

## Rules Dictionary Implementation

The dictionary was created by sequential pattern mining in the 26,255 abstracts classified as positive. Keywords related to the activity of anticancer agents were also used. First, abstracts with at least one association of recognized polyphenol-cancer or polyphenol-gene entities were selected and tokenized into sentences using the `sent_detect_nlp` function of the R `qdap` package. Only sentences containing at least 6 terms and 40 characters were considered. A total of 250,796 sentences were generated and 72,019 containing associations between entities (polyphenol-cancer, polyphenol-gene, gene-cancer) were selected.

Next, the 72,019 sentences were processed similar to the procedure described by (1), in which the entities were replaced with specific markers (CH& for the polyphenol entity; D&S for the cancer entity, and G&N for the gene entity). For example, the sentence “Curcumin inhibits invasion and migration of human pancreatic cancer via suppression of the ERK/NF-kB pathway” was changed to “CH& inhibits invasion and migration of human D&S via suppression of the G&N pathway” after processing. The sentences containing the substituted entities were pre-processed (removal of PubMed stopwords (PubMed HELP – Available at <[https://www.ncbi.nlm.nih.gov/books/NBK3827/pdf/Bookshelf\\_NBK3827.pdf](https://www.ncbi.nlm.nih.gov/books/NBK3827/pdf/Bookshelf_NBK3827.pdf)>. Accessed 6 November 2018), punctuation and single letters, cleaning of noise and special characters). Among the stopwords, prepositions were not removed since they are useful for searching associations between entities in the sentences (1,2).

After pre-processing, the 72,019 sentences were represented in the form of sequences for frequent sequential pattern mining, totaling 1,373,000 transactions (terms). The SPADE algorithm (3) available in the `cSPADE` function of the R `arulesSequences` package was used for mining. A support value of 0.0004 was adopted, i.e., only patterns that were present in at least 0.04% of all sequences ( $\approx 30$  sentences) were extracted. Patterns containing only one transaction and support  $\geq 0.01$  ( $\approx 720$  sentences), together with verbs (e.g., inhibited, suppressed, induces, active), keywords (e.g., inhibition, regulation, suppression, invasion, apoptosis, proliferation) and prepositions (e.g., of, by, in, on), related to anticancer activity, cancer markers and other more specific mechanisms (4-6) were used for the selection of patterns with associations between entities (e.g.,  $\langle \{CH\&\}, \{inhibited\}, \{proliferation\}, \{D\&S\} \rangle$ ,  $\langle \{CH\&\}, \{inhibits\}, \{invasion\}, \{D\&S\} \rangle$ ,  $\langle \{CH\}, \{increased\}, \{expression\}, \{GN\} \rangle$ ).

The selected patterns were analyzed manually and used for the elaboration of 25 categorized rules [anticancer activity of polyphenols (10 rules – R1 to R10), regulatory activity of genes and polyphenols (2 rules – R11 and R12), cancer markers (9 rules – HM12 to HM10), specific epigenetic markers (R13), novel drug (R14), cancer (R15), and conclusion/result (R16)] for the recognition of associations between entities in the sentences based on regular expressions (7-9). Supplementary Table 1 lists the rule specifications with the selected patterns and keywords, as well as the regular expressions created.

