BIOMARKER INDENTIFICATION USING BAYESIAN VARIABLE SELECTION BASED ON MARKER-EXPRESSION-PROTEOMICS DATA

Madhuchhanda Bhattacharjee
School of Mathematics and Statistics,
University of St Andrews, St Andrews, Scotland
+44 1334 46 1645
mb104st-andrews.ac.uk

Catherine H. Botting
School of Chemistry,
University of St Andrews, St Andrews, Scotland
+44 1334 46 7195
cb2@st-andrews.ac.uk

Mikko J. Sillanpää
Department of Mathematics and Statistics
University of Helsinki
Helsinki, Finland
+358 9 191 51512
mjs@rolf.helsinki.fi

ABSTRACT
Finding genetic biomarkers and a search of genetic-epidemiological factors, can be formulated as a statistical problem of variable selection, where from a large set of candidates a small number of trait-associated predictors are identified. We illustrate this by analyzing the data available for Chronic Fatigue Syndrome (CFS). CFS is a complex disease from several aspects, e.g. difficult to diagnose and difficult to quantify. From the clinical information subjects were classified in No-Fatigue (NF), Insufficient fatigue severity (IFS), Chronic Fatigue (CFS) and others. The additional clinical variables were used as stratifying factors to homogenize the study population. For identification of biomarkers microarray data and SELDI-TOF-based proteomics data were used. Genetic marker information for a large number of SNPs was also analyzed for an overlapping set of individuals. The objectives of the analyses were to identify markers specific to Fatigue which are also possibly exclusive to CFS. The WinBUGS software was used in implementation and parameter estimation of the proposed Bayesian models.

Keywords
Variable selection, Bayesian-hierarchical models, CFS, association analysis, high-throughput analysis.

1. INTRODUCTION
Finding genetic biomarkers and a whole-genome search of genetic-epidemiological predisposing factors, can both be formulated as a statistical problem of variable selection, where tiny number of trait-associated predictors (measurements from the genome) is selected out from the huge sets of candidates (Sillanpää and Bhattacharjee 2005, 2006; Hoti and Sillanpää 2006). A statistical variable selection plays an important role in personalized medicine, in the development of the models to predict disease state or drug-response in humans, and it is the first step in marker-assisted selection programs in plant and animal breeding. Moreover, statistical variable selection methods are also essential in different studies of cancer biology, immunogenetics, neurogenetics, behavior genetics, toxicology, and gastroenterology, which may like to use different crossing designs of mouse or rat data (animal models of human disease). In any case, the search of for new biomarkers for cancer diagnosis, prognosis and measures of response to therapy are already providing sufficient motivation for the modern statistical and computational work.

In CAMDA06, we presented a Bayesian joint disease-marker-expression analysis using data on CFS study (Bhattacharjee and Sillanpää 2007). However, full data was not utilized in that study. Here we consider the newly available SNP data, which is several times larger than the previous one. We also incorporate microarray data available from all individuals, unlike the previous study where only individuals with both SNP and microarray data were studied. Additionally we extend our previous work to simultaneously include also intensity at selected points, i.e. the mass-to-charge (m/z) values, of the proteomic profiles. Unlike SNP or microarray data, it is unclear what these peaks are in biological sense, these points could correspond to specific peptides or proteins or even mixtures of proteins.

In other words, we are considering multiple regression models where molecular markers and/or gene-expression measurements as well as intensity measurements from protein spectra serve as predictors for the outcome variable (trait or disease state) which is CFS, ISF, etc. Use of such models can be motivated, for example, by the search for new biomarkers for cancer diagnosis, prognosis and measures of response to therapy. Generally, for this we use Bayesian hierarchical modeling and Markov Chain Monte Carlo computation.

2. DATA CONSIDERED
In our previous analyses (Bhattacharjee & Sillanpää 2006) the objective was to form an all encompassing integrated model where clinical, SNP and expression data will be used. Thus it restricted us to utilize data on 164 individuals on whom all such data were available. However we also learned from that analysis that only expression showed some effect in joint analysis. Generally, it is likely that continuous expression measurements contain more information than discrete SNP markers. This property may be more noticeable when individual effect sizes are very small. Thus it may look like expressions "override" markers, even if in reality marker effects are too small to be detectable given the size of the current data sets.

