

Metabolomics database resolver

Rajmund Csombordi

Supervisor: Sara Yones

Subject Reader: Johan Viklund

# Abstract

Asd

MetaFetcher - accessing metabolomics data in a simplified way

Thesis documentation

Rajmund Csombordi

This text should be a popular science text, that is, directed to readers with no college education in bioinformatics or molecular biotechnology. This means you cannot use your usual abstract, but you must write with simpler words, avoid overly complicated sentences and even avoid - or explain - scientific terms that are not widely known by people in general. In many cases you can also write a more popular science adapted title, instead of the more scientific on of your report. You often need to reduce/select the content, avoiding too many details, and instead pick out the most important results or "highlights" from your studies.

The summary can be 1-2 pages. The main text should be adjusted to the left and Times New Roman 12 points is an appropriate font. The title and your name should be centred and made larger than the main text. It is not allowed to include pictures. Please note that the Swedish-speaking students are required to write this summary in Swedish. If you are not fluent in Swedish, you should write in English.

Some formal, general, information should be added to the bottom of the page in the same way as on the front of your report, and will vary with the type and length of your thesis. Use the automatically generated text from the front of the report as a template (see example below).

The table of contents should be placed on a right-hand side (odd page) in the document.

Degree project in bioinformatics, 2016  
Examensarbete i bioinformatik 30 hp till masterexamen, 2016   
Biology Education Centre and Biomedical Center  
Supervisor: Sara Yones

* Introduction
* Background
* Materials and methods
* Theory
* Results
* Discussion
* Conclusion
* Acknowledgement

Table of Contents

[Abstract 4](#_Toc39782314)

[Terminology and abbreviations 8](#_Toc39782315)

[Introduction and Background 9](#_Toc39782316)

[Metabolomics 9](#_Toc39782317)

[Problems with databases 9](#_Toc39782318)

[Design Principles 9](#_Toc39782319)

[Methods 10](#_Toc39782320)

[Internal Database 10](#_Toc39782321)

[Bulk insertions 10](#_Toc39782322)

[The queue algorithm 10](#_Toc39782323)

[Data sources 10](#_Toc39782324)

[Theory 11](#_Toc39782325)

[Results 11](#_Toc39782326)

[Installation 11](#_Toc39782327)

[Usage 11](#_Toc39782328)

[Resolving without an input dataframe: 12](#_Toc39782329)

[Coverage tests 12](#_Toc39782330)

[About the collected data 12](#_Toc39782331)

[1.1.1 Materials and Methods 13](#_Toc39782332)

[1.1.2 Results and Discussion 13](#_Toc39782333)

[1.1.3 Acknowledgement 14](#_Toc39782334)

[1.1.4 References in the text 14](#_Toc39782335)

[1.1.5 Tables and Figures 14](#_Toc39782336)

[1.2 Other parts outside the main text 15](#_Toc39782337)

[1.3 Layout 15](#_Toc39782338)

[2 Concluding part: References and appendix 15](#_Toc39782339)

[2.1 References 15](#_Toc39782340)

[2.2 Appendix 15](#_Toc39782341)

[3 Concerning writing 16](#_Toc39782342)

[4 What is required for an approved final report? 16](#_Toc39782343)

[X-heading 1 introductory part: Arial normal style 18 points, line spacing 1.15, 36 points before and 12 after the paragraph; not numbered in body 18](#_Toc39782344)

[X-heading 2: Arial normal style 14 points; line spacing 1.15; 18 points before and 12 after the paragraph; numbered 18](#_Toc39782345)

[X-heading 3: Arial bold style 12 points; line spacing 1.15; 10 points before and 0 after the paragraph; numbered 18](#_Toc39782346)

