Phenotype Classification Using Moment Features of Single-Cell Data

Cancer Informatics
Volume 17: 1–7
© The Author(s) 2018
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1176935118771701
SAGE

Chao Sima[1], Jianping Hua[1], Michael L Bittner[2], Seungchan Kim[3] and Edward R Dougherty[4]

[1]Center for Bioinformatics and Genomic Systems Engineering, Texas A&M Engineering Experiment Station, College Station, TX, USA. [2]Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ, USA. [3]Center for Computational Systems Biology, Department of Electrical and Computer Engineering, Prairie View A&M University, Prairie View, TX, USA. [4]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA.

ABSTRACT: Features for standard expression microarray and RNA-Seq classification are expression averages over collections of cells. Single cell provides expression measurements for individual cells in a collection of cells from a particular tissue sample. Hence, it can yield feature vectors consisting of higher order and mixed moments. This article demonstrates the advantage of using these expression moments in cancer-related classification. We use synthetic data generated from 2 real networks, the mammalian cell cycle network and a melanoma-related pathway network, and real single-cell data generated via fluorescent protein reporters from 2 cell lines, HT-29 and HCT-116. The networks consist of hidden binary regulatory networks with Gaussian observations. The steady-state distributions of both the original and mutated networks are found, and data are drawn from these for moment-based classification using the mean, variance, skewness, and mixed moments. For the real data, we only observe 1 gene at a time, so that only the mean, variance, and skewness are considered, the analysis being done for 2 genes, *EGFR* and *ERRB2*. For the synthetic data, classification improves as we move from just the mean to mean, variance, and skewness and then to these plus the mixed moments. Comparisons are done with 3, 4, or 5 features, using feature selection. Sample size effects are considered. For the real data, we only consider mean, variance, and skewness, with results improving when the higher order moments are used as features.

KEYWORDS: Classification, gene regulatory network, moment features, single-cell data

RECEIVED: December 1, 2017. **ACCEPTED:** March 21, 2018.

TYPE: Special Collection: Signal Processing Applications in Genomics

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work has been supported in part by the TEES-Agrilife Center for Bioinformatics and Genomic Systems Engineering (CBGSE).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Edward R Dougherty, Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77845, USA. Email: edward@ece.tamu.edu

Introduction

Transcriptome analysis is a powerful strategy to connect genotype to phenotype of cells. Essentially, all cells share the same genetic code inherited from their ancestor, but transcriptomes of individual cells characterize a subset of genes expressed to reflect their epigenetic status or their genetic regulatory system leading to specific phenotypes.[1,2] Hence, ideally, the transcriptome should be profiled for each individual cell; however, owing to technical limitations, until recently, most transcriptomic profiling has been done on bulk cells, yielding only average behavior of tens of thousands of cells. Recent advances in next-generation sequencing technologies have allowed in-depth investigation of the transcriptome at a single-cell resolution,[3] thereby opening avenues for innovative target discovery.[4] A recent theoretical analysis has compared phenotype classification based on single-cell expression trajectories with mean expression levels across multiple cells, as is the case with both ordinary RNA-Seq and expression microarrays.[5] Bulk expression measurement (multiple-cell averaging) destroys both intercell and dynamical information, and therefore using single-cell trajectories should be expected to achieve lower misclassification rates for phenotype classification. In the work by Karbalayghareh et al.,[5] cell trajectory versus average cell classification is studied in the context of Boolean networks with perturbation (BNp) and, except in some cases where the network attractors have special form,[6–8] single-cell trajectory data outperform average cell measurements. In practice, regulatory asynchronicity would lead to missing values.[9] Moreover, lower amounts of messenger RNA in individual cells can cause experimental issues that would also lead to dropouts.[10] Thus, the modeling in the work by Karbalayghareh et al.[5] assumes random missing values in trajectory readouts, with missing value rates as high as 50%.

Looking into the future, we should expect that nondynamical (nontrajectory) single-cell data in sufficient supply for cancer classification will become available before sufficient dynamical single-cell data. In this case, because a gene's expression is not static, expression measurements for a gene will possess a distribution over the cell collection and not simply from measurement error. Moreover, as genes interact, they will possess a joint expression distribution. Reducing this multivariate distribution to a set of averages entails a significant compression of information. If single-cell measurements were available, then each cell would yield an expression vector, the collection of cells would yield a sample of expression vectors, sample moments (not just averages) can be computed, and moment-based classification could proceed using higher order and mixed moments. The latter can be particularly useful because they reveal interaction and the alteration of signaling pathways can significantly alter gene interaction, thereby enabling phenotype discrimination.

