IMPLEMENTATION OF VIRTUAL WORKFLOWS IN KNIME FOR MEDICINAL CHEMISTRY

Jack DiTommaso
Grade 12, The Woodlands School, Peel District School Board (Mississauga ON)

ABSTRACT

This project demonstrates how two programs are created in KNIME - an open source data analytic, reporting and integration platform, are used to support research scientists in medicinal chemistry. The first application flags pan-assay interference compounds such as “promiscuous” compounds present in chemical libraries that recurrently behaves as false positive hits in screening campaigns. The second application adapted a previously published workflow, where it automatically scans the recently published scientific literature on a weekly basis, and identifies articles considered relevant to medicinal chemists focused on epigenetic mechanisms, a novel and promising field in drug discovery. These workflows are very important because they allow a user with relatively little training to be able to extract important data that would typically need a trained chemist for. The PAINS workflow performed adequately but data was problematic. This workflow and an online tool, used to compare results, flagged different, but overlapping sets of compounds. The PubMed alert workflow performed very well, being able to consistently identify new papers. These workflows have been implemented at the Structural Genomics Consortium, in Toronto. Both Workflows are available at http://sgc.utoronto.ca/ditommaso.zip The implementation of these workflows demonstrate that the process is viable, and paves the way for the implementation of more complex workflows.

Ce projet montre comment deux logiciels qui ont été créés en utilisant KNIME - une plate-forme open-source d’intégration et de reportage de data analytique, sont utilisées comme soutient pour les chercheurs dans le domaine de chimie médicale. La première application signale les composés d’interférence pan-essai (PAINS), par exemples des composés ‘libérés’ présents dans les chimiothèques, qui s’agissent souvent comme des fausses réactions positives pendant les campagnes de dépistage. La deuxième application, le système de workflow PubMed alert, a adapté un système de workflow développé auparavant qui parcourt rapidement la littérature scientifique publiée récemment une fois par semaine et identifie des articles qui sont pertinents pour des chimistes médicales qui étudient des mécaniques épigénétiques, un domaine novateur et prometteur dans les découvertes des drogues. Ces systèmes de workflow sont très importants car ils permettent un utilisateur avec relativement peu d’entraînement à soutirer des données importantes qui ont typiquement besoin d’être trouvées par les chimistes entraînés. Le système de workflow de PAINS a fonctionné suffisamment mais les données trouvées étaient problématiques. Le système et un outil en ligne utilisé pour la comparaison des résultats ont signalés des résultats différents, mais les résultats se sont débordés sur les unes les autres. Nous avons trouvés que le système de workflow PubMed alert a très bien fonctionné, car le système pouvait constamment identifier des nouveaux papiers scientifiques. Ces systèmes de workflow sont maintenant implémentés au Consortium Génomique Structurel (SGC) à Toronto. Les deux systèmes de workflow sont disponibles à http://sgc.utoronto.ca/ditommaso.zip . L’implémentation de ces systèmes de workflow montre que le procès est viable et ouvre la voie pour l’implémention des systèmes de workflow plus complexes.

KEY WORDS

Cheminformatics; KNIME; PAINS; PubMedn

INTRODUCTION

LThe Konstanz Information Miner (KNIME, version 2.12.0) is an open source program that allows for the creation of applications, known as workflows, used to mine data. There are several nodes in KNIME which can be used for cheminformatics, including learning and predicting models. (KNIME, 2016) RDKit collection. This collection contains many nodes used for cheminformatic such as a substructure filters, molecule fragmenters and more (RDKit, 2016). The second project automatically scans the research literature for documents of relevance to the user and applies many of the built in text processing, and model creation and prediction nodes.

PAINS

Pan-assay interference compounds (PAINS) are chemical substructures that produce strong signals in tests to measure a compound’s ability to bind to

a protein and affect its activity. One such example is a set of PAINS, which produces hydrogen peroxide under certain test conditions. The hydrogen peroxide deactivates the protein, making the original compound look like a strong inhibitor, when it is not actually binding to the protein (Baell, J, 2016). PAINS typically show very promising signal in screening assays, and introduce noise that dilute the signals associated with bona fide compounds occupying protein binding sites (i.e real hits).

Document Classifier

Papadatos et al. developed a workflow that can be used to create models which identify specific types of articles (Papadatos, G, 2014). The model is trained by providing the workflow with articles considered relevant and irrelevant to pre-defined fields of research. Once the model is generated, it can be used to determine whether other articles are relevant or not. For this project, ChEMBL, a public repository that compiles the biological activity of small molecules published in the scientific literature, was used as the training set to identify published articles of relevance to medicinal chemists. MEDLINE (a universal and exhaustive repository of abstracts for the medical research literature) was used as the background. The model generated was then used to predict the "ChEMBL-likeness" of new articles in order to determine if they were relevant. The workflows, models and data created by Papadatos et al. are freely available.

