2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

A Transdiagnostic Biotype Detection Method for Schizophrenia and Autism Spectrum Disorder Based on Graph Kernel

Yuhui Du*, Hui Hao, Ying Xing, Ju Niu, Vince D Calhoun

Abstract— **Psychiatric diagnoses based on clinical manifestations are prone to be inaccurate. Schizophrenia (SZ) and autism spectrum disorder (ASD) were historically considered as the same disorder, and they still have many overlaps of clinical symptoms in the current standard. Therefore, there is an urgent need to explore the potential biotypes for them using neuroimaging measures such as brain functional connectivity (FC). However, previous studies have not effectively leveraged FC in detecting biotypes. Considering that graph theory helps reveal the topological information in FC, in this paper, we propose a graph kernel-based clustering method to explore transdiagnostic biotypes using FC estimated from functional magnetic resonance imaging (fMRI) data. In our method, frequent subnetworks are identified from the whole-brain FCs of all subjects, and then the graph kernel similarity is computed to measure the relationship between subjects for clustering. Based on fMRI data of 137 SZ and 150 ASD subjects, we obtained meaningful biotypes using our method, which shows significant differences between the identified biotypes in FC. In brief, our graph kernel-based clustering method is promising for transdiagnostic biotype detection.**

Index Terms— Biotype, Transdiagnostic, Graph kernel, Schizophrenia, Autism spectrum disorder

I. Introduction

The reliability of psychiatric nosology has remained questionable due to the subjectivity of using self-reported symptom scores, which are not biologically based and show overlaps among disorders. Heterogeneity of the sophisticated brain may result in classifying mental diseases with different causal mechanisms into the same disorder [1]. This problem is especially obvious for mental disorders with similar symptoms, such as schizophrenia (SZ) and autism spectrum disorder (ASD). Both SZ patients and ASD patients have difficulties in communication and normal social behavior, which are highly heritable and fatal. The similarity in clinical phenotypes and genotype expressions in these two disorders can hinder precision treatment [2-4].

Recently, a promising avenue focusing on a transdiagnostic approach [5] to identify biotypes for psychiatry has been proposed to find a way out of the dilemma and even go beyond the existing diagnoses. In short, a transdiagnostic approach moves away from the currently defined nosology and explores the potential for biological types which can better inform us about mental disorders using neuroimaging measures. Clementz et al. divided three schizophrenia spectrum disorders into three biotypes using nine cognitive control and sensorimotor reactivity variables [1]. Ivleva et al. proved that biotypes developed from neurophysiologic measures showed stronger between-group separation than the conventional diagnoses [6]. Chand et al. used structural MRI and discovered two distinct biotypes for SZ using semi-supervised machine learning [7]. To date, how to effectively utilize brain functional connectivity (FC) to explore biotypes is unsolved yet.

FC estimated from the resting-state functional magnetic resonance imaging (fMRI) data is a useful measure to reflect functional interaction of the brain [8]. Graph theory works well in investigating FC [9]. In recent years, the graph kernel method has attracted increasing interest. A work mined the discriminative subnetworks and distinguished mild cognitive impairment patients from healthy controls based on the graph kernel similarity, which obtained competitive classification performance [10]. Similarly, Zhang et al. used a frequent-subnetwork-mining method to carry out the graph kernel-based classification for minimal hepatic encephalopathy and healthy controls [11]. However, there is no work applying the graph kernel approach to cluster subjects for biotype detection.

In this study, we propose a graph kernel-based biotype detection method by taking advantage of the graph-based substructure pattern mining (gSpan) technique [12] and the graph kernel similarity measure [13]. We apply the method to discover biologically distinct biotypes of SZ and ASD using FC features, and validate that there are significant FC differences between the identified biotypes, providing meaningful insights for the transdiagnostic regrouping.

II. Materials and Methods

A. Subjects and Functional Connectivity Estimation

In this work, we used the fMRI data of 137 SZ and 150 ASD patients that were from Functional Biomedical

*Research was supported by the National Natural Science Foundation of China (Grant No. 61703253 and 62076157, to YHD).

Yuhui Du is with School of Computer and Information Technology, Shanxi University, Taiyuan, China and Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, GA, USA (e-mail: duyuhui@sxu.edu.cn).

Hui Hao, Ying Xing, and Ju Niu are with School of Computer and Information Technology, Shanxi University, Taiyuan, China.

Vince D Calhoun is with Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, GA, USA.

