Toward a Unified Graph-Based Representation of Medical Data for Precision Oncology Medicine

Davide Belluomo[1], Tiziana Calamoneri[1][0000−0002−4099−1836], Giacomo Paesani[1][0000−0002−2383−1339], and Ivano Salvo[1][0000−0003−3111−701X]

[1]Computer Science Department, Sapienza University of Rome
dav.belluomo@gmail.com
{calamo,paesani,salvo}@di.uniroma1.it

Abstract. We present a new unified graph-based representation of medical data, combining genetic information and medical records of patients with medical knowledge *via* a unique knowledge graph. This approach allows us to infer meaningful information and explanations that would be unavailable by looking at each data set separately. The systematic use of different databases, managed throughout the built knowledge graph, gives new insights toward a better understanding of oncology medicine. Indeed, we reduce some useful medical tasks to well-known problems in theoretical computer science for which efficient algorithms exist.

Keywords: knowledge graph · precision oncology medicine · network medicine.

1 Introduction

One of the recent and numerous applications of graph theory is *network medicine* [1] that aims to identify, prevent, and treat diseases: graph-based approaches have offered an effective tool to systematically explore the intrinsic complexity of diseases, leading to the identification of disease specificity, disease-associated genetic mutations and a new way to assign treatments to patients. A new approach in the framework of network medicine is the so-called *personalized* or *precision* medicine, that is, the systematic use of individual patient characteristics to determine which treatment option is most likely to result in a better average outcome for the patient.

In particular, a new trend in precision oncology aims to shape drug treatments based on the specific gene mutation profile of the particular patient (see, for example, [16]). In the last 20 years, medical practitioners tested this new approach and obtained an improvement regarding its effectiveness. However, precision medicine developments have not been in line with the expectations. A common belief among oncologists and researchers in bioinformatics is that the discrepancy between real and expected performance can be partially explained by looking at the role of gene mutations in cancer evolution: the ambitious long-term goal of our research is to contribute to filling the gap in medical knowledge by examining and comparing genetic profiles of an extensive database of patients together with different treatment outcomes.

In computer science, graphs and networks are widely exploited data structures to store information as they can represent complex systems as sets of binary interactions or relations between various entities: in particular, it is possible to encode the information stored in different databases in a single graph.

Our research embraces this line of research: exploiting various databases at the same time, we represent all the available data regarding genetic information and medical records of a group of patients, together with medical knowledge in a unique *knowledge graph* and perform a guided analysis of some medical issues.

In particular, in Section 2, after giving some preliminary definitions, we formally describe how we construct our knowledge graph. Then, we show how some medical issues can be modeled as graph problems and solved through classical graph algorithms. Section 3 is devoted to showing the results of some preliminary experiments as a proof of concept. Finally, Section 4 concludes the paper.

2 Data and Methods

The main idea of this work is to collect together as much information as possible, coming from the structured data of different databases recording data from medical studies, and official documents of regulatory agencies, in order to study oncological diseases and, in particular, relationships among gene mutations, diseases, and treatment effectiveness, and try to infer vital information supporting the medical community.

In the following, we propose to use a graph (the *knowledge graph* H described in Subsection 2.2) to represent all such information in a uniform way. This approach has several advantages: we can give support to study and reduce medical issues by means of well-studied graph problems that, in turn, have well-known solutions based on efficient graph algorithms. Moreover, we can exploit the flexibility of graphs as a data structure, and easily support a function for quickly updating the information stored in the graph H; this is especially useful when new medical studies are published or when a new drug is individuated, and it is useful to inglobate this information in the graph. It is worth noting that this approach is in contrast to the static graphs created after a training phase and using them as predictive models.

2.1 Preliminary Definitions

We choose to favor intuition over formalism so, in this subsection, we informally give some basic definitions concerning graphs that will be useful in the following. The reader interested in a more formal setting can refer, for example, to [6].

A *graph* $G = (V, E)$ is constituted by a finite set V of elements, known as *nodes*, and a collection of *edges* E, connecting pairs of nodes, representing some kind of binary relation. It is possible to label some nodes and/or edges to annotate them with additional information. The set of nodes that are connected through an edge to the same node v constitute the *neighborhood* of v, and a *path* of G is a sequence of edges that joins a sequence of nodes.

A graph $G' = (V', E')$ is a *subgraph* of $G = (V, E)$ if $V' \subseteq V$, $E' \subseteq E$, and every edge in E' has both endpoints in V'. The *subgraph induced by a node set* $U \subseteq V$ is the graph whose nodes are all the nodes in U and whose edges are all the edges present in G such that both endpoints are in U.

A graph $G = (V, E)$ is *k-partite* if its node set V can be partitioned into $k > 1$ subsets, and no edge connects two nodes from the same subset. A 2-partite graph is also known as *bipartite*.

