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] Arcady Mushegian[1,2]

[1]Stowers Institute for Medical Research

[2]University of Kansas School of Medicine

7th June 2006

Outline

- CFS has no diagnostic clinical signs or laboratory abnormalities
- CFS is defined by symptoms and disability
- It is unclear if CFS represents single disease

- CFS has no diagnostic clinical signs or laboratory abnormalities
- CFS is defined by symptoms and disability
- It is unclear if CFS represents single disease

- CFS has no diagnostic clinical signs or laboratory abnormalities
- CFS is defined by symptoms and disability
- It is unclear if CFS represents single disease

- Characterize (define) CFS by clinical data + questionnaire
- microarray + clinical data $\Longrightarrow$
  (classify patients by clinical data, clustering, differentially expressed genes)
  heterogeneous illness & fundamental metabolic perturbations
  WHISTLER et al. 2003

Hypothesis

Pathways are important rather than 'genes'
$\Longrightarrow$ differentially expressed pathways, M. XIONG 2004

Questions

1. How to define pathways?
2. How to identify pathways?
3. How to compare pathways?

Definition

A pathway (directed graph) is an interconnected group of genes (variables) that regulates a biological process

Definition

A biological process is (hierarchically) defined by GO (gene ontology) terms

Definition

A pathway (directed graph) is an interconnected group of genes (variables) that regulates a biological process

Definition

A biological process is (hierarchically) defined by GO (gene ontology) terms

Used data

- Clinical Data (questionnaire + blood) $\Longrightarrow$ classify patients
- Gene Expression (peripheral blood mononuclear cells)
- GO database $\Longrightarrow$ classify genes

$\Longrightarrow$ reconstruct quasi-pathways (biological subprocesses)

Why quasi-pathways?

Central Dogma of Molecular Biology

- DNA - CHIP-chip
- RNA - microarray
- Protein - proteomics

Only partial information is used (available) to reconstruct the network

Why quasi-pathways?

Central Dogma of Molecular Biology

- ~~DNA - CHIP-chip~~
- RNA - microarray
- ~~Protein - proteomics~~

Only partial information is used (available) to reconstruct the network

Assumption

Patients participating are 'fair'

Result

Two groups of patients (classification)

1. non-sick
2. sick (chronic fatigue syndrome)

Assumption

GO database is correct (mega experiment)

Result

N groups of genes for N different biological processes (classification)

GO is a hierarchical database

- molecular function (7460)
- cellular component (1533)
- biological process (9384)

18377 GO terms

Examples of biological (sub)processes:

- regulation of cell cycle, GO:0000074
- DNA repair, GO:0006281
- circadian rhythm, GO:0007623
- endocytosis, GO:0006897
- ATP metabolism, GO:0046034

Examples of biological (sub)processes:

- regulation of cell cycle, GO:0000074, 791
- DNA repair, GO:0006281, 538
- circadian rhythm, GO:0007623, 44
- endocytosis, GO:0006897, 225
- ATP metabolism, GO:0046034, 14

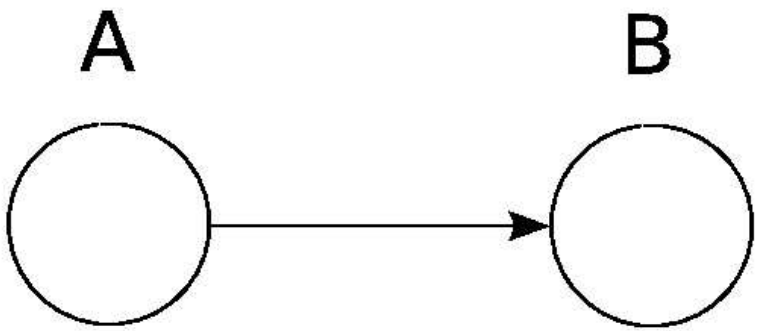
Expected disorder in biological processes

