# Molecular Clustering via Knowledge Mining from Biomedical Scientific Corpora

Panagiotis Hasapis, Dimitrios Ntalaperas, Christos C. Kannas, Aristos Aristodimou, Dimitrios Alexandrou, Thanassis Bouras, Christos Georgousopoulos, Athos Antoniades, Constantinos S. Pattichis and Andreas Constantinou

***Abstract*—In this paper, an architecture is presented that allows the extraction of argumentation clauses that might exist in publications, in order to perform molecular clustering on referenced molecules. Grammar rules are defined and used to identify sentences corresponding to argumentation being present in publications. The references of those molecules are then compiled as lists that include their structure definition in SMILES format. These lists are given as input to virtual screening tools and then to a molecular clustering tool, with the ultimate goal to classify molecules that are known to be prone to specific diseases, thus leading to the discovery of new drugs.**

## I. INTRODUCTION

In today's biomedical world, information is in abundance and in many forms. Large data repositories that contain information about molecules and compounds can be found online in many formats that are of use to biomedical researchers. These formats can range from simple CSV files to XML files using specific vocabularies designed for that purpose. Typically this information is in a semi or fully structured format stored in databases or structured files.

However, there is still a gap in the way unstructured information, such as publications, is handled. These pieces of data often contain a great amount of information that cannot be found in existing databases or in any other structured form. Examples include the identification of medical terms inside a document (i.e. names of drugs, molecules, compounds etc.) as well as arguments written for those molecules used in an in-silico experiment showing specific results.

Arguments relating one molecule to an organ (or disease) are of great interest to biomedical researchers for two main reasons: firstly, they acquire new knowledge for one specific compound in relation to an organ (or disease) and secondly, they provide them with a starting point for searching for alternative compounds with similar properties. In the fields of medicine, biotechnology and pharmacology, this process is part of drug discovery and it allows researchers to re-use other potentially interesting molecules in in-silico and in-vitro experiments. The in-silico experiments include among other a molecular clustering process.

In this paper a way of extracting argumentative propositions about drugs and the active ingredient molecule, in relation to organs and diseases is presented. The references of those drugs can then be used to retrieve additional information (such as their chemical structure in SMILE notation [1]) from the Linked Data Cloud [2] (also known as Web of Data) and then sent to a molecular clustering process to discover similar molecules. The outcome of the latter are groups of molecules with similar structural properties that can be used in the same biomedical processes.

## II. RELATED WORK

Text mining processes from unstructured data has been the focus of many researchers during the last decade. In [3] and [4], the authors introduce solutions that provide examples of infrastructure for information extraction based on GATE. In order to provide basic level of performance and allow easy bootstrapping of applications, their platforms are also equipped with an upper-level ontology and a knowledge base, that provides extensive coverage of entities of general importance or from a specific biomedical sub-domain. Their approaches however constitute of small data sets with general entities that can be recognized (namely cities, countries and so on) and the RDF triples produced, only connect the occurrence of such a word to a small local knowledge repository. It is also worth noticing that most work in this field has been done by taking into account parts of the document such as the abstract and not the whole text.

In [5], the authors present a web service which allows users to explore PubMed search results with the Gene Ontology (GO), a hierarchically structured vocabulary for molecular biology. This approach provides several important benefits. Firstly, it gives an overview of the literature abstracts by categorizing abstracts according to the GO and thus allowing users to quickly navigate through the abstracts by category. Secondly, it automatically shows general ontology terms related to the original query, which often do not even appear directly in the abstract. It also en ables users to verify its classification because GO terms are highlighted in the abstracts and each term is labeled with an accuracy percentage. Finally, exploring PubMed abstracts with this tool is useful as it shows definitions of GO terms without the need for further look up. This approach however refers to structured data repositories, and does not mine information from unstructured resources such as scientific publications.

