BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 08/15 pp107−130
DOI: 10.26599/BDMA.2023.9020007
Volume 7, Number 1, March 2024

Identification of Proteins and Genes Associated with Hedgehog Signaling Pathway Involved in Neoplasm Formation Using Text-Mining Approach

Nadezhda Yu. Biziukova, Sergey M. Ivanov, and Olga A. Tarasova*

Abstract: Analysis of molecular mechanisms that lead to the development of various types of tumors is essential for biology and medicine, because it may help to find new therapeutic opportunities for cancer treatment and cure including personalized treatment approaches. One of the pathways known to be important for the development of neoplastic diseases and pathological processes is the Hedgehog signaling pathway that normally controls human embryonic development. Systematic accumulation of various types of biological data, including interactions between proteins, regulation of genes transcription, proteomics, and metabolomics experiments results, allows the application of computational analysis of these big data for identification of key molecular mechanisms of certain diseases and pathologies and promising therapeutic targets. The aim of this study is to develop a computational approach for revealing associations between human proteins and genes interacting with the Hedgehog pathway components, as well as for identifying their roles in the development of various types of tumors. We automatically collect sets of abstract texts from the NCBI PubMed bibliographic database. For recognition of the Hedgehog pathway proteins and genes and neoplastic diseases we use a dictionary-based named entity recognition approach, while for all other proteins and genes machine learning method is used. For association extraction, we develop a set of semantic rules. We complete the results of the text analysis with the gene set enrichment analysis. The identified key pathways that may influence the Hedgehog pathway and their roles in tumor development are then verified using the information in the literature.

Key words: text-mining; data mining; Hedgehog pathway; neoplastic processes; enrichment analysis; pathology molecular mechanisms

1 Introduction

The Hedgehog (Hh) signaling pathway includes components that are key players in embryonic development, while it is mostly inactive in adults, excluding processes of tissue repair such as wound healing[1, 2]. However, multiple studies showed activation of proteins included in the Hh pathway in cancer development. In particular, some components of the Hh pathway may be upregulated in radio and chemo-resistant types of tumors, most of which are brain tumors[3]. Involvement of the Hh pathway proteins in cancer progression can be explained by the convergent functions of these proteins in embryonic development and tumor formation through the regulation of proliferation, differentiation, and migration[4].

• Nadezhda Yu. Biziukova and Olga A. Tarasova are with the Department of Bioinformatics, Institute of Biomedical Chemistry, Moscow 119121, Russia. E-mail: nad.smol@gmail.com; olga.a.tarasova@gmail.com.
• Sergey M. Ivanov is with the Department of Bioinformatics, Institute of Biomedical Chemistry, Moscow 119121, Russia, and also with Department of Bioinformatics, Pirogov Russian National Research Medical University, Moscow 117997, Russia. E-mail: smivanov7@gmail.com.
* To whom correspondence should be addressed.
Manuscript received: 2022-12-20; revised: 2023-04-11; accepted: 2023-04-25

There are multiple experimental studies aimed at the particular mechanisms leading to tumor formation, which involve activation or suppression of the Hh pathway genes leading to upregulation of the particular proteins[3, 5, 6]. For instance, the canonical Hh pathway includes interaction of the Hedgehog proteins (the so-called Hedgehog ligands: Shh, Ihh, and Dhh) with Ptch followed by motion of the Smo molecule to the primary cilium and phosphorylation of the C-terminus, leading to the activation of Gli transcription factors that, in turn, regulate the expression of particular genes[3, 7]. It is supposed that some specific nodes involved in the regulation of the canonical Hh pathway are not studied enough[7]. However, it is known that the Hh pathway is essential for the control of proper development, tissue homeostasis, cancer, and metabolism[2, 3, 7]. It is worth noting that the function of specific proteins in a particular type of cancer is still explored insufficiently[2]. Currently, many results on the Hh pathway regulation are available from multiple experiments yielding in several dozen thousands of publications. It allows exhaustive analysis of textual information for extracting associations between the components of the Hh pathway and other human genes or proteins, followed by further identification of the biological processes they control. Despite the fact that a large set of studies has been performed, there are some open questions about the role of the Hh pathway in various types of tumors. Also, the particular mechanisms of Gli transcription factors are widely discussed, and additional details of Gli-associated regulation of the Hh pathway can shed light on the mechanisms of cancer progression[8]. The processes regulating non-canonical activation of the Hh pathway and their regulation in particular types of tumors are of interest. The purpose of the study is to develop a computational approach for the identification of the proteins and molecular pathways associated with the key players of the Hedgehog pathway and to understand the role of the identified proteins and pathways in the development of tumors.

2 Experiment

We used a combined text- and data-mining approach and gene set enrichment analysis to identify the key proteins involved in the regulation of proteins and genes of the Hedgehog pathway and playing a role in the development of various types of tumors. We developed the workflow to study the details of the regulation of the Hh pathway. It includes the application of the text mining procedure followed by the gene set enrichment analysis. To identify proteins potentially interacting with the components of the Hh pathway, we (1) collected texts of publications relevant to the studies of the Hh pathway; (2) extracted from texts the names of proteins and genes; and (3) identified interactions between proteins and genes of the Hh pathway and those that were extracted from literature. It is known that malfunctioning of some proteins that are included in the Hh pathway is associated with the risk of tumor formation[9]. To identify common and rare tumors that can develop as a consequence of aberrant Hh pathway signaling, we extracted the names of neoplastic diseases and processes that are associated with the Hh pathway malfunction using text mining.

Then we applied gene set enrichment analysis for the identification of the genes and proteins involved in the direct or indirect regulation of the Hh pathway.

The scheme of the computational algorithm that includes information extraction for identifying the key proteins interacting with the nodes of the Hh pathway and gene set enrichment analysis that is used to study the regulation of the Hh pathway is provided in Fig. 1.

We assume that such an algorithm for information extraction can be used for complex analysis of the Hh pathway regulation based on the text and data retrieval.

2.1 Selection of texts by relevance to the analysis of the Hedgehog pathway

We collected from NCBI PubMed a set of texts relevant to the investigation of proteins known to be involved in the Hh pathway and interacting with it, namely “Hh interactions”. The extraction of relevant texts from PubMed is based on the use of Medical Subject Headings (MeSH). The selection of MeSH-terms for extracting relevant publications was carried out as follows: using simple keywords (“interaction”, “association”, etc.), we retrieved PubMed entries, followed by the manual selection of relevant publications since the formation of a request is crucial for further analysis. The set of MeSH-terms was selected for them. We focused only on the experimental approaches; therefore, we excluded reviews, case reports, etc. The set of MeSH-terms was then used for the automated collection of relevant publications.

To identify the neoplastic diseases and processes that

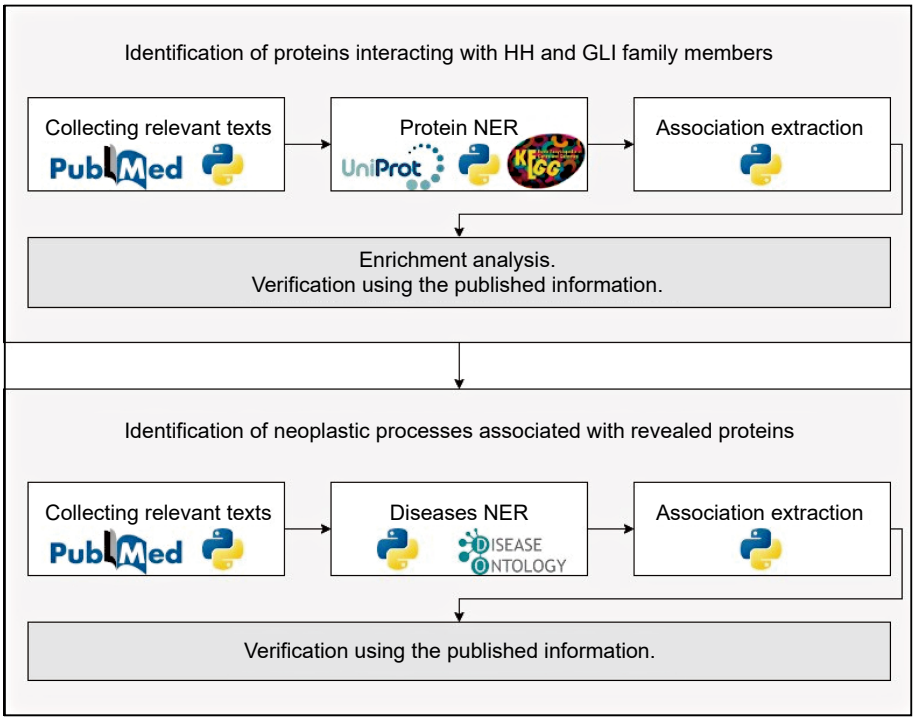

Fig. 1 General algorithm of information extraction on Hh proteins interactions and their associations with neoplastic processes from texts. NER is a named entity recognition.

are developing in various cases of aberrant Hh signaling, we carried out a text-based analysis of associations between neoplastic diseases and proteins either involved in the Hh pathway or interacting with its components. MeSH-terms were also used for the extraction of texts relevant to the Hh pathway function in the neoplasm formation. A set of texts (namely "Hh-onco") was compiled based on automatic queries to PubMed.

The texts for further analysis were collected using Python 3.10 script and Bio library (Entrez module).

2.2 Named entity recognition procedure

2.2.1 Recognition of protein and gene names

Proteins and genes names recognition was carried out using (1) a dictionary for proteins and genes of the Hh pathway and (2) the Named Entity Recognition (NER) procedure based on the Conditional Random Fields (CRF)[10–12] for those proteins and genes that are not included in the Hh pathway.

To recognize the names of proteins of the Hh pathway, we created a dictionary that includes all known names of the Hh pathway proteins. We did not use machine learning for NER of the Hh pathway components because we can enumerate all known names of these proteins and their synonyms that can be then found in a text. To create the dictionary, the names of proteins and genes of the Hh pathway were extracted from the KEGG database (https://www.genome.jp/kegg/, map 04340). Synonyms of proteins and genes included in the Hh pathway were collected using automated queries to the UniProt[13], ChEMBL[14], and KEGG[15] databases.

Names of proteins and genes that are not included in the Hh pathway were extracted from texts using a previously developed algorithm, which is based on the CRF[10–12] described in details earlier[13]. Briefly, the algorithm creates a sequence of elementary units (tokens) from texts. This process is called tokenization. Tokenization is an essential process for obtaining a high accuracy in the names recognition. We used tokenization, which is based on the usage of punctuation (dots, commas, semicolons, and spaces) that allowed achieving the best performance of the named entity recognition. Then, the sequence of tokens is transformed into text features, such as last and first symbols, belonging to stop-words, etc. We selected the features that provide for the highest accuracy in the names recognition. The set of features is presented in the Electronic Supplementary Materials (ESM) of the online version of this article (see Table S1).

Since several different synonyms of protein and gene

names may occur within the same text, we used requests to the UniProt database in order to obtain unique identifiers for each extracted name of a protein or a gene. The requests were carried out using the Python 3.10 script.

2.2.2 Recognition of diseases and disoders

Identification of neoplastic diseases that could be associated with the Hh pathway proteins and interacting with them was carried out using the dictionary-based method.

The dictionary of diseases and disoders was created based on the terms in the human Disease Ontology (DO)[16]. This ontology is a hierarchically arranged list of diseases including an additional description of a pathology, phenotypic characteristics and a list of synonyms. We chose the "disease of cellular proliferation" branch for the recognition of neoplastic diseases and their associated pathological processes. Entities belonging to this branch were extracted with their synonyms that are semantically associated with their "descendants". The named entity recognition was performed based on the direct comparison of strings in a lower register except for the abbreviations that were compared as they are in a higher register.

2.2.3 Recognition of miRNA names

To extract names of miRNAs, we used regular expressions since their names are used in the text in the common form.

To create a set of regular expressions, we analyzed a set of texts describing various investigations related to miRNAs functions, including reviews. Different forms of miRNA names were manually extracted and then grouped by their form of mention. Based on these groups, regular expressions were created. For example, "miR-182" and "miR-149" match the regular expression "miR-$\backslash d+\backslash b$" (Python "re" library syntax) which means that after the keyword "miR" we expect a hyphen character and one or more digits before the name ends.

2.3 Extraction of associations

2.3.1 Identification of the Hh pathway proteins interactors

To retrieve the associations between proteins, genes, and genes regulation by the proteins, we automatically identified the pairs of entities considering two points: (1) co-occurrence within the same sentence, and (2) the existence of a semantic relationship that can be expressed by one of the pattern phrases as we described previously[14]. Here, we consider a pattern phrase to be the one that specifies the type of interaction. Typically, these parts of texts are represented by verbs in various forms with a preposition, for example, "binds to" or "is regulated by". To collect pattern phrases, we analyzed one hundred texts related to the investigation of interactions between the proteins of the Hh pathway and other proteins and genes ("Hh interactions" corpus). This step was carried out automatically in part. For further manual analysis, we selected only the sentences that had two or more protein/gene named entities recognized automatically. Typically, the sequence of interacting objects in the text is specified by particular pattern phrases. Therefore, in most cases it was possible to identify the direction of interaction. For example, the aforementioned pattern phrase "is regulated by" has collocations with the name of a protein or a gene before and after it. In this example, the name after the phrase is the subject of interaction and the name before the phrase is the object of interaction. Analyzing the set of pattern phrases for extracting associations, we noticed that some pattern phrases varied from text to text, for instance, by verb forms and corresponding prepositions. So, we artificially expanded the list of phrases by changing the main verb form. For example, if the pattern phrase "regulates" was found in the texts, we transformed it into the forms, such as "regulate", "regulating", "is regulating", etc., taking into account the order of the named entities in the context and the possible direction of interaction. So, our approach for the relation extraction is based on the set of rules that consider pattern phrases, the order of the named entities, and the direction of interaction in some particular cases. A full list of the pattern phrases used is provided in Table S2 in the ESM.

In total, the list of 122 pattern phrases was created. Among them 54 phrases allow to identify the direction of interactions.

Since several different synonyms may occur within the same text, the extracted associations can be duplicated. We filtered out the duplicated associations based on the unique identifiers provided in UniProt and obtained for each recognized protein or gene name as described earlier.

