# Detecting Pathological Pathways of the Chronic Fatigue Syndrome by the Comparison of Networks

Frank Emmert-Streib[1], Earl F. Glynn[1], Christopher Seidel[1], Christoph L. Bausch[1] and Arcady Mushegian[1,2]
[1]Stowers Institute for Medical Research
Kansas City, 1000 E. 50th Street, MO 64110, USA
{fes|efg|cws|clb}@stowers-institute.org

[2]University of Kansas Medical Center
Kansas City, KS 66160, USA
arm@stowers-institute.org

## Abstract

In this paper we aim to identify biological processes affected by the chronic fatigue syndrome (CFS). So far, CFS has neither diagnostic clinical signals nor abnormalities that could be diagnosed by laboratory examinations. It is also unclear if the CFS represents one disease or can be subdivided in different categories. We use information from clinical trials, the gene ontology (GO) database as well as gene expression data to identify undirected dependency graphs (UDGs) representing biological processes according to the GO database. The structural comparison of UDGs of sick vs non-sick patients allows us to make predictions about the modification of pathways due to pathogenesis.

## Keywords

Chronic Fatigue Syndrome, Undirected Dependency Graph, Network Comparison, Microarray Data

## 1 Introduction

The chronic fatigue syndrome (CFS) is a disease that affects approximately one million people in the USA alone [9]. So far, the diagnosis and evaluation of CFS has been primarily based on the presence or absence of various symptoms over time [6] rather than, e.g., clear clinical signals [12]. Patients suffering from CFS experience persisting or relapsing fatigue of at least six month duration that is not substantially reduced by rest and causes a significant reduction in activities [8]. Currently, the definition of CFS is largely based on case studies such as Fukuda et al. [5].

In this article we do not aim to classify patients as sick (chronic fatigue) or non-sick nor we classify genes based on experimental data from various methodologies. Instead, we use this information as prior information to constrain our analysis, because we expect that the signature of this disease in experimental data of any kind will be vanishingly small due to our fundamental lack of understanding of this disease preventing an efficient design of experiments. This statement is confirmed by preceding studies of the CFS. Whistler et al. [12] found that microarray data could only be clustered meaningfully if the data of the patients were classified beforehand based on clinical data and the grouping of genes according to biological pathways. In this article we pick up this result and use clinical data that classify patients in various groups [8]. From these groups we use only the group of sick (CF) and non-sick people, because we

think that this distinction can be done reliably. We are aware of the speculation that the chronic fatigue disease is a hetero- rather than homogeneous illness which means that it is likely that the overall class of chronic fatigue patients can be further subdivided in smaller classes corresponding to different representatives of the CFS [12]. However, for our study this coarse classification is sufficient, because we do also not aim to detect these subclasses. In addition to the clinical data, we use the gene ontology (GO) database [2] as prior information to group genes according to their GO terms associating them with biological processes. For the resulting groups of genes, we construct an undirected dependency graph (UDG) from expression data of peripheral blood. This gives us, e.g., for the biological process of humoral immune response (GO:0006959) one UDG representing the stage 'sick' (CF) and one representing 'non-sick'. Our central hypothesis in this article is that the CFS has a signature detectable on a pathway level rather than on the expression level of single genes. More precisly, we suggest an extension of the concept of differentially expressed genes on a pathway level which means that the orchestra of expressed genes belonging all to a certain pathway is modified in response to the presence of a disease. Mathematically, we detect this modification by evaluating the similarity of the undirected dependency graphs representing pathways. We are aware of the fact, that the available data from microarray experiments may not be fully sufficient to reconstruct the molecular interactions of genes and their products. For this reason, the UDGs are an approximation of the true pathways capturing some but certainly not all structural information of the underlying interactions. Following Xiong et al. [**?**] we name these approximations quasi-pathways. The assumption behind this hypothesis is that in the simplest case the phenotype of an organism is affected by a single gene mutation that means there is a one-to-one correspondence between phenotype and a single gene. In general, however, this oversimplified view is wrong, because otherwise one could only observe $N$ different phenotypes, whereas $N$ corresponds to the number of genes of the genome of an organism. This implies, that combinations of genes and their controlled expression determine a phenotype. Due to the fact, that biological processes in general have a high intrinsic variability, not to confuse with a measurement error [4, 7], in contrast to, e.g., technical processes, it is further plausible that not only one expression pattern of genes results in a specific phenotype but multiple. This makes biological processes robust and stable against perturbations [4, 7] despite the sloppy functioning of its parts. This short excursion makes clear that the concept of differentially expressed genes loses its significance in the context of surjective phenotypes, which represent in our case the symptoms of CFS.

