Clustering Cancer Drugs According to their Mechanisms of Action

Syed Abdullah Ali and Riza Batista-Navarro

University of Manchester, Oxford Road, Manchester, M13 9PL, UK
syeabdullah.ali@postgrad.manchester.ac.uk, riza.batista@manchester.ac.uk

Abstract. This research investigates similarities between cancer-related drugs with respect to their mechanisms of action (MOA) guided by information extracted from two resources: scientific literature and ontologies. To find similarity between drug pairs, the Chemical Entities of Biological Interest (ChEBI) ontology and Gene Ontology (GO) were leveraged to compute drug and drug target features, respectively. A graph of drugs was formed based on drug pairs clustered using two unsupervised graph clustering algorithms: Chinese Whispers and Louvain Method. As a result of clustering, drugs that share the same dominant MOA were placed in the same cluster. Additionally, the most prominent drugs in the entire graph and within each cluster were identified according to graph centrality measures. Quality of the clusters was assessed by calculating silhouette coefficient values, ensuring consistency of results generated by the two algorithms, and employing the help of a domain expert who carried out manual evaluation.

Keywords: Mechanisms of Action, Drug Clustering, Chinese Whispers, Louvain Method, Cancer

1 Introduction

Cancer, the abnormal growth of cells due to gene mutations, is one of the top causes of mortality. In 2014 alone, cancer claimed 163,000 lives in the UK, which means that on the average, 450 die every day [24]. One in every two people who were born after 1960 in the UK will be diagnosed with cancer at some point in his or her life [2]. With over 200 different types, cancer is usually progressive, chronic and has a mild phenotype making it hard to diagnose.

The discovery of anti-cancer drugs has not been a successful endeavour. The rate at which new molecular entities (NMEs) are approved for clinical investigation is low. In the US, between the period of 2010 and 2014, the maximum number of approved NMEs was only 11 (in 2012) [14]. In addition to that, many drugs fail in clinical trials. The recorded success rate in cancer trials is three times lower than in clinical trials for cardiovascular diseases [13].

Numerous efforts have been carried out, based on different approaches, to facilitate anti-cancer drug discovery such as: proteomic analysis to highlight protein drug targets [3], metabolic control analysis to highlight likely pathological metabolism targets [6] or development of drug ontology databases [7, 8] for in-depth understanding of mechanisms of action (MOA) of known drugs. A MOA is a biochemical interaction that results in a drug producing its desired therapeutic pharmacological effect [1]. It depicts events happening at the chemical level, including drugs binding with drug targets, as well as reactions with enzymes or receptor sites [15]. To understand these mechanisms, different approaches have been proposed, e.g., analysis of chemical properties of drugs as well as of drug targets, and the extraction of drug-protein relationships from scientific literature. However, no effort has attempted to combine these approaches to group drugs according to their MOA. Clustering drugs which share similar MOA streamlines the drug discovery process as it constrains the number of drugs that need to be investigated with respect to their drug targets. It also provides the foundation for understanding and advancing combinatorial drug therapies for cancer.

To fill this research gap, we propose an approach that makes use of information from ontologies for clustering drugs in order to highlight similarities and differences between them according to a range of physiochemical properties. The approach is novel in that it seeks to identify groups or clusters of drugs which are similar in terms of three features: (1) chemical structure, (2) biological role and (3) drug target properties. In forming groups amongst a total of 831 drugs, we investigated the use of unsupervised graph clustering techniques.

2 Related Work

Clustering drugs based on similarity is a well-known approach to drug discovery, although there is great variation in terms of how similarity measures have been defined. Gemma et al. 2006 [9], for instance, clustered lung cancer drugs based on their gene expression profiles and sensitivity to multiple lung cancer cell lines. The results suggested that one of the drugs acted particularly different than the rest of the drugs; hence this drug might be useful in second-line chemotherapy if it was not administered to a patient initially. A similar attempt was made in Uhr et al. 2015 [23], whose research focussed on clustering 37 breast cancer drugs based on their sensitivity to 42 breast cancer cell lines. This resulted in six clusters which highlighted relationships between drugs and their sensitivities. Meanwhile, Jeon et al. 2011 [11] employed *k*-means clustering on gene expression data to understand changes in mitochondrial proteins brought about by mitochondrial DNA depletion. Their research revealed how cells compensate for mitochondrial DNA depletion and it also led to the identification of proteins that repair this depletion.

