EXTENDED ABSTRACT

Identifying Potential Biomarkers for Chronic Fatigue Syndrome via Classification Model Ensemble Mining

Ben Goertzel[1], Lucio Coelho[1], Cassio Pennachin[1]

[1]Biomind LLC
1405 Bernerd Place
Rockville MD 20851

We have analyzed the CAMDA 2006 data using a novel "classification model ensemble mining" (CMEM) methodology, in which genetic programming and heuristic search are applied to learn ensembles of classification rules that distinguish CFS from Control (one set of classifiers based on microarray data, and one based on SNP data), and these ensembles are then statistically analyzed to identify genes, gene categories, and combinations thereof that appear to play important roles in characterizing CFS. The results of this analysis include potential microarray and SNP based diagnostic rules for CFS, and also lists of SNP's, genes and gene categories that are potentially significant biomarkers for CFS (and are different from those found via simple statistical category-differentiation analysis). Overall, our results appear compatible with a system-theoretic view of CFS which views the disorder as a complex pattern of activity across the organism including interlinked disturbances in neural and endocrine systems.

Analysis of Microarray Data

The first step in applying CMEM to microarray data is to transform the microarray data profiles corresponding to individuals into numerical "feature vectors." To produce a feature vector from a microarray data profile, the data is first log-transformed and z-normalized, and then the microarray gene expression vector associated with an individual is extended into an "enhanced feature vector." The entries of an enhanced feature vector are either (transformed, normalized) gene expression values, or else values derived from these, each one corresponding to the average (transformed, normalized) expression of all the genes measured in an individual that belong to a given Gene Ontology or Protein Information Resource category.

Given a set of feature vectors divided into two categories, CMEM begins by using genetic programming [1] to learn an ensemble of classification models, each of which distinguishes the two categories using some learned rule. Given an ensemble of classification models, a calculation is done to determine, for each feature, the percentage of models that use that feature. (In the algorithm as currently implemented, we simply assume that all features in a given GP-learned model are equally important to that model.) This gives us a list of the features that are most useful for distinguishing the two categories – since they are frequently used as tools for building accurate classification models. The ordered list of most useful features may be subjected to qualitative biological relevance analysis: these are potential biomarkers, and potential components of combinational biomarkers.

The final step in CMEM is what we call MUTIC, or Model Utilization-based Clustering, which is a method for grouping together collections of features possessing interrelationships relevant to the categorization problem at hand. Toward that end, we associate each feature with a "utilization vector." The i'th entry of the utilization vector for the feature f is 1 or 0 depending on whether or not the feature f is used in the i'th model. The utilization vectors are then clustered, using a clustering algorithm called Omniclust, which is a variation on standard hierarchical clustering, with the cosine measure as the underlying vector similarity metric. These clusters may contain subsets that are combinational biomarkers.

```
div
 mul
  mul
   mul
    const 0.067531
    input AF044127
   mul
    const 0.002924
    const 0.023923
  mul
   mul
    input BC036349
    input AB037886
   sub
    const 0.480216
    input AK023090
 sum
  sum
   sum
    input BC033933
    input AF273052
   mul
    input NM_031954
    input BC001078
  sub
   sub
    input AK027884
    input L39833
   sub
    input AB025009
    input D86640
```

Figure 1. Example Microarray Classification Rule

Figure 1 shows an example classification rule produced via applying the CMEM methodology to the CAMDA 2006 microarray data (CFS versus NF=Non-fatigued). The Figure shows the rule in tree form, which is the form used internally by the GP learning algorithm. To evaluate the displayed rule on a given individual's gene expression profile, first the non-constant leaf nodes are replaced with that individual's gene expression values: e.g. NM_008149 is replaced with the expression value of that gene in the individual, GO:0019984 is replaced with the average expression value of genes in that GO category in the individual; FAM0008332 is replaced by the average expression value of genes coding for proteins in that PIR protein family. Then the arithmetic operations are performed, proceeding from the leaf nodes up. For instance, the final 7 lines of the rule are equivalent to the algebraic expression (AK027884 - L39833) - (AB025009 - D86640). This rule contains only arithmetic operators so it is equivalent to a rational function; one may also find high-quality rules involving logical operators and inequalities. Table 1 shows the confusion matrix corresponding to the 10x10 cross-validated classification run that produced that model (among others).

