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

Linear data mining the Wichita clinical matrix suggests sleep and allostatic load involvement in chronic fatigue syndrome

Brian M Gurbaxani[1†], *James F Jones*[1], *Benjamin N Goertzel*[2] & *Elizabeth M Maloney*[1]

†*Author for correspondence*
[1]*Centers for Disease Control and Prevention,*
600 Clifton Road, MS A-15,
Atlanta, GA 30333, USA
Tel.: +1 404 639 3699;
Fax: +1 404 639 2779;
E-mail: buw8@cdc.gov
[2]*Biomind LLC, Rockville, Maryland, USA*
E-mail: ben@goertzel.org

Objectives: To provide a mathematical introduction to the Wichita (KS, USA) clinical dataset, which is all of the nongenetic data (no microarray or single nucleotide polymorphism data) from the 2-day clinical evaluation, and show the preliminary findings and limitations, of popular, matrix algebra-based data mining techniques. **Methods:** An initial matrix of 440 variables by 227 human subjects was reduced to 183 variables by 164 subjects. Variables were excluded that strongly correlated with chronic fatigue syndrome (CFS) case classification by design (for example, the multidimensional fatigue inventory [MFI] data), that were otherwise self reporting in nature and also tended to correlate strongly with CFS classification, or were sparse or nonvarying between case and control. Subjects were excluded if they did not clearly fall into well-defined CFS classifications, had comorbid depression with melancholic features, or other medical or psychiatric exclusions. The popular data mining techniques, principle components analysis (PCA) and linear discriminant analysis (LDA), were used to determine how well the data separated into groups. Two different feature selection methods helped identify the most discriminating parameters. **Results:** Although purely biological features (variables) were found to separate CFS cases from controls, including many allostatic load and sleep-related variables, most parameters were not statistically significant individually. However, biological correlates of CFS, such as heart rate and heart rate variability, require further investigation. **Conclusions:** Feature selection of a limited number of variables from the purely biological dataset produced better separation between groups than a PCA of the entire dataset. Feature selection highlighted the importance of many of the allostatic load variables studied in more detail by Maloney and colleagues in this issue [1], as well as some sleep-related variables. Nonetheless, matrix linear algebra-based data mining approaches appeared to be of limited utility when compared with more sophisticated nonlinear analyses on richer data types, such as those found in Maloney and colleagues [1] and Goertzel and colleagues [2] in this issue.

Many traditional matrix algebra-based data mining techniques, such as principal components analysis (PCA), are good techniques to begin analyzing diverse, high-dimensional data sets. They often give a good idea of the basic shape, dimensionality and clustering of the data. However, these standard techniques often suffer from several problems in analyzing very high dimensional datasets (for example, microarray data), such as the 'curse of dimensionality' (i.e., a greater number of dimensions than subjects makes it hard to identify meaningful clusters in the sparsely filled space) [3], or the inappropriateness of a linear boundary to separate groups [4]. However, there is often something to be gained from analyzing almost any dataset using these data mining techniques, provided a good intuition for the data underlies the interpretation of results.

PCA is a powerful data mining technique that is widely used by itself or as the basis for many other widely used techniques, such as linear discriminant analysis (LDA) and factor analysis [5]. It is an excellent dimensionality reduction tool, and often reveals the essentially flat characteristics of many datasets, for example, that the data points do not fill the entire space they are embedded in. PCA has the advantage that it is computationally simple and robust (the results are not sensitive to removing data).

However, PCA has two principle drawbacks. The first is that the results are often hard for experts to interpret intuitively as PCA transforms the data into a rotated space, so that the data are embedded in a new coordinate system whose axes are linear combinations of the old coordinate system's basis vectors. To use the Wichita (KS, USA) clinical dataset as an example, whereas a physician could easily interpret a data mining result that says that chronic fatigue syndrome (CFS) cases have higher levels of blood glucose or thyroxine than controls, it would be harder to interpret if the discriminating criteria were on an axis

Keywords: allostatic load, chronic fatigue syndrome, data mining, feature selection

10.2217/14622416.7.3.455 © 2006 Future Medicine Ltd ISSN 1462-2416

that is composed of 15% glucose levels, 13% thyroxine, -7% diastolic blood pressure, 11% reaction time in a Cambridge Neuropsychological Test Automated Battery (CANTAB) test, and so on. However, this is what a PCA result might look like. The second drawback is that the new coordinate system resulting from PCA is organized along dimensions of maximum variance in the dataset, but the user is often interested in dimensions that maximally separate predefined groups of points, for example, CFS case and control. These are not necessarily the same thing.

Feature selection techniques do not suffer from either drawback. Feature selection techniques differ from many dimension reduction techniques in that they seek to reduce the number of parameters while preserving each parameter's meaning, i.e., they don't rotate the data into an abstract component space. Feature selection is a very active area of research in the computer science and bioinformatics communities. Currently, there is no consensus on the best method of performing feature selection – it varies by the dimension and type or types of data employed [6].

In this paper, we apply some standard data mining techniques and principles to the Wichita clinical data matrix (i.e., all of the data collected during the clinical evaluation, not including high throughput data such as single nucleotide polymorphism [SNP] and microarray) to learn its overall shape, dimensionality and characteristics with respect to CFS. We use feature selection to try and improve the quality of our data mining results, which were initially poor, on the purely biological (i.e., not self-assessed questionnaire-based) data.

