

RDF SKETCH MAPS - KNOWLEDGE COMPLEXITY REDUCTION FOR PRECISION MEDICINE ANALYTICS

NATTAPON THANINTORN¹, JUEXIN WANG², ILKER ERSOY¹, ZAINAB AL-TAIE²,
YUEXU JIANG², DUOLIN WANG², MEGHA VERMA¹, TRUPTI JOSHI^{2,3,4},
RICHARD HAMMER¹, DONG XU^{2,3}, DMITRIY SHIN^{1,3,2*}

¹*Department of Pathology and Anatomical Sciences*, ²*Department of Computer Science and Christopher S. Bond Life Sciences Center*, ³*MU Informatics Institute*, ⁴*Department of Molecular Microbiology and Immunology and School of Medicine – Office of Research, University of Missouri, Columbia, MO 65203, USA*

*Email: shindm@health.missouri.edu

Realization of precision medicine ideas requires significant research effort to be able to spot subtle differences in complex diseases at the molecular level to develop personalized therapies. It is especially important in many cases of highly heterogeneous cancers. Precision diagnostics and therapeutics of such diseases demands interrogation of vast amounts of biological knowledge coupled with novel analytic methodologies. For instance, pathway-based approaches can shed light on the way tumorigenesis takes place in individual patient cases and pinpoint to novel drug targets. However, comprehensive analysis of hundreds of pathways and thousands of genes creates a combinatorial explosion, that is challenging for medical practitioners to handle at the point of care. Here we extend our previous work on mapping clinical omics data to curated Resource Description Framework (RDF) knowledge bases to derive influence diagrams of interrelationships of biomarker proteins, diseases and signal transduction pathways for personalized theranostics. We present RDF Sketch Maps – a computational method to reduce knowledge complexity for precision medicine analytics. The method of RDF Sketch Maps is inspired by the way a sketch artist conveys only important visual information and discards other unnecessary details. In our case, we compute and retain only so-called RDF Edges – places with highly important diagnostic and therapeutic information. To do this we utilize 35 maps of human signal transduction pathways by transforming 300 KEGG maps into highly processable RDF knowledge base. We have demonstrated potential clinical utility of RDF Sketch Maps in hematopoietic cancers, including analysis of pathways associated with Hairy Cell Leukemia (HCL) and Chronic Myeloid Leukemia (CML) where we achieved up to 20-fold reduction in the number of biological entities to be analyzed, while retaining most likely important entities. In experiments with pathways associated with HCL a generated RDF Sketch Map of the top 30% paths retained important information about signaling cascades leading to activation of proto-oncogene BRAF, which is usually associated with a different cancer, melanoma. Recent reports of successful treatments of HCL patients by the BRAF-targeted drug vemurafenib support the validity of the RDF Sketch Maps findings. We therefore believe that RDF Sketch Maps will be invaluable for hypothesis generation for precision diagnostics and therapeutics as well as drug repurposing studies.

1. Introduction

Basic science discoveries coupled with tremendous advances in “omics” technologies have triggered a paradigm shift in today’s biomedicine. The idea of precision and personalized medicine is viewed by many as a solution to improve patient care by addressing disease complexity and heterogeneity [1]. It is especially evident in the direction that modern medical diagnostics and therapeutics, jointly coined as *theranostics*, is progressing. Pathway-based diagnostics is promising

to open up a view at internal biological mechanisms of complex interplay of clinical biomarkers, diseases, signal transduction and other processes to be able to more precisely describe differences in individual patient cases [2]-[10]. Generation of a mechanistic picture of such processes can help develop combinatorial therapies utilizing novel drugs, small molecules inhibitors, cytotoxic and differentiating agents and other interventional techniques. And, even though, precision theranostic approaches have not yielded significant advances yet due to limited drug options, the number of successful clinical cases using targeted therapies being reported is increasing [11]-[15]. For instance, in some cases deeper analysis of signal transduction pathways revealed an alternative activation of carcinogenic mechanisms, which mandated a use of novel combinatorial therapies. In other cases, unconventional drugs have been used to treat patient exhibiting no response to conventional regimens. For example, a successful unconventional therapy of Hairy Cell Leukemia (HCL) with vemuratenib, a drug usually associated with melanoma cancers and targeting BRAF proto-oncogene, has been reported in several clinical cases [16]-[19].

The great challenge here in our view is the difficulty of conducting a comprehensive precision theranostic study due to limitations of individual practitioners' knowledge of biological processes. An inter-expert collaboration, while being able to expand the knowledge space to a certain extent, is still not an effective solution. For instance, a number of reported cases indicate that current attempts to practice precision and personalized medicine reflect more *descriptive* rather than *predictive* approaches. Pathologists and oncologists are trying more to *describe* the successful application of unconventional drugs by analyzing biopsies and linking proteomic expression to signal transduction and known mutations rather to *predict* patient-specific disease mechanisms based on clinical omics data. We strongly believe that new methods for clinical hypothesis generation for precision theranostics are needed to increase the chance of having more successes similar to the use of vemuratenib in HCL.