Methods

For moment-based classification, suppose that for each tissue, n cells are collected and for each cell an expression vector is formed from g genes. This yields a sample of n expression vectors $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$, where $\mathbf{Y}_i = [Y_{i1}, Y_{i2}, \ldots, Y_{ig}]$. From these, we can compute empirical moments. A feature vector is composed of some set of moments. We focus on the first 3 moments, μ_i^1, μ_i^2, μ_i^3, for each gene expression $e_i = [Y_{1i}, Y_{2i}, \ldots, Y_{ni}]$, and the second-order mixed moments, μ_{ij}, $i, j = 1, 2, \ldots, g$, with $i \neq j$, for (e_i, e_j). This gives a total of $d(g) = 3g + (1/2)g(g-1)$ moments for each tissue and these form a feature vector $\mathbf{X} = (X_1, X_2, \ldots, X_{d(g)})$, where $X_1 = \mu_1^1, X_2 = \mu_1^2, X_3 = \mu_1^3, X_4 = \mu_2^1, \ldots, X_{3g} = \mu_g^3, X_{3g+1} = \mu_{12}, X_{3g+2} = \mu_{13}, \ldots, X_{4g-1} = \mu_{1k}, X_{4g} = \mu_{21}, \ldots, X_{d(g)} = \mu_{g-1,g}$. We will not consider additional moments due to the sample size being small, which is typical in genomics applications.

If there are N tissue samples, N_0 from phenotype 0 and N_1 from phenotype 1, then we have the training sample $S = S_0 \cup S_1$, with N_0 moment feature vectors from class 0 and N_1 from class 1. Because it is typically the case in biomedicine that sampling is separate and not random, meaning that tissues are chosen randomly from each class but not from the population as a whole, so that prior probabilities $c_0 = P(L = 0)$ and $c_1 = P(L = 1)$, where L is the class label, cannot be estimated from the data[11]; we assume that they are known.

Synthetic data via a gene regulatory network

If we assume a network model, then we can solve for the Bayes classifier and generate synthetic data to study classifier design and feature selection.[12] We shall assume Gaussian networks generated from hidden discrete networks, which we will take to be BNp[6] but which could also be probabilistic Boolean networks[8] or Bayesian networks.[13] Knowing the generating BNp allows us to study the effects of regulatory alteration, for instance, classifying between a nominal network and another resulting from mutation or drugs. Using Gaussian measurements allows us to model basal-level expressions and variability. We describe the network model for a single BNp and later return to classification with 2 BNps.

Consider a BNp with g genes. States are of the form $\mathbf{V} = (V_1, V_2, \ldots, V_g)$, where $V_i \in \{0,1\}$, and there are 2^g states. The $2^g \times 2^g$ transition probability matrix (TPM) can be analytically derived and the steady-state distribution π can be derived from the TPM. Let $\mathcal{M}$ be the $d(g)$ moment vector associated with π. We assume a Gaussian observation. The observation Y_i of the ith gene is normally distributed:

$$Y_i \sim \mathcal{N}(\lambda_i + \delta V_i, \sigma^2)$$

where λ_i is the mean expression when the ith gene is considered off, $\lambda_i + \delta$ is the mean expression when the ith gene is on, and σ^2 is the expression variance. There is no theoretical difficulty in making δ and σ depend on i, but to keep the results more transparent and the simulations less burdensome, we will not. $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_g)$ is the observation corresponding to the hidden state $\mathbf{V} = (V_1, V_2, \ldots, V_g)$. For a single subject, there are n cells observed. This yields observations $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$, where $\mathbf{Y}_j = (Y_{j1}, Y_{j2}, \ldots, Y_{jg})$, $Y_{ji} \sim N(\lambda_i + \delta V_{ji}, \sigma^2)$, $\mathbf{V}_j = (V_{j1}, V_{j2}, \ldots, V_{jg})$, and V_1, V_2, …, V_n are randomly drawn from π. A moment feature vector $\mathbf{X}$, where $\mathbf{X} = (X_1, X_2, \ldots, X_{d(g)})$ is defined earlier in the section, is calculated from $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$.

