ELSEVIER

Integrating domain knowledge with statistical and data mining methods for high-density genomic SNP disease association analysis

Valentin Dinu [a,b,*], Hongyu Zhao [c,d], Perry L. Miller [b,e,f]

[a] *Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA*
[b] *Center for Medical Informatics, Yale University School of Medicine, PO Box 208009, New Haven, CT 06520-8009, USA*
[c] *Department of Epidemiology and Public Health, Yale University, New Haven, CT, USA*
[d] *Department of Genetics, Yale University, New Haven, CT, USA*
[e] *Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT, USA*
[f] *Department of Anesthesiology, Yale University, New Haven, CT, USA*

Abstract

Genome-wide association studies can help identify multi-gene contributions to disease. As the number of high-density genomic markers tested increases, however, so does the number of loci associated with disease by chance. Performing a brute-force test for the interaction of four or more high-density genomic loci is unfeasible given the current computational limitations. Heuristics must be employed to limit the number of statistical tests performed.

In this paper we explore the use of biological domain knowledge to supplement statistical analysis and data mining methods to identify genes and pathways associated with disease. We describe Pathway/SNP, a software application designed to help evaluate the association between pathways and disease. Pathway/SNP integrates domain knowledge—SNP, gene and pathway annotation from multiple sources—with statistical and data mining algorithms into a tool that can be used to explore the etiology of complex diseases.
© 2007 Elsevier Inc. All rights reserved.

Keywords: False discovery rate (FDR); Data integration; Data mining; Genome-wide association (GWA); Pathway-based disease association; Single nucleotide polymorphisms (SNP)

1. Introduction

This paper describes Pathway/SNP, a software application that allows its user to utilize pathway data in the analysis of high-density genomic SNP data derived from disease association studies. Our goal in building Pathway/SNP is to allow the user to explore flexibly the etiology of disease using different statistical and data mining algorithms. The task of utilizing pathway data to help in the interpretation of genomic SNP data is complex. There are many statistical and data mining approaches that can be applied, and new approaches are being developed. There

is no single approach that is in some sense best. Different approaches may work better with different data sets and with different diseases. As a result, it is important to provide the user with a flexible set of tools that allow different approaches to be tried incrementally and interactively.

Indeed, the challenge of using pathway knowledge to assist in the interpretation of genomic SNP association data is one instance of a broader challenge. There are large amounts of very diverse biological knowledge being generated that can potentially be used integratively to help analyze high-throughput data. Examples include pathway information, tissue and cellular component localization, gene expression information, and regulatory networks. This knowledge can help drive the analysis of genomic data for many different purposes. In the case of Pathway/SNP, the purpose is to analyze the underlying etiology of disease

---

* Corresponding author. Fax: +1 203 737 5708.
E-mail address: valentin.dinu@yale.edu (V. Dinu).

through the integration of pathway information. At the same time, there is a wide range of useful statistical and data mining approaches that can be used when integrating biological knowledge in the analysis of such genomic data.

The goals of this paper are (1) to describe the current implementation of Pathway/SNP and its functionality, (2) to discuss some of the lessons learned during its implementation concerning the use of different statistical and data mining algorithms, and (3) to discuss some of the implications for building interactive tools in the future that allow such algorithms to be flexibly adapted to integrative genomic analysis.

2. Background

Over the past decade, many studies have used microarrays to analyze gene expression profiles associated with disease, most notably cancer [1–3]. Over the past 3 years, newly developed high-density single nucleotide polymorphism (SNP) microarray technology has brought within reach the promise of performing genome-wide association (GWA) studies to identify genomic mutations that are associated with a wide range of diseases. The results from several successful GWA studies have already been published, identifying mutations linked to age related macular degeneration (AMD) [4] and type 1 diabetes [5]. Other large-scale GWA studies are in progress, e.g., Welcome Trust Case Control Consortium's analysis of 19,000 DNA samples at 675,000 genome-wide SNPs [6].

Complex diseases, e.g., diabetes, hypertension, Alzheimer's disease, and AMD, are believed to be caused by the interaction of multiple genes and environmental factors. The number of mathematical operations required to assess the association between multiple interacting genomic loci and disease grows exponentially with the number of interacting SNPs. Simple arithmetic calculations (Appendix A) show that even with the most powerful supercomputers available today it is computationally impossible to perform a comprehensive test of association for four or more interacting SNPs when analyzing data sets for several hundred individuals and several hundred thousands SNPs. As a result, a variety of statistical, computational, heuristic, and knowledge-based approaches (often combined) must be developed to compensate for the inability to perform all potentially desirable computations.

Various statistical approaches to addressing the issue of multiple genomic loci association with quantitative traits (e.g., diseases) have shown interesting, if sometimes inconclusive, results. Storey and colleagues recently proposed a computationally efficient stepwise algorithm to identify genetic loci associated with gene expression in yeast [7]. In this algorithm, at each step the genomic locus with the strongest association to the trait is selected from the entire set of loci. The authors found, to their surprise, that the stepwise method was more powerful than the exhaustive method performing the association test over all pairs of loci. The authors acknowledged, however, that their method could miss locus pairs with primarily epistatic (and little individual locus) effect on the trait [7]. In another recent study, Marchini and colleagues [8] concluded that the determination of one single best strategy in a multi-locus disease association model might not be possible. Different strategies can perform better in different conditions based on varying parameters such as allele frequencies or locus interaction models (e.g., additive vs. multiplicative effects). Moreover, the authors determined that the same varying parameters can affect the ability to replicate the association of interacting loci across different studies. Using a set of simulated data, the authors found that in the largest fraction of configurations tested, the comprehensive testing of all paired loci interactions had the highest statistical power [8].

