Bioinformatics, 32(3), 2016, 432–440
doi: 10.1093/bioinformatics/btv585
Advance Access Publication Date: 9 October 2015
Original Paper

OXFORD

Data and text mining

Automatic semantic classification of scientific literature according to the hallmarks of cancer

Simon Baker[1,*], Ilona Silins[2], Yufan Guo[3], Imran Ali[2], Johan Högberg[2], Ulla Stenius[2] and Anna Korhonen[3]

[1]Computer Laboratory, University of Cambridge, Cambridge, UK, [2]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden and [3]Department of Theoretical and Applied Linguistics, University of Cambridge, Cambridge, UK

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on May 29, 2015; revised on September 11, 2015; accepted on September 28, 2015

Abstract

Motivation: The *hallmarks of cancer* have become highly influential in cancer research. They reduce the complexity of cancer into 10 principles (e.g. resisting cell death and sustaining proliferative signaling) that explain the biological capabilities acquired during the development of human tumors. Since new research depends crucially on existing knowledge, technology for semantic classification of scientific literature according to the hallmarks of cancer could greatly support literature review, knowledge discovery and applications in cancer research.

Results: We present the first step toward the development of such technology. We introduce a corpus of 1499 PubMed abstracts annotated according to the scientific evidence they provide for the 10 currently known hallmarks of cancer. We use this corpus to train a system that classifies PubMed literature according to the hallmarks. The system uses supervised machine learning and rich features largely based on biomedical text mining. We report good performance in both intrinsic and extrinsic evaluations, demonstrating both the accuracy of the methodology and its potential in supporting practical cancer research. We discuss how this approach could be developed and applied further in the future.

Availability and implementation: The corpus of hallmark-annotated PubMed abstracts and the software for classification are available at: http://www.cl.cam.ac.uk/~sb895/HoC.html .

Contact: simon.baker@cl.cam.ac.uk

1 Introduction

Cancer figures among the leading causes of mortality worldwide, with c. 14M new cases and 8.2M cancer-related deaths reported in 2012 (Stewart and Wild, 2014). The number of new cases is expected to rise by c. 70% over the next two decades, making it more important than ever to develop effective tools for prevention, detection and treatment of this disease. New research into cancer draws on existing knowledge reported in scientific literature. Relevant literature has grown rapidly in both size and complexity. There are over 3M citations related to 'cancer' in PubMed (www.ncbi.nlm.nih.gov/pubmed). As many as 151 872 were added over the past year. The high number of potentially relevant articles is a valuable source on to which new research can build but at the same time poses a challenge for scientists. While applications such as Google Scholar (scholar.google.com) and PubMed Advanced Search (www.ncbi.nlm.nih.gov/pubmed/advanced) can be of help, they fall short in providing all and only the information of interest. To collect relevant articles, keyword-based queries are the most common approach for literature retrieval. However, because of the complexity of scientific data in cancer research, the massive number of keywords, their synonyms and combinations exceeds what researchers can realistically handle. For example, a cancer researcher would not find all the literature about 'sustaining proliferative signaling' by searching for 'proliferative signaling'. Rather, the use of hundreds of search terms

(e.g. 'growth factors', 'growth factor receptors' and 'cell cycle') would be required, along with manual filtering of the search results. It is an extremely time-consuming task for researchers to read, interpret, select and structure data in an organized manner.

Text mining (TM) can provide more targeted methodology for identifying relevant information in scientific literature. Past decades have seen a great development in biomedical TM that has made large-scale information extraction (IE) and knowledge discovery from literature possible and yielded impressive results in real-life tasks (McDonald *et al.*, 2012; Simpson and Demner-Fushman, 2012). To date, the main emphasis of cancer-centric TM work has been on tasks such as text classification and IE [e.g. named entity recognition (NER), relation and event extraction] (Spasic *et al.*, 2014; Zhua *et al.*, 2011). Although only some of this work has been evaluated in context of real-life cancer research, its enormous promises are evidenced by studies which have revealed new scientific knowledge in text-mined information that are not observable by informal data inspection (Korhonen *et al.*, 2012).

In this article, we introduce a TM technique for supporting semantic classification of scientific literature for cancer research. Our classification is based on the *hallmarks of cancer*. First introduced by Hanahan and Weinberg in an article that has been cited over 20 000 times (Google Scholar, March 2015) (Hanahan and Weinberg, 2000), hallmarks are now widely employed in cancer research. A complex disease, cancer involves genetic and epigenetic alterations that affect a large number of genes, proteins and signaling networks during tumor progression (Marusyk *et al.*, 2012). Ten characteristics (i.e. hallmarks) of normal cells required for malignant growth have been proposed that provide an organizing principle to simplify the diversity of the biological processes leading to cancer. These include (i) sustaining proliferative signaling, (ii) evading growth suppressors, (iii) enabling replicative immortality, (iv) activating invasion and metastasis, (v) inducing angiogenesis, (vi) resisting cell death, (vii) deregulating cellular energetics, (viii) avoiding immune destruction, (ix) genome instability and mutation and (x) tumor-promoting inflammation) (Hanahan and Weinberg, 2011). Rationalizing the complexity in the underlying biological processes, hallmarks can help researchers gain a better understanding of the cellular events leading to cancer. The ability to identify important pathways within one or several hallmarks may also lead to the development of e.g. more effective cancer drugs (Hanahan and Weinberg, 2011).

Within the bioinformatics community, hallmarks have inspired harvesting interactions between genes and proteins and relations between environment and cancer from high-throughput omics data and scientific literature. The latter work has led into the development of ontology tools [e.g. OncoCL (Doland, 2014) and OncoSearch (Lee, 2014)] for discovery of information related to specific proteins, genes or cancers. Within biomedical TM, some research has been conducted on identifying hallmark-based processes (i.e. events) in biomedical literature (Pyysalo *et al.*, 2013). This work has been based on the initial hallmark classification of Hanahan and Weinberg (2000), and it has focused on a set of events known to be relevant for the cancer genetics domain. Since hallmarks involve complex processes, relevant scientific data are diverse and difficult to harvest using standard search techniques. What is ideally needed is powerful technology, which categorizes and ranks data in literature on the basis of their relevance for hallmarks. Exploration of the resulting semantically structured data could help scientists find relevant information faster, make links between otherwise unconnected articles and create summaries and novel hypotheses from the scientific literature.