**Supplementary Table 1. Rules created for information extraction**

Rules for Associations of Anticancer Activity of Polyphenols	
Sequences (patterns)	Rules and regular expressions
<{CH},{inhibited},{DS}> <{CH},{blocked},{DS}> <{CH},{suppressed},{DS}> <{CH},{attenuated},{DS}> <{CH},{reduced},{DS}> <{CH},{reduce},{DS}> <{CH},{prevent},{DS}> <{CH},{prevents},{DS}>	<b>Rule 1</b>  <b>Code: R1</b> <u>verb related to cancer inhibitory activity, detected by the regular expression:</u> <i>(inhibits inhibited inhibit blocks blocked suppress suppressed suppresses attenuate attenuated attenuates reduce reduces reduced prevent prevents prevented)[\w]*</i>
<{CH},{inhibited},{proliferation},{DS}> <{CH},{inhibit},{invasion}> <{CH},{suppressed},{growth},{DS}> <{CH},{inhibit},{growth}> <{CH},{inhibit},{migration}> <{CH},{inhibit},{angiogenesis}> <{CH},{inhibit},{metastasis}> <{CH},{reduced},{viability},{DS}> <{CH},{viability},{of},{DS}>	<b>Rule 2</b>  <b>Code: R2</b> <u>verb + term related to cancer inhibitory activity, detected by the regular expression:</u> <i>(inhibit inhibits inhibited suppress suppresses suppressed reduce reduced)[\w]*(.*)(invasion growth migration angiogenesis metastasis viability proliferation cancer carcinogenesis cells cell line tumor tumour neoplasm)[\w]*</i>
<{CH},{protective},{against}> <{CH},{activity},{against}> <{CH},{activities},{against},{DS}> <{CH},{effects},{against},{DS}> <{CH},{agent},{against},{DS}> <{CH},{protect},{against}> <{CH},{potent},{against},{DS}> <{CH},{effective},{against},{DS}> <{CH},{promising},{against},{DS}> <{CH},{cytotoxicity},{against},{DS}> <{CH},{chemopreventive},{against},{DS}>	<b>Rule 3</b>  <b>Code: R3</b> <u>term + term related to anticancer activity, detected by the regular expression:</u> <i>(protective activity activities effects agent protect potent effective promising cytotoxicity cytotoxic chemopreventive anticancer anti-cancer antitumor anti-tumor antiproliferative anti-proliferative inhibitory)[\w]*(.*)(against)[\w]*</i>
<{CH},{caused},{inhibition},{of},{DS}> <{CH},{resulted},{inhibition},{of},{DS}> <{CH},{mediated},{inhibition},{of},{DS}> <{results},{CH},{inhibition},{of},{DS}> <{CH},{induced},{inhibition},{of},{DS}> <{CH},{induces},{inhibition},{of},{DS}> <{CH},{induced},{death},{DS}> <{CH},{induces},{autophagy},{DS}> <{CH},{cell},{death},{DS}> <{CH},{mediated},{DS}> <{CH},{modulate},{DS}>	<b>Rule 4</b>  <b>Code: R4</b> <u>verb + term related to cancer inhibitory activity, detected by the regular expression:</u> <i>(cause causes caused resulted results mediates mediated induces induces induced enhance enhanced modulate modulated modulates mediates mediated promote promoted)[\w]*(.*)(inhibition reduction death arrest apoptosis autophagy suppression)[\w]*(.*)(of on in)[\w]*(.*)(cancer cell neoplasm malignancy tumor tumour malignant)[\w]*</i>
<{CH},{inhibition},{of},{proliferation}> <{CH},{inhibition},{of},{DS},{growth}> <{CH},{inhibition},{of},{growth},{DS}> <{CH},{growth},{inhibition},{DS}> <{CH},{inhibition},{of},{migration}> <{CH},{inhibited},{invasion},{DS}> <{CH},{inhibited},{invasion},{of}> <{CH},{inhibited},{migration},{DS}> <{CH},{inhibited},{angiogenesis}>	<b>Rule 5</b>  <b>Code: R5</b> <u>term + preposition + term related to cancer inhibitory activity, detected by the regular expression:</u> <i>(inhibition suppression reduction inhibited mediated reduced suppressed blocked)[\w]*(.*)(of in on)[\w]*(.*)(invasion growth migration angiogenesis metastasis viability proliferation cancer carcinogenesis cells cell</i>