P. Hasapis and C.Georgousopoulos are with INTRASOFT International, 2b Rue Nicolas Bove, Luxembourg, Luxembourg (e-mail: {Panagiotis.Hasapis, Christos.Georgousopoulos}@intrasoft-intl.com)

D. Ntalaperas , Th. Bouras and D. Alexandrou are with UBITECH Research, 429 Messogion Avenue, Ag. Paraskevi Square, 153 42 Athens, Greece (e-mail: {dntalaper, bouras, dalexandrou}@ubitech.eu)

C. C. Kannas, A. Aristodimou, A. Antoniadesand C. S. Pattichis are with the Department of Computer Science, University of Cyprus (e-mail:{kannas.christos,aarist07,athos,pattichi}@cs.ucy.ac.cy)

A. Constantinou is with the Department of Biological Sciences, University of Cyprus (andreasc@ucy.ac.cy)

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

Data mining clustering approaches are often used in drug discovery processes. Using clustering, molecule groups with similar structural and chemical characteristics can be created. Various clustering methods [6], [7], [8], [9] can be used for this procedure. In the majority of the cases hierarchical clustering is used [10]. Most commonly, clustering is based on the structural characteristics of the molecules, which is possible with the use of structural fingerprints. This procedure is possible with the use of LiSIs [11] and ChemBioServer [12] platforms, which will be presented in the next section.

## III. ARCHITECTURE

### A. Overall Process Presentation

The system is comprised by several components presented in Fig. 1. The first component, "GATE and Argument Extraction" is related to recognizing and extracting arguments from scientific corpora. These arguments come in RDF format and linked with Linked Data Entities URI (of molecules, diseases etc), provided that such references are made in the argument sentence. These Linked Data URI's are gathered by "Biomedical Triplestore Crawler". A compilation of a molecule list is created according to certain criteria (expressed via SPARQL), then compiled and sent to "LiSiS Molecular Clustering" procedure.

All components are presented in detail in the next sections.

### B. GATE and Argument Extraction

The first process in the pipeline is GATE (General Architecture for Text Engineering). GATE is open source software that can be used in a wide category of text processing problems. Its platform includes a desktop client for developers, a workflow-based web application, a Java library, architecture and a process. On top of the core functions GATE includes components for diverse language processing tasks (such as parsers, morphology, tagging etc). GATE Developer and Embedded are also supplied with an Information Extraction system which has been widely adapted.

Section IV provides a detailed description of the text mining procedure.

### C. Biomedical Triplestore Crawler

The biomedical triplestore RDF crawler is responsible for querying triple stores found in the Linked Data Space. A small list of several public biomedical triples stores (such as Bio2RDF) was compiled, and a simple mapping of basic concepts of interest was created. A batch process was initiated that queries those triple stores and downloads basic triples (e.g. a name, a SMILE string or a link to an SDF file for a molecule) for each of our main concepts of interest (Protein, Molecule, Drug, Organ, Disease). Then a local repository of triples was created that allows retrieving information about entity names that are used by the GATE Gazetteers. The RDF Crawler performs this task every month and updates the data found in the triple store.