Methods

We started with a 440 parameter by 227 subject identification (ID) matrix of clinical measurements that included self-assessed questionnaire data (SAQ) of various types (e.g., mental functioning, symptom inventory, sleep assessment, and so on), blood and urine measurements (cell counts, hormone and cytokine levels, and so on), mental functioning tests (CANTAB), lists of medications by function, and so on. Missing values in the matrix had been filled using mostly mean values for the remaining subjects for that variable. After examining the questionnaire data and determining that it was highly correlated with disease status, all questionnaire data, except the summary score for the Zung depression inventory and the Epworth sleepiness scale, were removed. Sparse areas of the matrix, such as the medications lists, were removed as they did not contain enough non-zero values to meaningfully differentiate between case and control. Parameters where a large percentage of values were filled with mean values and other parameters that showed no variation at all were also removed. In addition, subjects that did not meet the classifiaction criteria for either CFS, insufficient fatigue (ISF – i.e., those who partially meet the CFS criteria), or not ill (NI) by newly established criteria [7], and/or had major depressive disorder with melancholic features (MDDm) comorbidity or other psychiatric and medical exclusions, were removed. This left a reduced matrix of 183 parameters by 164 subjects.

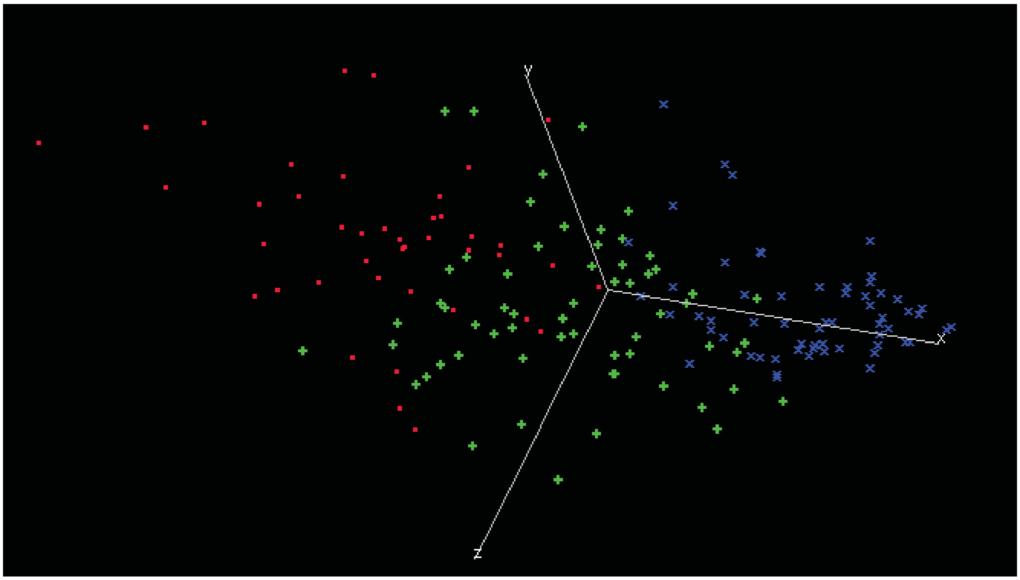
We performed a feature selection on this reduced matrix using two different methods. The most direct method is to difference the mean values for each parameter for the case group (i.e., the 43 CFS) and the control group (60 NI) and divide by the standard deviation based on the pooled variance of the two groups. This produces a type of Z-score indicating the signal to noise separation between the two means. The second method selected features using matrix linear algebra to compute how each parameter vector in a 'whitened' subject ID space ('people' space) projected onto vectors representing the case and control groups.

Calculations for whitening, vector projection, Z-scoring, and LDA were performed in Mathematica 5.1 (Wolfram Research, IL, USA). PCA and associated plots were done in JMP 5.1 (SAS Institute, NC, USA). All other data manipulations were performed using Excel.

Results & discussion

A naïve look at the original clinical data matrix (440 × 227) discussed in the methods section would produce a PCA that gives good separation between case and control groups. However, closer inspection reveals that the presence of the Short Form-36 (SF-36), multidimensional fatigue inventory (MFI), and Centers for Disease Control and Prevention (CDC) symptom inventory (SI) data in the dataset guarantees this will happen, given that the case classification rules are based on these questionnaire data [7]. Using data from the three questionnaires by themselves will separate cases from controls quite well, as can be seen in Figure 1. However, these are not the only questionnaires that are highly correlated with case classifications. Even the approximately 140 variables from the complete set of sleep questionnaires

Figure 1. Separation of CFS, ISF, and NF by rotation of the first three principle components of the combined SF-36, MFI and SI scores.