We present here a TM technique capable of such large-scale semantic classification of PubMed literature according to the current 10 hallmarks of cancer (Hanahan and Weinberg, 2011). We first introduce a corpus of 1499 cancer-related PubMed abstracts, which we have annotated according to the evidence they provide for hallmarks. We show that the annotations are accurate and that the corpus is representative. We then report experiments where the corpus is used as training and test data for automatic hallmark classification of literature. Our machine learning approach is based on support vector machines (SVM) and employs a rich set of features based on natural language processing (NLP) and existing resources. We present direct and task-based evaluation of the classification which demonstrates both the accuracy of the approach and its usefulness in supporting cancer research. We discuss future development and applications of our methodology.

2 Methods

The following two sub-sections describe the development of the hallmark corpus and the classifiers, respectively.

2.1 The hallmarks of cancer corpus

2.1.1 Evidence for hallmarks of cancer

Our starting point was to define the scientific evidence for hallmarks of cancer. Our primary resource was the two articles by Hanahan and Weinberg (2000, 2011), which describe examples of the cellular processes, proteins and genes involved in individual hallmarks. For example, 'apoptosis' can provide evidence for the 'resisting cell death' hallmark, and similarly 'caspase 3' because it is known to drive the apoptotic process. We also gathered additional evidence in literature during the annotation process. For example, articles studying specific cellular processes often also mention proteins or genes that can provide evidence for hallmarks. When needed, we used the KEGG pathways in cancer to confirm a proteins function and its role in cell signaling (http://www.genome.jp/kegg/disease/cancer.html).

2.1.2 PubMed literature retrieval

Abstracts were retrieved from PubMed journals representing subareas of biomedicine relevance to cancer research (e.g. molecular biology, public health and clinical medicine) using a set of search terms representative for each of the 10 hallmarks (Table 1). The terms and their synonyms appearing in Hanahan and Weinberg (2000, 2011) were employed, along with additional ones selected by a team of cancer researchers at Karolinska Institutet, Sweden. When needed, the term 'cancer' was added to filter out irrelevant abstracts (e.g. those concerning the 'immune response' without any obvious link to cancer). The PubMed searches were limited to years 1992, 2002 and 2012 to ensure coverage of varied data over time. The total number of retrieved abstracts per hallmark ranged from less than a hundred to several thousands. The abstracts were downloaded in the XML format.

2.1.3 Annotation

The annotation was conducted by an expert with 15+ years of experience in cancer research. The XUL-based annotation tool described in (Guo *et al.*, 2012) was used with its menu items customized to our hallmark task. The abstracts were chosen for annotation randomly, starting from the top of the list returned by PubMed search. Abstracts not containing information about hallmarks were left unannotated, as were those linked to review articles. Annotation

Table 1. Hallmarks and their search terms

Hallmark	Search term
1. Sustaining proliferative signaling (PS)	Proliferation Receptor Cancer 'Growth factor' Cancer 'Cell cycle' Cancer
2. Evading growth suppressors (GS)	'Cell cycle' Cancer 'Contact inhibition'
3. Resisting cell death (CD)	Apoptosis Cancer Necrosis Cancer Autophagy Cancer
4. Enabling replicative immortality (RI)	Senescence Cancer Immortalization Cancer
5. Inducing angiogenesis (A)	Angiogenesis Cancer 'Angiogenic factor'
6. Activating invasion & metastasis (IM)	Metastasis Invasion Cancer
7. Genome instability & mutation (GI)	Mutation Cancer 'DNA repair' Cancer Adducts Cancer 'Strand breaks' Cancer 'DNA damage' Cancer
8. Tumor-promoting inflammation (TPI)	Inflammation Cancer 'Oxidative stress' Cancer Inflammation 'Immune response' Cancer
9. Deregulating cellular energetics (CE)	Glycolysis Cancer; 'Warburg effect' Cancer
10. Avoiding immune destruction (ID)	'Immune system' Cancer Immunosuppression Cancer

was performed at sentence level, so that only sentences describing findings or conclusions of the study in question were included. A sentence was annotated when it contained clear evidence for one or several hallmarks of cancer. In the latter case, multiple labels were assigned to the sentence. Figure 1 shows annotated example sentences for different hallmarks, with hallmark evidence highlighted. The annotation labels for a given abstract are the combined set of labels of its individual sentences.

After the first round of annotation, additional annotation was required. For this, 10 supplementary sets of abstracts were retrieved from PubMed limited to the year 2010 using the search terms 'cell cycle', 'cellular energetics', 'DNA repair', 'glycolysis metabolism', 'immunosuppression', 'inflammation immune system cancer', ' inflammation oxidative stress cancer', 'necrosis cancer', 'cell cycle checkpoints' and 'contact inhibition'.

2.1.4 Statistics of annotated data

Table 2 lists the distribution of 1499 abstracts and sentences for each of the hallmark categories (for the shorthand notation of each hallmark category see Table 1). While we succeeded in finding a sufficient number of abstracts for most hallmarks, a few (e.g. CE) remained fairly low in frequency, most likely reflecting the lack of relevant scientific data in literature. To investigate the accuracy of annotations, we performed inter-annotator agreement analysis where a second expert annotator was asked to annotate a subset of 155 abstracts. The annotation was compared against that of the annotator who annotated the whole corpus. Good agreement was found between the two annotators with the average Cohen's Kappa of.81 for all the categories.