- immune cell activation , GO:0045321, 36
- positive regulation of apoptosis, GO:0043065, 42
- positive regulation of transcription, GO:0045941, 101
- circadian rhythm, GO:0007623, 44

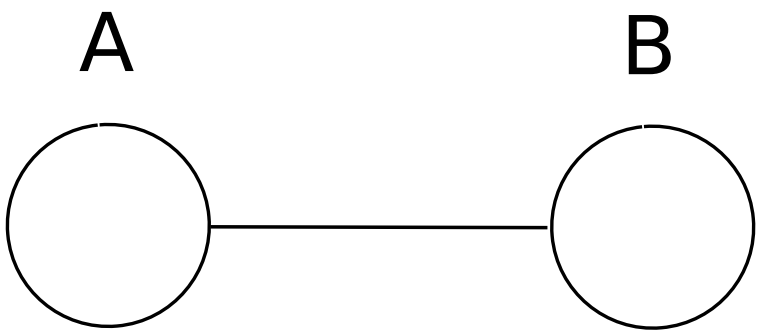
Expected order in biological processes

- housekeeping pathways, ???

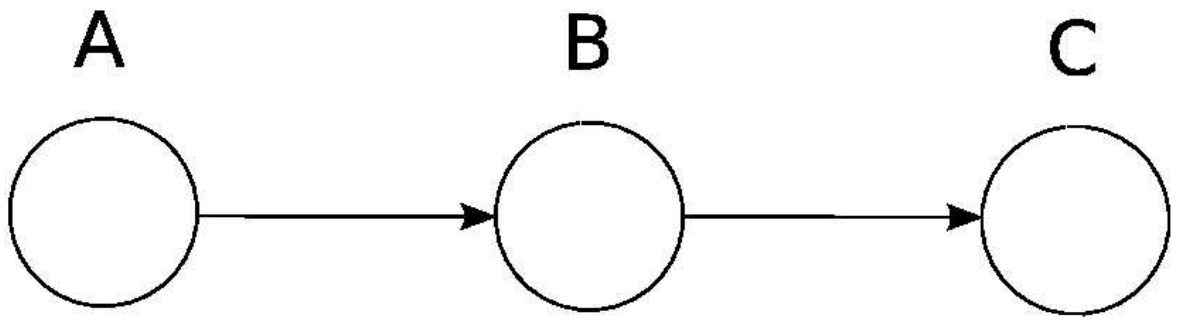
- correlation $\rho_{AC} \uparrow \Longrightarrow$ edge between A and B
- temporal ordering $\Longrightarrow$ direction



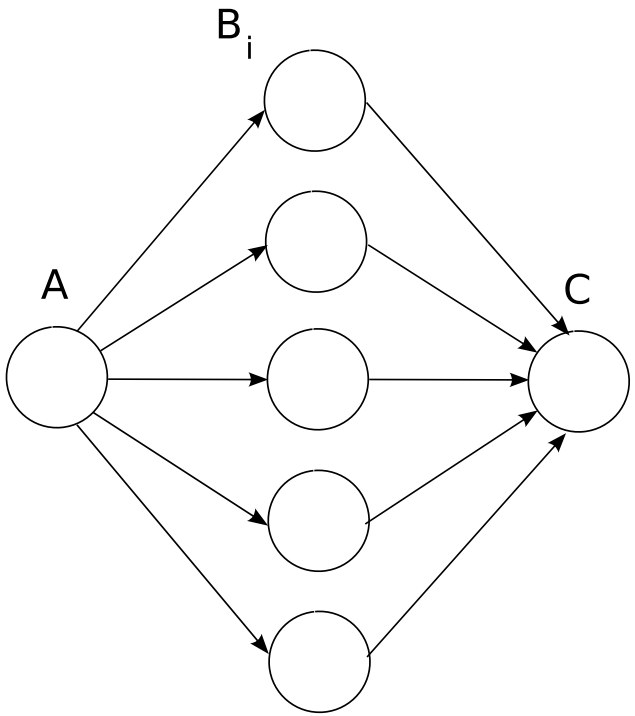
- correlation $\rho_{AC} \uparrow \Longrightarrow$ edge between A and B
- temporal ordering $\Longrightarrow$ direction



