Vladimir Rogojin | Keivan Kazemi | Krishna Kanhaiya | Eugen Czeizler | Ion Petre

NetControl4BioMed - Automatic discovery of combined drug therapy

Turku Centre for Computer Science

TUCS Technical Report
No 1162, June 2016

NetControl4BioMed - Automatic discovery of combined drug therapy

Vladimir Rogojin
Åbo Akademi University, Department of Computer Science
Domkyrkotorget 3, 20500 Turku, Finland
vrogojin@abo.fi

Keivan Kazemi
Åbo Akademi University, Department of Computer Science
Domkyrkotorget 3, 20500 Turku, Finland
kkazemi@abo.fi

Krishna Kanhaiya
Åbo Akademi University, Department of Computer Science
Domkyrkotorget 3, 20500 Turku, Finland
kkanhaiy@abo.fi

Eugen Czeizler
Åbo Akademi University, Department of Computer Science
Domkyrkotorget 3, 20500 Turku, Finland
eugen.czeizler@abo.fi

Ion Petre
Åbo Akademi University, Department of Computer Science
Domkyrkotorget 3, 20500 Turku, Finland
ipetre@abo.fi

TUCS Technical Report

No 1162, June 2016

Abstract

Recent results in network science have demonstrated that network control theory can lead to the development of novel therapeutic approaches for systemic diseases like cancer through the computational analysis of the structure of intracellular molecular interaction networks. These networks are a formal representation of relations between numerous components within cells that are used as a mean for formal holistic reasoning about biological structures. In particular, network controllability studies focus on discovering combinations of external interventions that can drive the biological system to a desired configuration. In practice, these studies can be translated into finding a combined multi-drug therapy in order to achieve a desired response from a cell. We develop a pipeline that finds a minimal set of nodes controlling a given set of targets within a network. The pipeline highlights those control nodes for which there are known FDA approved drugs. The network is generated automatically through quering of a number of pathway databases. The pipeline is deployed as an online web-service. We use these algorithms here to develop a bioinformatics pipeline and a web service that finds a minimal set of nodes that controls all the target nodes given by the user. Our pipeline queries a number of pathway databases and generates automatically molecular interaction networks. Then, it finds a minimal set of control nodes for the user-given set of target nodes. At user's request, the pipeline emphasizes the use of those control nodes for which there are known FDA approved drugs. We provide here both the source code of the pipeline as well as an online web-service based on this pipeline, with a web interface to interact with it.

Keywords: bioinformatics pipeline, combinatorial drug discovery, automatic discovery, network modeling, target network control, structural network control, intracellular molecular interaction networks, FDA approved drugs

TUCS Laboratory
Computational Biomodeling Laboratory (Combio Lab)

1 Scientific Background

Studies of biological networks by means of mathematical and computational modeling led to the development of innovative therapies and approaches in personalized medicine [3]. This is the reason for the high interest towards networks science among researchers in biology and medicine, focusing on understanding the dynamics and control features of various complex biological networks in association with matching experimental findings [5]. An efficient method to select a minimal set of driven nodes in a network in order to reach its full controllability (i.e., set of nodes through which via a finite number of cascading events one can control the behaviour of all the nodes in the network) was presented in [8]. However, it was shown through a number of computer-based experimental tests in [8] that in biological networks one may have to control as much as 80% of nodes of the whole gene-regulatory network in order to reach the full controllability. This makes the full network controllability impractical for biological and medical purposes. In many cases, it is more practical to control only a certain properly selected subset of the network's nodes (for instance, a disease-specific set of essential genes) in order to reach a desired overall behavior of the system [2]. This approach may lead, for instance, to an effective combined multi-drug therapy for a particular disease.

In particular, we focus here on directed biochemical interaction networks in human cells, where nodes are genes or proteins and directed edges represent such biochemical interactions as directed protein-protein interactions (activation/inhibition, phosphorylation, methylation, etc.) and gene expression regulatory interactions. Here, we consider sets of genes and proteins that are essential for malignant cells survival and proliferation [4] to be targets for the control. Our goal is to identify combinations of drug-target nodes that can lead to the control of these essential genes and proteins through cascading effects in our network. The mathematical background of this approach is based on network controlability and it is briefly explained in Section 2.1.

