Information-Preserving Techniques Improve Chemosensitivity Prediction of Tumours Based on Expression Profiles

E. G. Christodoulou[1], O. D. Røe[2], A. Folarin[3], I. Tsamardinos[1,4]

[1] Bioinformatics Laboratory, ICS-FORTH, Heraklion, Crete, Greece
[2] Dept. of Cancer Research and Molecular Med., NTNU
[3] University College London
[4] Computer Science Department, University of Crete

Abstract. Prior work has shown that the sensitivity of a tumour to a specific drug can be predicted from a molecular signature of gene expressions. This is an important finding for improving drug efficacy and personalizing drug use. In this paper, we present an analysis strategy that, compared to prior work, maintains more information and leads to improved chemosensitivity prediction. Specifically we show (a) that prediction is improved when the GI50 value of a drug is estimated by all available measurements and fitting a sigmoid curve and (b) application of regression techniques often results in more accurate models compared to classification techniques. In addition, we show that (c) modern variable selection techniques, such as MMPC result in better predictive performance than simple univariate filtering. We demonstrate the strategy on 59 tumor cell lines after treatment with 118 fully characterized drugs obtained by the National Cancer Institute (NCI 60 screening) and biologically comment on the identified molecular signatures of the best predicted drugs.
Keywords: chemosensitivity prediction, variable selection, feature selection, regression, classification

1 Introduction

Prior work shows that the sensitivity of a tumour to a drug can be predicted better than chance based on the gene-expressions of the tumour [1], [2]. This finding paves the way to personalized therapy models. However, the machine learning and statistical analysis employed in prior work processes the data in a way that reduces the available information with potential detrimental effects both on the models' prediction performance as well as the identification of the molecular signatures [1], [2], [3].

First the estimation of the response to a drug in prior work is sub-optimal [1], [2]. The response of a tumour depends of course, on the dosage. The National Cancer Institute has treated a panel of 60 cancer cell lines with several thousands drugs and has created a dosage-response profile for each combination of drug and tumour. Often, this profile is summarized with a single value such

as the $\log_{10} GI50$ where GI50 is the dosage (in μM) of the drug that reduces the natural tumour growth to 50% within 48 hours. NCI, in the majority of cases, estimates $\log_{10} GI50$ by piece-wise linear interpolation which are then employed by all prior work (e.g., [1], [4], [3]). In this paper, *we show that estimating the* $\log_{10} GI50$ *values by fitting a sigmoid to the dosage-response profile preserves more information about the effects of the drug that lead to statistically significantly improved predictive performance.*

Second, prior work typically quantizes the $\log_{10} GI50$ values to create classes of tumours: [1] and [2] categorize tumours as sensitive and resistant, while [3] and [4] as sensitive, intermediate, and resistant. This type of quantization allows the application of machine learning classification techniques, variable selection methods for classification tasks, and statistical hypothesis testing techniques for discrete outcomes. Our computational experiments demonstrate that maintaining the exact $\log_{10} GI50$ values and *employing regression analysis instead of classification is often preferable as it improves chemosensitivity prediction in approximately half of the cases.*

Third, prior work often employs simple methods for identifying molecular signatures such as selecting the top k genes that are mostly differentially expressed between different classes of tumours. *We show that more sophisticated methods such as the Max Min Parents and Children (MMPC) algorithm for multi-variate feature selection [5] select more predictive signatures for the same parameter k. We biologically interpret these signatures for the 5 better predicted drugs.*

2 Data and Problem Description

Data Description: Gene-expression profiles were obtained for the NCI-60 cell-line panel [6] (these actually contain expressions only for 59 of the 60 cell lines) representing nine types of cancers: 5 Breast, 6 Central Nervous System (CNS), 7 Colon, 6 Leukemia, 10 Melanoma, 8 Lung, 7 Ovarian, 2 Prostate, 7 Renal. The expressions were measured on AffymtrixU133plus2 array containing 54,675 probesets that correspond to about 47,000 transcript variants which in turn represent more than 39,500 of the best characterized human genes. We denote with X_i the vector of expressions for cell-line i, X_i^v the expression value for probeset v on cell-line i, and with $\mathcal{X} = \{X_i\}$ the matrix of expressions. The raw data have been subjected to GCRMA normalization before analysis as implemented in the BioConductor platform [7]. The drug-response data for all 59 cell-lines were obtained from the CellMiner database [8] for a panel of 118 drugs that are fully characterized. Specifically, for each combination of drug and cell-line, the data contain several pairs of $\langle d, r \rangle$, where d is the log_{10} drug dosage and r is the percentage of tissue that survived at 48 hours after treatment. We denote with $R_{i,j}$ the set of such pairs for cell-line i and drug j.