Ross et al. 2000 [20] presented a clustering approach which analysed 60 cell lines (NCI-60) based on microarray analysis, i.e., the parallel monitoring of gene expression levels in thousands of genes within the context of a specific biological process across different environments or tissue samples (e.g., tumours). Hierarchical clustering was performed to cluster cell lines and genes, separately. Their results showed that two of the breast cancer cell lines are similar to melanoma cell lines, suggesting a relationship between the two types of cancers.

A data-driven approach was presented in Udrescu et al. 2016 [22] where information on drug-drug interactions (DDI) from the DrugBank database was utilised to find communities based on Louvain Method [5]. Prominent drugs were ranked using graph centrality measures [16] such as degree, betweenness, closeness, eigenvector and PageRank centrality [17]. A total of 1,141 nodes (corresponding to drugs) and 11,688 links (corresponding to DDIs) were divided into nine different clusters, each of which represented a specific class of drugs. For example, one cluster represented drugs that targeted the immune system; another pertained to drugs for the nervous system, and so on. Effectively, the research identified functional drug categories and relationships.

While each of the related work presented above holds resemblance to our own research methodology in attempting to cluster drugs based on their chemical interactions, our work proposes a different set of features for highlighting similarities between drugs. Specifically, the work last mentioned above assumes all drug similarities to be of equal importance whereas we calculate similarity scores based on various features, described in the next section.

3 Methodology

Our approach to clustering drugs according to their mechanisms of action is based on their similarity with respect to the following types of information derived from ontologies: (1) chemical structure, (2) biological roles, and (3) drug target properties. Guided by a set of biomolecular events automatically extracted from scientific literature, we formed a list of drugs and drug targets of interest, whose features were calculated with the help of two ontologies: the Chemical Entities of Biological Interest (ChEBI) ontology [8] and the Gene Ontology (GO) [7]. The rest of this section describes in detail our processing pipeline, also depicted in Figure 1.

3.1 Information extraction

A set of biomolecular events, i.e., interactions between drugs and their targets, served as input to our clustering approach. These events were obtained using a text mining workflow developed using the Argo workbench [18] and employed in the work carried out by Zerva et al. [25]. Based on the processing of 6,529 full-text papers relevant to melanoma[1], a total of 3,168 events were extracted. Out of these, 293 pertained to interactions between drugs (which we will henceforth refer to as drug-drug interactions or DDIs) while 2,875 are drug-protein interactions (DPIs). A list of 831 drugs was formed upon combining all unique drugs in the DDI and DPI sets. For each drug in the DPI set, we also created a list of drug targets.

[1] Retrieved by calling the Europe PubMed Central API with a query containing “melanoma”, its synonyms and names of melanoma cell lines

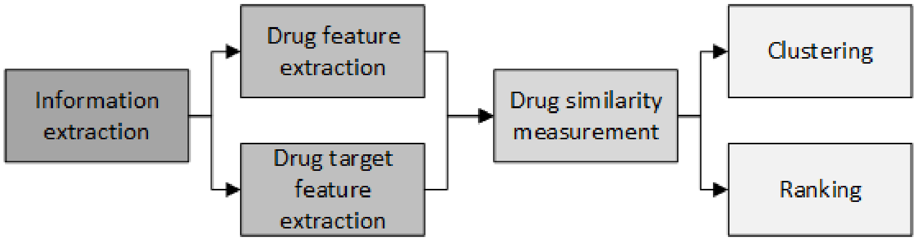


Fig. 1. Our processing pipeline for clustering drugs.

3.2 Drug feature extraction

ChEBI is an ontology that catalogues chemical structures and biological roles of drugs. For each drug of interest in our list, we obtained related terms from the "chemical structure" and "biological role" sub-trees of the ontology. Structural information among the tree terms was discarded and only terms appearing between the drug and three levels up were stored, corresponding to individual drugs[2].

3.3 Drug target feature extraction

The Gene Ontology (GO) contains information on cellular components and molecular functions of genes, as well as the biological processes they are involved in. For each of the proteins that a particular drug interacts with—according to our list of DPIs—we obtained a list of gene products from GO. For each gene product, we obtained all related terms in GO up to a tree depth of three levels up. The result is a combined list of GO terms associated with a particular drug. A depiction of the different types of features extracted for any given drug is shown in Figure 2.

