FROM GENE EXPRESSION TO DRUG RESPONSE: A COLLABORATIVE FILTERING APPROACH

Cheng Qian[‡] *Nicholas D. Sidiropoulos*[‡] *Magda Amiridi*[‡] *Amin Emad*[*]

[‡] Department of Electrical and Computer Engineering, University of Virginia, VA, USA
[*] Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada

ABSTRACT

Predicting the response of cancer cells to drugs is an important problem in pharmacogenomics. Recent efforts in generation of large scale datasets profiling gene expression and drug sensitivity in cell lines have provided a unique opportunity to study this problem. However, one major challenge is the small number of samples (cell lines) compared to the number of features (genes) even in these large datasets. We propose a collaborative filtering like algorithm for modeling gene-drug relationship to identify patients most likely to benefit from a treatment. Due to the correlation of gene expressions in different cell lines, the gene expression matrix is approximately low-rank, which suggests that drug responses could be estimated from a reduced dimension latent space of the gene expression. Towards this end, we propose a joint low-rank matrix factorization and latent linear regression approach. Experiments with data from the Genomics of Drug Sensitivity in Cancer database are included to show that the proposed method can predict drug-gene associations better than the state-of-the-art methods.

Index Terms— Gene expression, drug response, linear regression, collaborative filtering, Genomics of Drug Sensitivity in Cancer (GDSC).

1. INTRODUCTION

Selecting the right drugs is critical for cancer survival [1], but existing methods that predict a patient's response to a particular drug are not reliable enough. Resistance to chemotherapy is a major issue, as time is of essence in many cases. Therefore, it is of great interest to construct predictive models of chemotherapy response that physicians can use to prescreen the most promising treatment options. In recent years, the field of pharmacogenomics has emerged as a very promising area with challenging problems that can benefit from more attention from the signal processing community.

Several large-scale studies have been recently conducted to measure the gene expression (i.e. transcriptomic) profile of hundreds of cell lines and their sensitivity to tens to hundreds of different drugs [2–4]. The results of these studies, which are available in databases such as the Cancer Cell Line Encyclopedia (CCLE) [2], the Genomics of Drug Sensitivity in Cancer (GDSC) [3], and the Cancer Therapeutics Response Portal (CTRP) [4], bring predictive models linking gene expression to drug response closer within reach.

Numerous drug sensitivity prediction algorithms have been proposed to characterize the relationship between transcriptomic information and drug response [5–10]. Emad *et al.* recently proposed a gene prioritization method called Prioritization of Genes Enhanced with Network Information (ProGENI) to rank genes that are closely related to a phenotype [9]. With the ranked genes, the authors employed a kernel support vector machine (SVM) for drug sensitivity prediction, and showed that ProGENI–identified genes can better predict drug response compared to genes identified by other widely used prioritization methods such as Pearson correlation and Elastic Net (EN). In [10], through a collaborative effort between the National Cancer Institute (NCI) and the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project, a comparison of 44 different drug response prediction methods was undertaken, among which the Bayesian multitask multiple Kernel learning exhibited the best prediction performance. However, the training was based on just 35 samples, which seems very limited. To handle the cases where the number of genes is greater than that of cell lines, the prevailing methods rely on some sort of sparse regression for gene selection, to help resolve the underdeterminacy that arises in even the simplest linear prediction models [11, 12].

In this paper, we take a different approach. Motivated by the observation that the gene expression matrix is approximately (very) low-rank, instead of relying on gene selection to obtain a well-posed problem, we propose a collaborative filtering (CF) approach based on joint low-rank biased matrix factorization and linear prediction from the latent space. It is worth highlighting that unlike existing methods that ignore the bias in the expression of different genes, CF takes this bias into account, which results in a more accurate model. We provide preliminary results that corroborate the effectiveness of

Email: alextoqc@gmail.com (C. Qian), nikos@virginia.edu (N. D. Sidiropoulos), ma7bx@virginia.edu (M. Amiridi) and amin.emad@mcgill.ca (A. Emad)

the proposed method using real data from GDSC.