The interaction networks were visualized using the XGMML format in the Cytoscape application[17]. We used UniProt entry names as the names of the nodes.

2.3.2 Identification of neoplastic diseases related to Hh pathway proteins and their interactors

The workflow displaying full process for extraction of knowledge on neoplastic diseases regulation by Hh pathway proteins and their interactors is provided in Fig. 2. Texts of the "Hh-onco" corpus were split into sentences. The named entity recognition of neoplastic diseases was performed as previously described. Names of the Hh pathway proteins were extracted using dictionary. We used a CRF approach for the named entity recognition of genes and proteins that were not included in the Hh pathway. After the NER was performed, we unified the extracted entities by linking them to a UniProt identifier through the automated queries (Python 3.10). We performed a two-step filtering of the obtained proteins, excluding (1) the Hh pathway proteins in order to avoid duplicates, since these proteins had already been extracted by a dictionary-based approach and (2) proteins that do not interact with or participate in the regulation of the Hh pathway according to results of our approach described above (please see Section 2.3.1). Filtering was performed by comparing the UniProt identifiers. Associations were identified based on the co-occurrence of a protein and a neoplastic disease name in the same sentence.

For the most common protein-disease associations extracted from texts, we prepared visualizations of the interaction networks using the Cytoscape application[16]. Unification of the extracted protein names was performed through the previously described queries to UniProt.

2.3.3 Identification of miRNAs involved in cancer development and regulation

Since small non-coding RNAs are known to be involved in gene regulation, we performed a brief analysis of the associations between such molecules and cancer types[18].

To do so, we extracted a set of article abstracts from PubMed relevant to consideration of miRNAs as potential regulators of neoplastic processes. As for the protein-cancer associations extraction processes described above, we used MeSH-terms to define the relevance of publications. We included the MeSH-terms "MicroRNAs" in the request.

To extract miRNAs names, we used regular expressions and to extract neoplastic disease names, we used a dictionary as described above.

Associations were extracted based on the co-occurrence of both terms – a neoplastic disease and a miRNA – in the same text.

All the materials used for the information retrieval from text using described approach are available at: https://github.com/nad-smol/Materials.git.

2.4 Analysis of the text-mining approach results

2.4.1 Enrichment analysis

The lists of 49 proteins from the Hh pathway and 1285 interacting proteins obtained by text-mining were merged to perform a pathway enrichment analysis. The UniProt accession numbers were converted to Entrez gene identifiers using the UniProt ID mapping tool (https://www.uniprot.org/id-mapping). Enrichment analysis was performed using the clusterProfiler R package[19] for both the KEGG pathways[12] and the Gene Ontology Biological Processes (GO BPs). The pathways with at least two genes from the list and an adjusted p-value less than 0.05 as well as GO BPs with an adjusted p-value less than 10^{-5} were selected for further analysis.

The revealed KEGG pathways were grouped according to the pathway classification available on the KEGG pathway database page[12]. The identified biological processes were clustered according to the GO terms directed acyclic graph using a hierarchical classification with the average method and distances between the terms calculated by the method of Wang et al.[20] implemented in the rrvgo R package. The

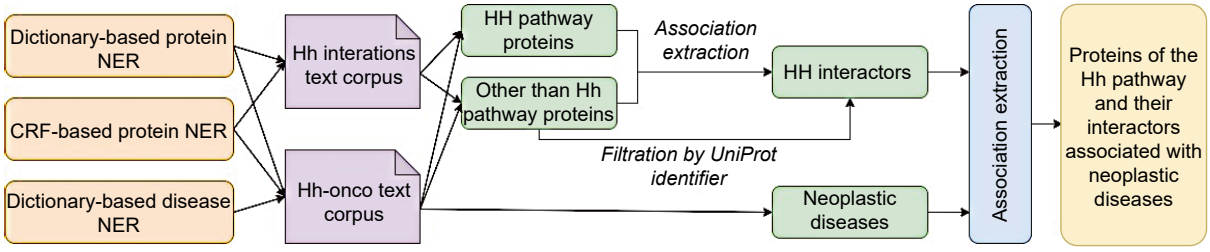

Fig. 2 Workflow for extracting associations between Hh proteins and their interacting proteins involved in neoplastic processes.

particular clusters were found using a hybrid method of the adaptive branch pruning of hierarchical clustering dendrograms approach implemented in the dynamicTreeCut R package.

2.4.2 Evaluation of the text-mining results according to the known associations between miRNAs and neoplastic diseases

To reveal miRNAs regulating the expression of the studied proteins, we performed a similar enrichment analysis using the experimental data on miRNA targets from miRTarBase[21]. We selected miRNAs that potentially regulate the expression of at least 2 out of the genes encoding 49 Hh pathway proteins and their 1285 interactors and the *p*-value to less than 0.05.

2.4.3 Comparison with proteins that were shown to be associated with neoplastic diseases according to experimental data

We used the OncoDB database[22, 23] for comparison of the results obtained using our method with the experimental data.

OncoDB provides analysis of the experimental data on changes in transcription of the genes in different Tumors included in the Cancer Genome Atlas program (TCGA)[24]. We assume that the change in gene expression in tumor compared to normal tissues indicates a close relationship between this gene and the pathology. Data on differential gene expression includes (1) cancer type designated as TCGA abbreviation, (2) NCBI gene ID, (3) adjusted *p*-value, (4) expression level in cancer and normal cell sample, (5) log2 fold change, (6) *p*-value, and (7) gene symbol.

In order to determine the list of genes associated with tumor processes, we used two criteria based on the data presented in OncoDB: (1) *p*-adjusted value should be less than 0.05, and (2) log2 fold change should be more than 0.5. Since our data obtained from texts included the human Disease Ontology IDentifier (DOID) for neoplastic diseases and the UniProt identifier for proteins, which are not the same as OncoDB data, we first normalized them in order to compare.

UniProt identifiers were compared with NCBI gene ID using the automated queries. OncoDB cancer types were linked to the human DO named entities using a direct comparison between the strings. If the name of a neoplastic disease belonging to the OncoDB cancer type completely coincided with one of the human DO entries, this type of cancer was assigned to a corresponding DOID. Otherwise, we performed the comparison manually. It is worth noting that most of the cancer types were matched automatically. The matching between the OncoDB cancer types and the human DO classification is provided in Table S3 in the ESM.

After normalization of identifiers, we compared the molecules extracted from texts, both included in the Hh pathway and interacting with its components with differentially expressed genes available in OncoDB.

2.4.4 Evaluation of the interactions between proteins revealed by text-mining using experimental data on protein-protein interactions

We performed network analysis to statistically evaluate whether the 1285 proteins found using our text-mining approach have significant interactions in the Protein-Protein Interaction (PPI) network with 49 Hh-proteins. We downloaded human PPIs from HIPPIE database (http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/) because it is the comprehensive databases containing more than 800 000 interactions between more than 19 000 human proteins. We downloaded only PPIs marked as "high-confidence" for further analysis. To calculate various characteristics of PPI networks, igraph R package was used. Before the analysis, we identified the largest connected network from HIPPIE data and removed self-loops and duplicated edges. It resulted in a PPI network with 14 515 nodes and 135 235 edges.

2.4.5 Validation of the obtained results by a literature search and analysis

In order to assess the quality of the study results, we conducted a review of the information published in the literature.

Our work covers a variety of associations of proteins with neoplastic diseases, including direct involvement in cascades of reactions and intracellular signal transduction, observations of changes in gene expression or protein concentration, usage as biomarkers of neoplastic diseases, and influence of the gene and genomic mutations on the pathology. Therefore, analyzing the literature, we aimed at classifying the extracted associations into the five indicated groups.

In the framework of the analysis, we also paid attention to the malignancy of the tumor. For cancer types, we provide survival statistics according to NCI's Surveillance, Epidemiology, and End Results Program (SEER)[25].

3 Result and Discussion

Although the relationships between various proteins and genes and neoplastic diseases development are studied deeply, our algorithm has its own place since it is based on the extraction of information from texts combined with a bioinformatics approach (enrichment analysis).

Nevertheless, great efforts have been made in order to collect and analyze existing knowledge on various cancers etiology, pathogenesis, consequences, diagnostics, and therapy.

For example, in the work of Li et al.[26], the role of proteins and genes in the development of ovarian cancer resistance to platinum-containing medications was discovered. The recognition of protein and gene named entities was performed based on a previously developed by Burr Settles algorithm that uses CRF[27]. Names of drugs were identified by the list of the same keywords as ovarian cancer subtype entities. The associations between extracted genes and the so-called platinum resistance were established by the presence of listed keywords and protein named entities in the relevant texts. The authors also performed an analysis of pathways that could be related to the resistance and revealed that various ovarian cancer subtypes could be associated with different mechanisms.

In the study by Min et al.[28], an Edge-group Sparse Principal Component Analysis (ESPCA) model was developed for the investigation of high-dimensional gene expression data. The validation of the approach was performed using two artificial datasets and two real biological datasets (TCGA pan-cancer expression data and ENCODE expression dataset). The results of validation showed superior performance of ESPCA over sparce PCA for variable selection. ESPCA is able to find more biologically relevant genes and contributes to their biological interpretation.

In another work[29], regulatory genes of hepatocellular carcinoma metastasis were evaluated through literature mining. Phenotype keywords (such as “metastasis” and “adhesion”) were selected as the criteria for metastatic processes. Protein and genes named entity recognition was performed using the ABNER app[27], which also uses CRF. Only proteins and genes that co-occurred with the established keywords of a metastatic process in the same text were considered. The authors of this work identified the genes that most frequently occurred in the associations and also discussed the role of several pathways according to the performed analysis.

Many efforts have been made to find associations between not only proteins or genes with a disease but also with various biological objects. COVID-KOP represents the knowledge base that integrates the literature data on COVID-19-related associations as graphs[30]. The authors used the COVID-19 Open Research Dataset, or CORD-19, which is a corpus of full-text articles related to any field of COVID-19 research[31]. CORD-19 was tagged to the known ontologies of chemicals, diseases and conditions, proteins and genes, etc. Associations between the edges were extracted by biological objects co-occurrences within a text and a sentence.

In this work, in contrast to the earlier developed approaches, we performed a deep analysis using text- and data-mining techniques in order to consider the molecular mechanisms of the Hh signaling pathway regulation in the development of neoplastic diseases. Some drawbacks of our approach are associated with a limited ability to gain abstracts relevant to a particular disease or disorder, and a risk of missing some relationships. Nevertheless, the automated knowledge extraction from publications allows quick and efficient obtaining of the information necessary for research in a particular area of science, for example, the investigation of molecular mechanisms of neoplastic diseases. Moreover, using enrichment analysis and manual verification of the results we identified several signal pathways associated with the Hh pathway that can be essential for the development of neoplastic diseases.

3.1 Identification of proteins interacting with the proteins of the Hh pathway

Based on a manual analysis of relevant texts, we identified keywords to construct queries to PubMed, that are present as follows: “<Gli1/Gli2/Gli3/Hedgehod pathway >AND ((regulat*) OR (mechanism) OR (signaling) OR (pathway) OR (transduction)) AND NOT (<exceptions by publication type>)”. To select topics related to neoplastic diseases, MeSH-term “Neoplasms” was used. The search for relevant texts containing any information about the Hh pathway proteins was supposed to be performed using the inclusion of the “Hedgehog Proteins” term. In total, we collected 9948 publications relevant to the Hh pathway analysis.

3.1.1 Extraction of protein and gene names

We collected all names and synonyms of the Hh pathway proteins from UniProt, ChEMBL and the KEGG database to construct a dictionary for extraction of the named entities from raw texts. Full dictionary is presented in Table S4 in the ESM. We also included names and synonyms of Gli3 protein subunits since such information was available in UniProt and KEGG.

Extraction of other protein named entities was performed using CRF as a machine-learning method[12]. The recognition accuracy was evaluated using five-fold cross-validation and manual analysis of the external test set (100 randomly selected texts). The results of the accuracy evaluation are given in Table 1.

According to the Table 1, we achieved the reasonable accuracy of proteins and genes names recognition that is comparable with earlier published approaches[32–35].

3.1.2 Identification of associations between proteins and genes

Using the created rules for all the proteins of the Hh pathway, we identified almost 9000 associations, where 896 proteins were involved in interactions with Hh pathway proteins. After filtering out the duplicated associations, 2335 unique associations were identified. Also, we have extracted more than 6000 associations that include Gli proteins separately, 1166 of them were unique. In total, 326 proteins, except for members of the Gli family, were involved in these associations. A full list of extracted associations of the Hh pathway proteins with other human proteins is presented in Table S5 in the ESM.

We evaluated the accuracy of the associations extraction. To do this, we used the recently developed bioRED[36] corpus, which includes 600 abstracts and annotations of the names of various biological objects and associations between them. Over 2500 protein-protein interactions in this annotated corpus were used for accuracy estimation. Cases where an association was recognized using our approach and was actually present in the annotated bioRED corpus were considered a true positive result. As a false positive result, we considered cases when our approach extracted an association from a sentence that was not annotated by the authors in the same sentence of the corpus. A false negative result was the presence of an annotated association in a corpus sentence, was not recognized in the same sentence by our approach.

Table 1 Accuracy of protein named entity recognition by a developed CRF-based approach.

Validation method	Precision	Recall	F1-score
5-fold cross validation	0.87	0.84	0.85
Manual annotation of external text set	0.84	0.79	0.81

The extraction of associations using the developed rules-based approach, taking the protein entities pre-annotated by the bioRED authors as input, was performed with F1-score equal to 0.84 (precision is 0.78, recall is 0.91). Based on this, we believe that the developed rule-based method allows the extraction of associations with reasonable accuracy comparable to state-of-the-art approaches for associations extraction[37, 38].

3.1.3 Visualization of associations

We built graphs indicating the associations between proteins and/or genes regulation using the XGMML format of the Cytoscape application.

Figure 3 shows the interaction network for fifty most frequently occurred proteins in associations including Gli proteins as some key transcription factors of the Hh pathway.

