11-2015

A novel approach to identify shared fragments in drugs and natural products

Ashkay Balasubramanya
University of Nebraska at Omaha, abalasubramanya@gmav.unomaha.edu

Ishwor Thapa
University of Nebraska at Omaha, ithapa@unomaha.edu

Dhundy Raj Bastola
University of Nebraska at Omaha, dkbastola@unomaha.edu

Dario Ghersi
University of Nebraska at Omaha, dghersi@unomaha.edu

A novel approach to identify shared fragments in drugs and natural products

Akshay Balasubramanya
Information Systems and
Quantitative Analysis Department
University of Nebraska at Omaha
Omaha, NE 68182

Ishwor Thapa
School of Interdisciplinary Informatics
University of Nebraska at Omaha
Omaha, NE 68182

Dhundy Bastola
School of Interdisciplinary Informatics
University of Nebraska at Omaha
Omaha, NE 68182

Dario Ghersi
School of Interdisciplinary Informatics
University of Nebraska at Omaha
Omaha, NE 68182
Email: dghersi@unomaha.edu

***Abstract*—**

Fragment-based approaches have now become an important component of the drug discovery process. At the same time, pharmaceutical chemists are more often turning to the natural world and its extremely large and diverse collection of natural compounds to discover new leads that can potentially be turned into drugs. In this study we introduce and discuss a computational pipeline to automatically extract statistically overrepresented chemical fragments in therapeutic classes, and search for similar fragments in a large database of natural products. By systematically identifying enriched fragments in therapeutic groups, we are able to extract and focus on few fragments that are likely to be active or structurally important. We show that several therapeutic classes (including antibacterial, antineoplastic, and drugs active on the cardiovascular system, among others) have enriched fragments that are also found in many natural compounds. Further, our method is able to detect fragments shared by a drug and a natural product even when the global similarity between the two molecules is generally low. A further development of this computational pipeline is to help predict putative therapeutic activities of natural compounds, and to help identify novel leads for drug discovery.

I. Introduction

A crucial factor for realizing the promises of personalized medicine is the availability of novel and safe drugs to modulate the increasing number of targets that are being identified. Of all the medical branches, oncology is posed to be among those that could benefit the most from a new array of therapeutics [1].

Despite substantial progress in understanding the molecular basis of human cancers, there is still a pressing need for more effective, rational and personalized treatments. A few drugs for specific cancer types have achieved a good degree of selectivity with relatively low toxicity, but for the vast majority of human cancers, standard chemotherapy regimens (with their related toxicity) remain the only viable option. However, the situation is rapidly changing.

Breakthroughs in cancer genomics are now leading to the identification of new actionable targets [2], opening up unprecedented opportunities for personalized treatment. As a result of our improved understanding of cancer biology, with some notable exceptions the search for "silver bullet" therapies has now largely been replaced by a quest for novel targets that can be simultaneously modulated by combinatorial therapies [1], [3], akin to what has been accomplished for the treatment of HIV infections [4]. As a result, a vast number of suitable new drugs will soon be required to modulate a large array of cancer targets.

Another area where the availability of new effective drugs is becoming a pressing need is the treatment of infectious diseases, as antibiotic-resistant bacteria are becoming more common and widespread and are a cause for serious concern [5].

The natural product derived structure plays a significant role in the discovery of novel pharmaceutical agents and/or bioactive molecules. The anti-diabetic activity in lupins has been attributed to quinozolidine alkaloids [6] and a review of the literature shows many such examples of natural products as sources of new drugs [7] including Paclitaxel, which is one of the most widely prescribed anticancer drugs on the market. Most of the natural products are biologically active and have favorable absorption, distribution, metabolism, excretion and toxicology properties. Plants are often the predominant source for the discovery of natural products due to the relative ease of access. However, more recently microbial as well as marine sources have been identified as alternative resources, particularly for antibiotics [8]. Several databases of natural products have been published and reviewed [9], [10], [11], [12].

