



For reprint orders, please contact:
reprints@futuremedicine.com

Combinations of single nucleotide polymorphisms in neuroendocrine effector and receptor genes predict chronic fatigue syndrome

Benjamin N Goertzel^{1,2†},
Cassio Pennachin²,
Lucio de Souza Coelho²,
Brian Gurbaxani³,
Elizabeth M Maloney³ &
James F Jones³

[†]Author for correspondence

¹Virginia Tech,
National Capital Region,
Arlington, VA, USA

²Biomind LLC,
Rockville, MD, USA

E-mail: ben@goertzel.org

³Centers for Disease Control
and Prevention,
Atlanta, GA, USA

Objective: This paper asks whether the presence of chronic fatigue syndrome (CFS) can be more accurately predicted from single nucleotide polymorphism (SNP) profiles than would occur by chance. **Methods:** Specifically, given SNP profiles for 43 CFS patients, together with 58 controls, we used an enumerative search to identify an ensemble of conjunctive rules that predict whether a patient has CFS. **Results:** The accuracy of the rules reached 76.3%, with the highest accuracy rules yielding 49 true negatives, 15 false negatives, 28 true positives and nine false positives (odds ratio [OR] 8.94, $p < 0.0001$). Analysis of the SNPs used most frequently in the overall ensemble of rules gave rise to a list of 'most important SNPs', which was not identical to the list of 'most differentiating SNPs' that one would calculate via studying each SNP independently. The top three genes containing the SNPs accounting for the highest accumulated importances were neuronal tryptophan hydroxylase (*TPH2*), catechol-*O*-methyltransferase (*COMT*) and nuclear receptor subfamily 3, group C, member 1 glucocorticoid receptor (*NR3C1*). **Conclusion:** The fact that only 28 out of several million possible SNPs predict whether a person has CFS with 76% accuracy indicates that CFS has a genetic component that may help to explain some aspects of the illness.

A significant portion of the genetic variation between individuals is due to the combined effect of a number of relatively easily detectable single-base changes in the genome, called single nucleotide polymorphisms (SNPs). While it is currently not economically feasible to sequence all the DNA in population studies, examination of the targeted and pathway-specific SNPs is possible.

One potential opportunity offered by SNP analysis is the identification of markers for genetic predisposition to disease. Marker identification may lead to novel diagnostic tests that provide information about genes or proteins suitable as targets for pharmaceutical intervention. The standard approaches for finding such biomarkers involve linkage analysis and association studies [1]. However, both of these methods have significant shortcomings. Linkage analysis requires the study of families with known pedigrees and disease histories, a type of data that is often not available in practice. Association studies do not require historical data but, like linkage analysis, are not generally able to handle phenomena involving phenotypic outcomes resulting from multiple loci effects. In order to identify combinations of SNPs that affect a phenotypic trait, a more sophisticated analytical approach is required.

The discipline of machine learning contains a variety of algorithms oriented toward searching complex datasets to find combinations of

features that predict an outcome [2]. For example, several researchers have recently applied a powerful machine learning algorithm, support vector machines (SVMs), to the problem of identifying combinations of SNPs capable of predicting susceptibility to diseases. Waddell and colleagues [3] have applied SVMs to predict susceptibility to multiple myeloma. Their work provided 71% accuracy on a dataset containing 40 cases and 40 controls. Listgarten and colleagues considered SNPs from 45 genes of potential relevance to breast cancer etiology in 174 patients, compared with matched normal controls [4]. These authors found 69% accuracy using SVMs as a learning algorithm, and almost equally good accuracy using simpler methods such as naive Bayes classification and decision trees. Overall, they concluded that multiple SNP sites from different genes over distant parts of the genome are better at identifying breast cancer patients than any one SNP alone.

We present a case study involving the application of machine learning algorithms to analyze this form of genetic data. We applied both SVMs and a simple enumerative search approach to a dataset consisting of SNP profiles for 43 chronic fatigue syndrome (CFS) patients and 58 non-fatigued control subjects, where cases and controls are defined as per Reeves and colleagues [5]. Enumerative search reveals numerous sets of 3–5 SNPs capable of serving as statistically significant

Keywords: chronic fatigue syndrome, single nucleotide polymorphism, supervised machine learning

future
medicine

classification rules. Significant classification rules involving larger sets of SNPs also exist but, according to our analyses, add very little predictive accuracy to that obtained with five-SNP sets. However, this conclusion must be considered tentative rather than definitive, since for sets larger than five we were only able to conduct heuristic rather than exhaustive searches of the space of possible sets due to computing resource limitations. Statistical analysis of the ensemble of significantly predictive SNPs yields a list of SNPs likely to be of interest in CFS, either in isolation or in their combination with other SNPs, and a list of genes containing these important SNPs. SVM-based analysis provides significantly lower accuracy than enumerative searches on this dataset.

