Improving the Prediction of Clinical Outcomes from Genomic Data Using Multiresolution Analysis

Pablo H. Hennings-Yeomans and Gregory F. Cooper

Abstract—The prediction of patient's future clinical outcome, such as Alzheimer's and cardiac disease, using only genomic information is an open problem. In cases when genome-wide association studies (GWASs) are able to find strong associations between genomic predictors (e.g., SNPs) and disease, pattern recognition methods may be able to predict the disease well. Furthermore, by using signal processing methods, we can capitalize on latent multivariate interactions of genomic predictors. Such an approach to genomic pattern recognition for prediction of clinical outcomes is investigated in this work. In particular, we show how multiresolution transforms can be applied to genomic data to extract cues of multivariate interactions and, in some cases, improve on the predictive performance of clinical outcomes of standard classification methods. Our results show, for example, that an improvement of about 6 percent increase of the area under the ROC curve can be achieved using multiresolution spaces to train logistic regression to predict late-onset Alzheimer's disease (LOAD) compared to logistic regression applied directly on SNP data.

Index Terms—Human genome, single nucleotide polymorphisms, multiresolution, pattern recognition, wavelets, prediction, clinical outcomes, genomics, SNPs.

1 Introduction

Using genomic information to estimate individual's risk for disease is an important objective of personalized medicine. Related objectives include diagnosis and the estimation of an individual's response to medication. Genome-wide sequencing platforms have advanced significantly during the last decade, which increases the opportunities for risk prediction and related objectives. Today, it is possible to measure millions of DNA loci (base pairs) within hours for a few hundred dollars. Such developments have led to genome-wide association studies (GWASs) which have discovered a variety of loci associated to common diseases, such as heart disease and late onset Alzheimer's disease (LOAD). These loci can be used as biomarkers to predict the disease risk of individuals. Most research has focused on searching for single loci that predict disease [1]. The search for sets of interacting multiple loci has been less explored [2], [3], [4], [5].

The current paper describes a multivariate approach for genomic pattern recognition. In particular, we treat the genome as a one-dimensional signal, such as voice, music, and ECGs, to which we apply multiresolution analysis. Our goal is to extract genomic features from sequences of single nucleotide polymorphisms (SNPs), which are the genomic variants that represent most of the variations in DNA. Instead of working directly with the input space of SNPs, we use multiresolution analysis to decompose the genome into multiresolution spaces, where we aim to find better features for prediction.

The main contribution of this paper is to propose that genomic feature extraction or selection be carried out in a multiresolution domain instead of the input SNP domain, which can lead to improve predictions of clinical outcomes. This approach has been used with success before in other fields of research, such as biometric recognition and bioimaging [6], [7], [8]. In genomics, the work of Hutter et al. [9] applies multiresolution analysis in the process of computing statistics from DNA sequence polymorphisms at the genome-wide level, but it does not train a model for the prediction of clinical outcomes from multiresolution features, as we propose here. To our knowledge, the current paper is the first investigation of multiresolution analysis with genomic data for the prediction of clinical outcomes.

The following section briefly explains background for this work. Section 3 describes the multiresolution classification framework applied here for prediction. Section 4 details the experimental setup and results. Section 5 discusses key points, and Section 6 provides conclusions.

2 Background

2.1 Genome-Wide Association Studies

With the discovery that polymorphic DNA sequences can be used as a reference to genetic markers [10], it is now possible to represent the human genome as a localized 1D sequence that can be read uniformly across populations. The most widely used polymorphisms today are single nucleotide polymorphisms.

Before the International HapMap Project [11], methods for detection of DNA polymorphisms produced few samples at a high cost. However, relatively recent developments in genomic technology allow the measurement of millions of

• *The authors are with the Department of Biomedical Informatics, University of Pittsburgh, Suite 500, 5607 Baum Boulevard, Pittsburgh, PA 15206-37041. E-mail: pablohy@ieee.org, gfc@pitt.edu.*

Manuscript received 1 Aug. 2011; revised 3 Apr. 2012; accepted 12 May 2012; published online 12 June 2012.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2011-08-0200. Digital Object Identifier no. 10.1109/TCBB.2012.80.

SNPs per individual for a few hundred dollars. An overview of genotyping technologies can be found in [12].

Genome-wide association studies aim to find disease-gene associations by mining the dense SNPs sequences of large groups of people. The studies often follow a case-control design in which a well-defined phenotype or clinical outcome is established, and the study is performed using an affected group (cases) and an unaffected group (controls). The analysis focuses on finding differences in the SNP sequences between these two groups.

Due to the large dimensionality of the input space, most GWAS studies have focused on univariate analyses for biomarker discovery. However, increasing effort is being exerted on multivariate approaches [4], [5].

2.2 Genomic Pattern Recognition

Prediction of clinical outcomes is typically posed as a binary classification problem. After quality control (QC) procedures are performed, the usual pattern recognition framework uses a feature extraction method to summarize the data in a space that will improve classification.

We use the χ^2 statistic as our ranking measure for feature selection. In particular, we use p-values from a chi-square analysis to rank the most univariately predictive SNPs, and select the top K SNPs to define a K-dimensional feature space, where K is found through cross-validation. One could use principal component analysis, alternatively, or other subspace methods as a way of feature extraction; we found ranking using χ^2 to give competitive results with genomic data.

Once a feature space is defined, a classification algorithm, such as logistic regression, random forests, or support vector machines can be trained [13].

