UNIVERSIDADE NOVA DE LISBOA

MScCBBi

MASTER IN
COMPUTATIONAL BIOLOGY & BIOINFORMATICS

SPECIALIZATION Multi-Omics for Life and Health Sciences

Maria Inês Nunes Vilar Gomes
BSc in Aquatic Sciences

Benchmarking Causal Reasoning Algorithms for enhanced Drug Discovery

September 2025

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

BENCHMARKING CAUSAL REASONING ALGORITHMS FOR ENHANCED DRUG DISCOVERY

SOME THOUGHTS ON THE LIFE, THE UNIVERSE, AND EVERYTHING ELSE

MARIA INÊS NUNES VILAR GOMES

BSc in Aquatic Sciences

Advisers: Dr. Filippo Ciceri
Senior Data Scientist, Clarivate

Dr. Cecilia Klein
Senior Manager Consulting, Clarivate

Co-adviser: Prof. Dr. Paula Maria Theriaga Mendes Bernardo Gonçalves
Associate Professor, NOVA University Lisbon

MASTER IN COMPUTATIONAL BIOLOGY AND BIOINFORMATICS
SPECIALIZATION IN MULTI-OMICS FOR LIFE AND HEALTH SCIENCES

NOVA University Lisbon
Draft: *January 16, 2025*

Abstract

Regardless of the language in which the dissertation is written, usually there are at least two abstracts: one abstract in the same language as the main text, and another abstract in some other language.

However, this order can be customized by adding one of the following to the file `5_packages.tex`.

```
\ntsetup{abstractorder={<LANG_1>,...,<LANG_N>}}
\ntsetup{abstractorder={<MAIN_LANG>={<LANG_1>,...,<LANG_N>}}}
```

For example, for a main document written in German with abstracts written in German, English and Italian (by this order) use:

```
\ntsetup{abstractorder={de={de,en,it}}}
```

Concerning its contents, the abstracts should not exceed one page and may answer the following questions (it is essential to adapt to the usual practices of your scientific area):

1. What is the problem?
2. Why is this problem interesting/challenging?
3. What is the proposed approach/solution/contribution?
4. What results (implications/consequences) from the solution?

Keywords: One keyword, Another keyword, Yet another keyword, One keyword more, The last keyword

Resumo

Independentemente da língua em que a dissertação está escrita, geralmente esta contém pelo menos dois resumos: um resumo na mesma língua do texto principal e outro resumo numa outra língua. No entanto, esse pedido pode ser personalizado adicionando um dos seguintes ao arquivo `5_packages.tex`.

```
\abstractorder(<MAIN_LANG>):={<LANG_1>,...,<LANG_N>}
```

Por exemplo, para um documento escrito em Alemão com resumos em Alemão, Inglês e Italiano (por esta ordem), pode usar-se:

```
\ntsetup{abstractorder={de={de,en,it}}}
```

Relativamente ao seu conteúdo, os resumos não devem ultrapassar uma página e frequentemente tentam responder às seguintes questões (é imprescindível a adaptação às práticas habituais da sua área científica):

1. Qual é o problema?
2. Porque é que é um problema interessante/desafiante?
3. Qual é a proposta de abordagem/solução?
4. Quais são as consequências/resultados da solução proposta?

Palavras-chave: Primeira palavra-chave, Outra palavra-chave, Mais uma palavra-chave, A última palavra-chave

Contents

List of Figures

Glossary

CBDD Computational Biology for Drug Discovery - citation [1] *(p. 7)*

computer An electronic device which is capable of receiving information (data) in a particular form and of performing a sequence of operations in accordance with a predetermined but variable set of procedural instructions (program) to produce a result in the form of information or signals. This is a test that adds a citation [1] to the glossary! *(p. 7)*

Acronyms

aaa acronym aaa *(p. 7)*

aab acronym aab *(p. 7)*

aba acronym aba *(p. 7)*

abbrev abbreviation of a longer text *(p. 7)*

bbb acronym bbb *(p. 7)*

xpto and extension of a xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto *(p. 7)*

Symbols

μ Mu *(p. 7)*

π the numerical value of pi *(p. 7)*

r the radius of a circle *(p. 7)*

1

Introduction

This section expounds the underlying motivation, rationale and goals for the study, emphasizing its significance in the field. It provides context by giving some background on the supporting company and the initiative. Furthermore, it outlies a read's guide of this thesis.

1.1 Motivation and Goals

The Research and development (R&D) of new drugs is a fast-growing area that has also experienced significant growth in complexity in recent years. Due to the time-consuming, costly and multidisciplinary nature of the process, drug discovery remains a challenging domain. Over half of clinical trial failures are attributed to inefficiency, underscoring the importance of identifying and validating pharmacological targets, and highlighting the lack of knowledge of the drug's mechanism of action (MoA) as one of the major barriers in drug discovery. The thorough understanding of the MoA represents a critical initial step in this process, and computational methods can accelerate this by providing more efficient and cost-effective alternatives to traditional approaches. These approaches can accurately do target identification and prioritization, thereby reducing the need for lengthy experimental trials.

A key to understanding a compound's MoA lies in transcriptomics data, which captures the molecular changes triggered by a perturbagen and reflects the system's changes in the gene expression profiles. While traditional RNA sequencing methods remain too costly for large-scale expression signatures, recent high-throughput technological advances, such as the L1000 assay, enable the cost-effective generation and analysis of large-scale omics datasets. Several existing databases provide public access to transcriptomic data from experiments involving diverse chemical and genetic perturbagens across different cell lines. These data can be exploited using various computational tools to establish the causes of specific gene expression changes in a biological system. Three primary approaches have emerged: causal reasoning, connectivity mapping and enrichment tools.