Thus in this analysis we will not attempt to form such an overall integrated model for the data, however we intend to provide

integrated prediction of relevant (bio and genetic) markers for the disease based on customized models for individual data types. A summary of data used is presented in Table-1 below.

Table 1. Distribution of subjects according to disease category and data availability

Empiric variable	SNP	Microarray	Proteomics
NF	58	40	40
ISF	59	39	42
CFS	43	39	36
Others	62	46	46
Total	222	164	164

2.1 Phenotype data

The variable "Empiric" was continued to be used as a comprehensive summary of the disease phenotype. Based on Reeves et al. (2005), variable empiric spans the space very similar to the first few principal components extracted from the original clinical variables. Therefore this phenotype can be seen as a linear combination of clinical variables. The other possible alternate was the "Cluster" variable which when compared shows a clear relationship with the Empiric-variable.

In our previous analyses of the CFS data we had utilized approximately two dozen clinical variables. These were used, mostly for the purpose of homogeneous stratification of the data. However other than a few like onset & gender the rest were not easily interpretable directly in the context of the disease, hence were discontinued for the present analysis.

2.2 Marker data:

The complete SNP data was utilized which is altogether available on 222 individuals. Altogether there were 168 SNPs on 39 genes, which included 123 SNPs pertaining to 29 new genes, 8 new SNPS of genes analysed previously and 37 SNPs of original 10 genes. We noted that 5 SNPs from the old data were discontinued in the currently available data set, which unfortunately contains one SNP previously found to be significant.

We obtained location information for both gene-regions (see Table-2) and the SNPs within those. This potentially can increase inferential powers by using models proposed by Sillanpää and Bhattacharjee 2005. Such model accounts for possible dependence in behavior between two closely placed SNPs and identify dependence structure.

Table 2. Location information of Gene-regions with SNP data

Gene	Location	Gene	Location
HTR6	1p36-p35	DBH	9q34
IL10	1q31-q32	HTR7	10q21-q24
HSD11B1	1q32-q41	SLC18A2	10q25
POMC	2p23.3	BDNF	11p13
IL1A	2q14	TH	11p15.5
IL1B	2q14	DRD2	11q23
HTR2B	2q36.3-q37.1	HTR3A	11q23.1
DRD3	3q13.3	HTR3B	11q23.1
SPP1	4q21-q25	TNFRSF1A	12p13.2
SLC6A3	5p15.3	IFNG	12q14
HTR1A	5q11.2-q13	TPH2	12q21.1
IL12B	5q31.1-q33.1	HTR2A	13q14-q21
NR3C1	5q31.3	SLC6A4	17q11.1-q12
HTR4	5q31-q33	CRHR1	17q12-q22
HTR1E	6q14-q15	ACE	17q23.3
CRHR2	7p15.1	COMT	22q11.21-q11.23
IL6	7p21	MAOB	Xp11.23
NOS3	7q36	MAOA	Xp11.3
HTR5A	7q36.1	HTR2C	Xq24
INDO	8p12-p11		

2.3 Expression data:

Of the 177 arrays five were excluded due to non-availability of clinical data on these subjects. The remaining 172 arrays included 8 replicate arrays on 8 subjects. Four such duplicate arrays were excluded after carrying out quality check between the two replicate arrays on an individual. For the remaining four individuals one array each were selected (the ones without "rep" in filenames) in order to maintain balance in information. It may however be mentioned that the models proposed for expression data analysis do not require design to be balanced.

The resulting 164 arrays were used for further analysis after carrying out quality check of the data contained. Of these data from 4 arrays were not satisfactory. These data were used carefully, for example, summary from the data were used as (hyper-)hyper-parameters in the model. While computing these summaries the data from the 4 unreliable arrays were not utilized. Cutoff intensity was set at 100. Spots were checked for missing data and only spots with at least 20 arrays with data for each Empiric group (viz. NF, ISF, CFS and Other) were selected for analysis. Thus of 20160, only 9953 spots meet this criteria and were used. As was done for the SNP data, for microarray data also we have gathered additional information in the locations of the genes and selected functionalities.