3.4 Drug similarity measurement

Based on the features obtained in the steps described above, we measure the similarity between drugs. Specifically, given any two drugs, two types of similarity scores were calculated, i.e., the number of shared ChEBI terms and the number of shared GO terms. From a total of 423,801 drug pairs, there were 92,700 and 208,499 pairs that have a non-zero number of shared ChEBI and GO terms, respectively. The similarity score (i.e., the number of shared terms) between a pair of drugs based on ChEBI terms varied from 1 to 49 while that based on shared GO terms ranged from 1 to 8,526. In order to normalise them, we divided each score by the respective maximum value observed, and then multiplied the result by 100 to get a percentage.

[2] The number of levels was constrained to three to ensure that succeeding steps will be computationally feasible.

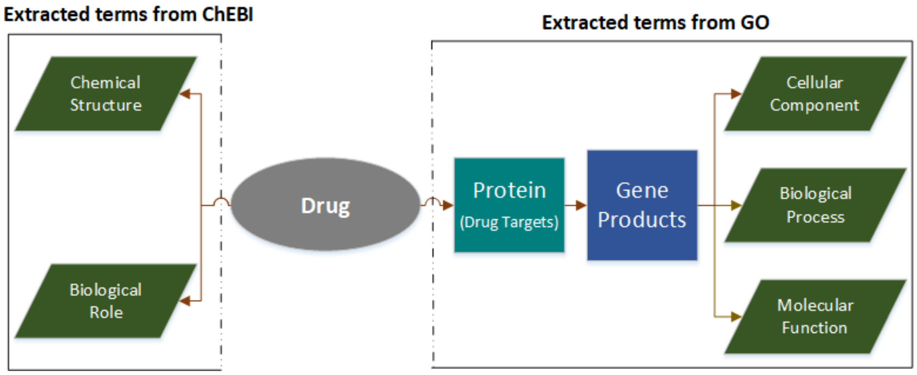


Fig. 2. Different types of features extracted from ontologies for any given drug.

There are however, many drug pairs whose similarity scores were very low, indicating insignificant similarity between drugs. To overcome this problem, we retained only drug pairs whose ChEBI and GO similarity scores are both at least 10%. The threshold was set to this level based on our analysis of the data distribution which informed our decision on an optimal cut off. A combined similarity score is finally calculated by taking the mean of the two scores. If only one of the scores exists for a drug pair (e.g., in cases where a score was obtained by counting the number of shared ChEBI terms but none based on shared GO terms, or vice-versa), we take half of the value of the lone score as the combined similarity but only if it is at least 50%. Otherwise, the pair is discarded. After this step, only 261 drug pairs remained, having 89 unique drugs.

The remaining drug pairs were combined to form an interconnected graph following the Yifan Hu representation [10]. In this graph representation, nodes represent drugs and weighted edges between nodes correspond to their similarity score. It is found that approximately around 20% of the nodes in the graph were not connected to the larger central interconnected graph. We consider them as pertaining to drugs which are extremely unique in terms of their MOA, or irrelevant to cancer but were spuriously included during the information extraction step. These nodes were therefore filtered out, retaining only 72 nodes.

3.5 Drug clustering and ranking

We investigated two algorithms for clustering the graph of drugs, namely, Chinese Whispers (CW) [4] and Louvain Method (LM) [5]. CW assigns a class label to each node based on the strongest weight of the neighbouring class, in iterative cycles. Meanwhile, LM optimises graph modularity, which is a measure of separation between densely connected regions of the graph.

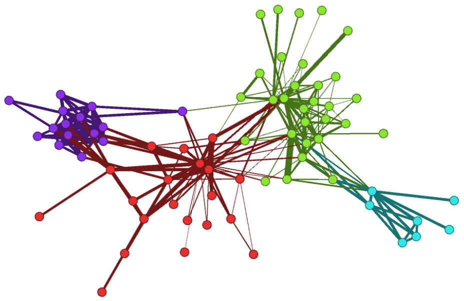
All drugs in the graph were then ranked with respect to their prominence[3] in the whole graph according to: (1) degree, the number of immediate neighbors;

[3] Pertains to importance based on their centrality within the graph

(2) closeness, the distance between each node and all other nodes in the graph; (3) betweenness, the number of times a node is traversed along the shortest distance between each node and all other nodes in the graph; (4) PageRank, the probability distribution based on the likelihood of stopping on the node while traversing connected nodes; and (5) eigenvector, the principal eigenvector in the adjacency matrix of the graph. Additionally, drugs were also ranked within each cluster using the same centrality measures.