When the BNp is perturbed, a different TPM will follow which results in a different steady-state distribution. We refer to the steady-state distribution for the unperturbed and perturbed BNps as π_0 and π_1, respectively. We are interested in studying whether including different moments will improve the classification. Therefore, we categorize 3 types of moment features: $\mathcal{M}_1$ for first moment features only, $\mathcal{M}_2$ supersetting $\mathcal{M}_1$ but also including the second and third moment features, and $\mathcal{M}_3$ supersetting $\mathcal{M}_2$ but also including mixed moments.

If there are N subjects, each with n cells in the sample, then this procedure yields a training sample $S = S_0 \cup S_1$, where $S_0 = \{(\mathbf{X}_{01}, L_{01}), (\mathbf{X}_{02}, L_{02}), \ldots, (\mathbf{X}_{0N_0}, L_{0N_0})\}$ and $S_1 = \{(\mathbf{X}_{11}, L_{11}), (\mathbf{X}_{12}, L_{12}), \ldots, (\mathbf{X}_{1N_1}, L_{1N_1})\}$, with N_0 and N_1 feature vectors based on the steady-state distributions π_0 and π_1, respectively, with L_{uv} being the observed labels, and $N = N_0 + N_1$. Randomness in $\mathbf{X}$ results from randomness in $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_n$ and randomness in the observations Y_{ji}.

In this study, we generate synthetic data using 2 pathway networks: Pathway Network 1 (PN1) is a mammalian cell cycle network and Pathway Network 2 (PN2) is melanoma-related pathway network (Figure 1). Both of these have been previously proposed and described.[14] Briefly, PN1 includes a few key genes in the mammalian cell cycle whose signals and controls play a critical role in cell growth, among which *P27* is active in the absence of the cyclins and blocks the action of *CycE* or *CycA*. When *P27* is mutated and always off, it introduces a mutated phenotype where the growth factors are inactive. On the other hand, PN2 focuses on gene *Wnt5a*, which has been found to be highly discriminating between cells with properties typically associated with high versus low metastatic competence, as validated in melanoma cells. A different type of perturbation is applied to PN2, where we added the regulatory predictor *Ret1* for *S1000p* (dashed arrow from *Ret1* in Figure 1) as a function wiring modification. A summary of these networks is shown in Table 1.

Real data from fluorescent protein reporters

Ideally, single-cell RNA-Seq technology could be used to profile the whole transcriptome of hundreds of cells from

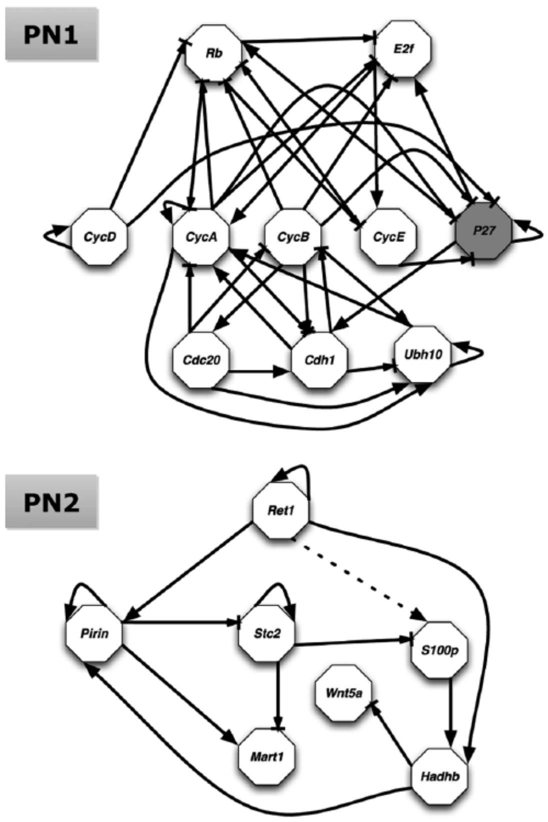

Figure 1. Logical regulatory network graphs for a mammalian cell cycle network (PN1) and a melanoma-related pathway network (PN2), modified from Figures 3 and 1 in Qian and Dougherty,[14] respectively. An arrow represents activation regulation, whereas an arrow ending with a bar represents inhibition. A different steady-state distribution resulted from *P27* stuck-at-0 change (shaded node in PN1) or regulatory change (dashed arrow in PN2). PN1 indicates Pathway Network 1; PN2, Pathway Network 2.