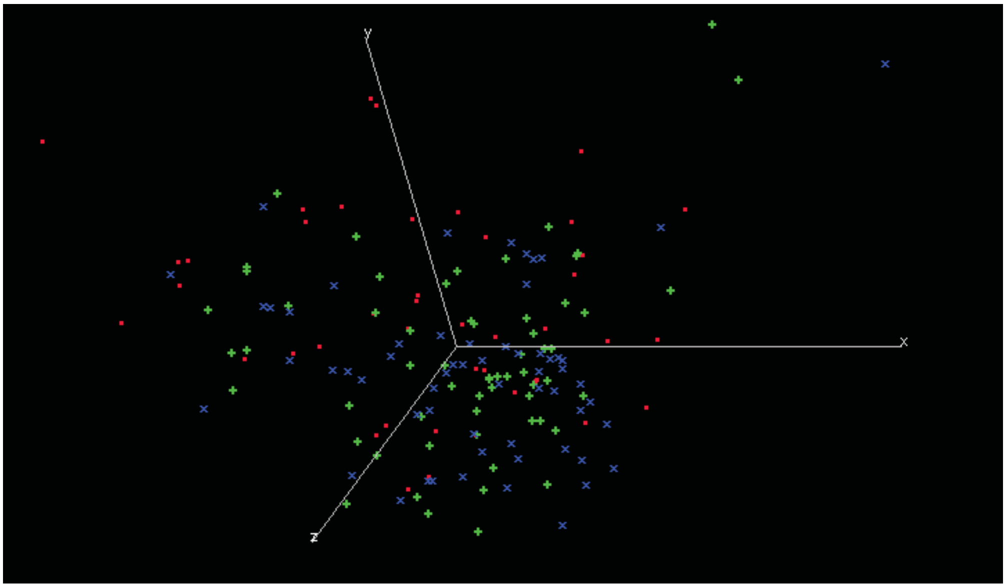
Red squares represent subjects with CFS, green crosses those with ISF, and blue letter x symbols are NF. CFS: Chronic fatigue syndrome; ISF: Insufficient fatigue; NF: Nonfatigued; MFI: Multidimensional fatigue inventory; SF-36: Short Form-36; SI: Symptom inventory.

– sleep inventories, SAQ's before and following sleep, and so on (these do not include physiological variables measured during sleep, which are part of the final biological data matrix) – can distinguish between disease status groups after a PCA quite well. However, if the questionnaire data is removed, PCA on the remaining variables becomes unrevealing, and the ability to cleanly separate cases from controls in a linear space appears to be reduced (Figure 2). While not discounting the validity of the self-reported symptoms that subjects were experiencing, the question remains whether there is anything in the biological data that provides an indication of disease status. The loss of separation between groups in Figure 2 demonstrates the need for feature selection in high-dimensional datasets.

We performed simple types of feature selection on the biological (reduced) dataset for the Wichita clinical data to demonstrate both the power and limitations of feature selection. The first technique was based on a data preprocessing technique known as 'whitening', as shown in Figure 3. Whitening is a matrix linear algebra technique which transforms data into a rotated space – much like PCA – for the purpose of making data columns, or rows, more independent/decorrelated and more similar in their distribution (hence, it is also referred to as 'sphering' the data [8]). This was desired so as to see more clearly the normalized impact of each parameter on the dataset. The reduced, 183 × 164 matrix of mostly biological variables was whitened in subject ID or 'people' space so that parameter definitions (i.e., features) could retain their meaning and not be rotated into an abstract parameter hyperspace. This also had the advantage of giving us a full set of eigenvalues to perform the transformation, since 183 points (parameters) in theory fill up a 164 dimensional space, whereas 164 points (subjects) will certainly be degenerate (flat) in a 183 dimensional space, leading to some zero eigenvalues. Vectors representing the average of the case and control vectors in the original space are then rotated into this white-space, and the projections (dot products) of each parameter onto the case and control averaged vectors are computed. Differences in the two dot products are then reported as a separation score.

Table 1 shows the top 20 parameters or features in the reduced dataset after the whitening-based feature selection is performed. The self-reported depression scale (SDS index), or Zung score, is the only questionnaire-based measure that was left in the reduced clinical matrix; this was done to demonstrate the high comorbidity of depression with CFS case classification. Three of these top 20 features are used to separate cases and controls in Figure 4. These three features provide a much better separation between groups in 3-dimensional (3D) space than the combined power of 183 different variables in a PCA, as shown in Figure 2. However, much of this separation is based on the inclusion

Figure 2. Separation of CFS, ISF and NF by rotation of the first three principle components of the remaining clinical data after the questionnaire data (except for Zung score) and medications were removed.



Red squares represent subjects with CFS, green crosses those with ISF, and blue letter x symbols are NF.
CFS: Chronic fatigue syndrome; ISF: Insufficient fatigue; NF: Nonfatigued.

of the Zung score in both Figure 2 and Figure 4. Without the Zung score (Figure 5), the boundary between groups appears more blurred, although it is hard to tell exactly in this simple 3D projection.

Table 2 shows the top 20 features selected by a simple Z-score computation, which provides results similar to a t-test, except that results are given in terms of a distance instead of a p-value. The Z-score is computed by differencing the means for cases and controls for a given variable, and then dividing by the pooled variance for that variable. Although they only share four features in common, Tables 1 and 2 have some similar themes. Both carry a lot of sleep related variables: five on

Figure 3. Computation of total separation between cases and controls for each parameter in a whitened 'people' space.