Methodology

In order to explore the hypothesis that combinations of SNPs are more strongly predictive of CFS compared with individual SNPs considered in isolation, we used computational algorithms to search for sets of SNPs that were present in CFS with significantly different frequency than in controls [101]. As a preliminary step before carrying out analysis, we removed the sex-linked genes from the dataset (monoamine oxidase [*MAO*]A and *MAOB*), and also removed all genes not in Hardy-Weinberg equilibrium according to a χ^2 test. This left the 28 SNPs shown in Table 1 as the basis for our computational analysis.

The methodology used was similar to that taken by several of the authors in Smigrodzki and colleagues [6], with the difference that in that prior work the patterns involved heteroplasmic mutations, rather than SNPs. There, a genetic algorithm was used to search the space of possible combinations of subsequences of mitochondrial genes corresponding to structural protein domains, searching for those with heteroplasmic mutation densities significantly different in the case and control groups. On the other hand, in the present study, since the total number of SNPs studied was not very large, it was feasible to proceed via an enumerative search rather than by utilizing a more complex search technique such as genetic algorithms or genetic programming [7]. To identify 1–5 SNPs of interest, we listed all sets of test SNPs with cardinality of less than 6 and evaluated each one as a potential classification rule for distinguishing CFS from controls.

Each SNP set was evaluated as a potential classification rule by being interpreted as a pattern strength classifier as detailed here (the software

use to perform this classification is not currently publicly available, but plans are underway to make an open-source version freely available via a collaboration with the National Institutes of Health [NIH] Clinical Center). Each pattern strength classifier is simply a list of SNPs and a threshold. For a given individual being evaluated by a given rule, the 'sum of SNP incidences' is computed in the following way: if the individual has a SNP (present in the SNP list of the rule) in homozygosity, then the value 2 is summed for *s*; if the SNP is present in heterozygosity, then 1 is summed; finally, if *s* is undetermined for that individual, then 0 is summed. This scoring rule was chosen empirically, based on its ability to yield good classification rules. After this sum is computed for all SNPs in the rule list, the value is compared with the rule threshold: if it is greater than the threshold, the individual is classified as CFS, otherwise control. For each SNP set, the threshold value is selected that allows the SNP set to achieve the maximum accuracy for distinguishing cases from controls.

In addition to finding effective rules for distinguishing CFS from controls, it is also interesting to ask which SNPs are most effective (important) in the context of the ensemble of classification rules found. Toward this end, we calculated the frequency of each SNP in the set of classification rules found, including in the calculation all conjunctive classification rules of any size, but restricting attention to classification rules with accuracy better than the frequency of the most frequent category (57.4%, in the case of the dataset used).

Results

For CFS versus controls, enumerative search in the range of 1–5 SNPs produced a number of reasonably high-quality rules as shown in Table 2. The threshold for each rule and a standard confusion matrix (with true negatives in the upper left, true positives in the lower right, and false positives and negatives in the upper right and lower left, respectively) are shown with the SNPs used in the simple additive rule.

To provide extra validation of these results we performed a shuffling (permutation) analysis. We created 115 shuffled versions of the data and performed a subset of the analyses described above on each one. The shuffling was performed by randomly relabeling the patients as CFS or control, while keeping the total number of patients in each category equal. The conclusion was that the top accuracies found from enumerative analysis are

Table 1. The 28 SNPs used in our analysis, after removal of sex-linked SNPs and SNPs not in Hardy-Weinberg equilibrium.