Roeder and colleagues [9] recently proposed integrative analysis using genetic linkage data to weigh the association p-values. The authors proposed dividing the association p-values by weights that are larger than 1 in regions previously linked with the disease and smaller than 1 in unlinked regions. Similarly, the use of biological knowledge, such as pathway information, can help researchers focus on genes involved in processes known to be related to disease. Such thinking underlies many candidate gene association analyses. For example, a recent investigation attempted to use pathway level information to model the association between various genes and disease and the interaction among genes, and between genes and environmental factors [10]. In another recent study, Schaid and colleagues proposed a non-parametric test of association between disease and a set of pathway genes using U-statistics [11]. The authors applied their method to investigate the association between prostate cancer and a group of several dozen SNPs selected from two metabolic pathways (androgen and estrogen response elements). In this example, the authors determined that their method could identify the combined disease association of multiple markers with small effects while a single locus association was not significant after Bonferroni correction for multiple testing. There are analogous efforts in functional genomic studies to use pathway information to better understand disease etiology [12,13].

A number of data mining approaches, such as dimensionality reduction [14,15], neural networks [16,17], random forests [18], and support vector machines (SVM) [19,20], have also been proposed for multi-locus association with traits. Data mining approaches are well-suited for identifying trends in large dimensional data sets. Some, however, can be computationally intensive (e.g., SVM, neural networks) and some may be hard to interpret (e.g., neural networks).

One conclusion that can be drawn from these previous studies is that the "one model fits all" approach is likely not optimal when analyzing the association between genomic loci and disease. We developed Pathway/SNP, a software tool that allows an exploratory approach to integrative association analysis, combining the use of biological pathway information and different statistical and

data mining algorithms. Pathway/SNP integrates pathway information, gene annotation, and SNP location to identify the pathways that are the most strongly associated with disease. The association between a pathway and disease can be quantified by *z*-scores generated by *U*-statistics, or by the percent of correct disease class assignments using data mining algorithms. The users can further drill down on specific pathways to explore which genomic loci contributed most to the pathway association score by performing association tests solely on one pathway's SNPs. The statistical significance can be assessed via permutation-based false discovery rate (FDR).

3. System description

This section describes the Pathway/SNP system and its various components. A later section then provides an example of how the system was used to analyze an example data set.

3.1. Design objectives

Pathway/SNP is designed as an exploratory tool. It must be able to perform basic operations such as performing an association test using a desired algorithm, displaying and sorting results. For more advanced exploration, users can modify algorithm parameters or filter the input SNPs to be analyzed (e.g., based on MAF) and/or the patients to be analyzed (e.g., based on disease stage, if such information is available).

The response time must be fast for simple queries such as single SNP or single pathway association tests. Since some algorithms can be computationally intensive (e.g., data mining algorithms such as neural networks, SVM or Random Forests, or FDR calculation using a large number of permutations), the users can choose to have those results saved to the file system and be notified via email about job completion.

The biological annotation data must be well curated and up to date. For example, since the reference human genome build changes over time, the SNP and gene annotations must be appropriately labeled with the different genome builds on which they are based.

3.2. Architecture

Pathway/SNP is written in Java and has a 3-tier architecture (Fig. 1): (1) presentation tier (GUI written in Java Server Pages (JSP) and accessible via a standard Internet browser); (2) logic tier (statistical and data mining algorithms written in Java); and (3) data tier (genotype, phenotype and annotation data stored in a heavily indexed relational database). The 3-tier architecture allows for flexibility in deployment. The application can be installed and run on a stand alone computer, or it can run on a parallel computing cluster to meet scalability requirements.

3.3. Biological data

The pathway annotation data are loaded in the application database. There are currently annotations for 561 pathways: 181 KEGG [21], 314 BioCarta [22] and 66 Gen-MAPP [23] human pathways. Some of these pathways overlap. For example, KEGG contains a "complement
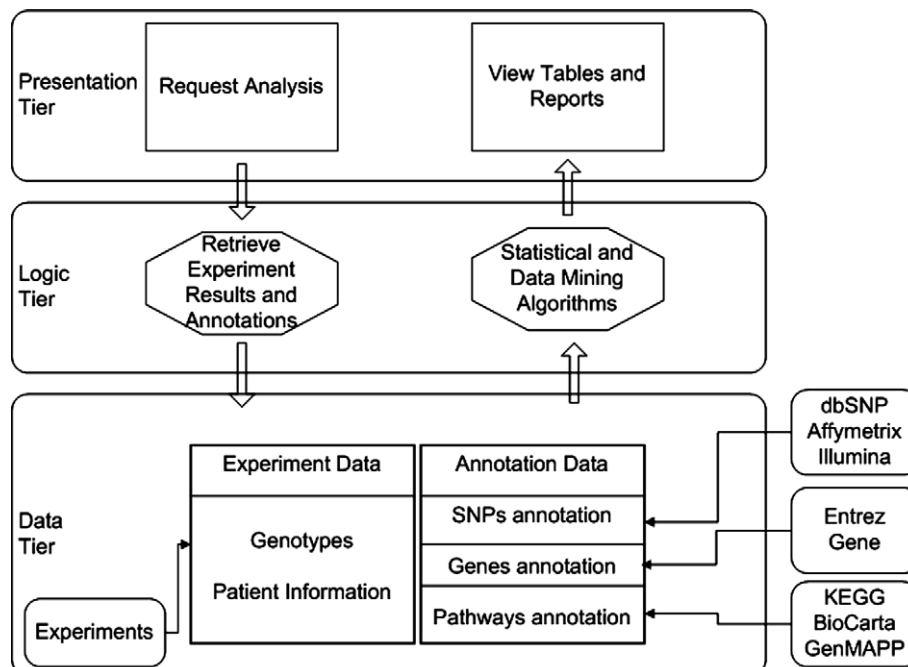