# Terminology and abbreviations

ATP adenosine triphosphate  
DNA deoxyribonucleic acid

**todo**

# Introduction and Background

## Metabolomics

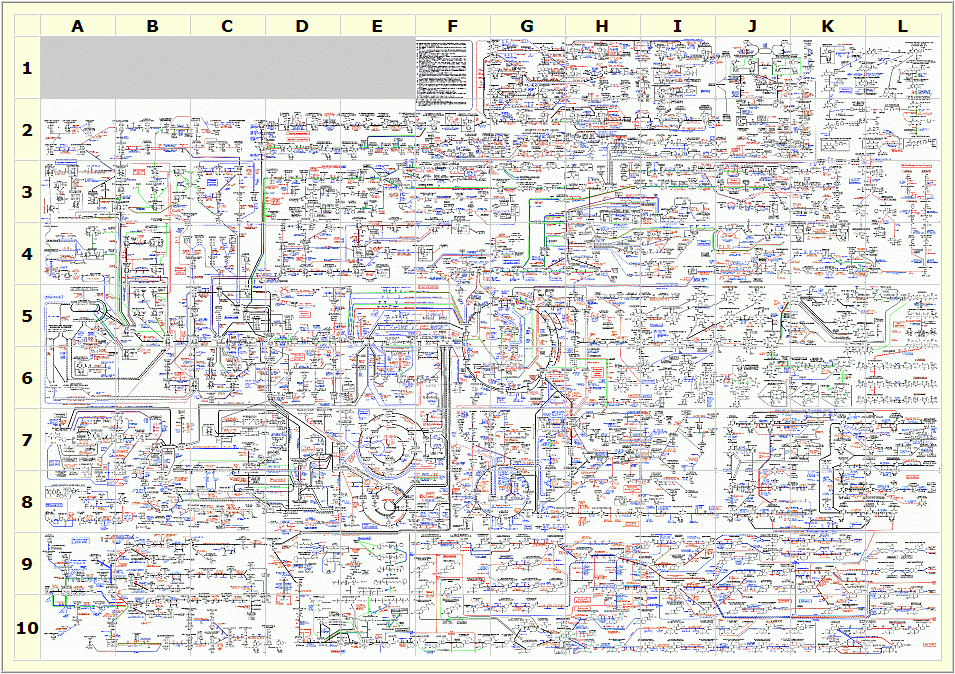
Metabolomics is the study of small molecule substrates and compounds that take part in metabolic processes. Many leading causes of death can be traced to metabolic disorders – which usually occur from an imbalance of metabolites – making it an especially important field of study. Metabolites are the substances occurring in metabolic pathways, they can be either an immediate product or the end result of a pathway.

Figure : an overview of Human metabolic pathways

Therefore, it is in the interest of the scientific community to identify, label and reliably store these substances and compounds in metabolome databases – in a similar manner to gene and protein databases.

## Problems with databases

Metabolome databases are quite numerous, during the work of this thesis we have addressed around 80 different databases. While databases in classical bioinformatics are well interconnected, curated and provide a consistent interface for users, the same can’t be said about metabolomics.

First of all, each database uses its own way of providing search, access and download features for researchers. Generally speaking the more databases we use in a research, the more cumbersome the start stage is, in which we develop scripts for data parsing and storage. This type work shouldn’t be the responsibility of a scientist.

To make things worse, metabolomics data is scarce and scattered. Different databases can have very little overlap in what metabolites they chose store. Some databases like PubChem have a great overlap with other databases, because they store a vast amount of compounds relative to other databases. A database record may reference other records from the same database or a foreign one. In rare cases this reference points to multiple records within the same database or points to a non-existing record, and sometimes it points to the wrong record within the database.

Unlike proteins, metabolites do not have a standardized identifier system, instead scientists are forced to rely on primary IDs from a database of their pick. Many of these databases do use chemical identifier systems such as SMILES or InChI, but these identify metabolites explicitly from the structure and fail to create an unambiguous, consistent system to identify molecules (see later chapters).

Some databases store multiple types of IDs, because they sort metabolites into different categories. Others have records with multiple IDs referring to the same metabolite record – these are referred as secondary IDs in this document.

## The package

All of these issues are making data preparation more difficult and discourages progress. To account for them we present our R package ‘MetaFetcher’ that is aimed to ease up preparation of a metabolomics research project. It’s primary purpose is resolving missing references between different kinds of metabolome databases. You can read the installation and usage guide in the Usage chapter of this document, or refer to the CRAN manual.

The package was designed to have a short run time whilst resolving IDs. This is achieved by relying on a local database on the user’s machine, serving both as a cache and a local copy of external databases.

# Methods

## Internal Database

The underlying component behind the package is the local database. The package gathers information from external metabolome databases, then store it in the local one. It also acts as a cache for api calls, serving as a way to reduce spamming external databases. The algorithm will not query the api if the underlying record is already found in the local database.

## Data sources

**Human Metabolome Database (HMDB)** stores approximately 114k metabolites found in the human body. They provide XML format for their web service and bulk download format.

**Chemical Entities of Biological Interest (ChEBI)** is EMBL-EBI’s database for small molecular entities that are involved in the processes of living organisms. They provide various formats for both web service and bulk downloading.

