

RESEARCH

Open Access



# A link prediction approach to cancer drug sensitivity prediction

Turki Turki<sup>1,2\*</sup> and Zhi Wei<sup>2\*</sup>

From The International Conference on Intelligent Biology and Medicine (ICIBM) 2016  
Houston, TX, USA. 08-10 December 2016

## Abstract

**Background:** Predicting the response to a drug for cancer disease patients based on genomic information is an important problem in modern clinical oncology. This problem occurs in part because many available drug sensitivity prediction algorithms do not consider better quality cancer cell lines and the adoption of new feature representations; both lead to the accurate prediction of drug responses. By predicting accurate drug responses to cancer, oncologists gain a more complete understanding of the effective treatments for each patient, which is a core goal in precision medicine.

**Results:** In this paper, we model cancer drug sensitivity as a link prediction, which is shown to be an effective technique. We evaluate our proposed link prediction algorithms and compare them with an existing drug sensitivity prediction approach based on clinical trial data. The experimental results based on the clinical trial data show the stability of our link prediction algorithms, which yield the highest *area under the ROC curve* (AUC) and are statistically significant.

**Conclusions:** We propose a link prediction approach to obtain new feature representation. Compared with an existing approach, the results show that incorporating the new feature representation to the link prediction algorithms has significantly improved the performance.

**Keywords:** Link prediction, Feature learning, Precision medicine, Cancer drug discovery, Applications in biology and medicine

## Background

Cancer has a significant global impact on public health; it is the second leading cause of death in the United States of America [1]. Cancer patients respond differently to potential drugs (i.e., chemotherapy) due to environmental causes, tumor heterogeneity, and genetic factors, making cancer drug discovery difficult [2–5]. The increasing number of deaths associated with cancer has attracted the attention of researchers from numerous domains, such as computational biology, machine learning, and data mining [6–9]. Costello et al. [10] assessed the performance of 44 drug sensitivity

prediction algorithms based on profiling datasets (i.e., genomic, proteomic, and epigenomic data) in breast cancer cell lines. The training set consists of 35 cell lines, in which each cell line is associated with 28 drug responses. The test set consists of 18 cell lines. The task of each prediction algorithm is to learn a model from the training cell lines and perform predictions on the test set. The predictions correspond to a ranking of the 28 drugs—from the most sensitive to the most resistant for each cell line on the test set. The top-performing approach [10] improved the performance by integrating several profiling datasets with improved representation with a probabilistic nonlinear regression model. The second-best performing approach employed random forest regression to make predictions on the test set. The prediction algorithms were evaluated using the weighted probabilistic *c*-index (wpc-index) and resampled Spearman

\* Correspondence: tturki@kau.edu.sa; zhiwei@njit.edu

<sup>1</sup>Department of Computer Science, King Abdulaziz University, P.O. Box 80221, Jeddah 21589, Saudi Arabia

<sup>2</sup>Bioinformatics Program and Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA

correlations [10]. The remaining prediction algorithms were not statistically different.

Geeleher et al. [11] proposed the following approach to drug sensitivity in which the input data are baseline expressions with drug  $IC_{50}$  values in cell lines and in vivo tumor gene expressions. The raw microarray data for the cell lines and clinical trials are processed separately and then combined and homogenized. The homogenized expression data consist of cell line expression data (i.e., baseline gene expression levels in the cell lines) and clinical trial expression data (i.e., baseline tumor expression data from the clinical trial). A learning algorithm is applied to the cell line expression data with the associated drug  $IC_{50}$  values for cell lines to learn a model. The resulting model is applied to clinical trial expression data to yield drug sensitivity predictions.

Two problems associated with the previous drug sensitivity prediction algorithms contribute to the degradation of the performance: (1) the poor quality of cell lines, especially when cell lines are not screened against all compounds [12]; and (2) the failure to adopt a new feature representation, because new feature representations provide a basis for improving the performance of learning algorithms [13–15].

In this paper, we model the cancer drug sensitivity as a link prediction problem, which is a classical research topic in computational social science [16–19] and biomedicine [20, 21]. Modeling the problem as link prediction enables us to exploit two link prediction algorithms: (1) the supervised link prediction algorithm, which aims to select better quality cancer cell lines; and (2) the extended supervised link prediction, which selects cancer cell lines and the top- $k$  genes (i.e., features) using state of the art CUR matrix decomposition [22]. Our experimental results indicate that the proposed link prediction algorithms outperform the baseline prediction algorithms proposed by Geeleher et al. [11].

The key contributions of our paper are as follows: 1) we represent cancer drug sensitivity as a link prediction problem, which to the best of our knowledge is the first robustly transfer cancer drug sensitivity prediction to link prediction, 2) we connect a social network domain to a health informatics domain for advancing health informatics, 3) we propose two link prediction algorithms, and 4) we perform an experimental study on clinical trial data to demonstrate the predictive power and stability of our proposed link prediction algorithms against the prediction algorithms that employ the current approach [11].

This paper is organized as follows: In Related works section, we review the relevant literature, which pertains to both link prediction and cancer drug sensitivity prediction. In Methods section, we describe how the cancer drug sensitivity problem can be modeled as a link

prediction problem. Then, we propose two link prediction algorithms that employ our link prediction approach: the supervised link prediction algorithm (A1) and the extended supervised link prediction algorithm (A2). Results and experiments section reports the experimental results and compares our link prediction algorithms against the baseline on the clinical trial data that pertains to breast cancer and multiple myeloma. Conclusions section summarizes our contributions in this paper.

## Related works

### Link prediction in gene regulatory networks

Given  $m$  genes, in which each gene has  $n$  expression values, we can denote their gene expression profiles by  $\mathbf{G} \in \mathbb{R}^{m \times n}$ , which contains  $m$  rows—each row corresponds to a gene—and  $n$  columns—each column corresponds to an expression value [23]. To learn a model, we need to know the regulatory relationships (i.e., labels) among the genes, which are stored in the matrix  $\mathbf{H} \in \mathbb{R}^{p \times 3}$ .  $\mathbf{H}$  contains  $p$  rows—each row shows a known regulatory relationship between two genes—and three columns. The first column shows the source gene (i.e., the transcription factor). The second column shows the target gene, and the third column shows the label, which is denoted as +1 (i.e., present link) when the source gene regulates the target gene or -1 (i.e., missing link) when the source gene does not regulate the target gene. Thus,  $\mathbf{H}$  represents the observed (i.e., known) gene regulatory network. To learn a model, we need to construct the training set  $\mathbf{D} \in \mathbb{R}^{p \times 2n+1}$ . The  $p$  examples in  $\mathbf{D}$  are constructed as follows: For each pair of genes with the associated label in matrix  $\mathbf{H}$ , the  $n$  expression values of each pair of genes in matrix  $\mathbf{G}$  are extracted, and the concatenation of the  $n$  expression values of each pair of genes and the corresponding label is performed. For example, consider the  $i$ th example in the training set  $\mathbf{D}$ , which is denoted by  $\mathbf{D}_i$  and defined as

$$\mathbf{D}_i = [g_i^1, g_i^2, \dots, g_i^n, g_j^1, g_j^2, \dots, g_j^n, y_i], \quad (1)$$

where  $g_i^1, g_i^2, \dots, g_i^n$  are the  $n$  expression values of  $\mathbf{g}_i$  (also called the expression profile of  $\mathbf{g}_i$ ),  $g_j^1, g_j^2, \dots, g_j^n$  are the  $n$  expression values of  $\mathbf{g}_j$ , and  $y_i \in \{1, -1\}$ . The  $i$ th example of the test set,  $\mathbf{T}$ , is denoted by  $\mathbf{T}_i$  and constructed as follows:

$$\mathbf{T}_i = [g_i^1, g_i^2, \dots, g_i^n, g_j^1, g_j^2, \dots, g_j^n], \quad (2)$$

where  $g_i^1, g_i^2, \dots, g_i^n$  are the  $n$  expression values of  $\mathbf{g}_i$ , and  $g_j^1, g_j^2, \dots, g_j^n$  are the  $n$  expression values of  $\mathbf{g}_j$ . These feature vector definitions have been used by the existing supervised inference of gene regulatory networks [23–28]. After constructing the feature vectors, the learning

algorithm is applied to  $\mathbf{D}$  to induce (i.e., learn) the model  $h$ . The resulting model is used to perform prediction on  $\mathbf{T}$ . The known regulations among genes enable using the induction principle to predict new regulations (i.e., labels): If gene  $\mathbf{g}_j$  has an expression profile that is similar to gene  $\mathbf{g}_i$ , which is known to be regulated by  $\mathbf{g}_i$ , then  $\mathbf{g}_j$  is likely to be regulated by  $\mathbf{g}_i$  [29]. Genes with similar expression profiles that are likely to be co-regulated have been used in the unsupervised clustering of expression profiles [30–32].

