Understanding Chronic Fatigue Syndrome (CFS) from CAMDA Data: A Systems Biology Approach

Vasyl Pihur, Susmita Datta, Somnath Datta[1]

Department of Bioinformatics and Biostatistics, University of Louisville, KY 40292, USA

Abstract

We start by constructing gene-gene association networks based on about 300 genes whose expression values vary between the groups of CFS patients (plus control). Connected components (modules) from these networks are further inspected for their predictive ability for symptom severity and genotypes of two single nucleotide polymorphisms (SNP) known to be associated with symptom severity. We use two different network construction methods and choose the common genes identified in both for added validation. Our analysis identified eleven genes which may play important roles in certain aspects of CFS or related symptoms. In particular, the gene WASF3 (aka WAVE3) possibly regulates brain cytokines involved in the mechanism of fatigue through the p38 MAPK regulatory pathway.

1 Introduction and a summary

Chronic Fatigue Syndrome (CFS) is a relatively rare, poorly understood, complex disorder that is characterized by severe and chronic physical and mental fatigue not attributable to other causes (diseases) which is sometimes accompanied by other symptoms such as weak immune response, digestive problems and depression. A great deal of effort has been put forth in recent years in collecting clinical, gene expression, gynotypic and proteomic data by the Chronic Fatigue Syndrome Group at CDC in an attempt to find a genetic basis of CFS. Even though these data have been analyzed by numerous researchers (and research teams) in the last two years resulting in a special issue of the journal Pharmacogenomics (Vernon and Reeves , 2006) and were also as part the CAMDA conference in 2006, the type of success has been mixed and limited.

Our attempt in analyzing these data as part of this year's CAMDA competition takes a systems biology approach where we study groups of genes (called modules) obtained from gene-gene association networks since genes do not act alone, especially, for a complex disorder. Thus, our approach is similar to that by Presson *et al.* (2006), although our network construction methods and the statistical analyses are different from theirs. At the end, we identify eleven "interesting" genes which may play important roles in certain aspects of CFS or related symptoms. In particular, the gene WASF3 (aka WAVE3) possibly regulates brain cytokines involved in the mechanism of fatigue through the p38 MAPK regulatory pathway.

2 CAMDA datasets

The CDC Chronic Fatigue Syndrome Research Group provided challenge datasets consisting of clinical, microarray, proteomics, and SNP data that were used for both CAMDA 2006 and CAMDA 2007 competitions. 227 subjects filled self-administered questionnaires and had their blood drawn for lab analysis. For many of them, microarray (163) and proteomics (60) data were also collected for the purpose of discovering biological (genetic) basis of CFS. In this work, we integrate clinical, microarray, and SNP data for our analysis.

[1]to whom correspondence should be addressed (somnath.datta@louisville.edu)

2.1 Microarray Data

Camda 2006 microarray data consists of 177 arrays, 9 of which were repeated twice at different times during the study. We discarded these 9 microarrays for multiplicity reasons and additional 5 arrays were excluded from this analysis due to the absence of clinical information on the subjects. Thus, we started our analysis with 163 arrays. Subtracted ARM (Artifact-removed) density column which is already adjusted for the background density was log-transformed to stabilize the variance.

2.2 Clinical Data

Clinical data contains extensive information on 227 subjects and can be linked to microarray and SNP data via the ABTID subject ID. The two pieces of clinical data that we made use of were the *Intake Classific* variable classifies patients into 5 categories and the *Cluster* variable provides information on the severity of the symptoms ("Worst", "Middle", "Least") for some patients.

2.3 SNP Data

Fortytwo Single nucleotide polymorphisms (SNP's) for 10 different genes were genotyped. For the purposes of this analysis, we selected two SNP's, hCV245410 (on gene TPH2) and hCV7911132 (on gene SLC6A4), which were previously identified (Presson *et al.*, 2006) to be associated with CFS severity.

Table 1: Gene association modules discovered by the PLS based network inference method. For each such module (in rows) we are listing the number of genes, relative association strength, gene names, and p-values for the three log-linear models as discussed in the text. Genes in red have also been included into modules by the PC method as well.