2. PROPOSED METHOD

One major challenge, even in large databases such as GDSC [3], is the large number of features (tens of thousands of genes) compared to the number of samples (hundreds of cell lines). Therefore, the prediction of drug response from gene expression is inherently under-determined. In the literature, the common way to deal with this problem is to judiciously select a small number of transcriptomic features through sophisticated feature selection methods such as sparse regression or other gene ranking strategies that utilize prior knowledge in the form of protein-protein interactions (PPI's) and genetic interactions [9]. The existing data sets contain both gene expression data of different cell lines and their response to different drugs, where the response of each drug is only measured for a subset of the cell lines. The experimentally measured gene expression data is naturally noisy and is not necessarily 'centered'. We propose to model the intrinsically low dimensionality of the gene expression data while taking noise and bias into consideration, using a new method based on collaborative filtering.

One way to tackle biases in the gene expression measurements is to model the gene expression level g_{ij} as

$$g_{ij} = \tilde{g}_{ij} + \beta_j + n_{ij} \tag{1}$$

where $\tilde{g}_{ij}$ denotes the actual gene expression corresponding to the jth gene of the ith sample (cell line), n_{ij} is the additive noise which here is assumed to be Gaussian distributed with zero mean, and β_j is the bias of the jth gene. The matrix form of (1) is given by

$$\mathbf{G} = \tilde{\mathbf{G}} + \mathbf{1}\boldsymbol{\beta}^T + \mathbf{N} \in \mathbb{R}^{M\times L} \tag{2}$$

where $(\cdot)^T$ is the transpose, $\mathbf{1}$ is a vector of length M with all elements equal to 1, M is the number of training samples, L is the number of transcriptomic features in a cell line, $\mathbf{G}(i,j) = g_{ij}$, $\tilde{\mathbf{G}}(i,j) = \tilde{g}_{ij}$ and $\mathbf{N}(i,j) = n_{ij}$.

To continue, we bring forth our motivation by an example shown in Fig. 1, where singular values of a gene expression matrix from the GDSC data set are plotted. This matrix contains the expression of $17,737$ genes in 1018 cell lines. Fig. 1 shows that the gene expression matrix is dominated by a few principal components, indicating that gene expressions of different cell lines are strongly correlated and the gene expression matrix is approximately low-rank. Thus, we have

$$\tilde{\mathbf{G}} \approx \mathbf{A}\mathbf{B}^T \tag{3}$$

where $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_M]^T \in \mathbb{R}^{M\times F}$ and $\mathbf{B} = [\mathbf{b}_1, \cdots, \mathbf{b}_L]^T \in \mathbb{R}^{L\times F}$ are low-rank factors with $F \ll \min(M, L)$. Therefore, (2) can be approximated by

$$\mathbf{G} \approx \mathbf{A}\mathbf{B}^T + \mathbf{1}\boldsymbol{\beta}^T + \mathbf{N}. \tag{4}$$

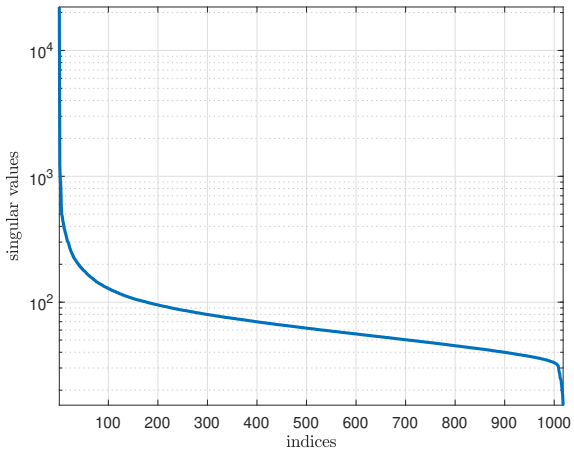

Fig. 1. Singular values of the gene expression matrix from GDSC

The above observation implies that it is not necessary to exploit all the transcriptomic features for drug response prediction. On the contrary, the dimension of $\tilde{\mathbf{G}}$ can be significantly reduced by a dimensionality reduction matrix $\mathbf{B}$. As a follow-up, we propose a novel joint dimensionality reduction and drug response prediction strategy, where the drug response is estimated from the latent space of the gene expression matrix—$\mathbf{A}$. Mathematically, we try to solve

