1832

Statistical Evaluation of Connectome Data using Object Oriented Data Analysis

Linda Larson-Prior[1,2], Patricio LaRosa[3,4], Terrence Brooks[3], Elena Deych[3], Berkley Shands[3], Fred Prior[1], William Shannon[3]

[1]Mallinckrodt Institute of Radiology, Washington University in St. Louis MO; [2]Department of Neurology, Washington University School of Medicine in St. Louis MO, [3]Department of Medicine, Washington University School of Medicine, [4]Predictive Analytic Research Gp, Monsanto, St. Louis MO

MIR Mallinckrodt Institute of Radiology
Washington University in St. Louis School of Medicine

INTRODUCTION

Complex biological systems may be modeled and analyzed as nested hierarchies of networks[1]. Complex disorders such as schizophrenia[2] may be the result of failures of functions at various levels of nested hierarchies of biomolecular and cellular networks. The Human Connectome project[3] and similar large-scale brain connectivity studies will generate a large, normative database of connectivity graphs. Future research using this normative data will depend on advanced statistical comparison techniques to enable new understanding of neuropsychiatric disorders[4].

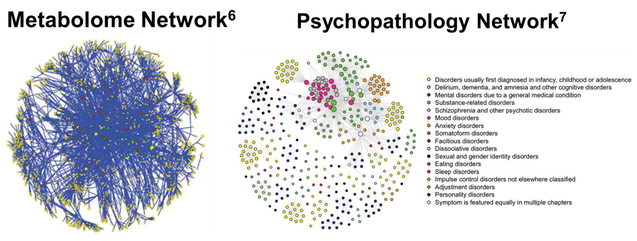
While in its infancy, a major thrust of the new field of connectomics lies in the identification of "typical" features identifying what have been termed connectopathies or pathconnectomics in clinically defined populations. A secondary goal lies in identifying subgraphs showing significant differences of potential clinical relevance. To meet these goals, an important corollary to the need for statistical methods for analyzing group data is the need to include important covariate information that may be provided by neuropsychological, genetic and/or biochemical data.

The development of statistical tools by which such large and complex data sets can be compared is essential to the eventual success of connectomics in providing new insights into human cognitive function. This paper presents a novel parametric statistical method from the emerging field of object oriented data analysis that applies the Gibb's distribution to sets of connectome graphs.

METHODOLOGY

The **Gibb's distribution** models populations of graphs, and it has been shown that it is a general framework that can describe all Markov networks and Bayesian networks conditions on evidence[5].

Figure 1. Any graph type can be analyzed



The distribution models graphs, including connectome graphs using two parameters:

- g* representing the central graph of the observed data
- τ representing the spread of the observed graphs

Let G be the finite set of graphs with elements *g* and *d*: G X G → $\mathbb{R}^+$ is an arbitrary distance metric on G. For a binary connectome graph, the Hamming metric $d(g_k, g_l)$ - which is the number of edge differences between two graphs - is used as the distance metric. The Gibb's distribution on G is defined by:

$$\mathbb{P}(\mathrm{g};\ \mathrm{g}^*, \tau) = c(\mathrm{g}^*, \tau)\exp\left(-\tau d(\mathrm{g}^*, \mathrm{g})\right),\ \forall \mathrm{g} \in \mathrm{G}, \quad (1)$$

A Maximum Likelihood estimate (MLE) approach is used to estimate (g*, τ) from a set of *N* connectome graphs. For binary connection graphs where *R* is the total number of possible edges in a graph *G*, the log-likelihood is given by:

$$\ln L(\mathrm{g}^*, \tau; \mathrm{G}_N) = -N\ln[1+\exp(-\tau)]^{-R} - \tau\sum_{i=1}^{N} d(\mathrm{g}_i, \mathrm{g}^*) \quad (2)$$

We want to find the MLE parameters $\mathrm{g}^* = \widehat{\mathrm{g}^*} \in \mathrm{G}$ and $\tau = \hat{\tau} \geq 0$ that maximize Eq. 2, which is obtained by computing the derivative of $\ln L(\mathrm{g}^*, \tau; \mathrm{G}_N)$ with respect to τ, $\frac{\mathrm{d}\ln L(\mathrm{g}^*,\tau;\mathrm{G}_N)}{\mathrm{d}\tau}$, and finding τ such that $\frac{\mathrm{d}\ln L(\mathrm{g}^*,\tau;\mathrm{G}_N)}{\mathrm{d}\tau} = 0$. This results in the following expression of $\hat{\tau}$:

$$\hat{\tau} = -\log\frac{(NR)^{-1}\sum_{i=1}^{N} d(\mathrm{g}_i,\ \mathrm{g}^*)}{1-(NR)^{-1}\sum_{i=1}^{N} d(\mathrm{g}_i, \mathrm{g}^*)} \quad (3)$$

The MLE of g* is a graph such that the likelihood is maximized at the solution of :

$$\widehat{\mathrm{g}^*} = \arg\max_{\mathrm{g}^*\in\mathrm{G}}\left\{-N[1+\exp(-\tau)]^{-R} - \tau\sum_{i=1}^{N} d(\mathrm{g}_i, \mathrm{g}^*)\right\} \quad (4)$$

To determine whether a set of connectomes comes from one or more distributions, a **finite mixture model** of Gibb's distributions is used.