To this end, we have been investigating advanced inference methods to map clinical biomarkers data to biological pathways to recreate interplay of signaling proteomic networks for individual patient cases [20]. Our new computational formalism called *Resource Description Framework (RDF)-induced Influgrams (RIIG)* has been shown in a recent proof-of-concept study to exhibit qualities sufficient to provide case-specific reasoning for theranostics [10]. RIIG takes advantage of vast amounts of publicly available *curated biological knowledge* represented as the RDF format. The importance and utility of use of RDF knowledge bases (KBs) in biomedicine have been demonstrated in a number of publications [21]-[24].

The application of RIIG on the set of all pathways involved can dramatically reduce RIIG performance and result in reduction of its practical utility in a medical setting. A number of studies related to biological pathway data processing have been focused on pathway curation [25], visualization [26] and analysis [3], [27], [28]. There have been also some studies to construct a skeleton from complex networks by pruning edges [29]. The general idea of maintaining most informative nodes by finding shortest path in a directed network has been explored in metabolic

engineering [30]-[32]. By searching all possible reactions between compounds, these methods output several minimum cost paths by defining different penalties of reaction type, compound type and atom mapping. However, simplifying networks in the context of precision medicine has not yet been investigated.

Here, we present the RDF Sketch Maps – a new computational method to reduce complexity of RDF-formatted knowledge networks to improve theranostic analyses in precision medicine settings. The method of RDF Sketch Maps is inspired by the way a sketch artist conveys only important visual information, while leaving out other unnecessary details. In our case, we compute and retain only so-called RDF Edges – places with highly important diagnostic and therapeutic information. To do this the method traverses knowledge networks and scores paths according to an objective function that incorporates information about a set of known diagnostic and therapeutic biomarkers (e.g. disease-associated genes and drug targets). The paths are then ranked by decreasing values of the scores. A set of *exploratory* genes that could possibly be useful in explaining patient-specific disease heterogeneity is used to compute the enrichment score for ranked paths for each version of the objective function. The top paths with high enrichment score are selected to form an RDF Sketch Map. The resulting maps are used for further analysis by computational methods or visualized for human analysis.

2. Methods

2.1. Construction of RDF knowledge base

For preliminary experiments we have constructed a knowledge base (KB) consisting of 35 signaling pathway maps from Kyoto Encyclopedia Genes and Genomes (KEGG), including pathways associated with molecular interactions, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases, and drug development [33]. We preferred KEGG maps over other pathway databases such as Biocarta [34] and Reactome [35] because of KEGG's inclusion of a variety of different types of signal transduction interactions (e.g. phosphorylation, methylation, ubiquitination, and glycosylation) that are relevant to cancer theranostics (Table 1).

KEGG however was initially designed as a set of manually-drawn pathway maps for human consumption. The electronic version of KEGG maps introduced later in the form of XML-like KGML files merely represents serialization of graphical artifacts. The high rate of inaccuracies and omissions (up to 30% comparing to graphical maps) in some KGML files makes them unacceptable for use in precision medicine applications, which requires high levels of accuracy of underlying facts and reliable knowledge provenance. We, therefore, set a goal to transform KEGG KGML files into highly accurate machine processable KB with inference capabilities. To do that we (i) designed a KEGG RDF ontology that models the relationships among biological entities and allows description logic inference, (ii) converted KGML files into RDF data set using in-house developed

graphical curation tool and a set of scripts, and (iii) added information about biological processes from Gene Ontology (GO)[36] and proteomic information from UniProt database [37].

Table 1. Modeled KEGG interactions.

KEGG RDF ontology specifies the type and constraints of interactions among biological entities as well as their class/sub-class hierarchical relationships. For instance, we decided to preserve specific paths of propagation of signal transduction in individual maps through the notion of a “gene instance”. The underlying reason behind it is that certain reactions (e.g. phosphorylation) occur under specific circumstances (e.g. presence of specific enzymes or involving specific protein domains). However, since gene instances coming from different maps are modeled as sub-classes of an “abstract gene”, we can combine individual maps into an integrated semantic “mash-up” KB. This allows one to potentially recreate a systems view of signal transduction in individual patient cases. A similar approach is utilized while modeling gene groups, which represents protein complexes at the proteomic level. The constructed KEGG RDF KB was loaded into AllegroGraph RDF store [38] for querying and processing. To optimize performance of running RDF Sketch Maps algorithm we use AllegroGraph’s internal SNA RDF graph processing only to resolve aliases and run description logic inference. The RDF Sketch algorithm is run on “static” graph serialization derived from the AllegroGraph KEGG RDF KB.

RDF predicate name	Modeling purpose
activates	Molecular interaction
binds_associates	Molecular interaction
changes_state	Molecular interaction
dephosphorylates	Molecular interaction
dissociates	Molecular interaction
expresses	Molecular interaction
glycosylates	Molecular interaction
indirectly_affects	Molecular interaction
inhibits	Molecular interaction
methylates	Molecular interaction
misses_interaction	Molecular interaction
phosphorylates	Molecular interaction
represses	Molecular interaction
ubiquitinates	Molecular interaction
deubiquitinates	Molecular interaction
phosphorylates_activates	Molecular interaction
phosphorylates_inhibits	Molecular interaction
dephosphorylates_activates	Molecular interaction
dephosphorylates_inhibits	Molecular interaction
ubiquitinates_activates	Molecular interaction
ubiquitinates_inhibits	Molecular interaction
deubiquitinates_activates	Molecular interaction
deubiquitinates_inhibits	Molecular interaction
methylates_activates	Molecular interaction
methylates_inhibits	Molecular interaction
glycosylates_activates	Molecular interaction
glycosylates_inhibits	Molecular interaction
indirectly_affects_activates	Molecular interaction
involved_in	Inference
is_part_of	Inference
contains	Inference
crosstalks_with	Inference

