Node Duplication in Disease Maps using Graph Neural Networks

Benjamin Moser

October 4, 2021

The harmony of the world is made manifest in Form and Number, and the heart and soul and all the poetry of Natural Philosophy are embodied in the concept of mathematical beauty.

– D’Arcy Wentworth Thompson

Preface

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Contents

List of Figures

List of Tables

1. Notation

1. Introduction

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.1. Biological Networks

Our understanding of the molecular mechanisms that are involved in biological systems is improving drastically. However, this knowledge is growing incrementally and is often scattered across individual scientific publications. The behaviour of a biological system is often defined by complex interactions, potentially across different levels of abstraction. We refer to biological entities as *species*. Possible species types are, for instance, proteins, genes, but possibly also abstract phenotype descriptions or drugs. Possible relations include chemical reactions such as state transition (e.g. phosphorylation), physical interaction between two proteins or the effect a drug has on the function of some protein. The set of species and their interactions (relationships) naturally form a network. To properly consider complex signalling effects such as activation/inhibition, crosstalk or feedback, it is of the essence to make this network structure accessible to both human cognition and computational analysis [1].

Choices of what species and relationships to consider yield different flavours of biological networks and analyses. Since the effective biological function of proteins is rarely defined solely by their identity but rather by their roles as enzymes, signalling molecules or structural components, it is worthwhile to study *Protein-Protein Interaction Networks* (PPI networks), for instance to infer the biological function of an unknown protein or gene, or groups of functionally similar proteins. Further, an organism's metabolism may be described as a set of chemical compounds (metabolites) and the set of chemical reactions or interactions between them, yielding a *metabolic model*. Computational methods on metabolic networks can be used to predict the growth of an organism under specific conditions, identify key intervention targets or decompose the network into relatively independent subsystems. However, species and relationships need not necessarily have a direct physical counterpart. Several works investigate the interactions between diseases, drugs, proteins and their annotated biological functions [2] [3]. Moreover, biological networks may serve as a scaffold to integrate data on, e.g., gene expression or reaction rates. This can be used to classify cancer or TODO.

1.2. Disease Maps

Biochemical pathways can be described by *process description diagrams* in which species (commonly metabolites, protein complexes and genes) are linked by chemical processes. An example is given in Figure **??**. Many diseases, however, affect not only a single mechanism. Rather, a systemic understanding of the involved subsystem and their relationships is required [4][5]. Moreover, knowledge on mechanisms contributing to a disease is obtained incrementally and scattered across individual publications or database entries. Particularly for visual, interactive exploration, experts have assembled *disease maps*, comprehensive diagrams combining all known mechanisms relevant for a given disease. Traditionally, such diagrams have been drawn as pixel- or vector-based graphics. Creating and updating such graphics requries a high amount of effort and renders the contained information practically inaccessible to computational methods. Formalised, digital representations that are both human- and computer-readable provide the following immediate advantages.

- The creation process, involving the extraction of knowledge from scientific publications or databases and finding an adequate layout, may be aided by computational tools from the areas of Data Mining and Graph Drawing.
- Entities in the diagram may be annotated with additional information such as links to research publication or database entries.
- A formalised representation enables the use of computational methods for analysis and interactive exploration (see Related Work for examples).
- Diagrams provide a formalized model that can serve as a scaffold for integrating *multi-omics* data.

Although their content is based on biological processes, disease maps differ in nature from other types of biological networks in the following aspects:

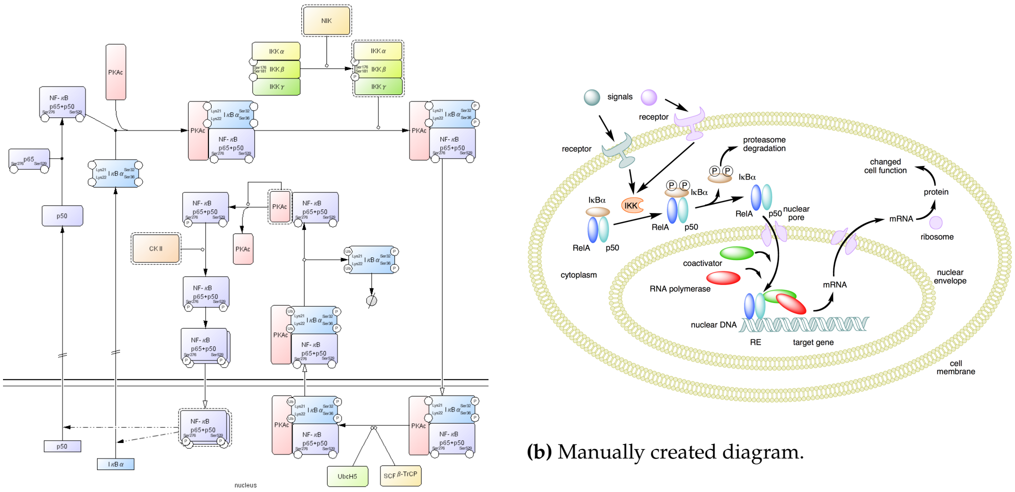
- The contents of a disease map are assembled based on the judgement of one or several curators. Only processes that are deemed relevant or informative to the given objective are included.
- A disease map is an actual visual diagram, i.e. visual representations of species and relationships have been laid out to optimally present the included information. Although several approaches for the drawing of large process diagrams exist (see TODO), it is still common practise to invest manual effort into the layout.

The above points also show that such maps are inherently subjective.

Recent disease maps contain up to several thousands of species and reactions and are very rich in information beyond the mere enumeration of species and their pairwise relationships. To give only a few examples: different types of species and relationships (as described above) are explicitly encoded. Further, species can be assigned relative positions to a biological compartment they reside in. Species can be assigned different states (e.g. "phosphorylated") and form *complexes* (groups). Further, species and relationships are often annotated with links to external databases such as Entrez Gene [6] or UniProt [7].

To date, disease maps have been created for a number of diseases, including Alzheimer's Disease (AlzPathway[8]), Parkinson's Disease (PDMap[9]), and recently COVID-19 [10]. Beyond serving as a platform for integrating existing knowledge, computational methods have been applied to, e.g., identify molecules and relations essential for the pathogenesis of Alzheimer's Disease [11]. Detailed information on the disease maps considered in this work can be found in Section **??**.

Several tools exist for the curation and exploration of such diagrams, including *CellDesigner* [12], *Minerva* [13], and *Cytoscape* [14] and *VANTED* [15]. We refer to TODO for a comprehensive comparison. One of the most common formats used for describing disease maps is an extension to SBML Level 2 given by *CellDesigner*. Further, SBML Level 3 now allows to attach layout information. Another prominent format is SBGN-ML.



(b) Manually created diagram.

(a) Diagram as created with CellDesigner.

Figure 1.1.: Two representations of prototypical mechanisms of NF-κB -signaling.

1.3. Drawing of Biological Networks / Disease Maps

1.4. Ontologies

1.5. Network Embedding & Neural Networks

Embeddings

Neural Networks

Graph Neural Networks

2. Related Work

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

2.1. Machine Learning in the Life Sciences

Applications of Graph Neural Networks

3. Methods

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3.1. Datasets & Preprocessing

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Datasets used for training and evaluation

Graph interpretation

Disease maps can be interpreted as bipartite graphs in a natural manner with the bipartite node sets being the set of reactions and the set of species, respectively.

Determining ground-truth Labels

Although numerous disease maps are publicly available, to the best of our knowledge none are explicitly annotated with a per-alias label indicating node duplication. In case we are given a sequence of reorganisation steps $(G_1, ..., G_k)$, we infer node labels by comparing successive steps G_t and G_{t+1}. In case we are given only a single disease map G , we first construct a collapsed version G_0 by collapsing any species aliases corresponding to the same species into a representative node and moving any edges incident to aliases to the corresponding representative. We then proceed by comparing G_0 and G like reorganisation steps. In order to make our results comparable to the work of Nielsen et al., we re-implement the algorithm employed therein and describe it in detail here for clarity.

Determining Predictors

3.2. Classification

Support Vector Machine

Graph Neural Network

4. Results

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

5. Discussion

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

6. First Chapter

Foo bar baz qux qoo. blblblbl

This is an equation:

$$f(x) = \mathbf{AX} \tag{6.1}$$

This is a reference to Equation 6.1.

