Automatically Navigating Protein Interaction Networks with a Software Product Line Approach

Daniel-Jesus Munoz
1 Instituto de Tecnología e Ingeniería del Software (ITIS)
2 CAOSD, Dpt. Lenguajes y Ciencias de la Computación
Universidad de Málaga, Andalucía Tech, Spain
danimg@lcc.uma.es

Dina Medina-Vera
1 Instituto de Investigación Biomédica de Málaga-IBIMA
Hospital Regional Universitario de Málaga
2 Facultad de Ciencias, Universidad de Málaga, Spain
dinamedina@uma.es

Abstract—Software Product Line Engineering (SPLE) would categorise protein-protein interaction (PPI) networks as highly configurable systems, and the main issue with those is the impracticability to manually analyse all network/system scenarios. SPLÉ solves this issue by means of automated reasoners – tools that analyse every solution of an SPLÉ system. Hence, if we apply and SPLÉ approach to analyse PPI networks, we can automatically navigate through every possible PPI pathway, including uncovering new ones with regards to the current literature, providing computer decision aid to both, bio-practitioners and researchers. While PPI networks are represented by the standard Systems Biological Graphical Notation (SBGN), product lines are represented by Variability Models (VMs). We present an approach where SBGN diagrams are transformed to SPLÉ VMs, providing compatibility between PPI networks and automated reasoners. We conjecture that protein artefacts are, in essence, variability features, and the signalling interactions are first-order logic relationships with certain cardinalities. We then analyse a reduced cell PPI (i.e., the Akt pathway) case study represented as a Clafer model with 9 features and 3 cross-tree constraints. The model is analysed with the chocosolver reasoner – the state-of-the-art multi-purpose and automated reasoner. The evaluation uncovered 84 pathways and was carried out in less than a millisecond using a RaspberryPi 3B+. We demonstrated how can be beneficial in biomedical research by supporting the creation, update, and re-usability of PPI network VMs.

Index Terms—software product line engineering, automatic reasoner, SBGN activity flow, PPI network, protein, Akt pathway, variability model, solution space, decision aid

I. INTRODUCTION

In cellular biology, networks are systems of interacting molecules that perform cellular tasks, such as the regulation of gene expression, protein-protein interaction, post-translational modifications, metabolism or intracellular signaling [1]. Among them, PPI networks represent one of the most challenged aspects of current molecular biology but also a powerful tool to understand the pathogenic mechanisms that trigger the onset of diseases [2]. Therefore, studies of PPI are fundamental to perceive their role within the cell. Consequently, the understanding of these networks facilitates the research for effectiveness biological and biomedical strategies.

Funding: Munoz work is partially supported by the projects TASOVA MCIU-AEI TIN2017-90644-REDT, MEDEA RTI2018-099213-B-I00 (co-financed by FEDER funds), LEIA UMA18-FEDERIA-157 and RHEA P18-FR-1081 (MCI/AEI/FEDER, UE). Medina-Vera work is partially supported by the project EU-LAC Health. FATZHEIMER (EULACH16/T010131).

However, the complexity and the large number of these PPI networks in a single cell, make difficult the understanding of the system as an integrated whole. Also, it makes impracticable to manually analyse all network/system scenarios or even possible alternatives. Some related works to this issue have been published. In [3] logic-based models alongside fuzzy logic determine the effects of protein over-expression or inhibition on phenotype, elucidating bio-chemical signalling network properties. The DeepLoc tool [4] aims to predict protein sub-cellular localisation relying only on sequence information by means of a deep neural networks prediction algorithm. A model-based approach is found in [5], where the authors analyse bio-chemical networks and predicts the effects of certain network parameters making use of dynamic modelling. A more sounding work is [6], in which the authors transforms knowledge of complex biological processes from sets of possible interactions and experimental observations to precise, predictive biological programs governing cell function. Their methodology is based on automated formal reasoning and machine learning.