SNP	NCBI ID
COMT_11804654	rs740603
COMT_2539273	rs933271
CRHR1-1570087	rs7209436
CRHR2_15872871	rs2267710
CRHR2_15960586	rs2284217
NRC1_1046361	rs6196
TPH2_11407441	rs1872824
CRHR1-2257689	rs242924
CRHR1-2544830	rs173365
POMC_3227244	rs12473543
TH_243542	rs4074905
TPH2_15836061	rs2171363
TPH2_1843075	rs2070762
5HTT-1841702	rs2066713
COMT_2538747	rs4633
NR3C1_1046353	rs6188
NR3C1_1046360	rs258750
NR3C1_8950998	rs852977
TPH2_8376146	rs4760750
TPH2_8376173	rs4760816
TPH2_8872233	rs1487280
COMT_2539306	rs165722
COMT-11804650	rs4646312
CRHR2_11823513	rs2267714
NR3C1_11159943	rs1866388
TPH2_8376042	rs1386486
COMT_2538746	rs6269
TH_245410	rs10784941

5HTT: 5-hydroxytryptamine transporter; COMT: Catechol-O-methyltransferase; CRHR: Corticotropin-releasing hormone receptor; NCBI: National Center for Biotechnology Information; NR3C1: Nuclear receptor subfamily 3, group C, member 1; NRC1: Nonpapillary renal carcinoma 1; POMC: Proopiomelanocortin; SNP: Single nucleotide polymorphism; TH: Tyrosine hydroxylase; TPH2: Tryptophan hydroxylase 2.

significant with $p < 0.01$, meaning that the odds of finding five or fewer SNP rules of this accuracy from a randomly shuffled version of the dataset are less than 1%. To illustrate the qualitative differences between the results obtained for shuffled and unshuffled data, Table 2 shows, for each accuracy value, the ratio between the percentage of rules found for the real data with this accuracy and the percentage of rules found for the shuffled data with this accuracy. More definitively, Table 3 shows the percentage of the shuffled datasets on which the best model found lay in each of a series of bins. As this illustrates, none of these shuffled datasets led to models with accuracy comparable

with that found on the real dataset, indicating that the results are clearly significant with $p < 5\%$ (and arguably $p < 1\%$). Figures 1 and 2 further illustrate the significance of the differences between the results on the unshuffled and the results on the shuffled data.

These 115 shuffled tests occupied roughly 2 days of compute time on two 2GHz computers. Of course, the compute time of applying these algorithms may vary significantly when applied to other problems. Generally, the compute time will increase linearly with the number of patients, but polynomially with the number of SNPs (for example, with N SNPs the number of quintuplets of SNPs is $O[N^5]$), which means that if there are too many SNPs one must resort to approaches more sophisticated than complete enumeration, such as genetic algorithms or genetic programming.

To establish which SNPs are most important for CFS overall, we calculated the frequency of each SNP in the set of enumerative models with accuracy greater than the frequency of the most frequent category (Table 3). Another interesting analysis of importance consists of computing the total importance of each gene involved in the study by simply summing up the importances of all SNPs belonging to that gene. This analysis is shown in Table 4.

In order to explore the possibility that the imperfect accuracy of these classifiers is due to the relatively simplistic nature of the learning algorithm, we also tried the SVM learning algorithm on the dataset. For the purposes of the algorithm, undetermined SNPs were assigned a value of 0.0, homozygotes for allele 1 a value of 0.33, homozygotes for allele 2 a value of 0.66, and heterozygotes a value of 1.0. However, it should be noted that these are not scores in the same sense as the pattern strength classifier described above. Rather, they simply define unique locations for each SNP in the multidimensional space. However, the performance of the SVM on this dataset was disappointing, yielding a maximum of 68% accuracy across 1000 runs with different parameter values. Table 5 compares the distribution of accuracies across the 1000 SVM runs with the distribution of accuracies of the classification models identified via complete enumeration on the dataset.

If we had to choose a minimal list of genes for follow-up analysis, based on the above results, it would be neuronal tryptophan hydroxylase (*TPH2*), *5HTT* and nuclear receptor subfamily 3, group C, member 1 glucocorticoid receptor

Table 2. Rules achieving high classification accuracy. A list of all rules achieving accuracies of 76.3 and 75.2% (the highest-accuracy rules found with up to five SNPs).

Accuracy	Confusion matrix		Threshold	SNPs
76.3%	49 15	9 28	6	rs1386486, rs1866388, rs6196, rs6188, rs2284217
75.2%	47 14	11 29	5	rs1386486, rs6196, rs6188, rs2284217
	47 14	11 29	6	rs4760750, rs1386486, rs6196, rs6188, rs2284217
	47 14	11 29	6	rs1386486, rs4633, rs6196, rs6188, rs2284217
	49 16	9 27	6	rs1386486, rs1866388, rs852977, rs6196, rs2284217
	47 14	11 29	6	rs1386486, rs852977, rs6196, rs6188, rs2284217

SNP: Single nucleotide polymorphism.