The Sonic hedgehog protein (Shh) is the most frequently occurring protein in associations with Gli proteins (see Fig. 3). It is not surprising, as Gli proteins and Shh protein are the well-known participants in the Hedgehog signaling pathway[39–41]. Also, nodes of the signaling pathway, Smo, and Ptch1 were included in the list of proteins most frequently found in associations[42–44]. Transforming growth factor (Tgf)-beta was shown to induce the expression of Gli protein family members[45], and we revealed its association with Gli3 based on text analysis.

Interaction network for the most common associations with the Hh pathway proteins and other human proteins interacting with this signaling pathway is presented in Fig. 4.

Looking at Fig. 4, it can be observed that Wnt is one of the proteins frequently interacting with the Hh pathway. Wnt protein is the key node in Wnt signaling pathway that is involved in many developmental processes, in particular, in organ development[46, 47]. Bmp2 is also one of the most frequently occurring protein in the associations found. Bmp2 is a key protein in Tgf-beta signaling. Together with the Hh and Wnt signaling pathways, it is involved in bone development[47]. Many of the extracted proteins are included in signaling pathways. For instance, Myc is included in the aforementioned Tgf-beta and Wnt pathways, Fgf and Tgf-alpha are involved in the Mapk

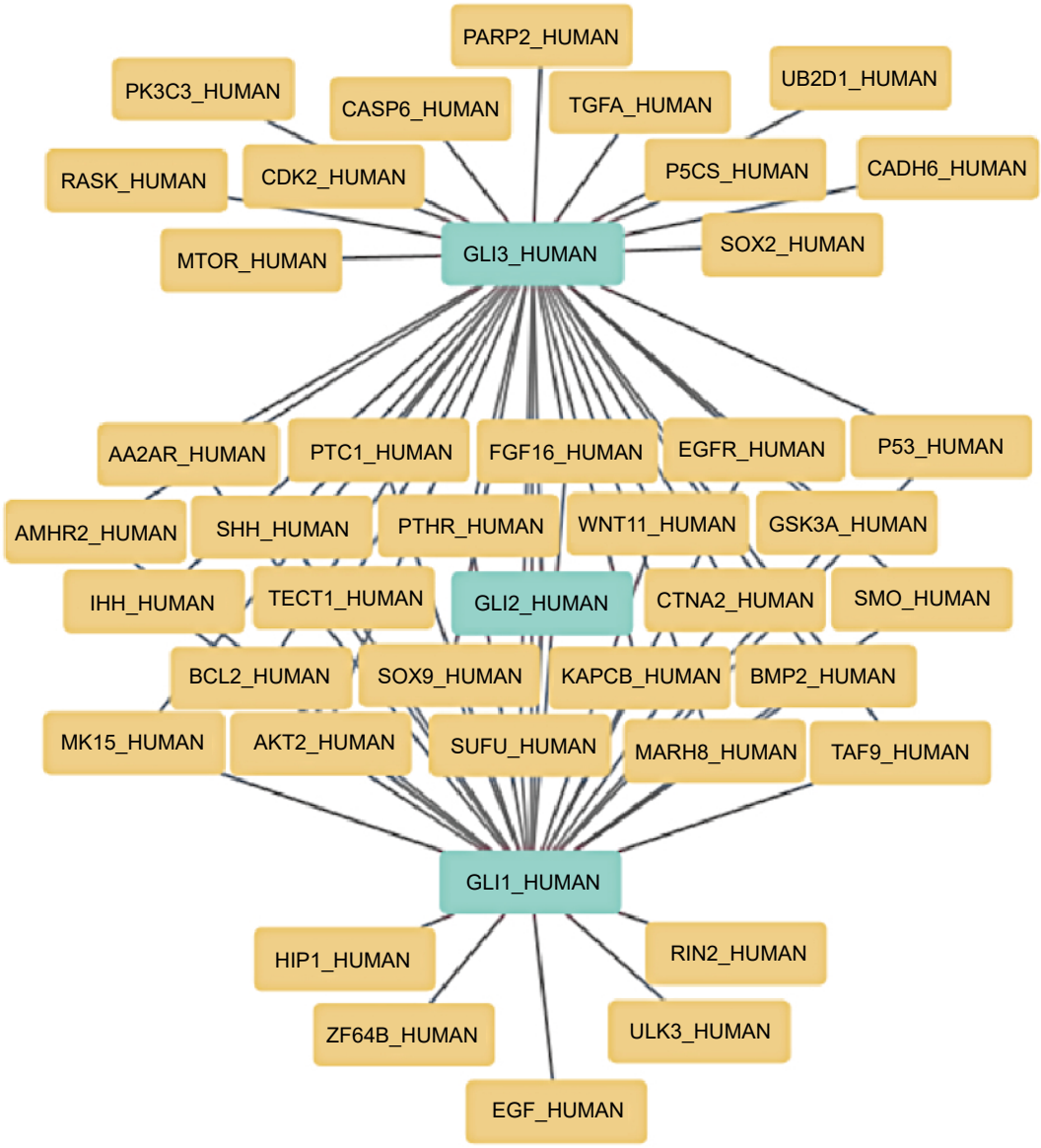

Fig. 3 Interaction network for the most frequently occurred key players of the Hh pathway (Gli proteins) within the considered collection of texts.

pathway, which, in turn, is associated with the Wnt pathway. Pi3k is included in Akt pathway having associations with the Mapk pathway. These observations lead us to the idea that the associations that were extracted from texts are connected with huge cascades of intracellular signaling in the human body.

3.2 Text-mining based search for protein associations with neoplastic diseases

For collecting "Hh-onco" corpus we used MeSH-based query that included the following terms: ("Neoplasms" [Mesh]) and ("Hedgehog Proteins" [Mesh]). Similar to "Hh interactions" text collection, reviews were excluded. In total, we were able to create a collection of 2280 texts relevant to the Hh pathways functioning in the neoplastic diseases and pathological processes.

To extract the relations between proteins involved in the identified interactions and pathological processes, we created a dictionary of neoplastic disease named entities (Disorders Named Entities, DNE) based on the human disease ontology and a CRF-based algorithm. The neoplastic disease dictionary included in total 2610 entities of diseases with 6466 names (approximately 3 synonyms per pathology).

As mentioned above (see Sections 2), the extraction of associations between proteins (genes) and neoplastic diseases and pathological processes was performed based on the co-occurrence of at least one protein and

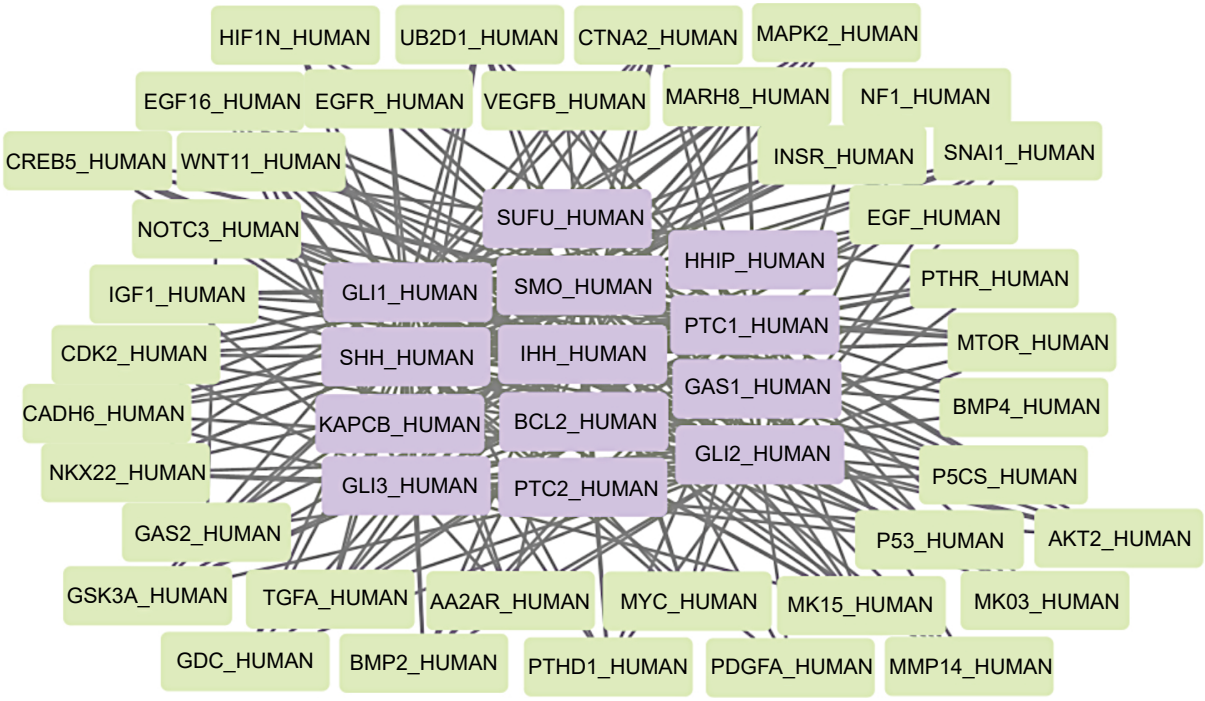

Fig. 4 Interaction network for the most frequently occurred associations of proteins with the Hh pathway. Purple nodes are the proteins of the Hh pathway, green nodes are not included in the Hh signalling pathway but interact with one of its proteins.

at least one disease name in the same sentence. In total, we extracted 3664 unique associations, and 281 of them include unique names of neoplastic diseases and pathological processes. We should note that among these 281 diseases mentioned there are also unspecified terms, for instance, "cancer". A full list of the extracted associations between the proteins of the Hh pathway, their interacting proteins, and neoplastic diseases and processes is provided in Table S6 in the ESM. We also provide a list of the unique extracted associations taking into account the frequency of each individual association in the corpus of texts (Table S7 in the ESM). Figure 5 provides a visualization of associations between the 20 most common neoplastic diseases and proteins found in the extracted associations.

3.3 Analysis of the obtained results

3.3.1 Enrichment analysis

3.3.1.1 Enrichment analysis for proteins associated with neoplastic diseases

To form a clearer picture of the extracted proteins' involvement in biological processes, we performed the enrichment analysis.

We merged the list of 49 proteins from the Hh pathway and with 1293 interacting proteins obtained by text-mining and converted UniProt accession numbers to 1334 NCBI Entrez gene identifiers. 1334 identifiers were used to obtain enriched GO Biological Processes (BPs), and the KEGG pathways. The revealed KEGG pathways are presented in Figs. 6 and 7, and the identified 26 clusters of GO BP terms are listed in Table S8 in the ESM.

Most of the identified GO BPs are related to processes of organ development and cell differentiation, proliferation and apoptosis, which is in agreement with the known role of the Hh pathway in the organism.

The revealed KEGG pathways can be divided into "normal" and "disease" pathways (see Figs. 6 and 7, respectively). The normal pathways obtained can be divided into two groups: (1) pathways associated with processes regulated by the Hh pathway, e.g., cell cycle, apoptosis, osteoclast differentiation, etc.; (2) pathways that contain genes and proteins, which either regulate or are regulated by the Hh pathway. The second group includes both pathways with well-known associations with the Hh pathway, e.g., Wnt, Tgf-beta, ErbB, Mapk, Ras signaling pathways[48, 49], and pathways whose relations with the Hh pathway are not well known, e.g., immune system pathways. We manually analyzed the positions of 1334 genes in pathway maps incuding direct interactions of Ptch1–2, Smo, and Gli1–3 with

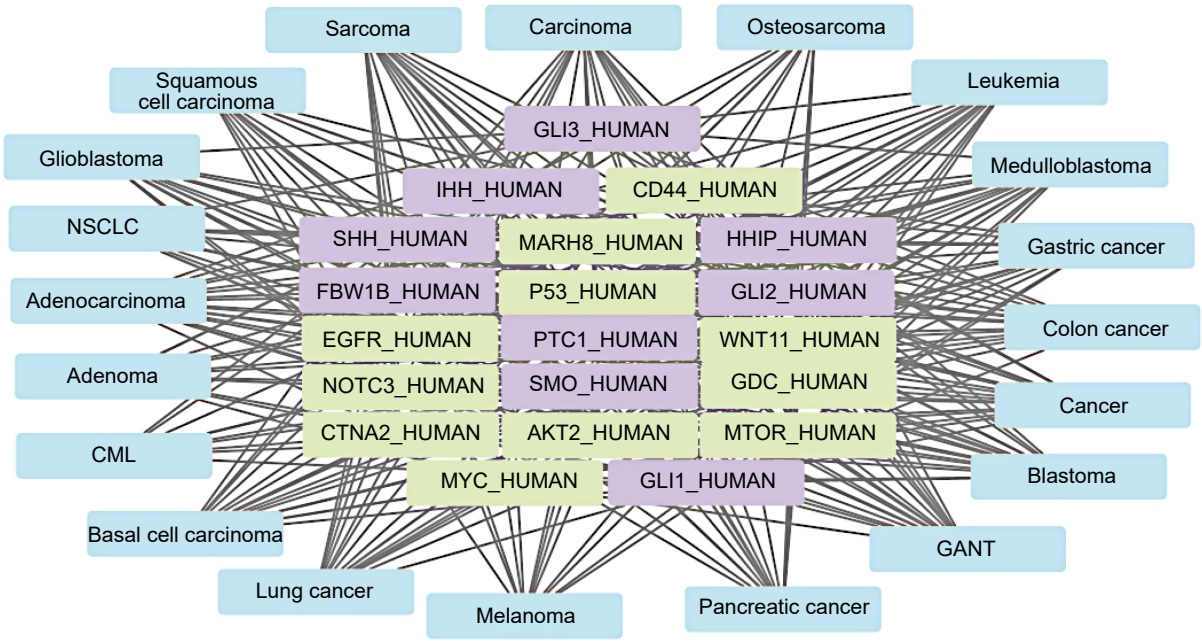

Fig. 5 Associations between the 20 most common neoplastic diseases and proteins. Blue nodes represent neoplastic diseases, purple are the proteins of the Hedgehog pathway, green nodes represent human proteins interacting with the Hedgehog pathway. Abbreviations for the names of neoplastic diseases are Gastrointestinal Autonomic Nerve Tumor (GANT), Chronic Myelogenous Leukemia (CML), and Non Small Cell Lung Cancer (NS CLC).