$$\min_{\mathbf{A},\mathbf{B},\mathbf{w},\boldsymbol{\beta},e} \left\|\mathbf{G} - \mathbf{A}\mathbf{B}^T - \mathbf{1}\boldsymbol{\beta}^T\right\|_F^2 + \rho\left\|\mathbf{A}\mathbf{w} - \mathbf{r} + e\mathbf{1}\right\|_2^2 \tag{5}$$

where $\|\cdot\|_2$ is the ℓ_2-norm, $\|\cdot\|_F$ is the Frobenius norm, $\mathbf{r}$ is the drug response in the training set, $\mathbf{w}$ and e are parameters for fitting the response from the latent space $\mathbf{A}$ such that $\mathbf{r} = \mathbf{A}\mathbf{w} + e\mathbf{1}$. The first term in (5) models the dimensionality reduction and bias cancellation, the second regularization fits the drug response from the latent space of $\mathbf{G}$, and ρ controls the strength of regularization. In (5), fixing any four variables, the problem for the remaining variable is linear least squares (LS). We therefore employ an alternating least squares (ALS) strategy to solve (5). Specifically, at each iteration, the subproblem w.r.t. $\mathbf{A}$ is

$$\min_{\mathbf{A}} \|\mathbf{Y}_1 - \mathbf{A}\mathbf{X}_1\|_F^2 \tag{6}$$

where $\mathbf{Y}_1 = [\mathbf{G} - \mathbf{1}\boldsymbol{\beta}^T \quad \sqrt{\rho}(\mathbf{r} - e\mathbf{1})]$, $\mathbf{X}_1 = [\mathbf{B}^T \quad \sqrt{\rho}\mathbf{w}]$ and the solution is

$$\mathbf{A} = \mathbf{Y}_1\mathbf{X}_1^T(\mathbf{X}_1\mathbf{X}_1^T)^{-1} \tag{7}$$

with $(\cdot)^{-1}$ being the matrix inverse.

Since $\mathbf{A}\mathbf{B}^T + \mathbf{1}\boldsymbol{\beta}^T = [\mathbf{A} \ \mathbf{1}][\mathbf{B} \ \boldsymbol{\beta}]^T$, $\mathbf{B}$ and $\boldsymbol{\beta}$ can be updated simultaneously. The associated subproblem is

$$\min_{\mathbf{B},\boldsymbol{\beta}} \left\|\mathbf{G}^T - [\mathbf{B} \ \boldsymbol{\beta}]\mathbf{X}_2\right\|_F^2 \tag{8}$$

where $\mathbf{X}_2 = [\mathbf{A} \ \ \mathbf{1}]^T$ and the minimum is reached at

$$[\mathbf{B} \ \ \boldsymbol{\beta}] = \mathbf{G}^T\mathbf{X}_2^T(\mathbf{X}_2\mathbf{X}_2^T)^{-1}. \tag{9}$$

Fixing $(\mathbf{A}, \mathbf{B}, \boldsymbol{\beta})$, the update for $\mathbf{w}$ and e is straightforward, i.e.,

$$\begin{bmatrix}\mathbf{w}\\ e\end{bmatrix} = (\mathbf{X}_2\mathbf{X}_2^T)^{-1}\mathbf{X}_2\mathbf{r}. \tag{10}$$

We iteratively update $\{\mathbf{A}, [\mathbf{B}, \boldsymbol{\beta}], [\mathbf{w}, e]\}$ until the algorithm converges. Convergence is monotonic in terms of the cost function, by virtue of the conditionally optimal updates of ALS.

So far, we have shown how to estimate the unknown variables in our model. However, it is still unclear how to use $(\mathbf{A}, \mathbf{B}, \boldsymbol{\beta}, \mathbf{w})$ to predict drug response for new patients. It is worth noting that $\mathbf{A}$ and $\mathbf{w}$ are not the parameters of interest, instead, $\mathbf{B}$ and $\boldsymbol{\beta}$ are the "meat" and "bread". To explain this point, let us first showcase how to use $\mathbf{B}, \boldsymbol{\beta}$ for dimensionality reduction of a new cell line $\mathbf{g} \in \mathbb{R}^{L\times 1}$ (note that we now switch to use a column vector for the cell line). We then solve

$$\min_{\mathbf{a}} \|\mathbf{g} - \boldsymbol{\beta} - \mathbf{B}\mathbf{a}\|_2^2 \tag{11}$$

resulting in the reduced dimension gene expression vector

$$\hat{\mathbf{a}} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T(\mathbf{g} - \boldsymbol{\beta}) \in \mathbb{R}^F. \tag{12}$$