(*NR3C1*) genes, which, as Table 4 shows, had greater occurrence in the high-quality classification models than all others. However, the combined importance of corticotropin releasing hormone receptor (*CRHR1*) and *CRHR2* was relatively high, suggesting a potentially significant role for corticotropin-releasing hormones.

Discussion

CFS has many clinical components that are universally present in many illnesses. This paper studies data derived from CFS subjects identified purely by clinical parameters, and explores the question of whether these subjects demonstrate genetic markers in regions of the genome that may be associated with illness expression. The clinical parameters used for CFS diagnosis initially consisted of physician interview data, which were subsequently supported by numbers derived from scoring clinical questions that identify subjects based on the core components of the 1994 definition, fatigue, impairment in functioning and symptom presence and severity [4]. Based on the study design where SNPs were chosen for analysis in preselected genes possibly associated with production of symptoms or proposed in previous studies of CFS, it is not surprising that the specific SNPs identified in this analysis have the appearance of biological plausibility. However, it was not inevitable that statistically significant classification performance would be achievable via combining any of the SNPs from these preselected genes. This is a nontrivial result achieved via the systematic application of automated pattern search techniques.

As noted above, other authors have reported significant success applying SVM to identify cancer susceptibility based on SNP data [3,4]. We were somewhat surprised that our relatively simple enumerative technique was able to outperform SVMs on the present CFS dataset. Our explanation for this phenomenon is a simple one. SVMs are much more powerful than an enumerative search for finding patterns that combine a large number of SNPs. However, if the most significant patterns in the data are ones that involve a small number of SNPs, then it is obvious that an enumerative search will outperform SVMs, since it is guaranteed not to miss any of these patterns.

Whether evaluated by most important genes or by most important SNPs, the identified genes are responsible for the synthesis or regulation of production of neurotransmitters, steroid hormones or their cognate receptors. These proteins are inextricably involved in multiple levels of brain function, but also participate in stress reactions, emotional responses, generation and maintenance of memory, and are thought to contribute to their disorders. They also participate in function and regulation of function in other organ systems and maintain cellular and organism integrity. As such, it is certainly plausible that defects in the specific genes indicated by the identified SNPs are causative in the pathogenesis of the CFS cases studied. However, we must keep in mind that we do not know if these SNPs are causative or merely correlative, serving as markers for nearby genes that truly are involved in disease pathogenesis.

Table 3. SNPs distinguishing CFS from control.

SNP	Model utilization		Allele frequency in CFS (%)				Allele frequency in NF (%)			
	Incidence	Frequency (%)	AA	BB	AB	NA	AA	BB	AB	NA
rs852977	28224	40.8	48.8	13.9	37.2	0	32.7	17.2	50	0
rs6188	27197	39.3	11.6	48.8	37.2	2.3	17.2	32.7	50	0
rs2284217	24430	35.3	2.3	58.1	39.5	0	1.7	62	34.4	1.7
rs6196	22531	32.5	79	6.9	13.9	0	63.7	3.4	29.3	3.4
rs933271	14139	20.4	48.8	0	51.1	0	53.4	6.8	39.6	0
rs1386486	13350	19.3	16.2	39.5	44.1	0	12	41.3	46.5	0
rs2267710	13330	19.2	11.6	41.8	44.1	2.3	3.4	43.1	53.4	0
rs12473543	13145	19	4.6	62.7	30.2	2.3	0	67.2	32.7	0
rs7209436	12331	17.8	27.9	13.9	55.8	2.3	27.5	15.5	55.1	1.7
rs242924	11722	16.9	27.9	16.2	53.4	2.3	31	15.5	51.7	1.7
rs173365	11029	15.9	16.2	27.9	53.4	2.3	15.5	29.3	51.7	3.4
rs4646312	10779	15.5	32.5	13.9	51.1	2.3	36.2	15.5	46.5	1.7
rs4760750	10393	15	13.9	39.5	46.5	0	17.2	39.6	43.1	0
rs4633	10393	15	20.9	30.2	48.8	0	25.8	29.3	44.8	0
rs1487280	9878	14.2	18.6	37.2	44.1	0	13.7	41.3	44.8	0
rs10784941	9625	13.9	18.6	27.9	53.4	0	27.5	25.8	46.5	0
rs165722	9579	13.8	27.9	20.9	51.1	0	31	22.4	46.5	0
rs1866388	9326	13.4	48.8	6.9	37.2	6.9	32.7	13.7	50	3.4
rs4760816	9203	13.3	13.9	39.5	46.5	0	17.2	39.6	43.1	0
rs2070762	8598	12.4	32.5	23.2	44.1	0	20.6	31	46.5	1.7
rs2171363	8241	11.9	13.9	44.1	39.5	2.3	17.2	34.4	39.6	8.6
rs740603	8131	11.7	20.9	16.2	62.7	0	25.8	22.4	51.7	0
rs6269	7630	11	16.2	32.5	51.1	0	15.5	37.9	46.5	0
rs1872824	7414	10.7	18.6	41.8	39.5	0	17.2	41.3	41.3	0
rs2066713	6706	9.6	13.9	30.2	55.8	0	20.6	39.6	39.6	0
rs4074905	6343	9.1	48.8	6.9	41.8	2.3	55.1	12	31	1.7
rs2267714	4430	6.4	16.2	32.5	51.1	0	10.3	43.1	44.8	1.7
rs258750	4307	6.2	48.8	9.3	39.5	2.3	31	18.9	46.5	3.4

