same way PIA and MDR do, and the ESNP2-model-option (ESNP2-Mx) tests the interaction against various two-locus genetic models. The version available here is the ESNP2-S.

ESNP2 standard uses the entropy to analyze which combinations better explain the disease. The strategy also analyzes all possible attribute combinations. The algorithm begins by calculating the entropy of each individual feature, using the following equation:

(eq1)

Then it proceeds selecting a combination of features and populating the genotype-phenotype table using case-controls status for all possible value that the combination can assume the same way PIA and MDR does.

Then the entropy of interactions of possible combinations is computed by the following equation:

(eq2)

The final score Delta R, that represents the interaction effects, is then given by:

(eq3)

**Parameter Details**

* Base Path: Defines the path and the name of the base in the user’s PC.
* Use Pathway: Check if the base contains information about pathway of each feature, for more details see session “Data Input Format”.
* Order: Interaction order. Ex: if the value passed is 2, then the algorithm will search for combination between 2 features.
* No. Top Results: The number of top combinations that will be shown in the result screen. If it is larger than the total number of combinations, the result will truncate to exhibit only the maximum number of combinations. The pathway analysis is restricted to that number of combinations.

**Results shown on this tool**

The first two light and dark blue tables are respectively the parameters used in the testing and the analysis time divided by each algorithm part and the total time.

Then it's shown the top combinations of features returned by each scoring function. If pathway analysis was requested then another table will be shown with the results for that analysis.

**MASS – Method Overview**

The MASS method is a novel approach designed based on the other methods capabilities and disabilities.

(Image 1)

MASS uses 4 different score metrics the same way PIA does, the four metrics are: Information Gain (based on entropy very similar to ESNP2), Chi-Square based on Contingency Table, Gini Index and APD (the same way PIA does). The final scoring metric (Overall) is calculated by normalizing the values of each score so that the higher is 100 and the lower is 0 and then summing the different metrics' score for each combination.

The MASS algorithmic process is similar to ESNP2 and the scores are similar to both PIA and ESNP2. The differences are summarized at Table 1. It describes the Chi-Square Scoring Metric and describes the new final formula to calculate the interaction effects considering all the individual effects instead of only the minimum (as ESNP2 does).

(Table 1)

**Parameter Details**

* Base Path: Defines the path and the name of the base in the user’s PC.
* Use Pathway: Check if the base contains information about pathway of each feature, for more details see session “Data Input Format”.
* Order: Interaction order. Ex: if the value passed is 2, then the algorithm will search for combination between 2 features.
* No. Top Results: The number of top combinations that will be shown in the result screen. If it is larger than the total number of combinations, the result will truncate to exhibit only the maximum number of combinations. The pathway analysis is restricted to that number of combinations.
* Ratsl: It’s a floating value that cut off some attribute combinations that does not satisfy the following criteria: Max(cases,controls) / Min(cases,controls) <= Ratsl. It is used to cut the combinations extremely affected by missing data.
* Simple Assignment: Defines the usage of only individual attribute's scores for the combination's scores (true) or the usage of all attribute's possible combinations (false).
* Use Gain: Defines whether or not the Information Gain will contribute to the Overall Score.
* Use Chi: Defines whether or not the Chi-Square will contribute to the Overall Score.
* Use Gini: Defines whether or not the Gini Index will contribute to the Overall Score.
* Use APD: Defines whether or not the APD will contribute to the Overall Score.

**Results shown on this tool**

The first two light and dark blue tables are respectively the parameters used in the testing and the analysis time divided by each algorithm part and the total time.

Then it's shown the top combinations of features returned by each scoring function. If pathway analysis was requested then another table will be shown with the results for that analysis.

**Data Input Format Description**

The data should follow a specific format. As the Image 1 at the end of this page shows, the first line is the header line. It should contain M values, where M is the number of features, describing the name of each feature, plus one string naming the last column that should contain the disease status. The second line (in the image 1) is optional. It is the pathway header and should contain M values containing the pathway associated with the SNP in question, plus one string naming the last column again. And then follows N lines where N is the population size. Each of these lines contain M+1 values, M values of the features that are integers or strings representing the category associated with that individual and the last value representing the disease status (0 for control and 1 for case). All the values can be separated by any of these characters, that should be used only to separate and cannot be used on the name or value strings: blank space, tabulation, “,”, “|”, “/” and “\”. Any of these characters can be used and many of them can be used on the same file, but each separation should contain only one (two empty spaces is wrong).

(image of example of the data)