2.3 Multiresolution Transforms

Building on properties such as time resolution and frequency localization, multiresolution transforms were originally proposed as a mathematical tool to overcome limitations of the Fourier transform. With Fourier transforms, the time-frequency bins to represent signals (sequences) is fixed, while with multiresolution transforms we can adapt the 2D inner tiling of this time-frequency plane to better support the sequence at hand [14]. In practice, this means that with multiresolution transforms, in most cases, the coefficients that represent our sequences will capture more precise information about our sequence than would Fourier coefficients.

A thorough exposition of multiresolution analysis can be found elsewhere [14], [15]. We cover here the discrete wavelet transform (DWT) and a particular multiresolution transform of the family of frames, named the undecimated discrete wavelet transform (UDWT), also found under other names, such as the shift invariant wavelet transform and the stationary wavelet transform (SWT).

From the perspective of linear algebra, both of these multiresolution transforms can be implemented as

$$\mathbf{u} = \mathbf{W}\mathbf{x}, \tag{1}$$

where $\mathbf{x}$ is an $N \times 1$ input vector. For the DWT, $\mathbf{u}$ is a vector of wavelet coefficients also $N \times 1$, while for the UDWT it is an $(L+1)N \times 1$ vector, where L is the number of levels (or scales) in the decomposition. This is because frames are *redundant*, while proper wavelet transforms are not. Then, for the DWT, $\mathbf{W}$ is $N \times N$, and for the UDWT it is an $(L+1)N \times N$ matrix. The matrix $\mathbf{W}$ consists of $L+1$ submatrices for both transforms

$$\mathbf{W} = [\mathbf{W}_1\ \mathbf{W}_2\ \ldots\ \mathbf{W}_L\ \mathbf{W}_{L+1}]^T, \tag{2}$$

however, for the DWT the dimensions of $\mathbf{W}_k$ are $N/2^k \times N$ for $k = \{1, 2, \ldots,\ L\}$ and $N/2^L \times N$ for $k = L+1$, while for the UDWT all $\mathbf{W}_k$ are $N \times N$. The columns of $\mathbf{W}_k$ are circularly shifted versions of the same vector. For $k = 1$ to L we obtain as output what are referred to as *detail* coefficients, while for $k = L+1$ we obtain *approximation* coefficients at the coarsest level of the decomposition. Every $\mathbf{W}_k$ implements an orthonormal projection of the input $\mathbf{x}$ onto a multiresolution space, or *subspace*. These subspaces are *orhtonormal* for the DWT, and there is a unique inverse transform to recover the input $\mathbf{x}$ from the subspace coefficients, $\mathbf{u}$. For the UDWT, and for frames in general, no unique inverse transform exists.

What makes multiresolution powerful for processing signals, such as voice, images and in our case, the genome, is that it can be implemented efficiently using filterbanks, regardless of the dimensionality of the signal [14]. Applying (1) directly takes $O((L+1)N)$ operations, however, using fast algorithms we only need $O(LN)$ computations.

From a digital signal processing perspective, (1) can be implemented iterating a single two-channel filterbank recursively. This *unit* filterbank implements a single-level multiresolution decomposition of the input. For an L-level multiresolution transform, the filterbank is applied L times recursively on the output channel of the approximation coefficients. The output $\mathbf{u} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{L+1}\}$ is the collection of detail coefficients from the filterbank branches where it was not iterated, plus the approximation coefficients of the last iteration. These filterbank outputs are the coefficients $\mathbf{u}_k = \mathbf{W}_k\mathbf{x}$.

An important generalization of multiresolution transforms comes from relaxing the requirement of iterating only on the approximation channel of the filterbank. By iterating on the detail channel we obtain more general multiresolution representations, which have been called wavelet packet trees. If we iterate on every output channel possible up to the Lth level, we have a *full* wavelet packet tree.

Wavelet transforms have been applied successfully in signal compression and coding. On the other hand, redundant multiresolution transforms, such as the UDWT, have proven useful in areas that include denoising and feature extraction. In this paper, we use multiresolution transforms that are wavelet packet trees and undecimated.

3 Multiresolution Framework

The approach we use to predict clinical outcomes from genomic data is shown in Fig. 1. It consists of expanding the original SNP space into multiple multiresolution spaces, in each of which a classifier is trained. A final decision rule combines the individual output scores (probabilities) of the multiresolution subspace classifiers into a final prediction decision.