Problem Definition: The analysis task we address is to predict the response to a drug of a tissue with expression vector X. The response of cell-line i to a drug j is often characterized with a single number that we denote with $GI50_{i,j}$ and corresponds to $\log_{10} GI50$. $GI50_{i,j}$ is typically not available in the raw data,

thus the value of $GI50_{i,j}$ is estimated from the data in $R_{i,j}$. Learning predictive models for $GI50_{i,j}$ given a vector X is a regression task. Additionally, we are interested in identifying minimal molecular signatures that are optimally predictive of response and that could provide insight into the molecular mechanisms of the drug.

3 Improving the estimation of *GI*50

The $GI50_{i,j}$ values in the publicly available NCI data are usually estimated as follows : The mean response $\bar{r}(d)$ for each dosage d is calculated and a piecewise linear function is interpolated through these mean values. The estimated GI50 value is the concentration that corresponds to $r = 50\%$ on this function, denoted as $GI50_{i,j}^{PLI}$. According to the official NCI-60 site [9], this is the methodology followed for estimating the 55% of the $GI50_{i,j}$ values. The remaining 45% were either approximated (manually, we presume) or chosen to be the highest concentration tested.

We now present an estimation method that employs all available measurements in $R_{i,j}$. We assume the dosage-response curve to have a sigmoid shape where at 0 dosage (i.e., its logarithm approaches $-\infty$) there is no reduction of the tumour (r=100%) and at infinity the tumour size is reduced to zero (r=-100%).

The equation of a sigmoid that ranges asymptotically between α and $\alpha + \beta$ and crosses the mid-range at γ is

$$r = \alpha + \frac{\beta}{1 + e^{(d-\gamma)\delta}} \tag{1}$$

where δ is a parameter controlling the slope of the function, r the response and d the dosage (expressed by its logarithm). Considering that asymptotically the drug has no effect at small dosages we set $\alpha = -100\%$; equivalently, at high dosages the tumour is eradicated completely which corresponds to -100% growth, and so we set $\beta = 200$. The remaining two parameters γ and δ were estimated using least-squares numerical optimization. Specifically, we used the function *nlinfit* of Matlab with initial values $\gamma = -5$ and $\delta = 1$. This function performs a number of steps towards the steepest descend direction for the parameters γ and δ in order to converge to a good value. In the cases where the procedure would not converge with these initial values, we repeated it 100 times with different initial values for the parameters γ and δ uniformly sampled within [-15 2] (the range of all concentrations in the data). Out of these 100 repetitions the parameter pair that led to the least mean squared error (MSE) was selected. The estimated $GI50_{i,j}$ values are found by setting r=50% and solving Eq. 1. In certain cases, fitting a sigmoid leads to extreme values. In order to detect the outliers we applied the matlab function *deletoutliers*. This implements iteratively the Grubbs Test that tests one value at a time [10]. If outliers are found they are trimmed to $\pm 2 * \sigma_j$, where σ_j is the standard deviation of all currently fitted values to drug j. We denote the final estimates as $GI50_{i,j}^{Sig}$. Figure 1(a) shows a graphical depiction of $R_{i,j}$ for the CCRF-CEM (Leukemia) cell-line and