Informatics Research Network (FBIRN) and Autism Brain Imaging Data Exchange II (ABIDE II), respectively. Similar to an earlier work, we estimated the functional network connectivity (FNC) matrix (i.e., FC among networks) for each subject [4, 14]. Guided by 53 brain functional network templates from a large-sample healthy population, the individual-level brain functional networks and their relevant time courses were automatically estimated using a spatially constrained ICA approach. Then, the FNC matrix (size: 53*53) was computed based on the time courses of networks for each subject. Due to the symmetry of the FNC matrix in which each element represents the connectivity strengths between two networks, the matrix is often converted into a vector containing only upper triangular elements for further analysis. However, such processing ignores the topological information of the FNC matrix. In our method, instead of pulling the FNC matrix into a vector, we threshold the matrix and work with it from a graph view.

B. *Similarity Computation between Subjects*

In this subsection, a similarity matrix reflecting relationships between subjects is obtained based on gSpan and graph kernel techniques. The similarity matrix will be used for clustering of subjects, aiming to explore potential biotypes.

1) Mining of Frequent Subnetworks and Reconstructing Functional Network Connectivity

Subnetwork (or subgraph) can reflect the local network structure. To mine the subgraphs that frequently exist across most subjects, many methods have been proposed, such as the priori-based algorithm [15], the frequent subgraph discovery algorithm [16], and gSpan [10, 17]. In this work, we extracted the frequent subgraphs from the FNC matrices of all SZ and ASD subjects via gSpan.

Given a set of undirected graphs in which a graph G represents thresholded FNC matrix of a subject, we mine frequent subgraphs which exist in most graphs based on gSpan. Here, nodes and edges of a graph respectively represent brain functional networks and connectivity between paired functional networks. Mining frequent subgraphs using gSpan mainly consist of three steps. 1) Constructing Depth-first search (DFS) trees for each graph, and each DFS tree can be coded by a unique DFS code. 2) Obtaining a lexicographic order to indicate the priority order of all codes derived from all graphs. 3) Mining frequent subgraphs in DFS trees corresponding to DFS codes with high priority order.

Based on the mined frequent subgraphs, each graph G is reconstructed to form a new sparse graph $\hat{G}$ to represent the FNC matrix of each subject. Specifically, the edges of all frequent subgraphs existing in the graph G constitute the edges of the $\hat{G}$.

2) Graph Kernel-based Similarity Computation

To measure the topological similarity of the paired reconstructed graphs, we use an iterative graph kernel method to perform similarity computation, i.e., Weisfeiler-Lehman (WL) subtree kernel. Given a reconstructed graph $\hat{G}$ with an original label set of nodes, we iteratively update the label of each node by combining its current label and labels of adjacent nodes, and then the extended label set is obtained. By exploiting the extended label set, the Weisfeiler-Lehman subtree kernel between two reconstructed graphs $\hat{G}_m$ and $\hat{G}_n$ in tth iteration can be defined as:

$$k^t < \hat{G}_m, \hat{G}_n >=< \phi^t(\hat{G}_m), \phi^t(\hat{G}_n) >, \qquad (1)$$

where $\phi^t(\cdot)$ counts the frequency of each label within the extended label set of a reconstructed graph to form a feature vector. The similarity between the two reconstructed graphs is measured by the dot product between two feature vectors derived from two graphs. It is worth noting that the iteration stops until the node label sets of two graphs are identical or the number of iterations reaches the maximum.

C. *Biotype Detection*

Based on the graph-kernel similarity, we use a hierarchical clustering method [18] to cluster the subjects for the biotype detection. Here, each cluster represents one biotype. In addition, we use a spectral clustering method, Normalized cut (Ncut), to verify whether the identified biotypes are stable. When using Ncut, we set the number of clusters to 2 and 3, respectively, to examine if the biotypes identified using Ncut coincide with that from the hierarchical clustering. Based on the results from different clustering methods, we evaluate the "purity" of the identified clusters. The purity assessment uses the Jaccard coefficient to reflect the similarity between clusters, and the Jaccard coefficient is defined as the ratio of the sample size (i.e., subject number) of the intersection between clusters and the sample size of the union of clusters.

In this work, we also compare our graph kernel-based method with the traditional similarity measure method in which we convert each FNC matrix into an FNC vector and calculate the Pearson correlation coefficient between the FNC vectors of subjects to build a between-subject similarity matrix. For an intuitive visualization of the results, we separately project the graph kernel-based similarity and the Pearson correlation similarity using original FNC features of subjects into a 2D plane using the t-Distributed Stochastic Neighbor Embedding (t-SNE) method [19], where the subjects are colored according to their cluster labels.