2.2 The Knowledge Graph H

The graph H, which is the core of this work, is built as the union of three graphs: the graph G, storing *Genetic information* of a group of patients (whose edges are colored in Green); the graph R, storing information from patient *medical Records* (whose edges are colored in Red), and the graph M, storing some general information that we will call *Medical knowledge* (whose edges are colored in Magenta).

While graphs G, R, and M share some (set of) nodes (for example, the set of patients), their edge sets are pairwise disjoint.

The green graph is bipartite and is defined as $G = (Pa \cup Mu, E_G)$ where:

– Pa is a set of encrypted recorded *patients* in the database; a mapping $\rho : Pa \to \mathbb{N}$ labels every patient $p \in Pa$ with their *survival period*, that is the time interval that spans between the diagnosis of a specific disease and the time of the study, if the patient is still alive, or the time of the patient's death, otherwise. Patients are also labeled with a boolean mapping $\alpha : Pa \to \{T, F\}$ that represents whether the patient is alive or not at the time of the study: for every $p \in Pa$, $\alpha(p) = T$ if p is alive and F, otherwise.

– Mu is a set of *gene mutations*. Observe that a gene can have more than one mutation;

– the set of green edges E_G contains an edge (p, m) if the patient p has the mutation m, and this edge is labeled with the variant allele frequency (VAF) associated with m for patient p;

The red graph is 3-partite and defined as $R = (Pa \cup Di \cup Dr, E_R)$ where:

– Pa is the set of patients as defined in graph G;

– Di is a set of *oncological diseases*, that affect patients in Pa;

– Dr is a set of *drugs* of interest, possibly labeled with a string β annotating adverse effects;

– the set of red edges E_R contains: • an edge (d, p) if the patient p is affected by the disease d; • an edge (p, f) if the patient p has been treated with the drug f; these edges are labeled with a pair of values (t, e) where $t \in \mathbb{N}$ is the number of treatments preceded it and $e \in \{p, u, r, n\}$ (standing for positive, unaltered, reduced, negative) represents the effectiveness of that drug; clearly, more drugs could have the same value of t when a cocktail of drugs has been administrated;

The magenta graph is 3-partite and defined as $M = (Mu \cup Di \cup Dr, E_M)$ where:

– Mu, Di, and Dr are defined as in graphs G and R;

– the set of magenta edges E_M contains: • an edge (d, m) if it is known that typically the disease d appears in presence of the gene mutation m; these edges can be labeled with a measure that quantifies the relevance of m with respect to d: in our experiments, we consider the so-called GDA Score. This score ranges between 0 and 1 and takes into account the number and type of sources (level of curation, model organisms), and the number of publications supporting the association between m and d; • an edge (m, f) if it is known that the drug f has some effect on the mutation m.

The graph $H = (V, E)$ (see Figure 1) is the union of its subgraphs G, R and M. Therefore the set of nodes is $V = Pa \cup Mu \cup Di \cup Dr$, and the set of edges is $E = E_G \cup E_R \cup E_M$, that is, the union of green, red, and magenta edges as defined above. Note that no edge of H has both endpoints in the same node subset; hence, H is 4-partite.

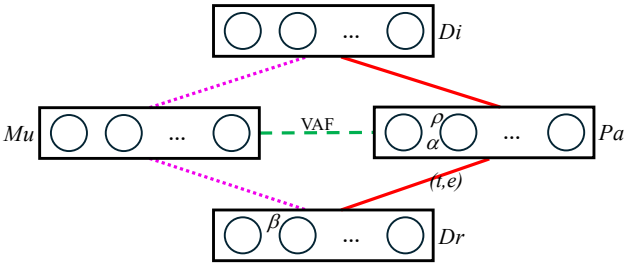

Fig. 1. A schematization of the knowledge graph H: green dashed lines represent edges of G (genetic information), red solid lines represent edges of R (information from medical records) and magenta dotted lines represent edges of M (medical knowledge). Di is the set of nodes representing the diseases, Pa the patients, Mu the genetic mutations, and Dr the drugs.

2.3 Experiment Design

In our experiments, we constructed the green graph G by exploiting database *cBioPortal* [2,3,7], from which we extract the genetic information of each patient. Gene mutations are derived from NGS (Next Generation Sequencing).

The information encoded in the red graph R can be obtained from the examination of a large number of medical records, which not only keeps track of the diseases affecting the patients, but also the treatment history, a quantitative estimation of the responses, and the survival period. In our experiments, we use cBioPortal again, to deduce this information.

At least in principle, the magenta graph M should store the vast and ever-evolving medical knowledge. In particular, in our experiments, we consider the *DisGeNET* database [21] and regulatory agencies databases (see, for example, *PharmGKB* [27,28]) for the information about the interaction between specific diseases and gene mutations, and for approved drugs targeting gene mutations, respectively.

