Motif Discovery in Brain Networks of Patients with Somatoform Pain Disorder

Annahita Oswald
University of Munich
oswald@dbs.ifi.lmu.de

Bianca Wackersreuther
University of Munich
wackersreuther@dbs.ifi.lmu.de

***Abstract*—Idiopathic chronic pain disorders constitude a large, clinically important health care problem that urgently needs deeper pathophysiological insight. The understanding which brain compartments are involved in such diseases, is therefore a very interesting research topic in neurological medicine. In this paper, we apply an efficient algorithm for motif discovery to time series data of somatoform patients and healthy controls. We find groups of brain compartments that occur frequently within the brain networks and are characteristic for patients with somatoform disorder.**

I. Introduction

Understanding the mechanisms that govern neural processes in the brain is an important topic for medical studies. Recently the emergence of an increasing number of clinical diagnostics approve a global analysis of complex processes and therefore allow a better understanding of numerous diseases. We focus on the mechanisms in the brain that are associated with somatoform pain disorder. Somatization disorders constitude a large, clinically important health care problem that urgently needs deeper insight [12].

Earlier studies have revealed that different subunits interact among each other, when particular stimuli are performed to the brain. Hence the different components form a network. From an algorithmic point of view, interacting subunits act as frequent subgraphs within the network. From a medical perspective, it is interesting to ask to which degree interacting subunits exert influence on medical disorders.

In this article, we study the question whether brain compartments of patients with somatoform pain disorder form different motifs than brain compartments of subjects that do not suffer from this disease. For this purpose we analyze task-fMRI scans (a special form of neuroimaging) of the brain of 11 subjects, 7 patients with somatoform pain disorder and 4 healthy controls that attended the study of [4]. In this study both groups underwent alternate phases of non-pain and pain stimuli during the fMRI scanning. We construct a network for each subject where each node represents a voxel in the fMRI image. Voxels are grouped together in 90 so called regions of interest using the template of Tzourio-Mazoyer et al. [10]. We apply the efficient heuristic approach GREW [8] to uncover frequent subgraphs in each of these networks. GREW is designed to operate on a large graph and to find patterns corresponding to connected subgraphs that have a large number of vertex-disjoint embeddings. This approach overcomes the limitations of existing complete or heuristic frequent subgraph discovery algorithms. We use GREW to find frequent subgraphs as it can efficiently operate on large networks. We demonstrate that patients with somatoform disorder show different activation patterns in the brain than healthy subjects.

The remainder of this article is organized as follows. Section 2 gives a brief survey of the previous work on frequent subgraph mining. Section 3 begins with basic definitions of graph theory and describes our proposed framework in the following. Experimental results of our method are given in Section 4 and Section 5 concludes our paper.

II. Related Work

Several algorithms have been defined for finding frequent subgraphs and their embeddings in one large graph or in a dataset of graphs. We distinguish algorithms for 'graph dataset mining' that work on a dataset of graphs, and algorithms for 'large graph mining' that discover frequent motifs in one large graph. While frequent subgraph algorithms that work on large graphs can directly be applied to datasets of graphs, the other direction is more complicated. However, by splitting a large graph into subgraphs, one can still use a graph dataset mining algorithm for frequent subgraph discovery on the large graph (albeit some subgraphs might be lost by the split).

A. Graph Dataset Mining

For the graph dataset mining tasks, approaches can be broadly divided into two classes, apriori-based and pattern-growth based. AGM (Apriori-based Graph Mining)[5] determines subgraphs G' in a dataset DS of graphs that occur in at least $minsup$ percentage of all graphs in the dataset. AGM works on graphs with edge and node labels. In principle, AGM uses the apriori principle of iterated candidate generation and candidate evaluation. Candidate generation means that candidate subgraphs are created by joining subgraphs that have been shown to be frequent in earlier iterations. In the candidate evaluation phase, these candidates are tested, i.e. it is checked whether their frequency is greater than $minsup$, and then the whole process is iterated until all frequent patterns have been found. AGM uses a canonical

1529-4188/09 $25.00 © 2009 IEEE
DOI 10.1109/DEXA.2009.48

form and a normal form to represent subgraphs to reduce runtime cost for subgraph isomorphism checking.