- correlation does not imply causality: $\rho_{AC} \uparrow$
- partial correlation of first order: $\rho_{AC.B} \downarrow$



- correlation does not imply causality: $\rho_{AC} \uparrow$
- partial correlation of first order: $\rho_{AC.B_i} \uparrow$
- partial correlation of higher order: $\rho_{AC.\{B_i\}} \downarrow$ (parallel pathways)



- correlation does not imply causality: $\rho_{AC} \uparrow$
- partial correlation of first order: $\rho_{AC.B_i} \uparrow$
- partial correlation of higher order: $\rho_{AC.\{B_i\}} \downarrow$

Example

N = 50, $n = |\{B_i\}| = 8$

$$\binom{N}{n} \sim 10^8$$

d-separation

$$x \perp\!\!\!\perp y | \{B_i\} \iff \rho_{xy.\{B_i\}} = 0 \tag{1}$$

VERMA et al. 1988, PEARL 1988, GEIGER et al. 1990, SPIRTES et al. 1998

variance

$$\sigma_x = E[(X - \mu_x)^2] \tag{2}$$

covariance

$$\sigma_{xy} = E[(X - \mu_x)(Y - \mu_y)] \tag{3}$$

Pearson correlation

$$\rho_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}} \tag{4}$$

partial Pearson correlation

$$\rho_{xy|z} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{(1 - \rho_{xz}^2)(1 - \rho_{yz}^2)}} \tag{5}$$

Definition (Undirected dependency graph (UDG) of first order)

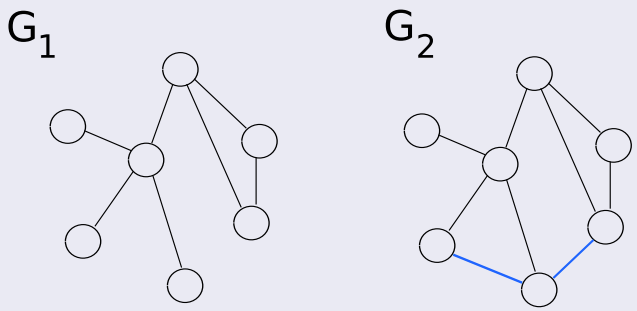
An UDG G of first order is an undirected, unweighted graph with N nodes (number of genes) that is obtained via the following procedure:

1. connect all nodes with an edge with each other
2. calculate the correlation between all profiles $\mathbf{x}_i$
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 $r_{\mathbf{x}_i\mathbf{x}_j|\mathbf{x}_k} < \Theta_{pc}$