Important differentiating SNPs are listed in decreasing order of importance. These SNPs occur most often in the set of classification rules for distinguishing CFS from control with accuracy greater than the frequency of the largest category (57.4%). The incidence column indicates how many models contain that SNP, out of 69,160 total models giving accuracy better than the frequency of the most frequent category; the percentage column gives this number as a ratio of 69,160. This doesn't mean that each of these SNPs on its own provides accurate differentiation between CFS and controls (though some of them do, to a moderate extent). It means that each of these SNPs, in combination with various other SNPs, is useful as a component in combinational rules for distinguishing CFS from controls.

CFS: Chronic fatigue syndrome; NF: Nonfatigued; SNP: Single nucleotide polymorphism.

Regarding specific SNP findings, the correlative findings of lowered urine cortisol and aldosterone levels in CFS subjects, as seen in the accompanying papers [8,9], and SNPs in glucocorticoid receptors are of interest as both hormones bind to these receptors. However, the single morning samples of epinephrine (E) and norepinephrine (NE) levels were not related to illness status, and point to the need of repeated measures of all of these proteins of interest in these patients. However, E and NE did contribute to the allostatic load, a summed physiological

variable that differed between fatiguing illness states. The other genes of possible interest associated with clinical illness, but not with other laboratory findings, are *TPH2* and *HTT* because of their putative roles in affective disorders [10–12] and catechol-*O*-methyltransferase (*COMT*) for its role in cognitive and behavioral aspects of schizophrenia [13].

The importance of intergene interrelationships here is highlighted by the fact that five-SNP set classification rules significantly outperform single SNP rules in terms of classification

Table 4. Importances of the genes based on their SNPs.

Gene	Short description	Incidence
NR3C1	Nuclear receptor subfamily 3, group C, member 1 glucocorticoid receptor	69054
TPH2	Neuronal tryptophan hydroxylase	67077
COMT	Catechol- <i>O</i> -methyltransferase	60651
CRHR2	Corticotropin-releasing factor 2	42190
CRHR1	Corticotropin-releasing hormone receptor 1	35082
NRC1	Nonpapillary renal carcinoma 1 growth mediator	22531
TH	Tyrosine hydroxylase	15968
POMC	Proopiomelanocortin	13145
5HTT	5-hydroxytryptamine transporter	6706

*The incidence number reports the number of rules found with accuracy greater than the frequency of the largest category, utilizing some SNPs in the gene.
SNP: Single nucleotide polymorphism.*

accuracy. The synergies observed in terms of classification accuracy as a result of grouping SNPs together reflect biological synergies between the systems associated with the genes in which the SNPs occur. This is not surprising from a biological perspective, since each of the systems associated with these genes responds to internal signals through the process of interoception [14]. Ultimately, we suspect a systems

biology-oriented modeling approach will be necessary in order to more fully understand the interactions and synergies involved.

