Bioinformatics, 33(16), 2017, 2601–2603
doi: 10.1093/bioinformatics/btx200
Advance Access Publication Date: 7 April 2017
Applications Note

Data and text mining

TIN-X: target importance and novelty explorer

Daniel C. Cannon[1], Jeremy J. Yang[1], Stephen L. Mathias[1], Oleg Ursu[1], Subramani Mani[1], Anna Waller[2], Stephan C. Schürer[3], Lars Juhl Jensen[4], Larry A. Sklar[2,5], Cristian G. Bologa[1,]* and Tudor I. Oprea[1]

[1]Translational Informatics Division, Department of Internal Medicine, University of New Mexico School of Medicine, Albuquerque, NM 87131, USA, [2]UNM Center for Molecular Discovery, University of New Mexico Comprehensive Cancer Center, University of New Mexico, Albuquerque, NM 87131, USA, [3]Department of Molecular and Cellular Pharmacology, Miller School of Medicine, University of Miami, Miami, FL 33136, USA, [4]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen N 2200, Denmark and [5]Department of Pathology, University of New Mexico, NM 87131, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on January 19, 2017; revised on March 16, 2017; editorial decision on March 31, 2017; accepted on April 6, 2017

Abstract

Motivation: The increasing amount of peer-reviewed manuscripts requires the development of specific mining tools to facilitate the visual exploration of evidence linking diseases and proteins.
Results: We developed TIN-X, the Target Importance and Novelty eXplorer, to visualize the association between proteins and diseases, based on text mining data processed from scientific literature. In the current implementation, TIN-X supports exploration of data for G-protein coupled receptors, kinases, ion channels, and nuclear receptors. TIN-X supports browsing and navigating across proteins and diseases based on ontology classes, and displays a scatter plot with two proposed new bibliometric statistics: Importance and Novelty.
Availability and Implementation: http://www.newdrugtargets.org
Contact: cbologa@salud.unm.edu

1 Introduction

Science builds upon past discoveries, traditionally communicated through scientific literature. However, scale-out of traditional and alternate publication modes has exceeded the limits of human processing (Hunter and Cohen, 2006), creating the need for computer-assisted methods. Text mining, ontologies, and interactive visualization are some of the emerging technologies that can alleviate information overload. We present a new method and software tool utilizing these technologies, to provide biomedical scientists with rankings and visualizations linking proteins and diseases. The information is derived from PubMed abstracts, text mined using previously published tools for named entity recognition (NER) of gene/protein and disease names (Pletscher-Frankild *et al.*, 2015). We propose two new bibliometric indices based on NER mentions, specifically devised to address research planning use cases. Our goal is to enable scientists to identify research subjects with sufficient evidence of importance, but not already well-studied, with focus on target prioritization for drug discovery. *Novelty* estimates the scarcity of publications about a protein target (see also Equation 1). *Importance* estimates the strength of the association between that protein target and a specific disease (see also Equation 2). By visualizing Novelty and Importance on a 2D plot, users can easily examine the strength of the evidence linking protein targets and diseases of interest.

TIN-X is a web application which allows users to navigate targets via the drug target ontology (DTO, http://drugtargetontology.org) and diseases via the disease ontology (DO) (Kibbe *et al.*, 2015). In its current implementation, TIN-X allows users to query, browse and display disease-target associations for the following protein families: ion channels, G-protein coupled receptors (GPCRs), nuclear receptors (NRs) and kinases. These plots are interactive, with built-in drill-down and link-out functionality for in-depth examination of selected targets. TIN-X was inspired by and developed for the Illuminating the Druggable Genome (IDG) Knowledge Management Center (KMC, http://targetcentral.ws), which aggregates and integrates protein-centric information, and seeks to identify and prioritize understudied genes and proteins for further investigation and validation as potentially novel drug targets.

2 Methods and implementation

TIN-X is implemented in Scala using the Lift web framework (http://liftweb.net). The back-end utilizes a PostgreSQL database derived from the IDG-KMC knowledge base known as Target Central Research Database (TCRD) (Nguyen *et al.*, 2017). TCRD compiles data from over 50 genomic, proteomic and drug-centric sources, and relies on the Disease Ontology, as well as in-network (IDG-KMC) developed tools such as DTO and disease—genes/protein mapping using a highly efficient dictionary-based NER application (Pafilis *et al.*, 2013) with dictionaries from the DISEASES (Pletscher-Frankild *et al.*, 2015) database, respectively. Data are extracted, transformed, aggregated and loaded into TCRD, with routine updates quarterly. All PubMed abstracts are downloaded from NCBI FTP site. Bibliometric statistics and the derived Novelty (N_i) and Importance (I_{ij}) scores are pre-computed for runtime performance, using the following formulae:

$$N_i = 1/\sum_k \frac{1}{T_k} \qquad (1)$$

$$I_{ij} = \sum_k \frac{1}{T_k \cdot D_k} \qquad (2)$$

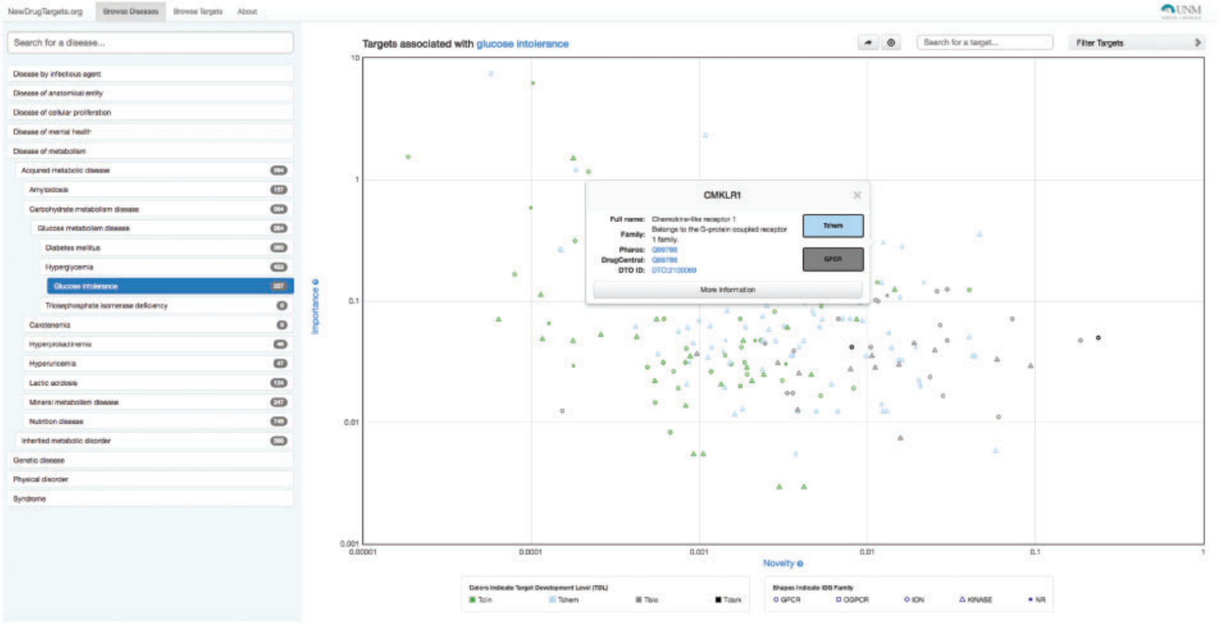
where T_k and D_k are the numbers of targets and diseases in abstract (k), respectively, and summation over all publications including target (i), and for importance, also including disease (j). Fractional counts are employed to reflect strength of association. For example, if an abstracts mentions three targets once, each receives a fractional count $1/T = {}^1/_3$.

3 Application example

In Figure 1, TIN-X displays targets for a selected disease, ‘glucose intolerance’. The DO hierarchy can be navigated in the left panel, and the targets are plotted with log–log Importance–Novelty axes. Searching by disease name and by target name (with auto-suggest) is supported. In the current layout, targets with stronger associations are in the upper part of the plot, while targets with a higher number of publications are on the left side of the plot. In general, more interesting associations will be on the upper right boundary of the plot, where data points represent non-dominated solutions to the multi-objective optimization maximizing both Importance and Novelty. Targets can be filtered based on Target Development Level, with Tclin representing mode-of-action drug targets (Santos *et al.*, 2017), and Tdark representing the understudied proteins (Nguyen *et al.*, 2017). Targets can be filtered by protein superfamily (e.g. GPCRs). Link outs to Pharos (targets) and DO (diseases) are provided for each association. Figure 2 illustrates the result of mouse-over actions for the target from Figure 1, ‘Chemokine-like receptor 1’ (CMKLR1), which links to 6 publications associating CMKLR1 and glucose intolerance. Titles and citations are displayed. Clicking on each article reveals full abstracts and links out to complete PubMed records, where users can examine the available evidence.