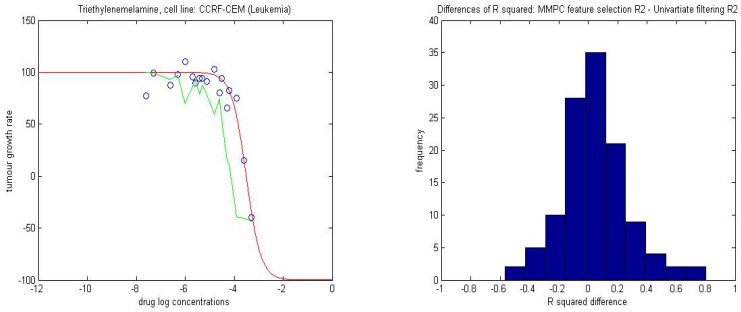

Fig. 1. (Left)The drug-response measurements for cell line CCRF-CEM (Leukemia) and Carmustine (BCNU). The $R_{i,j}$ values are shown, as well as the fitted sigmoid curve (red color) and the respective piece-wise linear interpolation segments (green color), (Right) Histogram of differences of leave-one-out cross validated R^2 when MMPC is used as the feature selection procedure minus the leave-one-out cross validated R^2 when univariate filtering is used as the feature selection procedure.

Carmustine (BCNU) with the fitted sigmoid superimposed. The corresponding piece-wise linear interpolation segments are also shown in the figure. We now show that this method of estimation leads to improvements in chemosensitivity prediction. The analysis includes the following steps:

Feature Selection: The most commonly applied feature (variable) selection method in the field of personalized medicine is to rank the genes according to their association with the class (equivalently the p-value) and select the top k. We call this method univariate filtering. In our work we additionally employed the Max Min Parents and Children algorithms (**MMPC**) [5] to select a minimal-size, optimally predictive set of probe-sets. MMPC is an algorithm that seeks to identify the neighbors of the variable-to-predict in the Bayesian Network capturing the data distribution by taking into account multivariate associations. It has been shown very effective in recent extensive experiments [11] against an array of state-of-the art feature selection methods. In this work, the causal explorer implementation of MMPC was used [12] with the default values for the parameters.

Regression: We employed SVM Regression to construct the predictive models [13]. In our implementation we used the Radial Basis kernel and all other parameters set to default(we have experimented with other values but they did not lead to better results).

Estimation of Performance: We used a leave-one-out cross validation protocol due to the small number of samples that we had (59). For each training set, the combination of MMPC and SVM regression produced a predictive model that was applied on the hold-out test sample. **Metric of performance:** The metric to measure prediction performance is the leave-one-out cross-validated R^2 (coefficient of determination), which is a conservative metric [14]. Specifically, for a given drug j, let $\mu_{\setminus i}$ be the mean value of GI50 in the data excluding cell line i (training data), $\widehat{GI50}_{\setminus i}$ the predicted GI50 by the model constructed excluding cell-line i, and $GI50_i$ the GI50 as estimated by the experiments in the

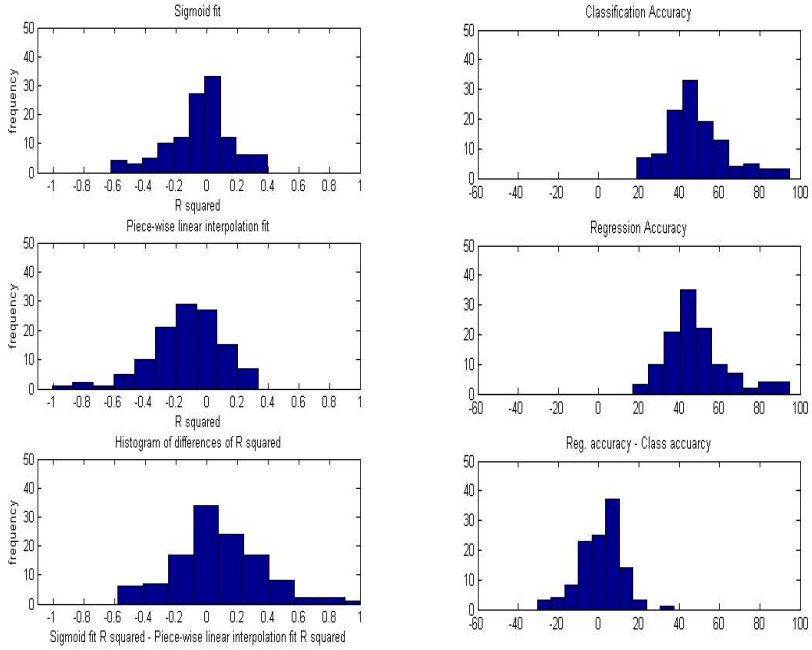

Fig. 2. (Left) Histogram R^2 for $GI50^{Sig}$ (GI50's fitted by a sigmoid), $GI50^{PLI}$ (standard estimation), and their difference, from top to bottom respectively. (Right) Histograms of cross validated classification accuracies A_j, discretized regression accuracies D_j and the $A_j - D_j$ differences, from top to bottom, respectively

corresponding cell line i for drug j. We define:

$$R_j^2 \equiv 1 - \frac{\sum_i (GI50_i - \widehat{GI50}_{\setminus i})^2}{\sum_i (GI50_i - \mu_{\setminus i})^2} \tag{2}$$