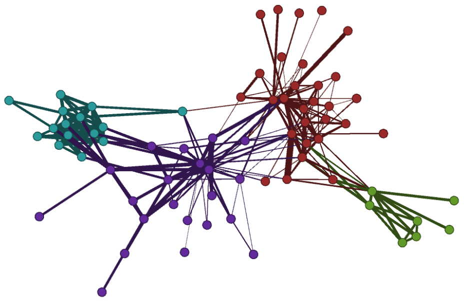
4 Experiments and Results

Upon the application of Chinese Whispers and Louvain Method on our data, we obtained the results shown in Figures 3 and 4, respectively, following the Yifan Hu graph representation. For simplicity, each cluster is assigned a cluster identifier and represented by a unique colour. Its distribution, i.e., the proportion of the nodes in the graph that belongs to the cluster, is also indicated.



Cluster ID	Color	Distribution
CW_C1	🟩	41.67%
CW_C2	🟥	27.78%
CW_C3	🟪	20.83%
CW_C4	🟦	9.72%

Fig. 3. Result of drug graph clustering based on the Chinese Whispers (CW) algorithm.



Cluster ID	Color	Distribution
LM_C1	🟥	40.28%
LM_C2	🟪	29.17%
LM_C3	🟩	20.83%
LM_C4	🟩	9.72%

Fig. 4. Result of drug graph clustering based on the Louvain Method (LM) algorithm.

Table 1. Ten most prominent drugs with respect to the entire graph, according to five centrality measures.

Rank	Degree	Closeness	Betweenness	PageRank	Eigenvector
1	L-threonine	L-gamma-glutamyl-L-cysteine	L-threonine	L-threonine	sirolimus
2	sorafenib	linifanib	sirolimus	Sorafenib	L-threonine
3	sirolimus	PD-153035	sorafenib	L-serine	sorafenib
4	curcumin	ibrutinib	curcumin	curcumin	curcumin
4	curcumin	ibrutinib	curcumin	curcumin	curcumin
5	L-serine	ceritinib	L-serine	Sirolimus	selumetinib
6	vemurafenib	GSK690693	tandutinib	tandutinib	vemurafenib
7	Glu-Phe-Val	desmosine	L-phenylalanine	selumetinib	L-serine
8	selumetinib	L-seryl group	selumetinib	Glu-Phe-Val	trametinib
9	Phe-Asn	L-tyrosine	ombitasvir	vemurafenib	PLX-4720
10	Thr-Ser	Glu-Met	trametinib	Phe-Asn	bortezomib

In both cases, the graph of 72 drugs was divided into four clusters of varying sizes. It is noticeable that similar clusters were produced by the two algorithms. We can consider the following as equivalent to each other: CW_C1 and LM_C1, CW_C2 and LM_C2, CW_C3 and LM_C3, and CW_C4 and LM_C4. While each of the first two cluster pairs has only a one-node difference, the latter two pairs correspond to exactly the same clusters.

Table 1 presents the most prominent drugs within the entire graph while Table 2 provides a list of the same but only within each cluster. In both tables, drugs are ranked according to prominence, calculated based on our chosen graph centrality measures, as mentioned in Section 3.5. The three highest ranked drugs across the different centrality parameters are exactly the same corresponding to clusters from CW and LM except for one difference, shown in Table 2.

5 Analysis and Discussion

The thresholds which were applied to ensure that only drugs with considerable similarity were retained—described in Section 3—significantly reduced the number of drugs which formed the graph. Out of 831 drugs in our initial list, only 72 were included in the graph for clustering. As both CW and LM clustering algorithms are non-deterministic, they were applied on the data several times; results were consistent over the different runs.