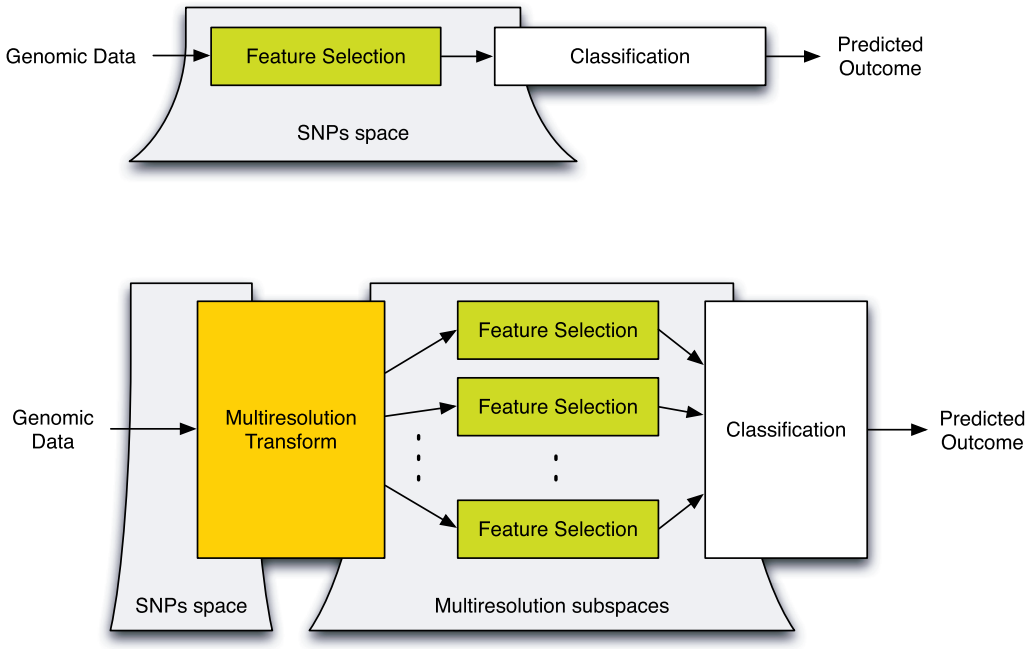


Fig. 1. *Top Panel:* Typical pattern recognition workflow for predicting clinical outcomes from genomic data. *Bottom Panel:* Pattern recognition workflow proposed to work with genomic data, using a multiresolution analysis block as a preprocessing step.

3.1 Training Algorithm

Define the genomic signal of an individual as a discrete set of samples $x = [x_1, x_2, \ldots, x_N]$ having a total of N base pairs or SNPs. For each loci, x_i, we count the number of minor alleles, i.e., $x_i \in \{0, 1, 2\}$. We refer to this N-dimensional space as the *SNP space*, or original space. For every data sequence x there is an outcome or class label, $y \in \{0, 1\}$, representing the membership of the sample to the unaffected or affected classes, respectively. Thus, we have a binary classification problem.

The training stage of the proposed approach is as follows: first a multiresolution transform is applied to the training set. We use a full decomposition transform and every subspace is kept for training, including the original SNP space. We have found that using a stationary wavelet transform gives better results than using orthonormal wavelets. In Section 5, we discuss our selection of a specific basis and decomposition.

We then continue in each multiresolution subspace with learning feature selection and training a classifier for each subspace. That is, a feature selection and standard classification algorithm are trained in each of the subspaces. The feature selection method used here ranks features (SNPs for the original space, but coefficients for the multiresolution subspaces) based on the χ^2 statistic [16]. In this way, we reduce each space from an N-dimensional input space to T-dimensional feature space, where the constant T can be learned by cross-validation. For each multiresolution subspace, a classifier is trained using its best T features. In this paper, we use logistic regression and random forests as classifiers [13], but any feature-extraction and classifier pairing may be used.

A final step in training requires learning the method to produce the final prediction probability or score by combing decisions from the subspace classifiers available. In our implementation we used a simple classifier combination rule, where the maximum score of a subset of subspace classifiers is output as the final prediction decision. To select which classifiers will be part of the subset, we rank each classifier using training cross-validation error, as measured using the area under the ROC curve (AUC). A discussion about the implementation of other classifier combination methods can be found elsewhere [13].

3.2 Prediction

The prediction or testing stage consists, first, on evaluating each subspace classifier as follows: for an input test sample, we perform multiresolution analysis to obtain the set of subspace coefficients necessary to test the subspace classifiers learned during training. From each multiresolution space, the corresponding T features are used to construct a feature vector, which is then input to the classifier of that subspace. The scores or probabilities of each of the subspace classifiers are computed and the maximum probability is output as the final decision and prediction score.

We compare this approach to the standard workflow of using the input SNP space only to obtain a final prediction score.

3.3 Computational Complexity

The computational cost of the multiresolution framework increases linearly with the number of subspaces in the multiresolution decomposition. We consider here costs for using an undecimated wavelet transform (UWT), but if an orthogonal discrete wavelet transform was to be used, costs would be much reduced.

For a wavelet decomposition of L levels, the number of subspaces is 2^L with L typically a small constant for

pattern recognition applications. For an input sequence of N samples, the number of operations of the UWT can be computed using a fast algorithm [15] requiring $O(LN)$ operations. We implement a full UWT decomposition, requiring $O(KN)$ operations, with $K = \sum_{n=0}^{L} 2^n$ and $2^L \ll N$ in our case. Since the computations of each multiresolution subspace can be computed independently from the others, these can be efficiently computed in parallel.

Besides the cost of the multiresolution preprocessing, we also need to consider the cost of implementing a classifier in each multiresolution subspace, and the method for combining the classifiers' decisions into a final prediction. Note that our interest lies in comparing the classification performance of an approach with multiresolution methods to an approach without it (standard classification). Since the multiresolution framework includes a classifier unit at every multiresolution space, and because such a classification process is the baseline of comparison, we can consider the computational costs of the multiresolution framework relative to the cost of such baseline classification process.

In training, let the process of learning feature selection and the classifier take $O(MP)$ time for a data set of M participants. Then, the total number of operations for K multiresolution spaces is $O(KMP + KMN)$. As mentioned above, to this amount we need to add the computational cost of learning a classifier combination method. When such combination method is a voting scheme or a simple combination rule among a subset of subspace classifiers, the process of selection or ranking of the classifiers will add computations to training. While using metrics obtained during training of the classifiers can save computations, alternatively, one can use performance metrics obtained using cross-validation. In either case, these computations can be parallelized.