Table 2 lists the features that occurred most often in the top 100 classification rules found. For each feature, it also shows the rank of that feature that obtains if one lists features ordered by their differentiation between CFS and Control. Full interpretation of the feature list in Table 2 requires a more in-depth analysis, but some aspects are fairly straightforward. For example, the potential relevance of the autoimmune regulator AB006684 is clear; and the top feature SARS2 is well-known to relate to caspase 3 [2], which is part of the neural apoptosis process [3], which has been implicated in CFS [5].

Finally, Table 3 shows one of the top-quality clusters obtained via MUTIC. Note the presence of genes related to glucocorticoid response and neurotransmitter metabolism, a relationship perhaps meriting future investigation, particularly in light of the prominent role of glucocorticoid response in the SNP results to be described below.

Expected	**Computed**	
	False	**True**
False	19	16
True	10	30

Table 1. Confusion Matrix for Classifying CFS vs. Control Using Microarray Data

Feature	**Description**	**Utility**	**Variance Rank**
BC001020	SARS2 - seryl-tRNA synthetase 2	0.901	11
AF365931	ZIM3 - zinc finger, imprinted 3	0.9	12
FAM0002474	Component genes:	0.899	33
	NM_001835: Homo sapiens clathrin, heavy polypeptide-like 1 (CLTCL1), transcript variant 1, mRNA.		
AB004064	TMEFF2 - transmembrane protein with EGF-like and two follistatin-like domains 2	0.898	22
AB006684	AIRE - autoimmune regulator (autoimmune polyendocrinopathy candidiasis ectodermal dystrophy)	0.898	34
BC038855	TMEM16K - transmembrane protein 16K	0.896	2
BC036207	C10orf48 - chromosome 10 open reading frame 48	0.896	8
AB020710	EHBP1 - EH domain binding protein 1	0.895	1
AF484964	SH2D1B - SH2 domain containing 1B	0.895	9

Table 2. Top 10 Features Based on Classification Model Utilization
The Utility column shows the percentage of classification models in the final population of a GP run that utilize the indicated feature.

Feature	Description
AK090939	WDR49 - WD repeat domain 49
GO:0002009	morphogenesis of an epithelium
AK097310	MYOC - myocilin, trabecular meshwork inducible glucocorticoid response
NM_004140	Homo sapiens lethal giant larvae homolog 1 (Drosophila) (LLGL1), mRNA.
SF001149	Thrombin
GO:0006903	vesicle targeting
NM_152291	Homo sapiens mucin 7, salivary (MUC7), mRNA
XM_032542	Homo sapiens FLJ41352 protein (FLJ41352), mRNA.
GO:0006875	metal ion homeostasis
NM_001132	Homo sapiens AFG3 ATPase family gene 3-like 1 (yeast) (AFG3L1),mRNA.
SF002108	fragile X mental retardation syndrome protein
GO:0042133	neurotransmitter metabolism
NM_018560	Homo sapiens WW domain containing oxidoreductase (WWOX), transcript variant 2, mRNA.

Table 3. Example MUTIC cluster with high cluster quality
This cluster also contained three features without descriptions in our database:
XM_114099, BC010433, NM_014158

Analysis of SNP Data

Analysis of the CAMDA SNP data using evolutionary learning and enumerative search algorithms resulted in the discovery of a number of SNP-combinations which give significant classification accuracy for distinguishing CFS from Control. The CMEM approach was then used to mine the discovered classification rules to identify "important SNPs." For these experiments, we labeled all pure CFS individuals as Case and all NF and ISF individuals as Control. Table 4 shows the best SNP classifiers found, together with their classification accuracies measured across the whole dataset.

