Reflections on: Finding Melanoma Drugs Through a Probabilistic Knowledge Graph

James P. McCusker[1], Michel Dumontier[4], Rui Yan[1], Sylvia He[1], Jonathan S. Dordick[2,3], and Deborah L. McGuinness[1,3]

[1] Department of Computer Science,
[2] Department of Chemical & Biological Engineering,
[3] Center for Biotechnology & Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY, US
[4] Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA, US

Abstract. We build a nanopublication-based knowledge graph of protein/protein, drug/protein, and protein/disease interactions to create a resource for exploring potential therapies for diseases. This is accomplished using Semantic Web standard tools like Blazegraph, SADI web services, and JSON-LD for integration with a Javascript-based web client. Metastatic cutaneous melanoma is an aggressive skin cancer with some progression-slowing treatments but no known cure. The omics data explosion has created many possible drug candidates, however filtering criteria remain challenging, and systems biology approaches have become fragmented with many disconnected databases. Using drug, protein, and disease interactions, we built an evidence-weighted knowledge graph of integrated interactions. Our knowledge graph-based system, ReDrugS, can be used via an API or web interface, and has generated 25 high quality melanoma drug candidates. We show that probabilistic analysis of systems biology graphs increases drug candidate quality compared to non-probabilistic methods. Four of the 25 candidates are novel therapies, three of which have been tested with other cancers. All other candidates have current or completed clinical trials, or have been studied in *in vivo* or *in vitro*. This approach can be used to identify candidate therapies for use in research or personalized medicine.

Keywords: melanoma, drug repositioning, knowledge graphs, uncertainty reasoning

1 Introduction

Metastatic cutaneous melanoma is an aggressive cancer of the skin with low prevalence but very high mortality rate, with an estimated 5 year survival rate of 6 percent [1] There are currently no known therapies that can consistently cure metastatic melanoma. Vemurafenib is effective against BRAF mutant melanomas [2] but resistant cells often result in recurrence of metastases

[8] Melanoma itself may be best approached based on the individual genetics of the tumor, as it has been shown to involve mutations in many different genes to produce the same disease [7]. Because of this, an individualized approach may be necessary to find effective treatments.

A knowledge graph is a compilation of facts and figures that can be used to provide contextual meaning to searches. Google is using knowledge graphs to improve its search and to analyze the information graph of the web; Facebook is using them to analyze the social graph. We built our knowledge graph with the goal of unifying large parts of biomedical domain knowledge for both mining and interactive exploration related to drugs, diseases, and proteins. Our knowledge graph is enhanced by the provenance of each fragment of knowledge captured, which is used to compute the confidence probabilities for each of those fragments. Further, we use open standards from the World Wide Web Consortium (W3C), including the Resource Description Framework (RDF) [6], Web Ontology Language (OWL) [12], and SPARQL [4]. The representation of the knowledge in our knowledge graph is aligned with best practice vocabularies and ontologies from the W3C and the biomedical community, including the PROV Ontology [9], the HUPO Proteomics Standards Initiative Molecular Interactions (PSI-MI) Ontology [5], and the Semanticscience Integrated Ontology (SIO) [3]. Use of these standards, vocabularies, and ontologies make it simple for ReDrugS to integrate with other similar efforts in the future with minimal effort.

We built a novel computational drug repositioning platform, that we refer to as ReDrugS, that applies probabilistic filtering over individually-supported assertions drawn from multiple databases pertaining to systems biology, pharmacology, disease association, and gene expression data. We use our platform to identify novel and known drugs for melanoma.

2 Results

We used ReDrugS to examine the drug-target-disease network and identify known, novel, and well supported melanoma drugs. The ReDrugS knowledge base contained 6,180 drugs, 3,820 diseases, 69,279 proteins, and 899,198 interactions.

We examined drug and gene connections that were 3 or less interaction steps from melanoma, and additionally filtered interactions with a joint probability greater or equal to 0.93. We identified 25 drugs in the resulting drug-gene-disease network surrounding melanoma as illustrated in Figure 1 .