**PubChem** is a massive database of approximately 103M chemical compounds. Due to its size we rely entirely on its web service for accessing compounds.

**LipidMaps** is a database containing ~44k lipids. They provide bulk download possibility in SDF format.

**Kyoto Encyclopedia of Genes and Genomes** offers a compound database as well. KEGG does not support bulk download for free, so we rely on its web service access solely.

## Bulk insertions

The resolving algorithm relies on API fetches and a local database acting as a cache. However, fetching new records in bulks slows down the execution time noticeably. To account for this problem the package provides a possibility to download all records from a remote database into the local cache. The databases have to support mass downloading for this, therefore, this option is only available HMDB, ChEBI and Lipidmaps. Each database have its own bulk output format which had to be accounted for when this feature was written.

Running the bulk insertions scripts on my local computer took around 1 hour in total for all three databases. This step is required for users, as the current version of the package does not support API fetching for databases that have bulk insertion option.

# Theory

In this chapter we will elaborate on the details and demonstrate a proof of the correctness of the queue-based discovery algorithm.

## The discovery algorithm

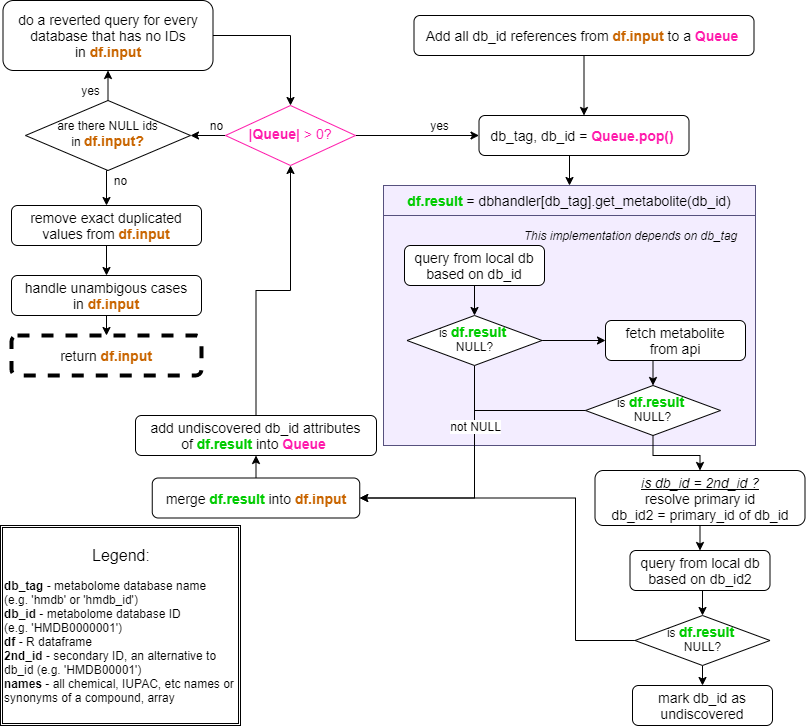
The algorithm resolves IDs by fetching the appropriate record and scanning it for additional database IDs. This is orchestrated via a queue-based algorithm that puts discovered IDs one by one and fetches the record on the top of the queue. The algorithm guarantees finding all relevant IDs by keeping track of already discovered records and not putting them into the queue twice.

The algorithm also supports reverse-querying records. After emptying the queue, the algorithm checks if there are external IDs that are null. Reverse-queries are then issued for each missing ID. A reverse-query means that the database is queried based on its foreign keys, for example:

SELECT chebi\_id FROM chebi\_data WHERE hmdb\_id = ’…’

The algorithm also supports resolving secondary IDs – IDs that do not have a record in the database, but instead these point to another ID. These can occur when a record is merged into another one, linking its primary ID to that of the other record’s.

Below is an overview flow chart that depicts the complete logic of the discovery algorithm.



# Results

## Installation

**todo** @later

## Usage

The main feature of the package is resolving missing IDs in a user-prepared dataframe. The user can chose which columns to include in the dataframe, please keep in mind that only these columns will be used as the bases of the discovery algorithm. For example, if the user excludes ''pubchem\\_id'', then pubchem won't be queried during discovery. The possible column names are:

* chebi\_id
* hmdb\_id
* lipidmaps\_id
* kegg\_id
* pubchem\_id
* inchi
* inchikey
* smiles
* names
* formula
* mass
* monoisotopic\_mass

For example, the user can rely on a CSV to resolve missing IDs the following way:

# discovery.csv:

hmdb\_id,chebi\_id,inchi,mass

,8337,,

HMDB0001008,,,

By loading this csv the user can add it straight into the resolve function, given that only the listed columns are present.

df.res <- read.csv("discovery.csv", stringsAsFactors=FALSE)

resp <- resolve(df.res)

The output will be a list containing the filled dataframe, and sets of unresolved and ambigous cases. For more information, please check the manual of the package.

