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 Cellier et al. (2015), 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[1], 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 (TARI et al., 2010; CELLIER et al., 2015).

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 (ZAKI, 2001) 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 (≈30 sentences) were extracted. Patterns containing only one transaction and support ≥ 0.01 (≈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 (BUNDELA; SHARMA; BISEN, 2014; BAKER et al., 2016; GUPTA et al., 2016) were used for the selection of patterns with associations between entities (e.g., <{CH&},{inhibited},{proliferation},{D&S}>, <{CH&}, {inhibits},{invasion},{D&S}>, <{CH},{increased},{expression},{GN}>).

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 (HOBBS; RILOFF, 2010; HUA et al., 2011; FANG et al., 2011). Table 1 lists the rule specifications with the selected patterns and keywords, as well as the regular expressions created.

[1] PubMed HELP – Available at <https://www.ncbi.nlm.nih.gov/books/NBK3827/pdf/Bookshelf_NBK3827.pdf>. Accessed 6 November 2018

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}>	**Rule 1** **Code: R1** verb related to cancer inhibitory activity, detected by the regular expression: *(inhibits|inhibited|inhibit|blocks|blocked|suppress|suppressed|suppresses|attenuate|attenuated|attenuates|reduce|reduces|reduced|prevent|prevents|prevented)[\\w]**
<{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}>	**Rule 2** **Code: R2** verb + term related to cancer inhibitory activity, detected by the regular expression: *(inhibit|inhibits|inhibited|suppress|suppresses|suppressed|reduce|reduced)[\\w]*(.*)(invasion|growth|migration|angiogenesis|metastasis|viability|proliferation|cancer|carcinogenese|cells|cell line|tumor|tumour|neoplasm)[\\w]**
<{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}>	**Rule 3** **Code: R3** term + term related to anticancer activity, detected by the regular expression: (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]**
<{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}>	**Rule 4** **Code: R4** verb + term related to cancer inhibitory activity, detected by the regular expression: *(cause|causes|caused|result|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]**
<{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}>	**Rule 5** **Code: R5** term + preposition + term related to cancer inhibitory activity, detected by the regular expression: *(*inhibition|*suppression|reduction|inhibited|mediated|reduced|suppressed|blocked)[\\w]*(.*)*(of|*in|on*)[\\w]*(.*)*(invasion|growth|migration|angiogenesis|metastasis|viability|proliferation|cancer|carcinogenese|cells|cell*

<{CH},{inhibited},{tumor}>	*line|tumor|tumour|neoplasm)[\\w]** *(invasion|growth|migration|angiogenesis|metastasis|viability|proliferation|cancer|carcinogenese|cells|cell line|tumor|tumour|neoplasm)[\\w]*(.*)(*inhibition|*suppression|reduction|inhibited|mediated|reduced|suppressed|blocked)[\\w]**
<{CH},{has},{anti}> <{CH},{anti},{DS},{activity}> <{CH},{anti},{DS},{effects}> <{CH},{anti},{DS},{properties}> <{CH},{anticancer},{properties},{DS}> <{CH},{anticarcinogenic},{DS}> <{CH},{antiinvasive},{DS}> <{CH},{antiproliferative},{DS}> <{CH},{antimetastatic}> <{CH},{protective},{effect}> <{CH},{chemopreventive},{activity}> <{CH},{chemopreventive},{agent},{DS} <{CH},{therapeutic},{agent},{DS}> <{CH},{anticancer},{agent},{DS}> <{CH},{promising},{agent},{DS}> <{inhibitory},{effect},{of},{CH}> <{inhibitory},{effects},{CH}> <{antitumor},{effect},{of},{CH}> <{anti},{effect},{of},{CH}> <{anticancer},{effect},{of},{CH}> <{antitumor},{effect},{of},{CH}> <{antiproliferative},{effect},{of},{CH}> <{CH},{potential},{agent},{treatment}> <{CH},{agent},{treatment},{DS}> <{CH},{agent},{therapy}> <chemopreventive},{agent},{CH}> <{CH},{potential},{treatment}>	**Rule 6** **Code: R6** <u>term + term related to anticancer activity, detected by the regular expression</u>: *(*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]** *(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]* *(*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 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]** *(effect|activity|activities|agent|propertie|properties|potential)[\\w]**(.*)*(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 apoptotic|anti-tumorigenic|antitumorigenic|inhibitory|cytotoxicity|cytotoxic*|chemopreventive|*promising|protective|therapeutic|chemotherapeutic|chemotherapy|preventive|treatment|therapy|therapies|radiotherapy|immunotherapy|prognosis|prognostic)*[\\w]*