### Cancer drug sensitivity prediction

The gene expression profiles denoted by  $X \in \mathbb{R}^{p \times n}$ , which contains  $p$  rows—each row corresponds to a cell line or a sample—and  $n$  columns—each column corresponds to a gene.  $Y = (y_1, \dots, y_p)^T$  consists of the corresponding real-value drug responses (i.e., drug  $IC_{50}$  values) to  $X$ , where  $Y \in \mathbb{R}^p$  (i.e., the  $p$ -dimensional column vector).  $IC_{50}$  is defined as the concentration of a compound that is required to produce 50% cancer cell growth inhibition after 48 h of treatment [33]. A training set is defined as  $\mathbf{D} = \{(\mathbf{g}_i, y_i)\}_{i=1}^p$ , where  $\mathbf{g}_i \in X$  and  $y_i \in Y$ . Let the  $i$ th example of the training set  $\mathbf{D}$ , denoted by  $\mathbf{D}_i$ , be defined as

$$\mathbf{D}_i = [g_i^1, g_i^2, \dots, g_i^n, y_i], \tag{3}$$

where  $g_i^1, g_i^2, \dots, g_i^n$  represent the  $n$  genes of the cancer cell line  $\mathbf{g}_i$  (also called the expression profile of  $\mathbf{g}_i$ ), and  $y_i \in \mathbb{R}$  is the drug response value. The  $i$ th example of the test set  $\mathbf{T}$ , denoted by  $\mathbf{T}_i$ , is constructed as follows:

$$\mathbf{T}_i = [g_j^1, g_j^2, \dots, g_j^n]. \tag{4}$$

These feature vector definitions have been used by existing supervised cancer drug sensitivity prediction algorithms [9–11, 33–36]. A learning algorithm is applied to  $\mathbf{D}$  to induce model  $h$ , which is subsequently used to perform predictions on  $\mathbf{T}$ . Known cancer cell lines with associated drug responses enabled the use of the induction principle: If tumor  $\mathbf{g}_j$  has an expression profile similar to  $\mathbf{g}_i$ , then  $\mathbf{g}_j$  is likely to have a drug response value closer to the drug response value associated with  $\mathbf{g}_i$ .

### Methods

The fundamental task of cancer drug sensitivity prediction is to correctly predict the response of a tumor to the drug. This prediction is typically achieved based on *how closely* this tumor (also referred to as the test example) is related to a known cancer cell line with the associated drug response. Proximity, which is a measure of closeness, lies at the heart of both link prediction in gene regulatory networks and cancer drug sensitivity prediction [29, 37].

### Feature vector construction

To bridge link prediction and cancer drug sensitivity, we transform the feature representations of Eqs. (3) and (4) to the corresponding Eqs. (1) and (2) as follows: Let  $\{(\mathbf{g}_i, y_i)\}_{i=1}^p \subseteq \mathbf{D}$  be the cancer cell lines, where  $\mathbf{D} \in \mathbb{R}^{p \times n+1}$ ,  $b = p$ .

- 1 Find the  $k'$  nearest neighbors  $\mathbf{g}_1^*, \mathbf{g}_2^*, \dots, \mathbf{g}_{k'}^*$  of each  $\mathbf{g}_i$  in  $\mathbf{D}$ . (In this study  $k' = 1$ .)
- 2 Generate synthetic cell lines along the lines between the randomly selected  $k'$  nearest neighbors and each  $\mathbf{g}_i$  using the following lines of code:
  - 2.1 for  $i = 1$  to  $p$ 
    - 2.1.1 for  $j = 1$  to  $k'$ 
      - 2.1.1.1  $b = b + 1$
      - 2.1.1.2  $\mathbf{g}_b = \mathbf{g}_i + (\mathbf{g}_i^* - \mathbf{g}_i)\lambda$
      - 2.1.1.3 Store  $[\mathbf{g}_i, \mathbf{g}_b, y_i]$  in  $\mathbf{G}$
    - 2.1.2 end for
  - 2.2 end for

where the index  $b$  refers to only those synthetic cell lines (e.g.,  $\mathbf{g}_{p+1}$  when the index  $b = p + 1$ ) that differ from the cell lines in  $\mathbf{D}$ , whose indexes run from 1 to  $p$ ,  $\lambda = 0.3$ , and  $\mathbf{G} \in \mathbb{R}^{p \times 2n+1}$  is the new feature representation of the cell lines of the training set. Step 2.1.1.2 creates the synthetic cell line  $\mathbf{g}_b$ . Let  $\mathbf{G}_i$  be the  $i$ th row of  $\mathbf{G}$ , defined as

$$\mathbf{G}_i = [g_i^1, g_i^2, \dots, g_i^n, g_{p+1}^1, g_{p+1}^2, \dots, g_{p+1}^n, y_i], \tag{5}$$

where  $g_i^1, g_i^2, \dots, g_i^n$  represent  $n$  genes of the cancer cell line  $\mathbf{g}_i$ ,  $g_{p+1}^1, g_{p+1}^2, \dots, g_{p+1}^n$  represent the synthetic  $n$  genes of the synthetic cancer cell line  $\mathbf{g}_{p+1}$ , and  $y_i \in \mathbb{R}$  denotes that both  $\mathbf{g}_i$  and  $\mathbf{g}_{p+1}$  are linked by sharing the same drug response value. Let  $\{(\mathbf{g}_i, y_i)\}_{i=1}^q \subseteq \mathbf{T}$  be the test set of tumors, where  $\mathbf{T} \in \mathbb{R}^{q \times n}$ . Note that Steps 1–2 are similar to the Synthetic Minority Oversampling Approach (SMOTE) [38, 39], However, Step 2.1.1.3 is a different core step in which we increase the dimensionality (i.e., the number of features) instead of the size, as SMOTE does. We then apply the previous steps (i.e., Steps 1 and 2—changing Step 2.1 to  $i = 1$  to  $q$  and Step 2.1.1.3 to Store  $[\mathbf{g}_i, \mathbf{g}_b]$  in  $\mathbf{G}$ ) to  $\mathbf{T}$  to obtain  $\mathbf{G}' \in \mathbb{R}^{q \times 2n}$ .  $\mathbf{G}'$  is the new feature representation of the clinical trial expression data of the test set. Let  $\mathbf{G}'_i$  be the  $i$ th row of  $\mathbf{G}'$ , which is defined as

$$\mathbf{G}'_i = [g_j^1, g_j^2, \dots, g_j^n, g_{p+2k'+1}^1, g_{p+2k'+1}^2, \dots, g_{p+2k'+1}^n]. \tag{6}$$

where  $g_j^1, g_j^2, \dots, g_j^n$  represent  $n$  genes of tumor  $\mathbf{g}_j$ , and  $g_{p+2k'+1}^1, g_{p+2k'+1}^2, \dots, g_{p+2k'+1}^n$  represent  $n$  synthetic genes of the synthetic tumor  $\mathbf{g}_{p+2k'+1}$ . A learning algorithm is

called on the training set,  $\mathbf{G}$  to induce the model  $h$ , which is subsequently used to perform predictions on the test set  $\mathbf{G}'$ . The logic behind the mechanism of the induction principle is as follows: If the expression profiles of the pair of tumors  $(\mathbf{g}_j, \mathbf{g}_{p+2k'+1})$  are similar to those of the cell lines  $(\mathbf{g}_i, \mathbf{g}_{p+1})$ , then  $(\mathbf{g}_j, \mathbf{g}_{p+2k'+1})$  is likely to have a drug response value closer to the drug response value associated with  $(\mathbf{g}_i, \mathbf{g}_{p+1})$ . In machine learning terms, let  $(\mathbf{g}_i, \mathbf{g}_{p+1}, y_i) \in \mathbb{R}^{2n+1}$  be a row feature vector that encodes information about the pair of cancer cell lines  $(\mathbf{g}_i, \mathbf{g}_{p+1})$ . Given a new pair of tumors encoded by  $(\mathbf{g}_j, \mathbf{g}_{p+2k'+1})$ , if  $(\mathbf{g}_j, \mathbf{g}_{p+2k'+1})$  has feature values similar to  $(\mathbf{g}_i, \mathbf{g}_{p+1})$ , whose label is  $y_i$ , then  $(\mathbf{g}_j, \mathbf{g}_{p+2k'+1})$  is more likely to have a closer response (i.e., label) value to  $y_i$ .

**Notations and algorithms**

**Notations**

To provide a better understanding of our proposed prediction algorithms, the notations used throughout the remainder of this paper are summarized as follows: Matrices are denoted by boldface uppercase letters, e.g., matrix  $\mathbf{X}$ . We denote the row vectors of a matrix by boldface uppercase letters with a subscript, e.g.,  $\mathbf{X}_j$  is the  $j$ th row of matrix  $\mathbf{X}$ . Vectors are denoted by boldface lowercase letters, e.g., vector  $\mathbf{v}$ . Vector entries are denoted by italic lowercase letters with a subscript, e.g.,  $v_i$  is the  $i$ th entry of vector  $\mathbf{v}$ . The number of entries of a vector is denoted by the cardinality symbol, e.g.  $|\mathbf{v}|$  is the number of elements of vector  $\mathbf{v}$ . Scalars are denoted