<{CH},{inhibited},{tumor}>	<p><i>line tumor tumour neoplasm)[\w]*</i></p> <p><i>(invasion growth migration angiogenesis metastasis viability proliferation cancer carcinogenesis cells cell line tumor tumour neoplasm)[\w]*(.*) (inhibition suppression reduction inhibited mediated reduced suppressed blocked)[\w]*</i></p>
<p>&lt;{CH},{has},{anti}&gt;</p> <p>&lt;{CH},{anti},{DS},{activity}&gt;</p> <p>&lt;{CH},{anti},{DS},{effects}&gt;</p> <p>&lt;{CH},{anti},{DS},{properties}&gt;</p> <p>&lt;{CH},{anticancer},{properties},{DS}&gt;</p> <p>&lt;{CH},{anticarcinogenic},{DS}&gt;</p> <p>&lt;{CH},{antiinvasive},{DS}&gt;</p> <p>&lt;{CH},{antiproliferative},{DS}&gt;</p> <p>&lt;{CH},{antimetastatic}&gt;</p> <p>&lt;{CH},{protective},{effect}&gt;</p> <p>&lt;{CH},{chemopreventive},{activity}&gt;</p> <p>&lt;{CH},{chemopreventive},{agent},{DS}&gt;</p> <p>&lt;{CH},{therapeutic},{agent},{DS}&gt;</p> <p>&lt;{CH},{anticancer},{agent},{DS}&gt;</p> <p>&lt;{CH},{promising},{agent},{DS}&gt;</p> <p>&lt;{inhibitory},{effect},{of},{CH}&gt;</p> <p>&lt;{inhibitory},{effects},{CH}&gt;</p> <p>&lt;{antitumor},{effect},{of},{CH}&gt;</p> <p>&lt;{anti},{effect},{of},{CH}&gt;</p> <p>&lt;{anticancer},{effect},{of},{CH}&gt;</p> <p>&lt;{antitumor},{effect},{of},{CH}&gt;</p> <p>&lt;{antiproliferative},{effect},{of},{CH}&gt;</p> <p>&lt;{CH},{potential},{agent},{treatment}&gt;</p> <p>&lt;{CH},{agent},{treatment},{DS}&gt;</p> <p>&lt;{CH},{agent},{therapy}&gt;</p> <p>&lt;{chemopreventive},{agent},{CH}&gt;</p> <p>&lt;{CH},{potential},{treatment}&gt;</p>	<p><b>Rule 6</b></p> <p><b>Code: R6</b></p> <p><u>term + term related to anticancer activity, detected by the regular expression:</u></p> <p><i>(anti)[\w]*(. *) (cancer tumor tumour neoplastic carcinogenic angiogenic angiogenesis tumorigenic metastatic metastasis proliferative oxidant invasive migration)[\w]*(. *) (effect activity activities agent propertie properties potential)[\w]*</i></p> <p><i>(effect activity activities agent propertie properties potential)[\w]*(. *) (anti)[\w]*(. *) (cancer tumor tumour neoplastic carcinogenic angiogenic angiogenesis tumorigenic metastatic metastasis proliferative oxidant invasive migration)[\w]*</i></p> <p><i>(anticancer anti-cancer anti-tumor antitumor anti-tumour antitumour anticarcinogenic anti-carcinogenic antineoplastic anti-neoplastic antiangiogenic anti-angiogenic antiangiogenesis anti-angiogenesis antimetastatic anti-metastatic antimetastasis anti-metastasis antiinvasive anti-invasive antiproliferative anti-proliferative antioxidant anti-oxidant antitumor anti-tumor proapoptotic pro-apoptotic pro apoptotic anti-tumorigenic antitumorigenic inhibitory cytotoxicity cytotoxic chemopreventive promising protective therapeutic chemotherapeutic chemotherapy preventive treatment therapy therapies radiotherapy immunotherapy prognosis prognostic)[\w]*(. *) (effect activity activities agent propertie properties potential)[\w]*</i></p> <p><i>(effect activity activities agent propertie properties potential)[\w]*(. *) (anticancer anti-cancer anti-tumor antitumor anti-tumour antitumour anticarcinogenic anti-carcinogenic antineoplastic anti-neoplastic antiangiogenic anti-angiogenic antiangiogenesis anti-angiogenesis antimetastatic anti-metastatic antimetastasis anti-metastasis antiinvasive anti-invasive antiproliferative anti-proliferative antioxidant anti-oxidant antitumor anti-tumor proapoptotic pro-apoptotic pro apoptotic anti-tumorigenic antitumorigenic inhibitory cytotoxicity cytotoxic chemopreventive promising protective therapeutic chemotherapeutic chemotherapy preventive treatment therapy therapies radiotherapy immunotherapy prognosis prognostic)[\w]*</i></p>