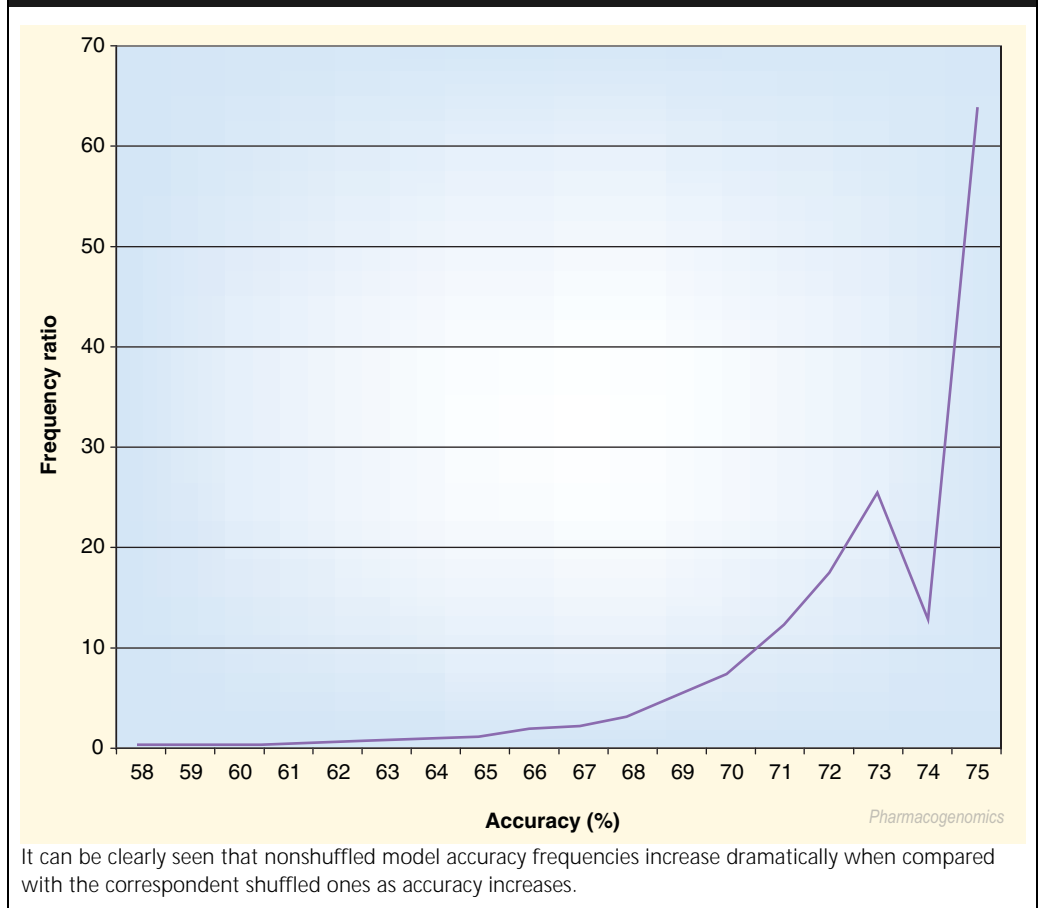
What is encouraging is that this rather mysterious and elusive illness called CFS appears to be finally yielding to attempts at biomarker discovery. SNPs with definite biological relevance to CFS are enriched in cases vs controls, and combinations of biologically relevant SNPs provide

Table 5. Accuracy distributions for enumerative search and a SVM 1000-fold test over real (unshuffled) data.

Accuracy (%)	Enumeration (%)	SVM (%)
58	23.21113649	
59	18.87449767	
60	14.88623551	
61	11.98357859	
62	8.919020498	
63	6.940067652	
64	4.378559658	
65	3.05154818	13.3257403189
66	2.372141432	67.6537585421
67	1.689843592	6.37813211845
68	1.274971812	12.6423690205
69	1.027783399	
70	0.604238342	
71	0.420653965	
72	0.209604209	
73	0.117089248	
74	0.027465379	
75	0.008673278	
76	0.002891092	

It is possible to see qualitatively that the enumerative distribution is prolonged into significantly high accuracy values unlikely to be achieved by chance. SVM tests, on the other hand, produced an accuracy distribution tightly concentrated in the 65–68% range, achieving no result nearly as good as the top ones from the enumeration approach. SVM: Support vector machine.

Figure 1. Ratio between the frequencies of models at various accuracies for real data versus shuffled data.