# resulting dataframe:

df.out <- revert\_df(resp$df)

### Resolving without an input dataframe:

There's an additional, simplified interface for the package. By calling "resolve\_single\_id" the user only has to provide one database ID to start the discovery algorithm from. This interface is encouraged for simpler use-cases. The function's output is the same.

# simplified interface:

resp <- resolve\_single\_id('hmdb\_id', 'HMDB0035495')

# resulting dataframe:

df.out <- revert\_df(resp$df)

## Coverage tests

**todo**

## About the collected data

To get a more complete understanding of the underlying data for each database, I've created scripts to measure the cardinality of the attributes and external references of metabolite records.

The below table shows how many records in each database has a cardinality larger than 1. 0 means a cardinality of 1.

Since I do not have direct access to the entire database of Kegg and Pubchem, I took a random sample of 4000 records and based the statistics on them.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **hmdb** | **chebi** | **pubchem** | **Kegg** | **lipidmap** |
| **synonyms** | - | 34346 | 795 | 1900 | - |
| **formula** | 0 | 668 | 0 | 0 | 0 |
| **smiles** | 0 | 0 | 800 | - | 0 |
| **inchi/key** | 0 | 1 | 0 | - | 0 |
| **Hmdb id** | 0 | 8 | 0 | - | 0 |
| **Chebi id** | 0 | 0 | 4 | 400 | 0 |
| **Pubchem id** | 0 | 2 | 0 | 0 | 0 |
| **Kegg id** | 0 | 109 | 0 | 0 | 0 |
| **Lipidmap id** | 0 | 13 | - | 23 | 0 |

This table is important because attributes with multiple possible values should be regarded as arrays. However, attributes where most records have a cardinality over 1 can be represented as scalar values in the database, while arrays can be stored in an optional extra column.

## Issues with databases

The issues with chemical structure formats

SMILES and InChI are two data formats that describe chemical structures with ASCII characters. It’d be intuitive to assume that the same metabolite gets the same SMILES or InChI, but this is not the case.

For example the Betulinic acid has these SMILES values in different dabatases:

**Chebi (3087):**

[H][C@]12CC[C@]3([H])[C@@]4(C)CC[C@H](O)C(C)(C)[C@]4([H])CC[C@@]3(C)[C@]1(C)CC[C@]1(CC[C@@H](C(C)=C)[C@]21[H])C(O)=O

**HMDB (HMDB0030094)**

[H][C@]12[C@@H](CC[C@@]1(CC[C@]1(C)[C@]2([H])CC[C@]2([H])[C@@]3(C)CC[C@H](O)C(C)(C)[C@]3([H])CC[C@@]12C)C(O)=O)C(C)=C

**LipidMaps (LMPR0106140004)**:

C1[C@@]2(C)[C@@]([H])(CC[C@]3(C)[C@]2([H])CC[C@@]2([C@@]4([C@](CC[C@@]32C)(C(O)=O)CC[C@H]4C(C)=C)[H])[H])C(C)(C)[C@@H](O)C1

**PubChem (64971):**

CC(=C)C1CCC2(C1C3CCC4C5(CCC(C(C5CCC4(C3(CC2)C)C)(C)C)O)C)C(=O)O,CC(=C)[C@@H]1CC[C@]2([C@H]1[C@H]3CC[C@@H]4[C@]5(CC[C@@H](C([C@@H]5CC[C@]4([C@@]3(CC2)C)C)(C)C)O)C)C(=O)O

This is because SMILES is non-unique – meaning that the same compound can be represented with various strings. In comparison, for all 4 databases the InChI value is the same:

1S/C30H48O3/c1-18(2)19-10-15-30(25(32)33)17-16-28(6)20(24(19)30)8-9-22-27(5)13-12-23(31)26(3,4)21(27)11-14-29(22,28)7/h19-24,31H,1,8-17H2,2-7H3,(H,32,33)/t19-,20+,21-,22+,23-,24+,27-,28+,29+,30-/m0/s1