<{CH},{shown},{activity},{DS}> <{CH},{shown},{effects},{DS}> <{CH},{exhibited},{activity}> <{CH},{exhibited},{effects},{DS}> <{CH},{exhibited},{potent},{DS}>	<b>Rule 7</b>  <b>Code: R7</b> <u>verb + term related to anticancer activity, detected by the regular expression:</u> (exhibit exhibited exhibits shown demonstrated present has have enhanced enhances reported possesses)[\w]*(.*) (effect activity activities potent propertie properties potential cytotoxicity cytotoxic inhibitory)[\w]*
<{promoted},{by}> <{apoptosis},{induced},{by},{CH}> <{death},{induced},{by},{CH}>	<b>Rule 8</b>  <b>Code: R8</b> <u>term + verb + preposition related to anticancer activity, detected by the regular expression:</u> (inhibition reduction death arrest apoptosis autophagy suppression)[\w]*(.*) (caused resulted mediated induced enhanced promoted)[\w]*(.*) (by)[\w]*
<{inhibited},{by},{CH}> <{inhibition},{by},{CH},{DS}> <{DS},{inhibition},{by},{CH}> <{DS},{suppressed},{by},{CH}> <{DS},{reduced},{by},{CH}> <{growth},{inhibited},{by},{CH}>	<b>Rule 9</b>  <b>Code: R9</b> <u>term + verb + preposition related to anticancer activity, detected by the regular expression:</u> (activity viability invasion growth migration angiogenesis metastasis viability proliferation cancer carcinogenesis cells cell line tumor tumour neoplasm)[\w]*(.*) (inhibited mediated reduced suppressed blocked)[\w]*(.*) (by)[\w]*
<{inhibition},{of},{DS},{by},{CH}> <{proliferation},{of},{DS},{by},{CH}> <{CH},{arrest},{of},{by}> <{CH},{apoptosis},{of},{by}> <{suppression},{of},{by},{CH}>	<b>Rule 10</b>  <b>Code: R10</b> <u>term + preposition + preposition related to anticancer activity, detected by the regular expression:</u> (inhibition reduction death arrest apoptosis autophagy suppression migration metastasis viability proliferation)[\w]*(.*) (of)[\w]*(.*) (by)[\w]*

#### Rules for Associations of Genes and Polyphenols

Sequences (patterns)	Rules and regular expressions
<{CH},{increased},{GN}> <{CH},{upregulate},{GN}> <{CH},{regulate},{GN}> <{CH},{reduced},{GN}> <{CH},{block},{GN}>	<b>Rule 11</b> <b>Code: R11</b> <u>verb related to the regulatory activity of genes, detected by the regular expression:</u> (disruption regulation abolished repressed stimulated regulate regulated regulates downregulate downregulates downregulated upregulate upregulated upregulates downregulate down-regulates down-regulated up-regulate up-regulated up-regulates down regulate down regulates down regulated up regulate up regulated up regulates reduce reduced reduces block blocks blocked increase increases increased decreases decreased decrease induce induced induces inhibit inhibited inhibits suppress suppressed suppresses enhanced attenuated activation)[\w]*
<{CH},{increased},{expression},{GN}> <{CH},{increased},{levels},{GN}> <{CH},{decreased},{levels},{GN}> <{CH},{decreased},{expression},{GN}>	<b>Rule 12</b> <b>Code: R12</b> <u>verb + terms related to the regulatory activity of genes, detected by the regular expression:</u>

<{CH},{downregulated},{expression}> <{CH},{decrease},{expression},{GN}> <{CH},{downregulate},{expression},{GN}> > <{CH},{reduced},{expression},{GN}> <{CH},{blocked},{activation},{GN}> <{CH},{induced},{activation},{GN}> <{CH},{inhibited},{activation},{of}> <{CH},{GN},{mediated},{activation}> <{CH},{enhanced},{expression}> <{CH},{resulted},{expression}> <{CH},{upregulated},{expression},{GN}> <{CH},{attenuated},{expression}> <{CH},{DS},{signaling}> <{signaling},{pathway}> <{signaling},{in},{DS}> <{induces},{signaling}> <{CH},{abolished},{GN}> <{CH},{repressed},{GN}> <{CH},{stimulated},{GN}> <{GN},{activity},{inhibited},{by},{CH}> <{GN},{expression},{by},{CH}> <{activation},{by},{CH}> <{induction},{by},{CH}> <{CH},{inhibition},{of},{GN},{signaling}> - <{CH},{inhibition},{of},{GN},{signaling}> <{CH},{inhibit},{signaling}> <{CH},{inhibition},{of},{GN},{pathway}> - <{CH},{inhibit},{pathway}> <{CH},{disruption},{of},{GN}>	(induction inhibition regulation abolished repressed stimulated regulate regulated regulates downregulate downregulates downregulated upregulate upregulated upregulates down-regulate down-regulates down-regulated up-regulate up-regulated up-regulates down regulate down regulates down regulated up regulate up regulated up regulates reduce reduced reduces block blocks blocked increase increases increased decreases decreased decrease induce induced induces inhibit inhibited inhibits suppress suppressed suppresses enhanced attenuated activation via pathway protein signaling kinase gene mRNA miRNA microRNA activity)[\\w]*  (expression levels activation via pathway protein signaling kinase gene mRNA miRNA microRNA activity)[\\w]*.(*) (induction inhibition regulation abolished repressed stimulated regulate regulated regulates downregulate downregulates downregulated upregulate upregulated upregulates down-regulate down-regulates down-regulated up-regulate up-regulated up-regulates down regulate down regulates down regulated up regulate up regulated up regulates reduce reduced reduces block blocks blocked increase increases increased decreases decreased decrease induce induced induces inhibit inhibited inhibits suppress suppressed suppresses enhanced attenuated activation disruption)[\\w]*
--	--