<{CH},{shown},{activity},{DS}> <{CH},{shown},{effects},{DS}> <{CH},{exhibited},{activity}> <{CH},{exhibited},{effects},{DS}> <{CH},{exhibited},{potent},{DS}>	**Rule 7** **Code: R7** verb + term related to anticancer activity, detected by the regular expression: (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}>	**Rule 8** **Code: R8** term + verb + preposition related to anticancer activity, detected by the regular expression: *(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}>	**Rule 9** **Code: R9** term + verb + preposition related to anticancer activity, detected by the regular expression: *(activity*|*viability*|*invasion*|*growth*|*migration*|*angiogenesis*|*metastasis*|*viability*|*proliferation*|*cancer*|*carcinogenese*|*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}>	**Rule 10** **Code: R10** term + preposition + preposition related to anticancer activity, detected by the regular expression: *(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}>	**Rule 11** **Code: R11** verb related to the regulatory activity of genes, detected by the regular expression: *(disruption*|*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*|*active*|*activation)[\\w]**
<{CH},{increased},{expression},{GN}> <{CH},{increased},{levels},{GN}> <{CH},{decreased},{levels},{GN}> <{CH},{decreased},{expression},{GN}>	**Rule 12** **Code: R12** verb + terms related to the regulatory activity of genes, detected by the regular expression:

<{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|active|activation|disruption)[\\w]*(.*)(expression|levels|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}>	**Rule 13** **(Epigenetic Marker)** **Code: R3** term related to epigenetic regulatory activity, detected by the regular expression: *(epigenetic|histone|methylation|methylated|mirna|microrna|mi-rna|phosphorylated|phosphorylation|acetylated|acetylation|acetylase)[\\w]**
<{CH},{novel},{strategy},{DS}> <{CH},{novel},{agent},{DS}> <{novel},{synthetic},{CH},{DS}> <{CH},{novel},{drug}>	**Rule 14** **(Novel Anticancer Drug Marker)** **Code: R14** *(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*