Although many pharmaceutical companies emphasize high throughput (HTP) screening of combinatorial libraries, natural products continue to provide enormous structural and chemical diversity to guide the careful design of drug-like leads. More importantly, the products of HTP screening often do not interact well with biomolecules and induce unexpected and possibly severe side effects. Therefore, over the years (since 1980)

only 2 drugs obtained through the HTP screening have been approved by the FDA, while over 85 drugs are either natural products-based or compounds derived from them [13], [14]. In the past decade, several databases focusing on the collection of medically important natural products and medicinal herbs have been established [10], [11] and the use of computer aided drug design including virtual screening of large databases has become an important part of the drug discovery process [15].

Pharmaceutically relevant natural products are of low molecular weight and often restricted to special plant families. While these compounds are not important for the primary metabolism of the plants, they are of great importance for their survival in a given environment. Therefore, medicinally important plants are often collected from the wild or their natural habitat and are more likely to be endangered due to severe over collection [16]. Unfortunately, we still have limited knowledge about plant secondary metabolism, its regulation, molecular mechanisms concerning gene expression and rate-limiting enzymes found within a diverse network of biosynthetic pathways in living organisms.

While molecular biology and biotechnology are being used to produce phytopharmaceuticals and natural pesticides, integration of different disciplines in plant sciences including computational strategies are necessary to unravel metabolic networks and to elucidate the biosynthesis of pharmaceutically relevant secondary metabolite. With the development of the fragment-based method described in this study, it is now possible to determine potentially important structures in natural products *in silico*, which may be investigated further to determine their pharmaceutical value as lead or intermediate compound, and potentially produced by cells cultivated *in vitro* utilizing plant biotechnology methods. To the best of our knowledge, this is the first time that a fragment-based approach using enrichment analysis is applied to identify potentially important chemical fragments in natural products.

II. Materials and Methods

A. Obtaining and representing drugs and natural products

The DrugBank database [17] (version 4.1) was used to obtain information on drugs that were approved for therapeutic use in at least one country. The initial set of drugs contained 1,554 molecules. Natural products were obtained from the SuperNatural II database [12], containing 325,508 molecules.

Drugs and natural products were represented using the SMILES system [18], a widely used notation that makes it possible to encode chemicals as ASCII strings. SMILES strings for drugs and natural products were directly obtained from the DrugBank and SuperNatural II databases, respectively.

B. Fragmenting the molecules

Both drugs and natural products were fragmented with the `fragment` program, part of the `molBLOCKS` suite [19], which breaks molecules along chemically important bonds and returns the corresponding fragments (or putative building blocks). The list of chemical bonds that were used by the program to fragment the molecules is shown in Figure 2, and is based on Lewell et al. [20]. The minimum size for a fragment was set to four atoms, and the fragmentation was carried out with the "extensive" flag turned on, which yields all possible fragments that can be generated given the list of chemical bonds of interest [19].

It is noteworthy to mention that the fragmentation rules are encoded as SMARTS (SMiles ARbitrary Target Specification), an extension to the SMILES notation created by `Daylight Chemical Information System, Inc.` and widely used in computational chemistry. Using SMARTS patterns the particular bonds that are to be cleaved are encoded as regular expressions, making it straightforward to add other cleavable bonds to the fragmentation rules.

C. Clustering fragments

Drug fragments obtained as described above were clustered with the `analyze` [19] program using standard parameters. In order to compute the fragment similarity for clustering, the program converts the fragment to a fingerprint representation, based on linear segments of up to 7 atoms in length (FP2 fingerprints [21]). The fingerprints are stored as bit arrays, where the presence or absence of a particular linear segment is represented by a 1 or 0, respectively. The FP2 fingerprint representation is obtained via the Open Babel library[1]. Then, the Tanimoto coefficient T_s between two fragments x and y is computed as:

$$T_s = \frac{\sum_i X_i \wedge Y_i}{\sum_i X_i \vee Y_i} \quad (1)$$

where X and Y are the bit array representations of the linear segments found in fragment x and y, respectively, and $\wedge$ and $\vee$ are the bitwise *and* and *or* operators.