Through knowledge graph H, we can answer many questions by exploiting graph algorithms. The underlying idea is that the involvement of multiple databases and subgraphs gives us robust and precise machinery to deduce meaningful information. Here, we list three examples: we describe the medical issue, we model it into a graph problem, propose an algorithmic solution, and highlight which part of the knowledge graph H is involved.

Comparing medical knowledge and data evidence. A natural issue is understanding whether the medical knowledge agrees with information that can be inferred from medical records. Using the information in the graph H, a possible question is the following: does data evidence from $G \cup R$ match the medical knowledge stored in M? An answer to this question can improve the understanding of the relationship between diseases and gene mutations.

On the one hand, thanks to the red edges of H, fixing a disease d, we consider set $Pa(d)$ of all patients affected by d, corresponding to all the nodes in the neighborhood of d in Pa. We define $Mu_{\cup}(d)$ as the set of all gene mutations affecting at least one patient in $Pa(d)$ and $Mu_{\cap}(d)$ as the set of all mutations affecting every patient in $Pa(d)$ (through green edges).

On the other hand, we determine through magenta edges the set $Mu_{|d}$ of gene mutations known to be involved in disease d, which results from medical knowledge. The set $Mu_{|d}$ can be computed by considering the neighbors of d in the magenta graph having GDA score close to 1.

Provided that the sample of patients is sufficiently broad, we have that $Mu_{|d}$ should be contained in $Mu_{\cup}(d)$: every gene mutation known to be involved in d necessarily occurs in some patient with disease d. Next, we compare the two sets of mutations $Mu_{|d}$ and $Mu_{\cap}(d)$ and distinguish the following cases: if $Mu_{|d} = Mu_{\cap}(d)$ then the medical knowledge perfectly matches with the experimental evidence for disease d; otherwise, either the medical knowledge is incomplete for disease d because there are gene mutations that are present in every patient with disease d but are not anticipated by the current medical knowledge, or the evidence is inconsistent for disease d because some patients with disease d do not have a predicted gene mutation, or a combination of them. In such cases, deeper examinations are suggested.

For each $d \in Di$ computing $Mu_{|d}$, $Mu_{\cup}(d)$ and $Mu_{\cap}(d)$ can be efficiently done through standard graph search algorithms, such as breadth-first search.

Partitioning patients into homogenous groups. Medical evidence shows (*e.g.*, see [22]) that the percentage of patients that positively react to treatments is less than expected, although drugs are chosen based on the patient specific gene mutation profile. The general feeling among experts is that concentrating on driver gene mutations is not enough.

To face this general problem, here we propose some possible approaches, all based on the idea of recognizing groups of patients that, for some reason, can be considered as similar. Then, we can propose to medical doctors a deep analysis of the gene mutations of patients in the same groups, so that they can look for the presence of specific gene mutations that the drug treatment has not

targeted: some of these gene mutations could inhibit the cure and be considered responsible for the treatment failure.

As a first approach we consider a *similarity function* between patient genetic profiles (see, for example, [15]); the idea is that if a patient has been successfully treated with a drug, patient with a similar genetic profile could be successfully administered with the same drug. More in detail, given a threshold k, exploiting green edges of H, we can determine patients with a genetic mutation profile distant less than k each other.

This approach is related to the so-called *agnostic paradigm*, in which patients with very similar genetic profiles are administered with the same drug, regardless of the tumor each patient has been diagnosed. Nevertheless, the agnostic paradigm, although biologically fascinating, did not produce significant effects apart from very few cases (*e.g.*, NTRK [17]).

As an example, note that in the previous section we implicitly considered groups of patients affected by the same disease. As a second aproach, we also propose to further partition patients with the same disease d according to their *survival period* (available from the magenta graph), clearly strongly related to the effectiveness of the administered drugs. The goal is to check if patients in the same group exhibit deeper similarities in the genetic profile with respect to the whole set of patients and provide evidence of some treatments' (in)effectiveness.

Optimized drug treatments. Here, we propose an algorithm joining the information obtained separately on the one hand from $G \cup M$ and on the other hand from R to suggest drug treatments optimizing the benefits and minimizing adverse effects for a specific patient.

For any gene mutation m, we can exploit magenta edges to deduce the set $Dr_{|m}$ of the drugs that have an effect on m. Hypothetically, administering to a patient p all the drugs in $\bigcup_{m_p} Dr_{|m_p}$ (where m_p is any gene mutation of p, selected through the green edges) would guarantee the best treatment for p.

Nevertheless, given the possible adverse effects of these drugs (possibly depending on their interactions), only few of them can be administered simultaneously to a patient, even at the cost of ignoring some gene mutations. Indeed, in practice, only very few mutations of a patient are being treated: current drugs are designed to deal with very specific gene mutations, known as *target*. Hence, given a patient p and a (small) subset Z of their gene mutations, the aim is to compute a drug subset W of minimum size that targets all gene mutations in Z.