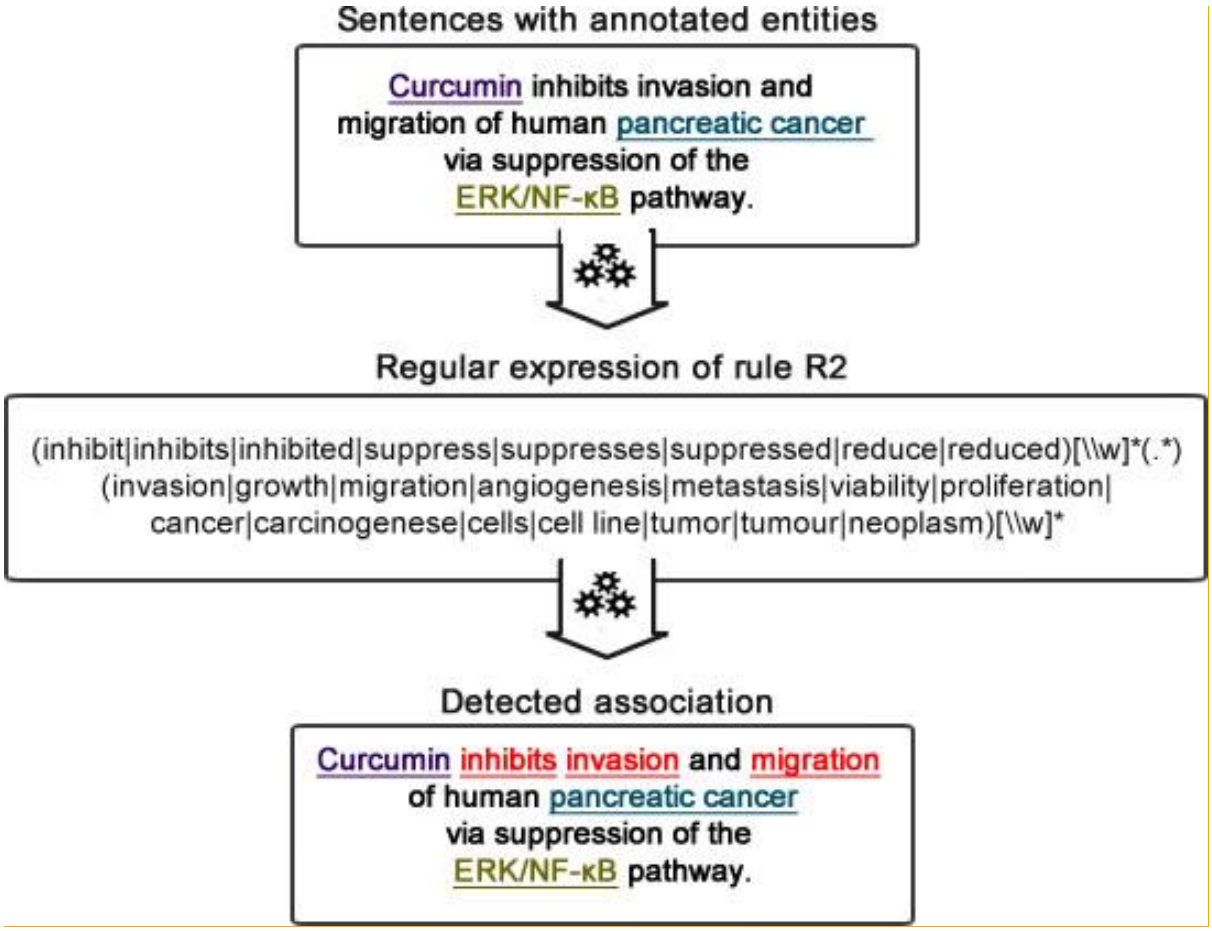
	*apoptotic|anti-tumorigenic|antitumorigenic|inhibitory|cytotoxicity|cytotoxic|chemopreventive|promising|protective|therapeutic|chemotherapeutic|chemotherapy|preventive|treatment|therapy|therapies|radiotherapy|immunotherapy|prognosis|prognostic|modulators)[\\w]**
<{CH},{DS},{cells}> <{CH},{cell},{lines}> <{CH},{tumor},{cells}> <{CH},{tumor},{cell}>	**Rule 15** **(Marker Cell Line)** **Code: R15** *(cell line|cells|tumor cell|neoplasm|leukemia|leukaemia|carcinogenesis|tumorigenesis|metastasis|sarcoma|carcinoma|blastoma)[\\w]**
<{findings},{suggest},{CH}> <{findings},{indicate},{CH},{of},{DS}> <{studies},{revealed},{CH}> <{CONCLUSION},{CH}>	**Rule 16** **(Marker Result/Conclusion)** **Code: R16** (*conclusion*)[\\w]**(results|our result|findings|conclusion|data|taken|together|study|studies)[\\w]*(.*)(suggest|indicate|suggested|indicated|illustrated|illustrate|revealed|reported|resulted|show)[\\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 **Code: HM12** *(inhibition|inhibits|inhibited|inhibit|blocks|blocked|suppression|suppress|suppressed|suppresses|attenuate|attenuated|attenuates|reduction|reduce|reduces|reduced|anti)[\\w]**(.*)(proliferation|growth factor|growthfactor|growth-factor|cellgrowth|cell growth|proliferative)[\\w]** *(proliferation|growth factor|growthfactor|growth-factor|cellgrowth|cell growth|proliferative)[\\w]**(.*)(inhibition|inhibited|blocked|suppression|suppressed|attenuate|attenuated|attenuates|reduction|reduced)[\\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 **Code: HM3** *(cause|causes|caused|pro|resulted|induces|induces|induced|enhance|enhanced|modulate|modulated|modulates|mediates|mediated|promote|promoted)*[\\w]*(.*)(apoptosis|apoptotic|autophagy|autophagic|necrose|necrosis)[\\w]** (*apoptosis|apoptotic|autophagy|autophagic|necrose|necrosis)[\\w]**(.*)*(caused|resulted|induced|induction|enhanced|modulated|mediated|promotion|promoted)*[\\w]* *(cause|causes|caused|pro|resulted|induces|induces|induced|enhance|enhanced|modulate|modulated|modulates|mediates|mediated|promote|promoted)*[\\w]*(.*)(cycle|cell)*[\\w]*(.*)(*arrest|death*)[\\w]* *(cycle|cell)*[\\w]*(.*)(*arrest|death*)[\\w]*(.*)*(caused|resulted|induced|induction|enhanced|modulated|mediated|promotion|promoted)*[\\w]*
<{CH},{senescence}> <{CH},{telomerase},{DS}> <{CH},{immortalized}>	4. Unlimited replication potential **Code: HM4** *(senescence|telomerase|immortalized)[\\w]**
<{CH},{angiogenesis},{DS}> <{CH},{inhibits},{angiogenesis}>	5. Induction of angiogenesis **Code: HM5**