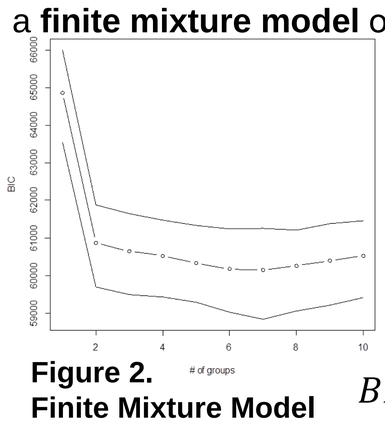


Figure 2. Finite Mixture Model

$$\mathbb{P}(\mathrm{g};\ \{\mathrm{g}_m^*, \tau_m, w_m\}_{m=1}^{M}) = \sum_{m=1}^{M} w_m c(\mathrm{g}_m^*, \tau_m)\exp\left(-\tau_m d(\mathrm{g}_m^*, \mathrm{g})\right) \quad (5)$$

To estimate the optimal number of Gibb's distributions *M*, the **Bayesian Information Criterion** (BIC) is used.

$$BIC_M = -2\ln L(\{\mathrm{g}_m^*, \tau_m, w_m\}_M) + M(2+R)\log(N) \quad (6)$$

Regression and Hypothesis Testing

We next develop a linear regression model for connectome data. Consider the problem of modeling a set of connectome graphs (g_j, j=1, …, N) as a function of some patient characteristic coded as X, (e.g. Xj = 0 for females and Xj = 1 for males). The parameters to be estimated are b_0, b_1 and τ, while $\oplus$ is the binary sum (XOR). The regression coefficients b_0 and b_1 are themselves connectomes and the predicted b_1 represents those connections that differ between the two groups of interest. The model is given as:

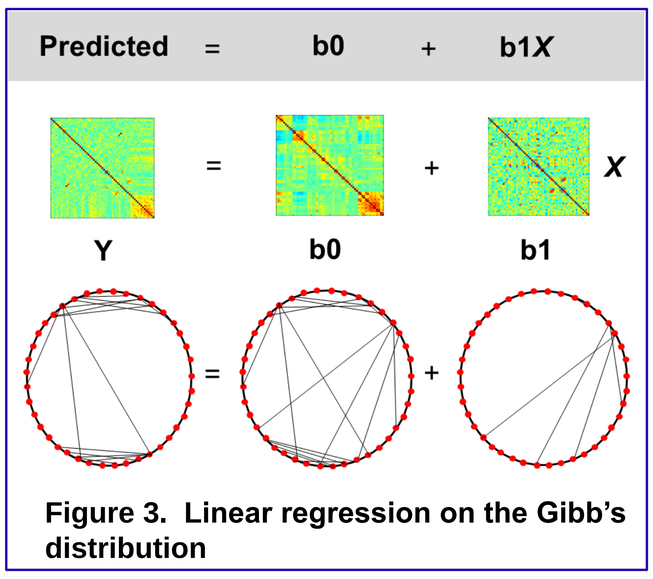


Figure 3. Linear regression on the Gibb's distribution

$$(\mathrm{g}_j|X_j \sim Gibbs(\mathrm{g}_j^* = \mathrm{b}_0 \oplus \mathrm{b}_1 X_j, \tau);\ j = 1, \ldots, N,) \quad (6)$$

A MLE approach is used to return estimates for τ, b_0 and b_1 for binarized graphs where $\hat{b}_0$ is the central graph of the group where X_j=0 and $\hat{b}_1$ is the difference between the central graphs of the two groups (Fig. 3 bottom panel).

Hypothesis Testing

To test the significance of $\hat{b}_1$, we test the hypothesis: H_0:$b_1 = 0$, H_A:$b_1 \neq 0$ where the null hypothesis H_0 is that there are no differences in edges between the two groups. This hypothesis is tested using a generalized likelihood ratio test (GLRT) defined as:

$$\lambda = 2\log\frac{\max_{H_A} \mathrm{L}(\{\mathrm{g}_j|X_j\}; \hat{\mathrm{b}}_0, \hat{\mathrm{b}}_1, \tau)}{\max_{H_0} \mathrm{L}(\{\mathrm{g}_j|X_j\}; \hat{\mathrm{b}}_0, \hat{\mathrm{b}}_1 = 0, \tau)} \quad (7)$$

A non-parametric permutation method was used to compute the distribution of the GLRT and obtain a p-value. Random sampling was used to define two groups, a regression model was fitted to each sample, and GLRT calculated to represent the distribution of this test statistic under the null hypothesis.