The `analyze` program computes pairwise similarities between fragments and converts them to a graph representation, where an edge between fragments indicates a pairwise Tanimoto greater than the chosen threshold, which was set to 0.7 in this study. Subsequently, the program extracts the connected components of the graph, and selects the representative element for each cluster as the fragment with the highest average similarity against all the other fragments in the cluster.

D. Extracting enriched fragments for each ATC code

In order to assign functional categories to drugs, we used the Anatomical Therapeutic Chemical (ATC) classification system[2], a widely used nomenclature that organizes drugs according to the organ or system which they modulate and their therapeutic properties. The ATC code system is hierarchically organized into five levels of increasing specificity. We considered the second level, which describes the therapeutic main groups. We note that a single drug can be annotated with multiple ATC codes, if it has multiple therapeutic indications. For this study, to get meaningful statistics we selected all the ATC codes that annotated at least 10 distinct drugs.

Enrichment analysis was carried out in order to identify the specific fragments (or clusters of fragments) that appear in a set of molecules more frequently than expected by chance,

[1] http://openbabel.org/wiki/Tutorial:Fingerprints

[2] http://www.whocc.no/atc/structure_and_principles/

given a background distribution. In this study the background was represented by the union of all approved drugs.

The `analyze` program uses the hypergeometric distribution to model the probability of obtaining a number of fragments (or clusters of fragments) equal to or greater than the observed by chance alone:

$$P(X \geq k) = \sum_{x=k}^{K} \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \quad (2)$$

where N is the total number of fragments; K is number of fragments of the given type; n is the total number of fragments in the main set; and x is the total number of fragments of the given type in the main set.

The program returns both uncorrected $p-$values and False Discovery Rate (FDR) corrected $p-$values, obtained with the procedure of Benjamini-Hochberg [22]. In this study we selected fragments that were enriched with an FDR < 0.05.

E. Comparing enriched fragments in the drug dataset against fragments from natural compounds.

The final step of the pipeline involves the comparison between enriched fragments from the drug dataset against fragments obtained from the natural compounds set. In order to calculate the pairwise similarity between each of the enriched drug fragments and each of the fragments from natural compounds we used the Tanimoto coefficient (see equation 1). To carry out the calculations we wrote an in-house program that uses the Python API [23] of the OpenBabel library [21], and retained the drug fragment–natural product fragment pairs that had a Tanimoto similarity > 0.9.

F. Computational requirements.

The most time-consuming step of the pipeline is represented by the pairwise fragment comparison, which took approximately 12 hours on a 24-core machine. Fragmentation of the 325,509 molecules found in the SuperNatural II database took approximately eight hours on a 24-core machine, bringing the entire analysis to roughly 20 hours.

III. Results

A. A computational pipeline to systematically compare functionally relevant drug fragments and natural products

We set out to systematically compare approved drugs obtained from the DrugBank database [17] against a large collection of natural products, assembled in the SuperNatural II database [12]. The novelty of our approach consists first in extracting the fragments that are statistically overrepresented in each pharmacological category, and then in comparing those fragments against the ones derived from the natural compounds.

The rationale behind this approach is twofold. On the one hand, chemical fragments capture important properties of the full molecules, and on the other hand they may be shared by otherwise globally dissimilar molecules, which might go undetected when using a global similarity measure. The pipeline