individual patients/cell lines. However, because this technology is still under intensive development and not widely adopted with a common protocol, current publicly available single-cell RNA-Seq data sets are generated to demonstrate the ability of a certain methodology, usually the amount and quality of cells profiled per patient/cell line. As a result, these data sets contain extremely small numbers of patient/cell line sample points per phenotype, usually 1 or 2, thus thwarting any effort for realistic classification based on such data. Thus, in this study, we use an in-house data set collected through a high-content imager that tracks a given gene's transcription level in individual cancer cells via fluorescent protein reporters.[15,16]

Prior to the introduction of single-cell RNA-Seq, fluorescent protein reporters, along with single-molecule fluorescent in situ hybridization and single-cell quantitative polymerase chain reaction, have been the most common approaches to examine transcriptional heterogeneity among cells.[17–20] In the fluorescent technology, the protein reporter is assembled by fusing the coding sequence of a fluorescent protein reporter with the promoter region of the target gene and then transfecting it into target cells. The abundance of the fluorescent protein indicates the transcriptional level and it can be captured by an epifluorescent microscope. In our experimental setup, each cell is transfected with just one reporter to follow the transcription of one specific target gene. The fluorescent images are commonly taken as 2-color image pairs with a blue channel for the nuclei and a green channel for the fluorescent reporters. Then, the expression levels of individual cells are extracted using an in-house software that first identifies the individual nuclei in the nuclei channel and then extracts the corresponding fluorescent protein signal in the other channel.

Table 1. A summary of the pathway networks in this study.

	PATHWAY NETWORK (PN1)	PATHWAY NETWORK (PN2)
Description	Mammalian cell cycle	Melanoma-related pathway
No. of genes	10	7
Perturbation	P27 mutated and stuck at 0	Adding regulatory predictor

Table 2. Number of wells/samples measured for every gene and cell line.

	HT-29	HCT-116
Egfr	43	24
Erbb2	24	24

Median number of cells per well: 247.

Compared with single-cell technology, the fluorescent protein reporter technology detects only 1 gene, rather than the whole transcriptome. An advantage of a high-content imager is that one can capture the transcriptional activities in many wells on the plate simultaneously, where each well is an independent sample point from the corresponding cell line. With this size, we can test the potential of moment classification, even with just a single gene. The drawback is that there are no second-order mixed moments. Nonetheless, we can test moment classification using the first 3 moments and demonstrate its advantage over classification using only the mean.

Our simulation study involves 2 cell lines, HT-29 and HCT-116, that are resistant and sensitive to the drug lapatinib, respectively. Lapatinib is a cancer drug that has been approved for treating breast cancer by inhibiting *Egfr* and *Erbb2*, 2 membrane-bound protein receptors commonly associated with cancer. Thus, we have selected *Egfr* and *Erbb2* as the 2 genes to be profiled. Because each cell has only 1 fluorescent protein, *Egfr* and *Erbb2* expressions are profiled separately. The number of wells (sample points) available for each (cell line, gene) combination from our experiment is shown in Table 2. The images are taken 2 hours before lapatinib is added.