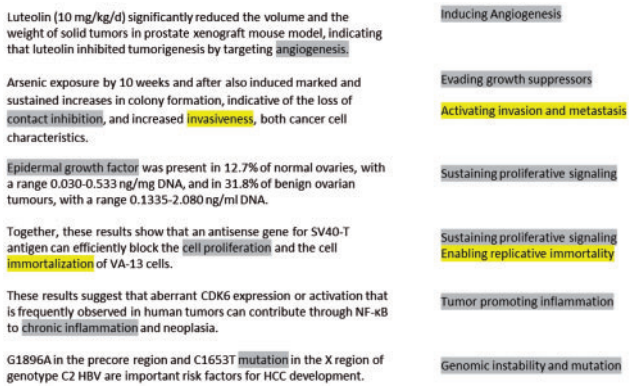

Fig. 1. Example sentences and color-highlighted evidence for hallmarks

Table 2. Distribution of data for the 10 hallmarks

Hallmark	No. abstracts	No. sentences
1. PS	462	993
2. GS	242	468
3. CD	430	883
4. RI	115	295
5. A	143	357
6. IM	291	667
7. GI	333	771
8. TPI	194	437
9. CE	105	213
10. ID	108	226

2.2 Hallmark classification

2.2.1 An overview of the classification process

Our methodology for hallmark classification consists of processing texts using an NLP pipeline, extracting a rich set of features from the resulting processed data and external resources and classifying the features using supervised machine learning. We have binary classifiers for each hallmark category, so that a given text can be classified under more than one category when each classifier is trained independently.

The NLP pipeline is illustrated in Figure 2. We start by tokenizing and Part-of-Speech tagging input text using the C&C tagger (Clark, 2002) which employs the Penn Treebank grammatical categories and is trained on biomedical texts. We lemmatize (stem) the output using the *BioLemmatizer* trained for biomedical texts (Liu *et al.*, 2012). The *C&C Parser* is then used to extract grammatical relations from lemmatized text. We trained the C&C Parser using available annotations from molecular biology (Rimell and Clark, 2009). Finally, named entities of relevance to hallmarks are extracted from parsed data using the state-of-the-art NER tool *ABNER* (Settles, 2005). ABNER is trained on the NLPBA and BioCreative corpora and achieves an *F*-score accuracy of 70.5% and 69.9% on these two corpora, respectively (Leitner *et al.*, 2010).

In the feature filtering (feature selection) stage of the pipeline, features that are deemed too rare or too common in the annotated corpus are filtered out, so that only the most discriminating features are used by the classifiers. The thresholds are set for each of the hallmarks by a process of trial and error, typically a minimum threshold value of 5, while the maximum threshold varies greatly depending on the feature type; usually a value larger than 500.

This improves both accuracy and reduces training time. This procedure is done separately for each of the hallmarks, i.e. we only select the features in the corpus that occur in abstracts annotated

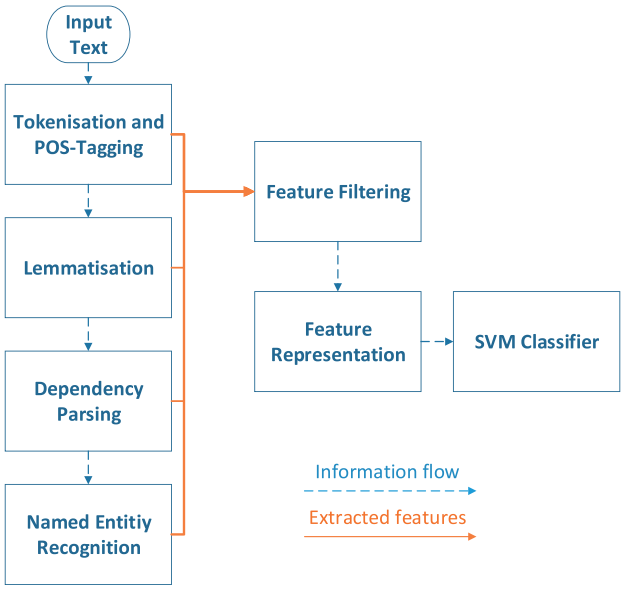

Fig. 2. Processing pipeline

Table 3. The final number of features for each hallmark after the feature filtering stage

Hallmark	LBoW	N-Bigrams	GR	VC	NE	MeSH	Chem	Total
1. PS	1471	355	435	121	302	280	129	3093
2. GS	863	176	189	119	108	156	55	1666
3. CD	1403	289	380	119	215	262	114	2782
4. RI	506	59	53	105	54	80	26	883
5. A	590	93	96	105	82	97	28	1091
6. IM	1052	220	253	114	161	172	43	2015
7. GI	1215	188	235	126	100	216	77	2157
8. TPI	843	122	152	111	99	121	34	1482
9. CE	410	54	53	103	31	68	20	739
10. ID	498	48	59	99	55	68	14	841

with the given hallmark. Therefore, each hallmark classifier has a unique set of selected features. Table 3 summarizes the number of features for each hallmark after feature filtering. The features are represented in a sparse binary format for each abstract, with a value 1 indicating that the given abstract contains this feature.

The binary features are then input into 10 classifiers [SVMs with radial basis function (RBF) kernels] that label each abstract with a binary label indicating its relevance for a particular hallmark.

2.2.2 Features and feature extraction

We experimented with seven feature types, chosen on the basis that many had performed well in previous (biomedical) text classification tasks.

1. **Lemmatized bag of words (LBow):** The simplest feature employs all the words occurring in input texts. We lemmatize the words to reduce feature sparsity.
2. **Noun bigrams (N-Bigrams):** Noun bigrams are used because they can be useful in capturing two word -concepts in texts (e.g. *Gene silencing*). No lemmatizing is employed to preserve the meaning of such concepts.
3. **Grammatical relations (GR):** We use the *Dobj* (direct object), *ncsubj* (non-clausal subject) and *iobj* (indirect object) relations in parsed data, taking into account their head and dependent words.
4. **Verb classes (VC):** Verb classes group semantically similar predicates together, providing the means to abstract away from individual verbs when faced with data sparsity. We used the hierarchical classification of 399 verbs by Sun and Korhonen (2009) which was automatically acquired from cancer risk assessment literature using clustering. We use all three levels of abstraction by allocating three bits in our feature representation for each concrete class (1 bit for each level of the abstraction hierarchy).
5. **Named entities (NE):** Named entities capture domain-specific concepts in texts, providing another way to group words into meaningful categories. We use five named entity types which are particularly relevant for cancer research: proteins, DNA, RNA, cell line and cell type. We store in the feature a pair of the entity type and the associated words or phrases.
6. **Medical subject headings (MeSH):** MeSH is a comprehensive controlled vocabulary for indexing journal articles and books in the life sciences. Most abstracts in our dataset contain an associated list of MeSH terms which we employ as features.
7. **Chemical lists (Chem):** Hallmark-related processes may involve chemicals. Since most abstracts in our corpus also contain, as metadata, a list of associated chemicals, we used these as features (a total of 3021 chemicals).