This problem is related to the well-studied *minimum hitting set problem*, defined as follows. Let U be a finite set and $\mathcal{U} = \{U_1, U_2, \ldots\}$ a collection of subsets of U. A *hitting set* for $\mathcal{U}$ is a subset U' of U such that $U_i \cap U' \neq \emptyset$, for every i. The minimum hitting set problem consists of determining a hitting set of minimum size. Computing the minimum hitting set is known to be computationally hard [23].

Our problem can be modeled in terms of the minimum hitting set problem as follows: U coincides with the set of the drugs Dr and each U_i is a $Dr_{|m}$, for

some $m \in Z$. Thus, solving the minimum hitting set problem on this instance gives a minimum-size drug treatment that targets all gene mutations in Z.

In the literature, some papers propose similar strategies. In particular, in [25], the authors solve the hitting set problem with a heuristic approach restricting to drug combinations of size at most three. Note that there are some papers that aim to solve the adequate drug treatment problem. The work of Johnson [14] highlights a polynomial-time heuristic approximation algorithm. Finally, in [18], it has been developed a statistical mechanics approach to attack this problem. We point out that these previous approaches are either non-deterministic or do not obtain exact solutions.

In contrast to the previous work, we propose a deterministic and exact algorithm to solve the adequate drug treatment problem, which can be generalized by taking into account the adverse effects of drugs, minimizing both them and the number of involved drugs, thus improving the precision and safety of drug treatments. The main idea is that we add a node weight on the drug nodes related to their toxicity and solve the problem by computing a minimum weight hitting set. Given that the set of target gene mutations is small in practice, the proposed is computationally reasonable.

3 Results

In this section, we show the results of some experiments. Due to the lack of some crucial information in the public databases and the difficulty of getting some part of it, we only partially address the objectives described in Subsection 2.2; nevertheless, we try to keep the flavor of the underlying idea.

We focus on three different medical studies: Metastatic Non-Small Cell Lung Cancer [13] with 930 patients, MSK MetTropism [20] with 24755 patients, and MSK-IMPACT Clinical Sequencing Cohort [29] with 7091 patients. We chose these studies because they consider a sequencing technology guaranteeing a 500-gene panel for each patient. Nevertheless, for the sake of brevity, we report only the results concerning the first study, but the interested reader can find all the other results in [**?**].

Analysis of data in public databases. Preliminarily, we observed that the databases we use as reference, CBioPortal and DisGeNET, are not coherent; indeed, while the former contains specific genetic mutations, the latter deals only with mutated genes. It follows that, in order to compare the extracted results, we have to downgrade the genetic mutations to simple mutated genes. In order to have an idea about how much information we are losing in this way, we compare the data extracted from CBioPortal, counting them in different ways. In Table 1, we show the following information for the study MSK MetTropism:

– the percentage of the 10 most frequent mutations with respect to the total number of mutations;

– the percentage of the 10 most frequent mutated genes, where all the mutations on the same gene are counted;

– the percentage of the 10 most frequent mutated genes, where multiple mutations on the same gene are counted as one.

gene mutations	%	genes (with mult.)	%	genes (w/o mult.)	%
KRAS_12_25398284_25398284	12.7	TP53	43.6	TP53	50.3
TERT_5_1295228_1295228	6.7	KRAS	21.7	KRAS	22.0
KRAS_12_25398285_25398285	4.8	PIK3CA	14.2	APC	14.7
PIK3CA_3_178936091_178936091	3.3	APC	9.8	PIK3CA	14.5
BRAF_7_140453136_140453136	3.2	TERT	9.6	TERT	11.5
PIK3CA_3_178952085_178952085	3.2	EGFR	4.8	ARID1A	10.2
TP53_17_7578406_7578406	2.7	BRAF	4.4	PTEN	7.2
PIK3CA_3_178936082_178936082	2.0	PTEN	3.4	KMT2D	7.1
TP53_17_7577120_7577120	1.7	ARID1A	3.1	EGFR	6.6
TP53_17_7577538_7577538	1.7	CDKN2A	2.8	BRAF	6.2

Table 1. Results of an experiment performed on the 24755 patients of MSK MetTropism study: we show the 10 most frequent mutations (first columns), mutated genes with multiplicity (second columns), and mutated genes without multiplicity (third columns) together with the corresponding percentages.

From this data, it is evident that not only it is completely different to consider percentages of mutated genes instead of specific mutations, but it is also different to take into account multiplicity instead of ignoring it. It follows that the setup used to extract each data from the knowledge graph H must be accurately detailed to medical doctors.

Comparing medical knowledge and data evidence: Lung Adenocarcinoma. We now consider only the patients characterized by the same disease $d \in Di$.

We show in Table 2 analogous data w.r.t. those shown in Table 1 when the patients are only the 3972 patients affected by one of the most frequent diseases included in the MSK MetTropism study, namely Lung Adenocarcinoma.