2.2. Computation of RDF Sketch Maps

The essential goal of RDF Sketch Maps method is to reduce knowledge complexity for theranostic analysis. In the case with the integrated RDF KEGG KB we have a “hairball” of myriad of molecular interactions that needs to be simplified. To do that we first define a model of a particular biological phenomenon. For our experiments in personalized theranostics we define a *cancer model* that reflects propagation of biological signal transduction from intercellular space through surface proteomic receptors all the way into the nuclear space where specific activated protein complexes regulate gene expression. In our cancer model we identify Start and End genes, with *Start* genes being surface receptors, proto-oncogenes and tumor suppressor genes and *End* genes being genes

associated with biological processes involved in carcinogenesis (Table 2). An example of such model is shown in Figure 1.

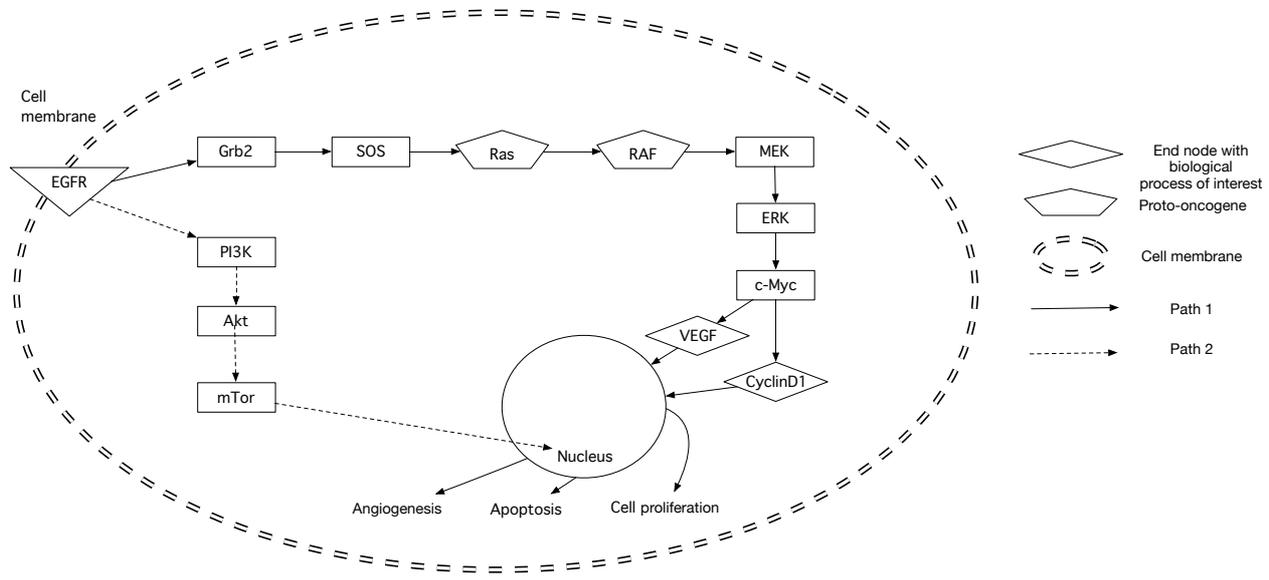


Figure 1. An example of cancer model.

We represent genes and their relationships in our KEGG RDF KB as a directed graph $G = (V, E)$, where V is a set of vertices representing KEGG genes and E as a set of edges representing gene-gene interactions. Adjacency matrix $A(i, j)$ describes whether there is a directed edge between vertices v_i and v_j . For the sake of simplicity, $A(i, j) \in \{0,1\}$. We define the *Start Gene* set as $SG = \{v_1, \dots, v_m\}$, where m is the total number of start genes. The *End Gene* set is defined as $EG = \{v_1, \dots, v_n\}$, where n is the total number of end genes. We also define a set of genes used as diagnostic, prognostic, and therapeutic biomarkers for specific cancer phenotypes. This gene set is called the *Confidence Gene* set CG . For each vertex $v \in V$, binary operator $confidence(v) \in \{0,1\}$ indicates membership of v in CG , i.e. $v \in CG$, if $confidence(v)=1$.

Table 2. Modeled biological processes.