2.4 Proteomics data:

We used the pre-processed proteomics data made available on 206 subjects of which because of high number of missing data we had to exclude several and data from 164 subjects were used for analysis. This data contains three measurements using three ProteinChip Array chemistries: Reversed Phase (H50), Metal Affinity Capture (IMAC30) and Weak Cation Exchange (CM10). The additional high stringency wash condition was used for the CM10, however this method resulted in high degree missingness data and hence was excluded from the analysis. Therefore for all three array chemistry, washed at lower stringency and two laser intensities were used for this analysis. This resulted in 895 m/z values (at different fractions) although all of which were not distinct. In several cases the same m/z values were observed at distinct different fraction, which would mean the pI of these proteins is very different although mass could be similar, which makes them different proteins. Hence such m/z was treated as distinct. The other m/z values that appeared in different array types at same fraction were treated as different since we are still uncertain about their pI, since from the fractionation policy we could only roughly estimate their pI.

Whilst SELDI data is ideal for profiling protein expression levels to determine patterns of expression that are associated with particular disease states (see Laronga et al 2003) it can not be used, on its own to identify biomarkers for CFS. SELDI reports only the peptide/protein molecular weights present in the (serum) sample. This approach is flawed on two counts. 1) SELDI molecular weights measurements (essentially MALDI-ToF mass measurements) are not accurate enough to uniquely identify a protein. 2) At present there is not a complete enough knowledge of the post-translational modifications that occur e.g. in the human body to produce a database of all the masses one would expect to find, for example, in serum. Swiss-Prot goes some way towards this by documenting the information known about signal sequences and propeptides removed post-translationally. This information is used by our programme of choice for searching for proteins of a given mass, ExPasy's TagIdent, before computing pI and Mw for each of the resulting chains. However, this does not address modifications which add to the mass of the protein e.g. phosphorylation, glycosylation etc.

However in absence of any other means to combine this data to the rest of the data we used predicted genes corresponding to the masses and pI information. These were then used to obtain similar annotations as SNP and microarray data. Objective would be to assess chromosome region enrichment (genomic overlap) between the suggested locations from different analyses as well as enrichment for a selected set of functionalities.

3. STATISTICAL MODELS AND ESTIMATION

3.1 Handling of missing values

In the association analyses models, we used missing data model 2 of Sillanpää and Bhattacharjee (2005) to handle missing values in the genotype data.

In case there were values missing in the stratifying variables the augmentation was carried out using posterior frequency distribution resulting from Uniform-Bernoulli prior assumption on the respective distribution.

For expression analysis the missing values are augmented through the integrated model for normalization and differential analysis. The augmentation is thus based on information of the location of a gene on the array, information about expression behavior of other neighboring genes and expression pattern of the same gene on other arrays and also overall expression pattern of all individual in the relevant treatment group.

For proteomics data a similar model based data augmentation is carried out.

3.2 Association analysis

The Bayesian association mapping models utilizes the location information of the gene-regions and the SNPs within them. These are similar to the one used in Sillanpää and Bhattacharjee (2005), where variable selection in the model is based on indicator variables controlling inclusion / exclusion of the genetic effects from the model. These models were applied to identify CFS related SNPs and by analyzing ISF associated SNPs we could then additionally identify SNPs specific to CFS but not other fatigues as captured by the ISF individuals.

A Markov-dependence model, similar to Sillanpää and Bhattacharjee (2005), was used to describe the dependence between the SNPS according to their map distance, a smoothing parameter and a stringency parameter describing essentials of the model. The shrinkage parameter S can be interpreted as the prior probability of selecting a candidate variable (that is, the corresponding indicator is one) in the model.

Stratification of the data using the clinical variable like onset and gender were carried out. These were found to be informative in our previous analysis and straightforward extension of Sillanpää and Bhattacharjee (2005) provided the necessary model setup.