The classifiers we used here are called "pattern-strength classifiers"; each one is simply a set of SNP's plus a threshold. Such a classifier is used to classify an individual as Case or Control based on the following approach. For a given individual being evaluated by a particular rule, a "sum of SNPs" is computed via the rule: if the individual has a SNP s (present in the SNP list of the rule) in homozygosis, then the value 2 is summed for s; is s is present in heterozygosis, then 1 is summed; finally, if s is undetermined for that individual, then 0 is summed. After this sum is computed for all SNPs in the rule list, the value is compared with the rule threshold: if it is greather than the threshold, the individual is classified as CFS, otherwise Control.

We used two approaches to find pattern-strength classifiers for the CAMDA SNP data: genetic algorithms (GA) [5] and enumerative search. In the GA approach, the GA was executed 10 times in order to create an ensemble of 10 pattern-strength classifiers (each one being the fittest from the final population of one run), and this ensemble was used as a single classifier by direct voting. This ensemble-building classification process was executed 100 times using 3x3 cross-validation. The best execution achieved 78.5% out-of-sample cross-validated accuracy.

The enumerative search approach, on the other hand, yielded individual pattern-strength classifiers with significant accuracy. We simply searched through all possible pattern-classifiers involving 4 or fewer SNP's and selected the ones with highest classification accuracy. The best classifiers found in this approach are shown in Table 4. The statistical significance of these enumerative search results was established via permutation analysis: enumerative search of shuffled versions of the dataset did not yield pattern-classifiers with comparably high accuracy.

Accuracy	Confusion Matrix		Threshold	SNPs
75.2%	45 12	13 31	7	TPH2_15836061 NR3C1_1046360 5HTT_7911143 CRHR2_11823513
	45 12	13 31	7	TPH2_1843075 NR3C1_1046360 5HTT_7911143 CRHR2_11823513

Table 4. Top-Accuracy SNP-Based Classifiers

Table 5 shows the top 5 SNP's used in all pattern-strength classifiers found for CFS vs. Control, and Table 6 shows the top 5 genes (the genes whose SNP's have the most classification model usage, all total).

SNP	Frequency	Frequency (%)
POMC_3227244	547	18.2
MAOB_15763403	412	13.7
MAOB_15959461	261	8.7
NR3C1_1046353	249	8.3
TH_243542	225	7.5

Table 5. Most Useful SNP's

Gene	Short Description	Frequency	Frequency (%)
NR3C1	glucocorticoid receptor	977	32.6
MAOB	monoamine oxidase B	884	29.5
TPH2	tryptophan hydroxylase 2	705	23.5
POMC	proopiomelanocortin	547	18.2
COMT	catechol-O-methyltransferase	530	17.7

Table 6. Most Useful Genes for SNP-Based Classification

Discussion

Chronic Fatigue Syndrome is a complex disorder and the data we have analyzed is almost surely not adequate for achieving a full understanding of its causes and dynamics, nor for arriving at completely accurate diagnostics. However, the results we have obtained provide some indications of genes, mutations and categories and combinations thereof that may be relevant to CFS and may potentially serve as CFS biomarkers.

We have found statistically significant classification rules predicting CFS vs. Control, separately based on both microarray and SNP data. Neither of these classification rules achieves the 90%+ accuracy that one often finds when learning classification rules for diseases such as cancer, but this is unsurprising given the greater complexity of CFS and the uncertainties involved with CFS case definition.

Analysis of microarray data reveals that some genes related to brain and immune function are important for distinguishing CFS from Control, which is not surprising in light of the diverse evidence pointing to a central role for neurological dysfunction in CFS. Analysis of SNP data reveals a number of important endocrine-related genes, mutations and mutation combinations. MUTIC clustering of the microarray data reveals clusters combining neural and endocrine related genes; in particular, glucocorticoid receptor related genes occur both in the MUTIC microarray results and the SNP results. Taken together, these results support the general concept that CFS may be a systemic disorder involving problems with both the brain and the endocrine system, and complex feedback dynamics between these two organs.

We suspect that ultimately the results we obtained from analyzing the CAMDA CFS data may be found compatible with a system-theoretic view of CFS as schematically depicted in Figure 2.