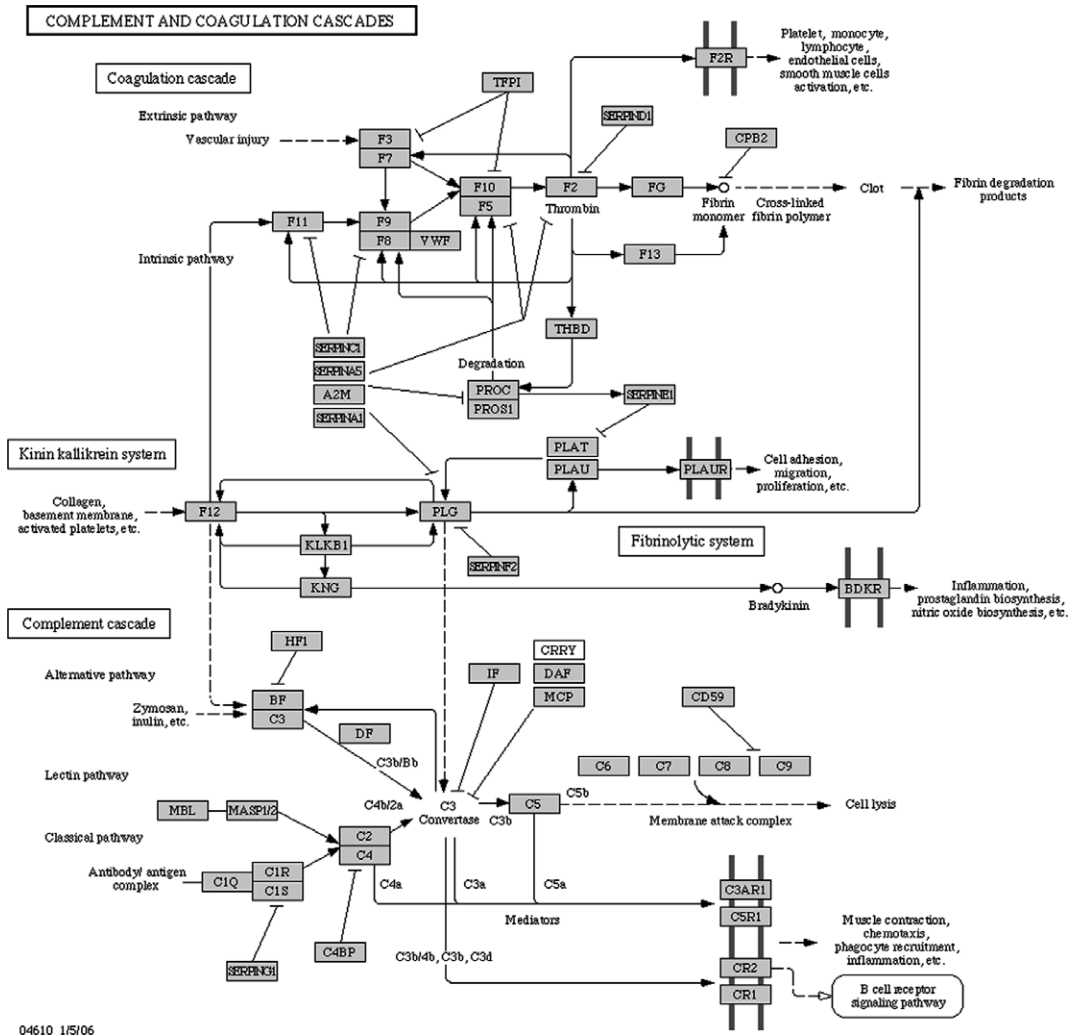Fig. 1. Pathway/SNP's 3-tier architecture.

Fig. 2. A schematic of the Complement and Coagulation Cascades Pathway. Source: KEGG Database website [21].

and coagulation cascades'' pathway (Fig. 2). The corresponding information is represented in BioCarta under multiple pathways: ''alternative complement pathway,'' ''classical complement pathway,'' ''lectin induced complement pathway,'' ''complement pathway'' (including the first three), ''fibrinolysis pathway,'' ''cells and molecules involved in local acute inflammatory response,'' etc. Some pathway annotation sources are constantly evolving (e.g., KEGG is re-built nightly, with new additions or modifications performed quite frequently), while others are less often updated. Pathway/SNP users have the option of connecting straight to the pathway annotation servers, such as KEGG's web service server [24], to use in analysis the most recent pathway information. Exercising this option can increase the total time of execution. It takes about 30 min to retrieve the annotation data for 181 pathways directly from the KEGG web service server using the SOAP API.

The gene annotation data are loaded locally from NCBI's Entrez Gene [25]. The Affymetrix 100k and 500k GeneChip microarray annotation files are preloaded in the database. Annotations for any additional microarray

platforms (e.g., Illumina HumanHap550) can also be loaded into the database. The annotations for genes and SNPs involve genomic positions, which change when new human genome references are released. As a result, one needs to keep track of the different genome releases to which the annotations apply in order to avoid mixing data from different genome builds.

By default, SNPs are considered relevant to a given biological pathway in Pathway/SNP if they are within 10,000 base pairs (bp) of a pathway gene's location. (Users may modify this default distance.) Using this default distance, for example, we identified 94 SNPs on the Affymetrix 100k GeneChip that were in the vicinity of 46 complement pathway genes.

Additionally, users can create their own sets of genes or SNPs to be used in analysis. They can store sets of genes as ''manually curated'' pathways in the database, or they can directly use sets of genes or sets of SNPs in the database queries. For example, the set of 46 complement pathway genes mentioned above is ''manually curated'' in the sense that we started with the genes in the

complement pathway in the KEGG database and we augmented it with genes selected from literature and Entrez Gene [26].

The genotype and patient information are stored in the database using the entity-attribute-value (EAV) data modeling approach. As a result, the database schema does not need to change when loading experimental data from studies that use different types of patient-descriptive data. The SNP information, however, is stored in conventional relational tables. As a result, the database uses a mixed schema data model [27].