GO ID	GO Definition
GO:0001525	Angiogenesis
GO:0006915	Apoptotic process
GO:0008150	Biological process
GO:0008283	Cell proliferation
GO:0008284	Positive regulation of cell proliferation
GO:0008285	Negative regulation of cell proliferation
GO:0016525	Negative regulation of angiogenesis
GO:0042127	Regulation of cell proliferation
GO:0042981	Regulation of apoptotic process
GO:0043065	Positive regulation of apoptotic process
GO:0043066	Negative regulation of apoptotic process
GO:0045765	Regulation of angiogenesis
GO:0045766	Positive regulation of angiogenesis
GO:0048518	Positive regulation of biological process
GO:0048519	Negative regulation of biological process
GO:0050789	Regulation of biological process

We then identify directed paths from the Start Genes to End Genes in the graph guided by the *Confidence Gene* set CG . In order to solve the problem, we formulate the problem as an $M-N$ problem, i.e., finding the optimum paths from M Start Genes to N End Genes. To divide and conquer, we also define the sub-problem of the $M-N$ problem as $I-I$ problem, which aims to find K best paths from one source gene to one sink gene in the graph, $K \geq 1$. In contrast to finding only one optimal

path, defining K optimal paths in the $I-I$ problem could provide much more depth in each single path, and these alternative paths could illustrate much more information incorporating all these paths

together in the $M-N$ problem. In the $1-1$ problem, comparing with the classical path finding problem in graph theory, which aims to find the shortest path defined by the adjacent matrix, the involvement of the *Confidence Genes* affects the path finding. The optimum paths we intend to find should be shortest and involve as many of *Confidence Gene* as possible. We could define a path with length l from a *Start Gene* $path(1)$ to an *End Gene* $path(l)$, $path(i) \in V (i=1 \dots l)$. In the classical shortest path finding problem, the objective function of the optimum path in the graph can be written as:

$$f(path) = \sum_{i=1}^{l-1} A(path(i), path(i+1)), \quad (1)$$

which only considers the topological distance. So, the problem is to find a path having $\min f(path)$. By including *Confidence Genes*, we could redefine the objective function of the optimum path as:

$$f(path) = \frac{\sum_{i=1}^l confidence(path(i))}{\sum_{i=1}^{l-1} A(path(i), path(i+1))}, \quad (2)$$

From Eq. (2), the path with the shortest distance and more *Confidence Genes* involved should be our optimum path, and the problem is redefined to find a path having $\max f(path)$. Hence, the $1-1$ problem is changed to the well-known k -shortest path problem with modified objective function. In our case, as the KEGG pathway graph contains many cycles, and since genes in such cycles might be important, we do not make the acyclic restriction. We implement Eppstein's algorithm [39] with a replaced objective function to solve the $1-1$ problem in polynomial time, which requires only computational complexity of $O(|E| + |V|\log|V|+K)$. The $M-N$ problem could be treated as an exhaustive combination of all possible $1-1$ problems with defined K in the graph. For each gene in the *Start Genes* and each gene in the *End Genes*, we obtain the K optimum paths on each pair of *Start* and *End Genes*. In total, we have M times N of $1-1$ combinations, which are $M \times N \times K$ paths. We then map these paths to the graph, and merge them together. We use Procedure 1 to solve the $M-N$ problem, as described below. In all of these paths, the importance of each path is evaluated using the objective function $f(path)$. Hence, the importance of each node in the graph is calculated by sum of the paths going through the node. The total computational complexity of the $M-N$ problem is $O(MN(|E| + |V|\log|V|+K))$.

M-N problem:

Input: Start Genes set SG , End Gene set EG , Confidence Genes set CG , directed adjacency matrix $A(i, j)$

Output: List of ranked paths by their decreasing objective function $f(path)$ values

Procedure 1:

- 1: For all $v_i \in SG$ do
- 2: For all $v_j \in EG$ do
- 3: Compute $path(i, j)$ by solving $1-1$ problem for (v_i, v_j, K)
- 4: End for
- 5: End for
- 6: Rank paths by $f(path)$ values
- 7: Output of top specified percentage of paths as RDF Sketch Map

According to the definition of the problem, the choice of objective function plays critical role in finding optimum paths. In practical usage, Eq. (2) may have limitations in favoring path with shorter length and having larger values of objective function. For instance, in a hypothetical case with a path of length 1 having a *Start Gene* to be a *Confidence Gene*, the value of objective function $f(path)$ will be maximal (i.e. 1). Such a path will be given preference over other, perhaps larger but more biologically important paths that can have many more *Confidence Genes*. To overcome this bias, we also define several other objective functions $f(path)$ as Eqs. (3-5).

$$f(path) = \frac{\sum_{i=1}^l confidence(path(i))}{\sum_{i=1}^{l-1} A(path(i), path(i+1))} - \frac{\sum_{i=1}^{l-1} A(path(i), path(i+1))}{\gamma} + \delta, \quad (3)$$

Comparing with Eq. (2), Eq. (3) adds a penalty term of the current path length divided by the maximum path length. γ is the estimated maximum path length, a predefined non-negative value. δ is a non-negative predefined value to guarantee $f(path) > 0$. In our case $\delta=1$.

$$f(path) = \frac{\sum_{i=1}^l confidence(path(i))}{\sum_{i=1}^{|\mathcal{V}|} confidence(i)} - \frac{\sum_{i=1}^{l-1} A(path(i), path(i+1))}{\gamma} + \delta, \quad (4)$$

To enforce the impact of confidence genes and reduce the redundancy of the multiple usage of path length in Eq. (3), Eq. (4) introduces the fraction of confidence genes included in the path in the left term at the right side of the equation. The path length information occurs only in the right term as in Eq. (3).