Comparing Tables 1 and 2, one can observe sensible differences: for example, the gene mutation KRAS_12_25398285_25398285 appears in only 4.8% of all patients while in 14.7% of those affected by Lung Adenocarcinoma. Moreover, the gene mutation TERT_5_1295228_1295228 appears in 6.7% of the patients in Table 1 while is negligible in Table 2. An even more notable discrepancy can be observed in the gene EGFR: only 6.6% of the total population of patients has this gene mutated, while the percentage increases to 29.4 for the patients affected by Lung Adenocarcinoma. These considerations are not meant to infer any conclusion at the medical level but, especially if joined with similar studies, aim to suggest a direction for further research.

One of the features of the DisGeNET database is an association between diseases and genes that is confirmed by the so-called GDA Score, that labels magenta edges of H.

gene mutations	%	genes (with mult.)	%	genes (w/o mult.)	%
KRAS_12_25398285_25398285	14.7	TP53	49.2	TP53	48.5
KRAS_12_25398284_25398284	13.6	EGFR	34.9	KRAS	33.9
EGFR_7_55259515_55259515	9.1	KRAS	34	EGFR	29.4
EGFR_7_55242465_55242479	5.0	STK11	13.9	STK11	16.1
EGFR_7_55242466_55242480	2.9	KEAP1	11.1	KEAP1	14.1
EGFR_7_55249071_55249071	2.9	RBM10	7.7	RBM10	11.7
U2AF1_21_44524456_44524456	2.0	PIK3CA	5.5	PTPRD	8.8
PIK3CA_3_178936091_178936091	1.8	BRAF	4.7	SMARCA4	8.2
ERBB2_17_37880981_37880982	1.6	CDKN2A	4.2	ATM	7.8
KRAS_12_25380275_25380275	1.5	SMARCA4	3.9	NF1	7.5

Table 2. Results of an experiment performed on the MSK MetTropism study on the 3972 patients affected by lung adenocarcinoma: we show the 10 most frequent mutations (first columns), mutated genes with multiplicity (second columns), and mutated genes without multiplicity (third columns) together with the corresponding percentages.

mutated gene	GDA Score	mutated gene	GDA Score	mutated gene	GDA Score
BRAF	1.0	FGFR2	0.85	MAP2K1	0.8
ALK	1.0	AKT1	0.85	CTNNB1	0.8
ROS1	1.0	MUC5AC	0.85	CDKN2A	0.8
KRAS	1.0	TYMS	0.85	RAF1	0.8
EGFR	1.0	CCND1	0.85	CHRNA3	0.8
ERBB2	0.95	STK11	0.8	FGFR3	0.8
PIK3CA	0.95	TERT	0.8	ATM	0.8
TP53	0.95	FGFR4	0.8	EGF	0.8
MYC	0.9	HRAS	0.8		

Table 3. Mutated genes with GDA score at least 0.8 in lung adenocarcinoma.

On the one hand, it is clear that no single gene mutation appears in all patients affected by lung adenocarcinoma. Therefore, $Mu_{\cap}(d)$ is trivially empty. On the other hand six of the genes appearing in Table 3, namely BRAF, KRAS, EGFR, STK11, ATM, and TP53, are also represented in Table 2 showing a level of agreement between medical knowledge on lung adenocarcinoma disease and the evidence collected on the patients. Anyway, some genes appearing in Table 3 with GDA score 1, namely ALK and ROS1, do not appear in Table 2 and hence in $Mu_{\cup}(d)$, and some genes appearing in Table 2 with frequency above 10% without multiplicity, namely KEAP1 and RBM10, do not appear in Table 3, showing some inconsistencies between medical knowledge and data that we have considered.

Partitioning patients into homogenous groups: survival period. Estimating the effectiveness of drug treatment is a difficult task because it takes into account different parameters. One of these parameters is the survival period.

We partition the patient population into three sets: the first one $Pa_{\geq 36} = \{p \in Pa \mid \rho(p) \geq 36\}$ contains all the patients whose survival period is of at least 36 months, the second one $Pa_{\leq 6} = \{p \in Pa \mid \rho(p) \leq 6 \wedge \alpha(p) = F\}$ contains

all the patients whose survival period is of at most 6 months and are marked as deceased, and the third set contains all the remaining patients.

gene mutations	%	genes (with mult.)	%	genes (w/o mult.)	%
KRAS_12_25398284_25398284	8.6	TP53	35.6	TP53	38.8
TERT_5_1295228_1295228	6.6	PIK3CA	16.9	PIK3CA	17
PIK3CA_3_178952085_178952085	4.7	KRAS	15.4	KRAS	15.7
KRAS_12_25398285_25398285	3.6	TERT	9.9	APC	13.7
PIK3CA_3_178936091_178936091	3.5	APC	9.7	TERT	11.5
BRAF_7_140453136_140453136	3.0	EGFR	6.6	ARID1A	9.3
PIK3CA_3_178936082_178936082	2.6	BRAF	4.8	EGFR	7.8
TP53_17_7578406_7578406	2.1	PTEN	4.7	PTEN	7.2
EGFR_7_55259515_55259515	1.8	ARID1A	3.7	FAT1	6.2
TP53_17_7577538_7577538	1.4	CTNNB1	3.0	PTPRT	6.1