Similar to AGM, FSG (Frequent SubGraph Discovery) [6] uses a canonical labeling based on the adjacency matrix. Canonical labeling, candidate generation and evaluation are sped up in FSG by using graph invariants and the Transaction ID principle, which stores the ID of transactions a subgraph appeared in. This speed-up is paid for by reducing the class of subgraphs discovered to connected subgraphs, i.e. subgraphs where a path exists between all pairs of nodes.

The most well-known member of the class of pattern-growth algorithms, gSpan (graph-based Substructure pattern mining), discovers frequent substructures efficiently without candidate generation [13]. Tree representations of graphs are encoded using a Depth First Search (DFS) code, amongst which a minimum DFS code is chosen according to some lexicographic order. Pre-order DFS-tree search is then conducted to find the complete set of frequent subgraphs in a set of graphs. gSpan is efficient, both with respect to runtime and memory requirements, making it one of the best state-of-the-art algorithms for graph dataset mining. CloseGraph [14] extends gSpan by limiting the search to frequent complete graphs, i.e. subgraphs without supergraphs that have the same support, thereby increasing the efficiency of mining substantially.

B. Large Graph Mining

Unlike graph dataset mining, large graph mining intends to find subgraphs that have $minsup$ embeddings in one large graph.

SUBDUE tries to minimize the minimum description length (MDL) of a graph by compressing frequent subgraphs. Frequent subgraphs are replaced by one single node and the MDL of the remaining graph is then determined. Those subgraphs whose compression minimizes the MDL are considered frequent patterns in the input graph. The candidate graphs are generated starting from single nodes to subgraphs with several nodes, using a computationally-constrained beam search.

GREW [8] and SUBDUE [1] are greedy heuristic approaches to frequent graph mining that deal speed for completeness of the solution. GREW iteratively joins frequent pairs of nodes into one super-node and determines disjoint embeddings of connected subgraphs by a maximal independent set algorithm. Similarly, vSIGRAM and hSIGRAM [7] find subgraphs that are frequently embedded within a large sparse graph, using "horizontal"(h) breadth-first search and vertical (v) depth-first search, respectively. They employ efficient algorithms for candidate generation and candidate evaluation that exploit the sparseness of the graph.

The algorithm by [11] uses a randomized enumeration strategy for sampling subgraphs. This approach exhaustively enumerates all subgraphs rather than randomly sampling from this enumeration. As the runtime effort for exhaustive enumeration grows exponentially in the size of the subgraphs we employ the heuristic approach GREW [8]. This enables us to find frequent subgraphs in large graph rather than to restrict ourselves to small frequent subgraph.

III. Method

In this section, we introduce the main concepts of this paper. We propose a framework for generating network data out of fMRI timeseries data and describe how to perform frequent subgraph mining on these networks.

A. Basics of Graph Theory

We start with a brief summary of necessary definitions from the field of graph mining.

Definition 1 (Labeled graph / network): A **labeled graph** is represented by a 4-tuple $G = (V, E, L, l)$, where V is a set of vertices (i.e. nodes), $E \subseteq V \times V$ is a set of edges, L is a set of labels, and $l : V \cup E \rightarrow L$ is a mapping that assigns labels to vertices V and edges E. If labels are not of decisive importance, we will use the *short* definition of a graph $G = (V, E)$. In the following we also use **network** as a synonym for graph.

Definition 2 (Subgraph): Let $G_1 = (V_1, E_1, L_1, l_1)$ and $G_2 = (V_2, E_2, L_2, l_2)$ be labeled graphs. G_1 is a **subgraph** of G_2 ($G_1 \sqsubseteq G_2$) if the following conditions hold: $V_1 \subseteq V_2$, $E_1 \subseteq E_2$, $L_1 \subseteq L_2$, $l_1 = l_2$. If G_1 is a subgraph of G_2, then G_2 contains G_1.