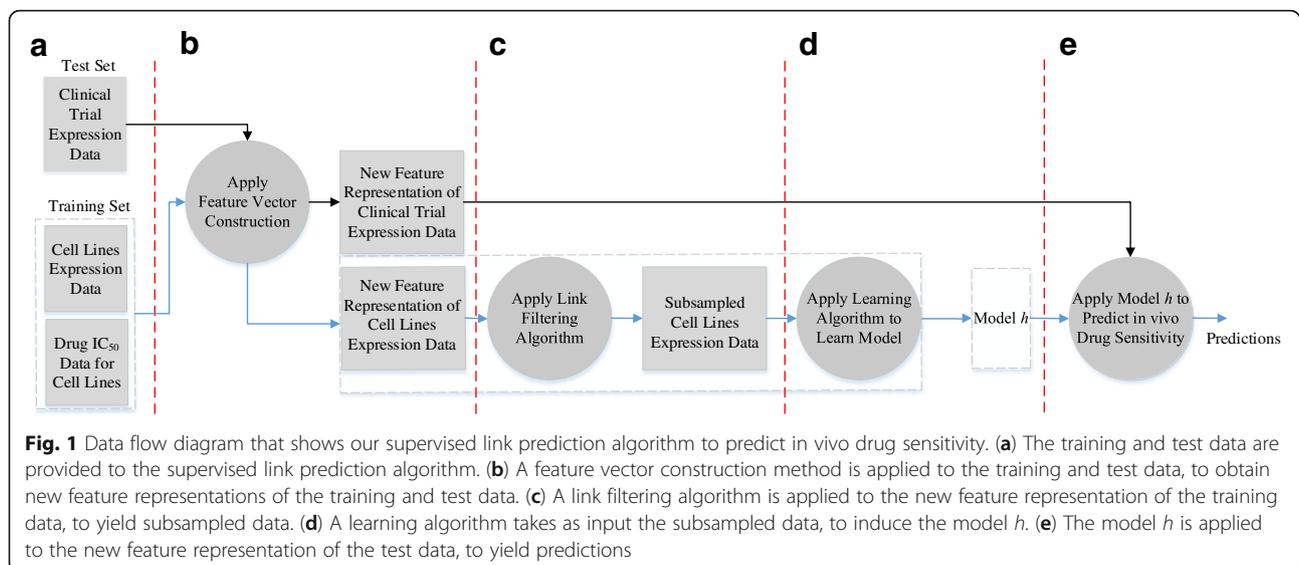
by italic lowercase letters, e.g.,  $m, f, f^*$ , and  $h$  are reserved letters, where  $f$  refers to a learning algorithm (e.g., SVR),  $f^*$  refers to an induced (i.e., learned) model, and  $h$  is an induced model used to perform predictions on the test set. We refer to specific learning algorithms and induced models using subscripts. For example,  $f_i$  ( $f_i^*$ , respectively) denotes the  $i$ th learning algorithm and induced model, respectively.

**The supervised link prediction algorithm (A1)**

Figure 1 outlines the supervised link prediction algorithm, which we designate A1, as follows. (a) Given a training set of cancer cell lines with associated drug responses  $\mathbf{D} \in \mathbb{R}^{p \times n+1}$  and a test set of tumors  $\mathbf{T} \in \mathbb{R}^{q \times n}$  that are described as in cancer drug sensitivity prediction subsection. (b) Transform  $\mathbf{D}$  and  $\mathbf{T}$  using the feature vector construction method described in feature vector construction subsection, to obtain a new feature representation  $\mathbf{G} \in \mathbb{R}^{p \times 2n+1}$  for the training set and a new feature representation  $\mathbf{G}' \in \mathbb{R}^{q \times 2n}$  for the test set. (c) Our link filtering method aims to select a better quality training set that works as follows: Each row (i.e., feature vector) in the new representations  $\mathbf{G}$  and  $\mathbf{G}'$  can be viewed as a cell line or tumor, represented by a  $2n$ -dimensional row vector when the drug responses of the training set  $\mathbf{G}$  are excluded. We weigh each cell line [40]  $\mathbf{g}_i$  in the training set  $\mathbf{G}$  by the minimum distance from the cell line  $\mathbf{g}_i$  to all tumors  $\mathbf{g}'_j$  in the testing set  $\mathbf{G}'$ :

$$w_i = \text{dist}(\mathbf{g}_i, \mathbf{g}'_{j^*}) \text{ with } j^* = \arg \min_{j \in \{1, \dots, q\}} \text{dist}(\mathbf{g}_i, \mathbf{g}'_j), \quad (7)$$

where  $\mathbf{g}_i \in \mathbb{R}^{2n}$ ,  $\mathbf{g}'_j \in \mathbb{R}^{2n}$ ,  $w_i$  is the weight assigned to  $\mathbf{g}_i$ ,



**Fig. 1** Data flow diagram that shows our supervised link prediction algorithm to predict in vivo drug sensitivity. (a) The training and test data are provided to the supervised link prediction algorithm. (b) A feature vector construction method is applied to the training and test data, to obtain new feature representations of the training and test data. (c) A link filtering algorithm is applied to the new feature representation of the training data, to yield subsampled data. (d) A learning algorithm takes as input the subsampled data, to induce the model  $h$ . (e) The model  $h$  is applied to the new feature representation of the test data, to yield predictions

and  $\text{dist}(\mathbf{g}_i, \mathbf{g}'_j)$  is the Euclidean distance. Let  $w = (w_1, w_2, \dots, w_p)$ . Then, we perform the following steps to select better quality training cell lines using our modified version of *Query by Committee* (QBC) [41–43]:

- 1 Let *med* be the median of the  $w$  vector of weights of each  $\mathbf{g}_i$  in  $\mathbf{G}$
- 2 Let  $\mathbf{X} = \{(\mathbf{g}_i, y_i) | (\mathbf{g}_i, y_i) \in \mathbf{G} \text{ and } w_i \leq \text{med}\}$
- 3 Let  $\mathbf{X}' = \{\mathbf{g}_i | \mathbf{g}_i \text{ in } \mathbf{G} \text{ and } w_i \leq \text{med}\}$
- 4 Let  $\mathbf{Z} = \{(\mathbf{g}_i, y_i) | (\mathbf{g}_i, y_i) \in \mathbf{G} \text{ and } w_i \geq \text{med}\}$
- 5 Let  $\mathbf{Z}' = \{\mathbf{g}_i | \mathbf{g}_i \text{ in } \mathbf{G} \text{ and } w_i \geq \text{med}\}$
- 6 Apply the learning algorithm,  $f_1$  or  $f_2$ , to  $\mathbf{X}$  or  $\mathbf{Z}$ , respectively, to induce the model  $f_1^*$  ( $f_2^*$ , respectively). (In this study, we chose ridge regression as the learning algorithm)
- 7 Apply the model  $f_1^*$  ( $f_2^*$ , respectively) to perform predictions on  $\mathbf{Z}'$  or  $\mathbf{X}'$ , respectively) and store predictions in  $\mathbf{v}$  or  $\mathbf{b}$  respectively)
- 8 Let  $q = |\mathbf{v}| = |\mathbf{b}|$
- 9 Let  $\mathbf{P} = (\mathbf{v}, \mathbf{b})^T$
- 10 Let  $\mathbf{r} = \{y_i | y_i \text{ in } \mathbf{Z}\}$  and  $\mathbf{e} = \{y_i | y_i \text{ in } \mathbf{X}\}$
- 11 Let  $\mathbf{R} = (\mathbf{r}, \mathbf{e})^T$
- 12  $j^* = \arg \max_{j \in \{1,2\}} \frac{1}{q} (\mathbf{P}_j - \mathbf{R}_j)^2$
- 13  $\mathbf{S} = \begin{cases} \mathbf{X} & \text{if } j^* = 1 \\ \mathbf{Z} & \text{otherwise} \end{cases}$
- 14  $\mathbf{U} = \begin{cases} \mathbf{Z} & \text{if } j^* = 1 \\ \mathbf{X} & \text{otherwise} \end{cases}$

QBC aims to partition the training set  $\mathbf{G}$  into  $\mathbf{S}$  and  $\mathbf{U}$ , where  $\mathbf{S}$  or  $\mathbf{U}$  is treated as the labeled or unlabeled set, respectively. QBC is accompanied by two major items: (1) the set of models (i.e., the committee) that are consistent with all labeled cell lines in  $\mathbf{S}$ ; and (2) given the unlabeled set,  $\mathbf{U}$ , the QBC applies the models (i.e., the committee) to  $\mathbf{U}$  to select the unlabeled tumor that maximizes the disagreement because it represents the most important tumor that will be added to  $\mathbf{S}$ , in addition to querying the drug response value associated with the tumor. The main obstacle of the first major step of QBC is to find models that agree on all the labels of set  $\mathbf{S}$  with reasonable computational complexity [43]. Thus, we relax the first major step according to Steps 1–14, where relaxation is practiced to address the first major step [41]. Steps 1–5 partition the training set into

$\mathbf{X}$  and  $\mathbf{Z}$  using the median as a threshold, where  $\mathbf{X}$  or  $\mathbf{Z}$  contains cell lines from  $\mathbf{G}$  that are near or far, respectively, from the test set  $\mathbf{G}'$ . Steps 6–14 aim to assign the set of cell lines where the model incurred fewer errors (or more errors, respectively) to  $\mathbf{S}$  or  $\mathbf{U}$ , respectively. The logic behind these steps (i.e., Steps 13–14) is that we want  $\mathbf{S}$  or  $\mathbf{U}$ , respectively, to contain the set of cell lines that are more or less, respectively, correctly labeled by one model (i.e., one member of the committee). Steps 1–14 are motivated by other QBC approaches [41–43], in which the success of the second major step of QBC is dependent on the first major step.