Finally, let the computations for testing a single patient case using the predetermined classification process be $O(q)$. Then, prediction of the outcome for a single case using a multiresolution decomposition, with K subspace classifiers in which individual decisions are combined using a voting scheme or a simple rule (as outlined in training), will be $O(Kq + KN)$.

4 Experiments and Results

4.1 Alzheimer's Data Set

We used a the data set collected for the study of late-onset Alzheimer's disease, available from the Translational Genomics Research Institute [17]. The data set contains information on 1411 individuals, with 861 being affected with LOAD (cases) and 550 being unaffected (controls).

An Affymetrix chip was used to genotype 502,627 SNPs for each person, and after applying quality controls a total of 312,316 SNPs were retained for analysis [17]. To this data set, two alleles were added for chromosome 19, namely, rs429358 and rs7412, that encode the $\epsilon 4$ APOE genotype. These alleles have been shown to be associated with LOAD [17], but for technical reasons they are difficult to measure using SNP arrays.

The genetic influences on LOAD are not completely understood, but strong association has been found to genetic factors. Specifically, the apolipoprotein E (APOE) gene has been consistently identified as a genetic risk factor for LOAD. From recent GWA studies, the $\epsilon 4$ APOE genotype, located in chromosome 19, has been associated with increased risk of developing Alzheimer's disease [17]. We, therefore, focused our use on chromosome 19 SNP data. We refer to this subset as TGEN19.

Missing values were imputed using fastPhase [18] by analysing nonoverlapping windows close to 1,000 SNPs long, without haplotype estimation. PLINK [19] was used to convert the genomic data to an additive genomic model.

4.2 Experiment Settings

In the proposed method, there are three parts that need specification: the multiresolution transform (MR), the feature extraction method for each subspace and the classification method used. The latter is actually a two-stage process, in which every subspace classifier produces a score or probability, and then these are combined to produce a final outcome prediction.

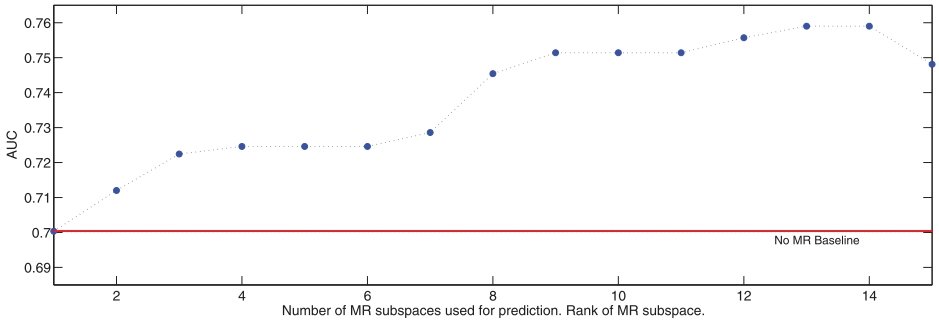
The multiresolution transform we used is the stationary wavelet transform [20], which we refer to as the undecimated discrete wavelet transform, and we use the Haar basis to perform multiresolution analysis. The feature selection method we used is a ranking procedure applying the χ^2 statistic [21]. For each MR subspace, we select the top 10 SNPs according to this ranking, which defines our feature space. Finally, a classifier is trained in each MR feature space, and we report on using either logistic regression or random forests as subspace classifiers. In this way, we have an output probability for each MR subspace from testing a single participant. As a classifier combination rule, the maximum of these output probabilities is used as the final classification score.

A variation to this algorithm is to include a multiresolution pruning algorithm that selects MR subspaces according to how well a classifier can be trained on them, or some measure of classifier performance. Instead of using all MR subspaces or selecting a particular subset of MR subspaces toward computing the final outcome score, we show performance results as we increase the number of MR subspaces that are included. The order on which these MR spaces are included (or pruned from taking part on the final classification decision) is learned through cross-validation in training using the area under the ROC curve.

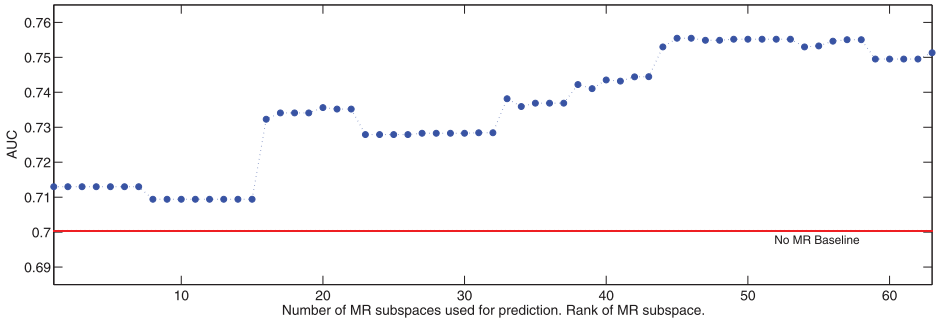
In these experiments we randomly held out about a fifth of the TGEN19 data set for testing and used the rest for training.