Table 4. Results of the experiment on the MSK MetTropism study on the 5295 patients with a survival period of at least 36 months: we show the 10 most frequent mutations (first columns), mutated genes with multiplicity (second columns), and mutated genes without multiplicity (third columns) together with the corresponding percentages.

gene mutations	%	genes (with mult.)	%	genes (w/o mult.)	%
KRAS_12_25398284_25398284	15.3	TP53	57.8	TP53	62.1
TERT_5_1295228_1295228	8.2	KRAS	27.9	KRAS	27.9
KRAS_12_25398285_25398285	7.1	TERT	12.1	TERT	13.3
BRAF_7_140453136_140453136	3.3	PIK3CA	11.4	PIK3CA	12.2
PIK3CA_3_178952085_178952085	2.8	APC	6.4	ARID1A	10.2
TP53_17_7578406_7578406	2.7	CDKN2A	6.0	APC	10.1
PIK3CA_3_178936091_178936091	2.6	BRAF	5.2	CDKN2A	9.5
PIK3CA_3_178936082_178936082	2.3	STK11	4.0	KEAP1	7.2
TP53_17_7577538_7577538	2.2	EGFR	3.7	STK11	6.9
TP53_17_7577094_7577094	2.0	SMAD4	3.3	RB1	6.8

Table 5. Results of the experiment on the MSK MetTropism study on the 2768 patients that have a survival period of at most 6 months: we show the 10 most frequent mutations (first column), mutated genes with multiplicity (second column), and mutated genes without multiplicity (third column) together with the corresponding percentages.

We selected the 5295 patients of the MSK MetTropism study in $Pa_{\geq 36}$ and the 2768 patients in $Pa_{\leq 6}$ and summarized the results in Tables 4 and 5, respectively. Comparing Table 1 with Tables 4 and 5, one can observe that the distribution of the percentages of their mutations completely changes. As an example, TP53, KRAS, and TERT dramatically increase their percentages, whereas PIK3CA, EGFR, and STR11 decrease significantly their percentages. This behavior can be explained by the medical awareness that certain combinations of mutations indicate either a different response to treatments or a different evolution of the disease. A deep study of these results should be performed by medical doctors, who could individuate interesting combinations of gene mutations, both

in the population of patients with a low survival period and in that one with a high survival period.

Partitioning patients into homogenous groups: genetic mutation profile. Moreover, to understand whether there are some groups of patients that can be considered similar, we implemented two different similarity measures based on the genetic profiles. The *Hamming distance* [9] between two patients counts the number of gene mutations that affect only one of them. The *Jaccard distance* [12] is a variation of the Hamming distance where the value is normalized by the total number of gene mutations affecting the two considered patients, taking into account the inequalities due to the possibly imbalanced number of observed gene mutations or different gene sequencing (*e.g.*, different number of checked genes). Clearly, two patients having a similar genetic profile are also very close with respect to the considered measures.

The overall idea is that patients who are grouped together, whether they have either Hamming or Jaccard distance small, might experience the same disease, similar disease evolution, and comparable responses to drug treatments.

Regardless of the similarity measure used, our experiments show that most of the patients are isolated, that is, the groups of similar patients are very often singletons. As an example, we report the results obtained from the MSK MetTropism study when considering patients at Hamming distance at most 10 from each other. Out of 24755 patients, there are only 6 groups that are not singletons which include a total of 14 patients. This means, at least considering the data at our disposal, that it is very unlikely that any two patients are similar from the genetic profile point of view.

As expected, these results confirm that genetic similarities are not enough to explain different behaviors of the human body with respect to oncology medicine. Since the techniques we use to aggregate patients are not the most sophisticated nor the most appropriate for this specific task, in the following we propose more advanced clustering techniques.

Paritioning Patient in homogenous groups: coexisting mutations. We wonder whether there exist some combinations of gene mutations that appear simultaneously in significant portions of the patient population: we compute all the (maximal) k-coexisting-mutation sets, *i.e.*, sets of mutations simultaneously present in at least k% of patients, and return these sets of patients.

From the evaluation of the data, we observe that k-coexisting-mutation sets are made of a single mutation, even for very small values of k. Considering the MSK MetTropism study again, there is only one k-coexisting-mutation set, with $k = 12$, which consists of the single mutation KRAS_12_25398284_25398284. It seems unrealistic to assume that every patient affected by KRAS_12_25398284_25398284 can be considered similar with respect to the affected disease, disease evolution, and response to drug treatments.