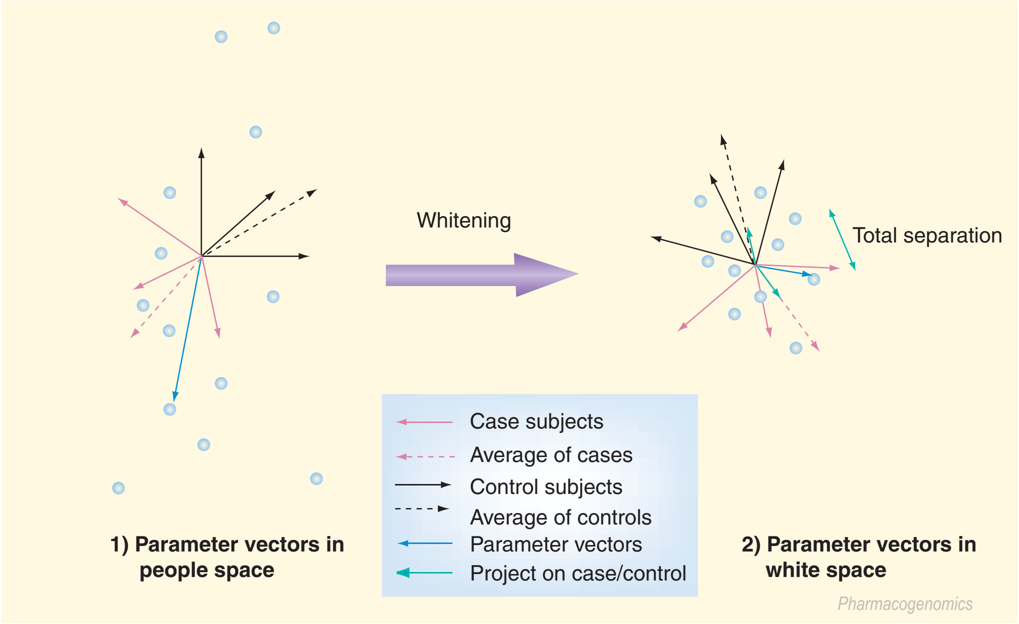


Table 1. Top 20 features of the reduced clinical matrix that separate CFS from NF controls in subject 'white space', using the vector projection method.

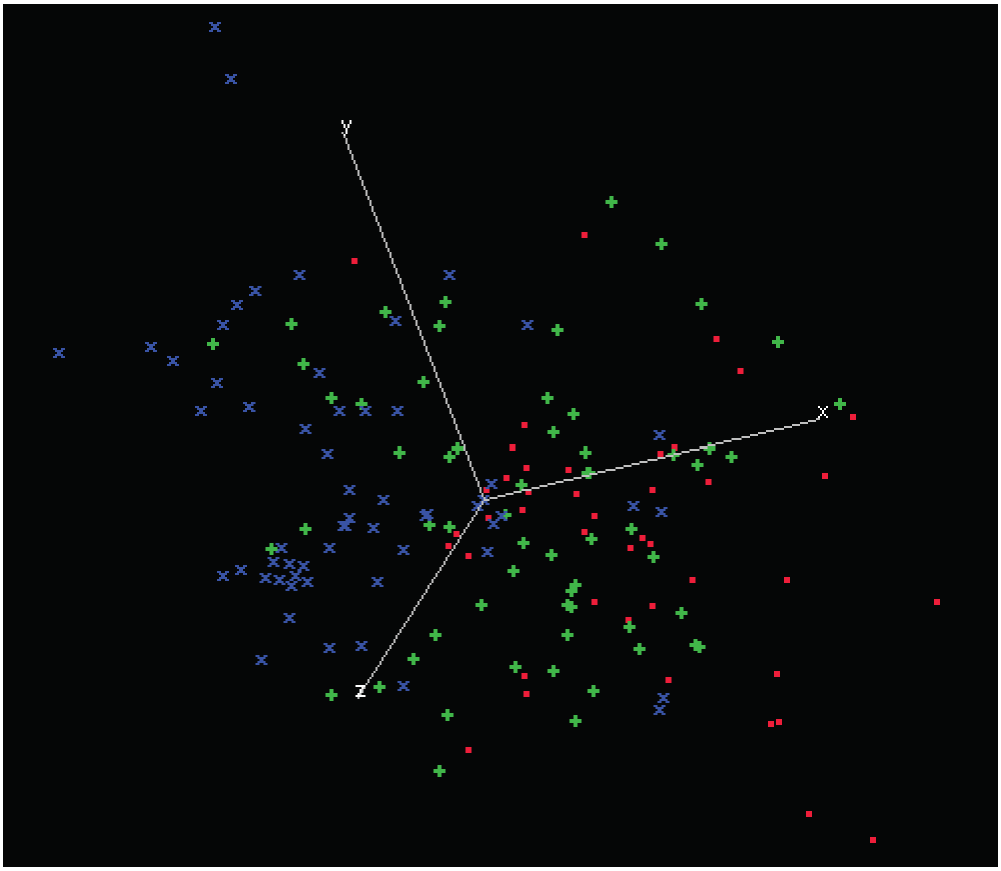
Clinical variable	Sep CFS–NF	Rank
SDS (Zung) index	0.626422	1
RR interval (heart rate)	0.298934	2
Hematocrit	0.238629	3
Number of basophils	0.233979	4
Waist:hip ratio	0.233042	5
Reaction five choice move time	0.228487	6
Specific gravity (urine)	0.221185	7
Glucose (serum)	0.212226	8
Reason index	0.212098	9
Total power (ECG)	0.210797	10
Progesterone	0.207702	11
Urine free cortisol (24hr)	0.200965	12
Triiodothyronine T3	0.199057	13
Spatial recognition memory mean correct latency	0.189989	14
Apnea-hypopnea total	0.189379	15
% basophils	0.184349	16
Aldosterone	0.181345	17
Flow limitation index	0.175478	18
Urine volume (24hr)	0.173496	19
Stage 2 total sleep time %	0.168036	20