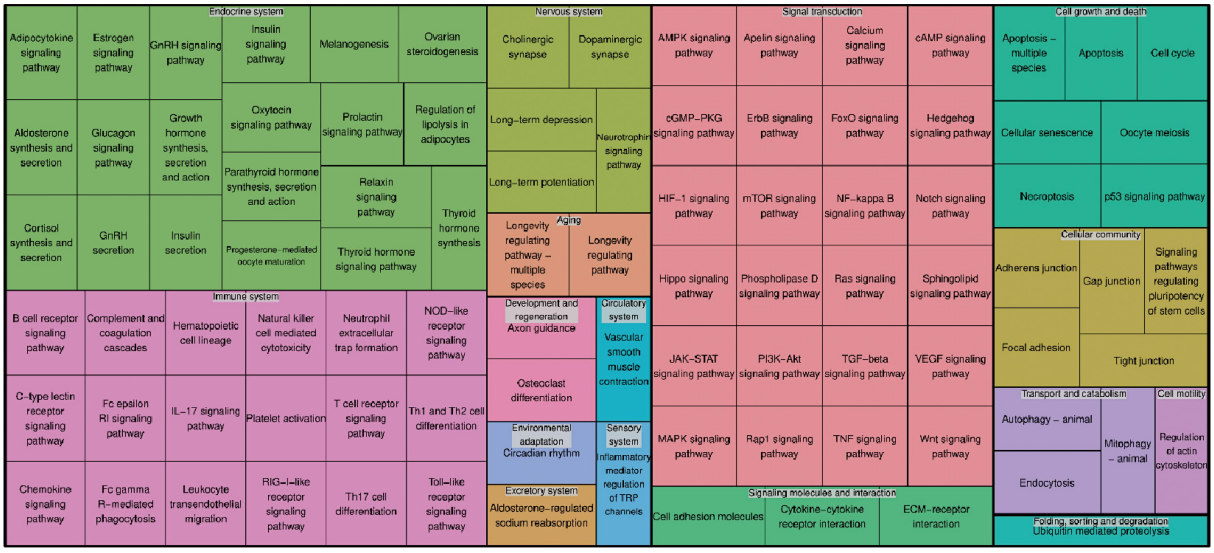

Fig. 6 KEGG normal pathways enriched by both proteins from the Hh pathway and proteins obtained by text-mining.

other proteins from the OmniPath database (https://omnipathdb.org) and some comprehensive reviews on Hh pathway regulation[2, 3, 48, 50–53]. We proposed that 1293 interacting proteins obtained by text-mining belong to several groups:

(1) Proteins directly interacting with proteins of the Hh pathway including kinases, acetyl- and methyltransferases, and ubiquitin-protein ligases. The corresponding proteins modify amino acids of Smo and Gli1-3 proteins that changes their activity and degradation rate;

(2) Transcription factors that form complexes with Gli1-3 and increase their transcription activity; Histone acetyltransferases and other proteins that increase transcription activity of Gli1-3;

(3) Transcription factors that increase the transcription of Hh pathway components;

(4) Target genes of Gli1-3 transcription factors;

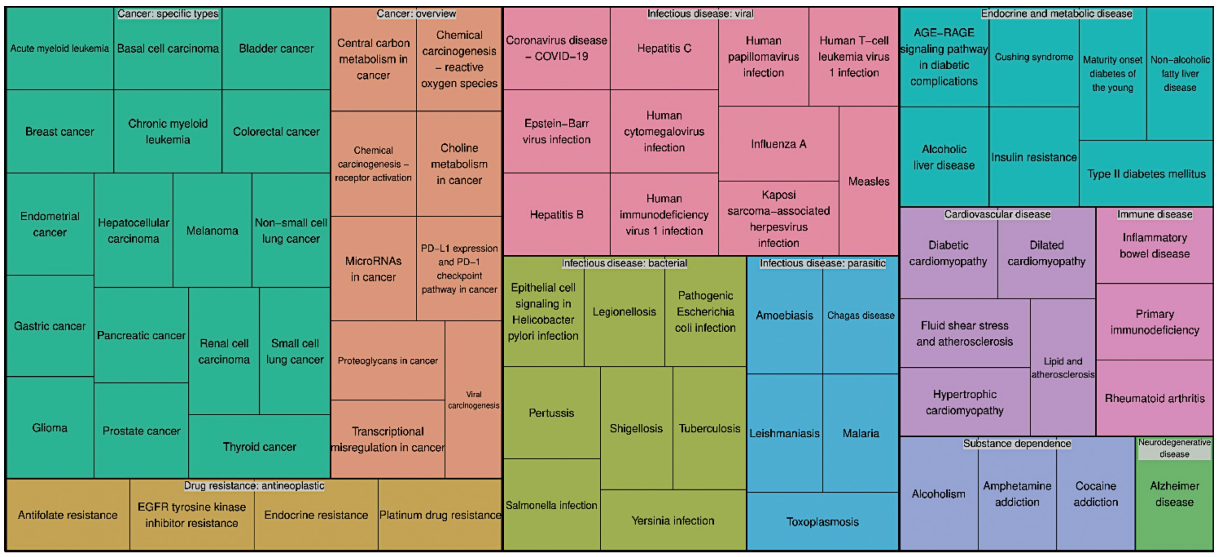

Fig. 7 KEGG "disease" pathways enriched by both proteins from the Hh pathway and proteins obtained by text-mining.

(5) Well known proteins involved in non-canonical activation of Gli1-3, which do not directly interact with them;

(6) Most of the 1293 interacting proteins that seem to activate/inhibit proteins (1–5) through interaction cascades.

The lists of proteins (1–5) are presented in Table S9 in the ESM. We considered these proteins as pathway endpoints, which immediately regulate the Hh pathway.

To show the relationships between groups of "endpoint" proteins and enriched signaling pathways, we created a bipartite graph containing only proteins belonging to at least two pathways (see Fig. 8). Figure 8 demonstrates that very different pathways regulate the Hh pathway activity through the same endpoints. The kinases Erk1 (Mapk3), Erk2 (Mapk1), and Akt1 are part of more than 30 pathways. It was shown previously that Mek1/2–Erk1/2 kinases activate Gli1 transcriptional activity[51]. The precise mechanism is not known yet, but it is possible that Erk kinases directly phosphorylate Gli1. Akt kinases also up-regulate Gli1 activity through various mechanisms: decreasing the Gsk3b activity, increasing the S6K1 kinase activity, which, in turn, phosphorylates and activates Gli1, and potentially directly phosphorylating Gli1[50, 54].

Along with the well-known pathways affecting the Hh signaling, we identified pathways related to both innate and adaptive components of the immune system (Fig. 6). For instance, the T cell signaling pathway is enriched by many of the Hh-interacting proteins revealed by text-mining (Fig. 8). It is currently known that the Hh pathway regulates proliferation and differentiation of T cell in the thymus[3]. It also influences various processes in mature peripheral T lymphocytes, such as negative regulation of Tcr-induced activation and proliferation[52], positive regulation of Il-2 and Ifn-gamma production by CD4 T-cells[53], and promotion of actin remodeling required for centrosome polarization and granule release by cytotoxic Cd8 T cells[55]. The observed immune modulation may be explained by the fact that transcriptional factors Gli1-3 altered the transcription of various gene groups, leading to modulation of T cell functions[56, 57]. In turn, the activation of Hh signaling by activation of Tcr may be a consequence of pathways overlapping at different levels: at the level of kinases, e.g., Erk1 (Mapk3), Erk2 (Mapk1), Gsk3b, and Akt1, or transcriptional factors, e.g., Ap-1 and NF-kB, which increase transcription of Hh components[54].

The Toll-like receptor (Tlr) signaling pathway is also known to be involved in negative regulation of the Hh pathway[58–60]. For instance, activation of the Tlr3 receptor in neural stem/precursor cells leads to a decrease in their proliferation and an increase in apoptosis by inhibiting Sonic Hedgehog signaling[56]. Activation of Tlr4 accelerates the senescence of placental mesenchymal stem cells via suppressing the Hh pathway that is associated with the development of preeclampsia[59].

The "disease" pathways are presented in Fig. 7. The

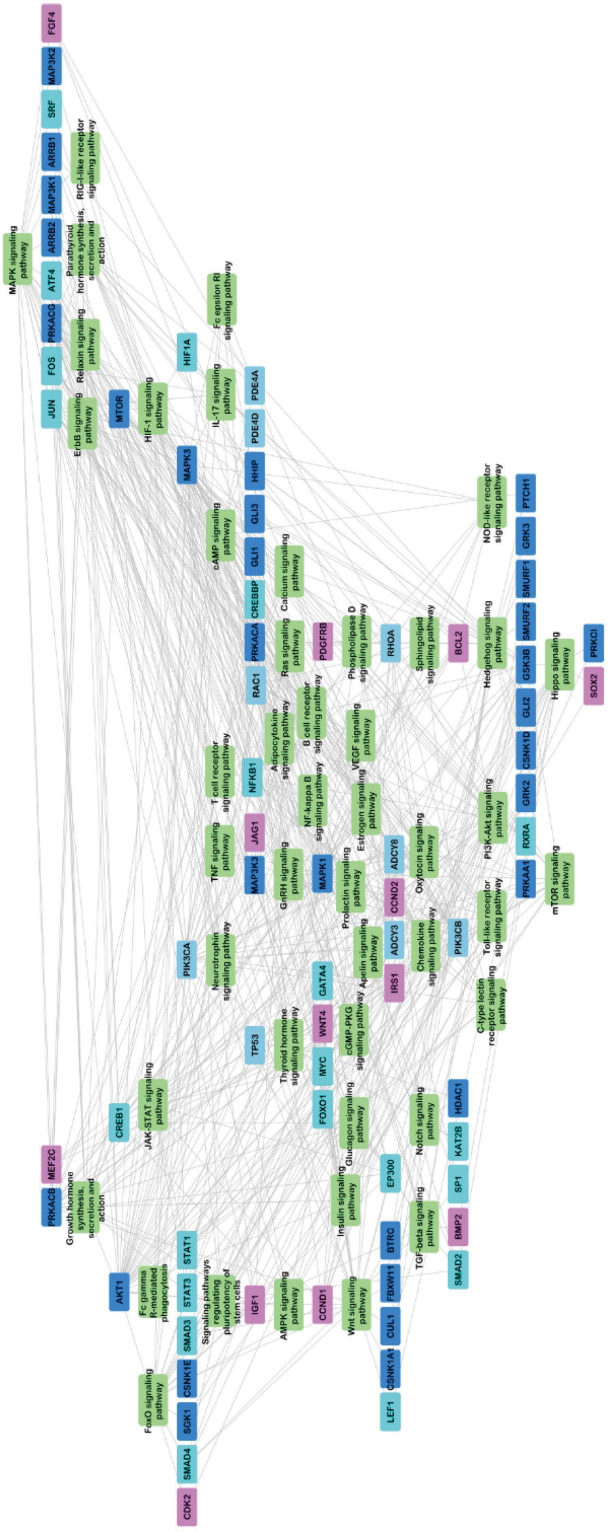

Fig. 8 Network of enriched pathways and proteins, which have evidence of either their involvement in regulation of the Hh pathway activity or in transcriptional regulation by Gli1-3. The dark blue nodes represent proteins of the Hh pathway and proteins that interact with them directly. The light blue nodes represent the proteins involved in non-canonical activation of Gli1-3 but not interacting with them directly. The cyan nodes represent transcription factors and other proteins that either increase the transcription of the Hh pathway genes or increase Gli1-3 transcription activity. Purple nodes represent target genes of Gli1-3 transcription factors.

role of the Hh pathway in cancer and developmental disorders is well known[54, 61–64], but involvement of the Hh pathway in infectious diseases is less studied[64–66].

Most evidence in the literature is related to the role of the Hh pathway in viral infections. We identified several virus-related pathways that were enriched by 1334 proteins (Fig. 7). Hepatitis B Virus (HBV), Hepatitis C Virus (HCV), Epstein-Barr Virus (EBV), Human PapillomaVirus (HPV), Human Immunodeficiency Virus (HIV), Human T-cell Leukemia Virus type I (HTLV-1), Influenza A Virus (IAV), and Kaposi's Sarcoma-associated HerpesVirus (KSHV) are known to induce pathological processes by regulating the expression, protein stability, and subcellular localization of Hh pathway components[64, 65]. For instance, the hepatitis B pathway has the highest number of nodes shared with Hh-interacting proteins (Fig. 7). The HBV and HCV are the main causes of HepatoCellular Carcinoma (HCC), since about 80% of HCC cases are related to HBV and HCV infections. The chronic infections of HBV and HCV are associated with an increase in expression of Hh pathway components and, in turn, with an increase in HCC tumor size, liver fibrosis, maintenance of cancer stem cells, initiation of epithelial-mesenchymal transition, metastasis, and drug resistance[59]. It is known that the HBV-encoded X protein accounts for HBV-induced activation of Hh signaling through promoting the protein stability and nuclear translocation of Gli1/2 by a direct protein-protein interaction, concomitant with increased transcriptional activity of Gli1/2[66, 67].

Bacterial infections, such as those induced by Salmonella enteritidis, Escherichia coli, and Helicobacter pylori are also known to use the Hh pathway to control the infection progression and the infected cell microenvironment[64]. Currently, there is no evidence in the literature of other bacterial and parasitic diseases having relationships with the Hh pathway. Since these pathways contain many of the 1334 Hh-related proteins, corresponding relationships can be revealed in the future.

3.3.1.2 Enrichment analysis for miRNAs associated with neoplastic diseases

Since miRNAs play an important role in regulation of the expression of protein-coding genes, we identified miRNAs that are associated with the Hh pathway. To reveal miRNAs regulating the expression of 1334 proteins, we performed the enrichment analysis using experimental data on miRNA targets from miRTarBase[21]. As a result, we identified 139 human miRNAs which may regulate Hh pathway components (Table S10 in the ESM). The 16 miRNAs enriched by at least 50 out of 1334 genes, along with their interactions with genes having evidence of involvement in regulation of the Hh pathway, are shown in Fig. 9.