In addition, we verify the stability of our method and the validity of biotypes by repeatedly performing the biotype detection on parts of the whole subjects. Specifically, we repeatedly conduct the clustering on 100 randomly selected subsets of subjects (here, one subset includes 90% of the subjects). For each subset of subjects, we evaluate the "purity" between the resulting clusters and the clusters obtained using the whole subjects. Finally, we average the "purity" across 100 runs (corresponding to 100 subsets) for a summary.

D. *Functional Connectivity Differences of Biotypes*

To evaluate the differences between the identified biotypes, we further perform a two-sample t-test on each FNC between the subjects clustered in any two biotypes, resulting in T-values for FNCs.

III. Results

A. Biotypes Extracted from SZ and ASD

Figure 1 displays the result of the hierarchical clustering using the graph kernel similarity, which reflects how the subjects are grouped into clusters and the relationship between clusters. It is seen that all subjects can be divided into two clusters, named Biotype 1_2 and Biotype 2_2. When grouping the subjects into three clusters, the biotypes are named Biotype 1_3, Biotype 2_3, and Biotype 3_3, and Biotype 2_2 included Biotype 1_3 and Biotype 2_3. Table I shows the purity measurement between the results from the hierarchical clustering and Ncut methods using graph kernel similarity, supporting that the results of the two clustering methods are highly consistent with each other, which verifies the stability of our method and the reliability of our results.

As mentioned above, we also compared our clustering results using the graph kernel similarity with that using traditional FC similarity. It is observed from Figure 2 that the graph kernel similarity worked well in identifying the biotypes, however, the traditional similarity method didn't generate clear clusters.

Table II displays the mean of "purity" (across 100 runs) between the clustering results using all subjects and using 90% of subjects. It is seen that the biotypes are stable since the clustering "purity" was high between using all subjects and using part of subjects.

B. Functional Connectivity Differences of Biotypes

Figure 3 shows the functional connectivity differences between different biotypes. Figure 3 (A) supports that the FC differences between Biotype 1_2 and Biotype 2_2 are mainly within the sub-cortical (SC) region. Figure 3 (B) shows that the significant differences between Biotype 1_3 and Biotype 2_3 mainly include FNCs within the SC regions, between the SC and auditory (AU) regions, and between the SC and cerebellar (CB) regions. Figure 3 (C) and (D) display significant FC differences between Biotype 1_3 and Biotype 3_3, and between Biotype 2_3 and Biotype 3_3, respectively. Compared to Biotype 3_3, Biotype 1_3 and Biotype 2_3 consistently show the significant differences of FNCs within the SC regions, between SC and CB regions, and between SC and AU regions.

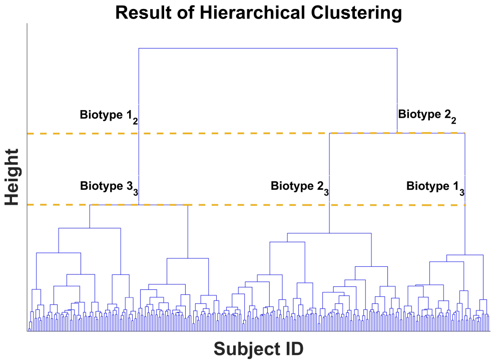

Figure 1. The result of the hierarchical clustering method based on the graph kernel similarity.

TABLE I. The "purity" measures of clusters obtained from different clustering methods

		Hierarchical Clustering				
		Biotype 1_2	***Biotype 2_2***	***Biotype 1_3***	***Biotype 2_3***	***Biotype 3_3***
Ncut	***Biotype 1_2***	0.9940	0	0.3253	0.6707	0
	Biotype 2_2	0.0035	0.9917	0	0.0043	0.9917
	Biotype 1_3	0.4345	0.0052	0.7067	0.1198	0.0052
	Biotype 2_3	0.5082	0.0075	0.0062	0.7077	0.0751
	Biotype 3_3	0.0037	0.8512	0	0.0046	0.8512

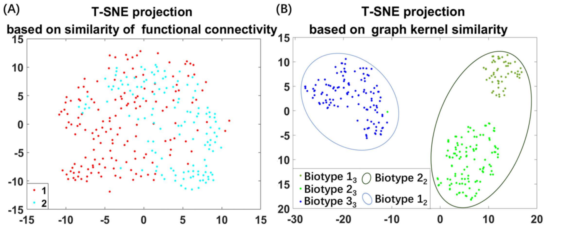

Figure 2. T-SNE projection of subjects for (A) Pearson similarity using original FC features and (B) graph kernel similarity. Here, the subjects are colored by their cluster labels.

