Algorithmic analysis of functional pathways affected by typical and atypical antipsychotics

Arsen Arakelyan, Anna Boyajian, Levon Aslanyan, David Muradian, and Hasmik Sahakyan
"Laboratory of Information Biology" Project of the Institute of Molecular Biology and Institute for Informatics and Automation Problems
National Academy of Sciences of Republic of Armenia
Yerevan, Armenia
e-mail: aarakelyan@sci.am

ABSTRACT

The advantages of atypical vs. typical neuroleptics have been demonstrated in a number of clinical trials. Differences in functional pathways affected by typical and atypical antipsychotics in the brain have been assessed using Gene Set Enrichment Analysis. Data on gene expression, obtained from Gene Expression Omnibus, is a numerical array of size ~4x17000, which can be treated directly neither by statistical approach nor by means of classification and pattern recognition regular theory. Extended logic-combinatorial scheme is designed for the treatment of this kind data set. Applied results show that atypical neuroleptics have less effect on pathways related to neurodegeneration, cognition, neuronal architectonics, as well as stimulation of inflammatory processes.

Keywords

Pattern Recognition, Functional Pathway, Gene Set Enrichment Analysis, Neuroleptics

1. INTRODUCTION

The functional pathway analysis affected by second-generation atypical antipsychotics (atypical neuroleptics; AN) over those from the first generation (typical neuroleptics; TN) become a hot research topic. Promising studies suggest that atypical antipsychotics have less pronounced extrapyramidal, anticholinergic, parkinsonian and dystonic side effects [3-5]. However, more detailed studies are needed for complete assessment of preponderance over the TN. Since the frontal cortex is one of the most important regions for antipsychotics action, in this study we compare the effects of treatment by typical and atypical neuroleptics on functional pathways in frontal cortex of the brain. The "Typical and atypical antipsychotic drugs effect on brain" dataset $\Im$ of Gene Expression Omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS775) has been used. This dataset contains gene expressions from frontal cortex of 13-week male mice treated for 28 days with antipsychotics. Chlorpromazine and thioridazine were used as typical antipsychotics, and olanzapine and quetiapine as atypical antipsychotics. Gene expression profiles in GEO dataset were obtained using Agilent 011978 Mouse Microarray G4121A (GEO Platform ID: GPL891, Agilent Inc, USA).

Generally the basic approach of diverse type of analysis of multidimensional experimental data sets is mathematical statistics (MS). Having satisfactory amount of experimental data (statistics) it helps to form conclusions that some properties and postulations take place in some probabilistic level. Simple correlation, regression and hypothesis estimation algorithms are components of the statistical approach.

A different situation appears in area of pattern recognition (PR). There is no satisfactory statistics in this case. These heuristics are more responsible and conditional. Learning set is given as a limited number of known classifications but it has to be large enough to describe the class properties in application area. A number of basic approaches are known in PR - Metric Algorithms, Logic Separation (LS), Neural Networks, etc. One of the well-known classes of metric algorithms is the voting (or estimation calculation) model [1]. This is an algorithmic model with a number of additional parameters, requiring optimization during the learning stage.

The dataset of gene expression, being considered, structurally obeys neither statistical requirements nor – of pattern recognition. Two classes are given, two learning examples in each. Instead, the features set is very large. All these raise a novel very specific situation for data analysis, when it is necessary to recover the limited and valuable knowledge contained in such structures. ANOVA methods are the typical tool being proposed to determine the gene sets that are diffrentially expressed over different experimental conditions. However, only a few studies have been concerned with the use of ANOVA when the number of genes is large and the number of observations is small. The strong normality, and independence assumptions, that traditional ANOVA imposes, makes it impractical and not powerful enough. Several improvements and alternative approaches were developed [8]. Biclustering or simultaneous clustering [9], where both genes and conditions is challenging particularly to find subgroups of genes and subgroups of conditions where the genes exhibit highly correlated activities over a range of conditions. Next to mention is the branch of pattern recognition name Logical Combinatorial Pattern Recognition [1], which works effectively with nonstandard classification problems. We design and extend logic-combinatorial scheme to overcome the difficulty raised by our practical problem. Elementary classifiers, cluster analysis, testing and greedy solvers are considered and applied.