$$f(path) = \left(\frac{\alpha + \sum_{i=1}^l confidence(path(i))}{\alpha + \sum_{i=1}^{|\mathcal{V}|} confidence(i)} \right) \left(\frac{\log \left(\frac{\sum_{i=1}^{l-1} A(path(i), path(i+1))}{\sum_i \sum_j A(i, j)} \right)}{\log \left(\frac{1}{\sum_i \sum_j A(i, j)} \right)} \right), \quad (5)$$

Like Eq.(4), the left term at the right side of Eq.(5) also describes how many confidence genes are included in the path. α is a small non-negative predefined value, such that $0 < \alpha < 1/|\mathcal{V}|$. α is defined to make sure the left term's values are within the interval $[0,1]$. The right term at the right side of Eq. (5) defines the influences of path length. Logarithm is used to favor large changes in short path length. The beneficial property of the objective function defined by Eq. (5) is normalization of the objective function values to the interval $[0,1]$. Even in the absence of any *Confidence Genes*, the algorithm can still be operational and compute the shortest paths.

3. Results and Discussion

Our preliminary experiments with RDF Sketch Maps method were performed on two sets of KEGG maps associated with signal transduction pathways related to leukemic cancers such as HCL and CML.

Table 3. Reduction of 7 KEGG maps of 1,597 nodes.

	Obj. function Eq. (3)		Obj. function Eq. (4)		Obj. function Eq. (5)	
	Uncollapsed	Collapsed	Uncollapsed	Collapsed	Uncollapsed	Collapsed
Top 10%	74	17	167	72	182	79
Top 30%	119	39	177	75	227	105

For each set we ran RDF Sketch Maps algorithm for three versions of the objective function $f(path)$, described by the Eqs. (3), (4), and (5). We then counted the number of nodes in the resulting graphs involving gene instances as well as in the transformed graphs where gene instances were collapsed and represented by their respective abstract genes. The number of nodes of RDF Sketch Maps representing the 7 KEGG maps' experiment consisting of 1,597 nodes is shown in Table 3 and the resulting graphs of top 10% of paths are shown in Figure 2. The number of nodes of RDF Sketch Maps representing the extended set of 18 KEGG maps consisting of 2,873 nodes is shown in Table 4 and the resulting graphs of top 30% paths are shown in Figure 3.



Figure 2. Reducing complexity of 7 integrated KEGG maps of 1,597 nodes. Top 10%.

Table 4. Reduction of 18 KEGG maps of 2,873 nodes.

	Obj. function Eq. (3)		Obj. function Eq. (4)		Obj. function Eq. (5)	
	Uncollapsed	Collapsed	Uncollapsed	Collapsed	Uncollapsed	Collapsed
Top 10%	666	220	596	188	669	224
Top 30%	778	298	661	220	791	283

It might be readily seen from the results that the overall reduction of nodes can reach 20 folds as in case with 7 KEGG maps' experiment and top 10% of collapsed gene instances using objective function of Eq. (3) (Table 3 and Figure 2). However, the practical utility of the RDF Sketch Maps is not defined by a mere reduction of the number of biological entities to be analyzed but by its retention of important entities that can explain subtle variations in patient-specific disease

mechanisms. The objective function $f(path)$ is biased toward inclusion of *Confidence Genes* in the resulting graphs. However, the *Confidence Genes*, as we noted before, are disease-associated biomarker genes that *are already known to be related to specific disease* for which analysis is performed. To uncover new, possibly unknown mechanisms, specific to individual patient cases, the resulting graphs should retain other important biological entities *not previously associated with the disease in question*.

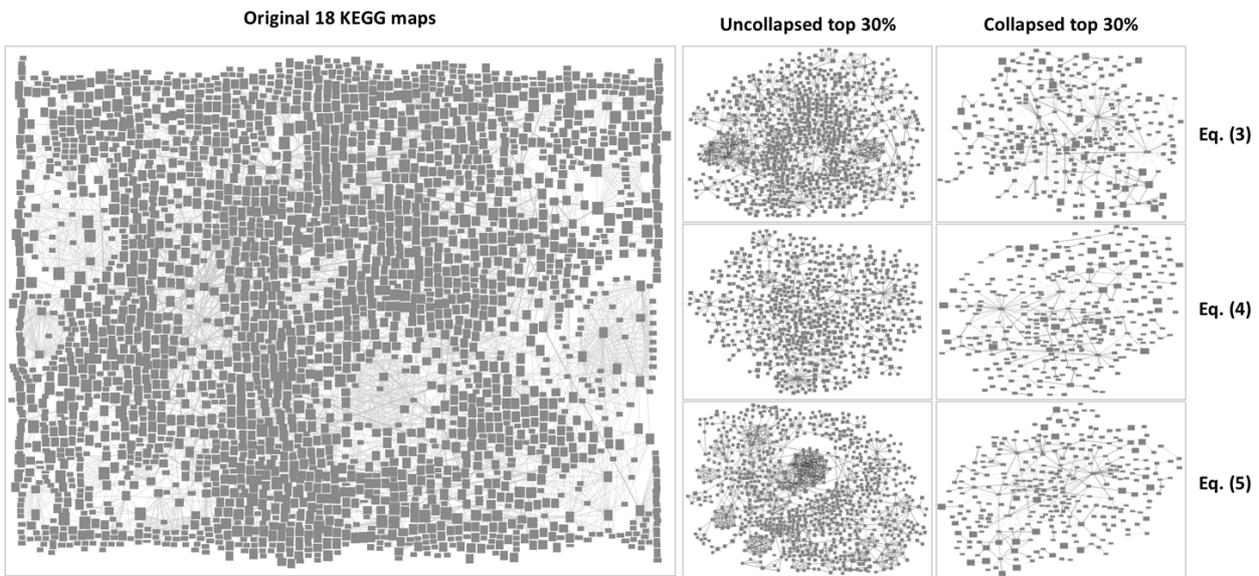


Figure 3. Reducing complexity of 18 integrated KEGG maps of 2,873 nodes. Top 30%.