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

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.

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 (TARI et al., 2010; CELLIER et al., 2015) which, in turn, contributed to the creation of regular expressions (HOBBS; RILOFF, 2010; HUA et al., 2011; FANG et al., 2011). 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 Gupta et al. (2016) 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, Sharma and Bisen (2014) and Baker et al. (2016) were important for the elaboration of rules HM12 to HM10. As reported by Tari et al. (2010) and Gupta et al. (2016), 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 (GUO et al., 2011; LEE et al.,2014).

REFERENCES

1. Baker, S. et al. (2016) Automatic semantic classification of scientific literature according to the hallmarks of cancer. Bioinformatics*,* Oxford, 32:3, 432-440.
2. Bundela, S., Sharma, A., Bisen, P.S. (2014) Potential Therapeutic Targets for Oral Cancer: ADM,TP53, EGFR, LYN, CTLA4, SKIL, CTGF, CD70. Plos One, 9(7), 1-15.
3. Cellier et al. (2015) Sequential pattern mining for discovering gene interactions and their contextual information from biomedical texts. Journal of Biomedical Semantics, 6:27, 1-12.
4. Fang, Y. at al. (2011) MeInfoText 2.0: gene methylation and cancer relation extraction from biomedical literature. BMC Bioinformatics, 12:471, 1-8.
5. Gupta, S. et al. (2018) DEXTER: Disease-Expression Relation Extraction from Text. Database, Oxford, 2018, 1-17.
6. Guo, Y. et al. (2011) A comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment. BMC Bioinformatics, 12:69, 1-18.
7. Hobbs, J.R., Riloff, E. (2010) Information extraction. In: Indurkhya N, Damerau FJ, editors. Handbook of Natural Language Processing. Cambridge, UK : Chapman and Hall/CRC, 2nd edition, 511 – 532.
8. Hua, Y. et al. (2011) Combination method of rules and statistics for abbreviation and its full name recognition. Advances in Intelligent and Soft Computing, 110, 707-714.
9. Lee, H.J. et al. (2014) OncoSearch: cancer gene search engine with literature evidence. Nucleic Acids Research*,* 1-6.
10. Tari, L. et al. (2010) Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. Bioinformatics, 26, i547-i553.
11. Zaki, M.J. (2001) SPADE: An efficient algorithm for mining frequent sequences. Machine Learning, 42, 1-2, 31-60.