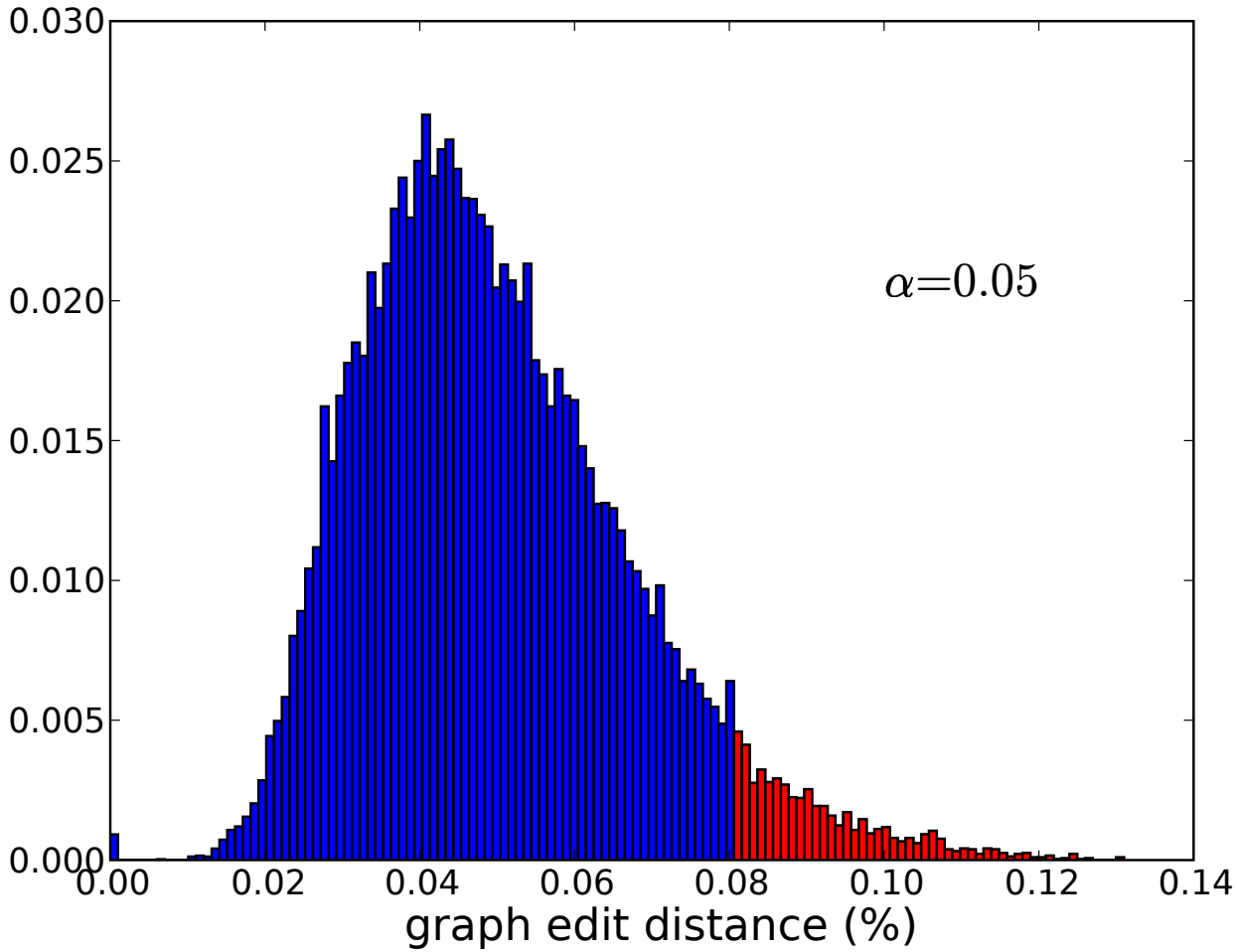
similar to PC-algorithm SPIRTES et al. 1991