- 15 Repeat  $k$  times
  - 15.1 Apply the learning algorithms  $f_1, f_2, \dots, f_t$  on  $\mathbf{S}$  to induce the models (i.e., committee)  $f_1^*, f_2^*, \dots, f_t^*$ . (In this study,  $t = 3$ , and the learning algorithms include support vector regression with a linear kernel (SVR + L), SVR with a polynomial kernel of degree 5, and SVR with a sigmoid kernel (SVR + S))
  - 15.2 Let  $w'_i$  be the weight of the  $i$ th model  $f_i^*$  where  $\mathbf{w}' = \sum_{i=1}^t w'_i = 1$ . (In this study,  $t = 3$  and  $w'_1 = w'_2 = w'_3 = \frac{1}{3}$ )
  - 15.3 For each  $\mathbf{g}_j$  in  $\mathbf{U}$ , let  $f'(\mathbf{g}_j) = \sum_{i=1}^t w'_i f_i^*(\mathbf{g}_j)$  where  $f_i^*(\mathbf{g}_j)$  is the prediction of the  $i$ th learned model on the  $j$ th cell line  $\mathbf{g}_j$ , and  $f'(\mathbf{g}_j)$  is the weighted ensemble average of the  $j$ th cell line  $\mathbf{g}_j$ .
  - 15.4 Find the cell line  $\mathbf{g}_{j^*}$  that maximizes the disagreement:
    - 15.4.1.  $j^* = \arg \max_{j \in \{1, \dots, |\mathbf{v}|\}} \sum_{i=1}^t w'_i (f_i^*(\mathbf{g}_j) - f'(\mathbf{g}_j))^2$
  - 15.5 Find the label  $y_{j^*}$  of  $\mathbf{g}_{j^*}$  in  $\mathbf{U}$
  - 15.6 Add the pair  $(\mathbf{g}_{j^*}, y_{j^*}) \in \mathbf{U}$  to  $\mathbf{S}$  and remove the pair  $(\mathbf{g}_{j^*}, y_{j^*})$  from  $\mathbf{U}$
  - 15.7 Update  $|\mathbf{v}| = |\mathbf{v}| - 1$
- 16 Return  $\mathbf{S}$

Steps 15.1–15.4.1 return the index of the cell line in set  $\mathbf{U}$  that maximizes the disagreement, where disagreement is defined in Step 15.4.1 [44]. Then,  $(\mathbf{g}_{j^*}, y_{j^*})$  is added to or removed from  $\mathbf{S}$  or  $\mathbf{U}$  respectively, as shown in Steps 15.5–15.6. (In this study,  $k = 5$ .) Step 15.7 updates  $|\mathbf{v}|$  as the size of  $\mathbf{U}$  is reduced after each iteration.  $\mathbf{S}$  (Step 16) is the returned set that will be used as the training set. (d) We apply a learning algorithm on  $\mathbf{S}$  to induce the model  $h$ . Finally (i.e., (e in Fig. 1)), we apply model  $h$  to perform predictions on the test set  $\mathbf{G}$  (i.e., the set of new feature representations of the clinical trial expression data). In the remainder of this

paper, we refer to the supervised link prediction algorithms that employ the following machine learning algorithms (SVR and RR) as: A1 + SVR + L, A1 + SVR + S, and A1 + RR (abbreviations are listed in Table 1).

#### The extended supervised link prediction algorithm (A2)

Figure 2 shows the data flow diagram of the extended supervised link prediction (A2). Steps (a), (b), and (c) are the same as Steps (a), (b), and (c) of the supervised link prediction algorithm. (d) Mahoney et al. [22] proposed CUR matrix decomposition as a dimensionality reduction paradigm that aims to obtain a low rank approximation of matrix  $\mathbf{S}$ , which is expressed in terms of the actual rows and columns of the original matrix  $\mathbf{S}$ :

$$\mathbf{S} \approx \mathbf{C}\mathbf{U}\mathbf{R}, \quad (8)$$

where  $\mathbf{C}$  consists of a small number of the actual columns of  $\mathbf{S}$ ,  $\mathbf{R}$  consists of a small number of the actual rows, and  $\mathbf{U}$  is a constructed matrix that guarantees that  $\mathbf{C}\mathbf{U}\mathbf{R}$  is close to  $\mathbf{S}$ . We select  $k$  genes based on their importance score (refer to Equation 9), which depends on matrix  $\mathbf{S}$  and the input rank parameter  $l$  (in this study, we used the default parameter value for  $l$  in CUR function [45].) If  $v_j^\xi$  is the  $j$ -th element of the  $\xi$ -th right singular vector of  $\mathbf{S}$ , then the normalized statistical leverage scores are equal to

$$\pi_j = \frac{1}{l} \sum_{\xi=1}^l \left( v_j^\xi \right)^2 \quad (9)$$

for all  $j = 1..2n$ , and  $\sum_{j=1}^{2n} \pi_j = 1$ . Statistical leverage scores have been successfully employed in data analysis to identify the most influential genes and outlier detection [22]. A high statistical leverage score for a given gene indicates that the gene is regarded as an important (i.e., influential) gene. A low statistical leverage score for a given gene indicates that the gene is regarded as a less

important gene. We store the indexes of the highest  $k$  leverage scores in  $\mathbf{I}$ ; these correspond to the positions of the  $k$  most influential genes in matrix  $\mathbf{S}$ . We select  $k$  genes from the training set  $\mathbf{S}$  using their positions in  $\mathbf{I}$  and store subsampled cell line expression data with  $k$  genes in  $\mathbf{S}'$ . (e) A learning algorithm is called on  $\mathbf{S}'$  to induce model  $h$ . (f) The  $k$  genes in the test set  $\mathbf{G}'$  are selected using their positions in  $\mathbf{I}$  and stored in  $\mathbf{G}''$ . Model  $h$  is applied on the test set  $\mathbf{G}''$  to perform predictions. We refer to the extended supervised link prediction algorithms that employ machine learning algorithms as A2 + SVR + L, A2 + SVR + S, and A2 + RR (see Table 1).

## Results

We empirically evaluate our proposed approach and compare it against the baseline approach proposed by Geeleher et al. [11] on clinical trial datasets. This section first describes the datasets and experimental methodology and presents the experimental results.

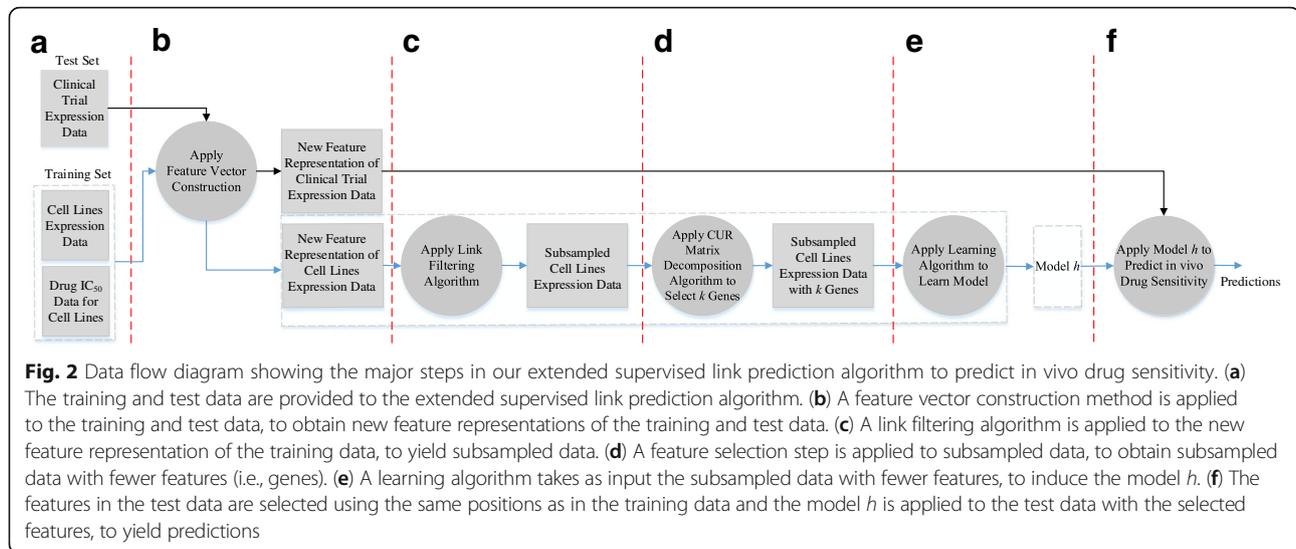
### Datasets

#### Data pertaining to breast cancer

The training set  $\mathbf{D} \in \mathbb{R}^{482 \times 6539}$  contains 482 cancer cell lines, 6538 genes, and drug  $IC_{50}$  values that correspond to a 482-dimensional column vector. The test set  $\mathbf{T} \in \mathbb{R}^{24 \times 6538}$  consists of 24 breast cancer tumors and 6538 genes. The drug  $IC_{50}$  values for docetaxel (a chemotherapy drug) [46, 47] were downloaded from (<http://gene.med.uchicago.edu/~pgeeleher/cgpPrediction/>). The cell line expression data were downloaded from the ArrayExpress repository [48] (accession number is E-MTAB-783, also available at <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-783/?query=EMTAB783>). The clinical trial data corresponding to the test set were downloaded from the Gene Expression Omnibus (GEO) repository (<http://www.ncbi.nlm.nih.gov/geo/>) with accession numbers GSE350 and GSE349 [49–51]. The data with accession numbers GSE350 and GSE349 contain

**Table 1** Abbreviations of the drug sensitivity prediction algorithms

Abbreviation	Prediction Algorithm
A1 + SVR + L	The supervised link prediction algorithm using support vector regression with a linear kernel
A1 + SVR + S	The supervised link prediction algorithm using support vector regression with a sigmoid kernel
A1 + RR	The supervised link prediction algorithm using ridge regression
A2 + SVR + L	The extended supervised link prediction algorithm using support vector regression with a linear kernel
A2 + SVR + S	The extended supervised link prediction algorithm using support vector regression with a sigmoid kernel
A2 + RR	The extended supervised link prediction algorithm using ridge regression
B + SVR + L	The baseline approach using support vector regression with a linear kernel
B + SVR + S	The baseline approach using support vector regression with a sigmoid kernel
B + RR	The baseline approach using ridge regression



10 and 14 samples, respectively. If the remaining tumor was  $<25\%$  or  $\geq 25\%$ , a breast cancer patient is considered to be sensitive or resistant, respectively, to docetaxel treatment. All the data were downloaded and processed according to the approach proposed by Geeleher et al. [11].