# of Genes	Average Scores(%)	Gene Symbols	Severity p-value	hCV245410 p-value	hCV7911132 p-value
4	100	MTA1, SRP68, XM_049568, CLDN10	0.7588	0.3869	0.0328
9	88	CREB5, MED6, UPP1, NP_775847.1, ABCG8, RNF25, XM_089436, THSD1, NM_022084	0.1645	0.2970	0.1271
5	85	HOXA1, NAGA, GAK, CK021_HUMAN, CDH23	0.4978	0.2636	0.6640
6	81	IER2, TCIRG1, XM67745, XM_065828, XM70678, HDAC7A	0.5051	0.5825	0.1689
*5	79	WASF3, NUP98, PRUNE, NP_079431.1, KIRREL3	**0.0154**	**0.0163**	0.1665
6	78	ZFYVE19, AK024757, CNGB1, WARS2, SPIN2, XM_069044	0.0081	0.7775	0.1203
5	77	PCDH21, ASS, GTF2I, ARID4B, TNFSF13B	0.3015	0.4987	0.0112
9	77	AB082528, HNRPLL, HBLD2, ZNF165, MOG, SORL1, VAT1, EPC2, NP_787114.1	0.0063	0.8304	0.1047
4	77	MTMR8, NP_076414.2, MLL3, XM_087606	0.0032	0.5670	0.2732
4	76	ZFYVE9, RAD51C, XM_085181, ZBTB11	0.2778	0.4110	9e-04
4	75	TNK2, EIF3S8, PMS2L5, TCP11	0.1286	0.6770	0.1270
4	73	MAP3K2, ATF5, AF107495, GALK2	0.0436	0.2459	0.1904
15	72	CDC2L5, PLP2, NR1H2, PLAUR, SPATA11, NP_060367.1, KCNQ5, COL9A1, AF173157, XM_067644, MAB21L1, CNR2, NP_054868.2, RAB32, ADAM9	0.0145	0.0960	0.2859
4	18	SLC1A4, F13A1, RGSL2, GUSB	0.0053	0.7451	0.4517

3 Statistical Analysis

The first step of the statistical analysis we performed was to identify a set of differentially expressed genes between different groups of subjects. Disease status of subjects came from the clinical portion of the CFS data (*Intake Classific* variable). All subjects included in the microarray study were classified into 5 different groups: *Ever CFS* - 45 subjects ever experiencing CFS, *Nonfatigues* - 34 controls who never experienced CFS, *Ever ISF* - 45 subjects who are fatigued but cannot be classified as CFS because of insufficient symptoms, *Ever ISF-MDDm* - 20 subjects experiencing ISF with melancholic depression, *Ever CFS-MDDm* - 19 subjects experiencing CFS along with melancholic depression.

ANOVA F-test for each probe was carried out to determine differentially expressed genes across the five groups. 286 probes were identified as differentially expressed (*p-values* < 0.01). Since we are not interested in determining the differentially expressed genes per se, multiplicity correction was not used. The reduced microarray data consisting of 286 probes and 163 samples (subjects) was used later for further statistical analysis as discussed below.

3.1 Network Construction and Identification of Associated Gene Sets

To better understand the relationships between the selected 286 probes in terms of interactions/ associations, we employ two computational network inference techniques. The first method is based on the Partial Least Squares regression (PLS) (Pihur *et al.*, 2007), while the second method is based on the Partial Correlations (PC) (Schäfer and Strimmer, 2005). A number of similar characteristics are shared by the two approaches, such as computing association scores whose magnitude reflects the strength of the interaction between genes and local false discovery rate (local fdr) Empirical Bayes procedure for multiplicity adjustment in testing multiple hypotheses.

Table 2: Gene association modules discovered by the PC based network inference method.