2. ALGORITHM

Pattern recognition deals with classes given by limited sets of classified examples and possibly by some hypotheses of the classes themselves. The main goal is to find an algorithm-classifier which extends the known classification to the area of unclassified objects. Formally, conditions of correct classification of all objects might be composed and then the problem of maximization of number of satisfied conditions appears. For linear hyperplane classifiers for example we receive systems of linear inequalities, - unnecessarily compatible in general. The question is in determining the maximal compatible subset of such systems, which is computationally a known NP hard problem. The situation with classes of typical and atypical antipsychotics given by data $\Im$ is relatively different. $\Im$, containing data on gene expression, is a numerical array of four ~1700 long numerical rows $\Im = \{S_1, S_2, S_3, S_4\}$. Classes consist of two members each: S_1, S_2 - typical, and S_3, S_4 - atypical. It is evident that almost any unique column $S_1(i), S_2(i)$, $S_3(i), S_4(i)$ of $\Im$ can correctly classify the two drug sets even using a simple hyperplane. And the number of such columns might be very large among the ~17000. The same time, it is realistic that different sets of columns are classifying the classes differently. Formally, a collection of subsets of the set $\{1,2,\mathrm{K}\ ,n\}$ is known as a set of support systems Ω [1]. Support system is the unit used in comparison of a pair of object descriptions. This is when a set of distances, - each by a member of Ω is defined. The application counterpart is that a set of features – not smaller and not larger than a support system is very effective in describing a particular classification. This brings us to the problems of determining the proper column subsets (support systems), which provide the maximal difference between classes (quality vs. accuracy of classification). In doing this we will eliminate the equivalent (in some sense) columns from one side; and will compose the sets of columns representing different equivalency subsets as approximations to the proper support systems. Last general note we bring is that for classes we consider, support systems are presented - by two vectors in each class. We connect these two vectors into the intervals and consider the best hyperplane separation of these two intervals. We receive a simplest geometry separation problem. The advantage is that we are able to compare support systems finding the most effective ones among them.

1. Classifiers

At first we define **Elementary classifiers**.

These are hyperplane classifiers by small number of columns. **1-classifier** is defined through a single column (let say the i th) and its expression values $S_1(i), S_2(i)$, and $S_3(i), S_4(i)$. Denote by $t_{av}(i)$ and $a_{av}(i)$ the average values on the intervals $(S_1(i), S_2(i))$ and $(S_3(i), S_4(i))$ respectively, and let $t_w(i)$ and $a_w(i)$ is the lengths of these intervals. **1-classifier** $c_1(i)$ by i th column, $i \in \overline{1,n}$ (n is the number of gene expressions), is defined as the balanced (by values $t_w(i)$ and $a_w(i)$) middle point of interval $(t_{av}(i), a_{av}(i))$, and $|t_{av}(i) - a_{av}(i)| = f_1(i)$ is called the power/force of $c_1(i)$.

2-classifier considers pair of genes and expression values. Logically 2-classifiers are to be composed by pairs of genes, higher ranked by corresponding **1-classifiers**. Arranging columns by decreasing order of values $f_1(i)$ we rank the gene expressions by their forces for differentiating two drug groups. 2-classifiers and in general k -classifiers consider any k columns, construct average values on corresponding intervals in classes (intervals by row vector pairs) and define structures $c_k(i_1,...,i_k)$ and $f_k(i_1,...,i_k)$. $c_k(i_1,...,i_k)$ defines the hyperplane, separating the average expressions by drug groups and gene collections, and $f_k(i_1,...,i_k)$ defines the quality of this separation.

Generally k -classifiers examine k columns, construct convex hulls in areas of two considered classes, consider the geometrical centers and balanced middle point, which serves as the value for classification. In our case convex hulls are just intervals of a multidimensional vector space. The force $f_k(i_1,...,i_k)$ is defined through the projections of intervals into the separating hyperplans. The projection area, divided on length of interval projections, provides a comparable for all k measure of force of separation.

2. Future Work – **Growing Support Systems**.

Among 2^n elementary classifiers defined above, we intend to find those ones for which corresponding subsets of genes are most differentially expressed by drug groups. The simplest way is to start by a 1-classifier, and growing it step by step to k -classifiers so that the forces are strictly increasing, with interruption in the k th step. Any k-classifier may be considered as a composition of one $k-1$-classifier together with one new column. Concepts $c_k(i_1,...,i_k)$ and $f_k(i_1,...,i_k)$ in this way introduce monotonity relation between gene sets, put into 1-1 correspondences to the vertices of an n dimensional unit cube. However, this might be rather hard to fulfill because of for some large values of k it will become impossible to consider all 2^k sub-classifiers. The search area for these subsets is very large, and appropriate heuristics to combat this complexity is necessary.