#### Data pertaining to multiple myeloma

The training set  $\mathbf{D} \in \mathbb{R}^{280 \times 9115}$  contains 280 cancer cell lines, 9114 genes, and drug  $IC_{50}$  values that correspond to a 280-dimensional column vector. The test set  $\mathbf{T} \in \mathbb{R}^{188 \times 9114}$  is composed of 188 multiple myeloma patients and 9114 genes. The drug  $IC_{50}$  values for bortezomib [52, 53] were downloaded from (<http://genemed.uchicago.edu/~pgeeheher/cgpPrediction/>), and the data for the cancer cell lines were downloaded from the ArrayExpress repository (accession number is E-MTAB-783 or available at <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-783/?query=EMTAB783>). The clinical trial data corresponding to the test set were downloaded from the Gene Expression Omnibus (GEO) repository (<http://www.ncbi.nlm.nih.gov/geo/>) with accession number GSE9782 [54]. The data were downloaded, processed and mapped according to Geeleher et al. [11].

#### Data pertaining to non-small cell lung cancer and triple-negative breast cancer

The training sets correspond to an  $258 \times 9508$  matrix and an  $497 \times 9621$  matrix for non-small cell lung cancer and triple-negative breast cancer, respectively. The test sets correspond to an  $25 \times 9507$  matrix (excluding labels) and an  $24 \times 9620$  matrix (excluding labels) for non-small cell lung cancer and triple-negative breast cancer, respectively. The data were downloaded from (<http://genemed.uchicago.edu/~pgeeheher/cgpPrediction/>) [11].

#### Experimental methodology

Kernel-based methods, such as SVM and support vector regression (SVR), are popular machine learning algorithms and exhibit state-of-art performance in many applications [55, 56], including biological fields [57]. Therefore, in our experiments, we used SVR with linear kernel (SVR + L) and sigmoid kernel (SVR + S) as machine learning algorithms, coupled with our proposed link prediction algorithms (A1 or A2). We also employed our proposed link prediction algorithms with linear ridge regression (RR). In total, we considered 9 drug sensitivity prediction algorithms, as summarized in Table 1.

Each prediction algorithm was trained on the same training set, whose labels are continuous to yield models (see Methods section). Then, each model is applied to the same test set to yield predictions, as discussed in Methods section. The test set consists of the clinical trial expression data of patients, including baseline tumor expression data from primary tumor biopsies prior to treatment with an anticancer drug. The responses (i.e., labels) of the test set are categorical (e.g., either “sensitive” or “resistant”). These labels were clinically evaluated by the degree of reduction in tumor size to the given drug [11].

To evaluate whether the proposed approach exhibits stable superior performance as the sample size changes, we gradually reduced the sample size for the training set by 1 to 4% in each run. That is, we have 5 runs with sample sizes of 482, 478, 473, 468, and 463 and 280, 278, 275, 272, and 269 for the two datasets, respectively.

The accuracy of the prediction algorithms is measured using the *Area Under the ROC Curve* (AUC), as shown in [11]. The higher AUC an algorithm has, the better performance that algorithm achieves. We denote the mean of the AUC values averaged over the five runs of the test set as the MAUC. A run of the test set is

defined as predictions of a learned model on the test set, such that the model is learned from the training set. The size of this training set is varied to assess the stability of prediction algorithms, in which a stable prediction algorithm is one for which the prediction accuracy on the test set does not change dramatically due to small changes in the size of the training set [58, 59]. This type of assessment is important in biological systems, in which the best prediction algorithm outperforms other algorithms many times in the conducted experiments. Statistical significance is measured between all pairs of the prediction algorithms.

The software employed in this study included support vector regressions with linear and sigmoid kernels in the LIBSVM package [60], ridge regression [11], gene selection using CUR and topLeverage functions in the rCUR package [45], and R code for processing the datasets and performance evaluation [11]. We used R to write the code for the link prediction algorithms and perform the experiments.

### Experimental results

Tables 2 and 3 show the AUC of 9 docetaxel and bortezomib, respectively, sensitivity prediction algorithms on clinical breast cancer or multiple myeloma trial data. For each variation in training set size the prediction algorithm with the best performance (i.e., the highest AUC) on the clinical trial data is shown in bold.

Table 2 shows that our prediction algorithms perform better than the baseline prediction algorithms (i.e., B + SVR + L and B + SVR + S) including B + RR, which is a prediction algorithm proposed by Geeleher et al. Row “m” and “d”, shows the number of cell lines or genes, respectively, in the training set that were provided to each prediction algorithm. We provided the same training set to

**Table 2** AUC scores of docetaxel sensitivity prediction algorithms in breast cancer patients on the test set

	m	478	473	468	463	MAUC
d	6538	6538	6538	6538	6538	–
A1 + SVR + L	0.878	0.864	<b>0.871</b>	0.857	0.871	0.868
A1 + SVR + S	0.871	0.857	0.814	0.828	<b>0.878</b>	0.849
A1 + RR	0.850	0.828	0.821	0.850	0.842	0.838
m + A1	246	244	242	239	237	–
d + A1	13,076	13,076	13,076	13,076	13,076	–
A2 + SVR + L	<b>0.892</b>	0.857	0.864	<b>0.864</b>	0.864	0.868
A2 + SVR + S	0.871	0.850	0.814	0.814	<b>0.878</b>	0.845
A2 + RR	0.857	0.842	0.835	0.835	0.835	0.841
m + A2	246	244	242	239	237	–
d + A2	13,000	13,000	13,000	13,000	13,000	–
B + SVR + L	0.835	0.814	0.800	0.821	0.835	0.821
B + SVR + S	0.842	<b>0.871</b>	0.864	0.857	0.857	0.858
B + RR	0.814	0.814	0.821	0.821	0.821	0.818

The algorithm with the highest AUC is shown in bold. MAUC = mean AUC

**Table 3** AUC scores of bortezomib sensitivity prediction algorithms in multiple myeloma patients on the test set

	m	278	275	272	269	MAUC
d	9114	9114	9114	9114	9114	–
A1 + SVR + L	0.668	0.669	0.665	0.663	0.656	0.664
A1 + SVR + S	0.638	0.623	0.637	0.642	0.662	0.640
A1 + RR	0.685	0.673	0.679	0.677	0.690	0.681
m + A1	145	144	143	141	140	–
d + A1	18,228	18,228	18,228	18,228	18,228	–
A2 + SVR + L	0.678	0.678	0.671	0.668	0.654	0.670
A2 + SVR + S	0.661	0.657	0.659	0.659	0.668	0.661
A2 + RR	<b>0.686</b>	<b>0.689</b>	<b>0.696</b>	<b>0.695</b>	<b>0.699</b>	0.693
m + A2	145	144	143	141	140	–
d + A2	9114	9114	9114	9114	9114	–
B + SVR + L	0.613	0.609	0.622	0.628	0.632	0.621
B + SVR + S	0.602	0.600	0.601	0.605	0.598	0.601
B + RR	0.614	0.611	0.603	0.607	0.606	0.608

The algorithm with the highest AUC is shown in bold. MAUC = mean AUC

each prediction algorithm. Rows “m + A1” and “m + A2”, or “d + A1” and “d + A2” show the number of selected cell lines or genes, respectively, that were used in the prediction algorithms that employed our approach for learning the models. The results of our prediction algorithms are dominant compared with the baseline prediction algorithms that employ clinical trial data of breast cancer in terms of the AUC of four runs and the MAUC. In contrast to the baseline prediction algorithms, the performance of our prediction algorithms on the test set outperforms in terms of the AUC when we reduce the training set size.

Table 3 shows that our prediction algorithms perform better than the baseline prediction algorithms (i.e., B + SVR + L and B + SVR + S) and B + RR, which is a prediction algorithm proposed by Geeleher et al. Row “m” or “d”, respectively, shows the number of cell lines or genes, respectively, in the training set that were provided to each prediction algorithm. We provided the same training set to each prediction algorithm. Rows “m + A1” and “m + A2” or “d + A1” and “d + A2” show the number of selected cell lines or genes, respectively, used in the prediction algorithms that employ our approach for learning the models. The results of our prediction algorithms are dominant compared with the baseline prediction algorithms on the multiple myeloma clinical trial data in terms of the AUC of each run and the MAUC. In particular, A2 + RR achieves the highest mean AUC (MAUC) of 0.693 and performed the best in all runs. In contrast to the baseline prediction algorithms, the performance of A2 + RR on the test results in the best AUC as we reduce the training set size, which indicates that A2 + RR has a stable performance.