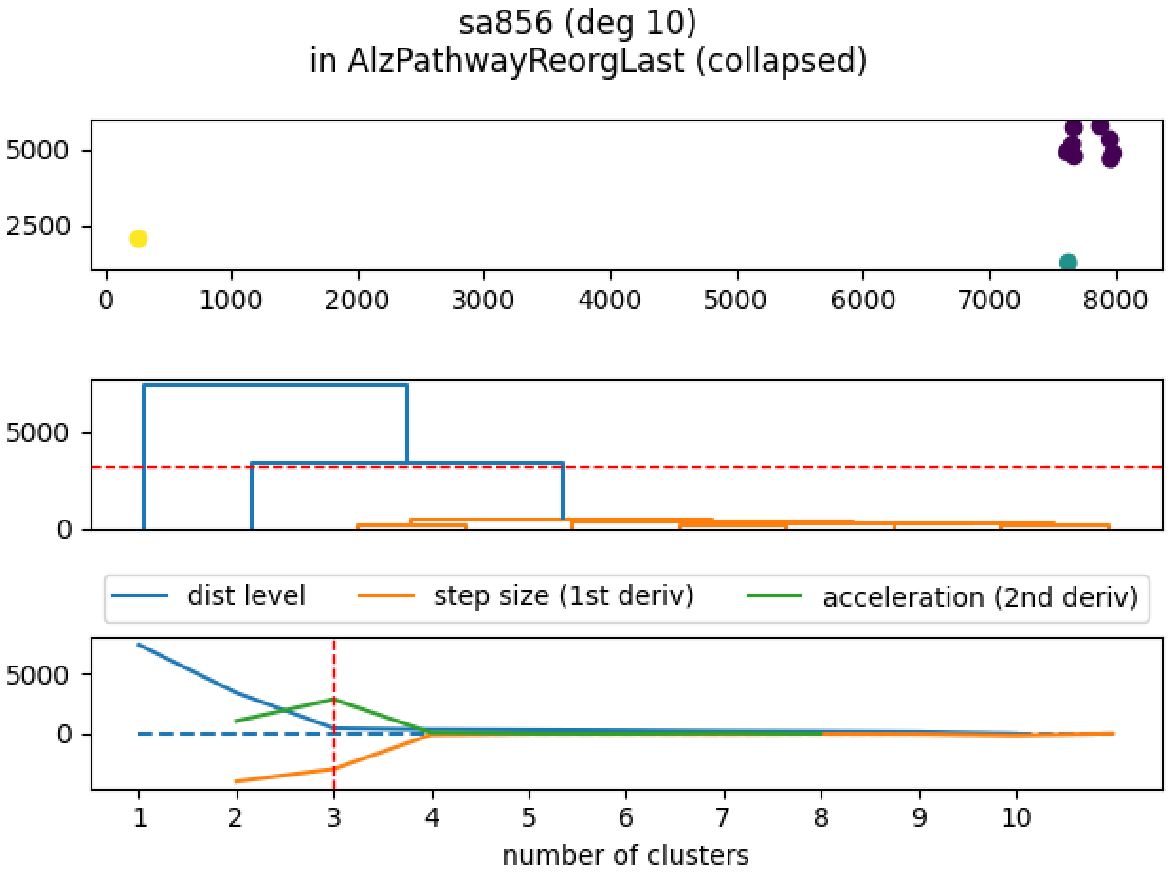


Figure 6.1.: This is some caption text

This is another reference to Figure 6.1

This is a bib ref [17].

This is an autor ref: Duvenaud et al.

7. Second Chapter

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Appendix

A. Some more blindtext

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Bibliography

Here are the references in citation order.

[1] Albert-László Barabási and Zoltán N. Oltvai. 'Network Biology: Understanding the Cell's Functional Organization'. In: *Nature Reviews Genetics* 5.2 (2 Feb. 2004), pp. 101–113. DOI: 10.1038/nrg1272. (Visited on 10/04/2021) (cited on page 1).

[2] Camilo Ruiz, Marinka Zitnik, and Jure Leskovec. 'Identification of Disease Treatment Mechanisms through the Multiscale Interactome'. In: *Nature Communications* 12.1 (1 Mar. 19, 2021), p. 1796. DOI: 10.1038/s41467-021-21770-8. (Visited on 03/22/2021) (cited on page 1).