The interpretation of R^2 is that it corresponds to variance explained (uncertainty) by the models, or the reduction of variance by the use of the gene-expression models.

We have computed R_j^2 for all 118 drugs both when the GI50 values are estimated using piece-wise linear interpolation as well as when fitting a sigmoid function, as described above. We denote the corresponding values as R_j^{PLI} and R_j^{Sig}. The results are shown in Figure 2(a).

The figure shows that GI50 values estimated by the sigmoid are better predicted using the protocol described above. Thus at least for the combination of MMPC and SVM Regression $GI50^{Sig}$ values facilitate the induction of predictive models vs. using the $GI50^{PLI}$. The 95% confidence interval for the median $R_j^{Sig} - R_j^{PLI}$ as estimated by a Wilcoxon signed-rank test is $[0.025, 0.122]$. Of course, one could argue that the results may not transfer to other feature selection or regression methods. The results however, corroborate our intuition

that the sigmoid estimation better preserves information in the $R_{i,j}$ measurements and given no evidence to the contrary, we would suggest this method of estimation in future analyses.

A second observation on the results is that several drugs are well predicted by the models. More specifically, for the Pearson correlation r between two quantities, Cohen gives the following interpretation guidelines: small effect size, r = 0.1 - 0.23; medium, r = 0.24 - 0.36; large, r = 0.37 or larger. Interpreting R^2 as r^2 and translating the values we get approximately the intervals $[0.01, 0.05)$, $[0.05, 0.13)$, $[0.13, 1]$. Under this interpretation for 21 drugs out of the 118, the tumour expression profiles have a large effect in predicting response; for 12 drugs they have a medium effect, and for 19 drugs they have a small effect. On the other hand, 65 out of 118 drugs have a negative size effect, meaning that our prediction does not improve much compared to the prediction by the mean value.

We finally compared the prediction performance of our models using MMPC and univariate filtering as feature selection methodology. We computed the cross-validated R^2 for both methods on all drugs using the same protocol as before. The k parameter is set to the number of genes returned by MMPC, so that both methods return the signatures of the same sizes. Figure 1(b) presents a histogram of the results. The figure shows that, on average, MMPC returns more informative signatures and thus, it should be preferred.

4 Comparison of Regression versus Classification

In all related prior work, to the best of our knowledge, classification models have been constructed for predicting GI50 values [1, 3, 4]. Given that the latter values are continuous, the authors have quantized them before applying any classifiers, as described in the previous sections. We now show that quantization is sometimes detrimental to performance and regression techniques have greater predictive power.

In this next set of computational experiments we pre-process the GI50 values of each drug to discretize them as described in [4]. Specifically, the class $C_{i,j}$ of a cell-line i and drug j is computed as sensitive, intermediate, or resistant if $GI50_{i,j}^{Sig}$ falls within $(-\infty, \mu_j - 0.5\sigma_j]$, $(\mu_j - 0.5\sigma_j, \mu_j + 0.5\sigma_j]$, and $[\mu_j - 0.5\sigma_j, \infty)$ respectively, where μ_j is the average GI50 value over all cell lines for drug j and σ_j the standard deviation.

To evaluate classification, we employed the same overall protocol described in Section 3 with the following modifications: we used multi-class SVM classification instead of SVM Regression [15]. SVMs have been very popular and successful classifiers, particularly in bioinformatics [16]. We used the *libsvm* implementation of SVMs with the Radial Basis kernel and all other parameters set to default. In addition, the metric of performance for classification is accuracy, i.e., the percentage of samples whose class is correctly predicted. We denote with A_j the leave-one-out cross-validated accuracy of the method on drug j.