While all the presented methods and tools have in common the goal of estimating signalling effects in order to guide simulations and research, they induce inherent predictive issues as requesting specific training sets to assure sufficient accuracy, scalability limitations for colossal networks as runtime or computational-power costs, etc. [7]. Nevertheless, we propose an approach that leaves-aside predictions and their issues, reusing the existing data independently of which, and purely aims to computationally aid to uncover every possible PPI pathway without considering their effects. While not tested, this approach has been suggested in [8], where mathematical models of molecular and gene networks are presented as a potential approach to uncover new synthetic entities. Last, the notions of organic software product lines are presented in [9], where a product line approach is applied to synthetic biology to uncover every possible DNA interaction.

In engineering, a product line is defined as a family of products designed to take advantage of its common aspects and predicted variabilities [10]. Product line in the software development community is called SPLÉ, and has become a best practice for modelling and managing families of highly configurable systems [11]. Configurable systems can be rep-
resented as VMs, where valid solutions are described by specifying relationships in terms of Boolean features, and constraints among them [12]. As it is infeasible for human brains to completely understand and analyse large solution spaces, in the SLE area solvers are being developed. A solver tool is an automated reasoner which, parting from a set of features, relationships, and cardinality constraints (i.e., a constraint satisfaction problem), is able to automatically count, generate, and navigate through all the possible solutions of the given model while discarding invalid paths [13]. Hence, SLE automated reasoners is a potential solution to automatically navigate throughout PPI networks solutions spaces uncovering every valid alternative interaction pathway.

In this work we explore the usage of SLE supporting bio-practitioners and researchers with a computerised aid for PPI network analysis. We conjecture that there are interaction pathways that consist of common and variable artefacts (i.e., proteins), and a set of Boolean relationships and cardinality (i.e. signalling interactions). Hence, PPI network models can be directly translated to SLE VMs in order to have access to automated reasoners. We test this approach with a reduced Akt pathway as a case study to evaluate this claim. The contributions of this work are:

1) A mapping of PPI networks models to SLE VMs which in turn provide access to automated reasoners.
2) A case study demonstrating the potential reuse and existence of both commonality and variability in the Akt pathway and showing that we can build VMs that have potential to help bio-practitioners and researchers.

II. METHODS

In our approach we work with two different feature models: SBGN activity flow diagrams and VMs.

Biological signalling pathway diagrams are graphically represented in different forms, being the SBGN the international standard [14]. SBGN activity flow diagrams puts emphasis on the activities performed by the biological entities, and their effects to other ones, showing influences such as stimulation/activation and inhibition. For example, a signal stimulates the activity of a receptor, and this activity in turn stimulates the activity of an intracellular protein. In the current literature, most signalling pathway diagrams, including PPI networks, essentially are activity flow diagrams, in which nodes symbolise biomolecules (e.g., proteins), and edges represent positive (activation) or negative (inhibition) influences. A positive influence in a SBGN activity flow is defined as an action that produces positive or activating effect from one activity to another, whose symbol is an arrow pointing to the target. On the other hand, a negative influence is defined as an action that produces a negative or inhibiting effect from one activity to another, whose symbol carries a bar perpendicular to a line. Those symbols are represented in Figure 1, and an example of an SBGN activity flow diagram is present in Figure 2-A.

A VM of a SLE system is a hierarchically arranged set of features [12]. A feature is an increment in system functionality. A solution is defined by a unique and limited combination of abstract features. The set of all legal feature combinations defines the set of valid solutions of the system. Current methodologies organise features into a tree, called a VM tree, which is used to declaratively specify product-line features [15]. VM trees relationships between parent feature and its child features are categorised as:

- **Optional**: The child can be present.
- **Mandatory**: The child must be present.
- **Alternative**: Only one child can be present.
- **Or**: One child or more can be present. Additionally, detailed cardinality can be specified as \([x : y]\) – a minimum of \(x\) and at most \(y\) children can be present.
- **And**: All children must be present.

