Software implementations allowing new approaches toward data analysis, modeling and integration / curation of biological knowledge for Systems Medicine

Subtitle: Title may as well change!

John Zobolas

August 2019

Contents

List of Tables

List of Figures

Structure

Chapters are currently split as:

- Work I have done (see Chapter 1)
- Future plans (see Chapter 2). This includes the list of papers for my PhD.
- For more experimental/future ideas see Chapter 3.

Keywords

curation/knowledge management, VSM, causal statements, DrugLogics pipeline (model parameterization/calibration and prediction of synergistic drug combinations), biomarker analysis, synergy assessment

PhD Plan chart

In my effort to *combine the technologies within our group, the expertise from other people and of course my own work* to produce (research) results, I have come with the following action plan + papers for an organically grown thesis:

Literature => Curation => Causal Statements => Models => Predicting Synergies => Finding mechanisms

Why and when?

Phd Midterm

Sections of presentation:

- Background
- Description
- Publication Plan
- Progress

Chapter 1

PhD work

This is a summary of all the work that I have done in my PhD until now. (mainly it's about software implementations related to the core technologies within the group). **To include in the thesis text**.

Note though that not all of these will be part of the main thesis (maybe include the rest in a section like 'Funny PhD side-quests').

1.1 Pipeline

- Lots of refactoring to increase the readability, maintainability and extendability of the source code (complete restructure of classes, addition of others). This has **RRI extensions**, because cleaning and re-structuring software code has a social aspect to it in the sense that other people can now contribute more easily, extend the code, use it (user perspective can bring changes and further improvements to software pipeline even though they may be used for research purposes) - how can you expect users to actually use a piece of code when it's not substantially documented and it's internal logics made obscure because nobody gave attention to detail and structure? How can anybody care for a (software and any) product that you have not cared enough so as to present it in an way that is acceptable, managable and proper?

- Bug fixing
- Enable maven packaging for easier source compilation, testing, installation, management and executing of the code
- Added tests to modules `gitsbe` and `drabme` using JUnit5, mockito and assertJ libraries
- Source code documentation + proper README files on `gitsbe`, `drabme` and `druglogics-synergy` modules
- Enabling *parallel simulations* in Gitsbe
- Added support for many features (ongoing work - see dev_plan_doc)
- druglogics-roc-generator: R shiny app to assess the performance of the Drabme results in the form of a ROC curve
- Export support using BioLQM: the initial model + best generation models can now be exported through configuration options to **GINML, SBML-Qual and BoolNet** community formats

1.2 VSM

Building VSM-dictionaries in order to connect/translate the data from various databases and ontology providers to proper VSM-terms. Most of this work is done in order to support Vasundra's causalBuilder Tool which is the first application of VSM after SciCura v1.

The vsm-dictionaries (code +documentation) can be found on the VSM Github page. They translate to VSM-terms data from BioPortal, UniProt, Ensembl, EnsemblGenomes, RNACentral and ComplexPortal. We have also released the respective packages on npmjs. See for example the npm package for BioPortal.

1.3 PSICQUIC

My work at the EBI with IntAct and Noemi Del Toro to extend the PSICQUIC web service to support the miTab 2.8 data format/standard. See the psicquic doc and the casualTab paper (Perfetto et al. 2019).

I also worked with Noemi on the update of the JAMI library to also support miTab 2.8 - this is the culmination of results from the BioHackathon 2018, in Paris and the Marseille GREEKC hackathon event.

1.4 Others

- Java Client for RSAT tool fetch-sequences

Chapter 2

PhD Tasks and Plans

2.1 Programming Tasks TODO

Tasks that I have promised that I will do to different people within the group. These tasks enable other workflows/collaborations, etc. so they are very important to finish before I move on to other work. You see only what's left of those:

- Pipeline (see the dev_plan_doc for what is left). Most important:
  - Full BioLQM support: stable state calculation and trap spaces
  - Do comparison between Aurelien's BioLQM stable state algorithm and BNReduction using M2 or without (Asmund already says that it BNReductions is faster but it's good to prove it once again)
- VSM
  - Make the `vsm-pub-dictionaries` module

2.2 Papers

Note that the titles and the details for each paper are liable to change though the core ideas behind should not.