We consider several heuristics:

➢ Sorting 1-classifiers by decreasing forces $f_1(i)$, and eliminating from the further treatment columns with forces lower than the threshold selected. Let the columns in sorted sequence are as $i_1, i_2, ..., i_k, ..., i_n$. An important property of this sequence is the first index i_0 so that forces $f_k(i_1,...,i_k)$ are increasing for $i_k < i_0$ and this increase interrupts at the point i_0. Besides, sorting may also be applied to the mixed sets of classifiers because of the note on comparability of forces for different k 's.

➢ Consider an arbitrary hyperplane elementary classifier $c_k(i_1,...,i_k)$. Compose n-dimensional binary vector, evaluating coordinates $i_1,...,i_k$ as 1. Completing by 0 all the coordinates, not used in $c_k(i_1,...,i_k)$ we create a 1-1 correspondence between classifiers and n-cube vertices. Applying hierarchical clustering in n-cube layers we split k-classifiers by the equivalency relation (after some cut of dendrogram). Similarity measure used is some correlation between the hyperplanes (their coefficient vectors). We consider the representatives sets of clusters. Some of them may give the same force of classifying drug groups by gene expressions as the whole descriptive table does. In this way we reduce the dimensionality combating the exponential explosion for large n.

➢ As it was mentioned, 1-classifiers might be directly sorted by their forces. Any k -classifier may be considered as a composition of one $k-1$-classifier $c_k(i_1,...,i_{k-1})$ together with one new column i_k. In terms of class vectors this change means concatenation of a new dimension in direction i_k. Concepts $c_k(i_1,...,i_k)$ and $f_k(i_1,...,i_k)$ in this way introduce monotonity relation between gene sets in the same way as the vertices of n dimensional unit cube which are in 1-1 correspondence to elementary classifiers. Considering subsets of different n-cube layers and taking into account monotonity we may apply the chain split technology [10] in finding the best separating gene sets. It is important to note that chain split (and other known frequent subsets growing algorithms of association rule mining) work with random objects otherwise with overall structure of all objects which is computationally hard. Instead, the representatives set mentioned above are a valuable heuristic that may help in reducing the computational complexity in growing.

➢ Consider the convex hull Ξ of all classifiers $c_j, j \in \overline{1,2^n}$ in n dimensional vector space. The volume and shape of Ξ appears as a sophisticated measure of drug groups' differences, characterized by the gene expressions. Approximation of Ξ by smaller groups of genes might be achieved in different ways. Such smaller subsets are effective candidates for separating the drug group-driven expression differences. These subsets might be compared to functional gene subsets describing the drug influences. A satisfactory approximation of Ξ by gene sets or by classifiers sets shows that these subsets keep the diversity of drug groups. The approximation we considered is greedy algorithm, given in [11].

3. APPLIED MODEL

To generate ranked gene list, first, average intragroup (TN and AN) expression levels for each gene were calculated. Then, for each gene the average level for TN group was subtracted from the average level of AN group. Finally, the gene list was sorted according to decrease in the average differences between groups (from largest to smallest). Further, Gene Set Enrichment Analysis (GSEA) [6] was applied to identify functional pathways (together with genes involved in each) affected by typical and atypical antipsychotic treatment.

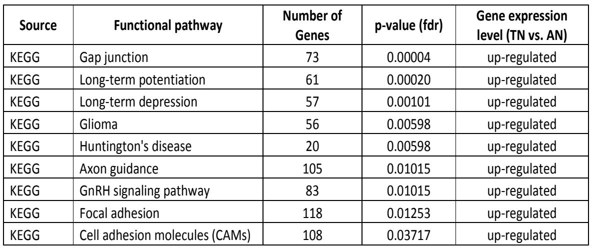
Given an a priori defined set of genes (e.g., genes encoding proteins of metabolic pathway, located in the same cytogenetic band, or sharing the same GO category), the goal of GSEA is to determine whether the members of set are randomly distributed throughout the ranked gene list or primarily found at the top or bottom. Whenever a gene belonging to the functional set is found, an enrichment statistic (ES) is increased by a certain amount, otherwise the ES is decreased. The enrichment score is the maximum deviation from zero in the random walk and corresponds to a support set elementary classifiers statistic. The minimum and maximum of this enrichment score are used to estimate the significance of the enrichment.