2.2.3 Classifiers

Given a set of training examples, each marked as belonging to one or more hallmark categories, an SVM training algorithm builds a binary model that predicts whether or not a new example falls into a particular category. An SVM model is a representation of the examples as points in space, mapped in a way such that the examples of the separate categories are divided by a clear gap that is as wide as possible. It constructs a hyperplane or a set of them in a high-dimensional space which can be used for classification or regression. The goal is to find the maximum-margin hyperplane which has the largest distance to the nearest data points of any class (Gunn *et al.*, 1998; Hsu *et al.*, 2003). SVMs have been applied widely in text classification over the past two decades (Joachims, 1998; Sebastiani, 2002) due to their relatively high performance in both cross-domain (Basu *et al.*, 2003; Sebastiani, 2002) and biomedical text classification tasks (Cohen and Hersh, 2005; Shatkay *et al.*, 2008).

We use the *LIBSVM* (Chang and Lin, 2011) in our experiments. It implements the Sequential Minimal Optimization Algorithm for kernelized SVMs. We have experimented using both a linear kernel and non-linear kernel such as the RBF kernel. On average, non-linear kernels such as the RBF performs around 5% higher in F-score accuracy.

3 Results

3.1 Intrinsic evaluation

We evaluate the classifiers intrinsically using precision: $\frac{\text{true positive}}{\text{true positive}+\text{false positive}}$, recall: $\frac{\text{true positive}}{\text{true positive}+\text{false negative}}$, accuracy: $\frac{\text{true positive}+\text{true negative}}{\text{total}}$ and F score: $\frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$ against manual annotations. The aforementioned measurements are typically expressed as percentages. We use standard cross-validation to avoid sampling bias. The data are divided into 4-folds, i.e. the model is trained with 75% of the data and tested with the remaining 25%, and this is done four times for full coverage of the dataset. The size of folds was selected based on the sparsity of the test data. Within the 75%

Table 4. Classification results using 10 independent SVM classifiers

Hallmark	No. abstracts	Precision (%)	Recall (%)	Accuracy (%)	*F* score (%)
1. PS	462	77.0	61.7	83.4	68.5
2. GS	242	73.5	59.5	90.5	65.8
3. CD	430	86.6	75.3	90.1	80.6
4. RI	115	92.9	68.7	97.3	79.0
5. A	143	90.2	70.6	96.6	79.2
6. IM	291	86.4	71.8	92.7	78.4
7. GI	333	88.2	76.3	92.8	81.8
8. TPI	240	81.6	64.6	92.4	72.1
9. CE	105	96.8	85.7	98.9	90.9
10. ID	108	81.6	65.7	96.6	72.8
Average		85.5	70.0	93.2	76.9

Table 5. Comparison of our approach to the Bag of Words (BoW) and keyword classification baselines

Hallmark	Our approach (%)	Baseline 1: BoW (%)	Baseline 2: keyword (%)
1. PS	68.5	63.2*	62.6**
2. GS	65.8	64.1*	64.5
3. CD	80.6	74.3	70.4**
4. RI	79.0	72.4	66.7**
5. A	79.2	75.2	74.1*
6. IM	78.4	71.2*	51.0*
7. GI	81.8	73.2	51.7*
8. TPI	72.1	67.4	58.6*
9. CE	90.9	78.4*	77.3**
10. ID	72.8	59.1**	44.7**
Average:	76.9	69.9	62.2

All numbers are *F* scores.
*Statistical significance level of $P < 0.05$ according to the McNemar test.
**$P < 0.001$ according to the McNemar test.

of the training data, we also perform another step of cross-validation for parameter tuning of the SVM kernels. Here, we do a 5-fold cross-validation, where we train with 80% of the data (for a given parameter configuration).

Table 4 lists the results for each of the 10 trained hallmark classifiers. The accuracy is impressive (93.2% on average), ranging between 83.4% and 98.9%. The average *F* score is 76.9%, with three categories scoring above 80% and from the remaining six categories only two scoring lower than 70%. The best results achieved are for the *Cellular energetics* (*CE*) category and the lowest for the *Evading growth suppressors* (*GS*) and the *Sustaining proliferative signaling* (*PS*) categories. The lower results are largely affected by recall (sensitivity) rather than precision (positive predictive value). They are not caused by data sparsity (as indicated by #*Abstracts*) but more likely by intrinsic difficulty of the categories in question. For example, the same underlying processes are relevant for both hallmarks and the difference is mainly apparent in the sets of proteins involved. A similar observation can be made with regards to *Avoiding immune destruction* (*ID*) and the *Tumor promoting inflammation* (*TPI*) categories, as they both represent immune response and inflammation, and involve overlapping processes. Further feature development such as named entity clustering could help and distinguish between such categories.

We compare the performance of our classifiers against two baselines:

Baseline 1: Bag of Words (*BoW*): We use the standard BoW baseline, where we count the occurrences of each word appearing in a given abstract (instance) and use these word-count pairs as features for the SVM classifier with an RBF kernel, fine-tuned across five cross-folds akin to our classifiers.

Baseline 2: Keyword-based classification: We compare our results to a simple keyword-match classification. We use the keywords in Table 1; if any of the keyword strings appears in the abstract text, it is classified under the corresponding hallmark(s).

The results presented in Table 5 show that our approach outperforms both baselines for all hallmarks, in most cases by a significant margin.

We conducted a leave-one-out feature analysis to determine which features contribute the most to the classification result. This involves removing one feature type (out of those outlined in Section 2.2.2) at a time and observing the change in results. The setup of this experiment is exactly the same as that described previously, with the exception of one feature type being left out. The cross validation and parameter turning are repeated for each feature leave-out iteration.

The analysis, shown in Table 6, shows that the most important feature type is the *lemmatized bag of words* (*LBoW*) which, when left out, results in a decrease of 9.8% basis points and (as our only feature type) decreases accuracy for all the 10 categories. The *verb clustering* (*VC*) feature performs the worst in this analysis, showing the smallest drop in *F*-score (0.2% on average). It is possible that the clusters learned from cancer risk assessment literature were not the best fit with our data and use of more relevant literature could improve performance. From the 10 categories, five benefited from all of the seven feature types.