The papers dictate my future work for this PhD (and in that order!).

Paper I: *emba*: an R package for ensemble boolean model biomarker analysis

Authors

John, Asmund

Idea

The idea here is to analyse the models produced by `Gitsbe` in order to find important nodes (biomarkers) responsible for either better performance (based on a metric score like TP or MCC) or for specific synergy(ies) prediction.

The R package emba (Zobolas 2019) is publishable by itself as an **application note paper**, but we decided with Asmund that is best to present with an analysis on some dataset to show it's use. The package is also used for analyses that will be included in Asmund's automated pipeline paper (on the cascade topology mostly).

An idea is to compare ML results with my method (on cascade/atopo results of the pipeline paper or other). Paper could be titled something along the lines of **"Ensemble model analysis vs Machine Learning for unraveling drug synergy mechanisms"**. (which depends only on me). Could also be that I combine Vasundra's and Barbara's computational work on the biomarker analysis with mine and the ML methods (so more generic even) to create a nice combo paper about **"Unraveling drug mechanisms on CRC cell lines using various computational methods"**.

A research question here (with similar kind of work) is about the **identification of optimal training data size**, in the sense that we can use various methods to derive a node (biomarker) set for the training data (with Eirini).

Paper II: *VSM-dictionaries*: common access to biological dictionaries

Authors

John, Steven, Vasundra, Martin

Idea/Implementation

A short **application note** paper for my work on VSM-dictionaries.

Paper III: From causal statements to building logical models compliant with biological knowledge

Authors

John, others (to decide)

Idea

This paper will be the main **research paper of my PhD**. This idea is like a continuation of the `causalBuilder` tool by Vasundra coupled with the need to have a more proper representation of complexes (and families) in our logical models (better models, better predictions). Asmund had *manually* changed some logical equations in his paper (Flobak et al. 2015), in order to make the model more compliant with biology knowledge and literature findings. One of them was about the beta-catenin complex and its constituents (connected with *AND's* instead of *OR's*) and the rest were about changing the link operators of the logical equations (from *AND NOT* to *OR NOT*). The latter is something that is enabled through the mutations introduced by the genetic algorithm of `Gitsbe`. The former depends on the dataset and the representation of complexes.[1]

[1] There is actually a mutation that can change this but not in the way that we want - i.e. all components of a complex should be connected with an *AND*

Only Signor has some complexes + interaction data but they are seperate files, making it thus difficult (and non-elegant computationally-wise) to integrate such knowledge/data to boolean models. Also Vasundra's experience with Reactome data in miTab2.8 showed us the difficulty to match binary interactions to a data model flexible enough to represent complexes and their internal components. Causal-JSON and the recursive schema that we thought allows the curator to put both the complex ID and it's constituents in the same data structure. And what's best than using a proper tool to annotate such causal statements like VSM!

Proposed Workflow

So, the general semi-automated web-application (web only if VSM comes into play) pipeline for this paper that I am thinking will be as follows:[2]

1. Get interaction + complexes/families data (Signor most probably or a form of Cascade + complexes???). Note that for the reason I explained above miTab 2.8 is out of the question, so the Signor data I am refering to is the .tsv files they offer (interaction data, complexes, families). And most probably I am referring to a **pathway interaction dataset** not the whole Signor data. For example, the Wnt Signaling pathway.
2. Build a small module that translates the (Signor) data to Causal-JSON.
3. Optionally[3] there can be support for (re-)curation of the data with VSM template(s) for proper annotation of complexes (the templates will be automatically pre-filled to match Signor's data model). Note that this implies causal-JSON => VSM-box JSON format, which is the reverse of what causalBuilder will do. It will be nice if such interface supports addition of one more template of the same thing to curate one new sentence and add it to the data. The thing is that I want to include VSM into this, but I don't yet know how to really support its place in this pipeline since the curation is already done from Signor. It can be part of a showcase though: an example causal statement from

[2]My goal here is to combine as much as possible different things within our group, so each point or package should be small, their combination effect large

[3]might actually take a lot more effort than what I am planning that's why it's optional

the pathway with a complex and how it could be cureated with VSM.