Definition 3 (Isomorphism): Two graphs are **isomorphic** if there exists a bijection f between the nodes of two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ such that $(v_{1a}, v_{1b}) \in E_1$ iff $(v_{2a}, v_{2b}) \in E_2$ where $v_{2a} = f(v_{1a})$ and $v_{2b} = f(v_{1b})$. If G_1 is isomorphic to G_2, we will refer to (v_{1a}, v_{1b}) and (v_{2a}, v_{2b}) as **corresponding edges** in the following.

The problem to decide whether two graphs are isomorphic, i.e. the *graph isomorphism problem*, is not yet known to be NP-complete or in P. Given two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, the *subgraph isomorphism problem* consists in finding a subgraph of G_2 that is isomorphic to G_1. This problem is known to be NP-complete [2].

Definition 4 (Embedding): If graph G_1 is isomorphic to a subgraph S of graph G_2, then S is referred to as an **embedding** of G_1 in G_2.

Definition 5 (Frequent subgraph / motif): A graph G_1 is a **frequent subgraph** of graph G_2 if G_2 contains at least t embeddings of G_1, where t is a user-set frequency threshold parameter. Such a frequent subgraph is often called a **motif**.

Definition 6 (Union graph): Given a time series of graphs G_{ts} with n states. Then the **union graph** $DG(G_{ts})$ of G_{ts} is defined as $DG(G_{ts}) = (V_{DG}, E_{DG}, \ell)$, where $V_{DG} = V_i$ for all $1 \leq i \leq n$ and $E_{DG} = \cup_{1 \leq i \leq n} E_i$.

An example for the transformation of a time series of graphs into a union graph is depicted in Figure 1. Note that the union of all edges of the time series is the set of edges of the union graph.

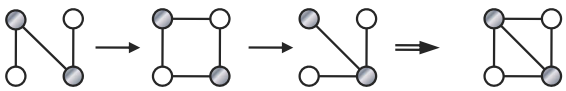


Fig. 1. Transformation from a time series of three labeled graphs into the corresponding union graph.

B. Construction of Brain Network Models Out of fMRI Time-series

We define the voxels of the fMRI image data as the vertex set of the brain network model. Edges between two vertices v_1 and v_2 stand for a similar level of activation of the two voxels. The measured value of a voxel indicates the degree of blood circulation in the particular brain region. The *darker* the voxel the *more* blood is present in the compartment, thus the *higher* the activation.

We are interested in topological patterns that are characteristic for patients with somatoform pain disorder. Therefore we want to perform graph mining methods on one graph that represents interacting regions of the brain. fMRI timeseries data can be transformed in the following manner.

We distinguish two categories of activation levels for each voxel v at each time point i, denoted by v_i. $a(i)$ stands for its activation level at time point i. We determined the $z - score_{median}$ for each voxel v_i by comparing the median activation level of v across the time series with $a(i)$ in order to assign activation categories.

$$z\text{–}score_{median}(v_i) = \frac{a(i) - median_{j \in \{1,..,n\}} a(j)}{median_{k \in \{1,..,n\}} \left| (a(k) - median_{j \in \{1,..,n\}} a(j) \right|}$$

We use a median-based $z\text{–}score$ rather than a mean-based as we want to detect unusually high activation levels. In contrast to the $z\text{–}score_{mean}$ a $z\text{–}score_{median}$ is more robust with respect to these extremes and better suited for detecting them, as validated in initial experiments (not shown here).

- **high activation**: $a(i)$ is significantly higher than the median activation level of v. ($z\text{–}score_{median}(v(i)) \geq 7.0$)
- **no significant activation**: $a(i)$ is not significantly higher than the median activation level of v. ($z\text{–}score_{median}(v(i)) < 7.0$).

Edges between vertices v_1 and v_2 are assigned if v_1 and v_2 both show high activation. We perform frequent subgraph mining on the resulting union graph.