TABLE II. The mean of "purity" between the clusters using all subjects and the clusters using 90% of subjects

		All subjects				
		2 clusters		***3 clusters***		
90% of subjects	***2 clusters***	0.9909	0.0050	-	-	-
		0.0040	0.9919	-	-	-
	3 clusters	-	-	0.9901	0	0
		-	-	0.0040	0.9783	0.0062
		-	-	0.0021	0.0042	0.9909

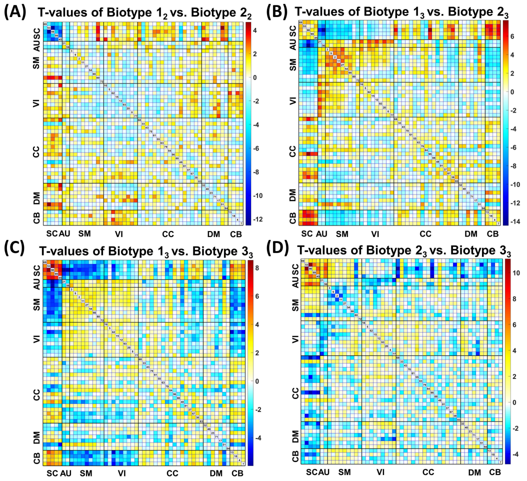

Figure 3. Differences between the biotypes in FNC. Here, the biotypes are identified by the hierarchical clustering based on graph kernel similarity. Figure 3 (A), (B), (C) and (D) show the T-values of two-sample t-tests on FNCs for (A) Biotype 1_2 vs. Biotype 2_2, (B) Biotype 1_3 vs. Biotype 2_3, (C) Biotype 1_3 vs. Biotype 3_3, and (D) Biotype 2_3 vs. Biotype 3_3, respectively.

IV. Conclusions

Due to the inaccurate diagnosis of mental disorders, biotype detection has attracted increasing attention for mental disorders with highly overlapping symptoms. Neuroimaging measure such as FC plays an important role to discover potential biotypes [20]. However, previous studies focusing on detecting biotypes using FC features ignore the graph structure in brain functional connectivity [21]. In this paper, we propose to use frequent subgraph mining combined with graph kernel-based similarity to discover the distinct biotypes via a clustering algorithm, and successfully apply our method to identify the biotypes in SZ and ASD.

The identified biotypes manifest that the graph-based functional connectivity measures can separate subjects with psychosis into subgroups that are neurobiologically distinctive and biologically meaningful. By analyzing the significant functional connectivity differences among biotypes, we highlight the important role of the SC regions in refining the nosology of two disorders. For the identified two biotypes, Biotype 2_2 shows higher connectivity than Biotype 1_2 within the SC domain. For example, compared to Biotype 1_2, Biotype 2_2 exhibits significantly increased connectivity between the hypothalamus and putamen. For the identified three biotypes, there are significant connectivity differences mainly in SC, AU, and CB regions, involving the putamen, the thalamus, the middle temporal gyrus, the superior temporal gyrus, and the cerebellum. In particular, we found that the strength of functional connectivity presents the following trend: Biotype 3_3 > Biotype 2_3 > Biotype 1_3 within AU regions, Biotype 1_3 > Biotype 2_3 > Biotype 3_3 within SC regions, Biotype 1_3 > Biotype 3_3 > Biotype 2_3 between SC and CB regions, and Biotype 2_3 > Biotype 3_3 > Biotype 1_3 between SC and AU regions. The significant differences among the detected biotypes demonstrate the meaning of transdiagnostic approaches, which requires more attentions [5, 22].

Taken together, we propose a new data-driven biotype detection method by effectively using the functional connectivity information at a higher level, although some parameters may be selected more automatically in the future. We believe our method can be applied to other mental disorders for benefiting the exploration of potential biotypes.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 61703253 and 62076157, to YHD and the National Institutes of Health grant R01MH123610 to VDC). FBIRN data acquisition was supported by the National Institutes of Health grant numbers: NIH 1U24 RR021992 (Function Biomedical Informatics Research Network) and NIH 1U24 RR025736-01 (Biomedical Informatics Research Network Coordinating Center). We acknowledge the contribution of all participants in FBIRN and ABIDE projects.

References

[1] B. A. Clementz, J. A. Sweeney *et al.*, "Identification of Distinct Psychosis Biotypes Using Brain-Based Biomarkers," *American Journal of Psychiatry*, vol. 173, no. 4, pp. 373-384, 2016.