We build here a data analysis pipeline and its web-based front-end in order to provide a web-based service for automatic generation of combined multi-drug therapies suggestions. The core of the pipeline consists of the implementation of the algorithm proposed in [2] that for a given set of target nodes calculates a minimal set of driven nodes through which one can control the target nodes. Based on the user's query, the pipeline generates automatically intracellular chemical interaction networks by combining the interactions between genes, proteins and other intracellular components from various public pathway repositories. Then, the resulting networks are subjected to the structural controllability analysis in order to identify the minimal set of driven genes [2]. The data from public drug repositories is used to maximize the use of drug-targetable genes and proteins as driven nodes, to increase the practical applicability of the approach. The results of this analysis are returned to the user in form of reports in *PDF* documents, *XML* files and files readable by *Cytoscape*.

2 Materials and Methods

We built a pipeline that integrates a number of different software tools such as: pathway and drug data imports from a number of public databases, our structural network controllability algorithm developed earlier in [2], and the network visualization software for generating *PDF* files with the visual annotated representation of the molecular interaction network with its driven and drug target nodes.

2.1 Mathematical setup

We consider discrete time-invariant linear dynamical systems as models of biological entities (genes, proteins) influencing each other. Such a system can be modeled by

$$x_{t+1} = Ax_t + Bu_t, \qquad\qquad y_t = Cx_t,$$

where A, B, C are matrices of size $n \times n$, $n \times m$, and $l \times n$, respectively, $x_t \in \mathbf{R}^n$, $u_t \in \mathbf{R}^m$ and $y_t \in \mathbf{R}^l$ are the state vectors, input vectors and output vectors, for all $t \in \mathbf{N}$. Matrix A describes the interactions *within* the system under scrutiny, B describes the influence of the m *driver nodes* over the internal nodes of the system, while C describes the l *output nodes* as a function of the internal nodes of the system. We call *driven node* any $i \in \{1, \ldots, n\}$ such that $B_{ij} \neq 0$, for some $j \in \{1, \ldots, m\}$. In other words, a driven node is any internal node linked to an external driver node through matrix B. We say that an output vector $y \in \mathbf{R}^l$ is *reachable* from an initial state $x_0 \in \mathbf{R}^n$ if there exists a finite sequence of inputs $u_0, u_1, \ldots, u_t \in \mathbf{R}^m$ such that $y_t = y$.

In this paper, we focus on target controllability, i.e., on the case when the focus is on controlling some selected subset of the internal nodes of the system. To capture this case, we consider matrices C with $l \leq n$ and such that on each row of matrix C there is a non-zero value; this effectively selects the internal nodes of interest as outputs of the dynamical system. We say that such a system is *target controlable* if any output vector is reachable from any input state. It is said that a system is target controlable if and only if

$$rank[CB, CAB, CA^2B, \ldots, CA^{n-1}B] = l,$$

see [2] and references therein. A related notion is that of *structural target controllability*, that refers to a system that becomes target controlable by changing the non-zero values of A and B with some arbitrary non-zero values (we call such matrices *equivalent*). Moreover, a system is structurally target controlable if and only if it is target controlable for all equivalent matrices A and B. This allows the problem to be redefined as a graph-theoretical problem since the target controllability depends on the *structure* of the system and not on its numerical setup (levels of interactions within the system and levels of influence of driver nodes onto the system's internal nodes). Due to space restrictions we skip all these details here and refer to [2] and references therein. We only mention that the problem may be reduced to the following problem on directed graphs: given a directed graph $G = (V, E)$ with n nodes and a subset $T \subseteq V$ with l nodes, decide if there exists a set of l directed paths in G such that each node in T is an end point of one such path and no two paths intersect at the same distance from their end points, see [7]. In an additional refinement of the problem, one may also be given a subset $D \subseteq V$ of driven nodes and require that the directed paths preferably start from nodes in D.

2.2 Structural network control

Our pipeline is based on the algorithm proposed in [2]. This algorithm is aimed to minimize the size of the set of driven nodes that can be used to control a given set of target nodes. The algorithm uses several heuristics strategies for a more efficient exploration of the search space, which leads to faster and better (smaller sets of driven nodes) results in comparison to [3]. The Python implementation of the algorithm is in [1].