3.2 Case studies

To evaluate the usefulness of the hallmark classification on unseen data, we performed four case studies. In the first two, we apply our approach to literature on selected tumor types and anticancer drugs. For well-studied tumor types and drugs, the most relevant and frequent hallmarks are known by experts. Whether the automatically generated literature distribution profiles confirm this existing knowledge can be a good indicator of the reliability of the classification and can complement intrinsic system evaluation. The results of these case studies were tested for statistical significance using χ^2 homogeneity test for each hallmark (using a 2×2 contingency table) followed by a Bonferroni correction for the entire profile's *P* values.

In the last two case studies, we evaluate our approach in the context of information retrieval. As described earlier, the hallmarks do not normally appear explicitly as literal strings in text; rather, they are latent in nature and retrieval of a comprehensive set of articles relating to these hallmarks requires using a large number of keywords like the ones presented in Table 1 and can result in a large number of false positives. Therefore, our goal is to show that we can identify a higher number of true instances than realistic using a standard keyword search, while keeping the number of false positives lower.

3.2.1 Case study 1

Basal cell carcinoma and melanoma are two types of human skin cancers with different biology and consequently differing degrees of malignancy. Melanoma is highly metastatic with high mortality, while the more common basal cell carcinoma rarely or never metastasizes and has lower mortality (Tomasetti and Vogelstein, 2015)

All PubMed abstracts available in December 2014, including 22 564 abstracts for basal cell carcinoma and 98 924 for melanoma

Table 6. Results for leave-one-out analysis

Hallmark	All (%)	GR (%)	VC (%)	NE (%)	MeSH (%)	Chem (%)	LBoW (%)	N-Bigram (%)
1. PS	68.5	66.1	67.1	64.4	66.5	66.2	66.3	65.8
2. GS	65.8	60.7	63.0	61.7	61.9	61.7	51.7	59.0
3. CD	80.6	78.6	80.5	80.2	78.3	80.4	70.4	80.5
4. RI	79.0	(80.2)	(81.2)	79.0	78.4	79.0	69.9	79.0
5. A	79.2	(81.4)	79.1	(81.1)	(79.4)	(81.1)	70.7	(80.3)
6. IM	78.4	78.1	(79.4)	77.6	76.9	(78.9)	68.7	77.5
7. GI	81.8	(82.1)	81.8	(82.0)	79.2	(83.0)	74.2	(81.9)
8. TPI	72.1	69.7	(72.5)	71.1	68.0	(71.6)	60.5	70.5
9. CE	90.9	88.8	89.8	88.1	88.7	88.7	84.5	88.1
10. ID	72.8	70.4	72.4	72.4	69.0	72.7	54.0	71.7
Average	76.9	75.6	76.7	75.8	74.6	76.3	67.1	75.4

Figures in parentheses show an improvement in accuracy when the given feature is removed from classifications. All figures are *F* scores.

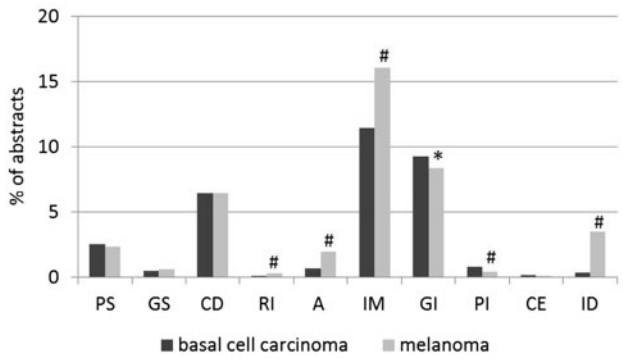

Fig. 3. Case study 1: the distribution of basal cell carcinoma and melanoma literature over the relevant hallmarks. *Statistical significance level ($P < 0.05$). #$P < 0.001$

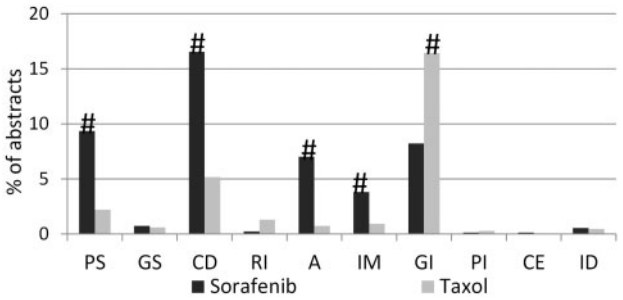

Fig. 4. Case study 2: the distribution of Sorafenib and Taxol literature over the relevant hallmarks. *Statistical significance level ($P < 0.05$). #$P < 0.001$

were used. The results of classification are shown in Figure 3. Out of a total of 121 488 abstracts from the original literature search; only 46 727 abstracts (38%) were classified as relevant, highlighting the time saving function of automatic classification.

Comparing the literature distribution for the hallmarks activating invasion and metastasis, a significant difference can be seen with higher numbers of abstracts for melanoma, reflecting the existing knowledge about the metastatic potential of melanoma (Akinci *et al.*, 2008; Fidler, 1995; Young *et al.*, 2008). A significantly higher number of abstracts for melanoma were also found for angiogenesis and avoiding immune destruction. Also these two parameters reflect malignancy. This classification pattern is in line with existing scientific knowledge, demonstrating the reliable performance of our approach. Our methodology structures a large amount of textual information (more than 121 000 abstracts for the two tumor types in the original PubMed search) according to hallmarks—a task that would be near impossible to conduct manually.

3.2.2 Case study 2

Sorafenib and taxol are two drugs that have been developed to treat cancer via different mechanisms. Sorafenib acts by inhibiting development of new blood vessels (anti-angiogenic) (Wilhelm *et al.*, 2006), while taxol inhibits cancer cell growth by inducing genomic instability (Schiff and Horwitz, 1980)

We used for investigation all the PubMed abstracts available in December 2014, including 3846 abstracts for sorafenib and 24 827 for taxol. The results of classification are illustrated in Figure 4. Out of a total of 28 673 abstracts in the original literature retrievals, only 8993 abstracts were classified as relevant for cancer hallmarks (31%), again highlighting the time-saving aspect of automatic classification

Comparing the literature distribution for sorafenib and taxol, there is a significant difference in the percentage of abstracts relevant for the aforementioned hallmarks: a significantly higher number of abstracts for sorafenib were found for the hallmarks inducing angiogenesis resisting cell death and sustaining proliferative signaling. This is in line with the anti-angiogenic effect of sorafenib and its effect in causing cell death (Wilhelm *et al.*, 2006). The most frequent hallmarks in taxol literature are 'genomic instability and mutation'. This corresponds to existing knowledge about taxol as a drug that interferes with microtubules and chromosomal segregation in a dividing cell and may lead to genetic instability (Abal *et al.*, 2003; Pihan and Doxsey, 1999). This study, again, reflects the accuracy of the literature distribution profiles and the reliability of the classification results.