[2] B. St Pourcain, E. B. Robinson *et al.*, "ASD and schizophrenia show distinct developmental profiles in common genetic overlap with population-based social communication difficulties," *Molecular Psychiatry*, vol. 23, pp. 263-270, 2018.

[3] A. Andriamananjara, R. Muntari *et al.*, "Overlaps in brain dynamic functional connectivity between schizophrenia and autism spectrum disorder," *Scientific African*, vol. 2, pp. 1-10, 2018.

[4] Y. Du, Z. Fu *et al.*, "NeuroMark: An automated and adaptive ICA based pipeline to identify reproducible fMRI markers of brain disorders," *Neuroimage Clin*, vol. 28, pp. 1-19, 2020.

[5] P. Fusar‐Poli, M. Solmi *et al.*, "Transdiagnostic psychiatry: a systematic review," *World Psychiatry*, vol. 18, no. 2, pp. 192-207, 2019.

[6] E. I. Ivleva, B. A. Clementz *et al.*, "Brain structure biomarkers in the psychosis biotypes: findings from the bipolar-schizophrenia network for intermediate phenotypes," *Biological psychiatry*, vol. 82, no. 1, pp. 26-39, 2017.

[7] G. B. Chand, D. B. Dwyer *et al.*, "Two distinct neuroanatomical subtypes of schizophrenia revealed using machine learning," *Brain*, vol. 143, no. 3, pp. 1027-1038, 2020.

[8] Y. I. Sheline and M. E. Raichle, "Resting state functional connectivity in preclinical Alzheimer's disease," *Biological psychiatry*, vol. 74, no. 5, pp. 340-347, 2013.

[9] E. T. Bullmore and D. S. Bassett, "Brain graphs: graphical models of the human brain connectome," *Annu Rev Clin Psychol*, vol. 7, pp. 113-140, 2011.

[10] F. Fei, B. Jie *et al.*, "Frequent and discriminative subnetwork mining for mild cognitive impairment classification," *Brain Connectivity*, vol. 4, pp. 347-360, 2014.

[11] D. Zhang, L. Tu *et al.*, "Subnetwork mining on functional connectivity network for classification of minimal hepatic encephalopathy," *Brain Imaging Behav*, vol. 12, no. 3, pp. 901-911, 2018.

[12] X. Yan and J. Han, "gspan: Graph-based substructure pattern mining," in *2002 IEEE International Conference on Data Mining*, pp. 721-724, 2002.

[13] N. Shervashidze, P. Schweitzer *et al.*, "Weisfeiler-Lehman graph kernels," *Journal of Machine Learning Research*, vol. 12, no. 9, pp. 2539-2561, 2011.

[14] Y. Du and Y. Fan, "Group information guided ICA for fMRI data analysis," *Neuroimage*, vol. 69, pp. 157-197, 2013.

[15] A. Inokuchi, T. Washio *et al.*, "Complete Mining of Frequent Patterns from Graphs: Mining Graph Data," *Machine Learning*, vol. 50, no. 3, pp. 321-354, 2003.

[16] M. Kuramochi and G. Karypis, "Frequent subgraph discovery," *2001 IEEE International Conference on Data Mining*, pp. 313-320, 2001.

[17] X. Cui, J. Xiang *et al.*, "Classification of Alzheimer's Disease, Mild Cognitive Impairment, and Normal Controls With Subnetwork Selection and Graph Kernel Principal Component Analysis Based on Minimum Spanning Tree Brain Functional Network," *Frontiers in Computational Neuroence*, vol. 12, pp. 1-12, 2018.

[18] L. Wang, W. Cheng *et al.*, "Association of specific biotypes in patients with Parkinson disease and disease progression," *Neurology*, vol. 95, no. 11, pp. e1445-e1460, 2020.

[19] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579-2605, 2008.

[20] L. Miranda, R. Paul *et al.*, "Functional MRI applications for psychiatric disease subtyping: a review." *arXiv preprint*, pp. 1-16, 2020.

[21] A. T. Drysdale, L. Grosenick *et al.*, "Resting-state connectivity biomarkers define neurophysiological subtypes of depression," *Nature Medicine*, vol. 23, no. 1, pp. 28-38, 2017.

[22] T. Dalgleish, M. Black *et al.*, "Transdiagnostic approaches to mental health problems: Current status and future directions," *Journal of consulting and clinical psychology*, vol. 88, no. 3, pp. 179-195, 2020.