Application to Connectome Data

A 52-node connectivity matrix was constructed in 20 subjects participating in a motor learning task. Six conditions were available: pre-task rest, task, post-task rest on each of two successive days with a full night of sleep intervening. The task was regressed from the data to provide a "resting state" connection matrix. Matrices were binarized and analyzed as described. g* calculated on these data were used to test the regression model with sample sizes N=10,25,50, and 100 (Fig 4). Table 1 (Fig 4) shows that the GLRT statistic correctly rejected H_0 for a sample size > 7 based on 100 Monte Carlo simulations. When computed on the actual data sets (pre-task and during-task), the null hypothesis was rejected (P=0.0264) for the 20 subject data set The null hypothesis. could not be rejected for comparisons of pre-task to post-task (P = 0.474 (day 1) and 0.409 (day 2)).

Figure 4.

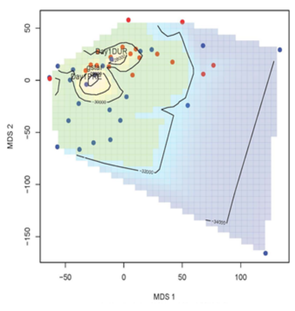


Table 1: Results of simulation studies to test the algorithms convergence rates on the regression parameters and the power to correctly reject the null hypothesis.

	Data (51 Nodes)	
	Day 1 Pre-Task vs During Task	
N	Hamming Distance Mean (Min, Max); MC=1000	Power to correctly reject the Null MC=100
5	63.7 (41, 91)	0%
6	25.9 (12, 43)	68%
7	28.9 (13, 49)	>99 %
8	12.1 (4, 24)	>99 %
9	13.2 (3, 24)	>99 %
10	5.7 (0, 15)	>99 %
15	1.3 (0, 7)	>99 %
20	0.1 (0, 3)	>99 %
25	0.033 (0, 2)	>99 %
50	0 (0, 0)	>99 %
100	0 (0, 0)	>99 %

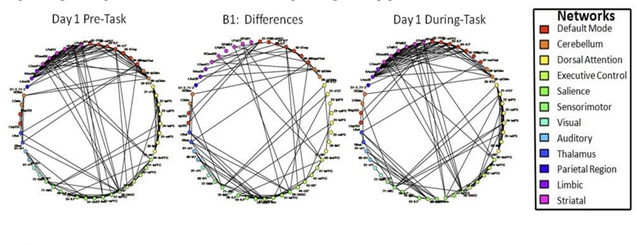


Figure 5. The "slope' connectome coefficient b1 is interpreted as the change in connectome connectivity between the groups. This edge 'switch' avoids represents connectivity differences between groups[8]

Conclusions

- A single P-value is calculated for connectome differences, avoiding massive univariate testing with multiple comparison adjustments
- The regression model can include multiple covariates of interest
- Edge differences are detailed in the b_1 connection graph

Software is available open source: http://cran.r-project.org/web/packages/bingat/index.html

REFERENCES

1. Alm E. and Arkin A.P. (2003), "Biological networks", Current Opinion in Structural Biology, vol 13, pp 193-202.
2. Lynall M-E., Bassett D.S., Kerwin R., McKenna P.J., Kitzbichler M., et al. (2010), "Functional connectivity and brain networks in schizophrenia", Journal Neuroscience, vol 30, pp 9477-87.
3. Van Essen D.C. and Ugurbil K. (2012), "The future of the human connectome", NeuroImage, vol 62, pp 1299-310.
4. Martin G. (2012), "Network analysis and the connectopathies: current research and future approaches", Nonlinear dynamics, psychology, and life sciences, vol 16, pp 79-90.
5. Koller D. and Friedman N. (2009) Probabilistic graphical models, principles and techniques. MIT Press.
   Patil, K.R. and Pers, T (2007), "Systems biology: looking beyond the genome", Forside biozoom, vol 514, e2285
6. Borsboom D., Cramer O.J., Schmittmann V.D., Epskamp S., Waldorp L.J. (2011), "The small world of psychpathology". PLoSone, vol 6, e27407.
7. Larson-Prior L.J. unpublished data
8. LaRosa P, Brooks T, Deych E, Shands B, Prior F, Larson-Prior, L, Shannon W, Gibbs distribution for statistical analysis of graphical data. Statistics in Medicine, in revision

Acknowlegements: Supported in part by NIH-R21MH1098223 (Shannon, PI)