3.2.3 Case study 3

This case study investigates whether our classification approach can identify a higher number of relevant abstracts and with fewer false positives than a standard keyword search approach. We compare the number of articles retrieved by PubMed keyword search to what our classifiers can identify for a given search topic. For our test topic, we use 'Melanoma'. We combine the search query 'Melanoma' with a single search string according to expert's best description of each hallmark name (see Table 7 for the search queries used). We can estimate the percentage of its false positives by ascertaining the relevance of the top 20 PubMed-retrieved search results for each hallmark using expert evaluation: an expert in cancer research is asked whether each of the top 20 retrieved abstract is really

Table 7. The search queries used to describe the 10 hallmarks while searching for the topic: melanoma

Hallmark	Search query
1. PS	melanoma AND proliferation
2. GS	melanoma AND 'growth suppression'
3. CD	melanoma AND 'cell death'
4. RI	melanoma AND immortalization
5. A	melanoma AND angiogenesis
6. IM	melanoma AND 'invasion metastasis'
7. GI	melanoma AND 'genomic instability mutation'
8. TPI	melanoma AND inflammation7
9. CE	melanoma AND 'warburg effect'
10. ID	melanoma AND 'immune destruction'

Table 8. Case study 3 results, comparing the number of abstracts retrieved from PubMed using the search queries in Table 7 and the number of classified abstracts out of a total of 98 924 abstracts using our approach

Hallmark	Keyword search		Our approach	
	No. retrieved	% False positives	No. classified	% False positives
1. PS	6958	0	1808	0
2. GS	105	35	472	0
3. CD	1813	0	4972	0
4. RI	23	25	198	5
5. A	2155	0	1514	0
6. IM	1954	0	12 424	0
7. GI	101	10	6478	0
8. TPI	1395	15	313	0
9. CE	20	15	80	5
10. ID	35	0	2674	0
Average	1456	10	3093	1

The % false-positive numbers are only of the top 20 retrieved abstracts using each of the search strings in Table 7 and not of the entire result set.

related to the given hallmark based on the abstract text. We can compare our classifiers' performance by retrieving articles from PubMed for 'Melanoma' (98 924 abstracts in total) and then run our classifiers over these articles to find which of the 'Melanoma' abstracts that are also associated with each of the hallmarks; therefore, we are identifying abstracts that are both relevant to the search topic 'Melanoma' and each of the hallmarks, thereby directly comparing retrieval performance with that of the PubMed keyword search queries (Table 7).

We then evaluate the output of our classifiers for the same list of 20 abstracts for each hallmark also against expert's judgment and compare the percentage of false positives of our classifiers, using the PubMed keyword search as a benchmark.

Table 8 summarizes the results. Overall, our classifiers managed to identify substantially more abstracts than the keyword-based approach in 7 of the 10 hallmarks, while having a much lower percentage of false positives for the sample of top 20 retrieved results for all of the hallmarks.

3.2.4 Case study 4

In the previous case study, we constrained the keyword search string to terms that best describe the 10 hallmarks. In this case study, we replicate the experimental setting from the previous case study, except that we do an unconstrained direct experiment, where an independent expert in cancer research is asked to search for papers about the topic (Melanoma) with a set of terms that they believe are associated with each of the hallmarks. That is the search terms do not need to be a description of the names of the hallmarks as in the previous case study, instead it is left to the expert to openly decide on any associated terms to search. Table 9 lists the resulting search queries selected by our expert volunteer.

Table 9. Search queries used by an independent user to retrieve documents about Melanoma, relating to each of the 10 hallmarks

Hallmark	Search query
1. PS	Melanoma AND 'growth factor'
2. GS	Melanoma AND 'cell cycle'
3. CD	Melanoma AND apoptosis
4. RI	Melanoma AND telomerase
5. A	Melanoma and 'angiogenic factor'
6. IM	Melanoma AND EMT
7. GI	Melanoma AND 'DNA damage'
8. TPI	Melanoma AND 'oxidative stress'
9. CE	Melanoma AND glycolysis
10. ID	Melanoma AND immunosuppression

Table 10. Case study 4 results, comparing the number of abstracts retrieved from PubMed using the search queries in Table 9 and the number of classified abstracts out of a total of 98 924 abstracts using our approach

Hallmark	Keyword search		Our approach	
	No. retrieved	% False positives	No. classified	% False positives
1. PS	3079	30	1808	0
2. GS	2512	0	472	0
3. CD	4834	0	4972	0
4. RI	191	85	198	0
5. A	89	0	1514	0
6. IM	1812	0	12 424	0
7. GI	868	0	6478	0
8. TPI	395	70	313	0
9. CE	142	35	80	10
10. ID	1399	0	2674	0
Average	1532	20	3093	1

The % false-positive numbers are only of the top 20 retrieved abstracts using each of the search strings in Table 9 and not of the entire result set.

The results (Table 10) show that overall, our classifiers identify substantially more abstracts than the keyword-based approach for the majority of the hallmarks, while having a much lower percentage of false positives for the sample of top 20 retrieved results when compared with the keyword-based search.

One should note the very high false-positives for Hallmarks 4 and 8 (RI and TPI) for the top 20 retrieved PubMed keyword search. This highlights some of the weaknesses of a keyword only search approach, where even expert users may not find the optimal search terms for each hallmark, it shows the need of carefully selecting the correct hallmark-related terms to avoid false positives in standard keyword search. In contrast, our classifiers can mitigate this risk since they take into account thousands of linguistic features instead of a small set of search terms.