Comparing regression vs. classification is not straightforward given that regression outputs a continuous prediction for $GI50_{i,j}^{Sig}$ while classification outputs

its class. To overcome this issue we discretize the output of the regression models to the three stated classes using the same intervals as above. This allows us to compute the cross-validated accuracy of the regression for each drug j, denoted as D_j. In other words, A_j are computed by first discretizing the data, then using classification, and measuring the accuracy of the output, while D_j is computed by using regression, then discretizing the predictions, and computing accuracy.

Figure 2(b) shows the histograms of A_j, D_j, and their difference $D_j - A_j$. In some cases regression accuracy scores higher than classification accuracy and in other cases the reverse happens. For example, for drug Vincristine-sulfate (NSC:67574) regression accuracy D is 55,9322 while classification accuracy A is 18,6441. On the other hand, for drug Guanazole (NSC:1895) A is 57,6271 while D is 27,1186.

5 Biological Interpretation of the Molecular Signatures

The biological correlates of these findings are quite interesting. Below we will discuss the target systems which may connect these genes with tumour resistance and tumour response. Due to space limitation we consider only the top five drugs with the highest R^2, i.e., the drugs for which we get the larger improvement in prediction when gene expressions are employed. These drugs and their signatures are shown in Table 1.

Compound	$\mathbf{R^2}$	**Selected Probesets and Their Gene Symbols**
Carmustine (BCNU)	0.3960	1558517_*s_at* (not mapped), 202574_*s_at* (CSNK1G2), 202957_*at* (HCLS1), 203987_*at* (FZD6), 207868_*at* (CHRNA2), 220176_*at* (NUBPL), 224391_*s_at* (SIAE), 227346_*at* (IKZF1), 227485_*at* (DDX26B)
Nitrogen mustard hydrochloride	0.3617	205739_*x_at* (ZNF107), 209694_*at* (PTS), 221679_*s_at* (ABHD6), 232543_*x_at* (ARHGAP9), 232909_*s_at* (BPTF), 235320_*at* (not mapped), 240458_*at* (not mapped), 242314_*at* (TNRC6C)
Chlorambucil	0.3564	209694_*at* (PTS), 225853_*at* (GNPNAT1), 232543_*x_at* (ARHGAP9), 240458_*at* (not mapped), 241935_*at* (SHROOM1), 243678_*at* (not mapped), 57516_*at* (ZNF764)
Amsacrine	0.3503	202538_*s_at* (CHMP2B), 203600_*s_at* (FAM193A), 204336_*s_at* (RGS19), 227539_*at* (not mapped), 229280_*s_at* (FLJ22536), 229389_*at* (ATG16L2) 230424_*at* (C5orf13), 241935_*at* (SHROOM1), 244551_*at* (not mapped)
Dichloroallyl-lawson	0.3071	200595_*s_at* (EIF3A), 203147_*s_at* (TRIM14), 208260_*at* (AVPR1B), 217917_*s_at* (DYNLRB1), 218130_*at* (not mapped), 222549_*at* (CLDN1), 223915_*at* (BCOR), 228124_*at* (ABHD12), 237726_*at* (not mapped), 243007_*at* (not mapped)

Table 1. The 5 best predicted drugs (highest R^2) along with their gene signatures.

Carmustine (BCNU) is a mustard compound and an alkylating and cross-linking agent, used in oncology since 1972, and still in use for tumours of the central nervous system, Hodgkin, non-Hodgkins disease and myelomatosis [17]. FDZ6 (Frizzled homolog 6) belongs to the family of catenin molecules and involved in various processes in developing and adult organisms, but is also implicated in carcinogenesis through the WNT/FZD signaling pathway. IKZF1 or Ikaros zinc finger 1 alteration status accurately predicts relapses [18] and a nine-gene signature including deletion of IKZF1 is predictive of chemotherapy response in pediatric acute lymphoblastic leukemia [19].