METHODS

PAINS identification workflow

To identify if a compound is a PAINS it must contain a specific substructure. This set of substructures was acquired in Simplified molecular-input line-entry system (SMILES) format from the blog of Rajarshi Guha, at http://blog.rguha.net/?p=850, which were converted into the correct format from the patterns from the paper written by Baell et. Al.(Baell, J, 2010). The three sets of data were merged to form a single input file. As well, the following PAIN substructure was missing and was added to the filter. This substructure was added because upon testing, the filter did not flag certain compounds that were deemed unwanted.

[#7,#6]1[cX3]2[cX3][cX3][cX3][cX3][cX3]2[SX2][cX3]2[cX3][cX3][cX3][cX3][cX3]12

<regId=phenothiazine>

A database to be tested in Structural Data File (SDF) format is the other input of the workflow. SDF is used to store a table containing the structural data for a series of compounds. The SDF reader node inputs the file into the program. The workflow was also developed to be used on individual molecules drawn by the user. In this case the marvin sketch node, from the marvin set of nodes, was used in place of the SDF reader (KNIME, 2016). The marvin sketch node allows the user to draw one or more molecules which are used as input.

Once the chemical database was submitted to the workflow, it was converted into the RDKit molecule type. This is an internal data type used by the RDKit nodes. The converted chemical database and the PAINS substructures were used as input for the molecule substructure filter node implemented in RDKit. This node separated the database into two tables, one which contained the flagged compounds and one which did not. A column was then added to each table, indicating whether or not it was flagged. These tables were combined, converted back into SDF format and written to an output file.

Pub Med Alert Workflow Model Generation

To create the model, the previously mentioned workflow created by Papadatos et al. was used. In this case a set of 47939 papers reporting chemical inhibition of protein activity were selected as the group of relevant papers. Another set of 500001 randomly selected papers were used as the background to create a set of irrelevant papers for the model. The size of the sample sets were selected to be roughly equal, and of arbitrary size

File Retrieval and Analysis

To retrieve the new papers to be analyzed, a script was written in python (version 3.5) that utilizes the Entrez Programming Utilities to access PubMed articles (Python, 2016), (NCBI, 2016). The script accesses the server and requests the IDs of research articles indexed in MEDLINE in the past seven days that fits the following query, focused on terms related to the field of epigenetics. The query is used to reduce the number of new papers from PubMed down to a reasonable number. The entirety of new articles in PubMed would be far too many to analyze, and most of the data would be irrelevant anyway.

epigenetics OR chromatin OR methyltransferase OR demethylase OR bromodomain OR PWWP OR histone OR acetyltransferase AND inhibitor NOT COMT

The majority of the resulting articles are unrelated to medicinal chemistry, and no simple keyword can be used to find such articles, which is why a more sophisticated machine learning approach, including model generation, is necessary. The script then requests the records and parses it to retrieve the articles. The articles are then written in .csv format (comma separated values format), so they can be used by the workflow.

The workflow created by Papadatos et al. was then modified so that it could be used on the new records acquired by the script. It was modified so it would complete the pre-processing and predict the relevancy of articles, using the previously created model. The table produced by the workflow contained the records which were flagged as being considered relevant or not.

RESULTS

PAINS Workflow

The set of approved drugs from drugbank.ca in sdf format was used(n=1582) to test the efficacy of the PAINS Flag workflow (Wishart DS, 2006). This file was then run through the workflow and an online smarts filter set to flag PAINS (Yang, J, 2016). The workflow flagged 94 compounds, 20 of which were not flagged by the website. The website flagged 88 compounds, 14 of which were missed by the workflow. (Table 1).

Table 1: Summary of number of PAINs flagged by workflow and website Pub Med Workflow.

	Workflow	Both	Website
Number Flagged	20	74	14

To determine the ability of the Pubmed alert workflow to classify new papers a control set was created and used as input in the work flow. This consisted of 20 PubMed records selected on the basis that their abstract contained information relevant to medicinal chemistry in epigenetics research. These articles are considered the relevant set. Another set of papers, 201 in total, were randomly selected in order to create a set of data known to be mostly irrelevant. This set was considered the irrelevant set. These two sets were merged to create the test data set(n=221). The size of each set was arbitrary, but kept small enough that it would be reasonable for one person to manually classify all articles.

This data set was submitted as input on the PubMed alert workflow. Any relevant articles that were flagged as relevant or irrelevant by the work flow were considered true positive or false negative respectively. Any irrelevant papers that were flagged as relevant, known as false positives, were examined to ensure they were actually irrelevant. The rest of the output papers from the workflow are all considered true negatives. Seventeen of the relevant papers and three of the irrelevant papers were flagged as relevant while the rest, four relevant and 177 irrelevant articles, were flagged as irrelevant (Table 2).

Table 2: Summary of Analysis of PubMed Alert Workflow When Tested with Control Set (n=221).

	Relevant	Irrelevant
Predicted Relevant	17 (True Positive)	4 (False Positive)
Predicted Irrelevant	3 (False Negative)	177 (True Negative)

DISCUSSION

In this work, the scientific workflow system implemented in KNIME was used to address two unrelated challenges commonly faced by biomedical research scientists. In the first project, a workflow was developed to rapidly and efficiently flag molecules with liable chemical features in chemical libraries. In the second project, a workflow was developed to automatically scan the research literature and alert medicinal chemists on a daily or weekly basis of recently published discoveries relevant to their work. As discussed below, the workflows can significantly increase the productivity of research scientists, and were adopted at the Structural Genomics Consortium, in Toronto. Our analysis also shows that each workflow comes with some limitations.