We then validated the set of 25 drugs by determining their position in the drug discovery pipeline for melanoma. Nearly all drugs uncovered by ReDrugS were previously been identified as potential melanoma therapies either in clinical trials or *in vivo* or *in vitro*. Of the 25 drugs, 12 have been in Phase I, II, or III clinical trials, 5 have been studied *in vitro*, 4 *in vivo*, 1 was investigated as a case study, and 3 are novel.

To further evaluate our system, we examined the impact of decreasing the joint probability or increasing the number of interaction steps. Figures 2 A and

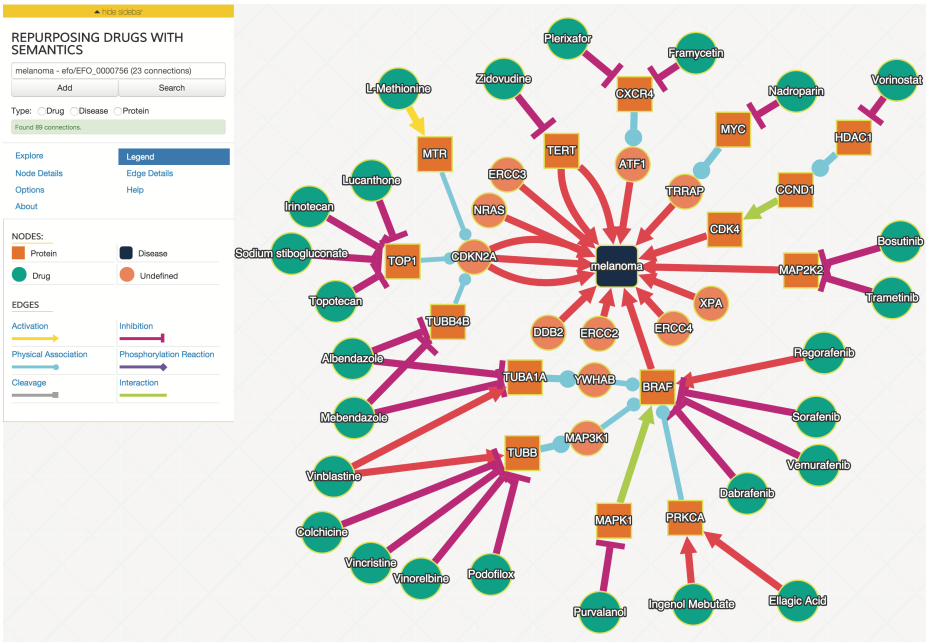

Fig. 1. The interaction graph of predicted melanoma drugs with a probability of 0.93 or higher and have three or fewer intervening interactions between drug and disease. The "Explore" tab contains the controls to expand the network in various ways, including the filtering parameters. Node and edge detail tabs provide additional information about the selected node or edge, including the probabilities of the edges selected. Users can control the layout algorithm and related options using the "Options" tab.

B show precision, recall, and f-measure curves while varying each parameter. Using these information retrieval performance curves we found that using a joint probability of 0.93 or greater with 3 or less interaction steps maximizes the precision and recall as shown in Figure 2.

By performing a literature search on hypothesis candidates with a joint probability of 0.5 or higher and 6 or fewer interaction steps, we were able to generate precision, recall, and f-measure curves for both cutoffs to find our cutoff of 0.93 with 3 or fewer interaction steps. The precision, recall, and f-measure curves are shown for varying joint probability thresholds in Figure 2 A and for varying interaction step counts in Figure 2 B.