The maximum possible separation is about 1.41.
CFS: Chronic fatigue syndrome; ECG: Electrocardiogram; NF: Nonfatigued; SDS: Self-reported depression scale.

each table, if one counts ECG (electrocardiogram) measures such as RR-interval, which were measured during the second night of sleep (to allow for acclamation). Sleep measures have ranks of 2, 10, 15, 18, and 20 and 5, 6, 15, 16, and 20 on Tables 1 and 2, respectively. After sleep variables, there are three allostatic load variables (refer to the companion paper by Maloney and colleagues in this issue [1]) at ranks 5, 12 and 17, and two CANTAB variables at ranks 6 and 14 on Table 1 (recent evidence has confirmed that cognitive functioning as measured by the CANTAB is impaired in CFS cases [7]). Both tables together contain three thyroid hormone-related variables. Table 2 contains several variables that are obviously correlated, for example, hematocrit, hemoglobin, mean corpuscular hemoglobin concentration, and red blood cell count, indicating that more could be done to reduce the redundancy of the clinical matrix.

Table 3 shows how allostatic load factors rank after whitening based feature selection and Z-score based feature selection are performed on the 183 × 164 matrix. Notice that the whitening, vector projection technique highlights the utility of the allostatic load factors, which are shown to be quite important in the accompanying articles by Maloney and colleagues and Goertzel and colleagues [1,10], much better than the Z-score does. Almost all of the 11 allostatic load factors used in Maloney and colleagues are in the top 50% of the biological features in the reduced clinical data matrix, and three are in the top 20, by the whitened vector projection calculation, vs only half in the top 50% and none in the top 20 by the Z-score calculation. Thus, if we believe the allostatic load results by Maloney and colleagues [1], a Z-score or t-test is not necessarily the best way to perform feature selection, as has also been noticed by other groups [11]. This might also be suspected by comparing Figure 5 to Figure 6, where PCA on the top 20 features (not including Zung) identified by Z-score does not separate the groups as definitively as a similar procedure on the top 20 whitening-based features. On the other hand, if one includes the total alloststic load index (ALI) score of Maloney and colleagues [1] in the 183 × 164 matrix and repeats the feature selection, ALI ranks higher than all of its components in Z-score-based feature selection, but in the middle of its components in the whitening-based

Figure 4. Separation of CFS (red), ISF (green), and NF (blue) after selection of Zung score (x-axis), RR interval (y-axis), and waist/hip ratio (z-axis) by feature selection.



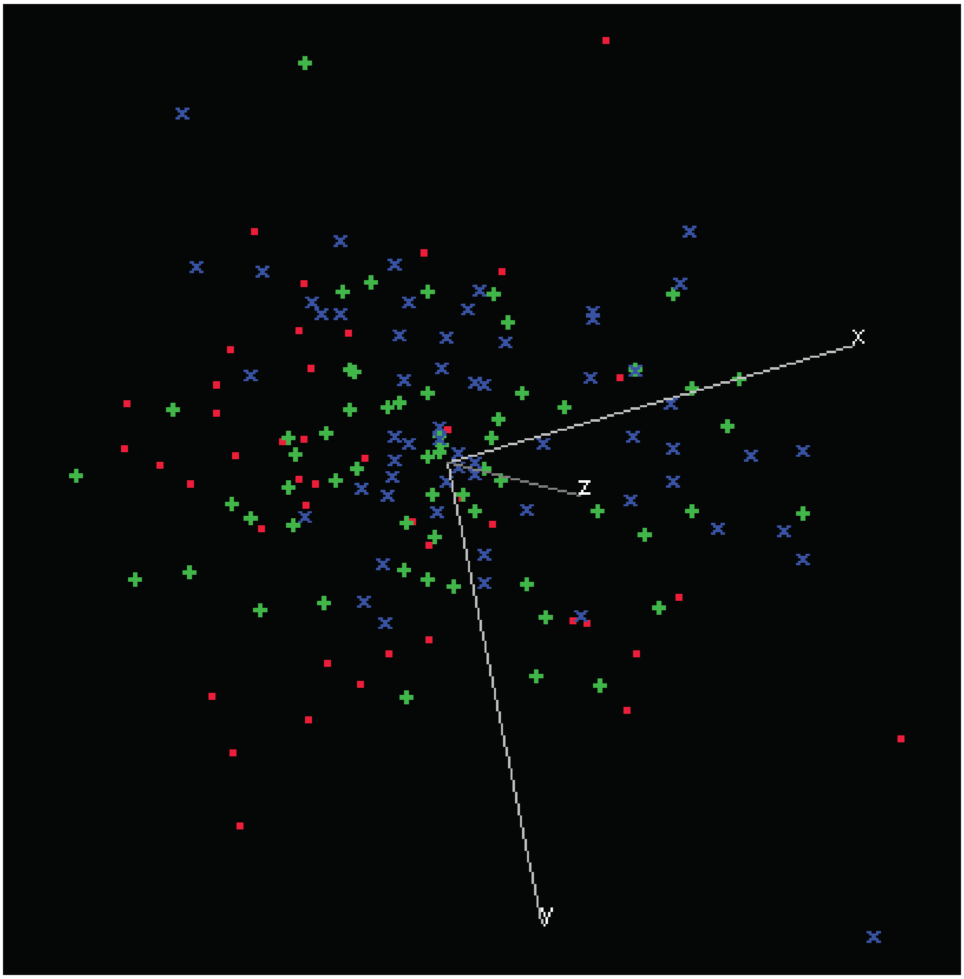
feature selection (last line of Table 3). However, both sets of 20 features perform equally well in a cross-validated LDA (Table 4).