Table 4 shows the *p*-values of the two-tailed Wilcoxon signed rank test [61, 62] to measure the statistical

**Table 4** *P*-values of Wilcoxon signed rank test (two-tailed) between all pairs of prediction algorithms

	A1 + SVR + S	A1 + RR	A2 + SVR + L	A2 + SVR + S	A2 + RR	B + SVR + L	B + SVR + S	B + RR
A1 + SVR + L	<b>0.0160</b>	0.5092	0.3077	0.0836	0.8807	<b>0.0051</b>	<b>0.0149</b>	<b>0.0051</b>
A1 + SVR + S	–	0.1675	<b>0.0208</b>	0.1282	0.0929	<b>0.0051</b>	0.1830	<b>0.0080</b>
A1 + RR	–	–	0.2846	0.5418	0.0672	<b>0.0051</b>	0.1388	<b>0.0076</b>
A2 + SVR + L	–	–	–	0.0587	0.5754	<b>0.0051</b>	<b>0.0207</b>	<b>0.0047</b>
A2 + SVR + S	–	–	–	–	0.1388	<b>0.0069</b>	0.0836	<b>0.0124</b>
A2 + RR	–	–	–	–	–	<b>0.0076</b>	0.1675	<b>0.0051</b>
B + SVR + L	–	–	–	–	–	–	0.5754	0.1609
B + SVR + S	–	–	–	–	–	–	–	0.2040

Values with statistical significance ( $p < 0.05$ ) are shown in bold

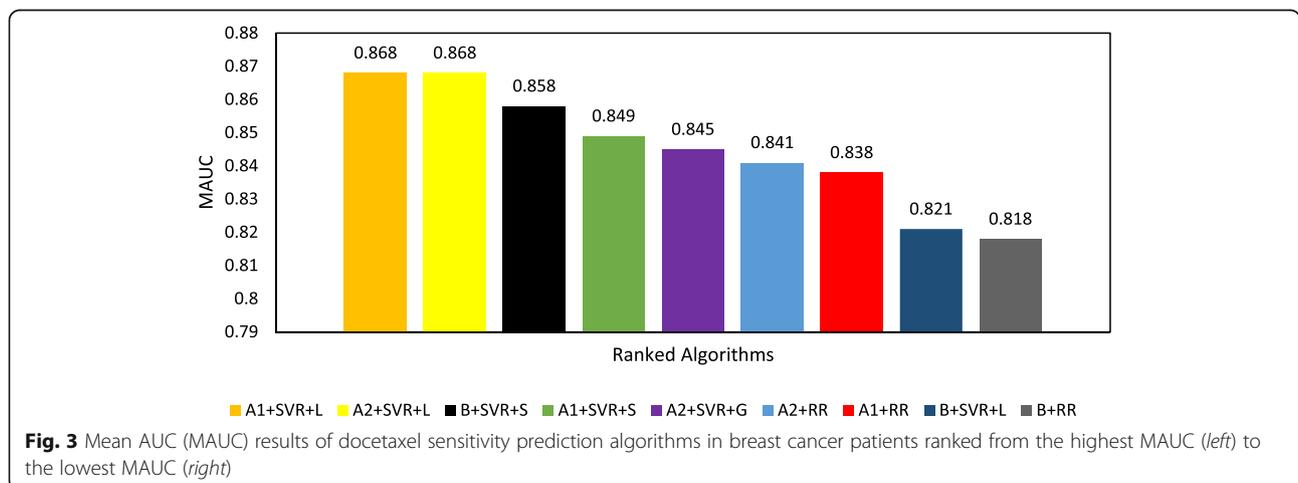
significance between the prediction algorithms using clinical trial data of breast cancer and multiple myeloma patients. The  $p$ -values indicate that our A1 + SVR + L and A2 + SVR + L prediction algorithms significantly outperformed the baseline prediction algorithms B + SVR + L, B + SVR + S, and B + RR. The remaining prediction algorithms that employ our approach are not statistically different from B + SVR + S.

Figures 3 and 4 show the ranking of all prediction algorithms from the highest to the lowest MAUC using clinical trial data pertaining to breast cancer and multiple myeloma patients, respectively. Each MAUC is calculated over the 5 runs of the clinical trial data. As shown in Figs. 3 and 4, our prediction algorithms outperform the baseline prediction algorithms [11] w.r.t the MAUC.

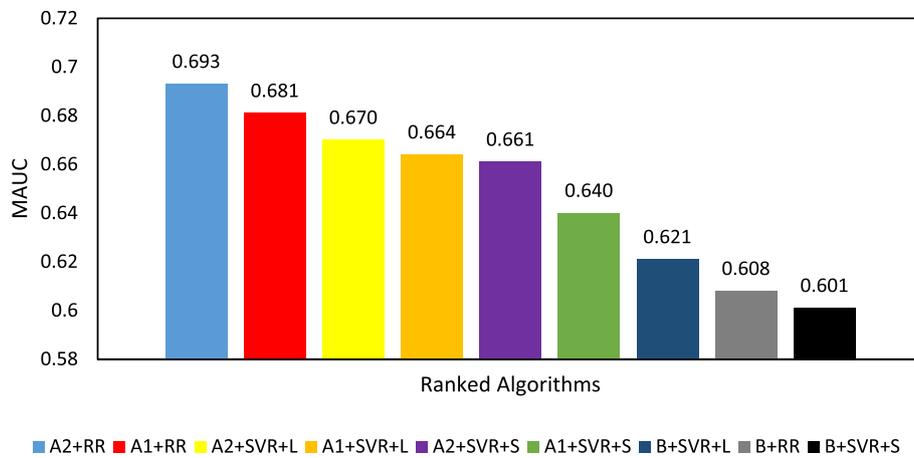
Figure 5 shows the predictions of three prediction algorithms on the test set (clinical data samples of 24 breast cancer patients) when the prediction algorithms were trained on a dataset with the size  $m = 482$  (i.e., the complete training set without any reductions). Figure 5a–c show the predictions of A2 + SVR + L, A1 + SVR + L and B + SVR + S, respectively. For A2 + SVR + L in Fig. 5a, the difference between the predicted drug

sensitivity in breast cancer patients was highly statistically significant ( $P=472 \times 10^{-6}$  from the result of a  $t$ -test) between the trial-defined sensitive and resistant groups. The result of A1 + SVR + L in Fig. 5b was also highly statistically significant ( $P=614 \times 10^{-6}$  from a  $t$ -test). B + SVR + S in Fig. 5c achieved statistical significance ( $P=1176 \times 10^{-6}$  from a  $t$ -test). Higher sensitivity or higher resistance, respectively, denote the greater or lesser effectiveness of the drug. In Fig. 5d, the ROC reveals AUC values of 0.892, 0.878 and 0.842 for A2 + SVR + L, A1 + SVR + L, and B + SVR + S, respectively, as shown in Table 2.

In Fig. 6, the predictions of three prediction algorithms are reported on the test set (clinical trial data of 188 multiple myeloma samples of patients) when prediction algorithms learned models from a training set of size  $m = 280$  (i.e., the training set without any reductions). Figure 6a–c show the predictions of the A2 + RR, A1 + RR, and B + RR, algorithms, respectively. For A2 + RR (Fig. 6a), the difference between the predicted drug sensitivity in multiple myeloma patients was highly significant ( $P=8 \times 10^{-6}$  from a  $t$ -test) between trial-defined responder groups and non-responder groups.



**Fig. 3** Mean AUC (MAUC) results of docetaxel sensitivity prediction algorithms in breast cancer patients ranked from the highest MAUC (left) to the lowest MAUC (right)

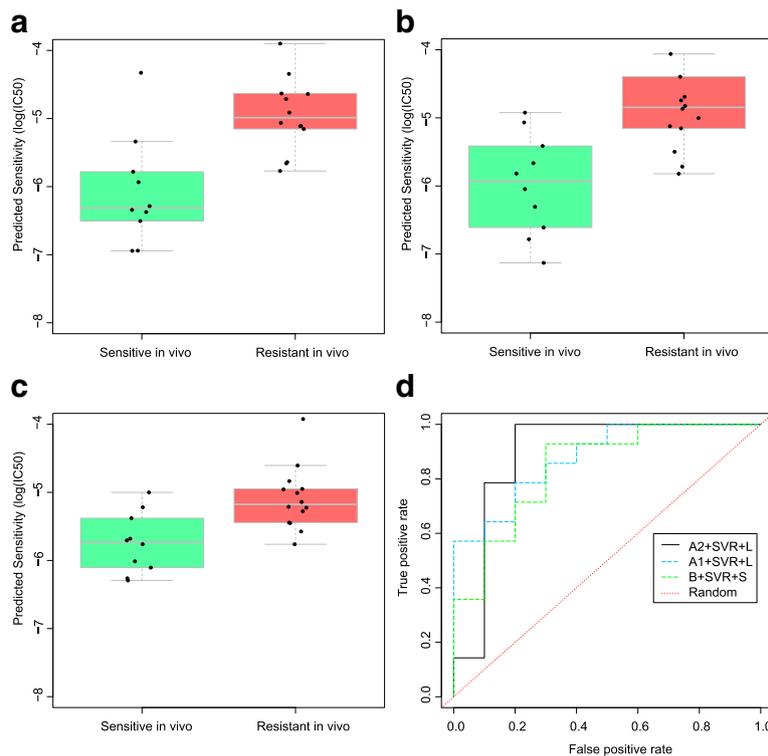


**Fig. 4** Mean AUC (MAUC) of bortezomib sensitivity prediction algorithms in multiple myeloma patients ranked from highest MAUC (left) to lowest MAUC (right)

