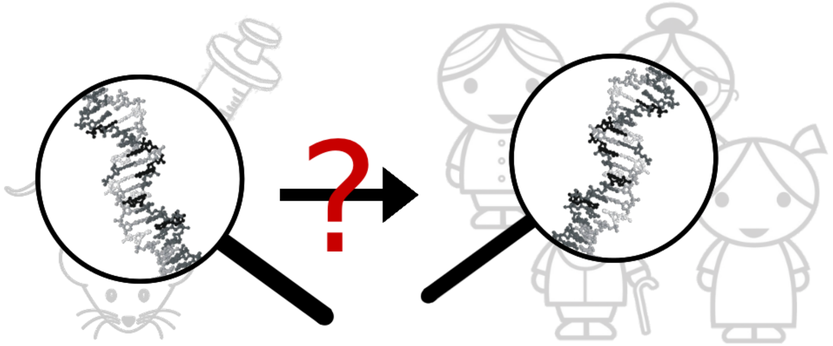
Cross-Species Alignment of Coexpression Networks

M.E. van der Wees[1,2], M. El-Kebir[1,2], U.K. Nandal[3], P.D. Moerland[3], A.H.C. van Kampen[3], J. Heringa[2], G.W. Klau[1]

1. Centrum Wiskunde & Informatica (CWI), Life Sciences Group, Amsterdam
2. VU University Amsterdam, Centre for Integrative Bioinformatics VU (IBIVU), Amsterdam
3. Academic Medical Center, Department of Clinical Epidemiology, Biostatistics and Bioinformatics (KEBB), Amsterdam

Biological Introduction

For both ethical and practical reasons, rodent models are commonly used to study human diseases. Experimental results, however, regularly fail to predict outcomes in humans. In this work we want to identify gene modules that are transferable across mouse and human.

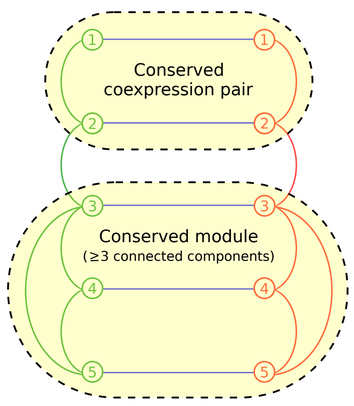


Data Networks

We use microarray data of individual versus pooled healthy liver samples in human and mouse. Correlations between expression profiles reflect whether genes might be regulated by a common mechanism. The coexpression networks contain 10185 and 16673 genes for mouse and human, respectively.

Bit scores from all against all BLAST alignments make up a third network. Depending on the score model, bit scores contribute to the score (continuous) or only generate a matching edge candidate list (discrete).

Conserved Modules

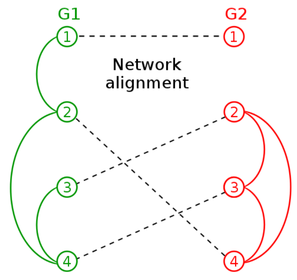


Modules of coexpressed genes that exist in both mouse and human indicate transferability. A conserved coexpression pair in the network alignment consists of two aligned gene pairs with correlation values above a threshold. Three or more coexpression pairs sharing genes make a conserved module.

Biological Validation

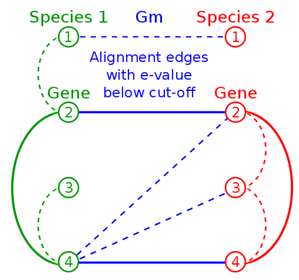
As a measure of quality, we calculate a Gene Ontology commonality score for each aligned gene pair *ik*. This score is based on the number of GO terms shared by both genes *i* and *k*, corrected for frequency based on hierarchy of the GO terms. Terms inferred by sequence are not taken into account.

Network Alignment



Given graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, alignment $a : V_1 \to V_2$ is a partial injective function from V_1 to V_2. Each possible alignment can be given a score $s(a)$, and the goal is to find the highest scoring alignment $a^* = \arg\max s(a)$.

In coexpression network alignment, nodes V_1 and V_2 represent genes, and edges E_1 and E_2 have coexpression values above threshold t. A macthing graph $G_m = (V_1 \cup V_2, E_m)$ has edges labeled with sequence alignment bitscores.



Scoring Function

Let $i, j \in V_1$ and $k, l \in V_2$. The score of alignment a is an addition of the sequence similarity score of genes i and k, and the coexpression similarity score of gene pairs ij and kl.

$$s(a) = (1-\beta) \cdot \sum_{i,k} b_{ik} x_{ik} + \beta \cdot \sum_{\substack{i,j \\ i<j}} \sum_{\substack{k,l \\ k \neq l}} w_{ikjl} x_{ik} x_{jl}$$

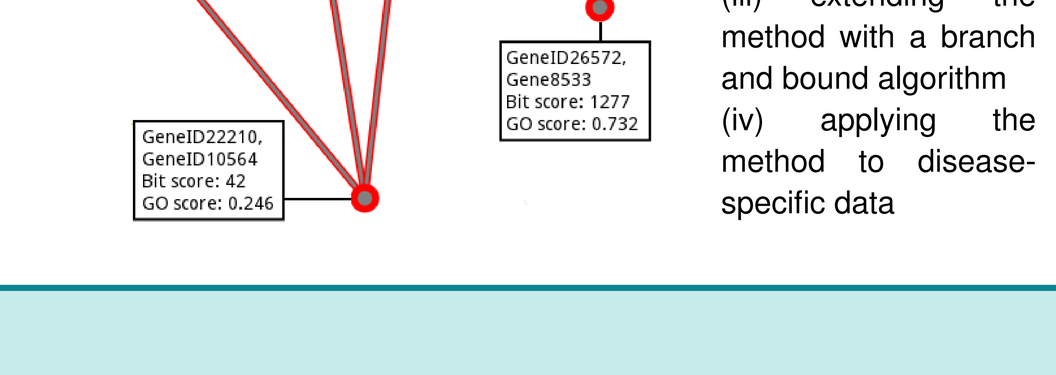
We used two score models:

(i) A discrete model with $\beta = 1$ and $w_{ikjl} = \{-1, 0, 1\}$, where bit scores are used to generate a matching edge candidate list

(ii) A continuous model with variable β, w_{ikjl} increasing for increasing coexpression values and b_k as normalised bit scores

Lagrangian Relaxation

Maximisation of the score is subject to matching constraints. Lagrangian relaxation of one or more constraints simplifies the problem by expanding the search space. Solutions to the relaxed problem *RP* are often infeasible for the problem *P*, but can be used as bounds to the optimal solution.



Preliminary Results and Future Work

Left: the continuous score model scores slightly better than the discrete model in terms of average GO scores for varying e-value cut-offs. Centre: gaps between the Lagrangian bounds are close to zero for all but the largest networks. Right: a graphical example of a conserved module, with edges corresponding to conserved coexpressions.

Future work includes:
(i) comparing GO scores with GO scores obtained from random network alignment
(ii) identifying non-transferable modules
(iii) extending the method with a branch and bound algorithm
(iv) applying the method to disease-specific data