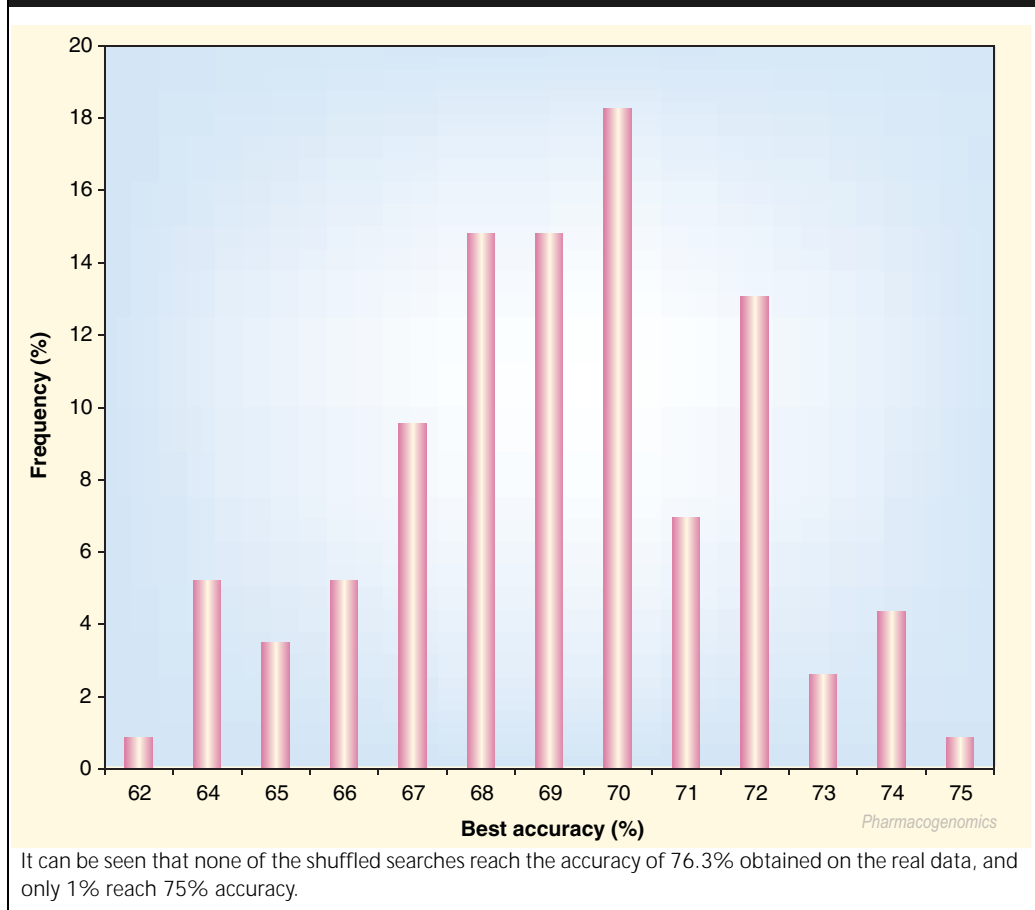
statistically significant classification accuracy for distinguishing CFS from controls. While the accuracies do not look spectacular compared with the accuracy of 57.4% obtainable by assigning all patients to the largest category or compared with the best accuracies of 65–70% found in a significant number of randomly shuffled datasets, there are also reasons to think the true accuracies may be better than the 76% reported. First, the case and control classifications that we used [4], although better than previous definitions, are almost assuredly inexact themselves in that no one is exactly sure how to define CFS or non-CFS, nor whether these are best considered as discrete categories, or rather on some sort of continuum. To illustrate the latter point, controls could have been defined more narrowly as those who are not currently suffering from unexplained, long-term fatigue and who have never suffered from such. Only 43 of the 58 controls used in the present study meet this more strict definition. When the analysis is performed using this more narrowly defined subset of controls,

the accuracy is increased to 78%. As for the case definition, it too is almost assuredly inexact, especially with regard to whether CFS is a single entity or many entities. The 76% accuracy observed is likely due in large part to a subset or subsets of CFS cases that are much more genetically linked to subtypes of the disease than to the larger group and would yield higher accuracies if considered alone.

Outlook

CFS is a syndrome and therefore shares complaints/symptoms and in some cases, laboratory and physical examination findings, with illness etiologies of diverse nature. The majority of human illnesses share, in some form or another, similar basic symptoms including: fatigue, cognitive problems, unrefreshing sleep, pain in a number of sites and decreased levels of functioning. CFS, due to its longevity and/or recurrences, presents an opportunity to address the mechanisms responsible for the production of these symptoms and their

Figure 2. Distribution of best accuracies of 115 enumerative searches on shuffled data.



consequences in subjects who do not have obvious pathogenic processes and who can be evaluated longitudinally.