To assess the quality of our clustering results, we firstly computed the value of the silhouette coefficient [21], a widely accepted metric for judging the quality

Table 2. Three most prominent drugs in each cluster obtained by Chinese Whispers and Louvain Method, according to five centrality measures. The cell corresponding to third rank of eigenvector-C2 is left empty as different results were obtained by CW and LM, i.e., 'L-threonine residue' and 'bortezomib', respectively.

Cluster ID	Degree	Closeness	Betweenness	PageRank	Eigenvector
C1	sorafenib	(-)-cubebin	sirolimus	sorafenib	sirolimus
	sirolimus	regorafenib	sorafenib	curcumin	sorafenib
	curcumin	cabozantinib	curcumin	sirolimus	curcumin
C2	L-threonine	desmosine	L-threonine	L-threonine	L-threonine
	L-serine	L-seryl group	L-serine	L-serine	L-serine
	L-phenylalanine	L-tyrosine	L-phenylalanine	L-phenylalanine	-
C3	Glu-Phe-Val	L-gamma-glutamyl-L-cysteine	ombitasvir	Glu-Phe-Val	Thr-Ser
	Phe-Asn	Glu-Met	Thr-Trp	Phe-Asn	Thr-Trp
	Thr-Ser	Ala-Asp-Pro	Thr-Ser	Thr-Ser	Glu-Phe-Val
C4	tandutinib	PD-153035	tandutinib	tandutinib	tandutinib
	afatinib	linifanib	afatinib	afatinib	afatinib
	ibrutinib	ibrutinib	ibrutinib	GSK690693	ibrutinib

of clusters where class labels are not predefined. Its value varies from -1 to 1, where positive values are taken to mean good clustering while anything less than zero is undesirable. Results obtained by CW and LM produced 0.294 and 0.292 as mean silhouette coefficient values, respectively. The cluster consistency is not very strong, but it is still substantial and falls within the acceptable range [12].

CW and LM use different techniques in generating clusters. While CW relies on random class distribution, LM is based on optimisation of modularity. We computed the value of Adjusted Rand Index (ARI), a measure of similarity between two clusterings, taking into account by-chance grouping [19]. The index can take a value that varies from -1 to 1, where -1 corresponds to lack of correlation while 1 pertains to a perfect match. We obtained 0.95 as the value of ARI for the clusterings produced by CW and LM, which is very high. This strengthens our confidence in the results, as it suggests that the drug clusters are consistent even across different algorithms.

Furthermore, our results were reviewed by a domain expert who was convinced by the results and suggested that it is worth pursuing further research into highlighting the dominant features for each cluster in order to uncover relevant mechanisms of action. He expressed confidence in the viability of the research and signified that it holds great potential.

6 Conclusion and Future Work

In this paper, we proposed a novel approach to highlighting similarities between drugs, according to features derived from scientific literature and ontologies. Two unsupervised clustering methods were employed, namely, Chinese Whispers and Louvain Method. Upon the application of our approach to 72 drugs, the same four clusters were independently produced by each of the clustering methods. We then obtained a list of the most prominent drugs within the entire graph of drugs as well as within each cluster.

The approach is a preliminary investigation and leaves room for improvement and future work. The most important next step is the experimental validation of the dominant MOA represented by each cluster using sources such as annotated data sets or manual annotation by domain experts. Other methods that could potentially improve the results include hierarchical clustering to explore sub-cluster relationships or soft clustering to take into account membership of a drug in multiple clusters. Feature engineering can also be extended, e.g., to increase tree depth when extracting terms from ontologies.

Acknowledgment. The authors would like to express their gratitude to the National Centre for Text Mining (NaCTeM) for sharing their event extraction results. The authors also thank Prof. Ross King for his help in evaluating the clustering results.

References

[1] M.P. Adams, C.Q. Urban, *Pharmacology: Connections to Nursing Practice* (Pearson Education, UK, 2015). ISBN 9780133896817. https://books.google.co.uk/books?id=usOgBwAAQBAJ

[2] A. Ahmad, N. Ormiston-Smith, P. Sasieni, Trends in the lifetime risk of developing cancer in great britain: comparison of risk for those born from 1930 to 1960. British journal of cancer **112**(5), 943 (2015)

[3] A.I. Archakov, V.M. Govorun, A.V. Dubanov, Y.D. Ivanov, A.V. Veselovsky, P. Lewi, P. Janssen, Protein-protein interactions as a target for drugs in proteomics. Proteomics **3**(4), 380–391 (2003)