To assess this quality of the RDF Sketch Maps method we define a set of *Exploratory Genes* – genes that are not directly implicated with disease in question but could possibly be useful in explaining its patient-specific disease heterogeneity (e.g. melanoma associated BRAF biomarker chosen as exploratory gene in an HCL case). We then estimate the inclusion of *Exploratory Genes* in the resulting maps.

In our preliminary experiments with leukemias the assessment procedure is done in the following way. A set of *Exploratory Genes EG* is defined as a set of genes implicated in other types of cancers. We then compute an enrichment score of *EG* genes in the resulting RDF Sketch Maps.

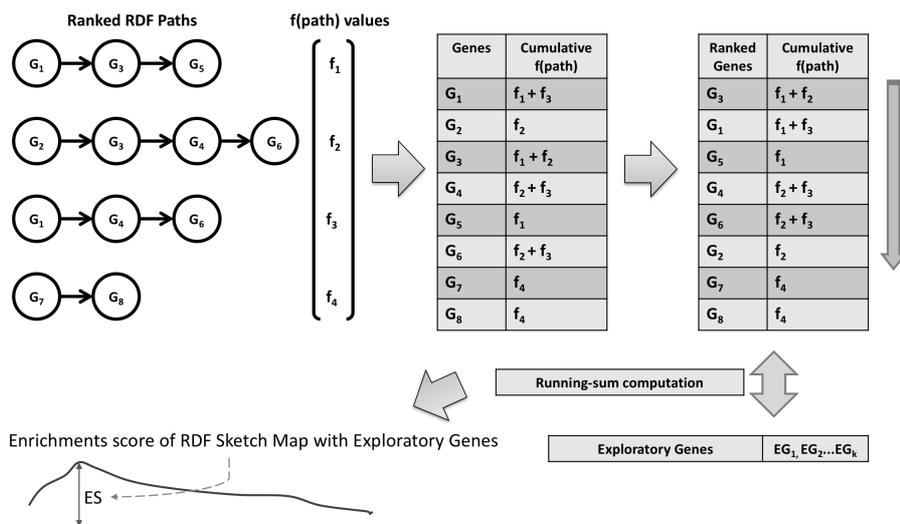


Figure 4. Extension of GSEA for RDF Sketch Maps.

To do that we extend Gene Set Enrichment Analysis (GSEA) [40]. First, we transform a ranked paths' list (see *Procedure 1*) into a ranked gene list according to their cumulative objective function $f(path)$ values (Figure 4). Then we compute a running-sum while walking down the ranked gene list in the fashion similar to the original GSEA, adding $f(path)$ value when current gene G is present in the *Exploratory Gene* set EG . The maximum deviation from zero is exported as an Enrichment Score for a specific RDF Sketch Map and *Exploratory Gene* set EG . The computed GSEA plot for objective function defined in Eq. (5) is shown in Figure 5. The non-normalized enrichment score is 0.728. The resulting RDF Sketch Map retained significant number of *Exploratory Genes* in top 10% of RDF Sketch Map paths.

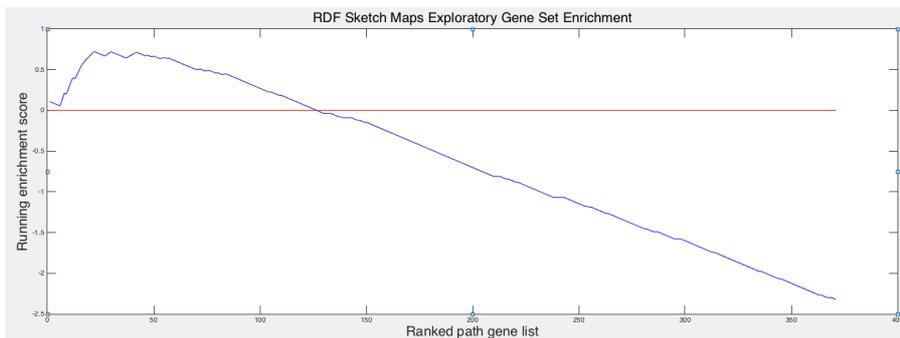


Figure 5. RDF Sketch Maps Exploratory Gene Set Enrichment Plot.

Another example of the exploratory power of RDF Sketch Maps is the fact that RDF Sketch Map of top 30% paths retained important information about signaling cascades leading to activation of proto-oncogene BRAF, which is usually

associated with a different cancer – melanoma being the prototype. An increased number of successful treatments of HCL patients by BRAF-targeted drug *vemurafenib* were recently reported. The mechanisms of the involvement of BRAF in leukemias and other tumors are now being studied. We argue here that similar hypotheses to the BRAF drug-repurposing case could be generated by using our method.