Chlorambucil and Nitrogen mustard hydrochloride compounds are derived from mustard gas and are used in chemical warfare since the first world war. They are both alkylating agents, that interfere with DNA replication and transcription of RNA, and ultimately result in the disruption of nucleic acid function. These compounds are very toxic as well as mutagenic, and thus a more tailored use by genomic signatures would be very important. The indications for Chlorambucil are chronic lymphocytic leukemia, Hodgkin's disease, and other lymphomas. Nitrogen mustard hydrochloride or chlorethamine is active on small cell, non-small cell and prostate cancer cell lines (NCI cell line database), and is indicated for the palliative treatment of Hodgkin's disease (Stages III and IV), other hematological malignancies as well as bronchogenic carcinoma. These two drugs share three genes of interest in their signatures: ARGHAP9 and zinc finger proteins (ZNF107 and ZNF764 respectively) and PTS. ARGHAP9 contains the RhoGAP domain and is a GTPase activator protein. It is responsible for the transduction of signals from plasma-membrane receptors and for the control of cell adhesion, motility and shape by actin cytoskeleton formation [20]. Zinc finger proteins are involved in DNA recognition, RNA packaging, transcriptional activation, regulation of apoptosis, protein folding and assembly, and lipid binding [21]. PTS, encoding 6-pyruvoyltetrahydropterin synthase is crucial in folate metabolism and neutransmitter synthesis [22]. This has not been associated to cancer previously.

Amsacrine is an aminoacridine derivative and a potent intercalating antineoplastic agent. It is effective in the treatment of acute leukemias and malignant lymphomas, but has poor activity in the treatment of solid tumors. It is frequently used in combination with other antineoplastic agents in chemotherapy protocols. The gene CHMP2B is involved in protein sorting and transport from the endosome to the vacuole/lysosome in eukaryotic cells. It affects the MVB sorting pathway, which plays a critical role in the decision between recycling and degradation of membrane proteins [23]. Individuals with mutations of this gene develop neurodegenerative disease, probably due to dysfunctional autophagy [24]. RGS19 belongs to the RGS (Regulator of G Protein Signalling) multi-functional, GTPase-accelerating proteins. Recently it was discovered that RGS19 overexpression can enhance cell proliferation [25]. Single nucleotide polymorphism in FLJ22536 was associated to clinically aggressive neuroblastoma, a childhood malignancy [26]. Interestingly, hypermethylation of the autophagy related gene ATG16L2 was associated with poorer prognosis in terms of molecular response to Imatinib treatment in acute lymphoblastic leukemia, which is also the target disease of amsacrine [27]. C5orf13 or P311 is involved in invasion and migration of glioblastoma cells [28].

Dichloroallyl lawson is not an anti-cancer drug yet, and is largely inactive in most cell lines screened. Interestingly, the eIF3a mRNA is found elevated in a number of cancer type and it shown that suppression of eIF3a expression can reverse the malignant phenotype and change the sensitivity of cells to cell cycle modulators [29].

In summary, four of the five signatures predict activity in drugs with a similar mechanism of action, that is DNA damage through alkylation, and mainly active in hematological diseases as lymphomas and central nervous system tumours. Something also worth noticing is that three of our top drugs are nitrogen mustards. The signatures include several genes that are not directly related to the classical DNA damage and repair systems and may infer novel associations and actions of these genes.

6 Conclusion

Predicting chemosensitivity of tumours from gene expressions is important for selecting treatment, understanding the molecular mechanisms of drug response, and selecting molecular signatures. In this paper, we show that predictive performance can be improved by employing a new method for estimating the GI50 (indication of response to drug), regression algorithms instead of classification, and state-of-the-art, multivariate feature selection. The signatures identified here have several links to cancer progression and resistance to chemotherapy, but the direct relations between the genes and the respective drugs are missing. Knowledge on these relations are still expanding and the methods used to identify those signatures may be important tool for novel biological hypotheses.

Acknowledgements

We would like to thank Matina Fragoyanni for parts of the code, Sofia Triantafyllou, Vincenzo Lagani and Angelos Armen for suggestions and feedback. Finally, we thank FORTH for funding and support of the work.