#### Rules for Specific Markers

Sequences (patterns)	Rules and regular expressions
<{CH},{epigenetic},{DS}> <{CH},{histone},{deacetylase}> <{CH},{histone},{acetylation}> <{acetylated}> <{CH},{phosphorylated},{GN}> <{CH},{phosphorylation},{GN}> <{CH},{methylation},{GN}> <{CH},{demethylation}> <{hypermethylation}> <{methylated}> <{microRNA}> <{CH},{miRNA}>	<b>Rule 13</b> <b>(Epigenetic Marker)</b> <b>Code: R3</b> <u>term related to epigenetic regulatory activity, detected by the regular expression:</u> (epigenetic histone methylation methylated miRNA microRNA miRNA phosphorylated phosphorylation acetylated acetylation acetylase)[\\w]*
<{CH},{novel},{strategy},{DS}> <{CH},{novel},{agent},{DS}> <{novel},{synthetic},{CH},{DS}> <{CH},{novel},{drug}>	<b>Rule 14</b> <b>(Novel Anticancer Drug Marker)</b> <b>Code: R14</b> (novel new protective potent effective promising)[\\w]*.(*) (compound agent drug compost synthetic strategy)[\\w]*.(*) (chemopreventive cells cell line tumor cell anticancer anti-cancer anti-tumor antitumor anti-fumour antitumour anticarcinogenic anti-carcinogenic antineoplastic anti-neoplastic antiangiogenic anti-angiogenic antiangiogenesis anti-angiogenesis antimetastatic anti-metastatic antimetastasis anti-metastasis antiinvasive anti-invasive antiproliferative anti-proliferative antioxidant anti-oxidant antitumor anti-tumor proapoptotic pro-apoptotic pro