3.3 Expression analysis

The normalization was done using the block-level-piecewise-linear-regression normalization method of Bhattacharjee et al. (2004). Therefore for every array and every block parameters necessary for carrying out five-piece-connected Bayesian linear regression were utilized (assuming known knot-points). All the arrays were normalized against the observed average intensities over all arrays for each spot. This can also be thought as utilizing the data at a hyper-hyper-parameter level, where mean of each gene for each disease category (say, μ^1_{ki}) is drawn from a gene-specific parameter (say μ^o_i) which in turn is described by a Normal distribution with given mean as overall sample average. This provides some identifiability to the prior distributions without influencing the model much (since these are hyper-hyper parameters). A joint normalization and expression analysis was done, followed by a regularized two sample Bayesian T-test (see. Baldi & Long 2001, Lewin et al 2005) to identify relevant genes.

3.4 Proteomics analysis

The principal outcome required from this data is very similar to that of identifying differentially expressed gene. Hence a similar model as above (without the normalization factors) was used to analyse this data at a first level. Here also Bayesian t-test was used for preliminary identification of differential protein masses. Every m/z which is found significant is then feed-in into the second-level model (and smoothed if necessary with the mz distance). That way if any m/z is only mariginally important it will still contribute to the second-level model.

The peaks in the protein spectra may occur as a real observed intensity difference between two groups, disease (CFS) and healthy (NF) individuals. However, due to noise, some differential peaks may be spurious and occur as a consequence of some other factor than the real change in the disease state. Potential reasons for spurious peaks include measurement errors and inaccuracies in alignment.

Thus, to control spurious peaks (false signals), the smoothing of peaks with respect to neighboring locations has been proposed for proteomic profiling (see Du et al. 2006). We present here the new smoothing approach based on two-level hierarchical modeling. Use of hierarchical modeling to subset selection and for more close inspection (special treatment) of most promising candidates in genomewide association study context has recently been proposed by Chen and Witte (2007). The two-level model is considered also here so that the first-level feed in candidates for the second-level, where the smoothing of intensity peaks with respect to neighboring locations (at the same m/z region) is done according to mz distance between the positions (cf. Sillanpää and

Bhattacharjee 2005). For each m/z ratio, first-level model checks if corresponding intensity is marginally differentially expressed between CFS & NF (similarly for ISF & NF). In the second level model, a logistic regression is performed jointly with all the differential peaks using mz distance and smoothing. This way multiple signals from same m/z region due to measurement error, inaccuracy in alignment, etc will be adjusted and overall signals can be identified. Note that in addition of adjusting dependence between neighboring mz:s, the joint explanation of mz:s with respect to phenotype is also accounted for. These two parts are modeled simultaneously in the single hierarchical structure, so in different MCMC iteration we would be proposing different sets of candidates (m/z values) as differentially expressed. This way we can adjust for decision error due to missing values in variable/feature selection.

Although the two-level model will be implemented as one large hierarchical structure, actually we want to treat them in some extent as two separate model parts. Technically we use "cut" function in WinBUGS to stop feedback from second-level model to first-level model. The reason for this is that the phenotype data is used twice (in both levels of model) and intensity data partly twice, which will bias some credible intervals. One interpretation of the proposed model is, we are identifying differential m/z:s marginally (one at a time, ignoring others) and then in the second-level model we are adjusting for dependence amongst the closely situated mzs and consider their joint explanation with respect to phenotype.

4. RESULTS

4.1 SNP analyses

It was noted previously that stratification seemed to improve our ability to find associations. It is not uncommon that behind a complex disease genetic mechanism may vary across subsets of individuals. For example previously when the clinical variable "Gender" was used to stratify the data a noticeable association was observed for particular SNP with the disease status. Unfortunately this particular SNP was discontinued for some reason in the present data. However since the current data was much larger than the previous one we attempted similar analysis to see if there are any other explanation to this disease. Quite encouragingly there was much consistency noticed between our previous analyses results and those obtained based on the current data.

Since we are able to use distance to model dependence between closely situated SNPs in a region we can accurately identify the SNP showing most variation in the disease groups. The gene region NR3C1 was found to be significant and it continued to be so. The SNP found is a different one than the analysis without distance, however the magnitude of variation remained quite comparable, which is highly significant since now these SNPs are identified amongst a much larger group of SNPs selected through a harder competition. Similarly the gene region TH was found relevant, in fact the same SNP (hCV1843075) showed up in both analyses.