4.3 Results

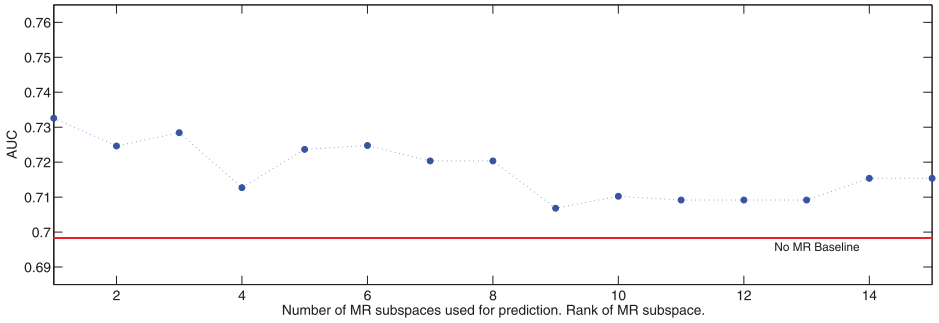
We evaluate performance on the TGEN19 data set using the area under the ROC curve (AUC) as a measure of performance. In general, we have four types of experiments: we explore training the proposed algorithm for two depths or levels of MR decomposition, $L = \{3, 5\}$, and we use either logistic regression or random forests as a subspace classifier. The effect of increasing the level of decomposition is simply that this makes more MR subspaces available, from which some may obtain better predictive performance than their parents. These results are to be compared to the case of standard classification on the original SNPs data.



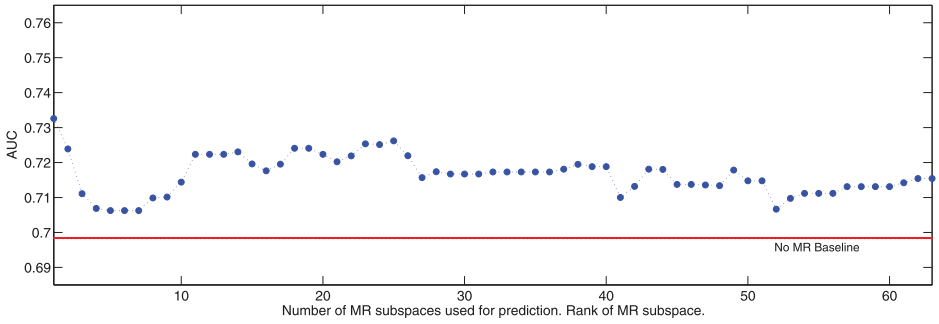
(a) *MR3LR*: Logistic regression MR classifiers ($L = 3$)



(b) *MR5LR*: Logistic regression MR classifiers ($L = 5$)



(c) *MR3RF*: Random forests MR classifiers ($L = 3$)



(d) *MR5RF*: Random forests MR classifiers ($L = 5$)

Fig. 2. AUCs obtained by the proposed MR algorithm when classifiers are trained in selected subspaces according to their rank, using an L-level full MR decomposition (Haar basis). The MR subspace classifiers are ranked in training, including the classifier trained at the original SNPs space. See Table 2 for a list of the MR subspaces ordered by their classifier's rank.

Fig. 2 shows the AUCs obtained by the proposed algorithm as more MR spaces are included in the final prediction process. Figs. 2a, and 2b show results for the case of logistic regression, when the MR level is $L = 3$ and $L = 5$, respectively. Similarly, Figs. 2c, and 2d, show results for the case of random forests. In these figures, the x-axis is the number of MR subspaces selected, from one to the total number of spaces available in the full packet decomposition, in the order they were ranked in training (using AUC performance of their classifiers in a cross-validation setting), including the original SNPs space. These rank orders are different for all the four cases, and the list of subspaces for each case is shown in

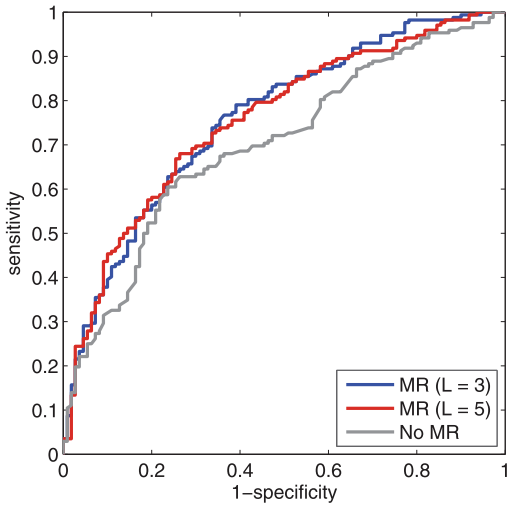


Fig. 3. ROCs for the case of classification with logistic regression: *(blue)* Proposed algorithm for the case of using a three-level decomposition Haar frame. In this case, using the top 13 MR spaces ranked in training gave the larger AUC. *(red)* Proposed algorithm for the case of using a five-level decomposition Haar frame. In this case, using the top 45 MR spaces ranked in training gave the larger AUC. *(gray)* Logistic regression applied on the SNPs space (using no multiresolution).

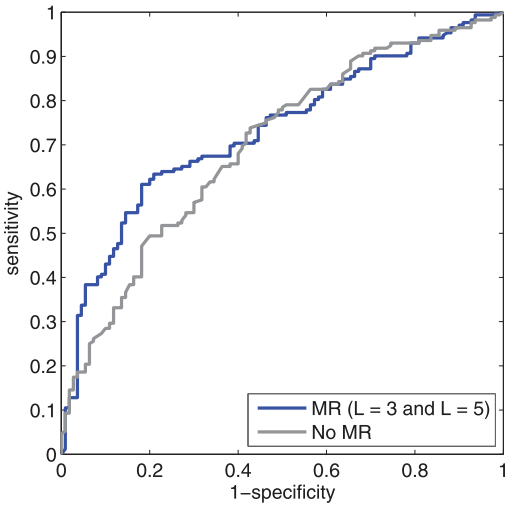


Fig. 4. ROCs for the case of classification with logistic regression: *(blue)* Proposed algorithm for the case of using a three-level decomposition Haar frame. In this case, using the top (rank-1) MR space, as ranked in training, gave the larger AUC for both cases, selecting spaces over a 3-level decomposition and a 5-level decomposition. *(gray)* Random forests applied on the SNPs space (using no multiresolution).