3 Discussion

We designed ReDrugS to quickly and automatically integrate and filter a heterogeneous biomedical knowledge graph to generate high-confidence drug repositioning candidates. Our results indicate that ReDrugs generates clinically plau-

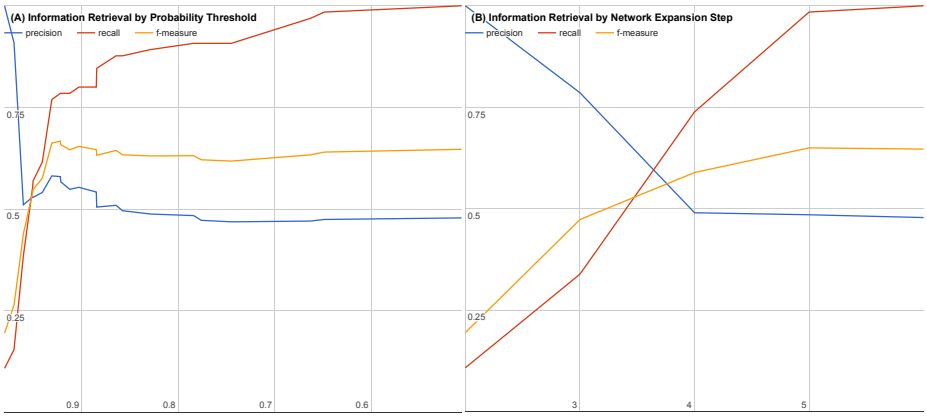

Fig. 2. Precision, recall, and f-measure by (A) varying thresholds for joint probability and (B) varying number of interaction steps. Precision is the percentage of returned candidates that have been validated experimentally or have been in a clinical trial (a "hit") versus all candidates returned. Recall is the percentage of all known validated "hits". F-measure is the geometric mean of precision and recall that provides a balanced evaluation of the quality and completeness of the results.

sible drug candidates, in which half are in various stages of clinical trials, while others are novel or are being investigated in pre-clinical studies. By helping to consolidate the three main datatypes - drug targets, protein interactions, and disease genes, ReDrugs can amplify the ability of researchers to filter the vast amount of information into those that are relevant for drug discovery.

3.1 Architecture

ReDrugS uses a fairly straightforward web architecture, as shown in Figure 3. It uses the Blazegraph RDF database backend. The database layer is interchangeable except that the full text search service needs to use Blazegraph-only properties to perform text searches as text indexing is not yet standardized in the SPARQL query language. All other aspects are standardized and should work with other RDF databases without modification. ReDrugs currently uses the Python-based TurboGears web application framework hosted using the Web Services Gateway Interface (WSGI) standard via an Apache HTTP server. TurboGears in turn hosts the SADI web services that drive the application and access the database. It also serves up the static HTML and supporting files.

The user interface is implemented with AngularJS and Cytoscape.js, which submits queries to the SADI web services using JSON-LD and aggregates results into the networked view. The software relies exclusively on standardized protocols (HTTP, SADI, SPARQL, RDF, and others) to make it simple to replace technologies as needed. The data itself is processed using conversion scripts as shown in Figure 4.

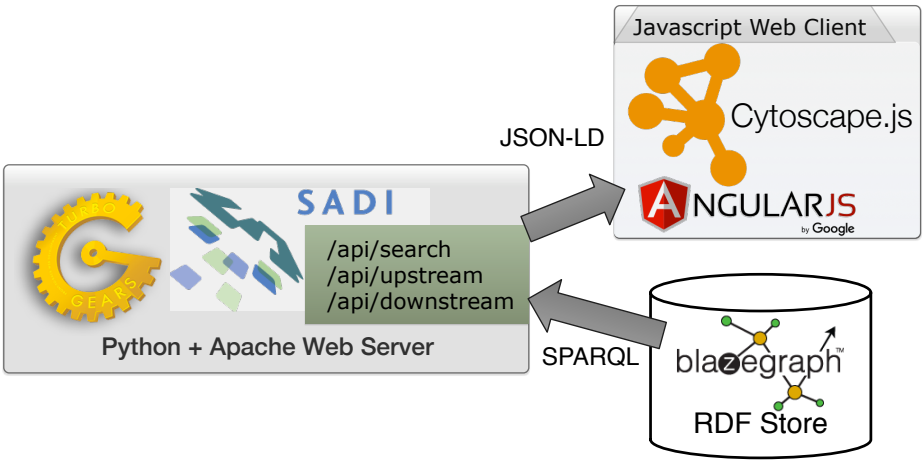

Fig. 3. The ReDrugS software architecture. Using web standards and a three layer architecture (RDF store, web server, and rich web client), we were able to build a complete knowledge graph analysis platform.