4. Conclusions and Future Directions

Our preliminary experiments have demonstrated that RDF Sketch Maps can be invaluable for hypothesis generation in precision diagnostics and therapeutics as well as for drug repurposing studies. However, we identified several directions for RDF Sketch Maps improvement. Other disease models need to be explored. Initial pruning of RDF KB networks might help to increase performance of the algorithm. Many diverse types of *Exploratory Genes* need to be investigated, such as potential drug targets. And finally, new variations of objective function need to be studied.

5. References

- [1] “FACT SHEET: President Obama’s Precision Medicine Initiative.”
- [2] H. Wang, H. Cai, L. Ao, H. Yan, W. Zhao, L. Qi, Y. Gu, and Z. Guo, “Individualized identification of disease-associated pathways with disrupted coordination of gene expression.,” *Briefings in bioinformatics*, p. bbv030, May 2015.
- [3] C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. M. Stuart, “Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM.,” *Bioinformatics*, vol. 26, no. 12, pp. i237–45, Jun. 2010.
- [4] Y. Drier, M. Sheffer, and E. Domany, “Pathway-based personalized analysis of cancer.,” *Proc Natl Acad Sci USA*, vol. 110, no. 16, pp. 6388–6393, Apr. 2013.
- [5] T. Ahn, E. Lee, N. Huh, and T. Park, “Personalized identification of altered pathways in cancer using accumulated normal tissue data.,” *Bioinformatics*, vol. 30, no. 17, pp. i422–9, Sep. 2014.
- [6] S. Ng, E. A. Collisson, A. Sokolov, T. Goldstein, A. Gonzalez-Perez, N. López-Bigas, C. Benz, D. Haussler, and J. M. Stuart, “PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis.,” *Bioinformatics*, vol. 28, no. 18, pp. i640–i646, Sep. 2012.
- [7] D. Bertrand, K. R. Chng, F. G. Sherbaf, A. Kiesel, B. K. H. Chia, Y. Y. Sia, S. K. Huang, D. S. B. Hoon, E. T. Liu, A. Hillmer, and N. Nagarajan, “Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles,” *Nucleic Acids Res*, vol. 43, no. 7, pp. gku1393–e44, Jan. 2015.
- [8] T. Kessler, H. Hache, and C. Wierling, “Integrative Analysis of Cancer-Related Signaling Pathways,” *Front. Physiol.*, vol. 4, Jun. 2013.
- [9] P. Khatri, M. Sirota, and A. J. Butte, “Ten years of pathway analysis: current approaches and outstanding challenges.,” *PLoS Comp Biol*, vol. 8, no. 2, p. e1002375, Feb. 2012.
- [10] D. Shin, G. Arthur, M. Popescu, D. Korkin, and C.-R. Shyu, “Uncovering influence links in molecular knowledge networks to streamline personalized medicine.,” *J Biomed Inform*, vol. 52, pp. 394–405, Dec. 2014.
- [11] R. E. Brown, “Morphogenomics and morphoproteomics: a role for anatomic pathology in personalized medicine,” *Arch Pathol Lab Med*, vol. 133, no. 4, pp. 568–579, Apr. 2009.
- [12] J. Liu and R. E. Brown, “Morphoproteomics demonstrates activation of mTOR pathway in anaplastic thyroid carcinoma: a preliminary observation,” *Ann Clin Lab Sci*, vol. 40, no. 3, pp. 211–217, 2010.
- [13] C. F. Streckfus, R. E. Brown, and J. M. Bull, “Proteomics, morphoproteomics, saliva and breast cancer: an emerging approach to guide the delivery of individualised thermal therapy, thermochemotherapy and monitor therapy response,” *Int J Hyperthermia*, vol. 26, no. 7, pp. 649–661, 2010.
- [14] J. Liu and R. E. Brown, “Morphoproteomics demonstrates activation of mammalian target of rapamycin pathway in papillary thyroid carcinomas with nuclear translocation of MTOR in aggressive histological variants,” Aug. 2011.
- [15] V. Subbiah, A. Naing, R. E. Brown, H. Chen, L. Doyle, P. LoRusso, R. Benjamin, P. Anderson, and R. Kurzrock, “Targeted morphoproteomic profiling of Ewing's sarcoma treated with insulin-like growth factor 1 receptor (IGF1R) inhibitors: response/resistance signatures,” *PloS one*, vol. 6, no. 4, p. e18424, 2011.
- [16] S. Dietrich, J. Hüllelin, M. Hundemer, N. Lehnert, A. Jethwa, D. Capper, T. Acker, B. K. Garvalov, M. Andrulis, C. Blume, C. Schulte, T. Mandel, J. Meissner, S. Fröhling, C. von Kalle, H. Glimm, A. D. Ho, and T. Zenz, “Continued Response Off Treatment After BRAF Inhibition in Refractory Hairy Cell Leukemia,” *JCO*, vol. 31, no. 19, pp. e300–e303, Jul. 2013.
- [17] G. A. Follows, H. Sims, D. M. Bloxham, T. Zenz, M. A. Hopper, H. Liu, A. Bench, P. Wright, M. B. van't Veer, and M. A. Scott, “Rapid response of biallelic BRAF V600E mutated hairy cell leukaemia to low dose vemurafenib,” *British Journal of Haematology*, vol. 161, no. 1, pp. 150–153, Apr. 2013.
- [18] F. Peyrade, D. Re, C. Ginet, L. Gastaud, and M. Allegra, “Low-dose vemurafenib induces complete remission in a case of hairy-cell leukemia with a V600E mutation,” ..., 2013.
- [19] L. Arcaini, S. Zibellini, E. Boveri, R. Riboni, S. Rattotti, M. Varettoni, M. L. Guerrero, M. Lucioni, A. Tenore, M. Merli, S. Rizzi, L. Morello, C. Cavalloni, M. C. Da Vià, M. Paulli, and M. Cazzola, “The BRAF V600E mutation in hairy cell leukemia and other mature B-cell neoplasms,” *Blood*, vol. 119, no. 1, pp. 188–191, Jan. 2012.
- [20] D. Shin, G. Arthur, C. Caldwell, M. Popescu, M. Petruc, A. Diaz-Arias, and C.-R. Shyu, “A pathologist-in-