Comparing (12) and (7), we see that $\mathbf{A}$ in (7) is updated differently since it is a regularized LS that involves $\{\mathbf{r}, \mathbf{w}, e\}$ while $\hat{\mathbf{a}}$ does not. Nevertheless, for real-data applications, the new gene expression may not rigorously follow the proposed model, which means that $\{\mathbf{w}, e\}$ estimated from ALS is not properly paired with the new cell line $\hat{\mathbf{a}}$. Thus, estimating the response from $(\hat{\mathbf{a}}\mathbf{w} + e)$ is not the best option.

To handle this issue, we need to recalculate $\mathbf{w}$ and e by using a refined $\mathbf{A}$ obtained in the same manner as (12), i.e.,

$$\hat{\mathbf{A}} = (\mathbf{G} - \mathbf{1}\boldsymbol{\beta}^T)\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}. \tag{13}$$

Hence, by minimizing $\|\hat{\mathbf{A}}\mathbf{w} + e - \mathbf{r}\|$, we have

$$\begin{bmatrix}\hat{\mathbf{w}}\\ \hat{e}\end{bmatrix} = (\hat{\mathbf{X}}_2\hat{\mathbf{X}}_2^T)^{-1}\hat{\mathbf{X}}_2\mathbf{r} \tag{14}$$

where $\hat{\mathbf{X}}_2 = [\hat{\mathbf{A}} \ \ \mathbf{1}]^T$. The drug response is then estimated through

$$\hat{r} = \hat{\mathbf{a}}^T\hat{\mathbf{w}} + \hat{e}. \tag{15}$$

The overall procedure is summarized in Algorithm 1.

Remark: As we can see from (9), $\boldsymbol{\beta}$ is not simply a gene expression bias vector; it actually plays an important role in finding an accurate dimensionality reduction matrix $\mathbf{B}$ and assisting drug response estimation from the latent space.

Algorithm 1 Collaborative Filtering

function CF($\mathbf{G}, \mathbf{r}, F, \rho$)
  Randomly initialize $\mathbf{B}$ and $\mathbf{w}$
  Set $\ell = 1$
  while stopping criterion has not been reached **do**
    $\mathbf{A} \leftarrow$ (7)
    $[\mathbf{B} \ \ \boldsymbol{\beta}] \leftarrow$ (9)
    $[\mathbf{w}^T \ \ e]^T \leftarrow$ (10)
    $\ell = \ell + 1$
  end while
  Refine $\mathbf{A}$ with (13) and $[\mathbf{w}^T \ e]^T$ with (14)
  Given a new cell line $\mathbf{g}$, reduce its dimension via (12) and then compute the drug response using (15)
end function

Table 1. RMSE Comparison over 10 drugs

drug name	CF	OMP	IHT	EN
SN-38	**0.2996**	0.3012	0.3678	0.4956
TAK-715	**0.2410**	0.2460	0.2519	0.4568
Ruxolitinib	**0.1967**	0.2073	0.2144	0.3751
Ispinesib Mesylate	**0.6228**	0.6438	0.6871	1.1992
BX-912	**0.4461**	0.4732	0.5012	0.8986
Avagacestat	**0.1241**	0.1294	0.1285	0.2640
XMD14-99	**0.1761**	0.1845	0.1890	0.3412
PHA-793887	**0.5477**	0.5533	0.5752	1.0619
XMD15-27	**0.1667**	0.1713	0.1774	0.3310
Quizartinib	**0.3568**	0.3634	0.3760	0.7285

3. RESULTS

We compare the performance of of our CF-inspired approach with state-of-the-art algorithms using real data from GDSC, which is fully accessible at `https://www.cancerrxgene.org/downloads`. We use the RMA-normalised basal gene expression profiles of the cell lines released on March 2, 2017. Tthe drug response data that we use was released on March 27, 2017, containing the biochemical half maximal inhibitory concentration (IC_{50}) values of different drugs for each cell line.