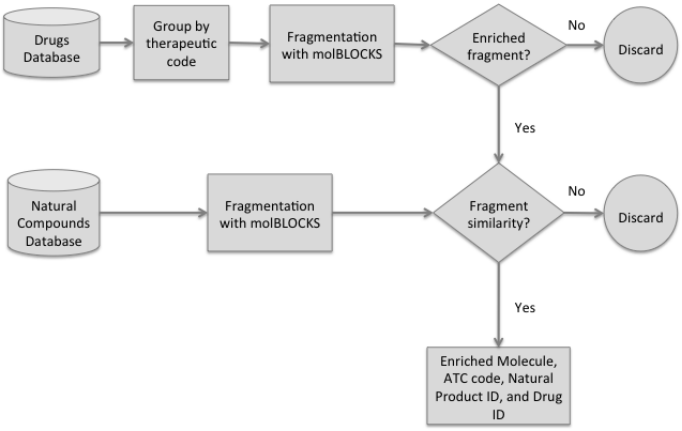


Fig. 1. **Simplified overview of the pipeline.** Each approved drug (obtained from Drugbank [17]) is assigned a therapeutic class using the ATC nomenclature. The drugs are then broken down into fragments using the `molBLOCKS` software [19], and enrichment analysis is performed on each therapeutic class to identify statistically overrepresented fragments (FDR < 0.05). Each overrepresented fragment is then compared against similarly obtained fragments from a database of natural compounds (SuperNatural II [12]), and (see Materials and Methods for further details).

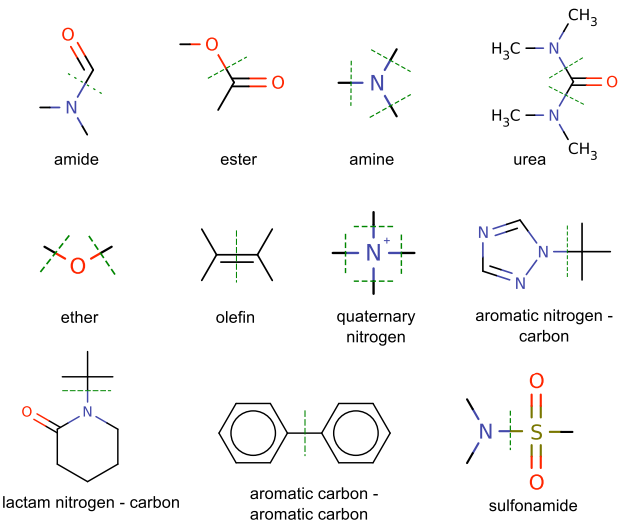
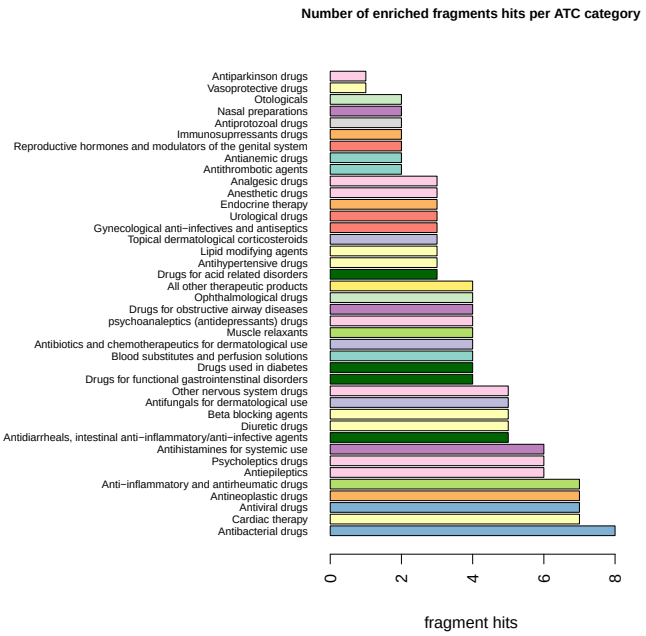