- the-loop IHC antibody test selection using the entropy-based probabilistic method.,” *Journal of pathology informatics*, vol. 3, no. 1, p. 1, 2012.
- [21] S. S. Sahoo, O. Bodenreider, J. L. Rutter, K. J. Skinner, and A. P. Sheth, “An ontology-driven semantic mashup of gene and biological pathway information: application to the domain of nicotine dependence.,” *J Biomed Inform*, vol. 41, no. 5, pp. 752–765, Oct. 2008.
- [22] S. S. Sahoo, K. Zeng, O. Bodenreider, and A. Sheth, “From “glycosyltransferase” to ‘congenital muscular dystrophy’: integrating knowledge from NCBI Entrez Gene and the Gene Ontology,” *Stud Health Technol Inform*, vol. 129, no. 2, pp. 1260–1264, 2007.
- [23] M. E. Holford, H. Rajeevan, H. Zhao, K. K. Kidd, and K.-H. Cheung, “Semantic Web-based integration of cancer pathways and allele frequency data.,” *Cancer Informatics*, vol. 8, pp. 19–30, 2009.
- [24] K. M. Livingston, M. Bada, W. A. Baumgartner, and L. E. Hunter, “KaBOB: ontology-based semantic integration of biomedical databases.,” *BMC Bioinformatics*, vol. 16, no. 1, p. 126, 2015.
- [25] A. R. Pico, T. Kelder, M. P. van Iersel, K. Hanspers, B. R. Conklin, and C. Evelo, “WikiPathways: pathway editing for the people.,” *PLoS Biol.*, vol. 6, no. 7, p. e184, Jul. 2008.
- [26] M. Streit, A. Lex, M. Kalkusch, K. Zatloukal, and D. Schmalstieg, “Caleydo: connecting pathways and gene expression.,” *Bioinformatics*, vol. 25, no. 20, pp. 2760–2761, Oct. 2009.
- [27] S. Ekins, Y. Nikolsky, A. Bugrim, E. Kirillov, and T. Nikolskaya, “Pathway mapping tools for analysis of high content data.,” *Methods Mol. Biol.*, vol. 356, pp. 319–350, 2007.
- [28] “Home - Ingenuity.”
- [29] F. Zhou, S. Mahler, and H. Toivonen, “Simplification of Networks by Edge Pruning,” in *Biosociative Knowledge Discovery*, vol. 7250, no. 13, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 179–198.
- [30] M. Latendresse, M. Krummenacker, and P. D. Karp, “Optimal metabolic route search based on atom mappings,” *Bioinformatics*, vol. 30, no. 14, pp. btu150–2050, Mar. 2014.
- [31] E. Pitkänen, P. Jouhten, and J. Rousu, “Inferring branching pathways in genome-scale metabolic networks,” *BMC Systems Biology*, vol. 3, no. 1, p. 103, Oct. 2009.
- [32] T. Blum and O. Kohlbacher, “MetaRoute: fast search for relevant metabolic routes for interactive network navigation and visualization.,” *Bioinformatics*, vol. 24, no. 18, pp. 2108–2109, Sep. 2008.
- [33] M. Kanehisa and S. Goto, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” *Nucleic Acids Res*, vol. 28, no. 1, pp. 27–30, 2000.
- [34] “BioCarta - Charting Pathways of Life,” *biocarta.com*. [Online]. Available: <http://www.biocarta.com/genes/index.asp>.
- [35] D. Croft, G. O’Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D’Eustachio, and L. Stein, “Reactome: a database of reactions, pathways and biological processes,” *Nucleic Acids Res*, vol. 39, no. Database issue, pp. D691–D697, 2011.
- [36] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, “Gene ontology: tool for the unification of biology. {T}he {G}ene {O}ntology {C}onsortium,” *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.
- [37] T. U. Consortium, “The Universal Protein Resource (UniProt) in 2010,” *Nucleic Acids Res*, vol. 38, no. 1, pp. D142–D148, Jan. 2010.
- [38] J. Aasman, “Allegro graph: RDF triple database,” 2006.
- [39] D. Eppstein, *Finding the k shortest paths*. IEEE, 1994, pp. 154–165.
- [40] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles,” *Proc Natl Acad Sci U S A*, vol. 102, no. 43, pp. 15545–15550, 2005.