References

1. A. Potti, H. K. Dressman, A. Bild, and R. F. Riedel et al. Genomic signatures to guide the use of chemotherapeutics, 2006.
2. C. K. Augustin, J. S. Yoo, A. Potti, and Y. Yoshimoto et al. Genomic and molecular profiling predicts response to temozolomide in melanoma. *Clinical Cancer Res*, 15(2), 2009.
3. J. E. Staunton, D. K. Slonim, and H. A. Coller et al. Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci.*, 98(19):10787–10792, 2001.
4. Y. Ma, Z. Ding, and Y. Qian et al. An integrative genomic and proteomic approach to chemosensitivity prediction. *Int. J. Oncol*, 34(1):107–115, 2009.
5. I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Journal of Machine Learning*, 65:31–78, 2006.
6. `http://dtp.nci.nih.gov/index.html`.
7. `http://www.bioconductor.org`.
8. U. T. Shankavaram, S. Varma, and D. Kane et al. Cellminer: a relational database and query tool for the nci-60 cancer cell lines. *BMC Genomics*, 10(277), 2009.
9. `http://dtp.nci.nih.gov/docs/compare/compare$\_$methodology.html`.

10. http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm.
11. C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *Journal of Machine Learning Research, Special Topic on Causality*, 11:171–234, 2010.
12. A. Statnikov, T. Tsamardinos, L. E. Brown, and C. F. Aliferis. Causal explorer: A matlab library of algorithms for causal discovery and variable selection for classification. *Challenges in Causality*, 1, 2009.
13. C. C. Chang and C. J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
14. R. G. D. Steel and J. H. Torrie. *Principles and Procedures of Statistics*. New York: McGraw-Hill, 1960.
15. B Boser, I. Guyon, and V. Vapnik. An training algorithm for optimal margin classifiers. In *In Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, 1992.
16. A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–643, 2005.
17. S. Egyhazi, J. Bergh, J. Hansson, P. Karran, and U. Ringborg. Carmustine-induced toxicity, dna crosslinking and o6-methylguanine-dna methyltransferase activity in two human lung cancer cell lines. *Eur J Cancer*, 27:1658–1662, 1991.
18. E. Waanders, V. H. van der Velden, C. E. van der Schoot, and F. N. van Leeuwen et al. Integrated use of minimal residual disease classification and ikzf1 alteration status accurately predicts 79% of relapses in pediatric acute lymphoblastic leukemia. *Leukemia*, 25:254–258, 2011.
19. Z. Zuo, D. Jones, and H. Yao et al. A pathway-based gene signature correlates with therapeutic response in adult patients with philadelphia chromosome-positive acute lymphoblastic leukemia. *Mod Pathol*, 23:1524–1534, 2010.
20. http://pfam.sanger.ac.uk/.
21. J. H. Laity, B. M. Lee, and P. E. Wright. Zinc finger proteins: new insights into structural and functional diversity. *Curr Opin Struct Biol*, 11(1):39–46, 2001.
22. D. Meili, J. Kralovicova, J. Zagalak, and L. Bonafe et al. Disease-causing mutations improving the branch site and polypyrimidine tract: pseudoexon activation of line-2 and antisense alu lacking the poly(t)-tail. *Hum Mutat*, 30:823–831, 2009.
23. M. Babst, D. J. Katzmann, E. J. Estepa-Sabal, T. Meerloo, and S. D. Emr. Escrt-iii: an endosome-associated heterooligomeric protein complex required for mvb sorting. *Dev Cell*, 3:271–282, 2002.
24. T. E. Rusten and A. Simonsen. Escrt functions in autophagy and associated disease. *Cell Cycle*, 7:1166–1172, 2008.
25. P. H. Tso, Y. Wang, S. Y. Wong, Poon L. S., and A. S. Chan et al. Rgs19 enhances cell proliferation through its c-terminal pdz motif. *Cell Signal*, 22:1700–1707, 2010.
26. J. M. Maris, Y. P. Mosse, and J. P. Bradfield et al. Chromosome 6p22 locus associated with clinically aggressive neuroblastoma. *N Engl J Med*, 358, 2008.
27. T. Dunwell, L. Hesson, T. A. Rauch, and L. Wang et al. A genome-wide screen identifies frequently methylated genes in haematological and epithelial cancers. *Mol Cancer*, 9(44), 2010.
28. L. Mariani, W. S. McDonough, D. B. Hoelzinger, C. Beaudry, and E. Kaczmarek et al. Identification and validation of p311 as a glioblastoma invasion gene using laser capture microdissection. *Cancer Res*, 61:4190–4196, 2001.
29. F. Saletta, Y. S. Rahmanto, and D. R. Richardson. The translational regulator eif3a: the tricky eif3 subunit! *Biochim Biophys Acta*, 1806(2):275–286, 2010.