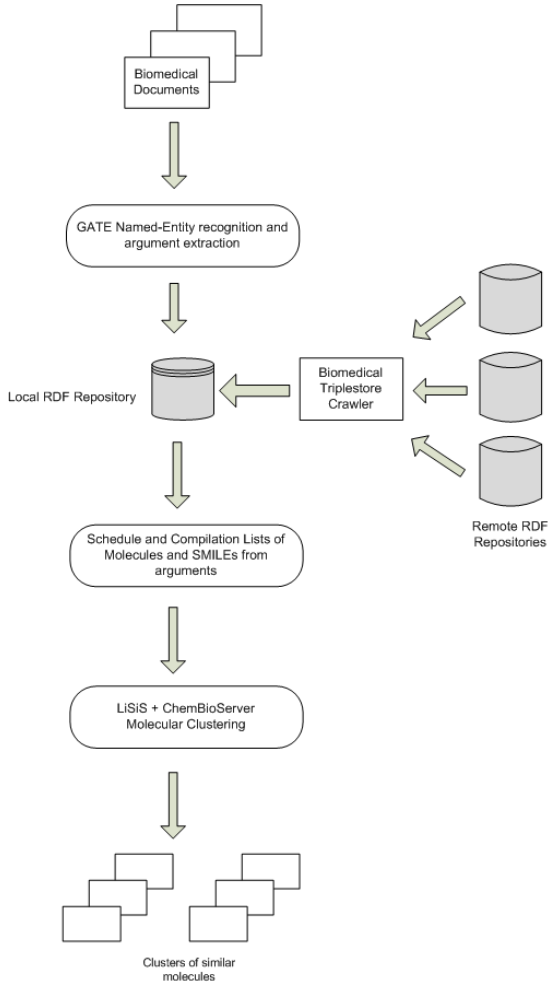


Fig. 1. Overall system process/pipeline

### D. RDF Repository

For the local triplestore installation, Jena Fuseki was chosen. The latter is an RDF server, which provides REST-style SPARQL HTTP Update, SPARQL Query, and SPARQL Update using the SPARQL protocol over HTTP.

### E. Compilation of drug /molecule lists

In order to compile a list of molecules a SPARQL query is created that allows finding arguments that show the relation between a molecule to a specific disease or organ, that a user specifies. From the returned arguments, the referenced compounds are selected, and another query is performed to retrieve the name and their respective SMILES. This list is compiled as a CSV file and it is sent to LiSIs.

### F. LiSIs and ChemBioServer Molecular Clustering

LiSIs (Life Science Informatics) is a scientific workflow platform for virtual screening. LiSIs provides the environment and the tools to design and execute workflows for screening in-silico molecules using their chemical and structural information. The structural information of molecules is represented by structural fingerprints, such as Morgan Fingerprints and MACCS 166 keys fingerprints. Fingerprints are used to find structural similarities among molecules. The most common measure used for molecular

similarity is the Tanimoto coefficient [13]. This measure gives values in the range of [0, 1] where 1 is for identical molecules and 0 for completely different ones. For the purposes of clustering, especially for methods that require a distance measurement between molecules, the Soergel coefficient is used [14]. The Soergel coefficient is in the range of [0, 1] where 0 is for identical molecules and 1 is for completely different ones. Soergel can be calculated as (1 – Tanimoto).

Clustering is performed on the ChemBioServer, a platform which facilitates computational compound screening and analysis tools, with the use of available R [15] hierarchical clustering methods. For the purposes of clustering, molecule structural information via the use of MACCS 166 keys fingerprints and the Tanimoto coefficient is used.

Section V provides a detailed description of the clustering procedure.

## IV. BIOMEDICAL CORPORA TEXT MINING PROCESS

The basic aim is for a grammar to be developed that generates sentences corresponding to argumentation, which is able to identify meaningful statements that express some sort of opinion. Authors of scientific publications tend to adopt commonly used norms when they express opinion or elaborate on some results or data. For the purposes of this work two main cases were distinguished, in which the keyword "suggest" is used as an example:

1. The author(s) use the keyword in its verb form, therefore expressing an argument, e.g. "We therefore suggest…" or "…according to our results it is suggested…"
2. The keyword is used as a noun. Argumentation is formed using phrases like "we make a suggestion", "a suggestion can be put forth"

Using these general rules, we can provide the resulting grammar rules that were used in a more formal way. By using a GATE Gazetteer (after performing the tokenization of the document) the sentences that contain the SynSet retrieved from WordNet can be extracted using a JAPE Transducer. Hence for each sentence extracted, the following checks are made:

*1) If the word from SynSet is a verb then the sentence is classified as an argument.*

*2) If the word is a noun, then the algorithm checks whether a verb in present tense appears before the word.*

i) If it does, then the sentence is classified as an argument.

ii) If it does not, a further check is made to see whether the part of the sentence after the word, forms a secondary sentence that refers to the conclusion of the first part. This check is made by checking whether there is a verb in past tense being present to that part of the sentence. If it does, then the sentence is classified as an argument.