Some combinations of gene mutations are particularly relevant for medical doctors: contemporary mutation in genes EGFR and KRAS is one of them. So, we extracted all the patients with both these two genes mutated, whose 61.3%

gene mutations	% patients	% living patients	% deceased patients
KRAS	100	61.3	38.7
EGFR	100	61.3	38.7
TP53	52.7	51.0	49.0
APC	45.2	76.2	23.9
ARID1A	38.7	83.3	16.7
KMT2D	35.5	93.9	6.1
PIK3CA	35.5	90.9	9.1
FAT1	35.5	81.8	18.2
ATM	33.3	77.4	22.6
PTEN	31.2	89.7	10.3

Table 6. Results of an experiment performed on the MSK MetTropism study on the 93 patients that have both EGFR and KRAS genes mutated: we show the 10 most frequent mutations (first column), the percentage of patients with that gene mutated (second column), and the percentages of living and deceased patients among those with that gene mutated (third and fourth columns).

of them were alive at the time of the study. It is natural to wonder whether there is an explanation for the alive patients to survive. For each analyzed gene, we computed the fraction (expressed as a percentage) of the patients having that gene mutated in three different patient populations: all patients, the living, and the deceased ones. It turns out that the patients with certain further mutations (such as ARID1A, PIK3CA, AT1, or PTEN) are much more likely to survive, as shown in Table 6, indeed the percentage of living patients is more than 80% in the presence of these gene mutations. This kind of table could be of interest to better understand whether there are special combinations of gene mutations that significantly increase the survival probability. The results of this experiment for other combinations of genes, such as EGFR and T790R, do not provide the same interesting output: for example, in all three considered studies, no patient had these two genes mutated at the same time.

4 Conclusions

In this paper, we designed a unified graph-based representation of medical data for precision oncology medicine and proposed three possible applications whose solutions exploit known results from theoretical computer science.

Our approach's novelty lies in how we store and deduce information. In particular:

– we develop a knowledge graph that exploits various databases to deduce fundamental information using graph-theoretic tools;

– we implement a deterministic framework to infer personalized medical information in contrast to past research strategies that have been using data aggregation [5,19,24], pattern recognition [4,10] and statistical performance [11,26];

– our knowledge graph model allows one for quick and efficient updates, whether there is a new node or some information has changed, in contrast to static models based on machine learning techniques (see for example [8]).

Acknowledgments. The authors would like to thank medical doctors Gennaro Daniele and Pasquale Lombardi for the exciting discussions on cancer handling from a medical point of view, Kilian Schulz for contributing to extracting data from databases during his honors program, Luciano Giacò and Federica Persiani for their helpful feedback on the databases to be used.

This work was supported by *Sapienza* University of Rome, project title: *Graph models for precision oncology medicine*, grant number: RM122181612C08BB.

References

1. Barabási, A.L., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. Nature reviews genetics **12**(1), 56–68 (2011). https://doi.org/10.1038/nrg2918
2. de Bruijn, I., Kundra, R., Mastrogiacomo, B., Tran, T.N., Sikina, L., Mazor, T., Li, X., Ochoa, A., Zhao, G., Lai, B., et al.: Analysis and visualization of longitudinal genomic and clinical data from the aacr project genie biopharma collaborative in cbioportal. Cancer research **83**(23), 3861–3867 (2023). https://doi.org/10.1158/0008-5472.CAN-23-0816
3. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al.: The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery **2**(5), 401–404 (2012). https://doi.org/10.1158/2159-8290.CD-12-0095
4. Cheng, T., Zhan, X.: Pattern recognition for predictive, preventive, and personalized medicine in cancer. EPMA Journal **8**, 51–60 (2017). https://doi.org/https://doi.org/10.1007/s13167-017-0083-9
5. Cirillo, D., Valencia, A.: Big data analytics for personalized medicine. Current Opinion in Biotechnology **58**, 161–167 (2019). https://doi.org/https://doi.org/10.1016/j.copbio.2019.03.004
6. Diestel, R.: Graph Theory, 4th Edition, Graduate texts in mathematics, vol. 173. Springer (2012)
7. Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al.: Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. Science signaling **6**(269), pl1–pl1 (2013). https://doi.org/10.1126/scisignal.2004088
8. Gong, F., Wang, M., Wang, H., Wang, S., Liu, M.: Smr: medical knowledge graph embedding for safe medicine recommendation. Big Data Research **23**, 100–174 (2021). https://doi.org/10.1016/j.bdr.2020.100174
9. Hamming, R.W.: Error detecting and error correcting codes. The Bell System Technical Journal **29**(2), 147–160 (1950). https://doi.org/10.1002/j.1538-7305.1950.tb00463.x
10. Huang, Y., Zhao, Y., Capstick, A., Palermo, F., Haddadi, H., Barnaghi, P.: Analyzing entropy features in time-series data for pattern recognition in neurological conditions. Artificial Intelligence in Medicine **150**, 102821 (2024). https://doi.org/https://doi.org/10.1016/j.artmed.2024.102821
11. Indrayan, A.: Personalized statistical medicine. IJMR **1157**(1), 104–108 (2023). https://doi.org/10.4103/ijmr.ijmr_1510_22
12. Jaccard, P.: The distribution of the flora in the alpine zone. 1. New phytologist **11**(2), 37–50 (1912). https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