Table 2—here we have used the common terminology of naming the MR space by the sequence of *unit*-filterbank branches applied to obtain it, namely, at each level one can obtain either detail coefficients, applying an "H"-branch, or approximation coefficients, applying an "L"-branch. The original SNPs space has been labeled "SNPs."

An important overall observation from these results is that regardless of the number of subspaces that are used to predict with the proposed algorithm, performance was the same or better than the baseline of using logistic regression or random forests trained on the original SNP space. However, when logistic regression is used as a classifier, the proposed algorithm is able to improve as more MR-subspace classifiers are considered for a final classification decision (based on ranking), for almost all subspaces. Nevertheless, this trend is not observed with random forests; the AUC performance does not increase after the first (rank-1) subspace is included. The ROC curves for the cases when we obtain best performance are shown in Fig. 3 for logistic regression, and Fig. 4 for random forests.

The maximum classification performance was obtained with an MR transform of 3-levels training logistic regression, $AUC = 0.759$, while for the case for random forests we obtain an $AUC = 0.732$. By comparison, with no multi-resolution we have $AUC = 0.700$ when using logistic regression, and an $AUC = 0.698$ when using random forests (the difference between the latter baseline algorithms was not found to be significant at the 0.05 level). Table 1 shows relevant AUC comparisons and their statistical significance using DeLong's test [22].

5 Discussion

5.1 Results

The MR framework with logistic regression achieved an improvement in predictive performance as measured by the area under the ROC curve (AUC). The difference between using an MR framework and classifying on the original SNP space is approximately a 6 percent increase in ROC area, while with random forests this increase was of about 3 percent (see Table 1). For the case of logistic regression MR classifiers, we are able to improve the performance starting from the rank-1 space and then adding MR spaces; however, this was not the case of random forests.

It is interesting that in the case of logistic regression the AUC trend is almost always increasing. Since this may not always be the case, as shown using random forests, this result emphasizes the importance of the strategy that is used for selecting MR spaces in the design of the algorithm. Specifically, the ranking order of the multiresolution spaces is performed in training (this order is shown in Table 2), where they are ranked on the basis of training AUC using a nested training cross-validation approach. It may be that

TABLE 1
Statistical Significance of AUC Difference of Main Results

Classifier	Description	AUC		Description	AUC	p-value
Logistic Regression	No MR (original SNPs space)	0.700	VS	MR 3-levels using top-13 ranked subspaces	0.759	0.00001824
Logistic Regression	MR 3-levels rank-1 subspace	0.700	VS	MR 3-levels using top-13 ranked subspaces	0.759	0.00001824
Logistic Regression	No MR (original SNPs space)	0.700	VS	MR 5-levels using top-45 ranked subspaces	0.755	0.00006917
Logistic Regression	MR 5-levels rank-1 subspace	0.713	VS	MR 5-levels using top-45 ranked subspaces	0.755	0.00035347
Random Forest	No MR (original SNPs space)	0.698	VS	MR 3-levels and 5-levels using rank-1 subspace	0.732	0.13870555

TABLE 2
Top 50 Multiresolution Spaces as Ranked in Training

Rank	MR3LR	MR5LR	MR3RF	MR5RF
1	**SNPs**	LHLHH	LL	LL
2	HH	HHLLH	HH	LLHH
3	LHH	LHHLH	**SNPs**	LHLLH
4	LL	LHHHL	LLL	LHHLL
5	HLH	HLLHH	HHH	HLLLH
6	HHL	HLHLH	L	HLLHL
7	HHH	HLHHL	HLH	HLHLL
8	H	LHHLL	HHL	LHLHL
9	LLH	HHLHL	H	HHLLL
10	LHL	LHLLH	LHH	LLHHL
11	HLL	HLLLH	LLH	HLHH
12	LLL	HLLHL	LHL	HHLH
13	LH	HLHLL	HLL	HHHL
14	HL	LLHHH	LH	LLLHH
15	L	HHHLL	HL	LLHHH
16		HHHHH		LLHLH
17		LLHHL		HHLLH
18		HHLLL		LHLLL
19		LLLHH		HLLLL
20		LHHH		HHLHL
21		LLHLH		HH
22		LHLHL		LHHH
23		**SNPs**		LHLL
24		HLHH		**SNPs**
25		HHLH		HHHLL
26		HHHL		LLL
27		HLHHH		HHHH
28		HHLHH		LHLHH
29		HHHLH		LLLH
30		HHHHL		LLHL
31		HH		HLLL
32		LHHHH		LHHLH
33		LHH		LHHHL
34		LLLL		HLLHH
35		LL		HLHLH
36		HLH		HLHHL
37		HHL		LHHL
38		LHHL		HHHHH
39		HHLL		LLLLH
40		HHH		LLLHL
41		LLHH		HHLL
42		HHHH		LLLL
43		LHLH		HHH
44		H		LHLH
45		HLLH		HLLH
46		HLHL		HLHL
47		LHLLL		LLHLL
48		HLLLL		LLLLL
49		LLH		L
50		LHL		HLH

during this process of training cross-validation, the amount of training data favored logistic regression over random forest. This would translate into logistic regression having a ranking of multiresolution spaces that is more representative of the (unknown) *true* ranking as compared to random forest. In general, when using SNP data sets, we have typically observed that logistic regression fares the same or better than random forest.