Causal reasoning, a topology-based method, utilizes a list of perturbation signatures and a biological interaction network to determine potential causes for the observed gene

expression profile. The network is defined as signed and directed graph describing relations between nodes (e.g., proteins). Efforts to assemble causal molecular relations have increased, resulting in several publicly accessible databases, such as OmniPath, which offers curated prior knowledge networks, however, some causal information remains commercially available, such as MetaBaseTM developed and curated by Clarivate.

The connectivity mapping (CMap) method stems from the efforts to collect and analyze perturbation signatures. It employs similarity scoring to compare a set of known MoA/compound reference signatures with a query gene expression signature resulting from a perturbagen. The principle behind CMap: the higher the similarity between the query and the reference signature, the more likely it is that the mechanism underlying the observed gene expression changes is related to a known perturbation.

On the other hand, enrichment tools take perturbation signatures as query input and utilize prior knowledge, such as regulon network or collections of perturbation-induced differentially expression genes (DEGs), as a reference. The primary function of these tools is to assess whether certain regulons or gene sets (e.g., those associated with transcription factors (TFs)) are significantly enriched in the perturbed data. Several algorithms have been developed based on this approach, each producing an enrichment score.

1.2 Scope

This project was conducted within the framework of the Algorithm Benchmarking Consortium (ABC), a subscription-based initiative led by Clarivate for pharmaceutical companies. ABC is dedicated to evaluating a wide range of computational tools for a variety of applications in the life sciences and healthcare field. The topic for this thesis is the development of the ABC's tenth use case – Causal Regulation – which focuses on benchmark and identify the most optimal tools tailored to specific needs within the drug discovery process by identifying key regulators from transcriptomics data and prior knowledge graphs.

1.3 Parallel Contributions

This study expands the state of art in causal reasoning using gene expression data and causal graphs by presenting a robust framework and methods for benchmarking various algorithms designed for this purpose. Beyond this primary focus, several parallel projects with real-world challenges and novel data were developed, and enriched the consultant experience allowing for an expansion in the expertise of bioinformatics. The parallel projects, spanning various domains of computational biology, include:

Skin Microbiome Atlas The skin microbiome atlas project involved extensive scientific literature review and dataset curation, followed by a systematic re-analysis of available datasets to ensure consistency and reliability. This was done by pre-processing

the raw sequencing data (. . . understand how deep I can go In terms of details – confidentiality issues)

More projects ... More projects ...

1.4 Structure

This study is organized in ? chapters. Chapter 2 introduces

2

Literature Review

This section details the

2.1 Drug discovery: the importance of the compound's mechanism of action

2.2 Causal Reasoning algorithms as an approach for target inference

3

Materials and Methods

3.1 Gene expression data

4

Adding Support to a New School (work in progress)

```
...
|
+-- nova
|   +-- Images
|   +-- fct
|   |   \-- Images
|   +-- ims
|   |   \-- Images
|   ...
|
\-- uminho
    +-- Images
    +-- ea
    |   \-- Images
    +-- ec
    |   \-- Images
    ...
```

The directory `uminho` contains the customization for all Schools of Universidade do Minho. This university is an example of the case where the regulations are defined at University level and all the schools apply the same thesis layout and organization. So, the all the customization is done in the file `uminho/uminho-defaults.ldf`, except the definition of the name and logo of each individual school.

As another example, the directory `nova` contains the customization for all Schools from NOVA University Lisbon. This university grants a lot of freedom in the definition of the thesis layouts. In some cases, they are defined at the School level (e.g., NOVA FCT), while is some other cases they are defined separately for each degree (e.g., NOVA IMS).

1. Try all the already supported schools and check which one is closer to your needs;
   a) Edit `Config/1_novathesis.tex` and near line 28 uncomment the line with key `\ntsetup{school=<SOMETHING>}`;
   b) For each school supported (see the comment), replace `<SOMETHING>` with the school name, e.g., make it `\ntsetup{school=ulisboa/fmv}`
   c) Recompile and check the document. Particularly, check the cover layout, the front-page (second cover) layout, the front-matter contents, the bibliography style;
   d) Repeat for the next school, until you find one close enough.
2. ...

This is the first occurrence of an abbreviation: abbreviation of a longer text (abbrev). And now the second occurrence of the same abbreviation: abbrev. And a new acronym with capital letter: And extension of a xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto xpto (xpto) and reused xpto. Let's also use a few other acronyms such as acronym aaa (aaa), acronym aab (aab), acronym aba (aba), acronym bbb (bbb) and xpto. In geometry, the area enclosed by a circle of radius r is πr^2. Here the Greek letter π is equal to the ratio of the circumference of any circle to its diameter. Lets add "computer" to the glossary! Be carefull with mathematical symbols in acronyms, please see the definition of μ.

CBDD

References

[1] *CBDD*. Web Page. URL: https://clarivate.com/life-sciences-healthcare/consulting-services/research-and-development-consulting/cbdd/ (cit. on p. vi).

[2] R. J. Dias et al. "Verification of Snapshot Isolation in Transactional Memory Java Programs". In: *Proceedings of the 26th European conference on Object-oriented programming (ECOOP'12)*. Springer-Verlag, 2012-06 (cit. on p. 11).

A

NOVAthesis covers showcase

This Appendix shows examples of covers for some of the supported Schools. When the Schools have very similar covers (e.g., all the schools from Universidade do Minho), just one cover is shown. If the covers for MSc dissertations and PhD thesis are considerable different (e.g., for FCT-NOVA and UMinho), then both are shown.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

A.1 A section here

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna

fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

B

Appendix 2 Lorem Ipsum

This is a test with citing something [2] in the appendix.

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

I

Annex 1 Lorem Ipsum

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum

wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

UNIVERSIDADE NOVA
DE LISBOA