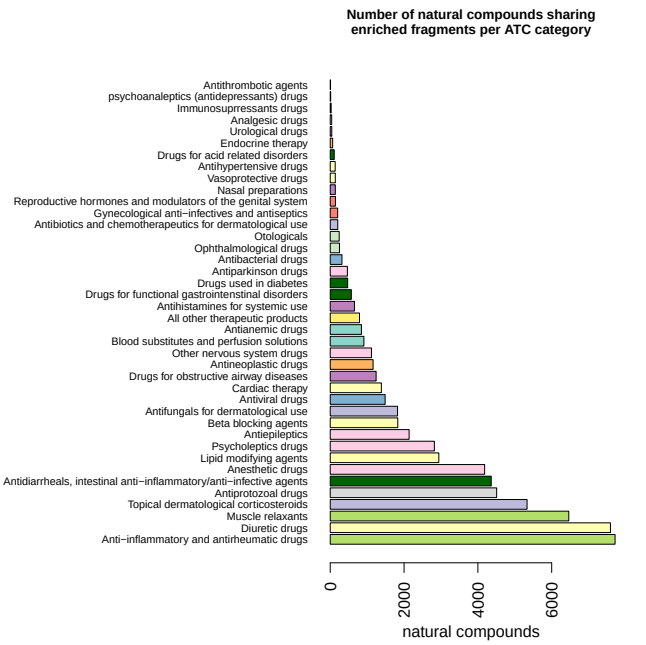
Fig. 2. **RECAP rules used to fragment drugs and natural products.** The 11 eleven types of chemical bonds depicted above (green dashed lines) indicate the potential sites that can be broken in the small molecules, resulting in smaller fragments. These 11 fragmentation rules were derived from Lewell et al. [20], and are believed to capture chemically relevant synthetic reactions that combine building blocks into more complex molecules.

is briefly outlined in Figure 1, which shows the main steps of the procedure. More details are found in the Materials and Methods section of the paper.

In order to fragment the molecules we used the `molBLOCKS` suite [19] with the RECAP rules [20] (Figure 2), which allow us to break small molecules apart along chemically important bonds. It is noteworthy to mention that in several cases no fragmentation rule applies to a small molecule, which is then left as it is and treated as a whole fragment. In our initial dataset of 1,543 approved drugs we were able to fragment 949 (62%) of the drugs. The remaining ones, for which no fragmentation RECAP rule applies, were treated as



(a)

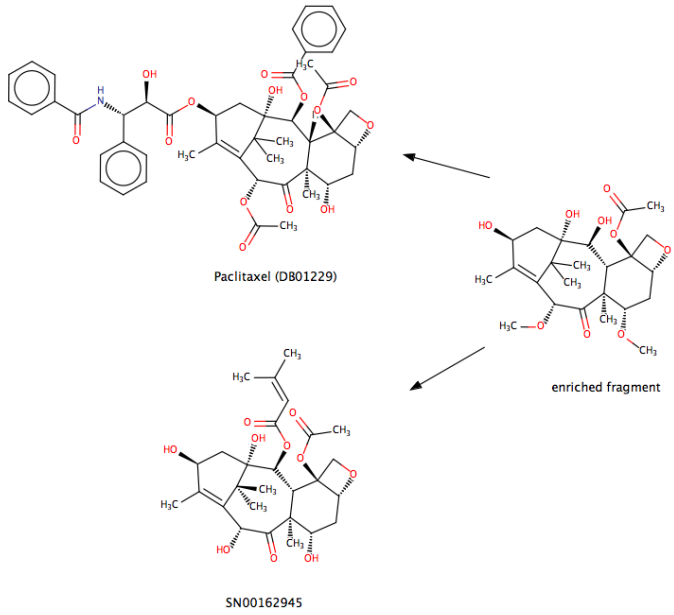


(b)