[4] C. Biemann, Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems, in *Proceedings of the first workshop on graph based methods for natural language processing*, Association for Computational Linguistics, 2006, pp. 73–80. Association for Computational Linguistics

[5] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), 10008 (2008)

[6] M. Cascante, L.G. Boros, B. Comin-Anduix, P. de Atauri, J.J. Centelles, P.W.-N. Lee, Metabolic control analysis in drug discovery and disease. Nature biotechnology **20**(3), 243–249 (2002)

[7] G.O. Consortium, et al., The gene ontology project in 2008. Nucleic acids research **36**(suppl 1), 440–444 (2008)

[8] K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, M. Ashburner, Chebi: a database and ontology for chemical entities of biological interest. Nucleic acids research **36**(suppl_1), 344–350 (2007)

[9] A. Gemma, C. Li, Y. Sugiyama, K. Matsuda, Y. Seike, S. Kosaihira, Y. Minegishi, R. Noro, M. Nara, M. Seike, et al., Anticancer drug clustering in lung cancer based on gene expression profiles and sensitivity database. BMC cancer **6**(1), 174 (2006)

[10] Y. Hu, Efficient, high-quality force-directed graph drawing. Mathematica Journal **10**(1), 37–71 (2005)

[11] J. Jeon, J.H. Jeong, J.-H. Baek, H.-J. Koo, W.-H. Park, J.-S. Yang, M.-H. Yu, S. Kim, Y.K. Pak, Network clustering revealed the systemic alterations of mitochondrial protein expression. PLoS computational biology **7**(6), 1002093 (2011)

[12] L. Kaufman, P.J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, vol. 344 (John Wiley & Sons, ???, 2009)

[13] I. Kola, J. Landis, Opinion: Can the pharmaceutical industry reduce attrition rates? Nature reviews. Drug discovery **3**(8), 711 (2004)

[14] D. Lu, T. Lu, M. Stroh, R.A. Graham, P. Agarwal, L. Musib, C.-C. Li, B.L. Lum, A. Joshi, A survey of new oncology drug approvals in the usa from 2010 to 2015: a focus on optimal dose and related postmarketing activities. Cancer chemotherapy and pharmacology **77**(3), 459–476 (2016)

[15] C. McQueen, *Comprehensive Toxicology* (Elsevier Science, ???, 2010). ISBN 9780080468846. https://books.google.co.uk/books?id=jzCAKsa2CpMC

[16] M. Newman, *Networks:An Introduction* (OUP Oxford, UK, 2009). ISBN 9780191637766. https://books.google.co.uk/books?id=7LmNAQAACAAJ

[17] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: Bringing order to the web., Technical report, Stanford InfoLab, 1999

[18] R. Rak, A. Rowley, W. Black, S. Ananiadou, Argo: an integrative, interactive, text mining-based workbench supporting curation. Database **2012** (2012)

[19] W.M. Rand, Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association **66**(336), 846–850 (1971)

[20] D.T. Ross, U. Scherf, M.B. Eisen, C.M. Perou, C. Rees, P. Spellman, V. Iyer, S.S. Jeffrey, M. Van de Rijn, M. Waltham, et al., Systematic variation in gene expression patterns in human cancer cell lines. Nature genetics **24**(3), 227 (2000)

[21] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)

[22] L. Udrescu, L. Sbârcea, A. Topîrceanu, A. Iovanovici, L. Kurunczi, P. Bogdan, M. Udrescu, Clustering drug-drug interaction networks with energy model layouts: community analysis and drug repurposing. Scientific reports **6** (2016)

[23] K. Uhr, W.J. Prager-van der Smissen, A.A. Heine, B. Ozturk, M. Smid, H.W. Göhlmann, A. Jager, J.A. Foekens, J.W. Martens, Understanding drugs in breast cancer through drug sensitivity screening. SpringerPlus **4**(1), 611 (2015)

[24] C.R. UK, *Cancer Statistics for the UK*. Accessed: 2017-10-05. http://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk

[25] C. Zerva, R. Batista-Navarro, P. Day, S. Ananiadou, Using uncertainty to link and rank evidence from biomedical literature for model curation. Bioinformatics, 466 (2017)