A VM tree example is can be seen in Figure 2-B. The physical entities represented by the features are called components, and the set of components and their relationships form the architecture space.

Unlike SBGN diagrams, VMs have access to automated reasoners. Hence, we propose the following SBGN diagram to VM transformation in order to provide automatic reasoning to biological signalling pathways:

- **SBGN nodes** are directly translated to VM features. Specifically, they are all direct children of the root feature, and their cardinality is \([2 : \ast]\) as at least 2 entities are needed in a interaction pathway. Additionally, they must follow the same sequential order as it intrinsically acts as the interaction direction information.
- **SBGN activation relationships** are translated to implication cross-tree constraints.
- **SBGN inhibition relationships** are translated to negated implications cross-tree constraints.

A transformation example can be seen in Figure 2.
Among the state-of-the-art reasoning ecosystems, Clafer suite [16] has proven to be the fastest [17]. VMs are defined in the ecosystem as a text file in Clafer modelling language, which combines structural modelling with behavioural formalisms, following an identical syntax as VM trees for the approach that we are presenting. The Clafer suite also integrates an automated reasoner; chocosolver solve problems by alternating constraint filtering algorithms with a configurable search mechanism. Technically speaking, chocosolver will compute every possible intermediate, complete and/or parallel pathways of a given PPI.

Fig. 2. Reduced Akt pathway SGBN model to VM transformation

III. RESULTS AND DISCUSSION

We conducted a case study to evaluate the feasibility of using an SPLE approach for PPI networks – concretely a reduced PPI, the Akt pathway. We chose this Akt model as it is the leading protein pathway study with regard to the literature, and yet continues to grow in areas as cancer or Alzheimer’s disease [18]. More precisely, Akt is a crucial protein for cell survival and proliferation which can be activated by many factors and other proteins. Here, we will just be focused on its activation pathway via the insulin-receptor. Figure 2-A shows a SBGN’s activity flow diagram in which insulin, acting as a physiological growth factors, activates the Akt route by...
binding to its receptors at the cell membrane. The adapter protein insulin receptor substrate 1 (IRS-1) activates PI3K. PI3K then activates Akt indirectly. At this point, two Akt downstream proteins are inhibited: glycogen synthase kinase 3 (GSK-3) and FOXO. The inhibition of GSK-3, prevents the inhibition of glycogen synthase (GS). The activation of Akt also leads an indirect activation of mTOR Complex 1 (mTORC1). On the other hand, the Akt pathway can be turned off at the beginning of the pathway through the phosphatase and tensin homolog (PTEN) or the protein phosphatase 2C (PP2C).

We ask the following research questions:

- **RQ1**: Do PPI networks have the characteristics of an SPLE? We need to show that SPLE is capable to integrate PPI network analyses.
- **RQ2**: Can VMs represent an SBGN of an existing PPI network? We need to prove that SBGN diagrams can be completely translated to SPLE VMs.
- **RQ3**: Can SPLE provide computerised aid to biopractioners? We need to analyse a PPI network with an SPLE automated reasoner.

A. **Do PPI networks have the characteristics of an SPLE?**

While Figure 2 shows a reduced Akt pathway, the complete human Akt system potentially comprises more than an hundred artefacts involved in more than a million valid pathways, being most of them yet undiscovered. These numbers are larger for the complete set of PPI networks. In an SPLE, living beings, proteins, and effects, are the re-usable components of the PPI network architecture space. PPI pathways, defined as a set of artefacts and their relationship, are solutions in the SPLE solution space, and programs once those solutions are physically deployed and executed. Last, but not least, SPLE PPI networks can be configured, as programs are; in an SPLE, a researcher can constrain the artefacts in his analysis which are exactly user requirements. This allows to restrict at analysis time the specific areas of the solution space in interest – Figure 2 is indeed a restricted diagram of the complete Akt pathway, and our case study. This SPLE PPI network vision is completely detailed in Figure 3 comprising: the PPI network VM in the SPLE Problem Space section, the case study requirements in the Requirement Analysis section, the proteins as physical artefacts in the Solution Space architecture, and the automated reasoner in the Product Derivation section which finally navigates throughout the possible pathways.