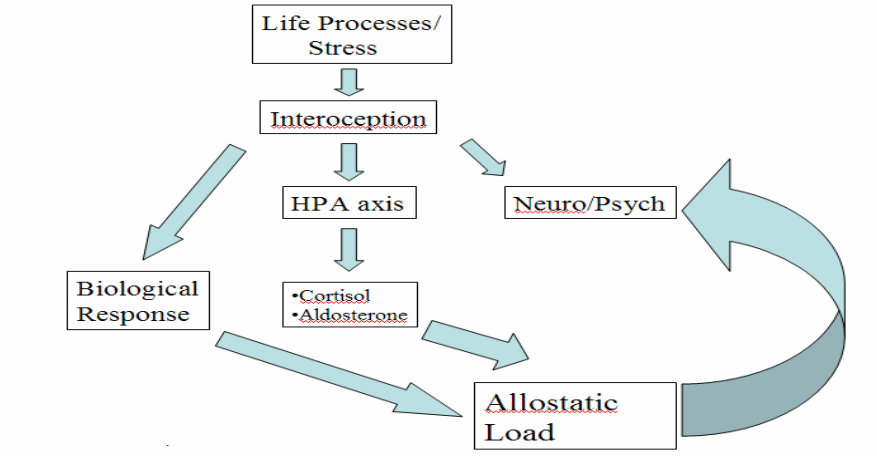

Figure 2. Hypothesis: CFS as a Systemic Disorder

Figure 2 integrates our present data analysis results with results from the literature that suggest a central role for the interoceptive process in CFS [6,7], as well as results recently obtained during collaborative research with Elizabeth Mahoney, Brian Gurbaxani and James Jones via analyzing other CFS-related clinical data not contained in the CAMDA dataset, indicating that an appropriately-defined concept of "allostatic load" can be used to help distinguish CFS from Control [8,9]. In this hypothetical interpretation, the distinctively CFS-relevant features we have observed in endocrine-related SNP's and neural and immune related genes reflect disturbances in particular parts of the body system which are part of the overall systemic disorder.

Potentially, future research may reveal that some of the neural and endocrine genes, SNPs and combinations found in this study may serve as biomarkers for CFS or for particular sub-syndromes of CFS. Among others, the glucocorticoid receptor related genes, and also SARS2 and AIRE, would seem to be candidates worth exploring.

REFERENCES

[1] J. Koza. *Genetic Programming*. MIT Press, 1991.

[2] C. Casas, J. Ribera, J. Esquerda. Antibodies against c-Jun N-terminal peptide cross-react with neo-epitopes emerging after caspase-mediated proteolysis during apoptosis. *Journal of Neurochemistry*, 2001-05-01

[3] G. Amarante-Mendes, K. Naekyung, L. Liu, Y. Huang, C. Perkins, D.R. Green. Bhalla K Bcr-Abl exerts its antiapoptotic effect against diverse apoptotic stimuli through blockage of mitochondrial release of cytochrome C and activation of caspase-3. *Blood*. 1998 Mar 1;91(5):1700-5 153-164

[4] World Patent # WO0215929; issued February 28, 2002; filed August 16, 2001; titled Methods and compositions for use in the diagnosis and treatment of chronic immune disease; Inventors: Englebienne P, DeMeirleir KL, Herst CVT; Applicant: R.E.D. Laboratories, N.V.

[5] D. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley, 1998

[6] N. Afari, D. Buchwald. Chronic fatigue syndrome: a review. *Am J Psychiatry* 2003;160: 221-36.

[7] P. White, J. Thomas, H. Kangro, W. Bruce-Jones, J. Amess, D. Crawford, et al. Predictions and associations of fatigue syndromes and mood disorders that occur after infectious mononucleosis. *Lancet* 2001;358: 1946-54.

[8] E Maloney, B. Gurbaxani, J. Jones, L. Coelho, C. Pennachin, B. Goertzel (2006). Chronic Fatigue Syndrome is Associated with High Allostatic Load, submitted for publication

[9] B. Goertzel, C. Pennachin, L. Coelho, E. Maloney, J. Jones, B. Gurbaxani (2006). Allostatic Load is Associated with Symptoms in CFS Patients, submitted for publication