Graph Edit Distance

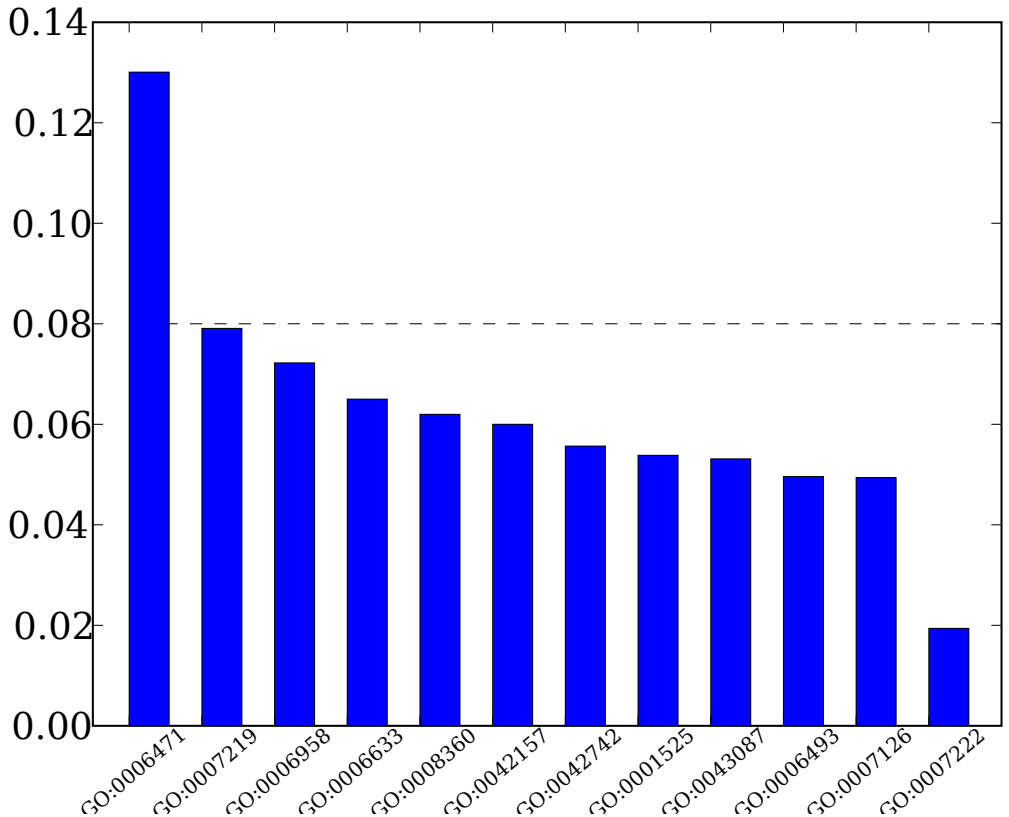
- Minimal number of edge deletions/insertions to transform graph G_1 to G_2
- Quasi-pathways:
  - compare only sick vs non-sick pathways $\Longrightarrow$ same number of genes
  - nodes are labeled (genes)



BUNKE 1997


0.030
0.025
0.020
0.015
0.010
0.005
0.000
α=0.05
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
graph edit distance (%)

GO term	name
GO:0006471	protein amino acid ADP-ribosylation (31)
GO:0007219	Notch signaling pathway (28)
GO:0008360	regulation of cell shape (22)
GO:0042157	lipoprotein metabolism (20)
GO:0007126	meiosis (36)
GO:0006958	complement activation, classical pathway (30)
GO:0007222	frizzled signaling pathway (19)
GO:0006633	fatty acid biosynthesis (37)
GO:0043087	regulation of GTPase activity (40)
GO:0042742	defense response to bacteria (32)
GO:0001525	angiogenesis (45)
GO:0006493	protein amino acid O-linked glycosylation (25)



GO:0006471 protein amino acid ADP-ribosylation
GO:0007219 Notch signaling pathway

Summary

- gene network represents biological (sub)process (pathway)
- comparison between normal (non-sick) and perturbed (sick) organism is reduced to the comparison between networks representing the corresponding biological processes
- conceptual generalization of differentially expressed genes to 'differentially' expressed biological processes (quasi-gene networks, M. XIONG et al. 2004)
- predicted pathways involved in CFS:
  GO:0006471 protein amino acid ADP-ribosylation
  GO:0007219 Notch signaling pathway

Acknowledgments

- Malcolm Cook
  Bioinformatics
  Stowers Institute for Medical Research, USA
- Matthias Dehmer
  Center for Integrative Bioinformatics
  Max F. Perutz Laboratories, Austria
- Galina V. Glazko
  Department of Biostatistics and Computational Biology
  University of Rochester, USA
- Daniel Thomasset
  Bioinformatics
  Stowers Institute for Medical Research, USA