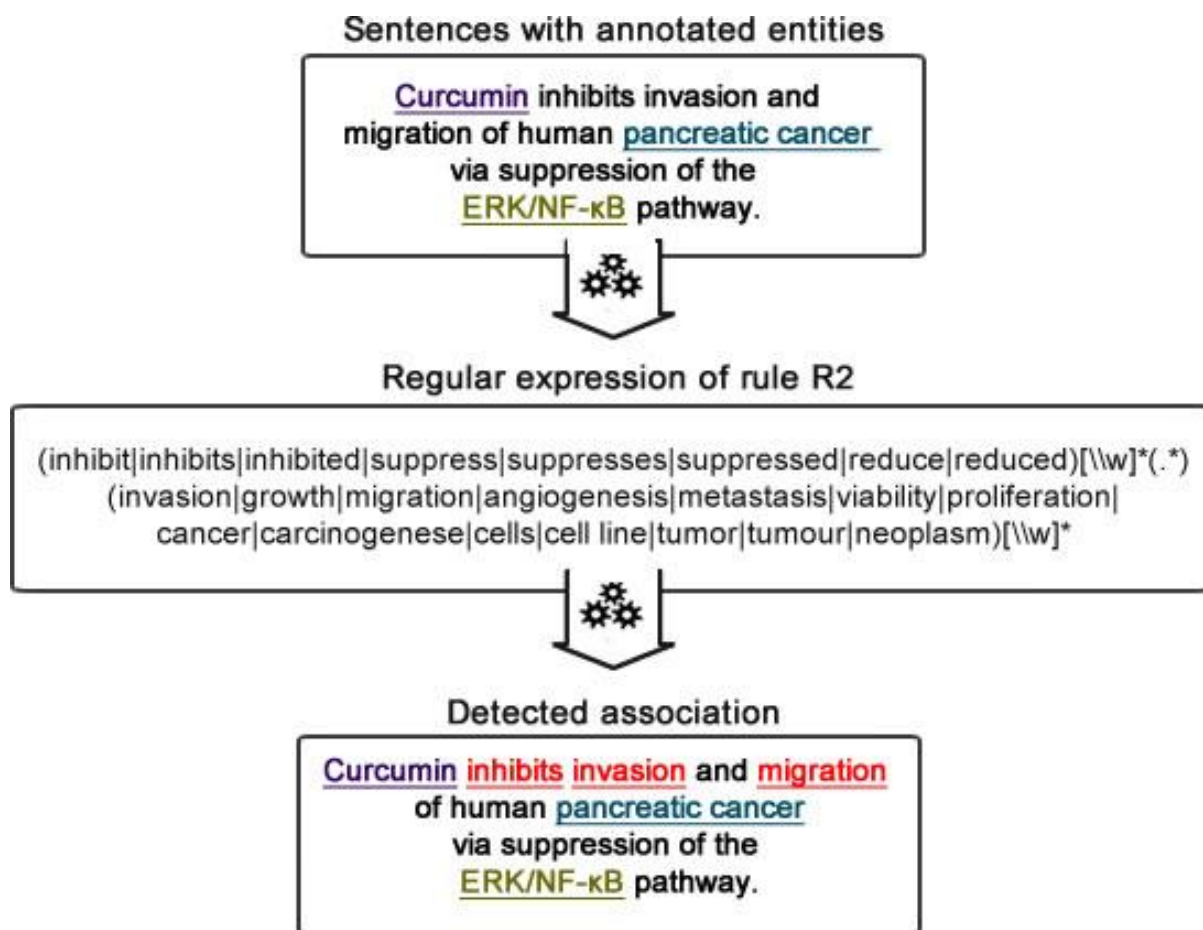
	<i>apoptotic anti-tumorigenic antitumorigenic inhibitory cytotoxicity cytotoxic chemopreventive promising protective therapeutic chemotherapeutic chemotherapy preventive treatment therapy therapies radiotherapy immunotherapy prognosis prognostic modulators </i> [\\w]*
<{CH},{DS},{cells}> <{CH},{cell},{lines}> <{CH},{tumor},{cells}> <{CH},{tumor},{cell}>	<b>Rule 15</b> <b>(Marker Cell Line)</b> <b>Code: R15</b> <i>(cell line cells tumor cell neoplasm leukemia leukaemia carcinogenesis tumorigenesis metastasis sarcoma carcinoma blastoma) </i> [\\w]*
<{findings},{suggest},{CH}> <{findings},{indicate},{CH},{of},{DS}> > <{studies},{revealed},{CH}> <{CONCLUSION},{CH}>	<b>Rule 16</b> <b>(Marker Result/Conclusion)</b> <b>Code: R16</b> <i>(conclusion) </i> [\\w]* <i>(results our result findings conclusion data taken together study studies) </i> [\\w]* <i>(.)*(suggest indicate suggested indicated illustrated illustrate revealed reported resulted show) </i> [\\w]*

#### Rules for Cancer Markers

Sequences (patterns)	Rules and regular expressions
<{CH},{proliferation},{DS}> <{cell},{proliferation},{DS}> <{growth},{factor}> <{CH},{DS},{cell},{growth}> <{CH},{growth},{DS}> <{CH},{growth},{of},{cells}> <{growth},{cells}> <{growth},{inhibition},{DS}> <{antiproliferative},{against},{DS}>	1. Self-sufficiency in growth signals 2. Insensitivity to antiproliferative signals <b>Code: HM12</b> <i>(inhibition inhibits inhibited inhibit blocks blocked suppression suppress suppressed suppresses attenuate attenuated attenuates reduction reduce reduces reduced anti) </i> [\\w]* <i>(.)*(proliferation growth factor growthfactor growth-factor cellgrowth cell growth proliferative) </i> [\\w]*  <i>(proliferation growth factor growthfactor growth-factor cellgrowth cell growth proliferative) </i> [\\w]* <i>(.)*(inhibition inhibited blocked suppression suppressed attenuate attenuated attenuates reduction reduced) </i> [\\w]*
<{CH},{apoptosis},{in},{DS},{cells}> <{CH},{cell},{cycle},{arrest},{DS}> <{CH},{cycle},{arrest},{DS}> <{apoptotic},{death},{cells}> <{apoptotic},{in},{DS}> <{CH},{apoptotic},{cell},{death}> <{CH},{proapoptotic},{DS},{cells}> <{CH},{necrosis}> <{CH},{autophagy},{in},{DS}> <{CH},{death},{of},{DS},{cells}>	3. Evasion of apoptosis <b>Code: HM3</b> <i>(cause causes caused pro resulted induces induces induced enhance enhanced modulate modulated modulates mediates mediated promote promoted) </i> [\\w]* <i>(.)*(apoptosis apoptotic autophagy autophagic necrose necrosis) </i> [\\w]*  <i>(apoptosis apoptotic autophagy autophagic necrose necrosis) </i> [\\w]* <i>(.)*(caused resulted induced induction enhanced modulated mediated promotion promoted) </i> [\\w]*  <i>(cause causes caused pro resulted induces induces induced enhance enhanced modulate modulated modulates mediates mediated promote promoted) </i> [\\w]* <i>(.)*(cycle cell) </i> [\\w]* <i>(.)*(arrest death) </i> [\\w]*  <i>(cycle cell) </i> [\\w]* <i>(.)*(arrest death) </i> [\\w]* <i>(.)*(caused resulted induced induction enhanced modulated mediated promotion promoted) </i> [\\w]*
<{CH},{senescence}> <{CH},{telomerase},{DS}> <{CH},{immortalized}>	4. Unlimited replication potential <b>Code: HM4</b> <i>(senescence telomerase immortalized) </i> [\\w]*
<{CH},{angiogenesis},{DS}> <{CH},{inhibits},{angiogenesis}>	5. Induction of angiogenesis <b>Code: HM5</b>