Some of the 16 miRNAs are involved in the regulation of Hh signaling and cancer development. For instance, hsa-miR-9 targets Ptch1 mRNA in glioblastoma and is involved in temozolomide resistance development. The inhibition of Ptch1 function by hsa-miR-9 also leads to Hh signaling pathway activation and neurological function recovery after brain trauma[68]. miR-26a decreases Fam98a, Shh, Smo, and Gli1 expression levels and suppresses cell proliferation, clone formation ability and metastasis of breast cancer[69]. miR-125b is one of the suppressors of the Hh pathway and is known to be down-regulated in medulloblastoma[70]. miR-125b might also be protective against liver fibrosis via inhibiting Gli3[71]. miR-375 inhibits the expression of Rac1, which, in turn, is involved in non-canonical activation of Glis. Expression of miR-375 is significantly down-regulated in liver fibrosis tissues, its overexpression inhibits the hepatic stellate cell viability and epithelial-mesenchymal transition and alleviates liver fibrosis[72]. miR-203a-3p inhibits the expression of Gli1 in Barrett's esophagus cells associated with a cell cycle arrest[73].

3.3.2 Comparison of obtained results with available experimental data

Comparison with experimental data available in the OncoDB[23] database has several features that are worth discussing. First of all, OncoDB uses a TCGA[24] classification to define cancer types. Since the TCGA initiative is aimed at studying the most socially significant cancer types it is obvious to conclude that not all the entities presented in Human Disease Ontology find a correlation with TCGA. Further, to simplify the narration, we will mention the types of cancer referring to OncoDB while discussing experimental data, to avoid the repetition of the term "TCGA". The human disease ontology that was used for extracting the names of diseases in our approach has far more levels of classification and includes a detailed classification of neoplastic diseases. Additionally, not all cancer types included in TCGA

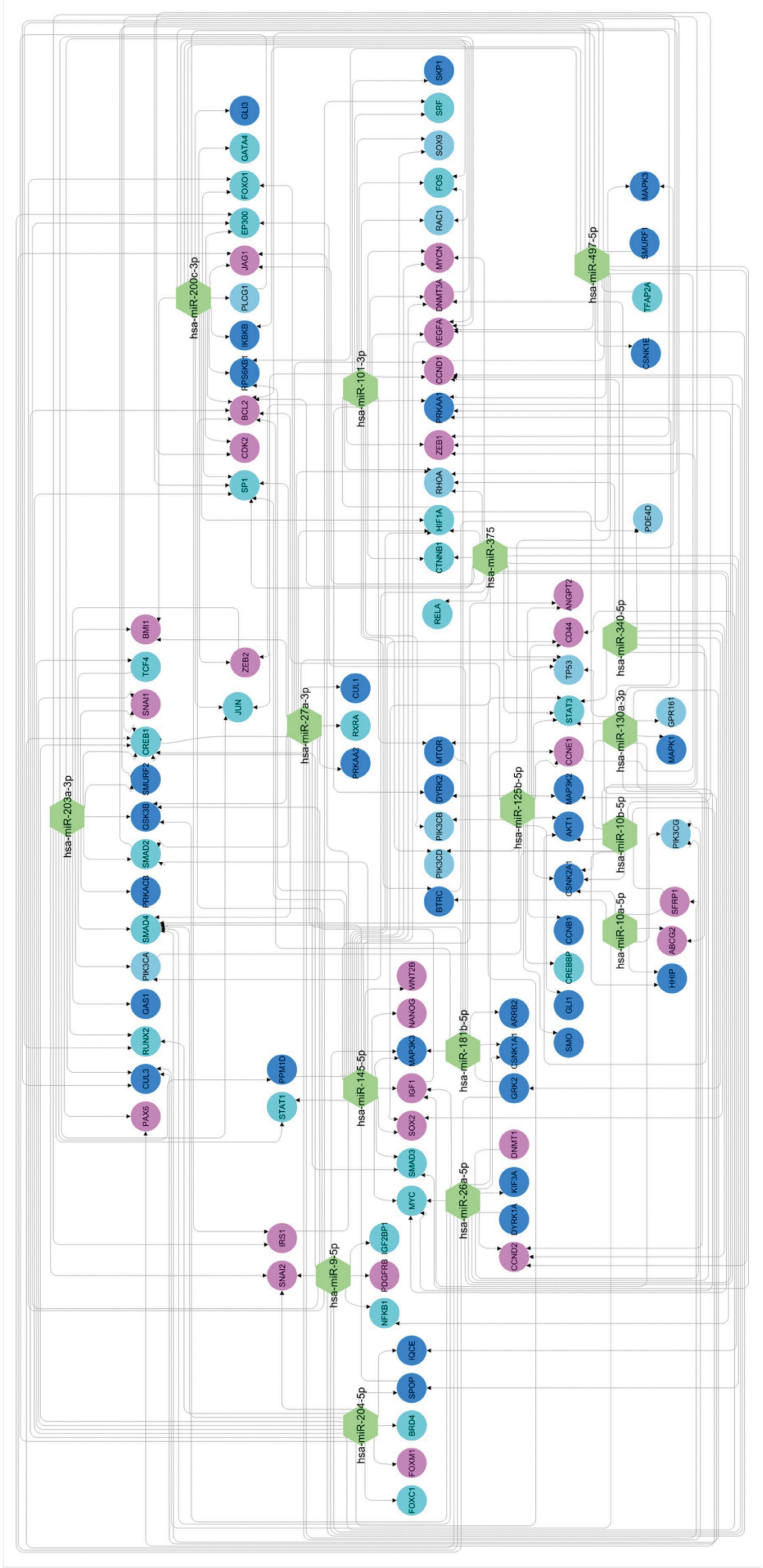

Fig. 9 Interactions between genes, which have evidence of involvement in regulation of the Hh pathway, and the top 16 miRNAs enriched by genes revealed by text-mining. The dark blue nodes represent the core proteins of the Hh pathway and proteins directly interacting with them. The light blue nodes represent proteins involved in non-canonical activation of Gli1-3 but not interacting with them directly. The cyan nodes represent transcription factors and other proteins that either increase the transcription of Hh pathway genes or increase the Gli1-3 transcription activity.

classification have data on differential expression in OncoDB. To this end, we only provide a comparison between our results and the OncoDB data, relying only on the matched types of cancer. In Table 2, we provide a number of mathches in proteins related to the particular neoplastic diseases according to the OncoDB cancer types. We also performed an analysis of experimental data based on type of genes (either belonging to the Hh pathway or not) and direction of changes in the expression (over- or under-expression). All the calculations carried out to analyze the experimental data and compare the results of text mining with them, we took out in the Supplementary Materials (Table S11 in the ESM).

In order to test the assumption about the existence of a relationship between the number of matches and the number of differentially expressed genes according to experimental data, we calculated the chi-square test of independence of variables for the columns "Total (matches)" and "Total (OncoDB)" of the Table 2. According to the obtained p-value, we can assume that these two variables are dependent ($p < 0.05$).

With the use of our approach 1334 unique genes and proteins, including Hh pathway participants, were found to be associated with neoplastic diseases. We were eager to compare this list of proteins with the experimental data in order to find other cancer types related to them in addition to those published in the literature. While comparing NCBI Entrez gene identifiers with those from UniProt, we were not able to unambiguously match the identifiers for 17 proteins, so in total we obtained the list of 1317 unique proteins, which was further used for matching OncoDB data with proteins extracted from texts. However, 1281 out of 1317 (97%) of extracted proteins were differentially expressed in at least one type of cancer according to OncoDB. The remaining 36 proteins were not found in OncoDB either due to the false positive results of our approach or due to the absence of corresponding data in OncoDB, which only focuses on a limited group of cancers.

3.3.3 Evaluation of the interactions between proteins revealed by text-mining using experimental data on protein-protein interactions

We proposed that if text-mining – derived 1293

Table 2 Comparison of data on associations between the Hh pathway proteins, their interactors, and neoplastic diseases obtained by text-mining with the experimental data on differential expression of the genes available in OncoDB.

Cancer type	Statistics on differentially expressed genes in OncoDB			Statistics on matching text-mining results with OncoDB data		
	Total (OncoDB)	Belong to the Hh pathway	Do not belong to the Hh pathway	Total (macthes)	Belong to the Hh pathway	Interacting with the Hh pathway
Bladder urothelial carcinoma (BLCA)	5570	21	5549	451	20	431
Breast invasive carcinoma (BRCA)	6111	21	6090	515	17	498
Cholangiocarcinoma (CHOL)	11 679	31	11 648	817	25	792
Colon adenocarcinoma (COAD)	6656	15	6641	510	13	497
Glioblastoma multiforme (GBM)	7869	26	7843	584	22	562
Head and neck squamous cell carcinoma (HNSC)	6435	17	6418	511	14	497
Kidney chromophobe (KICH)	7677	27	7650	553	24	529
Kidney renal clear cell carcinoma (KIRC)	7067	16	7051	521	16	505
Kidney renal papillary cell carcinoma (KIRP)	6364	16	6348	482	14	468
Liver hepatocellular carcinoma (LIHC)	7642	20	7622	530	15	515
Lung adenocarcinoma (LUAD)	6411	14	6397	478	13	465
Lung squamous cell carcinoma (LUSC)	8384	16	8368	611	15	596
Pancreatic adenocarcinoma (PAAD)	1062	4	1058	92	4	88
Pheochromocytoma and paraganglioma (PCPG)	5120	13	5107	356	10	346
Prostate adenocarcinoma (PRAD)	3710	10	3700	317	10	307
Rectum adenocarcinoma (READ)	6153	15	6138	459	13	446
Sarcoma (SARC)	2512	9	2503	207	9	198
Thyroid carcinoma (THCA)	4880	19	4861	364	17	347
Thymoma (THYM)	1475	2	1473	92	2	90
Uterine corpus endometrial carcinoma (UCEC)	7209	27	7182	582	25	557

proteins are not random and regulate 49 Hh-proteins, they will form dense connected PPI subnetwork. There were 41 out of 49 Hh-proteins and 1136 out of 1293 proteins in the PPI network constructed using data from HIPPIE database (1177 out of 1334 proteins in total) (see Section 2.4.5). We created a network from 1177 proteins and found that it contained 234 connected subnetworks. The largest connected subnetwork contained 936 nodes including 39 out of 41 Hh-proteins. We calculated connectivity index as the size of the largest connected subnetwork divided on the total size of network: 936/1177 = 0.795. Additionally, we found that the subnetwork of 936 proteins had 3633 edges. To define density of the subnetwork, we calculated the average number of 936 proteins interactors –7.8 and the average number of 39 Hh-proteins' interactors –10.7.

To estimate the significance of connectivity index and average numbers of protein interactors, we repeated the same analysis 100 thousand times using 41 Hh-proteins and 1136 random proteins from whole PPI network. The mean and maximal values of connectivity index on random data were 0.5 and 0.61, the mean and maximal average numbers of protein interactors were 3.2 and 4.4, the mean and maximal average numbers of Hh-protein interactors were 5.5 and 7.6. Thus, the obtained subnetwork characteristics are significantly higher than those obtained in randomization experiments.

3.4 Validation of the obtained results by literature search and analysis

3.4.1 Role of the extracted proteins in neoplastic diseases pathogenesis

The results of enrichment analysis were completed by the identification of proteins that are associated with neoplastic diseases and processes according to the literature data. We provide literature data for some of the extracted associations as an example of the results of our approach, taking into account the investigated role of a single protein in neoplastic processes (potential target of antitumor therapy, biomarkers, regulators, etc.).

Enrichment analysis and direct parsing of GO functions of the extracted proteins revealed that most of them belong to signaling pathways involved in the regulation of cell development and differentiation. Based on this fact we assumed that the extracted proteins might up- or downregulate processes in tumor cells.

In particular, the transcription factor 4 (Tcf4) (UniProt: P15884) identified by text-mining analysis is not involved in the Hh signaling. But the Hh signaling is associated with the Wnt signaling that includes Tcf4. A possible role of Tcf4 is that it belongs to the Wnt signaling pathway, which up-regulates the key players of the Hh pathway[74]. Tcf4 has both differentiation and transcription regulation functions. It was shown that the Wnt signaling pathway and, specifically, Tcf4 are involved in colon cancer, CML, and medulloblastoma[75–77]. SEER statistics do not provide a 5-year survival for medulloblastoma; but colon cancer and CML have a favorable prognosis in a relative majority of cases (5-year survival rates of 64 and 71%, respectively). Analysis of the transcription profiles of CD34+ cells in healthy people and patients with chronic myeloid leukemia revealed changes in the gene expression in many proteins involved in Wnt and Hh signaling. Carrara et al.[76] specifically described Tcf4 as a potential biomarker of chronic myeloid leukemia in patients. Also, specific mutations in this gene have been identified as being associated with the risk of developing medulloblastoma. It was shown that prenatal and postnatal knockdown of Tcf4 gene in mice lead to different effects. The authors discuss in their paper the role of Tcf4 as a suppressor of medulloblastoma and the linkage of low expression with a worse clinical outcome[76].

One of the most represented proteins in associations with proteins and genes of the Hh pathway is the bone morphogenetic protein 2 (Bmp2, UniProt ID: P12643). According to the literature, Bmp2 and the Hh signaling are closely connected during embryo- and organogenesis[78, 79] through some nodes. It was shown that the use of recombinant human Shh in mouse periodontal tissues resulted in the increased cell growth and the enhanced mineralization as well as in the increased expression of Bmp2. This leads to the suggestion that the activation of the Hh pathway results in a positive stimulation of Bmp2 and its functions. GO biological processes of Bmp2 also include functions related to transcription, gene expression, proliferation, and development of many organs in general. Using our approach, we established the associations between Bmp2 and various types of cancer. In the publication by Li et al.[80], a possible role of Bmp2 in the development of esophageal adenocarcinoma is described. Esophageal types of cancer are characterized

by poor outcomes: according to SEER statistics, the 5-year survival rate of patients with this diagnosis is only 20%. Kalal et al.[81] showed that Bmp2 may act as a target for esophageal adenocarcinoma treatment. They provide results on the inhibition of tumor cells by the developed antibodies that also act synergically with cisplatin. Bmp2 was considered as a target not only for therapeutic strategies against esophageal adenocarcinoma. Shin et al.[82] discussed the role of decreased expression of Shh in the development of invasive urothelial carcinoma. The stimulation of Shh by Bmp2 leads to an increase in cell differentiation. They provide information on clinical effects of pharmacological stimulation of Bmp2 signaling which is appeared in a significant reduction in tumor progression. Bmp2 could also be used as a biomarker of cancer and a feature for clinical outcome assessment. For example, analysis of transcription profiles in healthy people and patients with Breast Cancer (BC) revealed that the levels of Bmp2 expression in the BC group are higher than in the control group.