Figures 5 and 6 show visually how groups might separate using feature-selected variables; however, such visualizations cannot demonstrate the true accuracy of the separation in higher dimensional space. The optimal linear boundary in hyperdimensional space can be computed by an LDA, the results of which are actually better than one might surmise looking at both figures.

The initial LDA results for both feature selection methods was promising: classification accuracies (i.e., the total percentage of correctly called cases and controls) varied from 84% using just the top two features in either Table 1 or Table 2 (i.e., the boundary is a line in 2D space – picture a line on the x–y plane of Figure 4) to over 99% when using the top 100 features. However, such accuracies are typical when one tests and trains on the same data. Furthermore, we used the Zung score that, as can be seen in Figure 4, is quite effective at separating cases from controls almost by itself. To avoid testing the training data, and to test the true classification power of the biological data, we performed a bootstrap of a five times cross-validation on the top features excluding Zung score. We randomly chose 80% of the subjects and the top n features of data to train on, leaving the remaining 20% for test, and repeated this 1000 times, gathering statistics on false negative rate (cases misclassified as controls), false positive rate, and total error. 1000 times was chosen for the bootstrap as it gave repeatable results. The results are shown in Table 4.

Table 4 shows the true errors one might expect from a linear boundary separating cases and controls for the 182 variables of the Wichita biological data (no Zung score). As has been noted by other authors, there is often an optimal number of features that is much smaller than the total number of features which could be used – the motivation for feature selection [12]. In our case, that number is approximately 20 using either feature selection method. This was not true in the un-cross-validated analysis, where it seemed that the more features the better. It appears that

Figure 5. Rotation of the first three principle components of the 20 most important features by the whitening, vector projection method (excluding Zung score) as shown in Table 1.



the top 20 biological features listed in Tables 1 and 2 provide us with an approximately 95% chance of achieving a classification accuracy of 57% or better. This is slightly better than by chance. One curious aspect of Table 4 is that both the feature sets in Tables 1 and 2 perform comparably in a linear classifier, even though the features themselves and the methods used to derive them are quite different. Is this an indication that feature selection doesn't work, and any set of features will do? To test for this, the bootstrap was repeated using randomly selected features out of the 182 × 164 matrix for each run; the results are shown at the bottom of Table 4. As expected, the linear classifier performed close to randomly (i.e., approximately 50% false positive and false negative rates) on average when only five features were used. Due to the large proportion of uninformative features in the matrix, a random pick of five out of 182 features is very likely to produce a random classifier. One would expect such a random feature-based classifier to improve slightly as the number of features is increased, because the chances of randomly selecting good features would improve, and it does (unlike the feature-selected features, where one reaches diminishing returns at about 20 features). The data shows that feature selection is working; we can only conclude that multiple feature sets can separate the dataset using a linear boundary with similar accuracy. Perhaps these results also imply that we have not yet found the optimal set of features for a linear classifier.

Outlook

The results of this study clearly have some implications for further research into the pathogenesis of CFS, and should be pursued. As for pursuing

Table 2. Top 20 features of the reduced clinical matrix that separate CFS from NF controls by Z-score.

Clinical variable	Z-score	Rank
SDS (Zung) index	2.10262	1
Specific gravity (urine)	1.438295	2
Basal temp	1.153255	3
Mean corpuscular hemoglobin	0.978826	4
RR interval (heart rate)	0.963796	5
Epworth	0.785066	6
Chloride	0.721714	7
Red blood cell count	0.654418	8
Sodium	0.60304	9
Hematocrit	0.585339	10
Mucus index (urine)	0.582845	11
Hemoglobin	0.570832	12
Free thyroxine (t4)	0.555562	13
Potassium	0.482336	14
SD NN interval (ECG)	0.480514	15
Paradoxical breathing index	0.469425	16
Reverse t3	0.466955	17
IL-1b (cytokine)	0.466896	18
Bacteria index (urine)	0.461609	19
Hypopnea	0.441044	20

Scores can be thought of as Gaussian equivalent standard deviations. Notice that only Zung is significant by itself.
CFS: Chronic fatigue syndrome; ECG: Electrocardiogram; IL-1b:: Interleukin-1b; NF: Nonfatigued; SD: Standard deviation.

this particular dataset and these data types any further for biomarker validation, i.e., purely to discover combinations of the parameters presented here that separate CFS case from control with good sensitivity and specificity, it remains to be seen whether it is a worthwhile endeavor. Although many more sophisticated machine learning algorithms could be thrown at the data than were presented here, initial tests using the Biomind software [13], which employs genetic algorithms and support vector machines (SVMs) among other techniques, on some of the feature selected variables highlighted here were not promising. The truism to keep in mind is that “good data always trumps good math” ([Rob Grothe. PERS. COMMUN.] regarding a different type of data [14,15] but still relevant). Other data types, as is shown with the companion articles on SNPs and allostatic load, may provide the key to biomarker discovery for CFS.