From Fig. 2 we also see that this trend is similar for both transform level cases, namely, the 3-level decomposition and the 5-level decomposition, for each classification method. The maximum AUC achieved in each L-level decomposition experiment was not significantly different for that particular MR classifier. Therefore, to save computation time, a 3-level decomposition should be preferred over the 5-level decomposition, for predicting LOAD using this data set.

In summary, the proposed algorithm improved the AUC by using better *input-space* representations than the original SNPs. This is achieved by moving the feature space from the SNP space to a multiresolution space. These multiresolution spaces consist of multiresolution coefficients that encode the interaction between neighboring SNPs, since the coefficients are functions of SNPs. Using these coefficients, our algorithm classifies using functions of the interactions of neighboring SNPs, which happen to produce better prediction than using the raw SNPs. Finally, note that neighbor SNPs are usually highly correlated due to *linkage disequilibrium*. Thus, linkage disequilibrium is inherently exploited by the proposed algorithm. Our method shows that functions of linkage disequilibrium, as implemented by applying a multiresolution transform, can improve predictive performance.

TABLE 3
Top 10 SNPs of TGEN19 Data Set

Rank	SNP ID	χ^2	p
1	SNP ApoE1	229.9371	$\simeq 0$
2	SNP A-2236481	148.1748	$\simeq 0$
3	SNP ApoE2	36.8921	9.7493e-09
4	SNP A-1816239	17.4409	1.6321e-04
5	SNP A-2086668	15.3019	4.7560e-04
6	SNP A-1874859	15.2830	4.8010e-04
7	SNP A-2005924	15.1791	5.0572e-04
8	SNP A-2223004	15.0070	5.5114e-04
9	SNP A-1783210	14.9589	5.6456e-04
10	SNP A-1961684	13.0982	1.4314e-03

5.2 Genomic Markers

Using χ^2 statistics of the original SNP data, Table 3 shows the top 10 SNPs. The SNPs with strongest association among these are ApoE1, A-2236481, and ApoE2. SNPs ApoE1 and ApoE2, also known as rs429358 and rs7412, respectively, combine to define the ApoE4 variant of the ApoE gene, which has been repeatedly reported to be associated with high risk of LOAD [17].

The SNP A-2236481 (rs41377151/rs4420638), in the adjacent gene known as ApoC1, has also been associated with high risk of LOAD [23].

To the best of our knowledge, the remainder of the SNPs in Table 3 have not been previously reported as being significant predictors of LOAD.

5.3 Assumptions

There are several assumptions that we made to apply multiresolution analysis to genomic data as available from SNP data sets.

In the treatment of the genome as a 1D signal, we used the fact that SNPs are sampled sequentially using a 1D path (or as if it was a linear path, later assembled) that can be untangled into a discrete linear lattice. We assume that this linear arrangement of subsequent SNPs is an important element of the biology of the genome.

Also, SNPs are polymorphisms at loci that may not be evenly spaced. This depends on the genotyping technology; however most SNPs data sets now are genotyped using *tagging SNPs* to maximize genomic coverage, and these loci are not spatially uniform [24]. Moreover, due to quality control procedures, some SNP loci may be discarded. In this paper, we implicitly defined the SNP space after QC as the input to multiresolution analysis. The study of the effect of the nonuniform sampling of genotyping platforms is left as future work.

When working with genomic data, a genomic model has to be used to map each base pair to a number. Under an additive genomic model, the presumed map can be written as $AA = 0, AB = 1, BB = 2$, where A is the minor allele, and B is the major allele [21].

Finally, the multiresolution transformation is applied to all SNP data, person by person. From each multiresolution subspace we have selected the top 10 coefficients as our features for classification. There are as many classifiers as there are multiresolution subspaces and a final classification decision is obtained by taking the best classification score among them, as described in Section 3.2. We set the same number $N = 10$ for every space for simplicity and to show the robustness of classifying with multiresolution features. $N = 10$ worked best for both logistic regression and random forest on SNP data, which are our baseline for comparisons. An advanced version of the algorithm that we do not address here could search for a *best N* for each subspace in training, at an additional computational expense.

6 Conclusion

We have shown that expanding the original input SNPs space using a multiresolution transform and combing decisions of individual classifiers trained on subspace coefficients or features can improve predictive performance of LOAD.

A significant improvement is shown when logistic regression is trained to classify in every subspace as compared with standard prediction using logistic regression on the original SNPs space. When either logistic regression or a random forest classification strategy is implemented in the multiresolution framework, results show performance is improved significantly and is at least as good as classifying with no multiresolution.

The results in this paper support that multiresolution methods are a promising approach for improving the prediction of clinical outcomes from GWAS data. In future work, it will be important to evaluate the multiresolution framework using other GWAS data sets on a variety of clinical and biological outcomes.

Acknowledgments

The authors thank Dr. Michael Barmada for discussions in the early development of this work, Dr. Shyam Visweswaran for discussions related to the data set described in this paper, and Mr. Kevin Bui for providing technical support. This research was funded by grant R01-LM010020 from the National Library of Medicine of the National Institutes of Health. P.H. Hennings-Yeomans now works at the Ontario Institute for Cancer, Toronto, Ontario, Canada.

References

[1] M. Stephens and D.J. Balding, "Bayesian Statistical Methods for Genetic Association Studies," *Nature Rev. Genetics,* vol. 10, no. 10, pp. 681-690, Oct. 2009.