PAINS Work Flow Analysis

The benefit of implementing an automated system such as this workflow is that any researcher can easily check if a compound contains a PAINS substructure or not. This requires very minimal training to be used by any researcher while still yielding reliable results. This workflow would be used as a first step to identifying a PAINS. If a compound is flagged by the workflow it should not be considered for follow-up studies.

Both our workflow and the online tool flagged compounds that the other did not. The twenty compounds that were missed by the website could all by flagged with the same SMILES, as follows. At some point, these compounds were flagged by the website, suggesting an error or instability in the online program.

c : 2 (: c : 1 - [# 1 6] - c : 3 : c (- [# 7] (- c : 1 : c (: c (:c:2-[#1])-[#1])-[#1])-[$([#1]),$([#6](-[#1])(-[#1])-[#1]),$([#6](-[#1])(-[#1])-[#6]-[#1])]):c(:c(~[$([#1]),$([#6]:[#6])]):c(:c:3-[#1])-[$([#1]),$([#7](-[#1])-[#1]),$([#8]-[#6;X4])])~[$([#1]),$([#7](-[#1])-[#6;X4]),$([#6]:[#6])])-[#1]

<regId="het_thio_666_A(13)">

In total, fourteen compounds were missed by the workflow. Six of them could not be converted by RDKit, and eight failed for an unknown reason. This could have been caused by RDKit not recognizing the SMILES or some other unknown error.

PubMed Alert Workflow

The PubMed alert workflow performed reasonably well. It is important for medicinal chemists to keep up-to-date on novel chemical inhibitors of relevance to their own research. For example, a novel chemical inhibitor of protein A1 (encoded by gene A1), a distant homologue of the protein of interest (protein A2), could inform on the chemotypes that could be tested to inhibit protein A2. Another example is when the crystal structure of a chemical inhibitor bound to a protein is found. This information could be used to optimize the binding affinity of a different inhibitor to the same protein. Being able to have this information automatically retrieved means new papers or data are less likely to be missed in the constant stream of published data.

CONCLUSION

This work has shown that is viable to implement automated programs to complete tasks such as filtering certain compound, or collating relevant articles. The workflows created demonstrate the viability of replacing tasks with automated programs. There is potential for even more complex tasks to be replaced by workflows and programs.

Future Directions

The two applications developed in this work illustrate the power and limitation of informatics workflow technology applied to medicinal chemistry. Knime, the software used here, is freely available and rapidly improving. Novel virtual nodes contributed by the medicinal chemistry community both in academia and the pharmaceutical industry are constantly increasing the power, robustness and scope of this technology, and we expect that novel applications, as the ones presented here, will help increase the efficiency of early-stage drug discovery.

ABBREVIATIONS

Abbreviation	Full Form
KNIME	Knostanz Information Miner
PAINS	Pan-Assay Interference Compounds
SMILES	Simplified Molecular Input Line Entry System
SDF	Structure-Data File
CSV	Comma Separated Values
SGC	Structural Genomics Consortium

ACKNOWLEDGEMENTS

I would like to thank Dr. Matthieu Schapira and Dr. Renato Freitas. They have provided great assistance in editing the paper, as well as giving me the opportunity to work with them on several projects related to their field of work. They have given me the opportunity to work with more challenging material and discover what its like to work in this field. As well, I would like to thank Ms. McBryan for being there to support and facilitate our work, as well as being integral in making this whole course happen.

REFERENCES

1. KNIME | Open for Innovation. KNIME | Open for Innovation, https://www.knime.org/ (accessed Jan 2, 2016).
2. RDKit: Open-Source Cheminformatics Software. RDKit, http://www.rdkit.org/ (accessed Jan 2, 2016).
3. Baell, J.; Walters, M. A.; Introducing the PAINS. *Nature*. 2014, 513, 481-483.
4. Papadatos, G.; Westen, G. J. V.; Croset, S.; Santos, R.; Trubian, S.; Overington, J. P. A Document Classifier for Medicinal Chemistry Publications Trained on the ChEMBL Corpus. Journal of Cheminformatics. 2014, 6, 40.
5. Baell, J. B.; Holloway, G. A. New Substructure Filters For Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. J. Med. Chem. Journal of Medicinal Chemistry. 2010, 53, 2719–2740.
6. Free Marvin Chemistry Extensions. KNIME, https://www.knime.org/free-marvin-chemistry-extensions (accessed Jan 2, 2016).
7. Welcome to Python.org. Python.org, https://www.python.org/ (accessed Jan 2, 2016).
8. Entrez Programming Utilities Help. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/books/nbk25501/ (accessed Jan 2, 2016)
9. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D668-72. 16381955.
10. Yang, J. SmartsFilter. SmartsFilter, http://pasilla.health.unm.edu/tomcat/biocomp/smartsfilter (accessed Jan 2, 2016).