Acknowledgements

We would like to thank Eric Aslakson for supplying the initial scrubbed and ‘munged’ version of the 440 parameter by 227 subject Wichita clinical data matrix.

Disclaimer

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

Table 3. Allostatic load variables by separation scores and rank for two different feature selection methods.

Allostatic load variable	Separation	Rank	Z-score	Rank
SDS (Zung) index	0.63	1	2.10	1
RR interval (heart rate)	0.30	2	0.96	5
Waist:hip ratio	0.23	5	0.23	87
Urine free cortisol (24hr)	0.20	12	0.28	68
Aldosterone	0.18	17	0.38	34
DHEA sulfate	0.14	35	0.15	117
IL-6 (cytokine)	0.11	52	0.29	63
Epinephrine	0.096	67	0.21	93
BMI	0.090	72	0.053	162
Diastolic bp (recum 30min)	0.075	87	0.17	112
High sensitivity CRP	0.068	93	0.13	126
Albumin	0.067	95	0.37	39
Norepinephrine	0.032	141	0.15	120
Systolic bp (recum 30min)	0.019	161	0.11	138
Allostatic load index	0.079	82	0.46	18

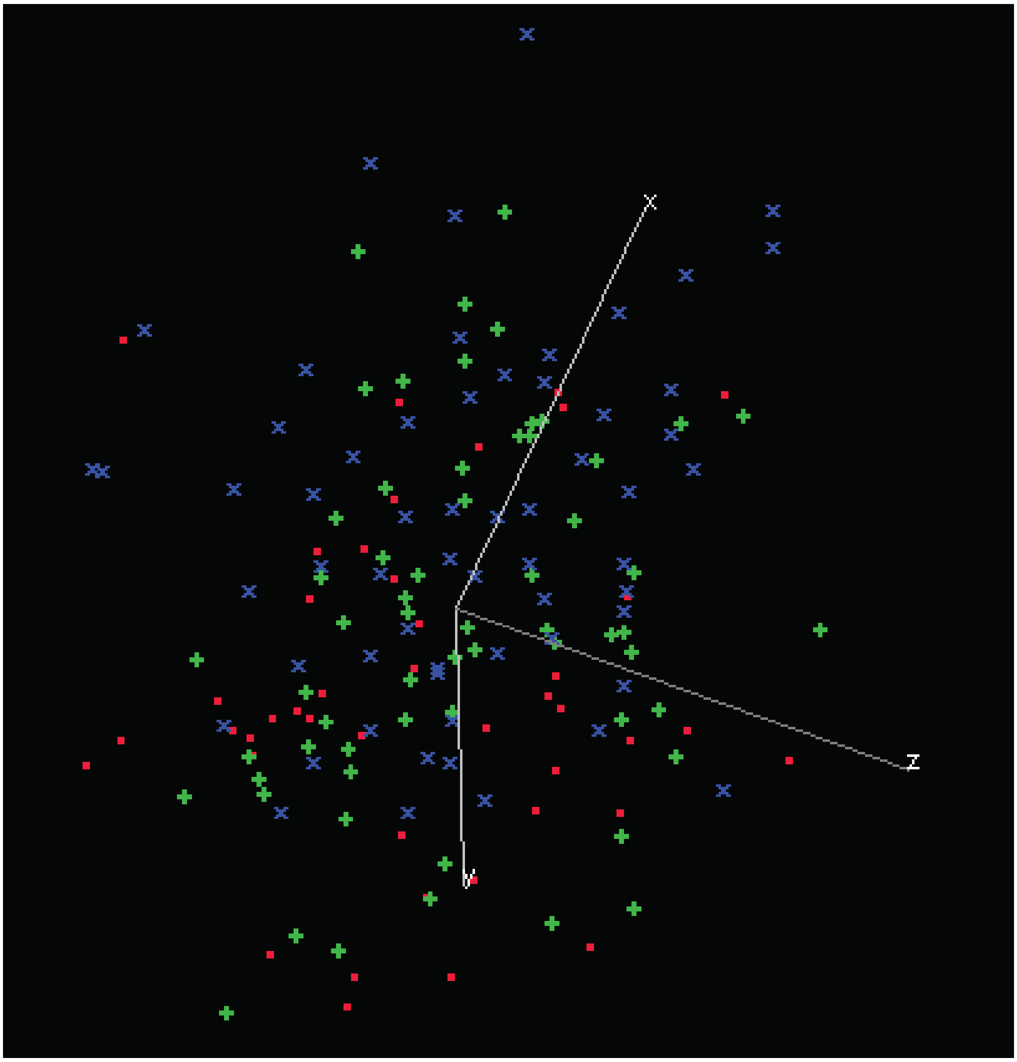
The first score shows the normalized separation of case and control vectors in subject white-space; the second shows Z-score separation based on pooled variance. The first two variables (Zung and RR interval) and BMI are shown for reference and are not part of the allostatic load index used in Maloney and colleagues [1]. *Ranks are out of 183 total variables studied. Scores and ranks for the allostatic load index itself were computed in separate runs.*
BMI: Body mass index; CRP: C-reactive protein; DHEA: Dehydroepiandrosterone; IL: Interleukin; SDS: Self-reported depression scale.