**RQ1 answer**: PPI networks comprises re-usable artefacts of a system which relationships can be modelled, architecture components, and analyses requirements. Hence, PPI networks present the characteristics of an SPLE.
B. Can VMs represent an SBGN of an existing PPI network?

For RQ2 we perform the A to B (i.e., SBGN to VM) transformation of Figure 2 as explained in Section II. With this approach, 9 features and 4 cross-tree constraints were detected. The concrete Akt pathway Clafer VM of Figure 2 can be downloaded for testing from our repository 1.

RQ2 answer: In short, proteins in SBGN are VM features with a cardinality \([2, \ast]\), and interactions are cross-tree constraint relationships.

C. Can SPLE provide computerised aid to bio-practitioners?

When running the automated reasoner choosolver with Figure 2 Clafer model as an input, 84 valid solutions are automatically generated in less than a millisecond using a Rapsberry Pi 3B+ micro-computer – a really small and low computational power device. Please consider that among the 84 solutions, larger ones can be considered final pathways, but intermediate ones are also generated as less granular solutions could be useful in research or simply to soften the understanding of the space (e.g., solution 1 is [IRS-1, Akt]). Alternatively, this could be easily configured in order to uncover just large (i.e., complete) routes, but it is out-of-the-scope of this paper. And just to point-it-out, training sets and/or predictive techniques are not involved in this approach.

RQ3 answer: We proved that SPLE reasoners automatically uncover every possible abstract or complete PPI pathway with low computation power devices in less than a millisecond.

IV. CONCLUSIONS AND FUTURE WORK

In this research paper we have shown how SPLE can potentially support bio-practitioners and researchers with a computerised aid for PPI network analysis. We then used the crucial Akt cell pathway as a PPI model to analyse: 1) whether PPI networks have the characteristics of an SPLE, 2) whether VMs can represent an SBGN of an existing PPI network, and 3) how common SPLE tools can be used to benefit bio-practitioners providing automated computerised decision aid. We answered RQ1 proving that PPI networks comprises re-usable artefacts of a system which relationships can be modelled, architecture components, all characteristics of an SPLE. RQ2 answer can be summarised as that proteins in SBGN are VM features with a cardinality \([2, \ast]\), and that interactions are cross-tree constraint relationships. In RQ3 we used the choosolver SPLE reasoner to automatically uncover the 82 abstract or complete PPI pathways with a RaspberryPi 3B+ in less than a millisecond. Additionally, we slightly detail the potential of navigating through every possible PPI route, including uncovering new ones, and how this can help research for effectiveness biological and biomedical strategies.

In future work we plan to investigate building a larger model while including the most complete PPI pathway based on the literature, living-species, pathways effects, and treatments, and introduce sampling and predictive techniques to uncover potentially unknown interactions, effects, while suggesting new treatments. Additionally, we will probably develop a web-tool and web-services as the ones in the SPLE HADAS Tool [17]. The idea is to provide a web-app with already modelled PPI networks in order to allow direct PPI network analysis supporting users case study requirements.