This article is organized in the following way. In the next section 2 we introduce our mathematical approach. Then we present in section 3 our results and conclude this article in 4 with a summary and conclusions.

# 2 Methods

The approach we suggest is based on three different sources of information. We use clinical data, gene expression data of peripheral blood mononuclear cells [1] and the gene ontology database.

## 2.1 Classification of patients

The clinical data provides us with a classification of the 227 patients participating in the study into nine categories [8]. From this study we use only data from patients from - chronic fatigue syndrome in the worst stage and non fatigue (NF) in the least stage. Loosely speaking, this corresponds to two classes - sick vs. non-sick. In the following we use this abbreviation to simplify the communication. We assume, that the distinction between people from either of these two classes can be done reliably provided the patients participated 'fair' in the sense that they answered the questionnaire correctly on which the classification was based.

## 2.2 Classification of genes

We obtain a meaningful classification of genes in biological processes by the GO database. GO is hi-

Table 1: Biological processes used in our analysis.

| GO term | name |
|---|---|
| GO:0006471 | prot. am. acid ADP-ribosylation |
| GO:0007219 | Notch signaling pathway |
| GO:0008360 | regulation of cell shape |
| GO:0042157 | lipoprotein metabolism |
| GO:0007126 | meiosis |
| GO:0006958 | complement activation, classical pathway |
| GO:0007222 | frizzled signaling pathway |
| GO:0006633 | fatty acid biosynthesis |
| GO:0043087 | regulation of GTPase activity |
| GO:0042742 | defense response to bacteria |
| GO:0001525 | angiogenesis |
| GO:0045892 | negative regulation of transcription, DNA-dependent |

erarchically organized in three major groups - biological process, molecular function or cellular component. Because we hypothesis that pathways should change significantly due to the influence of the disease we group genes according to biological processes they participate. Practically, this grouping is constrained by two factors. First, we are only looking for groups consisting of 20 to 50 genes, because the inference of meaningful connections between genes becomes more involved for larger groups [11]. Second, we are searching for gene groups that are present to a large extent on the microarray chip used. For this reason, we selected 12 biological processes fulfilling our conditions given in table 1.

## 2.3 Representing pathways as UDG

For the genes that are grouped according to the GO terms given in table 1 we calculate now the undirected dependency graph (UDG) with the expression profiles, $\mathbf{x}_i$ for $i \in \{1, \dots N\}$, of the genes from the microarray data. For example, we use the expression profiles of the genes from *regulation of GTPase activity* (GO:0043087) from the sick and non-sick patients according to section 2.1 and calculate two UDGs, one for each patient class.

**Definition 2.1 (UDG of first order)** *An UDG $G$ of first order is an undirected, unweighted graph with $N$ nodes (number of genes) that is obtained via the following procedure:*

1. *connect all nodes with an edge with each other*
2. *calculate the correlation between all profiles $x_i$*
3. *delete all edges connecting node $\mathbf{x}_i$ with $\mathbf{x}_j$ if $r_{\mathbf{x}_i\mathbf{x}_j} < \Theta_c$*
4. *calculate the partial correlation of first order for all triplets of nodes $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$ that have an edge between $\mathbf{x}_i$ and $\mathbf{x}_j$*
5. *delete all edges connecting node $\mathbf{x}_i$ with $\mathbf{x}_j$ if $r_{\mathbf{x}_i\mathbf{x}_j|\mathbf{x}_k} < \Theta_{pc}$*