Table 11. The number of hallmark labels per abstract in our corpus

Hallmarks per abstract	Frequency	Proportion (%)
1	951	60.2
2	450	28.5
3	141	8.9
4	31	2.0
5	5	0.3

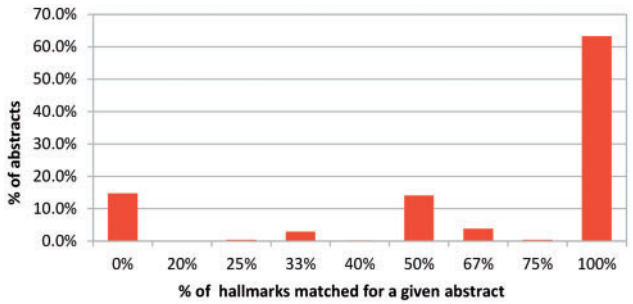

Fig. 5. The distribution of all abstracts according to the percentage of their correctly predicted hallmarks

4 Discussion

The evaluation reported above demonstrates the accuracy and the practical potential of hallmark-based text classification. We report here further analysis to gain insight into the errors made by the classifier and how to improve our approach in the future. We first looked into multi-labeling. As listed in Table 11, 40% of the abstracts in our annotated corpus are multi-labeled. Figure 5 displays the proportion of abstracts in the corpus according to the percentage of matched hallmark labels for a given abstract in the classifier output; 63.3% of abstracts have 100% of their labels correctly predicted by the classifier, while 14.8% have no matches (0% of their labels). The high percentage of 100% matched and 0% matched can be attributed to 60% of the abstracts having a single label.

We next examined the actual hallmark pair co-occurrences (i) in the annotated corpus (Table 12 where the diagonal line shows the number of occurrences for a given hallmark in the corpus independent of other co-occurring hallmarks) and (ii) as predicted by the classifier (Table 13). Looking at Table 12, hallmarks that most often co-occur with each other include 'sustaining proliferative signaling (PS)', 'resisting cell death (CD)' and 'evading growth suppressors (GS)', with 165 abstracts labeled as PS and CD and 120 abstracts labeled as PS and GS. These co-occurrences could be explained by the fact that they are all related to cell cycle regulation. For example, in the sentence: '*Moreover, harmine not only induced endothelial cell cycle arrest and apoptosis, but also suppressed endothelial cell migration and tube formation as well as induction of neovascularity in a mouse corneal micropocket assay*', the phrase: 'cell cycle arrest' is a good indicator of PS and GS, and the word 'apoptosis' is a good indicator of CD. This might be explained by overlapping capabilities (e.g. cell growth) and is likely to be the main reason for the lower classifier performance figures for these three hallmarks in Table 4. Looking at Table 13, for many hallmarks, the predicted co-occurrences are well-correlated with those in the annotated corpus. For example, PS and CD are co-classified 154 times which is relatively comparable to their 165 co-occurrences in the corpus. Similar observations can be made with regard to PS and GS. Our current approach is based on training 10 independent, binary classifiers to predict whether an abstract belongs to a given hallmark category.

Table 12. Hallmark co-occurrence distribution in the annotated corpus

	PS	GS	CD	RI	A	IM	GI	PI	CE	ID
PS	462	120	165	19	38	79	31	25	9	5
GS	120	242	86	15	10	28	31	9	2	0
CD	165	86	430	23	28	48	44	37	16	14
RI	19	15	23	115	2	6	28	4	3	2
A	38	10	28	2	143	42	0	14	2	3
IM	79	28	48	6	42	291	14	24	9	13
GI	31	31	44	28	0	14	333	27	7	6
PI	25	9	37	4	14	24	27	194	7	14
CE	9	2	16	3	2	9	7	7	105	0
ID	5	0	14	2	3	13	6	14	0	108

Table 13. Hallmark co-occurrence distribution as predicted by the classifier

	PS	GS	CD	RI	A	IM	GI	PI	CE	ID
PS	285	110	154	14	32	64	23	14	9	8
GS	113	142	82	6	8	24	27	3	2	0
CD	138	78	327	14	21	37	38	29	13	12
RI	13	16	21	80	1	7	26	4	3	1
A	31	8	19	1	105	32	1	10	2	2
IM	69	26	35	2	33	211	12	15	7	11
GI	23	26	40	24	0	12	258	19	6	5
PI	16	8	28	1	10	18	23	125	6	16
CE	8	6	18	1	0	7	1	4	86	0
ID	8	2	10	1	3	12	7	14	0	72

We could also experiment with models that allow the classifiers to work together, e.g. models based on joint inference (Poon and Vanderwende, 2010) or joint learning (Zang *et al.*, 2013). This type of methodology, which has been successfully applied to similar NLP tasks, is likely to improve performance as it provides the means to capture dependencies and interactions between co-occurring hallmarks.

The first two case studies provide additional evidence that our system correctly classifies literature over the hallmarks of cancer. The automatic system rapidly generated profiles that would have been difficult and very time-consuming to produce manually, which would facilitate overviews of scientific literature. In the future, the approach may be further developed to support the detection of novel patterns and research hypotheses in literature.

The last two case studies show that our approach can support information retrieval in comparison with a search string intersection query where the goal is to identify documents for a given topic, as well as articles that relate specifically to certain hallmarks. Our approach generally identifies more documents and has a smaller percentage of false positives than standard keyword-based search. This can perhaps be explained by the latent nature of the hallmarks in texts—the fact that they are rarely stated explicitly but rather via indirect correlation of terms that describe relevant biological processes, and therefore are not easily found by basic keyword search. Our case studies also demonstrate that experts selecting the wrong set of search terms may result in a high number of false positives and that our classifiers are not susceptible to this problem since they are trained on a large number of features and not on the occurrence of a single search term.

Our analysis also suggests that many hallmarks could be subdivided, e.g. according to the established pathways involved in

tumor development. During cancer development, aberrantly regulated intracellular signaling pathways tend to rearrange networks regulating cancer cells and the networks themselves can be divided into sub-circuits which regulate certain capabilities of cancer cells, e.g. viability circuit. We are currently developing such an enriched classification of hallmarks, so as to help cancer researchers navigate more easily to the literature of their specific interest. However, whether the more subtle differences between subcategories of hallmarks can be captured by machine learning is yet to be known, and we plan to investigate that in due time.