4 Materials and Methods

This research project did not involve human subjects. The ReDrugS platform consists of a graphical web application, an application programming interface (API), and a knowledge base. The graphical web application enables users to initiate a search using drug, gene, and disease names and synonyms. Users can then interact with the application to expand the network at an arbitrary number of interactions away from the entity of interest, and to filter the network based on a joint probability between the source and target entities. Drug-protein, protein-protein, and gene-disease interactions were obtained from several datasets and integrated into ontology-annotated and provenance and evidence bearing representations called nanopublications. The web application obtains information from the knowledge base using semantic web services. Finally, we evaluated our approach by examining the mechanistic plausibility of the drug in having melanoma-specific disease modifying ability. We evaluated a large number of possible drug/disease associations with varying joint probabilities and interaction steps to determine the thresholds with the highest F-Measure, resulting in our thresholds of three or less interactions and a joint probability of 0.93 or higher.

4.1 Semantic Web Services

We developed four Semantic Automated Discovery and Integration (SADI) web services [13] in Python to support easy access to the nanopubications (see Table 1) in ReDrugS. The four services are enumerated in Table 1.

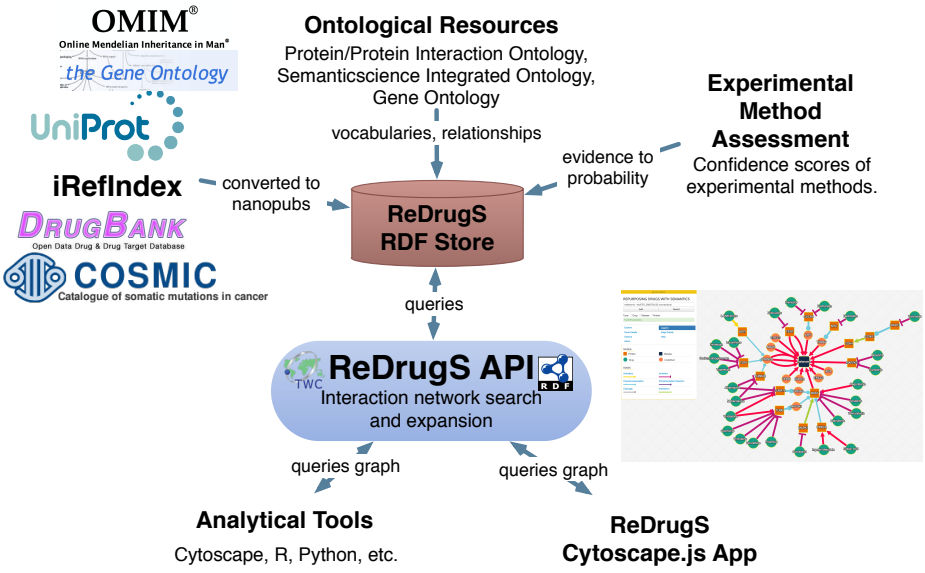

Fig. 4. The ReDrugS data flow. Data is selected from external databases and converted using scripts into nanopublication graphs, which are loaded into the ReDrugS data store. This is combined with experimental method assessments, expressed in OWL, and public ontologies into the RDF store. The web service layer queries the store and produces aggregate analyses of those nanopublications, which is consumed and displayed by the rich web client. The same APIs can be used by other tools for further analysis.

The first service is a simple free text lookup, that takes an *pml:Query*[5] [10] with a *prov:value* as a query and produces a set of entities whose labels contain the substring. This is used for interactive typeahead completion of search terms so users can look up URIs and entities without needing to know the details.