3) *The final step is to associate a term found in a argument sentence with a Linked Data URL (if that is possible).* The word lists created (before the initialisation of the prototype) associate a term (i.e. "Aspirin") with a specific Linked Data URI containing structured data associated with that particular term. For each such match, the triples are stored in a local RDF Triple Store.

```
Rule: Rule1
(
  {
    Sentence contains
    {Lookup.majorType ==     ArgumentVerb}
  }
)

Rule: Rule2a
(
  {Token.kind == "control"}
  ({Token})*
  {
   Lookup.majorType == "SecondaryVerb",
Token.category=="VB"
  }

  ({Token})*
  {Lookup.majorType == ArgumentWord}
    ({Token})*
    {Token.kind == "control"}
)

Rule: Rule2b
(
  {Token.kind == "control"}
  ({Token})*
  {Lookup.majorType == ArgumentWord}

  ({Token})*
  {
        Lookup.majorType == "SecondaryVerb",
              Token.category=="VB"
  }

  ({Token})*{Token.kind == "control"}
)
```

Fig 2. Grammar rules for argument extraction

A BNF description of the grammar is given in Fig. 2. Rule 1 checks whether a sentence contain a verb from the main Gazetteer which was constructed from the argumentation SynSet. Rule 2a must check the order in which different segments appear, so it parses each sentence in more detail. Condition '*Token.kind == "control"*' checks for sentence boundaries (character "."). Condition '*Lookup.majorType == "SeconcaryVerb"*' is used to identify words from the secondary Gazetteer. For the Rule 2a to hold, these verbs must also be in present tense. This is achieved by demanding that the condition '*Token.category=="VB"*' holds. Rule 3a works similarly, with the difference that the order of the *Lookup* annotations is reversed and the tense of the secondary verb is required to be in past tense.

## V. MOLECULAR CLUSTERING

This section describes the methodology that was used for molecular clustering using their chemical structure information. The procedure of the analysis was performed using the LiSIs platform and ChemBioServer. The process is as follows:

*1. Provide a file consisted of three columns;* the first column should be the Drug's Trade Name or Drug's Active Ingredient Name, the second column should be the trait of interest or if available multiple traits of interest and the last column contains one Linked Data Instance URI of that particular drug.

*2. By de-referencing the URI's the chemical information for each Drug in the data file is retrieved.* The chemical information needed is the SMILES chemical representation of the active ingredient for the drug.

*3. From the SMILES chemical representation it is possible to create an in-silico molecule via the use of the tools available at the LiSIs platform.* The in-silico molecules were used to calculate the structure based descriptors.

*4. From the in-silico molecules information, ChemBioServer generated structural descriptors fingerprints.* The fingerprints generated are MACCS 166 keys. These are used as the patterns, of each molecule, by the hierarchical clustering algorithm.

*5. In this step the molecular clustering takes place on ChemBioServer].* Clustering was performed using a Wards hierarchical clustering algorithm. The selection of clusters is based on a Tanimoto similarity threshold.

*6. Clusters are further processed to identify unique characteristics.* For each cluster a detailed analysis to identify the unique characteristics among its containing molecules takes place.

## VI. PRELIMINARY RESULTS

In our early prototype implementation we have extracted several arguments that are related to cancer chemo-preventive research. As an example of argument extraction, consider for example the following sentence: "These data might suggest an anti-cancer role of amiloride against gastric cancer with peritoneal metastasis, and might set the stage for a new therapy for gastric cancer."

The above sentence is recognized as an argumentation sentence (e.g. grammar rule 1). The sentence contains an instance of a molecule (amiloride) and a disease (gastric cancer) which are recognized by the corresponding gazetteers. These instances are linked to the argument sentence via the "appearsIn" property. For the "amiloride" molecule in particular, an RDF triple is composed of the form:"<http://www.chemspider.com/chemical-structure.15403.rdf> <argo:appearsIn> <:argument_102>", where argument_102 is the linked data instance created for the specific argument and http://www.chemspider.com/chemical-structure.15403.rdf is the URI corresponding to "amiloride". Queries can be made both against the arguments in order to see which of them contains references to the "amiloride" molecule and against "amiloride" to obtain arguments referring to it. As each argument contains several metadata the queries can be augmented to obtain more complex relations between molecules and entities such as authors, journal titles, year of publication etc.

