Classifying ^{31}P NMR Phospholipid Profiles from Postmortem Schizophrenic Brain: Multivariate Model Selection and Cross-Validation

J. A. Welge[1,2], and R. A. Komoroski[2]

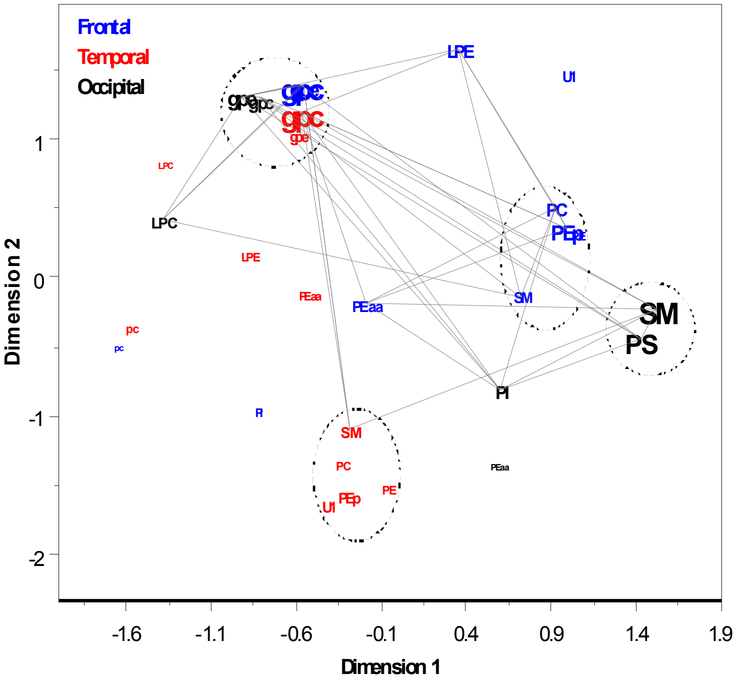
[1]Environmental Health, University of Cincinnati, Cincinnati, OH, United States, [2]Center for Imaging Research, University of Cincinnati, Cincinnati, OH, United States

Introduction: There is considerable evidence that cell membrane abnormalities associated with changes in phospholipid (PL) composition are implicated in schizophrenia pathogenesis. Using ^{31}P NMR spectroscopy we have compared the compositions of PLs and PL metabolites in postmortem brains from schizophrenics and matched controls (1,2). Univariate statistical analyses of individual PLs or metabolites revealed few differences between the groups (1,2). Here we report a multivariate approach for analysis of NMR data and apply it to our previous ^{31}P NMR results for schizophrenia.

Methods: Tissue concentrations of 4 aqueous PL metabolites (1) and 12 PLs (2) in 3 cortical brain regions (frontal, temporal, occipital) of 20 schizophrenic and 20 control postmortem left hemispheres were taken from previous work. Because the number of measurements [16 PLs (plus metabolites) x 3 brain regions = 48] is large relative to the number of samples (20 schizophrenics and 20 controls), it was not feasible to fit a full regression model, and the number of possible reduced models is massive. We applied model selection techniques based on information criteria to identify a set of regression models that optimally classified a randomly selected subset of the samples. We simultaneously performed a validation step on the remaining samples, and iterated this process to approximate the expected performance of the candidate models at classifying new cases.The branch and bound algorithm (3) as implemented in SAS's PROC LOGISTIC was used to identify the best initial 3000 models from all possible *k*-predictor models (500 models per model size for *k*=2 to 7). Because models with more predictors will generally provide better fit to the sample but may not generalize to new data, Akaike's Information Criterion [AIC] (4) was used to compare models of different sizes. AIC is a penalized likelihood measure that adjusts the usual log-likelihood for a model by a penalty term based on the number of parameters used, achieving a balance between goodness-of-fit and parsimony that is theoretically optimal for prediction in terms of Kullbeck-Liebler divergence (5).

Metabolite[a]	%
Occip SM	78.30
Occip PS	72.76
Frontal gpc	43.44
Temp gpc	42.34
Occip PI	30.74
Occip gpe	28.58
Occip gpc	23.82
Frontal PC	18.68
Frontal PEp	18.37
Frontal gpe	18.00
Occip LPC	17.08
Frontal LPE	16.90
Frontal PEaa	12.30
Frontal SM	10.83
Temp SM	9.39
Frontal U1	9.03
Occip PEaa	8.39
Frontal PE	8.25
Temp gpe	7.77
Temp PEp	7.08
Temp PE	5.99
Temp U1	5.29
Frontal PI	4.41
Temp PEaa	4.22
Temp LPE	4.02
Temp LPC	3.43
Frontal pc	3.11
Temp pc	3.01
Occip pc	2.92

[a]See refs. 1 & 2 for PL and metabolite identification



Results: The table shows the frequency of appearance of each metabolite in a focus set of models with differences from the AIC of the best model of less than five (this set of models has posterior probability no less than 1/10 the probability of the best model) for each random build-test split. The pattern of joint appearance of these metabolites in this set of models is illustrated in the figure, where proximity of individual metabolites is based on a two-dimensional scaling of the correlation matrix, symbol size is proportional to overall appearance frequency, and lines connect frequently co-occurring metabolites. Several clusters of highly correlated metabolites are apparent (gpc/gpe across all brain regions, and region-specific clusters composed of SM and PC), and the best predictive models with few exceptions contained at least one metabolite from each of these clusters. The model that appeared most frequently in the single most common model in the focus set contained temporal gpc and occipital SM, PI and PS, with an estimated out-of-sample classification accuracy of 79%. The model that was most frequently ranked as the single best contained these same metabolites plus frontal gpc and occipital gpe, and achieved an overall mean classification rate of 80%.

Discussion: The model selection procedure presented here identified a set of models that achieved classification accuracy well in excess of chance rates when tested on subsets of the data that had not been used to select the model. The combination of selection by AIC (which is designed to avoid overfitting to sample data) and internal cross-validation by random splitting (to produce reasonable estimates of predictive power in a new sample) allows exploration of the vast set of possible multivariate models that arise from data structures where the number of measurements approaches the number of samples (which is typical in metabolomic applications). Without use of such steps to control overfitting, it is easy to find superficially impressive results even when the data lack any true structure (6).

The present results suggest that occipital SM, PS, and PI, and frontal and temporal gpc, are important metabolites for our classification of schizophrenics and controls. Occipital SM (but not PS or PI) and frontal and temporal gpc were also suggested by univariate analysis (1,2). Our multivariate approach has provided a somewhat different perspective on the PL and PL metabolite differences observed between schizophrenic and control brains.

References: 1. Komoroski RA, Pearce JM, Mrak RE. Magn Reson Med 2008;59:469-474. **2.** Pearce JM, Komoroski RA, Mrak RE. Magn Reson Med 2009;61:28-34. **3.** Furnival GM, Wilson RW. Technometrics 1974;16:499-511. **4.** Akaiki H. IEEE Transactions on Automatic Control 1974;19:716-723. **5.** Burnham KP, Anderson DR. *Model Selection and Multi-Model Inference 2nd Ed.*, Springer: New York, 2000. **6.** Broadhurst DI, Kell DB. Metabolomics 2006;2:171-196.