However, InChI can also incorporate differences in isotopes, charges, stereochemical layer and other chemical properties as well. This makes InChI strings belonging to the same compound differ, such as in the case of neohesperidin:

**Chebi (59016):**

1S/C28H34O15/c1-10-21(33)23(35)25(37)27(39-10)43-26-24(36)22(34)19(9-29)42-28(26)40-12-6-14(31)20-15(32)8-17(41-18(20)7-12)11-3-4-16(38-2)13(30)5-11/h3-7,10,17,19,21-31,33-37H,8-9H2,1-2H3/t10-,17-,19+,21-,22+,23+,24-,25+,26+,27-,28+/m0/s1

**HMDB (HMDB0030748):**

1S/C28H34O15/c1-10-21(33)23(35)25(37)27(39-10)43-26-24(36)22(34)19(9-29)42-28(26)40-12-6-14(31)20-15(32)8-17(41-18(20)7-12)11-3-4-16(38-2)13(30)5-11/h3-7,10,17,19,21-31,33-37H,8-9H2,1-2H3

**LipidMaps (LMPK12140452):**

1S/C28H34O15/c1-10-21(33)23(35)25(37)27(39-10)43-26-24(36)22(34)19(9-29)42-28(26)40-12-6-14(31)20-15(32)8-17(41-18(20)7-12)11-3-4-16(38-2)13(30)5-11/h3-7,10,17,19,21-31,33-37H,8-9H2,1-2H3/t10?,17-,19?,21?,22?,23?,24?,25?,26?,27?,28?/m0/s1

**PubChem (442439):**

1S/C28H34O15/c1-10-21(33)23(35)25(37)27(39-10)43-26-24(36)22(34)19(9-29)42-28(26)40-12-6-14(31)20-15(32)8-17(41-18(20)7-12)11-3-4-16(38-2)13(30)5-11/h3-7,10,17,19,21-31,33-37H,8-9H2,1-2H3/t10-,17-,19+,21-,22+,23+,24-,25+,26+,27-,28+/m0/s1

Perhaps visiting the possibility of using InChI strings (with the exclusion the optional format extensions after the “/” symbol) for chemical-based searches in metabolites would be beneficial both for this package and in the construction of a non-structure based metabolome identifier as well in the future.

By comparing all records in HMDB, ChEBI and LipidMaps we find out that **XX**% of records do …

**Show: records that have the same pubchem\_chebi id if their SMILES isn’t**

**%-ban kifejezve!**

### Materials and Methods

After introducing the project, you describe how it was carried out. You must account for all the experiments you made and the basic rule is that it should be possible to repeat the experiments by reading the description. There are different traditions on how to describe experiments in different disciplines. Read Presenting Science (Rydin *et al.* 2014) and consult your supervisor or subject reader if you are unsure on how to describe the methods in your report.

### Results and Discussion

In the results section you write what you did and which results you got. Structure the results section in a way that makes the text easy to follow and understand. In the discussion section, put the results in a context and return to the questions you presented in the introduction. How does your work relate to previous research and development?

Also note whether there are other aspects of the work that may be of interest, but that may not always be obvious to people in the relevant field of research. Are there for example ethical aspects on what you have done, or what your work is aimed at? What role does your work have in society at large? Is your work a contribution to a more sustainable world? Mention relevant aspects appropriately in your report and make it more interesting for a wider readership. See Presenting Science (Rydin *et al.* 2014) for further instructions on how to write the results and discussion section.

### Acknowledgement

Sometimes there may be reasons to thank someone who contributed to the work. This is done at the end of the main body, before the references.

### References in the text

You need to provide references for all data in the report that is not your own, or commonly known. You can read about how to reference on the IBG's website (IBG 2016b) and in Presenting Science (Rydin *et al.* 2014). For Zotero there is a ready template for IBG's reference system to download from the IBG website (<https://www.ibg.uu.se/student-en/support-service/reference-management/>).

### Tables and Figures

Often it is convenient to describe methods, results and conclusions in the form of a figure or a table. Use, as far as possible, self-produced images and tables. Create your illustrations with software that is designed for it. If you use previously published material, you must obtain permission from the copyright holder. State this, and the source, last in the figure caption: "Illustration used with permission from ...". How to design tables and figures and how you refer to them in the text are described in detail in Presenting Science (Rydin *et al.* 2014). Make sure to refer to all figures and tables in the text, number them in the order they appear in the text, and insert them into the document in order. That is, Figure 1 is referenced to in the text before Figure 2 and also placed before Figure 2 in the document. An easy way to get this right is to use cross-references.