C. Performing Frequent Subgraph Mining on Brain Co-Activation Networks

In order to find frequent subgraphs in our network we have to group our nodes and assign each group a labeling. A meaningful grouping of nodes when considering brain networks is a mapping of the nodes to their corresponding brain compartments. Hence motifs in those networks represent compartments of the brain that show similar activation profile. We removed edges between nodes that share the same label, as a correlated degree of activation within one region is trivial, and we are interested in activity of different regions.
In the next step we applied GREW for finding motifs in our labeled graphs. As approaches that exhaustively enumerate all subgraphs would limit ourselves to motifs of length 3 and 4 we decided to employ a heuristic approach. This allowed us to find motifs of arbitrary size but missed out on some motifs.

D. Evaluation of Detected Motifs

To find motifs that are characteristic for a disease we have to analyze the motifs separately. Therefore we want to detect motifs that occur in patients but not in the control group and vice versa. Another class of motifs that might be interesting are motifs that occur in all subjects that attend a certain study.

Another aspect that should be considered is the label distribution across motifs. A label that is used for a high number of vertices has a higher probability to appear in a motif than a label that covers a small number of nodes. Hence we have to define the normalized frequency of a node label l, denoted by $Freq_{norm}(l)$.

$$Freq_{norm}(l) = \frac{freq_{m \in \{1,...,n\}}(l) \cdot \#Embeddings(m)}{freq_{Background}(l)}$$

$freq_{m \in \{1,...,n\}}$ stands for the number of occurences of label l in a motif m. This number has to be multiplied by the number of isomorphic subgraphs of m, its embeddings. The $freq_{Background}(l)$ describes the number of occurences of l with respect to alle vertices. In our case it refers to the *size* of a ROI.

IV. Experiments

Construction of the brain network models. We created networks for 11 subjects that attended the studies of [4]. The resulting networks comprise 39 to 440 nodes with 90 different classes of node labels and 278 to 13,788 edges. The network models indicate different number of edge types. The type of an edge is referring to the concatenation of the labels of the adjacent nodes. All edges are undirected because in the relationship 'both adjacent voxels show high activation' a direction makes no sense. We have measurements for 325 time steps. The exact statistics of each subject are depicted in table I.

TABLE I
STATISTICS ABOUT OUR 11 NETWORKS.

subject	nodes	different labels	edges	edge types
patient 1	39	18	49	12
patient 2	102	31	456	91
patient 3	185	32	1,997	84
patient 4	241	35	4,017	90
patient 5	263	46	7,407	293
patient 6	313	46	11,814	372
patient 7	440	58	13,788	475
control 1	66	10	358	15
control 2	99	33	408	111
control 3	109	18	1,093	13
control 4	202	35	2,243	133

Timeseries datasets. We used fMRI data (1.5 T MR scanner) of 7 female somatoform patients and 4 healthy controls. Standard data preprocessing including realignment, correction for motion artifacts and normalization to standard space have been performed using SPM2 (available http://www.fil.ion.ucl.ac.uk/spm/). In addition, to remove global effects the voxel time series have been corrected regressing out the global mean, as suggested in [9].

TABLE II
TOTAL NUMBER OF MOTIFS FOUND IN THE 11 NETWORKS.

subject	motifs
patient 1	3
patient 2	38
patient 3	706
patient 4	505
patient 5	752
patient 6	4,154
patient 7	4,256
control 1	15
control 2	32
control 3	133
control 4	236

Vertex labels. We labeled all nodes in our network model by regional parcellation of the voxels into 90 brain regions using the template of Tzourio-Mazoyer et al. [10].

Finding motifs. We searched for topological motifs of arbitrary size using GREW with a threshold parameter of $t = 5$ and a minimum number of one edge.

