k-PathA: *k-shortest* Path A*lgorithm*

(pronounce as "Qué Pasa")

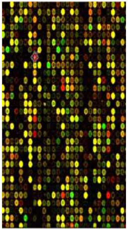
Alexander Ullrich and Christian V. Forst

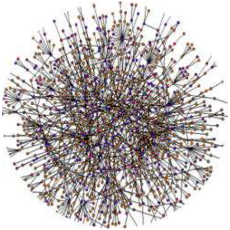
Chair for Bioinformatics
University of Leipzig

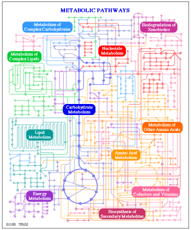
Department of Clinical Sciences
University of Texas Southwestern Medical Center

HiBi 2009, Trento, October 14-16

The Situation

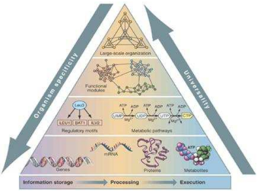


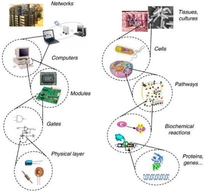




- Situation: Tons of available data on different levels
- Challenge: Derive useful information of structure and behavior of the biological systems

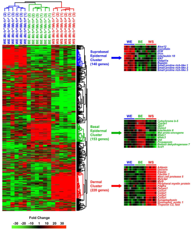
The Systems-Biology Approach





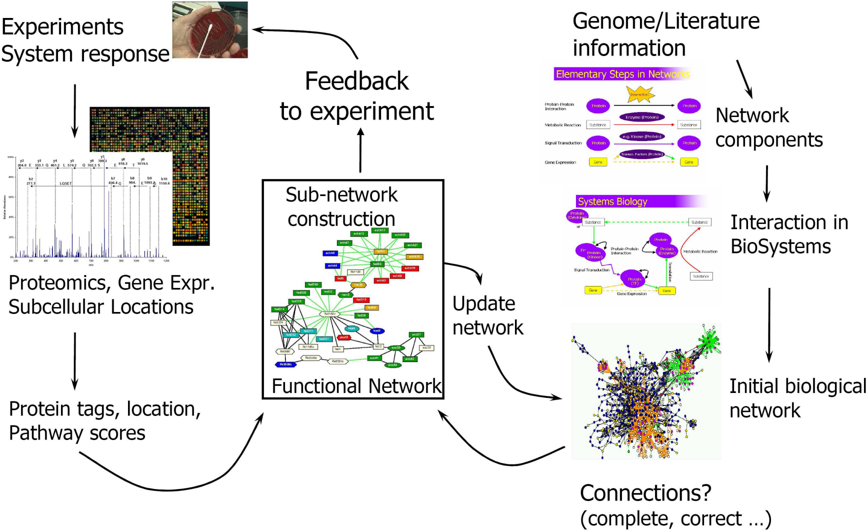
- Before SB: extensive research on single components
- SB says: interactions between components are more important
- SB looks beyond single genes, proteins, etc..
- SB looks at networks because they can give more insights
- SB combines data from different levels, components, conditions

Gene Expression



- Cell response to external conditions
- Expression Change = Control vs Experiment
- Biomarkers = Genes with high Expression Change
- Correlated genes are likely to share function
- Gene expression profiling, Gene clustering show some success
- But ...

Response Networks



Experiments
System response

Proteomics, Gene Expr.
Subcellular Locations

Protein tags, location,
Pathway scores

Feedback
to experiment

Sub-network
construction

Functional Network

Update
network

Genome/Literature
information

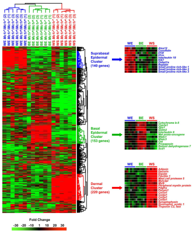
Network
components

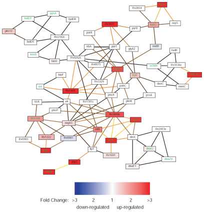
Interaction in
BioSystems

Initial biological
network

Connections?
(complete, correct ...)

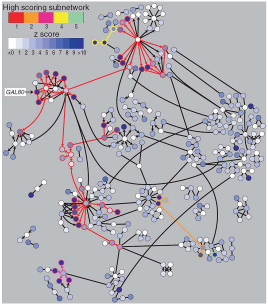
Gene Clusters vs Response Networks





- Genes in Response networks do not have to be active under all conditions
- Response networks underlie the constraints of the molecular interaction network
- Also non-correlated Genes can be included if they connect correlated genes
- Response networks of different conditions can be combined (response networks of new drugs can be compared to those of known drugs)