Detection of genes of interest, as identified by SNPs, is frequently applied in diseases, which unlike CFS, have clear phenotypes where specific processes are either understood or hypothesized. CFS, which is aggravated by excessive mental or physical activity in the world at large, as well as in experimental paradigms, is an interesting candidate for study of 'sickness' in general. Identification of genes by SNP analysis that allow predisposition to symptom production and biological changes following proscribed activity can be realized using this illness as a model system.

The biological complexity of CFS and the subtlety of its definition have implications for data analysis methodology. Machine learning approaches are particularly valuable in this context, as they provide powerful mechanisms for discovering the multiple multigenic patterns associated with CFS and its symptoms. Traditional statistical methods with their focus on understanding the effects of single SNPs may be less useful for exploring CFS and its symptoms than for understanding more straightforwardly defined disease.

Disclaimer

Findings and conclusions in this report are those of the authors and do not necessarily represent the views of the funding agency.

Highlights

- Chronic fatigue syndrome (CFS) is a complex illness with no currently discernable biological markers.
- Evaluation of selected single nucleotide polymorphisms (SNPs) by enumerative search techniques outperformed machine learning in identifying CFS versus nonfatigued control subjects.
- The genes containing the most important SNPs were nuclear receptor subfamily 3, group C, member 1 glucocorticoid receptor (*NR3C1*), tryptophan hydroxylase 2 (*TPH2*), and catechol-O-methyltransferase (*COMT*).
- SNPs in genes associated with response to stress and to manifesting emotions may be markers of CFS.

Bibliography

1. Kwok PW (Ed.): *Single Nucleotide Polymorphisms: Methods and Protocols*. AACCC Press, San Francisco, CA, USA (2002).
2. Mitchell T: *Machine Learning*. McGraw-Hill, New York, NY, USA (1997).
3. Waddell M, Page D, Zhan F, Barlogie B, Shaughnessy J: Predicting Cancer Susceptibility from Single-Nucleotide Polymorphism Data: A Case Study in Multiple Myeloma, Proceedings of BIOKDD '05, Chicago, IL, USA (2005).
4. Listgarten J, Damaraju S, Poulin B *et al.*: Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. *Clin. Cancer Res.* 10(8), 2725–2737 (2005).
5. Reeves WC, Wagner D, Nisenbaum R *et al.*: Chronic fatigue syndrome – a clinically empirical approach to its definition and study. *BMC Med.* 3, 19 (2005).
6. Smigrodzki R, Goertzel B, Pennachin C, Coelho L, Prosdociami F, Parker DW: Genetic algorithm for analysis of mutations in Parkinson's disease. *Artif. Intell. Med.* 35(3), 227–241 (2005).
7. Koza J: *Genetic programming*. MIT Press, Cambridge, MA, USA (1992).
8. Maloney EM, Gurbaxani BM, Jones JF, Coelho LdS, Pennachin C, Goertzel BN: Chronic fatigue syndrome and high allostatic load. *Pharmacogenomics* 7(3), 467–473 (2006).
9. Gurbaxani BM, Jones JF, Goertzel BN, Maloney EM: Linear data mining the Wichita clinical matrix suggests sleep and allostatic load involvement in chronic fatigue syndrome. *Pharmacogenomics* 7(3), 455–461 (2006).
10. Levinson DF: The genetics of depression: a review. *Biol. Psychiatry* [Epub ahead of print] (2005).
11. Brown SM, Peet E, Manuck SB *et al.*: A regulatory variant of the human tryptophan hydroxylase-2 gene biases amygdala reactivity. *Mol. Psychiatry* 10(9), 884–888 (2005).
12. Serretti A, Mandelli L, Lorenzi C *et al.*: Temperament and character in mood disorders: influence of *DRD4*, *SERTPR*, *TPH* and *MAO-A* polymorphisms. *Neuropsychobiology* 53(1), 9–16 (2005).
13. Craig AD: How do you feel? Interoception: the sense of the physiological condition of the body. *Nature Rev. Neurosci.* 3, 655–666 (2002).
14. Strous RD, Lapidus R, Viglin D, Kotler M, Lachman HM: Analysis of an association between the *COMT* polymorphism and clinical symptomatology in schizophrenia. *Neurosci. Lett.* 393(2–3), 170–173 (2005).

Website

101. Critical Assessment of Microarray Data Analysis 2006 Conference Contest Datasets. www.camda.duke.edu/camda06/datasets/