Note that although we are carrying out strata level modeling we actually are not losing power when it comes to identifying SNPs that could be significant irrespective of strata. In the above weighted genetic variances of all SNPs on chromosome 5 is plotted for each strata. Note that one particular SNP of NR3C1 (hCV11159943) highly relevant for both female and male strata while discriminating between CFS and NF individuals. As is expected we can identify SNPs that are specific to a strata, for example, HTR6 on chromosome 1 is highly significant in the male strata (see figure 1).

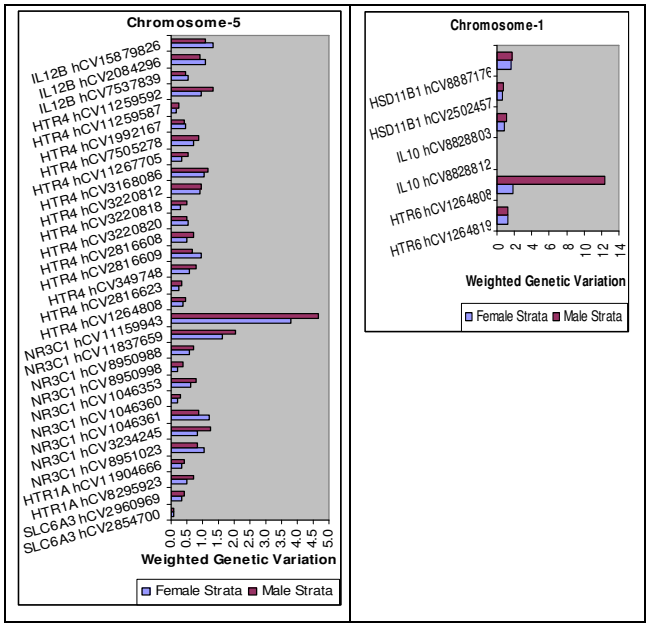

Figure 1. Gender-specific effects of SNPs on chromosomes-5 and 1 in disease-marker association analysis.

Similarly while comparing ISF-CFS (while also compring with NF individuals) we obtained significant SNP discriminating all (e.g. HTR2C-hCV339374). However, the mechanism is quite different between the two strata, with roles (or frequencies) of the alleles reversed in the two strata. For gene BDNF the SNP hCV12035465 has only "G" allele in the male strata for individuals with CFS, hence this is significant for NF-CFS styady in this strata, additionally the G allele proportion in male ISF is much lower than male NF population making it also relevant in ISF-NF comparison. Note again that behaviour of the SNP is opposite in the CFS & ISF cases. Thus, several SNPs were found to show the association signals in this analysis.

4.2 Expression analyses

Based on gene-expression data analysis as described before, several genes were identified as significant strongly. Some of these are presented in the following in Table 3.

Table 3. Genes showing high association with "Empiric" variable when analyzed using expression data

Rank	Gene	Location	Rank	Gene	Location
1	TACC2	10q26	11	DRG2	17p11.2
2	CDK7	5q12.1	12	C10orf48	10p12.1
3	PRO1853	2p22.2	13	SARDH	9q33-q34
4	PLAT	8p12	14	HSPA1L	6p21.3
5	EPHB2	1p36.1-p35	15	GNL1	6p21.3
6	C2orf32	2p14	16	APBA2	15q11-q12

7	SIRT5	6p23	17	CFLAR	2q33-q34
8	PURB	7p13	18	NEK6	9q33.3-q34.11
9	BTK	Xq21.33-q22	19	RUNX2	6p21
10	ZFY	Yp11.3	20	PRSS11	10q26.3

4.3 Proteomics analyses

The Bayesian t-test of the proteomics data while identifying CFS-specific m/z values produced some reasonably high but not many significant results. A less stringent test obviously picks up more bio-markers. In the following a few examples of estimated selection probabilities (under stringent tests) are presented for different pI values and all m/z values covered in the data (Fig-2).