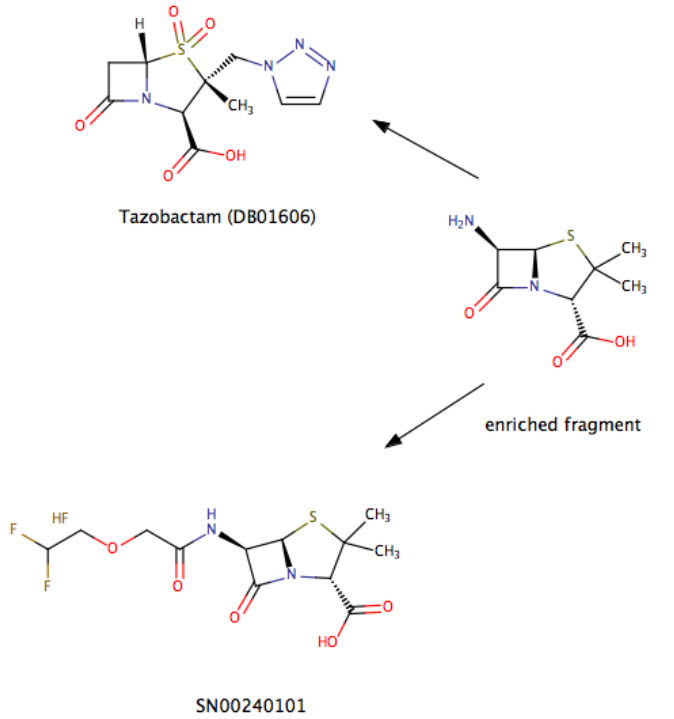
Fig. 3. **Distribution of enriched fragments and matching natural compounds per ATC code.** Panel (a) shows the number of drug fragments that are enriched in given therapeutic categories (ATC codes) that have at least one matching fragment in the set of natural compounds. Panel (b) shows the total number of natural compounds whose fragments match one or more of the enriched drug fragments in each therapeutic category.

one fragment. In the case of natural products, the fragmentation rules applied to 174,156 (54%), and the remaining molecules were treated as one fragment.

Subsequently, we grouped drugs by Anatomic Therapeutic



(a)



(b)

Fig. 4. **Examples of fragments shared by natural compounds and drugs in the absence of high global similarity.** The two examples shown here illustrate how a fragment-based approach can automatically detect commonalities between molecules that are globally different. Panel (a) shows a tetracyclic fragment present both in a natural compound and in an anti-cancer agent (Paclitaxel). In spite of the common core shared by the two molecules, the Tanimoto similarity between the drug and the natural compound is relatively low (0.56).
In panel (b), the beta-lactam ring is detected (which a small variation) in both an approved antibiotic (tazobactam) and a natural compound (SN0240101). However, the Tanimoto similarity between the natural compound and tazobactam is low (0.49).

Code, which gives the therapeutic group of a drug (e.g., "L01" stands for "Antineoplastic Agents", "C03" for "Diuretics", etc.). Multiple membership of a drug in several ATC groups was allowed if the drug was annotated in DrugBank with multiple ATC codes. We ended up with 40 ATC groups, each containing at least 10 distinct drugs. For each ATC group, we performed clustering of the fragment followed by enrichment analysis with the `molBLOCKS` suite, extracting statistically overrepresented fragments for each group, with an FDR < 0.05. The total number of enriched fragments across all therapeutic groups was 141.

In the last step of the pipeline, we systematically compared the enriched fragments from each ATC group against the fragments obtained from the natural compounds, and retained for further analysis all the pairs that had a Tanimoto similarity > 0.9.

B. Drugs and natural compounds are related at the fragment level in specific therapeutic groups

We considered the number of fragments for each therapeutic group with at least one matching fragment in the natural products dataset, obtaining the distribution shown in Figure 3(a). The top-ranking group was represented by the antibacterial drugs, followed by drugs active on the cardiovascular system, antiviral drugs, antineoplastic drugs, and anti-inflammatory drugs. The prominence of antibacterial drugs in this list is consistent with the importance that natural products have had in the development of antibiotics [24].

An alternative way of analyzing these data is to consider the number of natural products whose fragments match at least one of the fragments in each therapeutic group. The results are shown in Figure 3(b). The therapeutic group with the largest number of natural products is now the anti-inflammatory class, closely followed by diuretic drugs, muscle relaxants, and corticosteroids.

C. Case studies

In Figure 4(a) and 4(b) we show two examples of fragments shared by a drug and a natural product in the context of low global similarity. One of the advantages of our fragment-based approach is the automatic identification of common and chemically important building blocks among molecules that may be globally dissimilar.

A proof of concept is given by the anticancer drug Paclitaxel and the natural product SN00162945 (Figure 4(a)), which share a tetracyclic core but have different substituents. In fact, Paclitaxel itself was first isolated from the bark of a yew, and belongs to the taxane family, whose members all share the core fragment shown in the figure (or a closely related variation). However, because of the different substituents in the two molecules, the Tanimoto coefficient between Paclitaxel and SN00162945 turns out to be only 0.56.