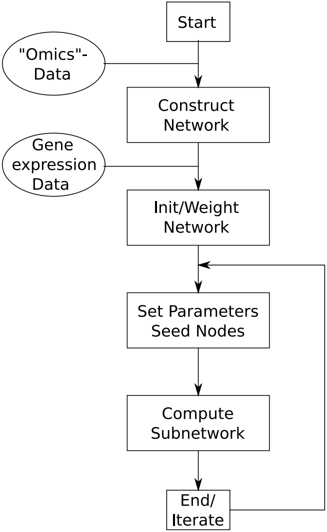
Response Networks - Computation



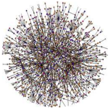
- Generated Sub-Networks have to be scored and compared against random Sub-Networks
- Computationally costly - Heuristics
- Still costly for large Networks - parallel Heuristics

k-PathA Approach - Overview

- Construction of Network
- Choice of Seednodes
- Generation of Sub-Network by spanning k-shortest Paths
- Scoring Sub-Networks
- Optimizing - Iterating

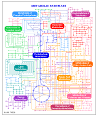


Network Construction



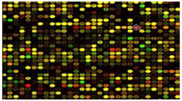
Nodes and Edges

+



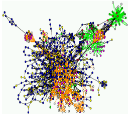
Edges

+



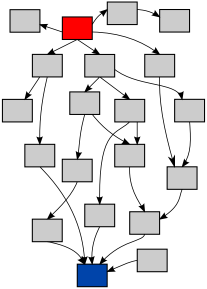
Node- and Edge-Weights

=



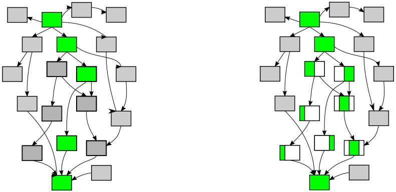
Network

Subnetworks - Preperation



- Choice of the Sub-Network borders = Seed Nodes
- Option 1: User Input
- Option 2: Randomly
- Option 3: Set of Nodes with highest weights
- In the Optimization step the set is changed

Sub-Networks - Generation

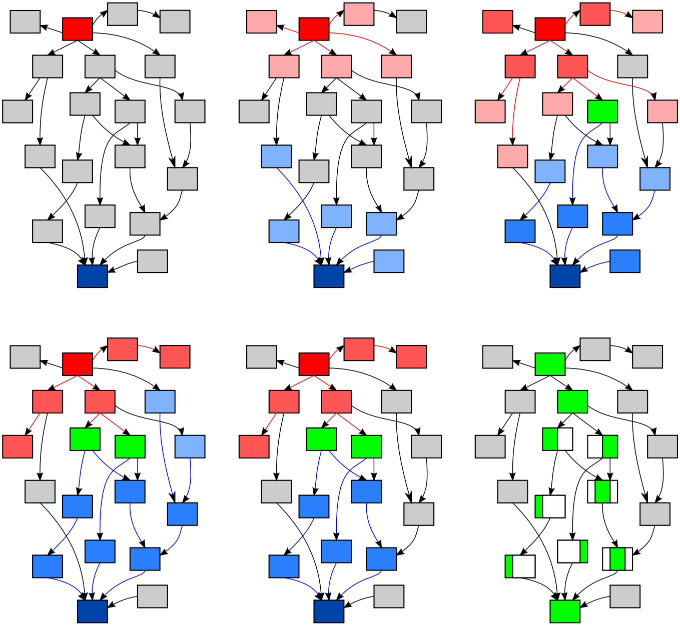


Definition

The k-th shortest simple path problem consists of the determination of a set of simple (loopless) paths between two nodes $P^k = \{p_1, p_2, \dots, p_k\} \subseteq P$, such that $\forall p \in P - P^k \wedge p^k \in P^k : cost(p^k) \leq cost(p)$.

- Spanning k-shortest simple (loopless) Paths
- Forward-, Backward Chaining Approach = Search from both Seeds
- Combination of Paths from both sides

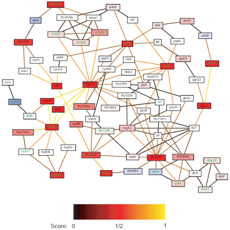
Sub-Networks - Generation

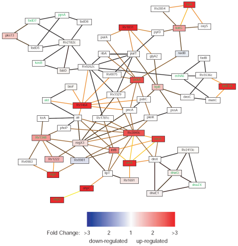


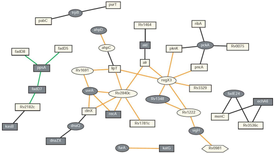
Sub-Networks - Scoring

- Expression values are normalized
- Nodeweights are ≥ 0
- < 1 down-regulated
- > 1 up-regulated
- $LinkCost(l_{x,y}) = \log(|x + y|)$
- $PathwayCost(p) = \frac{\sum_{l \in L_p} LinkCost(l)}{|L|^k}$