REFERENCES

[1] D. Bray, “Protein molecules as computational elements in living cells,” *Nature*, vol. 376, no. 6538, pp. 307–312, 1995.
[2] N. Safari-Alighirloo, M. Taghizadeh, M. Rezaei-Tavirani, B. Goliaei, and A. A. Peyvandi, “Protein-protein interaction networks (ppi) and complex diseases,” *Gastroenterology and Hepatology from bed to bench*, vol. 7, no. 1, p. 17, 2014.
[3] M. K. Morris, J. Saez-Rodriguez, P. K. Sorger, and D. A. Lauffenburger, “Logic-based models for the analysis of cell signaling networks; *Biochemistry*, vol. 49, no. 15, pp. 3216–3224, 2010.
[4] J. J. Almagro Armenteros, C. K. Sanderby, S. K. Sanderby, H. Nielsen, and O. Winther, “Deeploc: prediction of protein subcellular localization using deep learning,” *Bioinformatics*, vol. 33, no. 21, pp. 3387–3395, 2017.
[5] N. A. van Riel, “Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments,” *Briefings in bioinformatics*, vol. 7, no. 4, pp. 364–374, 2006.
[6] B. Yordanov, S.-J. Dunn, H. Kugler, A. Smith, G. Martello, and S. Emmott, “A method to identify and analyze biological programs through automated reasoning; *NPJ systems biology and applications*, vol. 2, no. 1, pp. 1–16, 2016.
[7] T. Pett, T. Thüm, T. Runge, S. Krieter, M. Lochau, and I. Schaefer, “Product sampling for product lines: The scalability challenge,” in *Proceedings of the 23rd International Systems and Software Product Line Conference-Volume A*, 2019, pp. 78–83.
[8] N. Le Novère, “Quantitative and logic modelling of molecular and gene networks,” *Nature Reviews Genetics*, vol. 16, no. 3, pp. 146–158, 2015.
[9] M. Cashman, J. Firestone, M. B. Cohen, T. Thanniwat, and W. Niu, “Dna as features: organic software product lines;” in *Proceedings of the 23rd International Systems and Software Product Line Conference-Volume A*, 2019, pp. 108–118.
[10] N. S. Staudemayer and D. E. Perry, “Session 5: Key techniques and process aspects for product line development;” in *10th International Software Process Workshop*, 1996.
[11] D. M. Weiss and C. T. R. Lai, *Software product-line engineering: a family-based software development process.* Addison-Wesley Reading, 1999, vol. 12.
[12] D.-J. Munoz, J. Oh, M. Pinto, L. Fuentes, and D. Batory, “Uniform random sampling product configurations of feature models that have numerical features,” in *Proceedings of the 23rd International Systems and Software Product Line Conference-Volume A*, 2019, pp. 289–301.
[13] J. Harrison, *Handbook of practical logic and automated reasoning.* Cambridge University Press, 2009.
[14] H. Mi, F. Schreiber, S. Moodie, T. Czauderna, E. Demir, R. Haw, A. Luna, N. Le Novère, A. Sorokin, and A. Vile`ge, “Systems biology graphical notation: activity flow language level 1 version 1.2,” *Journal of integrative bioinformatics*, vol. 12, no. 2, pp. 340–381, 2015.
[15] D. Batory, “Feature models, grammars, and propositional formulas,” in *International Conference on Software Product Lines*. Springer, 2005, pp. 7–20.
[16] M. Antkiewicz, K. Bak, A. Muraskhin, R. Olachea, J. H. Liang, and K. Czarnecki, “Clafer tools for product line engineering;” in *Proceedings of the 17th international software product line conference co-located workshops*, 2013, pp. 130–135.
[17] D.-J. Munoz, M. Pinto, and L. Fuentes, “Hadas: Analysing quality attributes of software configurations;” in *Proceedings of the 23rd International Systems and Software Product Line Conference - Volume B, ser. SPLC ’19*. New York, NY, USA: Association for Computing Machinery, 2019, p. 13–16. [Online]. Available: https://doi.org/10.1145/3307630.3342385
[18] R. M. Barker, J. M. Holly, K. M. Biernacka, S. J. Allen-Birt, and C. M. Perks, “Mini review: Opposing pathologies in cancer and alzheimer’s disease: Does the pi3k/akt pathway provide clues?” *Frontiers in Endocrinology*, vol. 11, p. 403, 2020.

1Clafers reduced Akt VM: [https://hadas.caosd.lcc.uma.es/reducedakt.txt](https://hadas.caosd.lcc.uma.es/reducedakt.txt)