3.4. Algorithms

The algorithms included in Pathway/SNP can be roughly divided into three categories: (1) single SNP association with disease, (2) pathway association with disease, and (3) permutation-based statistical significance inference. We started the development of Pathway/SNP by first implementing simpler algorithms, for single SNP association with disease, and gradually expanded to include more complex algorithms capable of incorporating biological domain knowledge. Additional algorithms could be added in the future, e.g., Fisher's exact test for single SNP association analysis, other data mining algorithms and/or packages.

3.4.1. Single SNP association with disease

We have included two algorithms for the exploration of association between single SNPs and disease: chi square and Armitage's trend test [28]. These algorithms do not utilize SNP, gene or pathway annotation and simply assess each individual's SNP association with disease using the basic genotype data, without regard to the SNP's genomic information such as neighboring SNPs or genes. These algorithms can be applied to one SNP at a time or to all genotyped SNPs at once.

The chi square test can be allele-based, using a 2×2 contingency table: (case, control) $\times$ (allele A count, allele B count), or it can be genotype-based, using a 2×3 contingency table: (case, control) $\times$ (genotype AA count, genotype AB count, genotype BB count). Since the allele-based chi square statistic has 1 degree of freedom, it is generally preferred over the genotype-based chi square statistic that has 2 degrees of freedom.

Armitage's trend test [28] is a more complex algorithm that allows the user to explore different models of association between a SNP and disease. For example, an additive interaction model can be specified by using the following multipliers: 0 for homozygous non-risk alleles, 1 for heterozygous alleles, 2 for homozygous risk alleles. Similarly, a dominant interaction model can be specified by using multipliers 0, 1, and 1, while a recessive model can use 0, 0, and 1, respectively. The trend test statistic has 1 degree of freedom.

Any statistical result obtained from using either of the association tests on multiple SNPs must then be adjusted in order to account for bias introduced by the multiple testing (e.g., there are $m = 116,204$ SNPs on the Affymetrix 100k GeneChip). One can apply this correction by using the Bonferroni adjustment (roughly, by multiplying the *p*-value by the value of markers tested, *m*), or by utilizing the FDR procedure described in Section 3.4.4.

3.4.2. U-statistics for pathway association with disease

We have implemented the non-parametric algorithm based on *U*-statistics proposed by Schaid and colleagues [11]. Briefly, the algorithm allows the use of different kernels (e.g., recessive, dominant and linear dosage) to explore the association between a set of SNPs (e.g., pathway SNPs) and disease. It also accounts for the correlation between genomic markers found in linkage disequilibrium (LD) by using weights for each marker. The weighted effects over all genetic markers are summed into a global statistic with one degree of freedom. The resulting *z*-scores can be used to rank pathways and also to calculate an approximate *p*-value (after adjusting for multiple testing). An alternative method provided for calculating statistical significance using permutation tests is described below in the FDR section. This algorithm can be used to explore the statistical likelihood that a given pathway might be associated with the disease being investigated.

3.4.3. Data mining

Data mining classifiers (e.g., tree-based, Random Forests, logistic, SVM) can be used to explore the association between pathways and disease. The "percent correct" classification of cases and controls estimated with the genotypes at the pathway SNPs can be used as a statistic for measuring the association between pathways and disease.

We have incorporated the data mining algorithms from the well known Weka data mining program [23,29]. The classifiers are run by default with a 10-fold cross validation. The users, however, can modify any of the default parameters used when running the data mining algorithms. As with most statistical and data mining tools, users would need to be well trained in machine learning/data mining techniques to properly tune the parameters of the data mining algorithms. More advanced users can use any of the Weka functionality that can be invoked from the command line, such as clustering, feature selection or "meta-learning" algorithms [29].

The statistical significance of the pathway association scores can be assessed by FDR, as described below.

3.4.4. Statistical significance using permutation-based FDR

The statistical significance of association scores between individual SNPs or pathways with disease can be estimated using permutation-based FDR. The algorithm can use any statistics such as scores generated by the chi square test, *U*-statistics, or the percent-correct classification scores generated by the data mining classifiers. The aim of the FDR

Table 1
Possible outcomes from *m* hypothesis tests

	Accept	Reject	Total
Null hypotheses true	U	V[a]	m_0
Alternative hypotheses true	Q[b]	S	m_1
Total	W	R[c]	m

[a] V, number of type I errors (false positives).
[b] Q, number of type II errors (false negatives).
[c] R, total number of null hypotheses rejected.

procedure [30] is to control at a desired level α (e.g., 0.05) the proportion of type I errors (false positives) among all significant results: $FDR = E(V/R)$ using the notation from Table 1. When multiple genes or pathways may be involved in disease etiology, this procedure may yield higher statistical power compared to family wise error rate (FWER) procedures, such as Bonferroni, which control at a desired level α the probability of having at least one type I error: $FWER = P(V \geqslant 1)$. For FDR, using a statistic T_i for each pathway, one chooses a cutoff point C and selects as significant all pathways satisfying $T_i > C$. C is chosen in order to control FDR at the desired level α.

The details of the FDR algorithm implementation in Pathway/SNP are provided in [26]. Briefly, the algorithm performs a large number (e.g., 1000) of random permutations of the patient labels and for each pathway the "random" scores ($-\log_{10}$ from the *p*-values derived using *U*-statistics, or the fraction of correct classifications using the data mining algorithms) are recorded. The original (using the unrandomized data) pathway scores are used as cutoffs and the fraction of random scores that are above each cutoff is calculated. This fraction is the FDR, or the so called *q*-value [31]. Pathways with a low FDR, e.g., below a threshold of 0.05, are considered significant.