2.3 Workflow engine: Anduril

The pipeline is developed for the *Anduril* workflow framework [10]. Anduril is an open source component-based pipeline engine for scientific data analysis. Anduril defines an API that allows to integrate rapidly a vast range of existing software analysis and simulation tools and algorithms into a single data analysis pipeline. An Anduril pipeline represents a set of interconnected executable programs (called components) through well-defined I/O ports. Upon the termination of the execution of an Anduril component, its output results are delivered as inputs to the other (downstream) components by means of connecting the output port of the component to the input ports of its downstream components.

2.4 Data: Moksiskaan, DrugBank, Cancer cell lines

Our pipeline uses the *Moksiskaan* platform [6] to generate molecular interaction networks based on the user's query for the further analysis. Moksiskaan integrates pathways, protein-protein interactions, genome and literature mining data into comprehensive networks for a given list of genes and proteins (so-called "seed nodes"). It combines the relations between genes and proteins from different known pathways in order to address the fact that pathways crosstalk and influence each other. In our pipeline, Moksiskaan constructs a comprehensive network for the list of seed nodes by using and combining all imported pathways in the following manner: it connects all seed nodes by all known paths of length not exceeding a certain user-defined

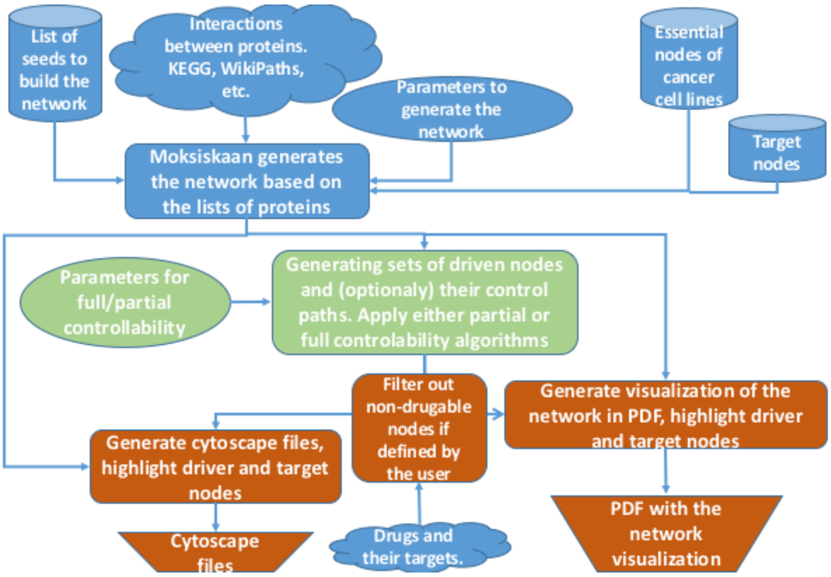
List of seeds to build the network
Interactions between proteins. KEGG, WikiPaths, etc.
Parameters to generate the network
Essential nodes of cancer cell lines
Target nodes
Moksiskaan generates the network based on the lists of proteins
Parameters for full/partial controllability
Generating sets of driven nodes and (optionaly) their control paths. Apply either partial or full controlability algorithms
Filter out non-drugable nodes if defined by the user
Generate visualization of the network in PDF, highlight driver and target nodes
Generate cytoscape files, highlight driver and target nodes
Cytoscape files
Drugs and their targets.
PDF with the network visualization

Figure 1: The general scheme of the *NetControl4BioMed* pipeline. The pipeline consists of three parts. In the first part we perform data input and preprocessing: we get from the user the list of seed nodes, the list of target nodes and possibly the predefined list of essential genes for a selected cancer cell line. Then, Moksiskaan generates the network. The second part of the pipeline deals with the network structural controllability analysis, where a minimal set of driven nodes is computed for the given set of target nodes. In the third part of the pipeline the post-processing is performed and the output is generated.

integer (so-called “gap” value). The intermediate genes and proteins from the paths need not necessarily belong to the given set of seed nodes.

The Moksiskaan platform defines a generic database schema to store the pathways from a number of different pathway databases and can be scaled to include the pathway data from new sources (such as new databases and user’s own data). Currently, Moksiskaan has built-in support for the integration of the pathway data from, among others, KEGG pathway database, Pathway Commons, and WikiPathways.

We use in our pipeline drug-target protein data from the open source DrugBank database. The DrugBank database combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information from bioinformatics and cheminformatics resources. For drug-target identifiers we have selected in total 1507 FDA approved drugs with known mechanisms.