For GSEA analysis of the generated ranked gene list the GeneTrail web-based software developed by Center of Bioinformatics of Saarland University was used [7]. For multiple testing adjustment Benjamini and Hochberg's false discovery rate (FDR) was used. P values less than 0.05 (after FDR correction) were considered as significant. GeneTrail covers a wide variety of biological categories and pathways, from which we chose KEGG.

4. RESULTS AND DISCUSSION

The results of this study showed different patterns of gene expression in frontal cortex of mice treated with typical and atypical antipsychotics. All identified functional pathways were up-regulated in TN group compared to AN group (table 1).

Table 1. Functional pathways affected by treatment with typical and atypical neuroleptics



Source	Functional pathway	Number of Genes	p-value (fdr)	Gene expression level (TN vs. AN)
KEGG	Gap junction	73	0.00004	up-regulated
KEGG	Long-term potentiation	61	0.00020	up-regulated
KEGG	Long-term depression	57	0.00101	up-regulated
KEGG	Glioma	56	0.00598	up-regulated
KEGG	Huntington's disease	20	0.00598	up-regulated
KEGG	Axon guidance	105	0.01015	up-regulated
KEGG	GnRH signaling pathway	83	0.01015	up-regulated
KEGG	Focal adhesion	118	0.01253	up-regulated
KEGG	Cell adhesion molecules (CAMs)	108	0.03717	up-regulated

The results obtained suggest that AN, as compared to TN, have less influence on regulatory pathways contributing to neurodegeneration (Huntington's disease) and neuronal architectonics (Axon guidance, Gap junction), cell proliferation (Glioma). Moreover, according to our findings, TN strongly affects GnRH signaling pathway, as well as immune response regulatory reactions (Focal adhesion and Cell adhesion molecules), whereas AN have very week influence on these processes. In addition, our study revealed that AN possess less pronounced effects on cognition, particularly related to learning memory (Long-term potentiation, Long-term depression) than TN do.

5. CONCLUSION

The benefit of AN, compared to TN, includes less side effects related to functional pathways of brain frontal cortex. Extended classification algorithms are designed to analyze the applied data which are of very specific structure.

REFERENCES

1. Yu. Zhuravlev, Selected research publications, Magistr, Moscow, 1998, 420p (in russian).
2. L. Aslanyan, J. Castellanos, Logic based Pattern Recognition - Ontology content (1), Int. Journal "Information Technologies and Knowledge", v.1, 2007.
3. R. Galili-Mosberg, et al. "Haloperidol-induced neuro-toxicity-possible implications for tardive dyskinesia", J. Neural Transm., 107 (4), pp. 479-490, 2000.
4. I. Gil-ad, et al. "Evaluation of the neurotoxic activity of typical and atypical neuroleptics: relevance to iatrogenic extrapyramidal symptoms", Cell. Mol. Neurobiol., 21 (6), pp. 705-716, 2001.
5. S. Hakobyanet al. "Classical pathway complement activity in schizophrenia", Neuroscience Letters 374, pp 35-37, 2005.
6. A. Subramanian, et al. "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles", PNAS, 102(43), pp. 15545-155502005, 2005.
7. C. Backes et al. "GeneTrail - advanced gene set enrichment analysis", Nucleic Acid Research, Web Server Issue 2007.
8. G.F. Von Borries, Partition clustering of high dimensional low sampling size data base on p-values, PhD dissertation, Kansas State University, 2008, p. 139.
9. F. Divina and J. S. Aguilar-Ruiz, Biclustering of expression data with evolutionary computation", IEEE Transactions on Knowledge and Data Engineering, vol. 18, pp. 590–602, 2006.
10. L. Aslanyan and H. Sahakyan, Chain split and computation in practical rule mining, Information Science and Computing, International book series no. 8., Classification, forecasting, data mining, 2009, pp.132-135.
11. H. Sahakyan, L. Aslanyan, Differential Balanced Trees and (0,1)-Matrices, International Journal "Information Theories and Applications", ISSN 1310-0513, 2003, Volume 10, Number 4, pp. 363-369.