For most of the cell lines in GDSC, the expression values of approximately $18,000$ genes are available, but the drug that has been measured using the largest number of samples includes approximately $1,000$ samples – which poses a challenge for training an accurate prediction model. Therefore, it is necessary to prudently select a smaller subset of informative features for training while excluding the irrelevant ones. Toward this end, we first employ the ProGENI algorithm [9] to rank the genes for each drug and select the top 500 genes to construct a smaller-size gene expression matrix. Then we choose 70% of the data samples of a tested drug for training, 10% for validation and 20% for testing.

In the first example, we compare CF with orthogonal matching pursuit (OMP), iterative hard-thresholding (IHT) and Elastic Net (EN) in terms of relative root mean square error (RMSE). We choose 10 drugs (i.e., SN-38, TAK-715, Ruxolitinib, Ispinesib Mesylate, BX-912, Avagacestat, XMD14-99, PHA-793887, XMD15-27, Quizartinib) to examine the performance of different competitors and report their RMSEs. Here, RMSE is averaged through 10 random permutations and is calculated as

$$\text{RMSE} = \frac{1}{10}\sum_{i=1}^{10} \|\mathbf{r} - \hat{\mathbf{r}}_i\|_2 / \|\mathbf{r}\|_2$$

where $\hat{\mathbf{r}}_i$ contains the drug response estimates of the testing cell lines from the ith permutation. Since the rank F and hyper-parameter ρ for CF are unknown, we vary F from 5 to 30 and ρ from 100 to 400, and choose $\{F, \rho\}$ that minimizes the Euclidean distance between the estimated and true drug response vector of the validation set. Simulation results are shown in Table 1, where CF has the smallest RMSE for all 10 drugs. Overall, OMP performs slightly better than IHT while EN has the worst performance.

In the second example, we examine the performance of CF on predicting drug sensitivity. We compare our method with OMP and three classifiers from MATLAB Statistics and Machine Learning Toolbox as baselines, i.e., logistic regression, kernel SVM (K-SVM) with Gaussian kernel function and linear SVM (L-SVM). Here, we do not include IHT and EN because their performance is worse than OMP. The drug selected for comparison is SN-38. This drug has the largest number of samples in the GDSC data set. There are 989 cell lines tested with this drug, but only 956 of them have associated transcriptomic data, which means that the total number of available samples is 965. Note that we define a threshold such that a cell line with IC_{50} value smaller than the threshold is identified as sensitive to the drug; otherwise, resistant. Thus, given a threshold, the data set can be divided into two parts corresponding to drug sensitive and resistant, respectively. The IC_{50} values for this drug range from -8.1319 to 1.4428, where the smaller the IC_{50} value, the more sensitive the cell line to this drug. Therefore, we vary the threshold from -5.5 to -2.5 to compare the prediction performance of different algorithms. Then given a threshold, we choose 70% of the data samples from each of the two parts for training, 10% for validation and 20% for testing, such that the percentage of either resistant or sensitive samples is fixed in training, validation and testing sets.

Similar to the previous example, we run 10 Monte Carlo simulations with randomly partitioned training/validation/testing sets and report the average prediction accuracy of the testing set defined as

$$\text{prediction accuracy} = \frac{\sum_{i=1}^{N} \delta(\ell_i, \hat{\ell}_i)}{N}$$

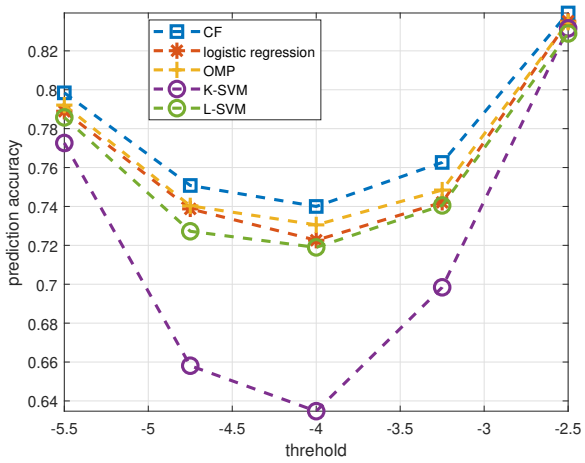

Fig. 2. Prediction accuracy comparison on drug sensitivity.

where ℓ_i is the drug response label, $\hat{\ell}_i$ is the estimated label, N is the size of the testing data and $\delta(a, b) = 1$ if $a = b$ and 0 otherwise. Also notably, for logistic regression and SVM, in consideration of the limited training samples, we employ OMP to further reduce the number of features by solving $\min_{\|\mathbf{s}\|_0 \le 20} \|\mathbf{G}\mathbf{s} - \mathbf{r}\|_2^2$, where the indices of the nonzero elements in $\mathbf{s}$ denote the selected features.