4. **Main:** Build a package that translates the causal-JSON data to a logical model with some filtering and parameterization included (e.g. filter based on cell line (so *cell-line* specific topologies), conditions on the biological state: 'by phoshorylation', exp. evidence, assertion/confidence score, species, compartment). So, **causal-JSON to .bnet files (logical equations)**, while substituting nicely[4] the complexes and families.
5. Showcase some small application of this logical model end-product:
    - use for example the colomoto notebook, do some small trapspace analysis and show that some results from literature or from previous logical or other models can now be reproduced with a better biological representation in the model itself
    - make many logical models of the pathways in Signor with simple attractor analysis and put them into the GinSim model repository for reference for the logical community.
    - extend `atopo` module to use the main (3) package and use it for finding drug combinations (comparing attractors or prediction results of automated topology building without complexes vs automated topology built from causal-JSON with complexes and families in each case). The main thing here would be of course better prediction performance results based on a better logical model representation. I could tweak `atopo` to choose actually not all of Signor's data but specific pathways to include in the analysis and this will help I believe to build smaller topologies for specific drug combinations that we want to test.

Interesting resources:

- Signor Pathway data
- CausalBuilder
- CancerGeneNet

[4]huge discussion here, but I already know what to do pretty much

2.3 Other Paper Ideas

Paper IV (Deprecated)

Title

Extending `SynergyFinder` for the use of multiple reference models for the assement of synergy in screening datasets

Idea

The core idea here is to extend an existing R package for calculating synergy reference models in order to include Wim's generalized Bliss method and the mean synergy score by Simone Laderer! Then I will test all the null reference models (Loewe, Bliss, ZIP, +2 new, others?) on dose-response matrix datasets (could be from Ladere's paper, from Asmund's paper, our own - Barbara's/Evelina's dataset, etc.) and see which is best at finding the synergies in each dataset.

R package to use:

- SynergyFinder from Finland group code here
- Also I have to check this software: R package COMBIA as well as others at the time I will be tackling this

Also, I should investigate if my own idea for a mathematical formulation of the volume-based synergy score as general method for describing 3-wise or more combinations as synergistic, could be part of this implementation (or could be another paper by itself).

Paper V (Why not?!)

Authors

John, others (to decide)

Title

On eye-balling synergy mechanisms: What is a ~~mountain~~ synergy?

Idea

I had the idea of writing a small paper (send to Science Signaling) that describes the *eye-ball* or *visual inspection* technique that is used so much in computational Biology and Medicine. It is used pretty much in any paper I have seen but nobody has actually defined or named it.

A characteristic example is the *call for synergy* when you see some experimental data in dose-response curves (show mostly when you see figures). Another is the threshold that data analysts put when defining output to classifiers or the parameterization that is used and the general human intuition/engineering that is shared in all these.

As Asmund once said:

> What is a mountain? What is a synergy?

So, I just want to discuss this in a small paper cause it seems to be a thing that everybody is aware of in their everyday job, but everyone keeps out when talking about results, etc and I haven't found a paper that discusses this one way or another! A very RRI-related paper...

Miscellaneous Stuff

Chapter 3

PhD ideas

Several ideas that I may do or not in my PhD but I still keep here for my future investigations!

3.1 Quantum logic formalism

My favourite! Investigate if instead of a logical modeling formalism, the idea of (quantum) logical gates can be used to represent and analyse protein interaction networks. The **core idea** makes sense: you don't know the state of a protein, but when you measure it, only then you really know what it is.

May also be worth to look at a game-theoritic approach to find attractors and such.

3.2 Compare fixpoint tools

Idea: Compare different tools that calculate fixpoints for logical modeling. Faster wins of course :)

Models used for testing could be of different types:

- self-contained

- varying the number of input nodes (1-n)
- small to large number of nodes
- small to large number of edges
- scale-free (boolnet generated) vs random (varying K connectivity)
- play with form of the boolean equations
- others ???