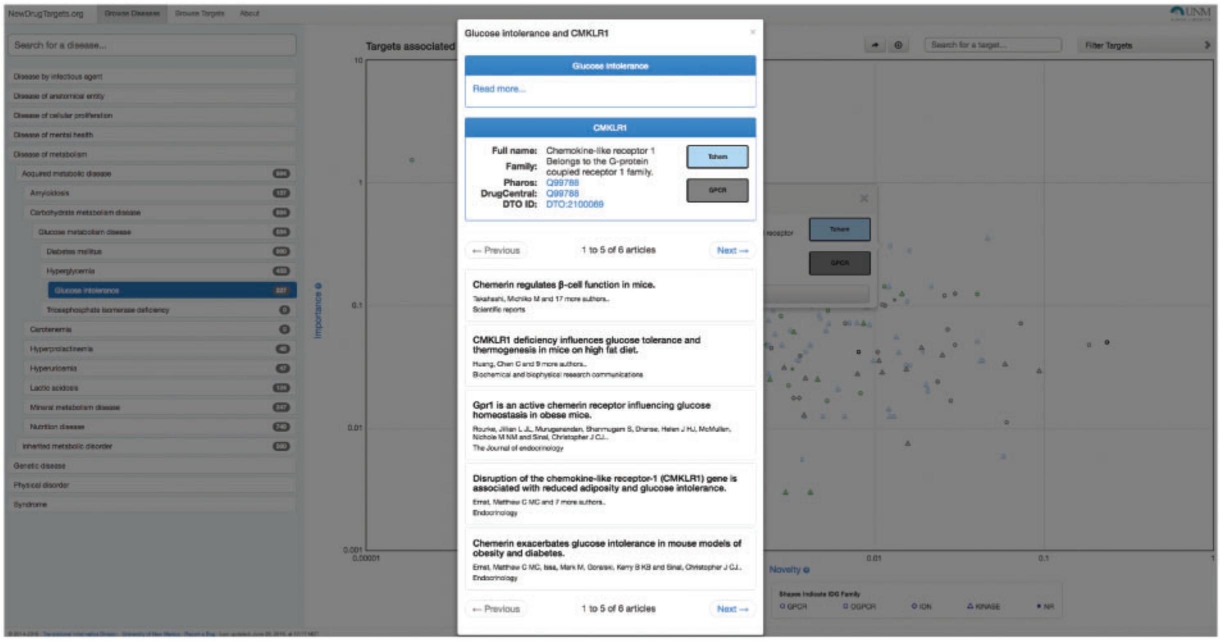
4 Conclusions and future directions

TIN-X provides an interactive visualization, ranking, and prioritization platform for scientists interested in exploring potentially novel drug targets, and examining the relationship between diseases, disease categories, proteins and protein classes, using automated text mining of biomedical literature. As with all current text mining,



Full name: Chemokine-like receptor 1
Family: Belongs to the G-protein coupled receptor 1 family.
Pharos: Q99788
DrugCentral: Q99788
DTO ID: DTO:2100069

Fig. 1. Screenshot of TIN-X. The disease, glucose intolerance, was queried. Mouse-over displays information about a specific data point, in this case, CMKLR1 (Color version of this figure is available at Bioinformatics online.)



Glucose intolerance and CMKLR1

Glucose intolerance

CMKLR1

Full name: Chemokine-like receptor 1
Family: Belongs to the G-protein coupled receptor 1 family.
Pharos: Q99788
DrugCentral: Q99788
DTO ID: DTO:2100069

1 to 5 of 6 articles

Chemerin regulates β-cell function in mice.
Takahashi, Michiko M and 17 more authors..
Scientific reports

CMKLR1 deficiency influences glucose tolerance and thermogenesis in mice on high fat diet.
Huang, Chen C and 9 more authors..
Biochemical and biophysical research communications

Gpr1 is an active chemerin receptor influencing glucose homeostasis in obese mice.
Rourke, Jillian L JL, Muruganandan, Shanmugam S, Dranse, Helen J HJ, McMullen, Nichole M NM and Sinal, Christopher J CJ..
The Journal of endocrinology

Disruption of the chemokine-like receptor-1 (CMKLR1) gene is associated with reduced adiposity and glucose intolerance.
Ernst, Matthew C MC and 7 more authors..
Endocrinology

Chemerin exacerbates glucose intolerance in mouse models of obesity and diabetes.
Ernst, Matthew C MC, Issa, Mark M, Goralski, Kerry B KB and Sinal, Christopher J CJ..
Endocrinology

1 to 5 of 6 articles

Fig. 2. Screenshot of TIN-X. Click on selected point shows the papers where the target protein (CMKLR1) might be relevant to the disease (glucose intolerance)

TIN-X cannot replace expert human readers and curators, yet, it is increasingly clear that automated bibliometry is essential given the accelerated pace and volume of publications. TIN-X is an important interactive tool in the IDG project, and future plans include adding functionality relating to known drugs and drug classes, and inclusion of additional protein classes. More information about Tclin, Tdark and other target development levels can be found at http://www.nature.com/nrd/posters/druggablegenome/

Funding

This work was supported by the National Institutes of Health [U54 CA189205-01], and the Novo Nordisk Foundation [NNF14CC0001].

Conflict of Interest: none declared.

References

Hunter,L. and Cohen,K.B. (2006) Biomedical language processing: what's beyond PubMed?. *Mol. Cell*, **21**, 589–594.

Kibbe,W.A. *et al.* (2015) Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.*, **43**, D1071–D1078.

Nguyen,D.-T. *et al.* (2017) Pharos: collating protein information to shed light on the druggable genome. *Nucleic Acids Res.*, **45**, D995–D1002.

Pafilis,E. *et al.* (2013) The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. *PLOS One*, **8**, e65390.

Pletscher-Frankild,S. *et al.* (2015) DISEASES: Text mining and data integration of disease–gene associations. *Methods*, **74**, 83–89.

Santos,R. *et al.* (2017) A Comprehensive Map of Molecular Drug Targets. *Nat. Rev. Drug Discov.*, **16**, 19–34.