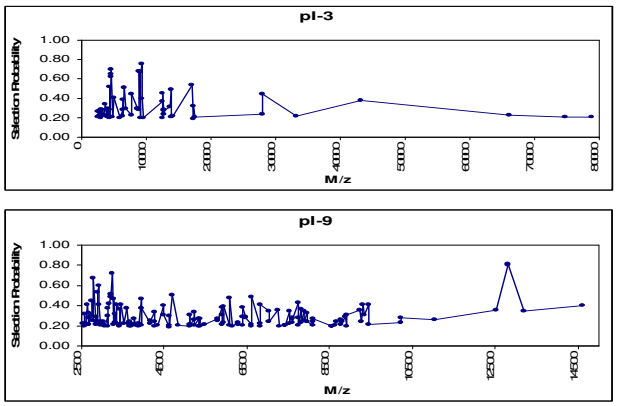

Figure 2. Estimated selection probabilities under stringent tests for m/z values, (for pI=3 in top panel and pI=9 in bottom panel).

4.4 Integrated analyses

From the proteomics data analysis, using TagIdent software we predicted genes for as many possible mz values at relevant pI levels as possible using a 10% tolerance level. For each gene we now utilize their selection probabilities from SNP, microarray and proteomics data analysis. These are then used on the genome level to identify possible genes region or functional enrichments.

5. DISCUSSION

We have formulated the problem of identifying genetic- and bio-markers in the framework of Bayesian variable selection. For each individual data type we have utilized advanced modeling techniques. These we have implemented on a much larger data than before. Wherever comparisons can be made with previous findings were done and results were consistent.

The methodological part of this manuscript contains some novel modeling, for example the two stage modeling of the proteomic data and also a very similar modeling for the integrated data. The integrated analysis thus can reflect uncertainty in analysis of individual data sets at the same time provide a comprehensive genome/functional level summarization of information in data.

We have to however admit, that implementing such model has not been easy. On standalone computers some of these models take 1 GB of RAM to run, some of these also take considerable time for the MCMC iterations. However we still were able to implement it on standard Bayesian software like WinBUGS, enabling us to implement numerous modeling options without having to create custom codes. This makes this method repeatable on demand as often as needed.

6. ACKNOWLEDGMENTS

Work of MS was supported by his research grant (202324) from the Academy of Finland.

7. REFERENCES

[1] Baldi P and Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inference of gene changes. *Bioinformatics 17: 509-519, 2001.*

[2] Bhattacharjee M and Sillanpää MJ. Bayesian joint disease-marker-expression analysis applied to clinical characteristics of chronic fatique syndrome. (*To appear in proceedings of CAMDA 2006*).

[3] Bhattacharjee M., Pritchard CC, Nelson PS, Arjas E. Bayesian integrated functional analysis of microarray data. *Bioinformatics 20: 2943-2953, 2004.*

[4] Chen GK and Witte JS. Enriching the analysis of genomewide association studies with hierarchical modeling. *Am. J. Hum. Genet. 81: 397-404, 2007.*

[5] Du, P, Kibbe, WA and Lin SM. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics 22: 2059-2065, 2006.*

[6] Hoti F, Sillanpää MJ. Bayesian mapping of genotype x expression interactions in quantitative and qualitative traits. *Heredity 97: 4-18, 2006.*

[7] Laronga C, Becker S, Watson P, Gregory B, Cazares L, Lynch H, Perry RR, Wright GL Jr, Drake RR, Semmes OJ, SELDI-TOF serum profiling for prognostic and diagnostic classification of breast cancers, *Disease Markers, 19(4-5), 229-38, 2003.*

[8] Lewin A, Richardson S, Marshall C, Glazier A, Aitman T. Bayesian modelling of differential gene expression. *Biometrics, 62:10-18, 2005.*

[9] Reeves WC, Wagner D, Nisenbaum R, Jones JF, Gurbaxani B, Solomon L, Papanicolaou DA, Unger ER, Vernon SD, Heim C J. Chronic Fatique Syndrome - A clinically empirical approach to its definition and study. In *BMC Medicine 3: 1-9,* 2005.

[10] Sillanpää MJ, Bhattacharjee M. Bayesian association-based fine mapping in small chromosomal segments. *Genetics 169: 427-439, 2005.*

[11] Sillanpää MJ and Bhattacharjee M. Bayesian association mapping of complex trait loci with context-dependent effects and unknown context variable. **Genetics 174: 1597-1611, 2006**.

[12] Spiegelhalter DJ, Thomas A, Best NG. WinBUGS Version 1.2 User Manual. *MRC Biostatistics Unit.*, 1999.