Another example that showcases the power of using fragments is shown in Figure 4(b). The antimicrobial Tazobactam contains a β-lactam ring, which is the building block of a highly important group of widely prescribed antibiotics, including penicillin, cephalosporins and carbapenems, and it occurs in several natural compounds. As in the example of Figure 4(a), Tazobactam has a low Tanimoto similarity (0.49) for the natural product SN00240101, in spite of the fact that they both share the β-lactam ring.

IV. Discussion and Conclusions

The natural world as a source of highly diverse and complex chemicals has always been of value to synthetic chemists, and is becoming even more relevant today, given the output slump of the pharmaceutical industry. The pipeline introduced here allows to automatically detect relationships between small molecules using a fragment-based approach. Using a fragment-based approach is motivated by the fact that natural products are often assembled from independent building blocks via a chain of enzymatic reactions. These processes are somewhat similar to what is common practice in synthetic chemistry.

By first extracting statistically overly represented fragments for each therapeutic class we reduced the complexity of the approved drugs to a handful of chemical fragments that are likely to be responsible (at least in part) for the pharmaceutical activity of the given drugs, or are important as chemical scaffolds. Comparing these fragments against the fragments obtained from a large library of natural products allowed us to establish potential relationships between drugs and natural products even in the absence of high global similarity between the molecules. As an analogy, we could compare this fragment-based approach to a local sequence alignment procedure, which can identify highly similar protein domains among globally different protein sequences.

As a note of caution, we should mention that the choice of the Tanimoto similarity thresholds or the stringency of the fragment clustering step would affect the final results, in that more or less matching fragments would be found depending on how stringent the parameters that control the similarity are set to be. Unfortunately, there are no hard and fast rules to guide the user in the choice of parameters. However, as it is often the case in bioinformatics applications, the results should be interpreted as a guide to help design further experiments or perform more thorough literature searches. In this context, our pipeline could be used to ask the question of whether a natural product that happens to share a fragment with an antihypertensive drug does in fact have pressure lowering activity. Alternatively, the pipeline could be used to investigate whether a natural product shows potential as a lead compound for a given therapeutic indication.

In the future we plan to further test the pipeline and extend it by including the sources of natural products. Although this may not be possible for all compounds, databases like the "Universal Natural Product Database" [11] (contained in SuperNatural II) do include source information for several compounds. Combined with metabolic information on plant and microbial pathways, this will yield a better understanding of natural product synthesis. As shown by a pioneering study by Runguphan et al. [25], this could eventually lead to co-opting natural systems for engineering better drugs.

Acknowledgments

The authors would like to thank the Bioinformatics group at UNO for useful discussions.

REFERENCES

[1] D. Hanahan, "Rethinking the war on cancer," *The Lancet*, vol. 383, pp. 558–563, 2014.

[2] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, J. Diaz L. A., and K. W. Kinzler, "Cancer genome landscapes," *Science*, vol. 339, no. 6127, pp. 1546–1558, 2013.

[3] B. Al-Lazikani, U. Banerji, and P. Workman, "Combinatorial drug therapy for cancer in the post-genomic era," *Nature Biotechnology*, vol. 30, pp. 679–692, 2012.

[4] E. J. Arts and D. J. Hazuda, "HIV-1 antiretroviral drug therapy," *Cold Spring Harbor Perspectives in Medicine*, vol. 2, 2012.

[5] J. M. A. Blair, M. A. Webber, A. J. Baylay, D. O. Ogbolu, and L. J. V. Piddock, "Molecular mechanisms of antibiotic resistance." *Nature reviews. Microbiology*, vol. 13, no. 1, pp. 42–51, Jan. 2015.