4. Sample results illustrating the user of Pathway/SNP

4.1. A hypothetical use case scenario

For illustration purposes, we start by describing a hypothetical use case scenario for the use of Pathway/SNP. A user, e.g., a genetic epidemiologist, intends to perform association analysis with a disease using 1000 controls and 1000 cases, genotyped with the Affymetrix 500k microarray. The data set has approximately 10^9 data points (500,000 genotypes $\times$ 2000 individuals).

The first analysis she performs is single SNP association with disease, aimed at identifying genomic loci that are significantly associated with the disease. No significant *p*-values are identified after adjusting for multiple testing.

She then performs a pathway based association with disease, aimed at identifying pathways that are significantly associated with the disease. She selects the J48 (Weka's implementation of C4.5) tree-based classification algorithm and chooses to find the statistical significance by performing 1000 permutations. The result displays a list of over 500 pathways, sorted by the statistical significance of the disease association scores. The top two pathways are statistically significant. One of the pathways is biologically relevant and there is literature pointing to the possible involvement of the pathway in the disease process.

Focusing on that pathway, the user then performs a single SNP disease association test only for the SNPs in that pathway to identify the SNPs and the genes that contribute to the high pathway association score.

4.2. Using Pathway/SNP to analyze the AMD data set

We used Pathway/SNP to analyze a data set used in a previously published GWA study [4]. The original analysis had identified a mutation in CFH, a complement pathway gene, as strongly associated with AMD [4], using the single SNP association test. This association has since been confirmed in many other studies and is believed to have an AMD population attributable risk of about 60% [32]. The data set consisted of 50 controls and 96 AMD cases (50 dry AMD, 46 wet AMD) that were genotyped using the Affymetrix 100k GeneChip. The details of our subsequent analysis can be found in [26].

We first performed the single SNP association test to replicate the results previously published. As illustrated in Fig. 3, the top 2 scoring SNPs (out of more than 100,000) were indeed the ones situated in the intron of CFH and previously identified as strongly associated with AMD.

We then performed a pathway based association analysis. We used *U*-statistics with five kernels: dominant, recessive, linear, quadratic, allele match; and four data mining algorithms: J48 (C4.5), Random Forests, SVM, and Naïve Bayes. Since the AMD patients were divided into two categories—wet AMD and dry AMD—we performed the tests by grouping the patients in various categories: control vs. all cases, control vs. wet AMD, control vs. dry AMD, dry AMD vs. wet AMD. We ran the computationally intensive data mining algorithms in a cluster environment. Fig. 4 illustrates a sample result from one of the pathway association tests, with pathways ranked according to the statistical scores. While the results were relatively different for the various *U*-statistics kernels and data mining algorithms, several pathways had significant statistical scores in multiple tests. Some of these pathways were relevant to AMD: complement pathway [33–35], mitochondrial fatty acid beta oxidation [36,37], calcium regulation and signaling [38,39].

We performed a more in depth analysis for the association between the complement pathway SNPs and AMD. We stratified the individuals based on the CFH SNP genotype and identified a striking pattern based on two complement pathway genes, C7 and MBL2, that can potentially further explain the difference between progressing to the less severe form of the disease, dry AMD, or to the more severe one, wet AMD [26].

Pathway/SNP

Trend Test for individual SNPs using Queries to select snps and patients.

116,204 items found, displaying 1 to 100.
[First/Prev] **1**, 2, 3, 4, 5, 6, 7, 8 [Next/Last]

SNP ID	Trend Test Control Vs Both AMD Types	Trend Test Control Vs Dry AMD	Trend Test Control Vs Wet AMD	Trend Test Dry AMD Vs Wet AMD	DBSNP Id	Chromosome	Position
4393	6.508319969816166	3.6324656196799965	6.386819412995011	1.0446804481590568	rs380390	1	193432708
8125	6.060368863001022	2.7184004978830387	6.488719412529549	1.3587788928342799	rs1329428	1	193434467
55642	4.755436442967263	3.476826255609935	3.0366154173145627	0.33691471387116567	rs10272438	7	33012074
59464	4.415528904406501	3.2276138692907628	3.174458259021607	0.14737166918131378	rs10254116	7	33010729
39635	4.340802070609787	2.477173464371491	4.299477571342766	0.3856443510257649	rs931798	5	155743401
40328	4.153331516532388	3.2506451656034137	2.3892345462844253	0.4426212491309568	rs4920799	5	84642284
50881	4.10653276791458	2.855507120953599	3.766463813357639	0.2315780017588782	rs8180608	6	89064414
43365	4.043286223829399	1.940906344664406	4.8844319278738	0.9266644997600094	rs970476	5	155734658
7012	4.041593917384488	3.184128012233245	2.0786036999258894	0.700399201893015	rs10495199	1	219215699
81381	4.011040457902311	3.873309213077327	3.1391651385958133	0.22642357826436763	rs7104698	11	36830141
106930	4.004323762886125	3.305746546591698	3.1233739164028336	0.00438104675841492	rs10513889	18	52162874
107190	4.004323762886125	3.305746546591698	3.1233739164028336	0.00438104675841492	rs727454	18	52196117
45455	3.9154234864022004	2.614028053758052	2.5241835282223373	0.08269233169321172	rs4292478	5	84634073
16179	3.913126768654347	2.1500591104515037	3.523382763473101	0.4087149738906286	rs1233255	2	190531598

Fig. 3. Trend test example with highest scoring SNPs in an AMD genome-wide association. The SNPs are sorted based on the second column, $-\log_{10}$ (*p*-value) for the association test between controls and both case types (dry AMD and wet AMD).