Also, the expression of Bmp2 is increased in cases of microcalcification. Thus, an increased level of Bmp2 expression can be used to diagnose breast cancer and to determine the outcome of the disease. In the study by Kalal et al.[81], the proteomics approach reveals changes in expression in mice with melanoma that were treated with Bmp2 inhibitor. Furthermore, subsequent bioinformatic analysis showed that inhibition of Bmp2 leads to changes in protein binding and catalytic activity in melanoma.

In general, most of the extracted proteins that could potentially be associated with the development of neoplastic diseases are involved in giant cascades of reactions consisting of multiple signaling pathways. It was proven by enrichment analysis: the aforementioned Wnt and Bmp2 signaling (which is also a specific part of Tgf-beta-signaling) were revealed. In particular, the Wnt protein, which plays one of the key roles in the Wnt signaling pathway, ultimately transmits molecular “information" to Tcf4. Indeed, among the extracted associations there are also associations between Tcf4 and neoplastic processes: lung cancer[83], melanoma[84], medulloblastoma[85], gastric adenocarcinoma[86], etc. It can also be considered as a target for the treatment of neoplastic diseases[83, 84, 87]. Similar to Tcf4 and Wnt, Myc is also involved in the same pathway and was extracted within the “neoplastic disease (process) protein” pairs. Myc directly affects the transcription. Text-mining analysis revealed its association with basal cell carcinoma[88], acute myeloid leukemia[89], medulloblastoma[90, 91], etc. Myc can also be considered a target for treatment development[89] or a biomarker based on its expression levels[88].

Using text mining analysis, we also revealed associations of protein jagged-1 (jagged1 or Jag1), which is a Tnf signaling acceptor. The Tnf signaling was found in the enrichment analysis as a pathway that has an impact on the Hh pathway. It was shown that higher expression of Jag1 in patients with endometrial cancer correlated with a reduced mortality risk[92]. Thus, we can assume that the expression level of Jag1 can be used to determine the outcome of endometrial cancer. Polychronidou et al.[92] considered Jag1 to be a part of the Notch signaling. They observed that the increased expression of Notch2 and Notch3, along with the increased expression of Jag1, was associated with a worse health status. Based on the knowledge that Jag1 is expressed on both tumor epithelial and endothelial cells, Steg et al.[93] made an assumption that this protein could be regarded as a target for therapy. They performed in vitro and vivo experiments in ovarian cancer targeting Jag1 by siRNA constructs. Targeting Jag1 induced apoptosis, reduced cell viability and allowed to avoid resistance to taxane through downregulation of Gli2. Jag1 was also shown to play a role in renal cancer[94], breast cancer[95], prostate cancer[96], and many other types of neoplasms.

Enrichment analysis allowed identifying associations of Mapk signaling pathways with the Hh pathway. Text-mining approach allows for revealing associations of beta-arrestin-1 (Arrb1) with neoplastic processes. Moreover, Arrb1 is involved in gonadotropin releasing hormone secretion regulation. This example may explain the presence of many hormonal-dependent processes in enrichment analysis. Dobson et al.[97] investigated the role of RE1-silencing transcription factor (Rest1) in medulloblastomas development. They revealed that Rest1 antagonizes Gli1 by regulating Arrb1 which is an inhibitor of Gli1. Also, Arrb1 is associated with gastric cancer: it was shown that a knockdown of the Arrb1 gene led to impaired cell proliferation[98]. Unidirectional changes in the expression of the Arrb1 gene may lead to divergent prognoses in different types of cancer[99]. Arrb1 can be used for the therapy of neoplastic processes. Direct

targeting of Arrb1 led to inhibition of cell proliferation and promoted apoptosis in non-small cell lung cancer[100]. Various roles of Arrb1 in prostate cancer[101] and gallbladder cancer[102] were described.

Enrichment analysis results are in accordance with some other results of text-mining on proteins that take part in signaling pathways. Both FoxO and Hippo signaling pathways included protein Smad 3. Text-mining analysis revealed an association of Smad 3 with basal cell carcinoma since this protein is also a part of Tgf-beta signaling pathway[86]. It was also shown that Smad 3 impaired cell proliferation and metastasis in bladder cancer. The obtained results were demonstrated by a knockdown of Hoxa 1 which increases the transcription of Smad 3[103].

3.4.2 Role of extracted miRNAs in neoplastic diseases pathogenesis

In total, we were able to collect 87 abstract texts relevant to the investigation of miRNAs' role in neoplastic processes. Association extraction allowed us to identify 329 miRNA-cancer relations unique within the abstract. The result is provided in the Supplementary Materials (Table S12 in the ESM).

During the described analysis, we found that the effect of microRNAs on such neoplastic processes as glioblastoma, medulloblastoma, and pancreatic cancer is most commonly described in the literature. MiR-200, miR-326, and miR-21 are some microRNAs, the relationships of which with neoplastic processes are studied intensively.

For example, miR-326 is associated with chronic myeloid leukemia. It was shown, that the reduced expression of miR-326 correlates with Smo upregulation in CD34+ cells in patients with a chronic myeloid leukemia diagnosis[104]. Moreover, miR-326 was shown to be associated with glioma cells sensitivity to curcumin through inhibition of the canonical Shh pathway[105]. By transcriptomic analysis, Minhee Park and colleagues[106] revealed overexpression of Hh pathway proteins in pancreas neoplasm tissue cells as well as associations of the miR-200 family and miR-192/215 with upregulated genes of the Hh pathway.

4 Conclusion

We develop and test a text-mining approach that provides the automated processing of a large number of scientific abstracts and the extraction from them of the proteins and genes names, followed by the identification of associations between proteins, genes, diseases, and pathological processes. Using the developed approach, we consider the extracted names of proteins and genes that can interact with the components of the Hedgehog signaling pathway, which is currently considered one of the pathways that, when disrupted, can play an important role in the development of neoplastic diseases. Using our approach, we collect a list of key proteins that might be involved in the development of particular tumors such as basal cell carcinoma, squamous cell carcinoma, medulloblastoma, and others. In particular, using the presented approach, one can observe the whole set of tumor types and reveal their potential molecular mechanisms. Using the gene set enrichment analysis, we show that the proteins and genes involved in the Hedgehog pathway or interacting with its nodes can also be a part of or regulate some other pathways essential for cell growth and differentiation. Some of the molecules identified via automated extraction from texts of scientific abstracts are identified in experimental studies as potential molecular targets for cancer treatment or as biomarkers that can be used for early diagnosis of a particular type of tumor.

The developed algorithm allows for the extraction of information relevant to the regulation of neoplastic processes that may be regulated by the Hh pathway and related proteins. An analysis of the biological functions of the identified proteins and signaling pathways demonstrates the applicability of the developed algorithm to the identification of processes and molecular mechanisms leading to the progression of a wide range of diseases.

Acknowledgment

This work was supported by the Ministry of Science and Higher Education of the Russian Federation within the framework of state support for the creation and development of World-Class Research Centers `Digital Biodesign and Personalized Healthcare' (No. 75-15-2022-305).

Electronic Supplementary Material

Supplementary material including

- Table S1: List of token features used for training CRF-based named protein and gene named entity recognition algorithm.
- Table S2: List of pattern-phrases and rules that are used for extraction of associations between proteins

and genes from the texts of publications.

- Table S3: Comparison of terms used for the classification of cancer types according to TCGA (OncoDB) with the human disease ontology.
- Table S4: List of terms that were used for the recognition of Hh proteins and genes among the texts.
- Table S5: Association between proteins of the Hh pathway and other human proteins that are extracted from the texts of the "Hh interactions" corpus.
- Table S6: Extracted from the texts associations between Hedgehog and related proteins and neoplastic disorders.
- Table S7: Unique associations between proteins of Hh pathway and their interactors, and neoplastic diseases, extracted from the texts, with their frequencies of occurrence among the texts of corpus "Hh-onco" where they were mentioned.
- Table S8: Identified clusters of GO BP terms based on enrichment analysis for proteins interacting with Hh-pathway.
- Table S9: List of proteins interacting with Hh pathway grouped by its' mechanism of action.
- Table S10: Human miRNAs revealed by enrichment analysis, which may regulate Hh pathway components.
- Table S11: Full comparison between data on the Hh pathway associations, and experimental data on differentially expressed genes in cancer tissues according to OncoDB.
- Table S12: Associations between miRNAs and neoplastic processes extracted from the texts.

It is available in the online version of this article at https://doi.org/10.26599/BDMA.2023.9020007.

References

[1] H. Le, R. Kleinerman, O. Z. Lerman, D. Brown, R. Galiano, G. C. Gurtner, S. M. Warren, J. P. Levine, and P. B. Saadeh, Hedgehog signaling is essential for normal wound healing, *Wound Repair Regen.*, vol. 16, no. 6, pp. 768–773, 2008.

[2] G. B. Carballo, J. R. Honorato, G. P. F. de Lopes, and T. C. L. de Sampaio e Spohr, A highlight on Sonic hedgehog pathway, *Cell Commun. Signal.*, vol. 16, no. 1, pp. 11, 2018.

[3] M. G. Smelkinson, The hedgehog signaling pathway emerges as a pathogenic target, *J. Dev. Biol.*, vol. 5, no. 4, pp. 14, 2017.

[4] E. Dessaud, A. P. McMahon, and J. Briscoe, Pattern formation in the vertebrate neural tube: a sonic hedgehog morphogen-regulated transcriptional network, *Development*, vol. 135, no. 15, pp. 2489–2503, 2008.

[5] S. S. Choi, S. Bradrick, G. Qiang, A. Mostafavi, G. Chaturvedi, S. A. Weinman, A. M. Diehl, and R. Jhaveri, Up-regulation of Hedgehog pathway is associated with cellular permissiveness for hepatitis C virus replication, *Hepatology*, vol. 54, no. 5, pp. 1580–1590, 2011.

[6] J. Fan, X. Zeng, Y. Li, S. Wang, P. Yang, Z. Cao, Z. Wang, P. Song, X. Mei, and D. Ju, A novel therapeutic approach against B-cell non-Hodgkin's lymphoma through co-inhibition of Hedgehog signaling pathway and autophagy, *Tumor Biol.*, vol. 37, no. 6, pp. 7305–7314, 2016.

[7] R. Teperino, F. Aberger, H. Esterbauer, N. Riobo, and J. A. Pospisilik, Canonical and non-canonical Hedgehog signalling and the control of metabolism, *Semin. Cell Dev. Biol.*, vol. 33, pp. 81–92, 2014.

[8] P. Niewiadomski, S. M. Niedzióa, Markiewicz, T. Uki, B. Baran, and K. Chojnowska, Gli Proteins: Regulation in development and cancer, *Cells*, vol. 8, no. 2, pp. 147, 2019.

[9] J. Wu, D. Di, C. Zhao, Y. Liu, H. Chen, Y. Gong, X. Zhao, and H. Chen, Role of Glioma-associated GLI1 oncogene in carcinogenesis and cancertargeted therapy, *Curr. Cancer Drug Targets*, vol. 18, no. 6, pp. 558–566, 2018.

[10] N. Biziukova, O. Tarasova, S. Ivanov, and V. Poroikov, Automated extraction of information from texts of scientific publications: Insights into HIV treatment strategies, *Front. Genet.*, vol. 11, pp. 618862, 2020.

[11] O. A. Tarasova, N. Y. Biziukova, A. V. Rudik, A. V. Dmitriev, D. A. Filimonov, and V. V. Poroikov, Extraction of data on parent compounds and their metabolites from texts of scientific abstracts, *J. Chem. Inf. Model.*, vol. 61, no. 4, pp. 1683–1690, 2021.

[12] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in *Proc. 18th Int. Conf. on Machine Learning*, San Francisco, CA, USA, 2001, pp. 282–289.

[13] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. L. Yeh, UniProt: The universal protein knowledgebase, *Nucleic Acids Res.*, vol. 32, pp. D115–D119, 2004.

[14] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, et al., The ChEMBL database in 2017, *Nucleic Acids Res.*, vol. 45, no. D1, pp. D945–D954, 2017.

[15] M. Kanehisa and S. Goto, KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, 2000.

[16] L. M. Schriml, J. B. Munro, M. Schor, D. Olley, C. McCracken, V. Felix, J. A. Baron, R. Jackson, S. M. Bello, C. Bearer, et al., The Human Disease Ontology 2022 update, *Nucleic Acids Res.*, vol. 50, no. D1, pp. D1255–D1261, 2022.