[6] B. Brunmair, Z. Lehner, K. Stadlbauer, I. Adorjan, K. Frobel, T. Scherer, A. Luger, L. Bauer, and C. Fürnsinn, "55p0110, a novel synthetic compound developed from a plant derived backbone structure, shows promising anti-hyperglycaemic activity in mice," *PLoS ONE*, 2015.

[7] D. J. Newman and G. M. Cragg, "Natural products as sources of new drugs over the 30 years from 1981 to 2010," *Journal of Natural Products*, vol. 75, no. 3, pp. 311–335, 2012.

[8] G. M. Cragg and D. J. Newman, "Natural products: A continuing source of novel drug leads," *Biochimica et Biophysica Acta*, vol. 1830, no. 6, pp. 3670–3695, 2013.

[9] C. Y. C. Chen, "TCM Database@Taiwan: The world's largest traditional Chinese medicine database for drug screening In Silico," *PLoS ONE*, vol. 6, no. 1, 2011.

[10] A. B. Yongye, J. Waddell, and J. L. Medina-Franco, "Molecular scaffold analysis of natural products databases in the public domain," *Chemical biology & drug design*, vol. 80, no. 5, pp. 717–724, 2012.

[11] J. Gu, Y. Gui, L. Chen, G. Yuan, H. Z. Lu, and X. Xu, "Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology," *PLoS ONE*, vol. 8, no. 4, 2013.

[12] P. Banerjee, J. Erehman, B.-O. Gohlke, T. Wilhelm, R. Preissner, and M. Dunkel, "Super Natural II–a database of natural products." *Nucleic acids research*, vol. 43, no. Database issue, pp. D935–9, Jan. 2015.

[13] Y.-W. Chin, M. J. Balunas, H. B. Chai, and A. D. Kinghorn, "Drug discovery from natural sources," *The AAPS journal*, vol. 8, no. 2, pp. E239–E253, 2006.

[14] F. Ntie-Kang, J. A. Mbah, L. M. Mbaze, L. L. Lifongo, M. Scharfe, J. N. Hanna, F. Cho-Ngwa, P. A. Onguéné, L. C. O. Owono, E. Megnassan *et al.*, "Cammednp: Building the cameroonian 3d structural natural products database for virtual screening," *BMC complementary and alternative medicine*, vol. 13, no. 1, p. 88, 2013.

[15] F. E. Koehn and G. T. Carter, "The evolving role of natural products in drug discovery," *Nature reviews Drug discovery*, vol. 4, no. 3, pp. 206–220, 2005.

[16] V. Sarasan, G. C. Kite, G. W. Sileshi, and P. C. Stevenson, "Applications of phytochemical and in vitro techniques for reducing over-harvesting of medicinal and pesticidal plants and generating income for the rural poor," *Plant cell reports*, vol. 30, no. 7, pp. 1163–1172, 2011.

[17] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, "Drugbank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic Acids Res*, vol. 34, no. Database issue, pp. D668–72, 2006.

[18] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, 1988.

[19] D. Ghersi and M. Singh, "MolBLOCKS: Decomposing small molecule sets and uncovering enriched fragments," *Bioinformatics*, vol. 30, no. 14, pp. 2081–2083, 2014.

[20] X. Q. Lewell, D. B. Judd, S. P. Watson, and M. M. Hann, "Recap–retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry," *J Chem Inf Comput Sci*, vol. 38, no. 3, 1998.

[21] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open babel: An open chemical toolbox," *J Cheminform*, vol. 3, p. 33, 2011.

[22] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, pp. 289–300, 1995.

[23] N. M. O'Boyle, C. Morley, and G. R. Hutchison, "Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit." *Chemistry Central journal*, vol. 2, p. 5, 2008.

[24] D. G. Brown, T. Lister, and T. L. May-Dracka, "New natural products as new leads for antibacterial drug discovery." *Bioorganic & medicinal chemistry letters*, vol. 24, no. 2, pp. 413–8, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24388805

[25] W. Runguphan, X. Qu, and S. E. O'Connor, "Integrating carbon-halogen bond formation into medicinal plant metabolism." *Nature*, vol. 468, no. 7322, pp. 461–464, 2010.