The thresholds $\Theta_c$ and $\Theta_{pc}$ are obtained from randomization tests by generating random profiles $\hat{\mathbf{x}}_i$ by randomly assigning expression values to the components of $\hat{\mathbf{x}}_i$ from all available expression data. These thresholds correspond to $P$ values with $P = \alpha$. The significance level $\alpha$ was set to 0.05. For the correlation $r_{\mathbf{x}_i\mathbf{x}_j}$ we use Pearson correlation

$$r_{\mathbf{x}_i\mathbf{x}_j} = \frac{C_{\mathbf{x}_i\mathbf{x}_j}}{\sigma_{\mathbf{x}_i}\sigma_{\mathbf{x}_j}} \tag{1}$$

with the covariance $C_{\mathbf{x}_i\mathbf{x}_j} = E[(\mathbf{x}_i - \mu_{\mathbf{x}_i})(\mathbf{x}_j - \mu_{\mathbf{x}_j})]$. The partial Pearson correlation of first order is obtained by

$$r_{\mathbf{x}_i\mathbf{x}_j|\mathbf{x}_k} = \frac{r_{\mathbf{x}_i\mathbf{x}_j} - r_{\mathbf{x}_i\mathbf{x}_k} r_{\mathbf{x}_j\mathbf{x}_k}}{\sqrt{(1 - r^2_{\mathbf{x}_i\mathbf{x}_k})}\sqrt{(1 - r^2_{\mathbf{x}_j\mathbf{x}_k})}} \tag{2}$$

The reason, why we use partial correlation in addition is exemplified in Fig. 1. Suppose we measure the correlation between node $A$ and $C$ then we will receive for the upper situation a high correlation, because $A$ influences $C$ along some path depicted as dashed line. However, if along the path from $A$ to $C$ there is another node $B$ then $A$ does not influence $B$ directly. If we would estimate the underlying graph structure only based on correlation the resulting graph would have an additional edge from $A$ to $C$ not present in the graph in Fig. 1. In contrast, the partial correlation $r_{\mathbf{x}_A,\mathbf{x}_C|\mathbf{x}_B}$ is zero, because $B$ blocks the information flow. Hence, using partial correlation of first order is a first step to estimate the underlying causal structure of interactions [10]

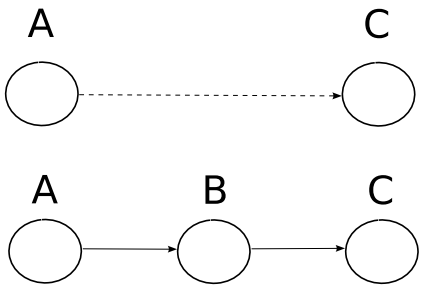


Figure 1: Visualization of direct and indirect cause from a hypothetical gene $A$ to gene $C$.

### 2.4 Similarity of pathways

Finally, we need to define a similarity measure to evaluate the modifications of UDGs from a sick vs a non-sick pathway. Due to the fact, that we compare only graphs with the same order (number of nodes) whose nodes are labeled (names of genes) we use the well known *graph edit distance* by Bunke [3]. In this special case this measure counts the number of edge deletions and insertions one needs to transform, e.g., graph $G_A$ into graph $G_B$. Again, to decide if two graph are significantly un-similar we calculate numerically a distribution of similarity values of random UDGs based on randomized profiles from the microarray data and obtain a threshold $\Theta_s$ corresponding to a $P$ value with $P = \alpha$. Similarity values larger than $\Theta_s$ have $P$ values less than $\alpha$ and, hence, are statistically significant indicating to reject the null hypothesis.

## 3 Results

The graph edit distances for the biological processes in table 1 are shown in Fig. 2. The results are ranked in descending order. The horizontal dashed line corresponds to a $P$ value of $\alpha = 0.05$. Graph edit distances above this line are significant with a $P$ value less than $\alpha$. One can clearly see, that the first bar is much higher than all other bars and the threshold $\Theta_s$. The biological process of this pathway is protein amino acid ADP-ribosylation (GO:0006471).

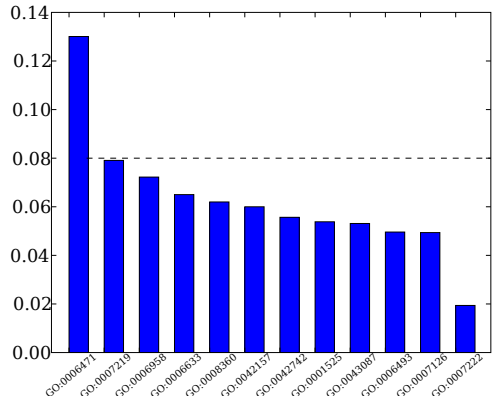


Figure 2: Similarity values of networks obtained by the comparison of the same pathway from sick vs non-sick patients. The similarity between the networks was calculated with the graph edit distance [3]. The dashed line corresponds to a $P$ value of 0.05.