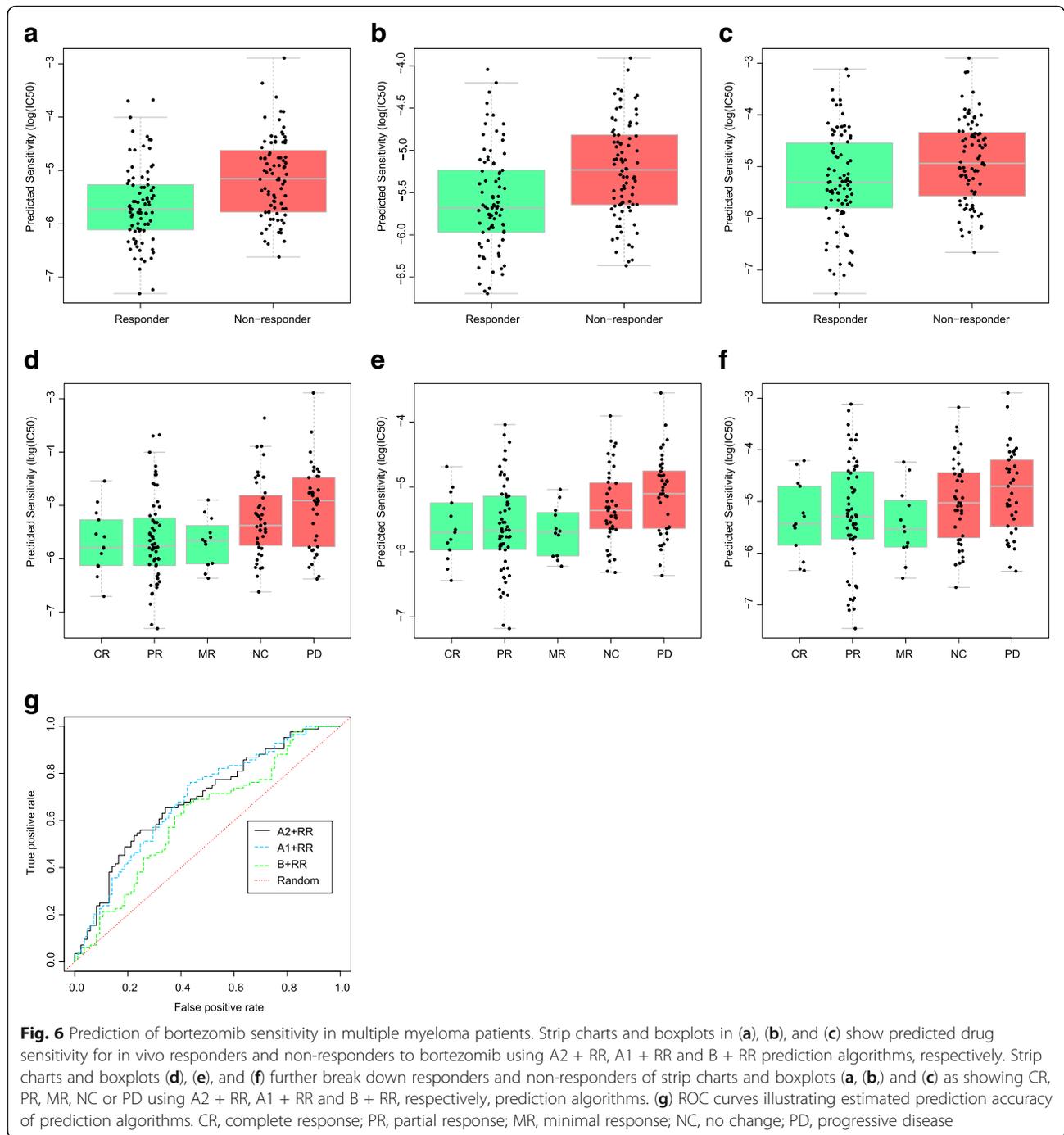
The result of A1 + RR was also highly significant ( $P=11 \times 10^{-6}$  from a *t*-test), while B + RR achieved statistically significant result ( $P=2612 \times 10^{-6}$  from a *t*-test). Figure 6d–f break down the responders and non-responders of Fig. 6a–c, respectively, to CR, PR, MR, NC or PD. In Fig. 6g, The ROC reveals AUCs of 0.686,

0.685, and 0.614 for A2 + RR, A1 + RR, and B + RR, respectively, as shown in Table 3.

We also evaluated the performance of prediction algorithms on the clinical trial data pertaining to non-small cell lung cancer patients and the triple-negative breast cancer patients. We observed similar results that our



**Fig. 5** Prediction of docetaxel sensitivity in breast cancer patients. Strip charts and boxplots in (a), (b), and (c) show the differences in predicted drug sensitivity for individuals who are sensitive or resistant to docetaxel treatment using the prediction algorithms A2 + SVR + L, A1 + SVR + L and B + SVR + S, respectively, while (d) shows the ROC curves of prediction algorithms, revealing the proportion of true positives compared to the proportion of false positives. ROC = receiver operating characteristics



prediction algorithms noticeably outperform the baseline prediction algorithms (See Additional file 1: Tables S1 and S2).

It is worth mentioning that we also assessed the performance of other machine learning algorithms, including random forests [63], support vector regression with a polynomial kernel of degree 2, and support vector regression

with a Gaussian kernel. Moreover, we applied other dimensionality reduction methods such as principal component analysis (PCA) [64] based on the prcomp package in R [65], sparse PCA [66, 67], non-negative and sparse cumulative PCA, and negative and sparse PCA [68, 69]. However, they did not exhibit acceptable predictive performance; consequently, their results are not included in this paper.

## Discussion

Gene (feature) selection is important to the success of the proposed method. After many years of biomedical research, some signaling pathways have been known for being implicated in various cancers. It is tempting to exploit this pathway information for feature selection. For example, we might consider adding the signaling pathways as a constraint to get reliable feature sets. Consequently, we assessed the performance of the proposed prediction algorithms using only the genes in the signaling pathways that are known to the cancers. We obtained inferior results (See Additional file 2 for details). It is noted that the current pathway information is limited. If we consider only those signaling genes, we may miss those important genes not identified yet by domain knowledge. This may hurt the overall performance as shown in our case. Therefore, a better strategy may be to include all genes but assign more weights to those signaling pathway genes. This is an interesting direction, and we leave it to our future work.

## Conclusion

In this paper, we introduce a link prediction approach to cancer drug sensitivity prediction. The benefit of introducing a link prediction approach is to obtain satisfactory feature representation for better prediction performance. We propose two algorithms that employ the link prediction approach: (1) A supervised link prediction algorithm, which selects better quality training cancer cell lines using a modified version of QBC; and (2) An extended supervised link prediction, which selects both better training cancer cell lines and a subset of important genes using state of the art CUR matrix decomposition.

In our study, the link prediction algorithms use two machine learning algorithms: support vector regression and ridge regression. The experimental results demonstrate the stability of the proposed link prediction algorithms, which outperform drug sensitivity prediction algorithms of an existing approach as measured by their higher and statistically significant AUC scores.

## Additional files

**Additional file 1:** Performance evaluation of prediction algorithms on clinical trial data pertaining to non-small cell lung cancer patients and triple-negative breast cancer patients. (DOCX 31 kb)

**Additional file 2:** Performance of prediction algorithms using signaling pathways as a constraint to get reliable feature set. (DOCX 25 kb)

## Abbreviations

A1: The supervised link prediction algorithm; A2: The extended supervised link prediction algorithm; AUC: Area under curve; B: The baseline approach; GEO: Gene expression omnibus; IC50: Half-maximal inhibitory concentration; MAUC: Mean area under curve; QBC: Query by committee; ROC: Receiver operating characteristic; RR: Ridge regression; SVR + L: Support vector regression with a linear kernel; SVR + S: Support vector regression with a sigmoid kernel

## Acknowledgements

The authors thank anonymous reviewers for their valuable comments of the manuscripts submitted to ICIBM 2016, which helped improve the paper considerably. TT thanks King Abdulaziz University for their scholarship and Saudi Arabian Cultural Mission for their academic and financial support.

## Funding

Publication charges for this article have been funded by the first corresponding author.

## Availability of data and materials

All datasets used in this study are publicly available at <http://genomed.uchicago.edu/~pgeeher/cgpPrediction/>

## About this supplement

This article has been published as part of *BMC Systems Biology* Volume 11 Supplement 5, 2017: Selected articles from the International Conference on Intelligent Biology and Medicine (ICIBM) 2016: systems biology. The full contents of the supplement are available online at <https://bmcsystbiol.biomedcentral.com/articles/supplements/volume-11-supplement-5>.

## Authors' contributions

TT and ZW conceived the study. TT and ZW designed the algorithms. TT implemented the algorithms and conducted the experiments. TT and ZW wrote and approved the final manuscript.

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published: 3 October 2017

## References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. *CA Cancer J Clin.* 2015;65(1):5–29.
2. Kamb A, Wee S, Lengauer C. Why is cancer drug discovery so difficult? *Nat Rev Drug Discov.* 2007;6(2):115–20.
3. Marx V. Cancer: A most exceptional response. *Nature.* 2015;520(7547):389–93.
4. Turner NC, Reis-Filho JS. Genetic heterogeneity and cancer drug resistance. *Lancet Oncol.* 13(4):e178–85.
5. Roden DM, George AL Jr. The genetic basis of variability in drug responses. *Nat Rev Drug Discov.* 2002;1(1):37–44.
6. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet.* 2015;16(6):321–32.
7. Sanchez-Garcia F, Villagrana P, Matsui J, Kotliar D, Castro V, Akavia U-D, Chen B-J, Saucedo-Cuevas L, Rodriguez Barrueco R, Llobet-Navas D, et al. Integration of Genomic Data Enables Selective Discovery of Breast Cancer Drivers. *Cell.* 159(6):1461–75.
8. Zhang P, Brusci V. Mathematical modeling for novel cancer drug discovery and development. *Expert Opin Drug Discov.* 2014;9(10):1133–50.
9. Covell DG. Data Mining Approaches for Genomic Biomarker Development: Applications Using Drug Screening Data from the Cancer Genome Project and the Cancer Cell Line Encyclopedia. *PLoS One.* 2015;10(7):e0127433.
10. Costello JC, Heiser LM, Georgii E, Gonen M, Menden MP, Wang NJ, Bansal M, Ammad-ud-din M, Hintsanen P, Khan SA, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol.* 2014;32(12):1202–12.
11. Geeleher P, Cox NJ, Huang RS. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol.* 2014;15(3):R47.