[2] J. Couzin-Frankel, "Major Heart Disease Genes Prove Elusive," *Science,* vol. 328, no. 5983, pp. 1220-1221, June 2010.

[3] H.J. Cordell, "Detecting Gene-Gene Interactions that Underlie Human Diseases," *Nature Rev. Genetics,* vol. 10, no. 6, pp. 392-404, June 2009.

[4] C.J. Hoggart, J.C. Whittaker, M. De Iorio, and D.J. Balding, "Simultaneous Analysis of All SNPs in Genome-Wide and Re-Sequencing Association Studies," *PLoS Genetics,* vol. 4, no. 7, p. e1000130, July 2008.

[5] Y. Zhang and J.S. Liu, "Bayesian Inference of Epistatic Interactions in Case-Control Studies," *Nature Genetics,* vol. 39, no. 9, pp. 1167-1173, Aug. 2007.

[6] A. Chebira, "Adaptive Multiresolution Frame Classification of Biomedical Images," PhD dissertation, Carnegie Mellon Univ., 2008.

[7] P. Hennings, J. Thornton, J. Kovačević, and B.V. Kumar, "Wavelet Packet Correlation Methods in Biometrics," *Applied Optics,* vol. 44, no. 5, pp. 637-646, 2005.

[8] P. Phillips, "Matching Pursuit Filters Applied to Face Identification," *IEEE Trans. Image Processing,* vol. 7, no. 8, pp. 1150-1164, Aug. 1998.

[9] S. Hutter, A.J. Vilella, and J. Rozas, "Genome-Wide DNA Polymorphism Analyses Using VariScan," *BMC Bioinformatics,* vol. 7, article 409, Sept. 2006.

[10] D. Botstein, R.L. White, M. Skolnick, and R.W. Davis, "Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms," *Am. J. Human Genetics,* vol. 32, no. 3, pp. 314-331, May 1980.

[11] R. Gibbs, "The International HapMap Project," *Nature,* vol. 426, no. 6968, pp. 789-796, Dec. 2003.

[12] J. Ragoussis, "Genotyping Technologies for Genetic Research," *Ann. Rev. Genomics and Human Genetics,* vol. 10, no. 1, pp. 117-133, Sept. 2009.

[13] C.M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[14] M. Vetterli, J. Kovačević, and V. Goyal, *The World of Fourier and Wavelets: Theory, Algorithms and Applications,* http://FourierAndWavelets.org, 2011.

[15] S. Mallat, *A Wavelet Tour of Signal Processing.* Academic Press, 1999.

[16] R.L. Plackett, "Karl Pearson and the Chi-Squared Test," *Int'l Statistical Rev.,* vol. 51, no. 1, pp. 59-72, Apr. 1983.

[17] E.M. Reiman et al., "GAB2 Alleles Modify Alzheimer's Risk in APOE E4 Carriers," *Neuron,* 54, no. 5, pp. 713-720, June 2007.

[18] P. Scheet and M. Stephens, "A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase," *The Am. J. Human Genetics,* vol. 78, no. 4, pp. 629-644, Apr. 2006.

[19] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, and P.C. Sham, "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses," *The Am. J. Human Genetics,* vol. 81, no. 3, pp. 559-575, Sept. 2007.

[20] R.R. Coifman and D. Donoho, *Translation-Invariant De-Noising,* pp. 125-150, Springer-Verlag, 1995.

[21] C.M. Lewis, "Genetic Association Studies: Design, Analysis and Interpretation," *Briefings in Bioinformatics,* vol. 3, no. 2, pp. 146-153, 2002.

[22] D.M.D. Elizabeth, R. DeLong, and D.L. Clarke-Pearson, "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach," *Biometrics,* vol. 44, no. 3, pp. 837-845, Sept. 1988.

[23] J.H. Lee et al., "Analyses of the National Institute on Aging Late-Onset Alzheimer's Disease Family Study: Implication of Additional Loci," *Archives of Neurology,* vol. 65, no. 11, pp. 1518-1526, Nov. 2008.

[24] M.L. Metzker, "Sequencing Technologies - The Next Generation," *Nature Rev. Genetics,* vol. 11, no. 1, pp. 31-46, Jan. 2010.

Pablo H. Hennings-Yeomans received the BS degree in electronics and communications engineering and the MS degree in electronic systems, both from ITESM, Monterrey Campus, in 1998 and 2002, respectively, and the PhD degree in electrical and computer engineering from Carnegie Mellon University in 2008. He has worked as a postdoctoral researcher and consultant for the Center for Bioimage Informatics at Carnegie Mellon. From 2009 until 2011, he was a postdoctoral associate in the Department of Biomedical Informatics in the University of Pittsburgh, where his work focused on genome-wide association studies. His research interests include biomedical imaging, biometrics, genomic pattern recognition, computational genomics and cancer bioinformatics. In 2011, he joined the Ontario Institute for Cancer Research as a bioinformatics research scientist. He is a member of the IEEE.

Gregory F. Cooper received the BS degree in computer science from MIT in 1977, the PhD degree in medical information sciences from Stanford University in 1985, and the MD degree from Stanford in 1986. He is a professor in the Department of Biomedical Informatics at the University of Pittsburgh. His research interests include the use of decision theory, probability theory, Bayesian statistics, and artificial intelligence to address biomedical informatics research problems, with current work on automated methods for predicting patient outcomes from clinical and genome-wide data, clinical alerting based on machine learning, and detecting disease outbreaks (biosurveillance) and other complex patterns from clinical data.