Workflow for this includes:

- support BNReduction data format by Veliz-Cuba in BioLQM
- add support for calculating the fixpoints using the Colomoto docker (python interface) + BNReduction
- then comparison between **BioLQM, Pint, MABOSS and BNReduction** could be done then in a Jupiter colomoto-enabled notebook!

Further extension/comparisons could be:

- (Akutsu, Hayashida, and Tamura 2009) - Integer programming method
- (Dubrova and Teslenko 2009) - SAT-based

3.3 Reasoning with VSM

The idea here is to use VSM to annotate sentences about some knowledge area, store this information in a format like RDF or something else (graph database?) and then ask questions that will enable you to learn stuff that you didn't know before.

One goal would be to show the superiority of *connection-based reasoning* (humans and what VSM encapsulates) vs *logic reasoning* (OWL).

Another thing I thought was to just translate the VSM-data to PROLOG and then ask questions using that logical language framework. It is a way to show that you *learned* something using the VSM-supported curation but I don't know where to go from there... this whole knowledge semantics and reasoning stuff seem to be a PhD on its own :) There are a lot of things that should be investigated for this idea to materialize properly (lots of reading).

An idea by Steven:

Mapping / inferring different forms of representing interactions (molecular and/vs. causal) using a VSM sentence presentation as a graph diagram. A first step to reasoning would be to make some rules like, if you have VSM-sentences A and B, then you can infer C from that.

3.4 Use Logical modeling to predict single-drug data

Asmund's proposal idea that he sent to my email once. Has to do about *mechanistic drug response prediction analysis*:

- Automate drug target profile annotation from:
  - (Klaeger et al. 2017)
  - mrc ppu
  - (Davis et al. 2011)
- Omics data (rna, cnv etc)
  - COSMIC
  - CCLE
- Drug screen data
  - Single drug
    - COSMIC/GDSC
    - CCLE
  - Combo
    - (O'Neil et al. 2016)
    - (Holbeck et al. 2017)

My idea is more like this:
Predict drug-response curves from drug combination datasets (GDSC, CCLE), using logical modeling for singaling network analysis or translation from logical to ODE modeling. Also try to predict drug combinations datasets (dose-response matrices?). Pretty much what is done in this paper (Fröhlich et al. 2018) with help from (Wittmann et al. 2009) for converting boolean

models to continuous.

3.5 Druglogics-Pipeline related

3.5.1 Harmony Search

Nice idea because it's related to music! Investigate if this algorithm could be used for optimizing the boolean equations for `gitsbe` - thus opening the stage for `JazzLogics`.

3.5.2 Train models to cell-specific proliferation

Concept is that training models to proliferate provides a wider variance of models than the cell-specific trained ones in `gitsbe`: main directive is **proliferation**, not just fitting to a steady state pattern. So a hybrid training approach should be way more advantageous.

3.5.3 A bottom-up model building for drug prediction

Start with a model and some observed synergies. Build/train/produce models that predict the first observed synergy (using Harmony Search?), from them the next one, etc. You end up with many models that can predict all the observed synergies or you try to find out why that cannot happen for example (e.g. contrasting synergies?). Do the latest models' stable states or attractors correspond to activity of proteins from literature?

3.5.4 Simulate cancer resistance

For example, you have some models that predict some (observed) synergies or you just find some synergistic drug comibnations for these models or per model. Then, you modify these models in order to be resistant to these drugs, simulating thus the cancer rewiring process! Then, you apply (n+1) drug

combinations to win over the resistance (and you do this procedure at more levels to suggest 3-way, 4-way drug combos and why there might be cancer models that can 'win' over these models and continue the proliferation). You end up with super cancer resistant models and methods to achieve them or reasons why this cannot happen at all.