In our pipeline, we provide the user with a number of predefined sets of target genes associated to some specific cancer cell lines. These target genes are cancer-specific essential genes. We have collected the data for all three types of cancer from the COLT-Cancer database [4]. In particular, we considered 29, 23 and 15 cell lines respectively for breast, pancreatic and ovarian cancer. The collected data follows the GARP (Gene Activity Rank Profile) and GARP-P value of corresponding proteins mentioned in the database. The previous studies showed that proteins with lower GARP score are more essential and directly associated with oncogenesis [9]. Therefore, we have selected only those essential genes whose GARP value is in the negative range, and moreover, whose GARP-P value is less than 0.05. Following the above criteria, we identified genes for breast, pancreatic and ovarian cancer respectively.

2.5 Pipeline

Our pipeline currently accepts the following inputs from the user:

1. **Seed genes**: List of genes and proteins that will be used as seed nodes by Moksiskaan to generate the network. This input can be any protein/gene ID of *Homo sapiens*.

2. **Cancer Cell Lines**: A cancer cell line whose set of essential genes will be used as target nodes for the network controllability algorithm. These nodes can act also as seed nodes if the user decides so. The user has also the option not to include any of the cell lines. However, in this case the next field should not be empty.

3. **Additional target genes**: A set of target nodes defined in addition to those in the "Cancer Cell Lines". This input can be left empty if the previous field is set to a cancer cell line. These nodes can act also as seed nodes if the user decides so.

4. **Gap**: The gap parameter used by Moksiskaan to generate the network.

5. **Include drug information**: Should the pipeline include also the drug target information for the driven nodes. If so, then the driven nodes for which there exist FDA approved drugs will be specifically highlighted in the output of the pipeline.

Our pipeline consists of the following three parts (see Figure 1):

1. **DATA IMPORT**: Integrate the user's defined input into the pipeline. Moksiskaan generates the network basing on the user's defined input. Target genes are imported for the specified cancer cell line if it is defined.

2. **NETWORK CONTROLLABILITY**: Compute the minimal set of driven nodes for the given target genes in the network generated by the Moksiskaan at the previous step.

3. **POSTPROCESSING AND OUTPUT**: Highlight those driven nodes that can be targeted by FDA approved drugs. Generate the network file (*GRAPHML*, *Cytoscape* and *PDF*) from the original network created by Moksiskaan and by adding additional annotations to

the nodes representing selected driven genes/proteins, drug-targetable driven genes/proteins if any and target genes. Generate CSV tables with the information about the driven genes/proteins, if they are drug-targetable and the list of target genes.

3 Results

The pipeline source code and the web-service based on it are available at [1]. The back-end of the service is the pipeline itself, while the front-end is its web-interface.

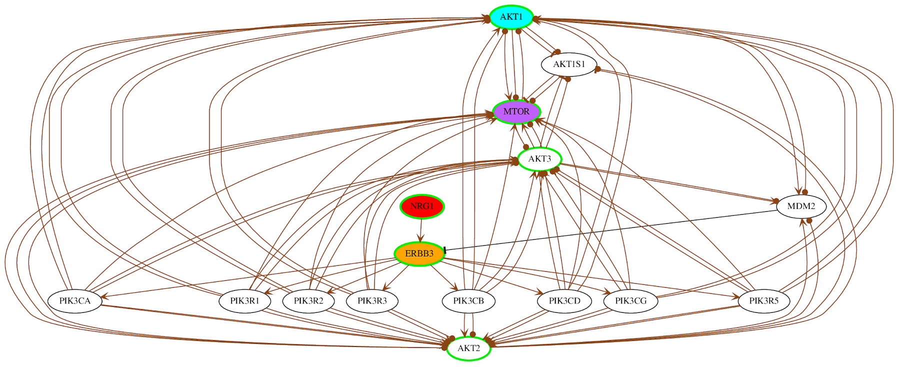

Figure 2: A network view generated by the pipeline. The nodes with green circles are “seed nodes”, light green nodes are drug target nodes which control essential genes. Red nodes are non-drug target nodes controlling essential genes. Purple nodes are essential genes controlled by drug-target nodes, and yellow are the genes controlled by non-drug target nodes.