# of Genes	Average Scores(%)	Gene Symbols	Severity p-value	hCV245410 p-value	hCV7911132 p-value
4	100	SRP68, MTA1, XM_049568, CLDN10	0.7588	0.3869	0.0328
5	85	ABCG8, NP_775847.1, UPP1, NM_022084, THSD1	0.0329	0.2552	0.1428
9	85	CASP3, XM72572, TMEM5, XM14557, CANT1, XM_033654, FOXF1, VCPIP1, PRUNE	0.1299	0.5498	0.2732
*24	84	CHST3, SIP1, TNK2, CLIC2, AK097480, NP_065988.1, XM_065828, EIF3S8, HES1, HOXA1, PMS2L5, KCNH2, XM66160, TNFRSF14, EFEMP1, KCNQ2, WASF3, Q8N8I1_HUMAN, MYPN, HDAC7A, WDR32, NP_620310.1, GPR41, MAP3K2	**0.0169**	**0.0586**	0.6642
6	84	NP_060367.1, SPATA11, XM_058846, CDC2L5, RAB32, NP_054868.2	0.0315	0.3867	0.1895
6	83	NAGA, CDH23, GAK, NP_061934.2, CK021_HUMAN, ZFYVE9	0.1886	0.8259	0.08
14	82	CHST4, CDR2, NP_114416.1, NP_056318.1, IKBKAP, KIRREL3, FAS, ZNF77, B3GALT3, MST1R, XM71032, PNLIPRP1, OPRD1, MRPL50	1e-04	0.9611	0.4225
4	81	VIPR1, CFLAR, SPTA1, ZNF7	0.0105	0.5447	0.7448
8	79	CNGB1, KRT20, TCIRG1, PGLYRP3, PRSS12, SMPX, XM_085181, XM70678	0.0932	0.6724	0.4883
4	78	CHD3, AK075566, XM14294, NP_062550.2	0.0536	0.838	0.9018

The results from applying the PLS and PC network reconstruction techniques to the reduced microarray data are summarized in the first three columns of Tables 1 (for PLS) and 2 (for PC). The actual visual representation of the networks themselves can be found on the supplementary website at http://www.somnathdatta.org/Supp/CamdaCFS/supp.htm. Both Tables 1 and 2 have the same structure. The first column shows the number of genes in distinct gene association modules (connected components) within each network. Gene association modules were defined to be clusters of 4 or more connected genes such that genes in two distinct components are not connected by an edge. Thus, it differs from the definition used in Presson *et al.* (2006). The tables are sorted by the second column which displays the percentages of each module's average association score when compared to the module with the largest average association score (the first module in each table). The exact definition of association scores are dependent on the method used. As for example, for the PC method, the association score of an edge is the partial correlation between the connected gene pair. Finally, in the third column we list all the genes belonging to each individual module. Genes shown in red are the genes that appear in both tables.

3.2 Regression of Symptom Severity

After identifying clusters of associated/interacting genes, we investigate the ability of each module to predict the CFS severity level. For that purpose, we fit a log-linear model for each gene module to regress the clinical variable *Cluster* on the set of expression profiles of genes included in the module. The overall predictive ability of the CFS severity by a given module can be judged on the basis of the likelihood ratio test which compares the full model (all genes in a module included as covariates in the model) and the null model which includes no covariates. The p-values obtained from the tests are shown in the fourth column of Tables 1 and 2. Small p-values indicate that gene association modules are effective in predicting the symptom severity categories.

Table 3: Common genes from the two PLS and PC clusters identified as predictive of disease severity status and SNP hCV245410 genotype. GO annotations and pathways were available from existing literature.

Gene	GO Process	Pathways	Description
WASF3	Cell Organization and Biogenesis, Metabolism	Adherens Junction	Actin-binding WH2
NUP98	Cell Organization and Biogenesis, Transport, DNA Replication	RAN regulation	
PRUNE	Energy production and conversion	Purine metabolism	Glycoside hydrolase, Phosphoesterase
KIRREL	Signal Transduction, Cell Adhesion		Integral to membrane, protein binding
TNK2	Cell Organization and Biogenesis, Signal Transduction, Protein amino acid phosphorylation	Regulation of CDC42 activity, Regulation of RAC1 activity	PAK-box/P21-Rho-binding, Protein kinase
EIF3S8	Protein Biosynthesis		Translation initiation factor activity
HOXA1	Transcription	p44/42 MAP kinase	Sequence-specific DNA binding
PMS2L5	DNA Repair		ATP binding, damaged DNA binding
HDAC7A	DNA Metabolism, Transcription		Histone deacetylase 7A
GPR41	Signal Transduction	p53/Bax pathway	G Protein-Coupled Receptor
MAP3K2	Protein amino acid phosphorylation	Mapk signaling, Gap Junction	Mitogen-activated protein kinase