When referencing the figure in the text you use insert-cross-reference and select figure, and “only label and number” as within this parenthesis (Figure 1). If a figure is inserted before another in the text, just mark the entire text after inserting the figure and figure text, and update the field. Then all figure numbers and cross-references in the text will be updated.

## Other parts outside the main text

In addition to tables and figures it might be appropriate to use other elements in the report that do not belong to the text, for example equations, chemical reactions or programming code. Make sure to then follow the practice of the subject area in which you have done your work. Consult your supervisor or subject reader if you want advise.

## Layout

A report with good layout gives a good impression. The basis is that you use this template. Other things to consider are:

* make page breaks at appropriate places to minimize empty spaces in the report,
* gather the body text - avoid single lines of running text above or below tables and figures,
* keep figures and tables within the margins used for the text,
* make figures and tables as uniform as possible with respect to font, font size and colors,
* avoid naked headings, that is, when there is no text between two headings at different levels.

# Concluding part: References and appendix

At the end, create a list of references, and include an appendix if necessary. Use the style **X-1 heading concluding part** for headings.

## References

Every scientific journal has its own standard for references. At IBG the requirement is that you follow the instructions in Presenting Science (Rydin *et al.* 2014). This you can do, for example, by using Zotero and the available template for the reference system.

## Appendix

In the appendix, you can account for results that are too extensive to be included in the main report, but motivated to show. The appendix always starts on a new page and is provided as Appendix A, Appendix B, etc. Each appendix begins on a new page.

# Concerning writing

If you have problems with the language, the subject reader may refer you to the Language Workshop ([www.sprakverkstaden.uu.se/](http://www.sprakverkstaden.uu.se/)), which then gives you support in the writing process. You can also contact the Language Workshop yourself if you want.

# What is required for an approved final report?

The supervisor and subject reader will provide feedback on your report. When the subject reader has approved the report, send it to the coordinator. We read it and make the final assessment. What we are mainly looking at is if:

* the work is scientifically accurate and understandable for a person with general knowledge in biotechnology or bioinformatics,
* outline and layout follow the instructions in this document and Presenting Science (Rydin *et al.* 2014),
* the work is set in its social context, that is, that a person outside the research field understand why the work has been done and what the purpose was,
* the work has been carried out in a responsible manner, for example, that there are permits for any animal testing, and that you have considered the potential ethical aspects.

If you meet all the criteria your report will be approved.

References

IBG. 2016a. Degree project - Disciplinary Domain of Science and Technology - Uppsala University, Sweden. WWW document 26 October 2016: [http://teknat.uu.se/student-en/programmes-courses/degree-project/](https://teknat.uu.se/student-en/programmes-courses/degree-project/). Accessed 26 October 2016.

IBG. 2016b. Reference management - Biology Education Centre - Uppsala University, Sweden. WWW document 4 May 2016: <http://www.ibg.uu.se/student-en/support-service/reference-management/>. Accessed 26 October 2016.

Rydin H, Carlson K, Berglund A, Svensson BG. 2014. Presenting Science. Biology Education Centre

TEKNAT. 2012. Guidelines and instructions for the degree project course within the technical programs at Uppsala University.

Appendix A

If you do not use the template, use the following formatting for the different parts of the report:

# X-heading 1 introductory part: Arial normal style 18 points, line spacing 1.15, 36 points before and 12 after the paragraph; not numbered in body

X-heading 1 main part: Arial normal style 18 points, line spacing 1.15, 36 points before and 12 after the paragraph; numbered in body

## X-heading 2: Arial normal style 14 points; line spacing 1.15; 18 points before and 12 after the paragraph; numbered

### X-heading 3: Arial bold style 12 points; line spacing 1.15; 10 points before and 0 after the paragraph; numbered

X-main text: Times New Roman normal style 12 points; line spacing 1.15; 0 points before and 12 after the paragraph

* X-bulleted list: Times New Roman normal style 12 points; line spacing 1.15; 0 points before and 12 after the paragraph

1. X-numbered list: Times New Roman normal style 12 points; line spacing 1.15; 0 points before and 12 after the paragraph

X-heading 1 concluding part: Arial normal style 18 points, line spacing 1.15, 36 points before and 12 after the paragraph; not numbered in body