13. Jee, J., Lebow, E.S., Yeh, R., Das, J.P., Namakydoust, A., Paik, P.K., Chaft, J.E., Jayakumaran, G., Rose Brannon, A., Benayed, R., et al.: Overall survival with circulating tumor dna-guided therapy in advanced non-small-cell lung cancer. Nature medicine **28**(11), 2353–2363 (2022). https://doi.org/10.1038/s41591-022-02047-z
14. Johnson, D.S.: Approximation algorithms for combinatorial problems. Proc. STOCS 1973 pp. 38–49 (1973). https://doi.org/10.1145/800125.804034
15. Kosman, E., Jokela, J.: Dissimilarity of individual microsatellite profiles under different mutation models: Empirical approach. Ecology and Evolution **9**(7), 4038–4054 (2019). https://doi.org/https://doi.org/10.1002/ece3.5032
16. Langreth, R., Waldholz, M.: New era of personalized medicine: targeting drugs for each unique genetic profile. The oncologist **4**(5), 426–427 (1999). https://doi.org/0.1634/theoncologist.4-5-426
17. Marchetti, A., Ferro, B., Pasciuto, M.P., Zampacorta, C., Buttitta, F., D'Angelo, E.: NTRK gene fusions in solid tumors: agnostic relevance, prevalence and diagnostic strategies. Pathologica **114**, 199–216 (2022). https://doi.org/10.32074/1591-951X-787
18. Mézard, M., Tarzia, M.: Statistical mechanics of the hitting set problem. Physical Review E **76**(4), 041124 (2007). https://doi.org/10.1103/PhysRevE.76.041124
19. Moscatelli, M., Manconi, A., Pessina, M., Fellegara, G., Rampoldi, S., Milanesi, L., Casasco, A., Gnocchi, M.: An infrastructure for precision medicine through analysis of big data. BMC Bioinformatics **19**(Suppl 10) (2018). https://doi.org/10.1186/s12859-018-2300-5
20. Nguyen, B., Fong, C., Luthra, A., Smith, S.A., DiNatale, R.G., Nandakumar, S., Walch, H., Chatila, W.K., Madupuri, R., Kundra, R., et al.: Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. Cell **185**(3), 563–575 (2022). https://doi.org/10.1016/j.cell.2022.01.003
21. Piñero, J., Ramírez-Anguita, J.M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., Furlong, L.I.: The disgenet knowledge platform for disease genomics: 2019 update. Nucleic acids research **48**(D1), D845–D855 (2020). https://doi.org/10.1093/nar/gkz1021
22. Plana, D., Palmer, A.C., Sorger, P.K.: Independent drug action in combination therapy: Implications for precision oncology. Cancer Discov. **12**(3), 606–624 (2022). https://doi.org/10.1158/2159-8290.CD-21-0212
23. Shi, L., Cai, X.: An exact fast algorithm for minimum hitting set. Proc. IJCCSO 2010 **1**, 64–67 (2010). https://doi.org/10.1109/CSO.2010.240
24. Ullah, A., Azeem, M., Ashraf, H., Alaboudi, A.A., Humayun, M., Jhanjhi, N.: Secure healthcare data aggregation and transmission in iot - a survey. IEEE Access (2021). https://doi.org/10.1109/ACCESS.2021.3052850
25. Vazquez, A.: Optimal drug combinations and minimal hitting sets. BMC Systems Biology **81**(3), 1–6 (2009). https://doi.org/10.1186/1752-0509-3-81
26. Venkatraman, D.L., Pulimamidi, D., Shukla, H.G., Hegde, S.R.: Tumor relevant protein functional interactions identified using bipartite graph analyses. Scientific Reports **11**(1), 21530 (2021). https://doi.org/10.1038/s41598-021-00879-2
27. Whirl-Carrillo, M., Huddart, R., Gong, L., Sangkuhl, K., Thorn, C.F., Whaley, R., Klein, T.E.: An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. Clinical Pharmacology & Therapeutics **110**(3), 563–572 (2021). https://doi.org/10.1002/cpt.2350
28. Whirl-Carrillo, M., McDonagh, E.M., Hebert, J.M., Gong, L., Sangkuhl, K., Thorn, C.F., Altman, R.B., Klein, T.E.: Pharmacogenomics knowledge for personalized medicine. Clinical Pharmacology & Therapeutics **92**(4), 414–7 (2012). https://doi.org/10.1038/clpt.2012.96

29. Zehir, A., Benayed, R., Shah, R.H., Syed, A., Middha, S., Kim, H.R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S.M., et al.: Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nature medicine **23**(6), 703–713 (2017). https://doi.org/10.1038/nm.4333