3.3 Regression of SNP

Carrying out a similar analysis as in the previous section, we study how effectively each gene cluster (module) can predict the genotypes of the two SNP's, hCV245410 and hCV7911132, which have been identified by Presson *et al.* (2006) to be associated with symptom severity. Again, we fit multiple

log-linear models and compute the p-values for the likelihood ratio tests. The p-values for both SNP's are shown in columns 5 and 6.

4 Discussion of Results

Two gene association modules (indicated by asterisks) are of interest based on their predictive ability of symptom severity and of, at least, one of the SNP genotypes. The first cluster comes from the PLS reconstructed network and the other one from the PC reconstructed network. Table 3 lists the eleven genes that are in common between these two gene modules. The GO annotations listed in the table were mined from the BioGrid online repository (Stark *et al.*, 2006) and the pathway analysis was conducted using the DAVID webtools (Dennis *et al.*, 2003) in addition to mining existing literature.

It is plausible that these genes are responsible for certain aspects of CFS or its symptoms. As for example, the first gene on the list WASF3 (aka WAVE3) is thought to take part in the p38 MAPK regulatory pathway (Sossey-Alaoui *et al.*, 2005). On the other hand, in recent animal model studies (Katafuchi *et al.*, 2006), it has been demonstrated that regulation of brain cytokines through p38 MAPK pathway is involved in the in the central mechanisms of fatigue and therefore may play a role in the pathogenesis of the CFS. The list also includes autoimmune response gene NUP98 and genes related to tumor activities (PRUNE, TNK2, HOXA1). Gene expression of HDAC7A has been shown to be correlated with unexplained fatigue in a past study (Whistler *et al.*, 2006). The gene GPR41's role in autoimmune disorders including CFS has been hypothesised in Staines (2005).

References

Dennis, G., J., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. A. (2003). David: Database for annotation, visualization, and integrated discovery. *Genome Biol*, **4**(5), P3.

Katafuchi, T., Kondo, T., Take, S. and MEGUMU Yoshimura, M.(2006). Brain cytokines and the 5-HT system during Poly I:C-induced fatigue *Ann. N.Y. Acad. Sci.*, **1088**, 230237.

Pihur, V., Datta, S., and Datta, S. (2007). Reconstruction of genetic association networks from microarray data: A partial least squares approach. (under revision for *Bioinformatics*).

Presson, A., Sobel, E., Papp, J., Lusis, A., and Horvath, S. (2006). Integration of genetic and genomic approaches for the analysis of chronic fatigue syndrome implicates forkhead box n1. (CAMDA 2006 Conference Paper).

Schäfer, J. and Strimmer, K. (2005). An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**(6), 754–64.

Sossey-Alaoui, K., Ranalli, T. A., Li, X., Bakin, A. V., Cowell, J. K.(2005). WAVE3 promotes cell motility and invasion through the regulation of MMP-1, MMP-3, and MMP-9 expression. *Experimental Cell Research*, **308**(1), 135–145.

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). Biogrid: a general repository for interaction datasets. *Nucleic Acids Res*, **34**(Database issue), D535–9.

Staines, D. (2005). Are vasoactive neuropeptide autoimmune fatigue-related disorders mediated via G protein-coupled receptors? *Medical Hypothses*, **65**, 29–31.

Vernon, S. D. and Reeves, W. C. (2006). The challenge of integrating disparate high-content data: epidemiological, clinical and laboratory data collected during an in-hospital study of chronic fatigue syndrome. *Pharmacogenomics*, **7**, 345–354.

Whistler, T., Taylor, R. *et al.* (2006). Gene expression correlates of unexplained fatigue. *Future Medicine*, **7**,(3) 395–405.