[17] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, Cytoscape: A software environment for integrated models of biomolecular interaction networks, *Genome*

Res., vol. 13, no. 11, pp. 2498–2504, 2003.
[18] W. Min, T. H. Chang, S. Zhang, and X. Wan, TSCCA: A tensor sparse CCA method for detecting microRNA-gene patterns from multiple cancers, *PLoS Comput. Biol.*, vol. 17, no. 6, pp. e1009044, 2021.
[19] T. Wu, E. Hu, S. Xu, M. Chen, P. Guo, Z. Dai, T. Feng, L. Zhou, W. Tang, L. Zhan, et al., clusterProfiler 4.0: A universal enrichment tool for interpreting omics data, *Innovation*, vol. 2, no. 3, pp. 100141, 2021.
[20] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C. F. Chen, A new method to measure the semantic similarity of GO terms, *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, 2007.
[21] H. Y. Huang, Y. C. D. Lin, S. Cui, Y. Huang, Y. Tang, J. Xu, J. Bao, Y. Li, J. Wen, H. Zuo, et al., miRTarBase update 2022: An informative resource for experimentally validated miRNA-target interactions, *Nucleic Acids Res.*, vol. 50, no. D1, pp. D222–D230, 2022.
[22] G. Tang, M. Cho, and X. Wang, OncoDB: An interactive online database for analysis of gene expression and viral infection in cancer, *Nucleic Acids Res.*, vol. 50, no. D1, pp. D1334–D1339, 2022.
[23] OncoDB, https://oncodb.org/index.html, 2022.
[24] National Institutes of Health, Center for Cancer Genomics at the National Cancer Institute, The Cancer Genome Atlas Program (TCGA), https://www.cancer.gov/ccg/research/genome-sequencing/tcga, 2022.
[25] National Institutes of Health, National Cancer Institute, The Surveillance, Epidemiology, and End Results (SEER) Program, https://seer.cancer.gov/, 2022.
[26] H. Li, J. Li, W. Gao, C. Zhen, and L. Feng, Systematic analysis of ovarian cancer platinum-resistance mechanisms via text mining, *J. Ovarian Res.*, vol. 13, no. 1, pp. 27, 2020.
[27] B. Settles, ABNER: An open source tool for automatically tagging genes, *Bioinformatics.*, vol. 21, no. 14, pp. 3191–3192, 2005.
[28] W. Min, J. Liu, and S. Zhang, Edge-group sparse PCA for network-guided high dimensional data analysis, *Bioinformatics*, vol. 34, no. 20, pp. 3479–3487, 2018.
[29] C. Zhen, C. Zhu, H. Chen, Y. Xiong, J. Tan, D. Chen, and J. Li, Systematic analysis of molecular mechanisms for HCC metastasis via text mining approach, *Oncotarget*, vol. 8, no. 8, pp. 13909–13916, 2017.
[30] C. Bizon, S. Cox, J. Balhoff, Y. Kebede, P. Wang, K. Morton, K. Fecho, and A. Tropsha, ROBOKOP KG and KGB: Integrated knowledge graphs from federated sources, *J. Chem. Inf. Model.*, vol. 59, no. 12, pp. 4968–4973, 2019.
[31] The Allen Institute for Artificial Intelligence, CORD-19: COVID-19 Open Research Dataset, https://allenai.org/data/cord-19, 2020.
[32] S. K. Hong and J. G. Lee, DTranNER: Biomedical named entity recognition with deep learning-based label-label transition model, *BMC Bioinformatics*, vol. 21, no. 1, pp. 53, 2020.
[33] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, J. Wang, and H. Lin, A neural network approach to chemical and gene/protein entity recognition in patents, *J. Cheminform.*, vol. 10, no. 1, pp. 65, 2018.
[34] S. Kaewphan, K. Hakala, N. Miekka, T. Salakoski, and F. Ginter, Wide-scope biomedical named entity recognition and normalization with CRFs, fuzzy matching and character level modeling, *Database* (*Oxford*), vol. 2018, pp. 1–10, 2018.
[35] H. Zhou, S. Ning, Z. Liu, C. Lang, Z. Liu, and B. Lei, Knowledge-enhanced biomedical named entity recognition and normalization: Application to proteins and genes, *BMC Bioinformatics*, vol. 21, no. 1, pp. 35, 2020.
[36] L. Luo, P. T. Lai, C. H. Wei, C. N. Arighi, and Z. Lu, BioRED: A rich biomedical relation extraction dataset, *Brief. Bioinform.*, vol. 23, no. 5, pp. bbac282, 2022.
[37] Y. Zhang, H. Lin, Z. Yang, J. Wang, S. Zhang, Y. Sun, and L. Yang, A hybrid model based on neural networks for biomedical relation extraction, *J. Biomed. Inform.*, vol. 81, pp. 83–92, 2018.
[38] H. Fei, Y. Zhang, Y. Ren, and D. Ji, A span-graph neural model for overlapping entity relation extraction in biomedical texts, *Bioinformatics*, vol. 37, no. 11, pp. 1581–1589, 2021.
[39] S. Kuijper, H. Feitsma, R. Sheth, J. Korving, M. Reijnen, and F. Meijlink, Function and regulation of *Alx4* in limb development: Complex genetic interactions with *Gli3* and *Shh*, *Dev. Biol.*, vol. 285, no. 2, pp. 533–544, 2005.
[40] P. Mill, R. Mo, M. C. Hu, L. Dagnino, N. D. Rosenblum, and C. C. Hui, Shh controls epithelial proliferation via independent pathways that converge on N-Myc, *Dev. Cell*, vol. 9, no. 2, pp. 293–303, 2005.
[41] T. T. Turner, H. J. Bang, S. A. Attipoe, D. S. Johnston, and J. L. Tomsig, Sonic hedgehog pathway inhibition alters epididymal function as assessed by the development of sperm motility, *J. Androl.*, vol. 27, no. 2, pp. 225–232, 2006.
[42] M. C. Hu, R. Mo, S. Bhella, C. W. Wilson, P. T. Chuang, C. C. Hui, and N. D. Rosenblum, GLI3-dependent transcriptional repression of *Gli1*, *Gli2* and kidney patterning genes disrupts renal morphogenesis, *Development*, vol. 133, no. 3, pp. 569–578, 2006.
[43] O. Nolan-Stevaux, J. Lau, M. L. Truitt, G. C. Chu, M. Hebrok, M. E. Fernández-Zapico, and D. Hanahan, *GLI1* is regulated through Smoothened-independent mechanisms in neoplastic pancreatic ducts and mediates PDAC cell survival and transformation, *Genes Dev.*, vol. 23, no. 1, pp. 24–36, 2009.
[44] E. J. Tolosa, M. G. Fernandez-Barrena, E. Iguchi, A. L. McCleary-Wheeler, R. M. Carr, L. L. Almada, L. F. Flores, R. E. Vera, G. W. Alfonse, D. L. Marks, et al., GLI1/GLI2 functional interplay is required to control Hedgehog/GLI targets gene expression, *Biochem. J.*, vol. 477, no. 17, pp. 3131–3145, 2020.
[45] S. Dennler J. André, I. Alexaki, A. Li, T. Magnaldo, P. ten Dijke, X. J. Wang, F. Verrecchia, and A. Mauviel, Induction of sonic hedgehog mediators by transforming growth factor-beta: Smad3-dependent activation of *Gli2* and *Gli1* expression *in vitro* and *in vivo*, *Cancer Res.*, vol. 67, no. 14, pp. 6981–6986, 2007.
[46] S. L. Wild, A. Elghajiji, C. Grimaldos Rodriguez, S. D.

Weston, Z. D. Burke, and D. Tosh, The canonical Wnt pathway as a key regulator in liver development, *Genes* (*Basel*)., vol. 11, no. 10, pp. 1163, 2020.

[47] A. Guasto and V. Cormier-Daire, Signaling pathways in bone development and their related skeletal dysplasia, *Int. J. Mol. Sci.*, vol. 22, no. 9, pp. 4321, 2021.

[48] S. Pietrobono, S. Gagliardi, and B. Stecca, Non-canonical hedgehog signaling pathway in cancer: Activation of GLI transcription factors beyond smoothened, *Front. Genet.*, vol. 10, pp. 556, 2019.

[49] J. H. Kong, C. Siebold, and R. Rohatgi, Biochemical mechanisms of vertebrate hedgehog signaling, *Development*, vol. 146, no. 10, pp. dev166892, 2019.

[50] V. Montagnani and B. Stecca, Role of protein kinases in hedgehog pathway control and implications for cancer therapy, *Cancers* (*Basel*), vol. 11, no. 4, pp. 449, 2019.

[51] J. Pyczek, N. Khizanishvili, M. Kuzyakova, S. Zabel, J. Bauer, F. Nitzki, S. Emmert, M. P. Schön, P. Boukamp, H. U. Schildhaus, et al., Regulation and role of GLI1 in cutaneous squamous cell carcinoma pathogenesis, *Front. Genet.*, vol. 10, pp. 1185, 2019.

[52] N. J. Rowbotham, A. L. Furmanski, A. L. Hager-Theodorides, S. E. Ross, E. Drakopoulou, C. Koufaris, S. V. Outram, and T. Crompton, Repression of hedgehog signal transduction in T-lineage cells increases TCR-induced activation and proliferation, *Cell Cycle*, vol. 7, no. 7, pp. 904–908, 2008.

[53] G. A. Stewart, J. A. Lowrey, S. J. Wakelin, P. M. Fitch, S. Lindey, M. J. Dallman, J. R. Lamb, and S. E. M. Howie, Sonic hedgehog signaling modulates activation of and cytokine production by human peripheral CD4+ T cells, *J. Immunol.*, vol. 169, no. 10, pp. 5451–5457, 2002.

[54] J. Y. Chai, V. Sugumar, M. A. Alshawsh, W. F. Wong, A. Arya, P. P. Chong, and C. Y. Looi, The role of smoothened-dependent and -independent hedgehog signaling pathway in tumorigenesis, *Biomedicines*, vol. 9, no. 9, pp. 1188, 2021.

[55] M. de la Roche, A. T. Ritter, K. L. Angus, C. Dinsmore, C. H. Earnshaw, J. F. Reiter, and G. M. Griffiths, Hedgehog signaling controls T cell killing at the immunological synapse, *Science*, vol. 342, no. 6163, pp. 1247–1250, 2013.

[56] A. L. Furmanski, A. Barbarulo, A. Solanki, C. I. Lau, H. Sahni, J. I. Saldana, F. D'Acquisto, and T. Crompton, The transcriptional activator Gli2 modulates T-cell receptor signalling through attenuation of AP-1 and NFκB activity, *J. Cell. Sci.*, vol. 128, no. 11, pp. 2085–2095, 2015.

[57] V. S. F. Chan, S. Y. Chau, L. Tian, Y. Chen, S. K. Y. Kwong, J. Quackenbush, M. Dallman, J. Lamb, and P. K. H. Tam, Sonic hedgehog promotes CD4+ T lymphocyte proliferation and modulates the expression of a subset of CD28-targeted genes, *Int. Immunol.*, vol. 18, no. 12, pp. 1627–1636, 2006.

[58] K. Yaddanapudi, J. De Miranda, M. Hornig, and W. I. Lipkin, Toll-like receptor 3 regulates neural stem cell proliferation by modulating the Sonic Hedgehog pathway, *PLoS One*, vol. 6, no. 10, pp. e26766, 2011.

[59] Y. Zhong, Y. Zhang, W. Liu, Y. Zhao, L. Zou, and X. Liu, TLR4 modulates senescence and paracrine action in placental mesenchymal stem cells via inhibiting hedgehog signaling pathway in preeclampsia, *Oxid. Med. Cell. Longev.*, vol. 2022, pp. 7202837, 2022.

[60] S. J. Matissek, M. Karbalivand, W. Han, A. Boutilier, E. Yzar-Garcia, L. L. Kehoe, D. S. Gardner, A. Hage, K. Fleck, V. Jeffers, et al., A novel mechanism of regulation of the oncogenic transcription factor GLI3 by toll-like receptor signaling, *Oncotarget*, vol. 13, pp. 944–959, 2022.

[61] A. N. Sigafoos, B. D. Paradise, and M. E. Fernandez-Zapico, Hedgehog/GLI signaling pathway: Transduction, regulation, and implications for disease, *Cancers* (*Basel*), vol. 13, no. 14, pp. 3410, 2021.

[62] S. Fattahi, M. P. Langroudi, and H. Akhavan-Niaki, Hedgehog signaling pathway: Epigenetic regulation and role in disease and cancer development, *J. Cell. Physiol.*, vol. 233, no. 8, pp. 5726–5735, 2018.

[63] M. Niyaz, M. S. Khan, and S. Mudassar, Hedgehog signaling: An Achilles' Heel in cancer, *Transl. Oncol.*, vol. 12, no. 10, pp. 1334–1344, 2019.

[64] S. Iriana, K. Asha, M. Repak, and N. Sharma-Walia, Hedgehog signaling: Implications in cancers and viral infections, *Int. J. Mol. Sci.*, vol. 22, no. 3, pp. 1042, 2021.

[65] Y. Zhou, J. Huang, B. Jin, S. He, Y. Dang, T. Zhao, and Z. Jin, The emerging role of hedgehog signaling in viral infections, *Front. Microbiol.*, vol. 13, pp. 870316, 2022.

[66] T. Yoshida, A. Hamano, A. Ueda, H. Takeuchi, and S. Yamaoka, Human SMOOTHENED inhibits human immunodeficiency virus type 1 infection, *Biochem. Biophys. Res. Commun.*, vol. 493, no. 1, pp. 132–138, 2017.

[67] H. Y. Kim, H. K. Cho, S. P. Hong, and J. Cheong, Hepatitis B virus X protein stimulates the Hedgehog-Gli activation through protein stabilization and nuclear localization of Gli1 in liver cancer cells, *Cancer Lett.*, vol. 309, no. 2, pp. 176–184, 2011.

[68] Z. HajiEsmailPoor, P. Tabnak, B. Ahmadzadeh, S. S. Ebrahimi, B. Faal, and N. Mashatan, Role of hedgehog signaling related non-coding RNAs in developmental and pathological conditions, *Biomed. Pharmacother.*, vol. 153, pp. 113507, 2022.

[69] T. Liu, Z. Wang, M. Dong, J. Wei, and Y. Pan, MicroRNA-26a inhibits cell proliferation and invasion by targeting FAM98A in breast cancer, *Oncol. Lett.*, vol. 21, no. 5, pp. 367, 2021.

[70] E. Ferretti, E. De Smaele, E. Miele, P. Laneve, A. Po, M. Pelloni, A. Paganelli, L. Di Marcotullio, E. Caffarelli, I. Screpanti, et al., Concerted microRNA control of Hedgehog signalling in cerebellar neuronal progenitor and tumour cells, *EMBO J.*, vol. 27, no. 19, pp. 2616–2627, 2008.

[71] Z. Hu, L. Li, J. Ran, G. Chu, H. Gao, L. Guo, and J. Chen, miR-125b acts as anti-fibrotic therapeutic target through regulating Gli3 in *vivo* and in *vitro*, *Ann. Hepatol.*, vol. 18, no. 6, pp. 825–832, 2019.