Fig. 2 shows the results, from which we see that CF has the highest prediction accuracy under different thresholds. Its performance is followed by OMP, logistic regression and L-SVM. However, the K-SVM does not work well.

4. CONCLUSION

A novel CF algorithm has been proposed for drug response prediction from gene expression. Simulations validated that CF works better than many sparse regression methods (e.g., OMP, IHT and EN) and classical linear and nonlinear classification algorithms (e.g., logistic regression and SVM).

The CF method estimates the IC_{50} values rather than the labels of drug sensitive/resistant. This can be valuable for imputing missing IC_{50}'s of cell lines for which the drug response is not measured, and for understanding the relationship between gene expression and drug response. Another important advantage of our CF-based method is that it can be used in the case of incomplete gene expression measurements – i.e., even when the matrix $\mathbf{G}$ has many missing entries. The only change in this case is that one needs to use weighted LS or stochastic gradient updates within the main ALS algorithm, as is well-known in the collaborative filtering literature.

5. REFERENCES

[1] J.C. Chang, *et al.*, "Gene expression profiling for the prediction of therapeutic response to docetaxel in pa-

tients with breast cancer," *The Lancet*, vol. 362, no. 9381, pp. 362-369, 2003.

[2] J. Barretina, *et al.*, "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603-607, 2012.

[3] W. Yang, *et al.*, "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells," *Nucleic acids research*, vol. 41, no. D1, pp. D955-D961, 2012.

[4] M.G. Rees, *et al.*, "Correlating chemical sensitivity and basal gene expression reveals mechanism of action," *Nature chemical biology*, vol. 12, no. 2, pp. 109, 2016.

[5] J. Lamb, *et al.*, "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease," *Science*, vol. 313, pp. 19291935, 2006.

[6] J. Barretina, *et al.*, "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, pp. 603607, 2012.

[7] M.J. Garnett, *et al.*, "Systematic identification of genomic markers of drug sensitivity in cancer cells," *Nature*, vol. 483, pp. 570575, 2012.

[8] M.P. Menden, F. Iorio, M. Garnett, U. McDermott, C.H. Benes, P.J. Ballester and Saez-Rodriguez, "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties," PLoS one, vol. 8, no. 4, pp. e61318, 2013.

[9] A. Emad, J. Cairns, K. R. Kalari, L. Wang and S. Sinha, "Knowledge-guided gene prioritization reveals new insights into the mechanisms of chemoresistance," *Genome biology*, vol. 18, no. 1, pp.153, 2017.

[10] J.C. Costello, *et al.*, "A community effort to assess and improve drug sensitivity prediction algorithms," *Nature biotechnology*, vol. 32, no. 12, pp. 1202-1211, 2014.

[11] X. Yu, I. Weber and R. Harrison, "Sparse representation for HIV-1 protease drug resistance prediction," *Proc. SIAM Int. Conf. Data Mining. Society for Industrial and Applied Mathematics*, pp. 342-349, 2013.

[12] A. Basu, R. Mitra, L. Han, S.L. Schreiber and P. A. Clemons, "RWEN: response-weighted elastic net For prediction of chemosensitivity of cancer cell lines," *Bioinformatics*, vol. 34, no. 19, pp. 3332-3339, 2018.

[13] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," *Proc. The Twenty-Seventh Asilomar Conf. Signals, Systems and Computers*, pp. 4044, Pacific Grove, CA, 1993.

[14] T. Blumensath, M. E. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Computat. Harmon. Anal.*, vol. 27, no. 3, pp. 265-274, 2009.

[15] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J.R. Statist. Soc. B.*, vol. 67, pp. 301320, 2005.