Pathway/SNP

Pathway Based Association Analysis using U-Statistics

554 items found, displaying 1 to 100.
[First/Prev] **1**, 2, 3, 4, 5, 6 [Next/Last]

Pathway	Score
Lectin Induced Complement Pathway(BC)|hsalectinPathway|BIOCARTA	4.40062194
Valine leucine and isoleucine degradation - Homo sapiens (human)|path hsa00280|KEGG	4.305863617
Alternative Complement Pathway(BC)|hsaalternativePathway|BIOCARTA	4.124592052
WNT Signaling Pathway(BC)|hsawntPathway|BIOCARTA	3.79219417
Map Kinase Inactivation of SMRT Corepressor(BC)|hsaegfr smrtePathway|BIOCARTA	3.76933069
BCR Signaling Pathway(BC)|hsabcrPathway|BIOCARTA	3.72587612
MAPKinase Signaling Pathway(BC)|hsamapkPathway|BIOCARTA	3.682392108
Nitric Oxide Signaling Pathway(BC)|hsanos1Pathway|BIOCARTA	3.493648962
IL-7 Signal Transduction(BC)|hsail7Pathway|BIOCARTA	3.376932349
CARM1 and Regulation of the Estrogen Receptor(BC)|hsacarm-erPathway|BIOCARTA	3.310405375
Benzoate degradation via hydroxylation - Homo sapiens (human)|path hsa00362|KEGG	3.165484649
Cyclin E Destruction Pathway(BC)|hsafbw7Pathway|BIOCARTA	3.145337028
E2F1 Destruction Pathway(BC)|hsaskp2e2fPathway|BIOCARTA	3.145337028
Regulation of p27 Phosphorylation during Cell Cycle Progression(BC)|hsap27Pathway|BIOCARTA	3.145337028
Multi-step Regulation of Transcription by Pitx2(BC)|hsapitx2Pathway|BIOCARTA	3.098806779

Fig. 4. Pathway-based association analysis example output. The pathways are sorted based on *z*-scores obtained from *U*-statistics with the linear dosage kernel.

5. Current status and future directions

We have used Pathway/SNP to perform pathway-based GWA for several diseases. Loading data and updating the indexes into the database is a complex process and some of the data loading processes are still performed from the command line. We are working on creating a GUI-based data loading mechanism. Since some of the data mining algorithms can be very computationally intensive, we are investigating the deployment of Pathway/SNP in a high performance computing (HPC) cluster environment. (We have used Pathway/SNP in a HPC environment from the command line.)

We are also interested in further developing several analysis areas of Pathway/SNP, by including:

- environmental factors, such as smoking status, into the analysis,
- linkage regions from previous studies, as previously suggested [9],
- algorithms, such as Pareto ranking [40], that can rank pathways by their performance across different models,

- additional knowledge about the pathways in tool—such as recessive/dominant effect of mutations and chemical reaction rates,
- functional relevance of SNPs (e.g., non-synonymous coding SNPs).

The code will be made available at http://www.dinuinformatics.info.

6. Lessons learned

6.1. The potential need for high performance computation to support a tool like Pathway/SNP

As described previously, one might attempt to take a completely brute force approach to analyzing genome-wide SNP, which would be computationally intractable (as discussed in Appendix A). Indeed, the various statistical and data mining approaches described previously in this paper attempt in different ways to avoid such brute force methods. At the same time, there is clear value to being able to take advantage of HPC. It is useful to look at the different capabilities within Pathway/SNP with regard to the potential need for high performance computation (e.g., the use of parallel clusters of workstations to solve compute-intensive problems).

Table 2 indicates the approximate amount of compute time required by various Pathway/SNP computations for the AMD dataset described above. (These figures were determined using a single desktop PC with a 1.8 GHz processor.) While some of the algorithms complete in a few minutes, others can be very time consuming, e.g., random forest with a large number of random trees. When analyzing the statistical significance of the scores through permutation-based FDR, the execution times must be multiplied by the number of permutations. For example, while pathway ranking using U-statistics completes in 1 min, calculating the FDR with 1000 permutations will actually complete in about 1000 min, or about 16.5 h.

As a result, to make a system like Pathway/SNP as robustly functional as possible, it will be extremely useful if it could be linked to a HPC facility that could be utilized for the compute-intensive analyses, as outlined above.

Table 2
Approximate execution times for different statistical and data mining algorithms when applied to the AMD data set (146 individuals genotyped at 116,204 SNPs; 561 total pathways)

Algorithm	Execution time (min)
Single SNP genome-wide association	5
Pathway ranking using U-statistics	1
Pathway ranking using data mining	
J48 (C4.5)	8
Naïve Bayes	2
SVM	8
Random forests 200 random trees	400

6.2. The need for permutation testing to evaluate the results of the analysis

Another issue that became clear in the implementation of Pathway/SNP is that while it is important to allow a variety of different algorithms to be used in analyzing the data, it is equally important to allow permutation testing using those algorithms to test the statistical significance of the results obtained. Indeed, there are different algorithms that can be used for the permutation testing. Pathway/SNP currently uses the FDR algorithm, but others are possible.

An interesting feature of the permutation testing is that it involves running the analysis algorithm many times, each time using a different set of randomly permuted data based on the original data set. This shuffling of the original data is commonly performed 1000 or more times. As a result, from the standpoint of HPC, even algorithms that make only relatively modest computational demands for the initial analysis may require HPC to do the required subsequent permutation testing. Clearly, when using analysis algorithms that are already quite computationally intensive, it will often be infeasible to do permutation testing without HPC.

Pathway/SNP computes permutation-based FDR by randomly shuffling the labels of the individuals and recalculating the association statistic using the randomized data. A potentially interesting alternative approach, which we did not perform, would be to randomly select sets of SNPs throughout the genome and calculate a baseline distribution of scores using these randomly selected SNP sets. The first approach (shuffle individual labels) measures how well a single collection of SNPs, selected in a scientifically well founded way—namely, membership in the same pathway—can be associated with the disease. The second approach would assess whether arbitrary sets of SNPs could result in a strong association with the disease.