<{antiangiogenic},{CH}>	<i>(inhibition inhibits inhibited inhibit blocks blocked suppression suppress suppressed suppresses attenuate attenuated attenuates reduction reduce reduces reduced anti)[\w]*.*)(angiogenesis angiogenic)[\w]*</i>  <i>(angiogenesis angiogenic)[\w]*.*)(inhibition inhibited blocked suppression suppressed attenuate attenuated attenuates reduction reduced)[\w]*</i>
<{CH},{metastatic},{DS}> <{CH},{metastasis},{DS}> <{CH},{invasion},{metastasis}> <{antimetastatic},{DS}> <{CH},{migration},{invasion},{DS}> <{CH},{migration},{of},{DS}> <{migration},{by},{DS}>	6. Activation of invasion and metastasis <b>Code: HM6</b> <i>(inhibition inhibits inhibited inhibit blocks blocked suppression suppress suppressed suppresses attenuate attenuated attenuates reduction reduce reduces reduced anti)[\w]*.*)(metastasis motility invasion migration metastatic migratory invasive invasiveness)[\w]*</i>  <i>(metastasis motility invasion migration metastatic migratory invasive invasiveness)[\w]*.*)(inhibition inhibited blocked suppression suppressed attenuate attenuated attenuates reduction reduced)[\w]*</i>
<{DNA},{damage}> <{DNA},{damage},{DS},{cells}> <{DNA},{damage},{cells}> <{damage},{cells}> <{CH},{DNA},{fragmentation},{DS}>	7. Mutation and genomic instability <b>Code: HM7</b> <i>(mutation damage fragmentation)[\w]*</i>  <i>(dna cell)[\w]*.*)(repair damage fragmentation)[\w]*</i>
<{CH},{has},{antiinflammatory}> <{CH},{antiinflammatory},{DS}> <{CH},{antioxidant},{DS}> <{antioxidant},{in},{DS}> <{antioxidant},{in},{cells}> <{antioxidant},{cells}> <{antiinflammatory},{activities}> <{CH},{oxidative},{stress}>	8. Tumor-induced inflammation <b>Code: HM8</b> <i>(inhibition inhibits inhibited inhibit blocks blocked suppression suppress suppressed suppresses attenuate attenuated attenuates reduction reduce reduces reduced anti)[\w]*.*)(inflammation inflammatory oxidative oxidation oxidant)[\w]*</i> <i>(inflammation inflammatory oxidative oxidation oxidant)[\w]*.*)(inhibition inhibited blocked suppression suppressed attenuate attenuated attenuates reduction reduced)[\w]*</i>
<{CH},{metabolism},{DS}> <{CH},{metabolic},{DS}> <{CH},{glycolysis}> <{CH},{mitochondrial}>	9. Cellular energy dysregulation <b>Code: HM9</b> <i>(metabolism metabolic glycolysis mitochondrial)[\w]*</i>
<{CH},{immune},{response}> <{CH},{DS},{immune}>	10. Immune evasion <b>Code: HM10</b> <i>(immune immunosuppression)[\w]*</i>

Supplementary Figure 1 illustrates as an example the recognition of rule R1 in the sentence “Curcumin inhibits invasion and migration of human pancreatic cancer via suppression of the ERK/NF-kB pathway”, through a regular expression. The regular expression created for the detection of sentences associated with rule R2 searches for sentences that contain “verbs related to inhibitory activity + terms related to carcinogenesis”. In this example, the sentence was detected using the verb “inhibits” and the terms “invasion” and “migration”.