Papers

Papers that are already published and I am in the list of authors:

- (Perfetto et al. 2019)

Papers that will probably be published and I will probably be in the list of authors:

- The Biohackathon 2018 paper
- Asmund's automated pipeline paper: *A novel and versatile drug synergy prediction pipeline with single sample resolution for discovery of potent drug combinations in pre-clinical and clinical datasets*
- Vasundra's causalBuilder tool paper

Bookdown useful links

- Bookdown github repo
- Bookdown package reference
- Writing Thesis with Bookdown
- Bookdown workshop slides

References

Akutsu, Tatsuya, Morihiro Hayashida, and Takeyuki Tamura. 2009. "Integer programming-based methods for attractor detection and control of boolean networks." In *Proceedings of the 48h Ieee Conference on Decision and Control (Cdc) Held Jointly with 2009 28th Chinese Control Conference*, 5610–7. IEEE. https://doi.org/10.1109/CDC.2009.5400017.

Davis, Mindy I, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. 2011. "Comprehensive analysis of kinase inhibitor selectivity." *Nature Biotechnology* 29 (11): 1046–51. https://doi.org/10.1038/nbt.1990.

Dubrova, Elena, and Maxim Teslenko. 2009. "A SAT-Based Algorithm for Computing Attractors in Synchronous Boolean Networks." https://arxiv.org/pdf/0901.4448.pdf https://ieeexplore.ieee.org/document/5958722/.

Flobak, Åsmund, Anaïs Baudot, Elisabeth Remy, Liv Thommesen, Denis Thieffry, Martin Kuiper, and Astrid Lægreid. 2015. "Discovery of Drug Synergies in Gastric Cancer Cells Predicted by Logical Modeling." Edited by Ioannis Xenarios. *PLOS Computational Biology* 11 (8): e1004426. https://doi.org/10.1371/journal.pcbi.1004426.

Fröhlich, Fabian, Thomas Kessler, Daniel Weindl, Alexey Shadrin, Leonard Schmiester, Hendrik Hache, Artur Muradyan, et al. 2018. "Efficient Parameter Estimation Enables the Prediction of Drug Response Using a Mechanistic Pan-Cancer Pathway Model." *Cell Systems* 7 (6): 567–579.e6. https://doi.org/10.1016/J.CELS.2018.10.013.

Holbeck, Susan L, Richard Camalier, James A Crowell, Jeevan Prasaad

Govindharajulu, Melinda Hollingshead, Lawrence W Anderson, Eric Polley, et al. 2017. "The National Cancer Institute ALMANAC: A Comprehensive Screening Resource for the Detection of Anticancer Drug Pairs with Enhanced Therapeutic Activity." *Cancer Research*. https://doi.org/10.1158/0008-5472.CAN-17-0489.

Klaeger, Susan, Stephanie Heinzlmeir, Mathias Wilhelm, Harald Polzer, Binje Vick, Paul-Albert Koenig, Maria Reinecke, et al. 2017. "The target landscape of clinical kinase drugs." *Science (New York, N.Y.)* 358 (6367): eaan4368. https://doi.org/10.1126/science.aan4368.

O'Neil, Jennifer, Yair Benita, Igor Feldman, Melissa Chenard, Brian Roberts, Yaping Liu, Jing Li, et al. 2016. "An Unbiased Oncology Compound Screen to Identify Novel Combination Strategies." *Molecular Cancer Therapeutics* 15 (6): 1155–62. https://doi.org/10.1158/1535-7163.MCT-15-0843.

Perfetto, L, M L Acencio, G Bradley, G Cesareni, N Del Toro, D Fazekas, H Hermjakob, et al. 2019. "CausalTAB: the PSI-MITAB 2.8 updated format for signalling data representation and dissemination." Edited by Jonathan Wren. *Bioinformatics*, February. https://doi.org/10.1093/bioinformatics/btz132.

Wittmann, Dominik M, Jan Krumsiek, Julio Saez-Rodriguez, Douglas A Lauffenburger, Steffen Klamt, and Fabian J Theis. 2009. "Transforming Boolean models to continuous models: methodology and application to T-cell receptor signaling." *BMC Systems Biology* 3 (1): 98. https://doi.org/10.1186/1752-0509-3-98.

Zobolas, John. 2019. *Emba: Ensemble Boolean Model Biomarker Analysis*. https://github.com/bblodfon/emba.