[3] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. 'Network Medicine: A Network-Based Approach to Human Disease'. In: *Nature Reviews Genetics* 12.1 (1 Jan. 2011), pp. 56–68. DOI: 10.1038/nrg2918. (Visited on 10/04/2021) (cited on page 1).

[4] Marek Ostaszewski et al. 'Community-Driven Roadmap for Integrated Disease Maps'. In: *Briefings in Bioinformatics* 20.2 (Mar. 25, 2019), pp. 659–670. DOI: 10.1093/bib/bby024. (Visited on 06/08/2021) (cited on page 1).

[5] Alexander Mazein et al. 'Systems Medicine Disease Maps: Community-Driven Comprehensive Representation of Disease Mechanisms'. In: *NPJ systems biology and applications* 4 (2018), p. 21. DOI: 10.1038/s41540-018-0059-y (cited on page 1).

[6] Donna Maglott et al. 'Entrez Gene: Gene-Centered Information at NCBI'. In: *Nucleic Acids Research* 33 (Database issue Jan. 1, 2005), pp. D54–58. DOI: 10.1093/nar/gki031 (cited on page 2).

[7] The UniProt Consortium. 'UniProt: The Universal Protein Knowledgebase in 2021'. In: *Nucleic Acids Research* 49.D1 (Jan. 8, 2021), pp. D480–D489. DOI: 10.1093/nar/gkaa1100. (Visited on 10/04/2021) (cited on page 2).

[8] Soichi Ogishima et al. 'AlzPathway, an Updated Map of Curated Signaling Pathways: Towards Deciphering Alzheimer's Disease Pathogenesis'. In: *Methods in Molecular Biology (Clifton, N.J.)* 1303 (2016), pp. 423–432. DOI: 10.1007/978-1-4939-2627-5_25 (cited on page 2).

[9] Kazuhiro A. Fujita et al. 'Integrating Pathways of Parkinson's Disease in a Molecular Interaction Map'. In: *Molecular Neurobiology* 49.1 (Feb. 1, 2014), pp. 88–102. DOI: 10.1007/s12035-013-8489-4. (Visited on 06/14/2021) (cited on page 2).

[10] Marek Ostaszewski et al. 'COVID-19 Disease Map, Building a Computational Repository of SARS-CoV-2 Virus-Host Interaction Mechanisms'. In: *Scientific Data* 7.1 (May 5, 2020), p. 136. DOI: 10.1038/s41597-020-0477-8 (cited on page 2).

[11] Satoshi Mizuno et al. 'Network Analysis of a Comprehensive Knowledge Repository Reveals a Dual Role for Ceramide in Alzheimer's Disease'. In: *PloS One* 11.2 (2016), e0148431. DOI: 10.1371/journal.pone.0148431 (cited on page 2).

[12] Akira Funahashi et al. 'CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks'. In: *Proceedings of the IEEE* 96.8 (Aug. 2008), pp. 1254–1265. DOI: 10.1109/JPROC.2008.925458 (cited on page 2).

[13] Piotr Gawron et al. 'MINERVA—a Platform for Visualization and Curation of Molecular Interaction Networks'. In: *npj Systems Biology and Applications* 2.1 (1 Sept. 22, 2016), pp. 1–6. DOI: 10.1038/npjsba.2016.20. (Visited on 10/04/2021) (cited on page 2).

[14] Paul Shannon et al. 'Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks'. In: *Genome Research* 13.11 (Jan. 11, 2003), pp. 2498–2504. DOI: 10.1101/gr.1239303. (Visited on 05/04/2021) (cited on page 2).

[15] Hendrik Rohn et al. 'VANTED v2: A Framework for Systems Biology Applications'. In: *BMC Systems Biology* 6.1 (2012), p. 139. DOI: 10.1186/1752-0509-6-139. (Visited on 04/24/2019) (cited on page 2).

[16] Sune S. Nielsen et al. 'Machine Learning to Support the Presentation of Complex Pathway Graphs'. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2019), pp. 1–1. DOI: 10.1109/TCBB.2019.2938501 (cited on page 6).

[17] David Duvenaud et al. *Convolutional Networks on Graphs for Learning Molecular Fingerprints*. Nov. 3, 2015. URL: http://arxiv.org/abs/1509.09292 (visited on 01/21/2021) (cited on page 9).