One possible complication related to the latter approach of picking random sets of SNPs is that, especially in the case of high density SNP arrays, the correlation structure between SNPs might be lost. For example, when one selects SNPs that are in the vicinity of pathway genes, some of the SNPs are close genomic neighbors and are likely to be in linkage disequilibrium and have correlated genotypes. This correlation between neighboring SNPs would be lost when selecting random sets of SNPs, randomly dispersed throughout the genome. To avoid this problem when applying this approach of selecting random sets of SNPs, one would need, for example, to use a gene-based approach with similar gene set sizes.

6.3. Dealing with different versions of the biological data and knowledge

Another important issue that arises when contemplating using a program like Pathway/SNP on an ongoing basis is that the various sources of biological data evolve over time.

A well known example of this phenomenon is that the "official" sequence of the human genome has been updated several times over the past several years. Pathway data is undergoing the same evolution, as new pathways are added and as existing pathways are refined. Indeed, when developing Pathway/SNP, one apparent anomaly in the analysis arose because several genes that previously had been in a pathway had been removed in a later version of the pathway database. It will therefore be important (1) to continually update the various versions of the different sources of biological data and knowledge that are used, (2) to record which versions were used for each analysis performed, and (3) to retain previous versions so that previous calculations can be checked and verified.

6.4. Why different analysis algorithms might work better with different data sets and different diseases

It is well established in the data mining community that different data mining algorithms are frequently found to work better for different data sets in the same basic domain. This occurs because different data sets can exhibit different basic structures that are best uncovered by different algorithms. For example, in one data set only one attribute might be sufficient to understand the structure of the data. In another data set, several attributes might contribute independently to produce an observed effect. In another data set, a linear combination (or a more complex mathematical relationship) involving a subset of several attributes might be contributing.

Looking at biological processes and disease, one can hypothesize a number of potential underlying reasons for such variability in optimal analysis. Some genes act in a recessive fashion, some in a dominant fashion, and others in a mixed mode. Sometimes multiple genes must operate together (in an "epistatic" fashion) to produce dysfunction. In addition, the exact structure of a pathway (e.g., whether two genes are "in series" or "in parallel" in a pathway, or are more complexly interrelated) may affect the nature of the data and the corresponding success of different analytic approaches. Indeed, one interesting research issue for the future will be to explore how best to utilize our knowledge of pathway structure to assist in the more accurate statistical and data mining analysis of genome-wide SNP association data.

6.5. The complexity of the "Clinical Phenotype"

An additional issue in building a system like Pathway/SNP concerns the complexity of the clinical information included about the patients in a disease association study. The initial studies that are currently being performed frequently focus on quite extreme manifestations of a disease in the hope that this may help in producing a strong statistical result. As a result, a study might compare patients with extreme or malignant hypertension or patients who had myocardial infarctions before age 40 compared to normal individuals. Indeed, our AMD dataset focused on patients with quite large amounts of drusen buildup in the retina combined with evidence of sight-threatening dry or wet AMD, which represents a quite extreme form of the disease. In these cases, the relevant clinical information about the patient is by design very limited.

As researchers become more familiar with GWA analysis and start applying the technique to much larger sample sizes and to common diseases presumed to be multigenic in etiology, the "clinical phenotype" data relevant to the analysis may become much more complex. For example, the clinical phenotype might involve (1) a fairly detailed history of the progression of a disease with clinical findings as various points along the way, and/or (2) a spectrum of presentations of signs and symptoms involving several organ systems each with multiple degrees of severity. For example, the severity of AMD can be classified on a scale with five stages, from 0 (no AMD) to 4 (most advanced form) [41]. To complicate matters, there are additional AMD grading scales [42,43], which can make comparing study results particularly challenging.

For such studies with complex phenotypes, the amount and complexity of the clinical data that will need to be include in a system like Pathway/SNP will be much greater than with current data sets. In addition, the statistical and data mining analyses of the data will need to be structured to take advantage of a much more complex clinical phenotype. This challenge will raise a host of additional research issues for the future.

7. Summary

In this paper we introduced Pathway/SNP, a software program that can perform pathway-based association analysis between genome-wide high density SNPs and disease. The software tool integrates domain knowledge—pathway information, gene and SNP annotation—with statistical and data mining algorithms. Pathway/SNP can be used to explore the etiology of complex diseases in a flexible, interactive, incremental fashion.

Acknowledgments

Appendix A

The computational complexity of performing a brute-force "full-scan" interaction analysis between all possible combinations of n genomic markers and a disease (or trait) is exponential in n. Using a set of about $m = 100,000$ genomic markers, such as the Affymetrix 100k SNP GeneChip, a full-scan for $n = 2$ marker interaction would require performing $C(m, n) = 5 \times 10^9$ tests; 1.66×10^{14} tests for three markers;

4.16×10^{18} tests for four markers; and 8.33×10^{22} tests for five markers (number of subsets of size n in a set of size m). For comparison, a modern PC (with a clock frequency of a few GHz) can perform a few 10^9 flops/s (flops = floating point operations), while the fastest supercomputer can perform about 3.67×10^{14} flops/s (and doubling about each year) [44]. A simple arithmetic computation shows that performing a comprehensive scan for association between disease and four or more interacting markers is virtually impossible with the current available technology. For example, we found that, using a 1.8 GHz PC, it takes 5 min to run a GWA analysis for single SNP disease association for 116,204 SNPs and 146 individuals. Using the same data set and the fastest current supercomputer, we estimate that a test for the interaction of two markers would take about 1 min; 1 month for three markers; 2000 years for four markers; and 40,000 millennia for five markers.

References

[1] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.

[2] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95:14863–8.

[3] Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 2002;99:6567–72.

[4] Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science 2005;308:385–9.

[5] Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, Smink LJ, et al. A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nat Genet 2006;38:617–9.

[6] The Wellcome Trust Case Control Consortium website, web site: http://www.wtccc.org.uk/, date accessed: August 16, 2006.

[7] Storey JD, Akey JM, Kruglyak L. Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biol 2005;3:e267.