The pipeline generates as the result of the computation a *zip*-archive with the following files. Table *driven.csv* contains the drug-targetable driven nodes and the number of cancer essential genes they control. Table *extra.csv* contains the non-targetable driven nodes (no FDA approved drug is known to be targeting the node) and the number of cancer essential genes they control. In *details.txt* the first line indicates the heuristics which was used for obtaining the result in the file. A blank line follows, then the names of the driven nodes, each on a separate line. After another blank line, the control path for each target is provided. File *graph.xml* contains the generated network and can be visualized in Cytoscape or Gephi. The archive also contains visualization of controlled graph from GraphML software as a *PDF*.

Figure 2 contains an example of a generated network with the drug-targetable and non-targetable driven nodes for some defined target nodes. We used *EGF*, *NRG1*, *ERBB3* and *MTOR* as seed nodes to generate the network. The pipeline discovered that target *MTOR* is controllable by *AKT1*, that can be targeted by an FDA approved drug. *ERBB3* is targeted by

NRG1, for which we have no known FDA approved drug. According to [4] *MTOR* and ERBB3 are associated to a number of different cancer subtypes.

4 Conclusion

The structural network controllability approach allows to get a better insight into a system represented as a directed graph: for a set of target nodes it is possible to identify a set of driven nodes through which one can control the target nodes by an external intervention through using the internal "wiring" of the network. We use here a recently developed algorithm [2] for structural targeted network controllability that identifies the minimal set of driven nodes for a user-given set of target nodes. We have demonstrated the practical applicability of this algorithm [1] through the development of the pipeline (that can be downloaded and installed as a stand-alone software) and of the related online service (i.e., a publicly available web interface for an instance of the pipeline installed on our servers). The pipeline performs an automatic generation of intracellular molecular interaction networks (by combining publicly available pathway data) and identification of driven nodes (that also can be targeted by FDA approved drugs) for a set of target genes/proteins defined by the user.

In this paper we address an interesting problem of using the controllability approach for combination of data on FDA approved drug targets and data on gene-essentiality for different types of disease. We anticipate that further developments on our pipeline have the potential in suggesting novel therapeutic strategies by using currently known drugs.

Acknowledgments

This work was supported by the Academy of Finland through grant 272451, and by the Finnish Funding Agency for Innovation through grant 1758/31/2016.

References

[1] COMBIO. Network controlability project. http://combio.abo.fi/research/network-controlability-project/, 2016.

[2] Eugen Czeizler, Cristian Gratie, Wy Kai Chiu, Krishna Kanhaiya, and Ion Petre. Target controllability of linear networks. Technical Report 1157, Turku Centre for Computer Science, 2016.

[3] Jianxi Gao, Yang-Yu Liu, Raissa M. D'Souza, and Albert-László Barabási. Target control of complex networks. *Nat Commun*, 5, 2014.

[4] Judice L. Y. Koh, Kevin R. Brown, Azin Sayad, Dahlia Kasimer, Troy Ketela, and Jason Moffat. COLT-cancer: functional genetic screening resource for essential genes in human cancer cell lines. *Nucleic Acids Research*, 40(D1):D957–D963, nov 2011.

[5] Walter Kolch, Melinda Halasz, Marina Granovskaya, and Boris N. Kholodenko. The dynamic control of signal transduction networks in cancer cells. *Nature Reviews Cancer*, 15(9):515–527, aug 2015.

[6] Marko Laakso and Sampsa Hautaniemi. Integrative platform to translate gene sets to networks. *Bioinformatics*, 26:1802–1803, 7 2010.

[7] Ching-Tai Lin. Structural controllability. *IEEE Transactions on Automatic Control*, 19(3):201–208, jun 1974.

[8] Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási. Controllability of complex networks. *Nature*, 473(7346):167–173, May 2011.

[9] Richard Marcotte et al. Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer Discovery*, 2(2):172–189, dec 2011.

[10] Kristian Ovaska et al. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Medicine*, 2(9):65+, September 2010.

TURKU
CENTRE *for*
COMPUTER
SCIENCE

Joukahaisenkatu 3-5 A, 20520 TURKU, Finland | www.tucs.fi

University of Turku
Faculty of Mathematics and Natural Sciences
- Department of Information Technology
- Department of Mathematics and Statistics

Turku School of Economics
- Institute of Information Systems Sciences

Åbo Akademi University
- Computer Science
- Computer Engineering

ISBN 978-952-12-3423-1
ISSN 1239-1891