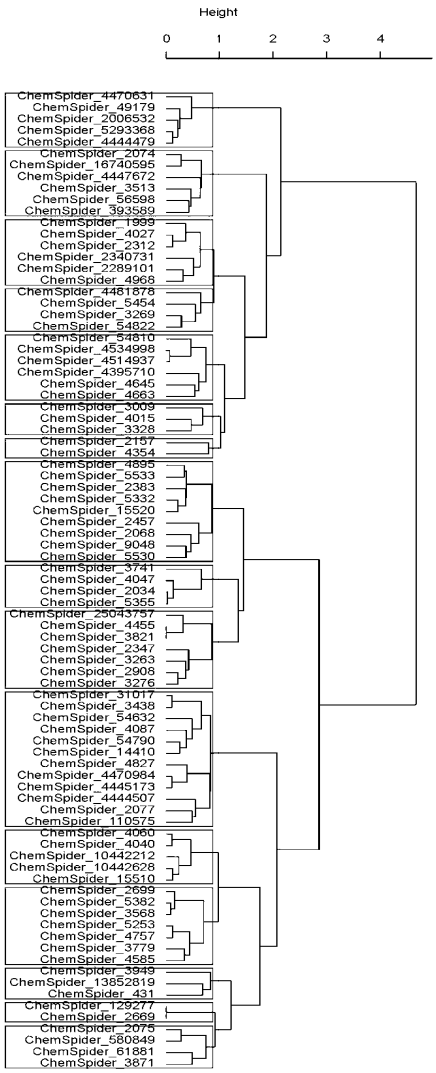


Fig. 3. Clusters of drugs identified using Wards hierarchical clustering. The rectangles indicate the drugs that belong to the same cluster.

For the preliminary clustering results, LiSIs and ChemBioServer were used as described in Section V. In Fig. 3, a list of ~90 drugs used in cardiology and respiratory disorders were used for a preliminary clustering, with the goal to identify which of those drugs have similar structural characteristics. A hierarchical algorithm (Wards) from ChemBioServer platform was used for clustering as described in Section V. In fig. 3 the results of the hierarchical clustering are illustrated. The Tanimoto similarity threshold (height axis in fig. 3) selected for separating the drugs into clusters was set to 0.85 in order to have a balance between the number of clusters created and the similarity of the molecules in each cluster. The rectangles in the figure indicate which molecules belong to the same cluster.

The results of this clustering are undergoing additional examination to identify unique characteristics among the molecules of each cluster. For each of these clusters a detailed analysis based on specific 'traits of interest' is currently undergoing.

## VII. DISCUSSION

In this paper a method of combing existing knowledge from argumentative discussion about molecules into data clustering processes was proposed. In the last few years, a great amount of research has been developed regarding the extraction of such information from publications and its linking to structure knowledge that is already available in the Web of Data. By connecting the argumentative sentence of a publication that refers to a compound or molecule, to the actual RDF representation of such a molecule, additional knowledge can be extracted, perform queries, and request the appropriate properties of that molecule which are not referred directly to the publication (such as its molecular weight or SMILE). By employing such techniques, the opinion about the properties of a molecule made by scientists in an in-vitro experiment is connected with other information about that molecule found available on the Web.

Molecular clustering is a methodology that helps researchers, identify commonalities among molecules based on chemical and/or structural characteristics. In the preliminary results shown in this paper, an initial clustering was performed for identifying molecules with similar chemical structures. Once molecules with such similarities are identified, such information can be used for testing for associations with specific traits of interest. An example is to test if molecules with certain similarities are associated with certain adverse events. Hence the actual chemical structures that might be associated with such adverse events can be identified and use this knowledge at the development of new drugs.