The other three SADI services look up interactions that contain a named entity. Two of them look at the entity to find upstream and downstream connections, and the third service assumes that the entity is a biological process and finds all interactions that related to that process. The services return only one interaction for each triple (source, interaction type, target). There are often multiple probabilities per interaction, and more than one interaction per interaction type. This is because the interaction may have been recorded in multiple databases, based on different experimental methods. To provide a single probability score for each interaction of a source and target, the interactions are combined. A single probability is generated per identified interaction by taking the geometric mean of the probabilities for that interaction. However, this

[5] PML 3, in development: https://github.com/timrdf/pml. This includes PML 2 constructs that are not covered in PROV-O.

Service Name	Description	URL	Input	Output
Resource text tearch	Look up resources using free text search against their RDFS labels. This service is optimized for typeahead user interfaces.	search	*pml:Query*	*pml:AnsweredQuery*
Find interactions in a biological process	Find interactions whose participants or targets also participate in the input process.	process	*sio:Process*	*sio:Process*
Find upstream participants	Find interactions that the input entity is a target of in and have explicit participants.	upstream	*sio:MaterialEntity*	*sio:Target*
Find downstream targets	Find interactions that the input entity participates in and have explicit targets.	downstream	*sio:MaterialEntity*	*sio:Agent*

Table 1. The API endpoint prefix is `http://redrugs.tw.rpi.edu/api/`.

method is undesirable when combining multiple interaction records of the same type. We instead combine the interaction records using a form of probabilistic voting using composite Z-Scores. This is done to model that multiple experiments that produce the same results reinforce each other, and should therefore give a higher overall probability than would be indicated by taking their mean or even by Bayes Theorem. We do this by converting each probability into a Z Score (aka Standard Score) using the Quantile Function ($Q()$), summing the values, and applying the Cumulative Distribution Function ($CDF()$) to compute the corresponding probability:

$$P\left(x_{1\ldots n}\right) = CDF\left(\sum_{i=1}^{n} Q\left(P\left(x_i\right)\right)\right)$$

These composite Z Scores, which we transform back into probabilities, are frequently used to combine multiple indicators of the same underlying phenomena, as in [11].

4.2 User Interface

The user interface was developed using the above SADI web services and uses Cytoscape.js,[6], angular.js,[7] and Bootstrap 3.[8] An example network is shown

[6] http://cytoscape.github.io/cytoscape.js
[7] https://angularjs.org
[8] http://getbootstrap.com

in Figure 1 Users can search for biological entities and processes, which can then be autocompleted to specific entities that are in the ReDrugS graph. Users can then add those entities and processes to the displayed graph and retrieve upstream and downstream connections and link out to more details for every entity. Cytoscape.js is used as the main rendering and network visualization tool, and provides node and edge rendering, layout, and network analysis capabilities, and has been integrated into a customized rich web client.

In order to evaluate this knowledge graph, we developed a demonstration web interface[9] based on the Cytoscape.js[10] JavaScript library. The interface lets users enter biological entity names. As the user types, the text is resolved to a list of entities. The user finishes by selecting from the list, and submitting the search. The search returns interactions and nodes associated with the entity selected, which are added to the Cytoscape.js graph. Users are also able to select nodes and populate upstream or downstream connections. Figure 1 is an example output of this process.