12. Yadav B, Gopalacharyulu P, Pemovska T, Khan SA, Szwajda A, Tang J, Wennerberg K, Aittokallio T. From drug response profiling to target addiction scoring in cancer cell models. *Dis Model Mech.* 2015;8(10):1255–64.
13. Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. *IEEE Trans Pattern Anal Mach Intell.* 2013;35(8):1798–828.
14. Mohri M, Rostamizadeh A, Talwalkar A. *Foundations of machine learning.* MIT press; 2012.
15. Coates A, Ng AY. Learning feature representations with k-means. In: *Neural Networks: Tricks of the Trade.* Springer; 2012. p. 561–80.
16. Leskovec J, Faloutsos C. Sampling from large graphs. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.* Philadelphia: ACM; 2006. p. 631–6.
17. Getoor L, Diehl CP. Link mining: a survey. *SIGKDD Explorations.* 2005;7(2):3–12.
18. Hasan MA, Zaki MJ. A Survey of Link Prediction in Social Networks. In: *Social Network Data Analytics;* 2011:243–275.
19. Lü L, Zhou T. Link prediction in complex networks: A survey. *Physica A.* 2011;390(6):1150–70.
20. Barzel B, Barabási A-L. Network link prediction by global silencing of indirect correlations. *Nat Biotechnol.* 2013;31(8):720–5.
21. Clauset A, Moore C, Newman ME. Hierarchical structure and the prediction of missing links in networks. *Nature.* 2008;453(7191):98–101.
22. Mahoney MW, Drineas P. CUR matrix decompositions for improved data analysis. *Proc Natl Acad Sci.* 2009;106(3):697–702.
23. Turki T, Wang JTL. A New Approach to Link Prediction in Gene Regulatory Networks. In: *Intelligent Data Engineering and Automated Learning – IDEAL 2015: 16th International Conference, Wroclaw, Poland, October 14–16, 2015, Proceedings.* Edited by Jackowski K, Burduk R, Walkowiak K, Woźniak M, Yin H. Cham: Springer International Publishing; 2015. p. 404–15.
24. Gillani Z, Akash MS, Rahaman MM, Chen M. CompareSVM: supervised, Support Vector Machine (SVM) inference of gene regularity networks. *BMC bioinformatics.* 2014;15(1):395.
25. Cerulo L, Elkan C, Ceccarelli M: Learning gene regulatory networks from only positive and unlabeled data. *BMC Bioinformatics.* 2010;11(1):1.
26. De Smet R, Marchal K. Advantages and limitations of current network inference methods. *Nat Rev Microbiol.* 2010;8(10):717–29.
27. Patel N, Wang JTL. Semi-supervised prediction of gene regulatory networks using machine learning algorithms. *J Biosci.* 2015;40(4):731–40.
28. Turki T, Bassett W, JTL W. A Learning Framework to Improve Unsupervised Gene Network Inference. In: *Perner P, editor. Machine Learning and Data Mining in Pattern Recognition: 12th International Conference, MLDM 2016, New York, NY, USA, July 16–21, 2016, Proceedings.* Cham: Springer International Publishing; 2016. p. 28–42.
29. Mordelet F, Vert J-P. SIRENE: supervised inference of regulatory networks. *Bioinformatics.* 2008;24(16):i76–82.
30. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. *Nat Genet.* 1999;22(3): 281–5.
31. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci.* 2000;97(22):12182–6.
32. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* 2007;5(1):e8.
33. Riddick G, Song H, Ahn S, Walling J, Borges-Rivera D, Zhang W, Fine HA. Predicting in vitro drug sensitivity using Random Forests. *Bioinformatics.* 2011;27(2):220–4.
34. Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, Saez-Rodriguez J. Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. *PLoS ONE.* 2013;8(4):e61318.
35. Jang IS, Neto EC, Guinney J, Friend SH, Margolin AA. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. In: *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing;* 2014. NIH Public Access: 63.
36. Falgreen S, Dybkær K, Young KH, Xu-Monette ZY, El-Galaly TC, Laursen MB, Bødker JS, Kjeldsen MK, Schmitz A, Nyegaard M. Predicting response to multidrug regimens in cancer patients using cell line experiments and regularised regression models. *BMC Cancer.* 2015;15(1):235.
37. Chiluka N, Andrade N, Pouwelse J. A link prediction approach to recommendations in large-scale user-generated content systems. In: *Advances in Information Retrieval.* Springer; 2011. p. 189–200.
38. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res.* 2002;321–57.
39. Turki T, Wei Z. A greedy-based oversampling approach to improve the prediction of mortality in MERS patients. In: *2016 Annual IEEE Systems Conference (SysCon);* 18–21 April 2016 2016. 1–5.
40. Turki T, Wei Z. IPRed: Instance Reduction Algorithm Based on the Percentile of the Partitions. In: *MAICS: 2015.* 181–185.
41. Settles B. Active learning literature survey. *Univ Wis Madison.* 2010;52(55–66):11.
42. Melville P, Mooney RJ. Diverse ensembles for active learning. In: *Proceedings of the twenty-first international conference on Machine learning;* 2004. ACM: 74.
43. Gilad-Bachrach R, Navot A, Tishby N. Query by committee made real. In: *Advances in neural information processing systems;* 2005. 443–450.
44. Krogh A, Vedelsby J. Neural network ensembles, cross validation, and active learning. *Adv Neural Inf Proces Syst.* 1995;7:231–8.
45. Bodor A, Csabai I, Mahoney MW, Solymosi N. rCUR: an R package for CUR matrix decomposition. *BMC Bioinformatics.* 2012;13:103.
46. Joensuu H, Kellokumpu-Lehtinen P-L, Bono P, Alanko T, Kataja V, Asola R, Utriainen T, Kokko R, Hemminki A, Tarkkanen M, et al. Adjuvant Docetaxel or Vinorelbine with or without Trastuzumab for Breast Cancer. *N Engl J Med.* 2006;354(8):809–20.
47. Aujla M. Chemotherapy: Treating older breast cancer patients. *Nat Rev Clin Oncol.* 2009;6(6):302.
48. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, et al. ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 2003;31(1):68–71.
49. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002;30(1):207–10.
50. Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez MC, Elledge R, Mohsin S, Osborne CK, Chamness GC, Allred DC, et al. Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet.* 362(9381):362–9.
51. Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez MC, Tham Y-L, Kalidas M, Elledge R, Mohsin S, Osborne CK, et al. Patterns of Resistance and Incomplete Response to Docetaxel by Gene Expression Profiling in Breast Cancer Patients. *J Clin Oncol.* 2005;23(6):1169–77.
52. Neubert K, Meister S, Moser K, Weisel F, Maseda D, Amann K, Wiethe C, Winkler TH, Kalden JR, Manz RA, et al. The proteasome inhibitor bortezomib depletes plasma cells and protects mice with lupus-like disease from nephritis. *Nat Med.* 2008;14(7):748–55.
53. Paramore A, Frantz S. Bortezomib. *Nat Rev Drug Discov.* 2003;2(8):611–2.
54. Mulligan G, Mitsiades C, Bryant B, Zhan F, Chng WJ, Roels S, Koenig E, Fergus A, Huang Y, Richardson P, et al. Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib. *Blood.* 2007;109(8):3177–88.
55. Bermolen P, Rossi D. Support vector regression for link load prediction. *Comput Netw.* 2009;53(2):191–201.
56. Wu Z, Ch L, Jky N, KRph L. Location Estimation via Support Vector Regression. *IEEE Trans Mob Comput.* 2007;6(3):311–21.
57. Balfer J, Bajorath J. Systematic Artifacts in Support Vector Regression-Based Compound Potency Prediction Revealed by Statistical and Activity Landscape Analysis. *PLoS One.* 2015;10(3):e0119301.
58. Bousquet O, Elisseeff A. Stability and generalization. *J Mach Learn Res.* 2002;2:499–526.
59. Poggio T, Rifkin R, Mukherjee S, Niyogi P. General conditions for predictivity in learning theory. *Nature.* 2004;428(6981):419–22.
60. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. *ACM Trans Int Syst Technol (TIST).* 2011;2(3):27.
61. Kanji GK. 100 statistical tests. Sage; 2006.
62. Japkowicz N, Shah M. Evaluating learning algorithms: a classification perspective. Cambridge University Press; 2011.
63. Breiman L. Random Forests. *Mach Learn.* 2001;45(1):5–32.
64. Jolliffe I. *Principal component analysis;* Wiley Online Library; 2002.
65. Hothorn T, Everitt BS. *A handbook of statistical analyses using R;* CRC press; 2014.

66. Witten D, Tibshirani R, Gross S, Narasimhan B. PMA: Penalized Multivariate Analysis (2011). URL <https://cran.r-project.org/web/packages/PMA/index.html> package version, 1(9).
67. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*. 2009;kxp008.
68. Sigg CD, Buhmann JM. Expectation-maximization for sparse and non-negative PCA. *In.*: 2008: 960–967.
69. Sigg C, Sigg MC: Package 'nsprcomp'. 2013.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