Sub-Networks - Combination

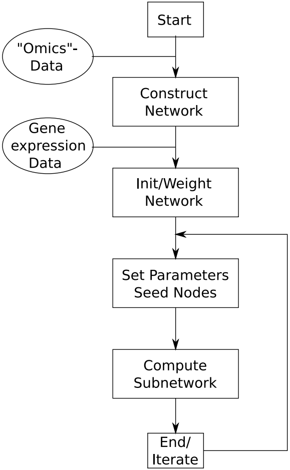






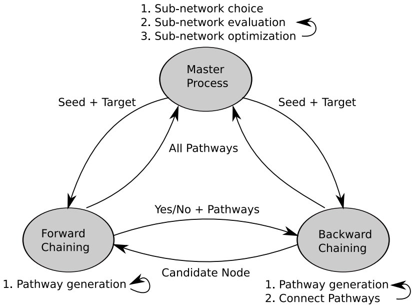
- Combine Sub-Networks for different Conditions
- Can show Similarities of Drug Responses
- $PC(X,Y) = \frac{(E(XY) - E(X)E(Y))}{\sqrt{Var(X)Var(Y)}}$
- $FPC(X,Y) = PC(X^{(2,n)}Y^{(1,n-1)})$
- $BPC(X,Y) = PC(Y^{(2,n)}X^{(1,n-1)})$

Sub-Networks - Optimization



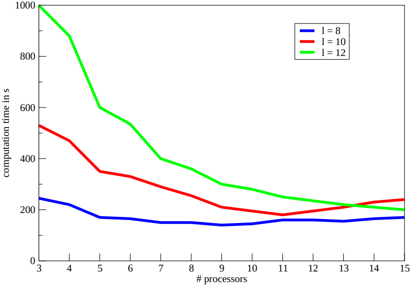
- Change of Seed nodes in every iteration
- Genetic algorithm: e.g. hill climbing

Parallel Implementation



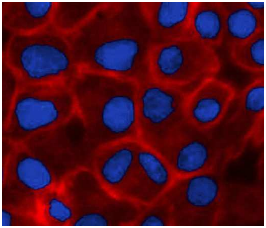
- Communication through message passing
- No shared Memory necessary
- 3 types of Nodes
- Master - Sub-Network Handling
- Forward - Path search and Common Node Checking
- Backward - Path search and Combination of Paths

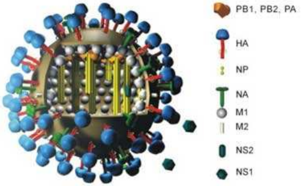
Parallel Implementation Behavior



- For larger Networks or high k and l: it scales linear
- For small Networks: communication is big part

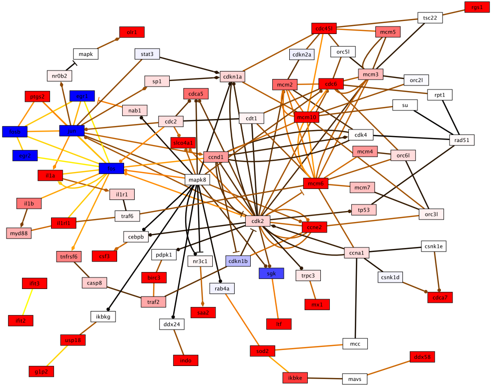
Application





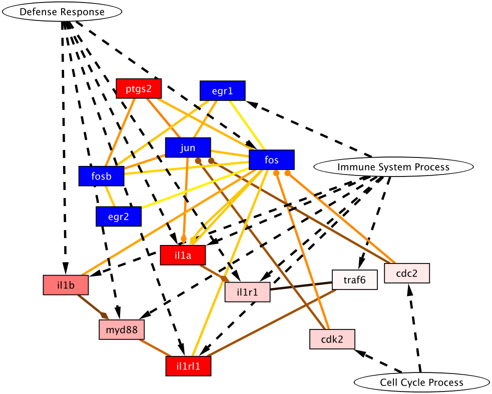
- Human-hybrid Network: 45,041 nodes + 438,567 edges
- Human bronchial cells
- Influenza (H5N1) infection (24h)
- Control: Mock infection (24h)
- $k = 3$, $l = 13$

Application



- Two well known processes affected by viral infections are present in the response network
- Cell-Cycle Genes (cdk2 and cdc's) are up-regulated
- Transcription factors known as early responders (fos and jun) are down-regulated

Application



- Comparison with other viral infection (LMCV, Djavani et al)
- Cell Cycle Process is modulated to induce apoptosis in the same way
- Immune Response is targeted similarly
- Moderate differences in the Host Response

Conclusions

- Computationally feasible through Parallel Implementation
- No shared Memory needed
- Flexible because new kinds of interactions can be introduced
- Flexible because of different scoring functions
- Response Networks for different conditions
- Comparison with other drug or viral responses possible

Acknowledgements

Christian V. Forst

Electra Sutton
Lawrence Cabusora

Money: Volkswagen Stiftung

For your Attention