References

1. Barth, A., Wanek, L., Morton, D.: Prognostic factors in 1,521 melanoma patients with distant metastases. J Am Coll Surg **181**, 193–201 (Sep 1995)
2. Chapman, P.B., Hauschild, A., Robert, C., Haanen, J.B., Ascierto, P., Larkin, J., Dummer, R., Garbe, C., Testori, A., Maio, M., Hogg, D., Lorigan, P., Lebbe, C., Jouary, T., Schadendorf, D., Ribas, A., O'Day, S.J., Sosman, J.A., Kirkwood, J.M., Eggermont, A.M., Dreno, B., Nolop, K., Li, J., Nelson, B., Hou, J., Lee, R.J., Flaherty, K.T., McArthur, G.A.: Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. New England Journal of Medicine **364**(26), 2507–2516 (jun 2011). https://doi.org/10.1056/nejmoa1103782, http://dx.doi.org/10.1056/NEJMoa1103782
3. Dumontier, M., Baker, C.J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N.R., Duck, G., Furlong, L.I., Keath, N., Klassen, D., McCusker, J.P., Queralt-Rosinach, N., Samwald, M., Villanueva-Rosales, N., Wilkinson, M.D., Hoehndorf, R.: The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery. Journal of Biomedical Semantics **5**(1), 14 (2014). https://doi.org/10.1186/2041-1480-5-14, http://dx.doi.org/10.1186/2041-1480-5-14
4. Harris, S., Seaborne, A., Prudhommeaux, E.: SPARQL 1.1 query language. W3C Recommendation **21** (2013)
5. Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S.G.N., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C., Apweiler, R.: The hupo psi's molecular interaction formata community standard for the representation of protein interaction data. Nature biotechnology **22**(2), 177–183 (2004)

[9] http://redrugs.tw.rpi.edu
[10] http://cytoscape.github.io/cytoscape.js

6. Klyne, G., Carroll, J.J.: Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation (2005)
7. Krauthammer, M., Kong, Y., Bacchiocchi, A., Evans, P., Pornputtapong, N., Wu, C., McCusker, J., Ma, S., Cheng, E., Straub, R., Serin, M., Bosenberg, M., Ariyan, S., Narayan, D., Sznol, M., Kluger, H., Mane, S., Schlessinger, J., Lifton, R., Halaban, R.: Exome sequencing identifies recurrent mutations in NF1 and RASopathy genes in sun-exposed melanomas. Nat Genet **47**, 996–1002 (Sep 2015)
8. Le, K., Blomain, E.S., Rodeck, U., Aplin, A.E.: Selective RAF inhibitor impairs ERK1/2 phosphorylation and growth in mutant NRAS vemurafenib-resistant melanoma cells. Pigment Cell Melanoma Res **26**(4), 509–517 (apr 2013). https://doi.org/10.1111/pcmr.12092, http://dx.doi.org/10.1111/pcmr.12092
9. Lebo, T., Sahoo, S., McGuinness, D.: PROV-O: The PROV Ontology. http://www.w3.org/TR/prov-o/ (2013)
10. McGuinness, D.L., Ding, L., Silva, P.P.D., Chang, C.: PML 2: A Modular Explanation Interlingua. In: Proceedings of the AAAI 2007 Workshop on Explanation-Aware Computing. pp. 22 – 23 (2007), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.186.8633
11. Moller, J., Cluitmans, P., Rasmussen, L., Houx, P., Rasmussen, H., Canet, J., Rabbitt, P., Jolles, J., Larsen, K., Hanning, C., Langeron, O., Johnson, T., Lauven, P., Kristensen, P., Biedler, A., van Beem, H., Fraidakis, O., Silverstein, J., Beneken, J., JS, G.: Long-term postoperative cognitive dysfunction in the elderly: ISPOCD1 study. The Lancet **351**(9106), 857–861 (Mar 1998). https://doi.org/10.1016/s0140-6736(97)07382-0, http://dx.doi.org/10.1016/S0140-6736(97)07382-0
12. Motik, B., Patel-Schneider, P.F., Cuenca Grau, B.: OWL 2 Web Ontology Language: Direct Semantics (2009), http://www.w3.org/TR/2009/REC-owl2-direct-semantics
13. Wilkinson, M., Vandervalk, B., McCarthy, L.: SADI Semantic Web Services - 'cause you can't always GET what you want! In: 2009 IEEE Asia-Pacific Services Computing Conference (APSCC). Institute of Electrical & Electronics Engineers (IEEE) (dec 2009). https://doi.org/10.1109/apscc.2009.5394148, http://dx.doi.org/10.1109/APSCC.2009.5394148