[8] Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 2005;37:413–7.

[9] Roeder K, Bacanu SA, Wasserman L, Devlin B. Using linkage genome scans to improve power of association in genome scans. Am J Hum Genet 2006;78:243–52.

[10] Conti DV, Cortessis V, Molitor J, Thomas DC. Bayesian modeling of complex metabolic pathways. Hum Hered 2003;56:83–93.

[11] Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, Thibodeau SN. Nonparametric tests of association of multiple genes with human disease. Am J Hum Genet 2005;76:780–93.

[12] Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 2003;34:267–73.

[13] Pang H, Lin A, Holford M, Enerson BE, Lu B, Lawton MP, et al. Pathway analysis using random forests classification and regression. Bioinformatics 2006;22:2028–36.

[14] Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. Bioinformatics 2003;19:376–82.

[15] Ritchie MD, Motsinger AA. Multifactor dimensionality reduction for detecting gene–gene and gene–environment interactions in pharmacogenomics studies. Pharmacogenomics 2005;6:823–34.

[16] Lucek PR, Ott J. Neural network analysis of complex traits. Genet Epidemiol 1997;14:1101–6.

[17] Motsinger AA, Lee SL, Mellick G, Ritchie MD. GPNN: power studies and applications of a neural network method for detecting gene–gene interactions in studies of human disease. BMC Bioinform 2006;7:39.

[18] Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Van Eerdewegh P. Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol 2005;28:171–82.

[19] Yoon Y, Song J, Hong SH, Kim JQ. Analysis of multiple single nucleotide polymorphisms of candidate genes related to coronary heart disease susceptibility by using support vector machines. Clin Chem Lab Med 2003;41:529–34.

[20] Yu R, Shete S. Analysis of alcoholism data using support vector machines. BMC Genet 2005;6(Suppl. 1):S136.

[21] Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P. A haplotype map of the human genome. Nature 2005;437:1299–320.

[22] BioCarta website, web site: http://www.biocarta.com/, date accessed: August 16, 2006.

[23] GenMAPP website, web site: http://www.genmapp.org/, date accessed: August 16, 2006.

[24] SOAP/WSDL interface for the KEGG system, web site: http://www.genome.jp/kegg/soap/, date accessed: August 20, 2006.

[25] Entrez Gene website, web site: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/, date accessed: August 20, 2006.

[26] Dinu V, Miller PL, Zhao H. Evidence for association between multiple complement pathway genes and AMD. Genetic Epidemiology 2007;31:224–37.

[27] Dinu V, Nadkarni P. Guidelines for the effective use of entity-attribute-value modeling for biomedical databases. Int J Med Inform 2006.

[28] Armitage P. Tests for linear trends in proportions and frequencies. Biometrics 1955;11:375–86.

[29] Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics 2004;20:2479–81.

[30] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc. Series B (Methodol) 1995;57:289–300.

[31] Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA 2003;100:9440–5.

[32] Thakkinstian A, Han P, McEvoy M, Smith W, Hoh J, Magnusson K, et al. Systematic review and meta-analysis of the association between complementary factor H Y402H polymorphisms and age-related macular degeneration. Hum Mol Genet 2006.

[33] Gold B, Merriam JE, Zernant J, Hancox LS, Taiber AJ, Gehrs K, et al. Variation in factor B (BF) and complement component 2 (C2) genes is associated with age-related macular degeneration. Nat Genet 2006;38:458–62.

[34] Donoso LA, Kim D, Frost A, Callahan A, Hageman G. The role of inflammation in the pathogenesis of age-related macular degeneration. Surv Ophthalmol 2006;51:137–52.

[35] Bora PS, Sohn JH, Cruz JM, Jha P, Nishihori H, Wang Y, et al. Role of complement and complement membrane attack complex in laser-induced choroidal neovascularization. J Immunol 2005;174:491–7.

[36] Liang FQ, Godley BF. Oxidative stress-induced mitochondrial DNA damage in human retinal pigment epithelial cells: a possible mechanism for RPE aging and age-related macular degeneration. Exp Eye Res 2003;76:397–403.

[37] Beatty S, Koh H, Phil M, Henson D, Boulton M. The role of oxidative stress in the pathogenesis of age-related macular degeneration. Surv Ophthalmol 2000;45:115–34.

[38] Hoffmann S, He S, Jin ML, Masiero L, Wiedemann P, Ryan SJ, et al. Carboxyamido-triazole modulates retinal pigment epithelial and choroidal endothelial cell attachment, migration, proliferation, and MMP-2 secretion of choroidal endothelial cells. Curr Eye Res 2005;30:103–13.

[39] Kindzelskii AL, Elner VM, Elner SG, Yang D, Hughes BA, Petty HR. Toll-like receptor 4 (TLR4) of retinal pigment epithelial cells participates in transmembrane signaling in response to photoreceptor outer segments. J Gen Physiol 2004;124:139–49.

[40] Komuro R, Ford ED, Reynolds JH. The use of multi-criteria assessment in developing a process model. Ecol Model 2006;197:320.

[41] Ferris FL, Davis MD, Clemons TE, Lee LY, Chew EY, Lindblad AS, et al. A simplified severity scale for age-related macular degeneration: AREDS Report No. 18. Arch Ophthalmol 2005;123:1570–4.

[42] Bird AC, Bressler NM, Bressler SB, Chisholm IH, Coscas G, Davis MD, et al. An international classification and grading system for age-related maculopathy and age-related macular degeneration. The International ARM Epidemiological Study Group. Surv Ophthalmol 1995;39:367–74.

[43] Klein R, Davis MD, Magli YL, Segal P, Klein BE, Hubbard L. The Wisconsin age-related maculopathy grading system. Ophthalmology 1991;98:1128–34.

[44] top500, web site: http://www.top500.org, date accessed: August 16, 2006.