Evaluation of the motifs. Altogether we found 10,530 different motifs in somatoform patients and healthy controls. 10,173 different motifs were detected among patients, 413 within the group of healthy subjects.
For validation we divided the subjects into three classes. Class (1) contains only the somatoform patients, class (2) consist of the controls exclusively and class (3) composes the union of class(1) and (2). Figure 2 shows typical representatives of each class. The two motifs on the left occur in 57% of the patients but in no healthy subject. The middle motif arises in 50% of the class (2)–subjects but in no patient. The upper motif on the right side was found in 50% of the control group

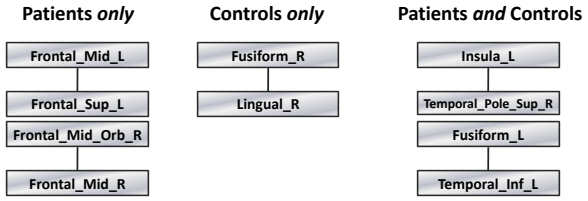


Fig. 2. Typical representatives of motifs found in the groups of somatoform patients, healthy controls respectively and the group of all subjects.

and in 14% of the patients, the lower motif in 25% of the control group and in 43% of the patient group.
The largest motifs (highest number of vertices and edges) of class (1) were found in subject 'patient 6'. They consist of 28 vertices and 29 edges, five different brain compartments are involved in this motif. A total of 34 motifs of this kind were found in this subject. The largest motifs in class (2) were detected in subject 'control 3'. We found two motifs that comprise 12 nodes with two different labels and 17 edges. An example of the largest motifs found in class (1) and the two largest motifs of class (2) are shown in figure 3.
It has to be mentioned that there is no motif that occurs in all subjects.

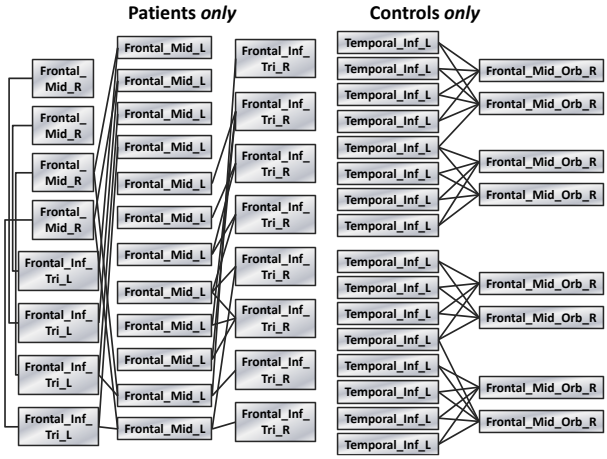
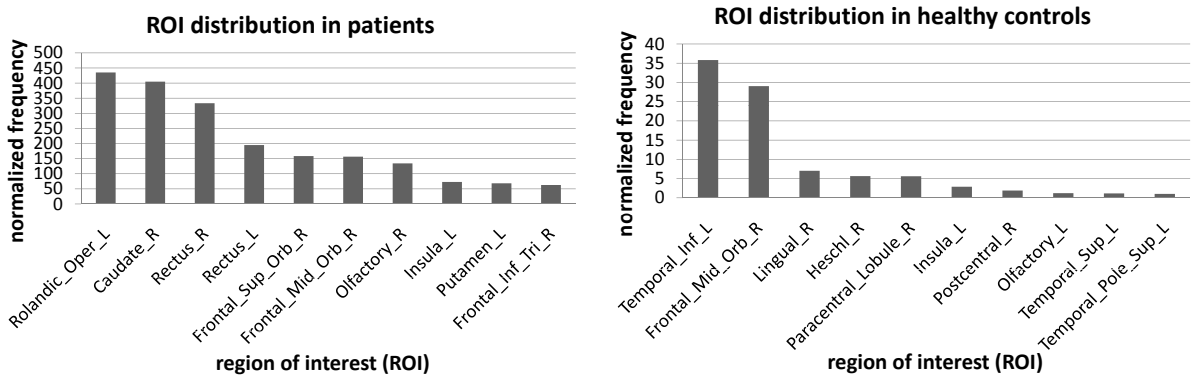


Fig. 3. Largest motifs found in the groups of somatoform patients and healthy controls.

Evaluation of ROIs. We found motifs that can discriminate very well between somatoform patients and controls. In the next step we determined the normalized frequencies of the ROIs in patients and controls, respectively. The results are depicted in Figure 4. Our results are consistent with a previous study [4]. They report different activation pattern in the regions *Insula_L* and *Frontal_Mid_Orb_R* of patients and controls. Our results of different activation of the parahippocampal cortex in patients and controls (not depicted in Figure 4) supports a recent study that suggests that patients with posttraumatic stress disorder showed also an altered activation pattern in the parahippocampal cortex in comparison to healthy controls when subjected to painful heat stimuli [3].