**Source:** The author.

Supplementary Figure 1 – Rule R2 being recognized in a sentence containing a polyphenol-cancer association (curcumin-pancreatic\_cancer)

Sequential pattern mining was important for identifying the order in which the terms, verbs, entities, and prepositions occur in the sentences (1,2) which, in turn, contributed to the creation of regular expressions (7-9). The absence of stemming during pre-processing of the sentences allowed to identify specific patterns with the appropriate variations of words (e.g., inhibit, inhibited, inhibits, inhibition). Frequent unique patterns with a support higher than 0.01, such as cells (0.396), expression. (0.144), apoptosis (0.125), activity (0.098), induced (0.094), effects (0.091), growth (0.082), treatment (0.074), inhibition (0.069), activation (0.066), inhibited (0.063), proliferation (0.060), pathway (0.049), potential (0.043), against (0.042), and levels (0.042) were important as keywords in the search for patterns of entity associations (polyphenol-cancer, polyphenol-gene) and for the creation of rules. The verbs related to the gene-microRNA molecular interaction in the study of (6) were also important as keywords in the search for patterns and for the elaboration of rules R1, R2, R4, R8, R9, R11 and R12.

The keywords related to cancer markers obtained from the studies of Bundela et al. (5) and Baker et al. (4) were important for the elaboration of rules HM12 to HM10. As reported by



Tari et al. (2) and Gupta et al. (6), the use of prepositions (“of”, “by” and “on”) was important for the elaboration of the rules used to extract sentences with active (R4 and R10) and passive (R8 and R9) associations between entities. Among the specific rules, R15 was created to minimize some perceived problems resulting from inconsistencies in the NER process (e.g., in PMID 24300195, PubTator did not identify the A549 cell line [lung cancer]). Thus, rule R15 identifies sentences containing some cell line or cancer-related term. Rule R16 was created to capture sentences related to the “results, discussion or conclusion” sections of abstracts since these sections are more likely to contain information (10,11).

## REFERENCES

1. Cellier, P., Charnois, T., Plantevit, M., Rigotti, C., Cremilleux, B., Gandrillon, O., Klema, J. and Manguin, J.L. (2015) Sequential pattern mining for discovering gene interactions and their contextual information from biomedical texts. *J Biomed Semantics*, **6**, 27.
2. Tari, L., Anwar, S., Liang, S., Cai, J. and Baral, C. (2010) Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. *Bioinformatics*, **26**, i547-553.
3. Zaki, M.J. (2001) SPADE: An Efficient Algorithm for Mining Frequent Sequences. *Machine Learning*, **42**, 31-60.
4. Baker, S., Silins, I., Guo, Y., Ali, I., Hogberg, J., Stenius, U. and Korhonen, A. (2016) Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, **32**, 432-440.
5. Bundela, S., Sharma, A. and Bisen, P.S. (2014) Potential therapeutic targets for oral cancer: ADM, TP53, EGFR, LYN, CTLA4, SKIL, CTGF, CD70. *PLoS One*, **9**, e102610.
6. Gupta, S., Dingerdissen, H., Ross, K.E., Hu, Y., Wu, C.H., Mazumder, R. and Vijay-Shanker, K. (2018) DEXTER: Disease-Expression Relation Extraction from Text. *Database (Oxford)*, **2018**.
7. Fang, Y.C., Lai, P.T., Dai, H.J. and Hsu, W.L. (2011) MeInfoText 2.0: gene methylation and cancer relation extraction from biomedical literature. *BMC Bioinformatics*, **12**, 471.
8. Hobbs, J.R. and Riloff, E. (2010) *Information extraction*. 2 ed. Chapman & Hall/CRC, Cambridge, UK.
9. Hua, Y., Yu, H., Zhenwei, H., Jianmin, Y., Mingming, Z. and Yanhui, F. (2011) Combination Method of Rules and Statistics for Abbreviation and Its Full Name Recognition. *Advances in Intelligent and Soft Computing*, **110**, 707-714.
10. Guo, Y., Korhonen, A., Liakata, M., Silins, I., Hogberg, J. and Stenius, U. (2011) A comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment. *BMC Bioinformatics*, **12**, 69.
11. Lee, H.J., Dang, T.C., Lee, H. and Park, J.C. (2014) OncoSearch: cancer gene search engine with literature evidence. *Nucleic Acids Res*, **42**, W416-421.