[72] Z. Liang, J. Li, L. Zhao, and Y. Deng, miR-375 affects the hedgehog signaling pathway by downregulating RAC1 to inhibit hepatic stellate cell viability and

epithelial-mesenchymal transition, *Mol. Med. Rep.*, vol. 23, no. 3, pp. 182, 2021.

[73] Y. Hou, Q. Hu, J. Huang, and H. Xiong, Omeprazole inhibits cell proliferation and induces G0/G1 cell cycle arrest through up-regulating miR-203a-3p expression in Barrett's esophagus cells, *Front. Pharmacol.*, vol. 8, pp. 968, 2018.

[74] B. N. Singh, M. J. Doyle, C. V. Weaver, N. Koyano-Nakagawa, and D. J. Garry, Hedgehog and Wnt coordinate signaling in myogenic progenitors and regulate limb regeneration, *Dev. Biol.*, vol. 371, no. 1, pp. 23–34, 2012.

[75] G. R. van den Brink, S. A. Bleuming, J. C. H. Hardwick, B. L. Schepman, G. J. Offerhaus, J. J. Keller, C. Nielsen, W. Gaffield, S. J. H. van Deventer, D. J. Roberts, et al., Indian Hedgehog is an antagonist of Wnt signaling in colonic epithelial cell differentiation, *Nat. Genet.*, vol. 36, no. 3, pp. 277–282, 2004.

[76] R. de Cássia Viu Carrara, A. M. Fontes, K. J. Abraham, M. D. Orellana, S. K. Haddad, P. V. B. Palma, R. A. Panepucci, M. A. Zago, and D. T. Covas, Expression differences of genes in the PI3K/AKT, WNT/b-catenin, SHH, NOTCH and MAPK signaling pathways in CD34+ hematopoietic cells obtained from chronic phase patients with chronic myeloid leukemia and from healthy controls, *Clin. Transl. Oncol.*, vol. 20, no. 4, pp. 542–549, 2018.

[77] M. Hellwig, M. C. Lauffer, M. Bockmayr, M. Spohn, D. J. Merk, L. Harrison, J. Ahlfeld, A. Kitowski, J. E. Neumann, J. Ohli, et al., TCF4 (E2-2) harbors tumor suppressive functions in SHH medulloblastoma, *Acta Neuropathol.*, vol. 137, no. 4, pp. 657–673, 2019.

[78] D. Hu, N. M. Young, X. Li, Y. Xu, B. Hallgrímsson, and R. S. Marcucio, A dynamic *Shh* expression pattern, regulated by SHH and BMP signaling, coordinates fusion of primordia in the amniote face, *Development*, vol. 142, no. 3, pp. 567–574, 2015.

[79] W. J. Bae, Q. S. Auh, H. C. Lim, G. T. Kim, H. S. Kim, and E. C. Kim, Sonic hedgehog promotes cementoblastic differentiation via activating the BMP pathways, *Calcif. Tissue Int.*, vol. 99, no. 4, pp. 396–407, 2016.

[80] S. Li, S. J. M. Hoefnagel, M. Read, S. Meijer, M. I. van Berge Henegouwen, S. S. Gisbertz, E. Bonora, D. S. H. Liu, W. A. Phillips, S. Calpe, et al., Selective targeting BMP2 and 4 in SMAD4 negative esophageal adenocarcinoma inhibits tumor growth and aggressiveness in preclinical models, *Cell. Oncol.* (*Dordr.*), vol. 45, no. 4, pp. 639–658, 2022.

[81] B. S. Kalal, P. K. Modi, D. Upadhya, P. Saha, T. S. K. Prasad, and V. R. Pai, Inhibition of bone morphogenetic proteins signaling suppresses metastasis melanoma: A proteomics approach, *Am. J. Transl. Res.*, vol. 13, no. 10, pp. 11081–11093, 2021.

[82] K. Shin, A. Lim, C. Zhao, D. Sahoo, Y. Pan, E. Spiekerkoetter, J. C. Liao, and P. A. Beachy, Hedgehog signaling restrains bladder cancer progression by eliciting stromal production of urothelial differentiation factors, *Cancer Cell*, vol. 26, no. 4, pp. 521–533, 2014.

[83] J. Y. Zhu, X. Yang, Y. Chen, Y. Jiang, S. J. Wang, Y. Li, X. Q. Wang, Y. Meng, M. M. Zhu, X. Ma, et al., Curcumin suppresses lung cancer stem cells via inhibiting Wnt/*β*-catenin and sonic hedgehog pathways, *Phytother. Res.*, vol. 31, no. 4, pp. 680–688, 2017.

[84] G. Liang, M. Liu, Q. Wang, Y. Shen, H. Mei, D. Li, and W. Liu, Itraconazole exerts its anti-melanoma effect by suppressing Hedgehog, Wnt, and PI3K/mTOR signaling pathways, *Oncotarget*, vol. 8, no. 17, pp. 28510–28525, 2017.

[85] J. Rodriguez-Blanco, L. Pednekar, C. Penas, B. Li, V. Martin, J. Long, E. Lee, W. A. Weiss, C. Rodriguez, N. Mehrdad, et al., Inhibition of WNT signaling attenuates self-renewal of SHH-subgroup medulloblastoma, *Oncogene*, vol. 36, no. 45, pp. 6306–6314, 2017.

[86] Y. Tajima, T. Murakami, T. Saito, T. Hiromoto, Y. Akazawa, N. Sasahara, H. Mitomi, T. Yao, and S. Watanabe, Distinct involvement of the sonic hedgehog signaling pathway in gastric adenocarcinoma of fundic gland type and conventional gastric adenocarcinoma, *Digestion*, vol. 96, no. 2, pp. 81–91, 2017.

[87] T. Ueda, H. Tsubamoto, K. Inoue, K. Sakata, H. Shibahara, and T. Sonoda, Itraconazole modulates hedgehog, WNT/*β*-catenin, as well as akt signalling, and inhibits proliferation of cervical cancer cells, *Anticancer Res.*, vol. 37, no. 7, pp. 3521–3526, 2017.

[88] J. M. Bonifas, S. Pennypacker, P. T. Chuang, A. P. McMahon, M. Williams, A. Rosenthal, F. J. De Sauvage, and E. H. Epstein Jr, Activation of expression of hedgehog target genes in basal cell carcinomas, *J. Invest. Dermatol.*, vol. 116, no. 5, pp. 739–742, 2001.

[89] Y. Xu, P. Wang, M. Li, Z. Wu, X. Li, J. Shen, and R. Xu, Natural small molecule triptonide inhibits lethal acute myeloid leukemia with FLT3-ITD mutation by targeting Hedgehog/FLT3 signaling, *Biomed. Pharmacother.*, vol. 133, pp. 111054, 2021.

[90] S. Huq, N. V. Kannapadi, J. Casaos, T. Lott, R. Felder, R. Serra, N. L. Gorelick, M. A. Ruiz-Cardozo, A. S. Ding, A. Cecia, et al., Preclinical efficacy of ribavirin in SHH and group 3 medulloblastoma, *J. Neurosurg. Pediatr.*, vol. 27, no. 4, pp. 482–488, 2021.

[91] V. Kumar, Q. Wang, B. Sethi, F. Lin, V. Kumar, D. W. Coulter, Y. Dong, and R. I. Mahato, Polymeric nanomedicine for overcoming resistance mechanisms in hedgehog and Myc-amplified medulloblastoma, *Biomaterials*, vol. 278, pp. 121138, 2021.

[92] G. Polychronidou, V. Kotoula, K. Manousou, I. Kostopoulos, G. Karayannopoulou, E. Vrettou, M. Bobos, G. Raptou, I. Efstratiou, D. Dionysopoulos, et al., Mismatch repair deficiency and aberrations in the Notch and Hedgehog pathways are of prognostic value in patients with endometrial cancer, *PLoS One*, vol. 13, no. 12, pp. e0208221, 2018.

[93] A. D. Steg, A. A. Katre, B. Goodman, H. D. Han, A. M. Nick, R. L. Stone, R. L. Coleman, R. D. Alvarez, G. Lopez-Berestein, A. K. Sood, et al., Targeting the notch ligand JAGGED1 in both tumor cells and stroma in ovarian cancer, *Clin. Cancer Res.*, vol. 17, no. 17, pp. 5674–5685, 2011.

[94] D. Danielpour, S. Corum, P. Leahy, and A. Bangalore, Jagged-1 is induced by mTOR inhibitors in renal cancer cells through an Akt/ALK5/Smad4-dependent mechanism, *Curr Res. Pharmacol. Drug Discov.*, vol. 3, pp. 100117, 2022.

[95] J. P. Liu, Y. T. Shi, M. M. Wu, M. Q. Xu, F. M. Zhang, Z. Q. He, and M. Tang, JAG1 promotes migration, invasion, and adhesion of triple-negative breast cancer cells by promoting angiogenesis, (in Chinese), *J. Southern Med. Univ.*, vol. 42, no. 7, pp. 1100–1109, 2022.

[96] L. Rios-Colon, J. Chijioke, S. Niture, Z. Afzal, Q. Qi, A. Srivastava, M. Ramalinga, H. Kedir, P. Cagle, E. Arthur, et al., Leptin modulated microRNA-628-5p targets Jagged-1 and inhibits prostate cancer hallmarks, *Sci. Rep.*, vol. 12, no. 1, pp. 10073, 2022.

[97] T. H. W. Dobson, R. H. Tao, J. Swaminathan, S. Maegawa, S. Shaik, J. Bravo-Alegria, A. Sharma, B. Kennis, Y. Yang, K. Callegari, et al., Transcriptional repressor REST drives lineage stage-specific chromatin compaction at *Ptch1* and increases AKT activation in a mouse model of medulloblastoma, *Sci. Signal.*, vol. 12, no. 565, pp. eaan8680, 2019.

[98] H. Yu, M. Wang, T. Zhang, L. Cao, Z. Li, Y. Du, Y. Hai, X. Gao, J. Ji, and J. Wu, Dual roles of β-arrestin 1 in mediating cell metabolism and proliferation in gastric cancer, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 119, no. 40, pp. e2123231119, 2022.

[99] Y. Ye, H. Jiang, Y. Wu, G. Wang, Y. Huang, W. Sun, and M. Zhang, Role of ARRB1 in prognosis and immunotherapy: A Pan-Cancer analysis, *Front. Mol. Biosci.*, vol. 9, pp. 1001225, 2022.

[100] Y. Jiang, P. Zhu, Y. Gao, and A. Wang, miR-379-5p inhibits cell proliferation and promotes cell apoptosis in non-small cell lung cancer by targeting β-arrestin-1, *Mol. Med. Rep.*, vol. 22, no. 6, pp. 4499–4508, 2020.

[101] H. T. Purayil, Y. Zhang, J. B. Black, R. Gharaibeh, and Y. Daaka, Nuclear β Arrestin1 regulates androgen receptor function in castration resistant prostate cancer, *Oncogene*, vol. 40, no. 14, pp. 2610–2620, 2021.

[102] X. Zhang, Z. Kong, X. Xu, X. Yun, J. Chao, D. Ding, T. Li, Y. Gao, N. Guan, C. Zhu, et al., ARRB1 drives gallbladder cancer progression by facilitating TAK1/MAPK signaling activation, *J. Cancer*, vol. 12, no. 7, pp. 1926–1935, 2021.

[103] Q. Fan, M. He, T. Sheng, X. Zhang, M. Sinha, B. Luxon, X. Zhao, and J. Xie, Requirement of TGF β signaling for SMO-mediated carcinogenesis, *J. Biol. Chem.*, vol. 285, no. 47, pp. 36570–36576, 2010.

[104] S. Babashah, M. Sadeghizadeh, A. Hajifathali, M. R. Tavirani, M. S. Zomorod, M. Ghadiani, and M. Soleimani, Targeting of the signal transducer Smo links microRNA-326 to the oncogenic Hedgehog pathway in CD34+ CML stem/progenitor cells, *Int. J. Cancer*, vol. 133, no. 3, pp. 579–589, 2013.

[105] S. Yin, W. Du, F. Wang, B. Han, Y. Cui, D. Yang, H. Chen, D. Liu, X. Liu, X. Zhai, et al., MicroRNA-326 sensitizes human glioblastoma cells to curcumin via the SHH/GLI1 signaling pathway, *Cancer Biol. Ther.*, vol. 19, no. 4, pp. 260–270, 2018.

[106] M. Park, M. Kim, D. Hwang, M. Park, W. K. Kim, S. K. Kim, J. Shin, E. S. Park, C. M. Kang, Y. K. Paik, et al., Characterization of gene expression and activated signaling pathways in solid-pseudopapillary neoplasm of pancreas, *Mod. Pathol.*, vol. 27, no. 4, pp. 580–593, 2014.

Nadezhda Yu. Biziukova is a PhD candidate at Institute of Biomedical Chemistry, Moscow, Russia, and is studying under the program “Mathematical Biology, Bioinformatics”. She is also a junior researcher at Laboratory of Structure-Function Based Drug Design (Department of Bioinformatics) of IBMC, Russia. She is currently involved in the development of text-mining algorithms for biomedical information retrieval from the literature. Her research interests include text-mining, bioinformatics, and analysis of virus-host interactions

Sergey M. Ivanov received the PhD degree in bioinformatics from the Institute of Biomedical Chemistry, Moscow, Russia in 2014. He is currently a senior researcher at Laboratory of Structure-Function Based Drug Design of IBMC, Russia. He is also an associate professor at Department of Bioinformatics, Pirogov Russian National Research Medical University, Moscow, Russia. His research interests include bioinformatics, cheminformatics, computer-aided drug design, toxicology, transcriptomics data analysis, and systems biology.

Olga A. Tarasova graduated from the Pirogov Russian State Medical University in 2008, and received the PhD degree in mathematical biology and bioinformatics from the Institute of Biomedical Chemistry, Moscow, Russia. Currently, she is a senior researcher at Laboratory of Structure-Function Based Drug Design (Department of Bioinformatics) of IBMC (Moscow, Russia). Her research interests include bioinformatics, cheminformatics, computer-aided drug design, virology, analysis of virus-host interactions, text, and data mining.