(a) patients

(b) healthy controls

Fig. 4. Frequencies of ROIs in motifs of patiens and controls respectively

In addition, we found that patients show increased activation in *Rolandic_Oper_L*, *Caudate_R* and *Rectus_R* whereas the control group is activated in the regions *Temporal_Inf_L*, *Heschl_R* and *Lingual_R* to a higher degree. Also the olfactory region shows alterations in the activation of patients and controls. Whereas *Olfactory_R* occurs to a much higher degree in motifs found in patients, motifs found in the networks of controls are labeled more often with *Olfactory_L*.

V. Conclusion

In this paper, we have applied an efficient algorithm for frequent subgraph discovery to time series data of 7 somatoform patients and 4 healthy controls. These motifs represent groups of brain compartments that covary in their activity during the process of pain stimulation.

We evaluate the appearance of motifs for both groups. Our results let suspect that somatoform brain disorder is caused by an additional pathogenous activity, not by a missing physiological activity.

So far we care about the topology in the network of interactions, but ignore the temporal order of these interactions. When studying network topology, it is important to bear in mind that the network models currently available are simplified models of the systems that govern cellular processes. While these processes are dynamic, the models we consider so far are all static. In future research, we will look into the temporal order of the motifs, thus we want to determine dynamic motifs.

Additional authors

Alexander v. Kalckreuth, Christian Sorg, Till Sprenger, Thomas R. Tölle, Michael Valet, Claus Zimmer, (Technical University of Munich, email: kalckreuth|c.sorg|sprenger|toelle|valet|claus.zimmer@lrz.tum.de)
and Harald Gündel (Medizinische Hochschule Hannover, email: psychosomatik@mh-hannover.de).

References

[1] D. J. Cook and L. B. Holder. Substructure discovery using minimum description length and background knowledge. *J. Artif. Intell. Res. (JAIR)*, 1:231–255, 1994.

[2] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Series of Books in Mathematical Sciences. W. H. Freeman, 1979.

[3] E. Geuze, H. Westenberg, A. Jochims, C. de Kloet, M. Bohus, E. Vermetten, and C. Schmahl. Altered pain processing in veterans with posttraumatic stress disorder. *Arch. Gen. Psychiatry*, 64:76–85, Jan 2007.

[4] H. Gündel, M. Valet, C. Sorg, D. Huber, C. Zimmer, T. Sprenger, and T. R. Tölle. Altered cerebral response to noxious heat stimulation in patients with somatoform pain disorder. *Pain*, 137(2):413–421, July 2008.

[5] A. Inokuchi, T. Washio, and H. Motoda. Complete mining of frequent patterns from graphs: Mining graph data. *Machine Learning*, 50(3):321–354, 2003.

[6] M. Kuramochi and G. Karypis. Frequent subgraph discovery. In *ICDM*, pages 313–320, 2001.

[7] M. Kuramochi and G. Karypis. Finding frequent patterns in a large sparse graph. In *SDM*, 2004.

[8] M. Kuramochi and G. Karypis. Grew-a scalable frequent subgraph discovery algorithm. In *ICDM*, pages 439–442, 2004.

[9] G. E. Sarty. *Computing Brain Activity Maps from fMRI Time-Series Images*. Cambridge University Press, 2007.

[10] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *NeuroImage*, 1(15):273–289, January 2002.

[11] S. Wernicke. Efficient detection of network motifs. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 3(4):347–359, 2006.

[12] S. Wessely, C. Nimnuan, and M. Sharpe. Functional somatic syndromes: one or many? *Lancet*, 354:936–939, September 1999.

[13] X. Yan and J. Han. gspan: Graph-based substructure pattern mining. In *ICDM*, pages 721–724, 2002.

[14] X. Yan and J. Han. Closegraph: mining closed frequent graph patterns. In *KDD*, pages 286–295, 2003.