The proposed architecture is capable of combining the aforementioned methodologies and their merits. It creates an automatic way of scanning the biomedical scientific corpora, links the arguments extracted with the available digital representations and creates pipelines of in-silico experiments, thus drastically reducing eliminating the time needed for finding new promising compounds with medical applications.

## ACKNOWLEDGMENT

This work has been co-funded by the GRANATUM project, a European Commission research program under Contract Number FP7-270139 and by the Linked2Safety project, a European Commission research program under Contract Number FP7-288328.

## REFERENCES

[1] D. Weininger, “SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules,” J. Chem. Inf. Model., vol. 28, no. 1, pp. 31–36, 1988

[2] C. Bizer, T. Heath, and T. Berners-Lee. "Linked data-the story so far." International Journal on Semantic Web and Information Systems (IJSWIS) , pp. 1-22, 2009

[3] H. Cunningham et al., “GATE: an Architecture for Development of Robust HLT Applications”, ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 168-175, 2003

[4] H. Cunninghamet al., “Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics”, PLoS computational biology 9.2, 2009

[5] Andreas Doms et al., "GoPubMed: exploring PubMed with the gene ontology", Nucleic acids research 33.suppl 2: W783-W786, 2005

[6] J. W. Raymond, C. J. Blankley, and P. Willett, “Comparison of chemical clustering methods using graph- and fingerprint-based similarity measures,” Journal of Molecular Graphics and Modelling, vol. 21, no. 5, pp. 421–433, Mar. 2003.

[7] J. MacCuish, C. Nicolaou, and N. E. MacCuish, “Ties in Proximity and Clustering Compounds,” J. Chem. Inf. Comput. Sci., vol. 41, no. 1, pp. 134–146, Jan. 2001.

[8] R. A. Jarvis and E. A. Patrick, “Clustering Using a Similarity Measure Based on Shared Near Neighbors,” IEEE Trans. Comput., vol. C–22, no. 11, pp. 1025–1034, Nov. 1973.

[9] H. L. Morgan, “The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service.,” J. Chem. Inf. Model., vol. 5, no. 2, pp. 107–113, 1965.

[10] A. Böcker, S. Derksen, E. Schmidt, A. Teckentrup, and G. Schneider, “A Hierarchical Clustering Approach for Large Compound Libraries,” J. Chem. Inf. Model., vol. 45, no. 4, pp. 807–815, Jul. 2005.

[11] C. C. Kannas, K. G. Achilleos, Z. Antoniou, C. A. Nicolaou, C. S. Pattichis, I. Kalvari, I. Kirmitzoglou, and V. J. Promponas, “A workflow system for virtual screening in cancer chemoprevention,”, 2012 IEEE 12th International Conference on Bioinformatics Bioengineering (BIBE), pp. 439–446, 2012

[12] E. Athanasiadis, Z. Cournia, and G. Spyrou, “ChemBioServer: A web-based pipeline for filtering, clustering and visualization of chemical compounds used in drug discovery,” Bioinformatics, Sep. 2012

[13] J. W. Godden, L. Xue, and J. Bajorath, “Combinatorial Preferences Affect Molecular Similarity/Diversity Calculations Using Binary Fingerprints and Tanimoto Coefficients,” J. Chem. Inf. Comput. Sci., vol. 40, no. 1, pp. 163–166, Jan. 2000

[14] A. R. Leach and V. J. Gillet, An introduction to chemoinformatics, Rev. Ed. Springer, 2007.

[15] “The R Project for Statistical Computing.” Available: http://www.r-project.org/. [Accessed: 17-Jul-2012]

[16] Luis Tari et al. “Discovering drug-drug interactions: a text mining and reasoning approach based on properties of drug metabolism”, Bioinformatics 26.18, 2010, pp 547-553, 2010

[17] C. Sotriffer, Ed., Virtual Screening: Principles, Challenges, and Practical Guidelines. 2011.

[18] C. Tempich et al., “An argumentation ontology for distributed, loosely-controlled and evolving engineering processes of ontologies (DILIGENT)”, The Semantic Web: Research and Applications. Springer Berlin Heidelberg, pp. 241-256, 2004