## 4 Conclusions

In this paper we introduced a novel method to detect differences in biological pathways of a sick vs non-sick organism. We exemplified our method on expression data from peripheral blood mononuclear cells from patients suffering from chronic fatigue syndrome and control patients who are healthy. The basic idea of our method consists in the estimation of an 'interaction strength' between genes corresponding biologically to a causal influence the genes have on each other. Because the biological processes under investigation are complex we did not use the absolute value of the estimated strengths but simplified the situation by allowing only two values of an edge - zero or one. This conservative approach allows us to calculate numerically $P$ values for the rejection or acceptance of an edge. That means, we did not aim to estimate the real networks representing biological pathways, but quasi pathways representing some but certainly not all biological interactions present or absent in a living cell. By comparing undirected dependency graphs of the same biological quasi pathway from chronic fatigue and control patients we found that protein amino acid ADP-ribosylation (GO:0006471) is signif-

icantly changed due to the influence of the disease.

We want to emphasize that we did not aim to detect the differentially expression of single genes between chronic fatigue and healthy people, but to detect modifications due to the pathogenesis of complete pathways the genes participate. This conceptual understanding was already introduced in a similar form by XIONG et al. [13], however, by using a different mathematical framework. We believe that complex diseases, as the chronic fatigue syndrome, needs to be understood on the systems level which can be represented as networks rather than on the level of single genes and we are already curious about the outcome of further experiments using our prediction, about the involvement of the protein amino acid ADP-ribosylation pathway in the chronic fatigue syndrome, as starting point.

# 5 Acknowledgments

We would like to thank Galina V. Glazko and Piotr Kozbial for fruitful discussions and Malcolm Cook and Dan Thomasset for computer support.

# References

[1] Camda 2006 conference datasets. 2006.

[2] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1):25–29, May 2000.

[3] H. Bunke. On a relation between graph edit distance and maximum common subgraph. *Pattern Recogn. Lett.*, 18(9):689–694, 1997.

[4] H. B. Fraser, A. E. Hirsh, G. Glaever, J. Kumm, and M. B. Eisen. Noise minimization in eukaryotic gene expression. *PLOS Biology*, 2(6):e137, 2004.

[5] K. Fukuda, S. E. Straus, I. Hickie, M. C. Sharpe, J. G. Dobbins, and A. Komaroff. The chronic fatigue syndrome: A comprehensive approach to its definition and stydy. *Annals of Internal Medicine*, 121:953–959, 1994.

[6] L. A. Jason, C. P. King, E. L. Frankenberry, and K. M. Jordan. Chronic fatigue syndrome: Assessing symptoms and activity level. *Journal of Clinical Psychology*, 55(44):411–424, 1999.

[7] H. H. McAdams and A. Arkin. Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. USA*, 94:814–819, 1997.

[8] W. C. Reeves, D. Wagner, R. Nisenbaum, J. F. Jones, B. Gurbaxani, L. Solomon, D. A. papanicolaou, E. R. Unger, S. D. Vernon, and christine Heim. Chronic fatigue syndrome - a clinically empirical approach to its definition and study. *BMC Medicine*, 3:19, 2005.

[9] K. J. Reynolds, S. D. Vernon, E. Bouchery, and W. C. Reeves. The economic impact of chronic fatigue syndrome. *Cost effectiveness and resource allocation*, 2(4), 2004.

[10] B. Shipley. *Cause and Correlation in Biology.* Cambridge University Press, 2000.

[11] P. J. Waddell and H. Kishino. Cluster inference methods and graphical models evaluated on nci60 microarray gene expression data. *Genome Informatics*, 11:129–140, 2000.

[12] T. Whistler, E. R. Unger, R. Nisenbaum, and S. D. vernon. Integration of gene expression, clinical and, epidemiologic data to characterize chrinic fatigue syndrome. *Journal of Translational Medicine*, 1:10, 2003.

[13] M. Xiong, J. Li, and X. Fang. Identification of genetic networks. *Genetics*, 166:1037–1052, 2004.