Table 4. False negatives*, false positives and total error‡ for a 1000× bootstrap of a 5× cross-validation study§ using linear discriminant analysis.

Top number features	FN %	1σ	FP %	1σ	Error %	1σ
Whitening						
5	33	19	31	15	31	10
20	27	18	20	14	23	10
100	35	19	26	14	30	11
Z-score						
5	31	18	24	14	27	10
20	27	18	19	13	22	10
100	38	20	31	16	34	11
Random						
5	50	22	41	18	45	13
20	45	21	36	16	40	12
100	42	21	35	16	39	12

**Cases misclassified as controls; ‡Total % of incorrectly classified cases and controls.*
§Train on 80% and test 20% of the data, repeated 1000 times.
Standard deviations for each metric are shown to the right. The top N features by both feature selection methods were used in the LDA. Notice that 20 seems to be the optimal number of features for LDA on this dataset. LDA for N randomly selected features is shown for comparison.
FN: False negative; FP: False positive; LDA: Linear discriminant analysis.

Figure 6. Rotation of the first three principle components of the 20 most important features (excluding Zung score) by the Z-score method.



Highlights

- Chronic fatigue syndrome (CFS) appears to be not just a matter of self-assessed variables, as important as they are, but there appear to be real physiological differences between CFS cases and age, race, sex, and body mass index (BMI)-matched controls, although most of these differences do not appear to be individually statistically significant .
- Some of the physiological findings in this study, that are not addressed in Maloney and colleagues [1] as part of the allostatic load hypothesis, need to be pursued further (for example, implicated sleep and electrocardiogram variables).
- Feature selection is a critical component of high-dimensional data mining on diverse datasets such as the Wichita (KS, USA) clinical dataset.
- Linear matrix algebra techniques, such as principal component analysis and linear discriminant analysis, can only go so far on this dataset. On the other hand, more sophisticated machine learning techniques might improve classification accuracy only slightly, and further obscure biological interpretation of the data. Hypothesis-driven approaches based on a medical and/or molecular understanding of the problem, such as was done in the companion articles on SNPs and allostatic load in this issue [1,2], might provide a clearer result and make more biological sense.

Bibliography

1. 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).
2. GoertzelBN, Pennachin C, Coelho LdS, Gurbaxani BM, Maloney EM, Jones JF: Combinations of single nucleotide polymorphisms in neuroendocrine effector and receptor genes predict chronic fatigue syndrome. *Pharmacogenomics* 7(3), 475–483 (2006).
3. Dougherty ER: Feature-selection overfitting with small-sample classifier design. *IEEE Intelligent Systems* 20(6), 64–66 (2005).
4. Hastie T, Tibshiriani R, Friedman J: *The Elements of Statistical Learning*. Springer, New York, NY, USA, 371–406 (2001).
5. Hastie T, Tibshiriani R, Friedman J: *The Elements of Statistical Learning*. Springer, New York, NY, USA, 485–491 (2001).
6. Liu H: Evolving feature selection. *IEEE Intelligent Systems* 20(6), 64 (2005).
7. 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).
8. Hastie T, Tibshiriani R, Friedman J: *The Elements of Statistical Learning*. Springer, New York, NY, USA, 91 (2001).
9. Capuron L, Welberg L, Heim C *et al.*: Cognitive dysfunction relates to subjective report of mental fatigue in patients with chronic fatigue syndrome. *Neuropsychopharmacology* [Epub ahead of print] (2006).
10. Goertzel BN, Pennachin C, Coelho LdS, Maloney EM, Jones JF, Gurbaxani BM: Allostatic load is associated with symptoms in chronic fatigue syndrome patients. *Pharmacogenomics* 7(3), 485–494 (2006).
11. Berens M, Liu H, Yu L: Fostering biological relevance in feature selection for microarray data. *IEEE Intelligent Systems* 20(6), 71–73 (2005).
12. Forman G: Feature selection: we've barely scratched the surface. *IEEE Intelligent Systems* 20(6), 74–76 (2005).
13. Smigrodzki R, Goertzel B, Pennachin C, Coelho L, Prosdocimi F, Parker WD: Genetic algorithm for analysis of mutations in Parkinson's disease. *Artif. Intell. Med.* 35(3), 227–241 (2005).
14. Diaz-Avalos R, Long C, Fontano E *et al.*: Cross-β order and diversity in nanocrystals of an amyloid-forming peptide. *J. Mol. Bio.* 330(5), 1165–1175 (2003).
15. Nelson R, Sawaya MR, Balbirnie M *et al.*: Structure of the cross-β spine of amyloid-like fibrils. *Nature* 435(7043), 773–778 (2005).