5 Conclusions

We have introduced a TM technique capable of large-scale semantic classification of PubMed literature according to the evidence they provide for the hallmarks of cancer (Hanahan and Weinberg, 2011). Our evaluation demonstrates both the accuracy of the approach and its usefulness in supporting cancer research. In the future, we plan to improve and refine our classification approach as discussed in the previous section. Given the prominence of hallmarks in recent and current cancer research, we expect that the resulting methodology will offer a highly useful literature analysis tool for cancer researchers. The ability to organize literature semantically according to the hallmarks of cancer can help researchers summarize known and find novel information in literature faster. It can support both basic and applied research into cancer, including cancer drug development, prevention strategies, biomarker discovery and identification of previously unknown associations between genes, proteins, signaling networks, tumor types, drug, chemicals and other entities.

Funding

This work was supported by the Commonwealth Scholarship Commission and the Cambridge Trust (to S.B.), by Vinnova (to I.S.) and by MRC grant MR/M013049/1.

Conflict of Interest: none declared.

References

Abal,M. *et al.* (2003) Taxanes: microtubule and centrosome targets, and cell cycle dependent mechanisms of action. *Curr. Cancer Drug Targets*, **3**, 193–203.

Akinci,M. *et al.* (2008) Metastatic basal cell carcinoma. *Acta Chirurgica Belgica*, **108**, 269.

Basu,A. *et al.* (2003) Support vector machines for text categorization. In: Proceedings of the 36th Annual Hawaii International Conference on System Sciences, 2003. IEEE, pp. 7.

Chang,C.-C. and Lin,C.-J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 27.

Clark,S. (2002) Supertagging for combinatory categorial grammar. In: Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6), pp. 19–24.

Cohen,A.M. and Hersh,W.R. (2005) A survey of current work in biomedical text mining. *Brief. Bioinform.*, **6**, 57–71.

Doland,M.E. (2014) Capturing cancer initiating events in OncoCL, a cancer cell ontology. In: *AMIA Joint Summits on Translational Science*.

Fidler,I.J. (1995) Melanoma metastasis. *Cancer Control*, **2**, 398–404.

Gunn,S.R. *et al.* (1998) Support vector machines for classification and regression, Vol. 14. *ISIS Technical report*, University Of Southampton.

Guo,Y. *et al.* (2012) CRAB reader: a tool for analysis and visualization of argumentative zones in scientific literature. In: *Proceedings of COLING 2012: Demonstration Papers*, pp. 183–190.

Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.

Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.

Hsu,C.-W. (2003) A practical guide to support vector classification. National Taiwan University, Taipei, Taiwan., www.csie.ntu.edu.tw/cjlin/papers/guide/guide.pdf.

Joachims,T. (1998) *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. Springer, Berlin Heidelberg.

Korhonen,A. *et al.* (2012) Text mining for literature review and knowledge discovery in cancer risk assessment and research. *PLoS One*, **7**, e33427.

Lee,H.-J. (2014) Oncosearch: cancer gene search engine with literature evidence. *Nucleic Acids Res*, **2**(Web Server issue), W416–W421. doi: 10.1093/nar/gku368.

Leitner,F. *et al.* (2010) An overview of biocreative ii. 5. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **7**, 385–399.

Liu,H. *et al.* (2012) Biolemmatizer: a lemmatization tool for morphological processing of biomedical text. *J. Biomed. Semantics*, **3**, 3.

Marusyk,A. *et al.* (2012) Intra-tumour heterogeneity: a looking glass for cancer? *Nat. Rev. Cancer*, **12**, 323–334.

McDonald,D. *et al.* (2012) Value and benefits of text mining. *JISC Digital Infrastructure*.

Pihan,G.A. and Doxsey,S.J. (1999) The mitotic machinery as a source of genetic instability in cancer. In: *Seminars in Cancer Biology*, **Vol. 9**. Elsevier, pp. 289–302.

Poon,H. and Vanderwende,L. (2010) Joint inference for knowledge extraction from biomedical literature. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 813–821.

Pyysalo,S. *et al.* (2013) Overview of the cancer genetics (cg) task of bionlp shared task 2013. In: BioNLP Shared Task 2013 Workshop.

Rimell,L. and Clark,S. (2009) Porting a lexicalized-grammar parser to the biomedical domain. *J. Biomed. Inform.*, **42**, 852–865.

Schiff,P. and Horwitz,S.B. (1980) Taxol stabilizes microtubules in mouse fibroblast cells. *Proc. Natl. Acad. Sci. USA*, **77**, 1561–1565.

Sebastiani,F. (2002) Machine learning in automated text categorization. *ACM Comput. Surv.*, **34**, 1–47.

Settles,B. (2005) ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, **21**, 3191–3192.

Shatkay,H. *et al.* (2008) Multi-dimensional classification of biomedical text: toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, **24**, 2086–2093.

Simpson,M.S. and Demner-Fushman,D. (2012) Biomedical text mining: a survey of recent progress. In: *Mining Text Data*, pp. 465–517. Springer.

Spasic,I. *et al.* (2014) Text mining of cancer-related information: review of current status and future directions. *Int. J. Med. Inform.*, **83**, 605–623.

Stewart,B. and Wild,C.P. (2014) *World Cancer Report 2014*. IARC, Lyon, France.

Sun,L. and Korhonen,A. (2009) Improving verb clustering with automatically acquired selectional preferences. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2*, pp. 638–647. Association for Computational Linguistics.

Tomasetti,C. and Vogelstein,B. (2015) Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, **347**, 78–81.

Wilhelm,S. *et al.* (2006) Discovery and development of sorafenib: a multikinase inhibitor for treating cancer. *Nat. Rev Drug Discov.*, **5**, 835–844.

Young,L. *et al.* (2008) Evidence that dysregulated DNA mismatch repair characterizes human nonmelanoma skin cancer. *Br. J. Dermatol.*, **158**, 59–69.

Zang,Z. *et al.* (2013) Learning classifier system with average reward reinforcement learning. *Knowl. Based Syst.*, **40**, 58–71.

Zhua,F. *et al.* (2011) Biomedical text mining and its applications in cancer research. *J. Biomed. Inform.*, **46**, 200–211.