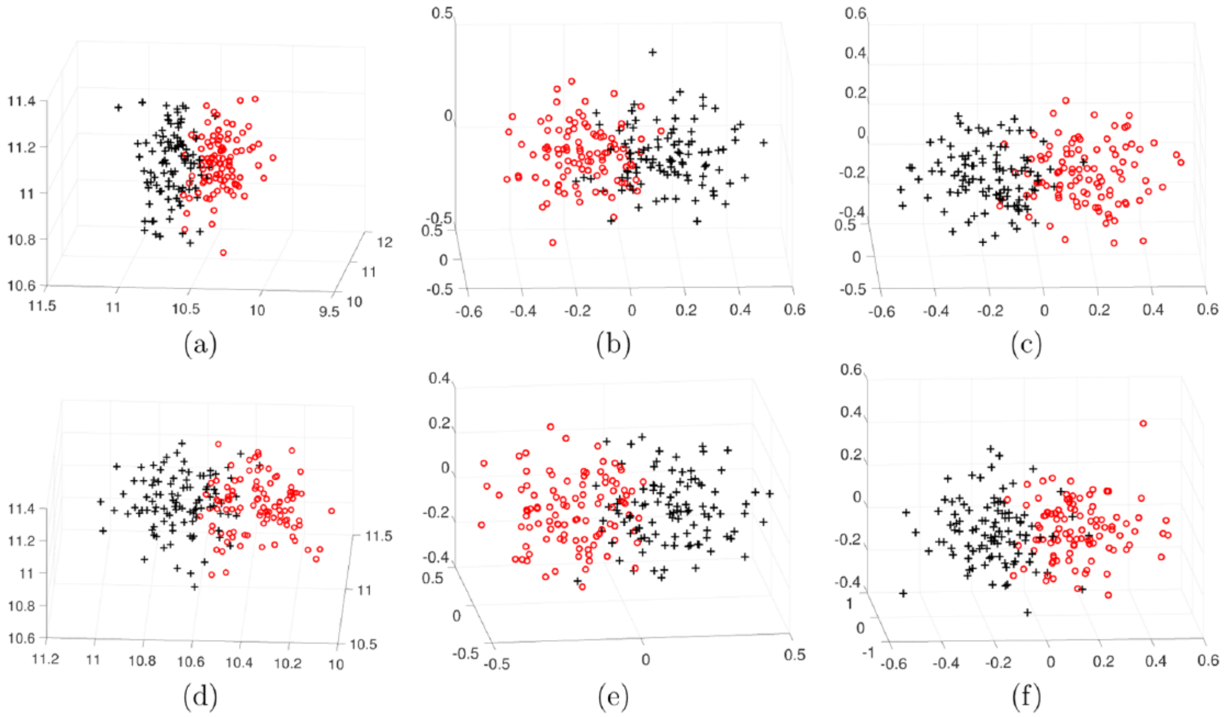

Figure 2. The 3-dimensional scatterplots for network PN1 sample points, with sample size $N = 200$, for $k = 3$ ((a) and (d)), $k = 4$ ((b) and (e)), and $k = 5$ ((c) and (f)). For $k = 4,5$, multidimensional scaling has been used to reduce the plot to 3 dimensions. For each value of *k*, there are 2 data plots arising from different samples: one possessing low LDA error ((a)-(c)) and the other possessing high LDA error ((d)-(f)). LDA indicates linear discriminant analysis; PN1, Pathway Network 1.

Results and Discussion

Synthetic data

To compare classification error rates using features from different moments, we repeatedly sample $R = 200$ times from both π_0 and π_1. For sample $r = 1,2,\ldots,200$, a sequential forward search[21] is used to find $k \in \{3,4,5\}$ features from $\mathbf{X}$, estimated with $\sim 200(\pm 25)$ cells. For $k = 3,4,5$, the error rates $\varepsilon_1^{r,k}, \varepsilon_2^{r,k}, \varepsilon_3^{r,k}$ are computed from a large set of test data for each category of moment features, $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$, respectively, and average error rates $\hat{\varepsilon}_1^k, \hat{\varepsilon}_2^k, \hat{\varepsilon}_3^k$ are computed from the R samples. We have computed error rates for linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), a support vector machine (SVM) with linear kernel, and a shallow feedforward neural network with hidden layer of size 10 (NNet). We use sample sizes $N = 100$ and $N = 200$, which are representative of many studies in genomics.

Figure 2 shows data plots for network PN1, with sample size $N = 200$, for $k = 3,4,5$, where for $k = 4,5$, multidimensional scaling[22] has been used to reduce the plot to 3 dimensions. For each value of k, there are 2 data plots arising from different samples: one possessing low LDA error (top figures) and the other possessing high LDA error (bottom figures). In all figures, the data appear to be compatible with linear discrimination, an observation that is born out in the error rates. Table 3 shows the average error rates for feature sets $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$, feature counts $k = 3,4,5$, and sample sizes $N = 100, 200$ for networks PN1 and PN2.

A considerable amount of insight can be gleaned from Table 3. Focusing first on PN1 with $N = 200$, we see the kind of behavior one might expect regarding the relation between $\mathcal{M}_1, \mathcal{M}_2$, and $\mathcal{M}_3$. For all values of k, the errors decrease from $\mathcal{M}_1$ to $\mathcal{M}_2$ to $\mathcal{M}_3$. Hence, using single-cell measurement is important. Most critically, the decrease from $\mathcal{M}_2$ to $\mathcal{M}_3$ is much greater than the decrease from $\mathcal{M}_1$ to $\mathcal{M}_2$. This means that adding mixed moments has a more significant effect than adding higher order single-variable moments. This behavior is important for genomics: the mixed moments capture gene-gene interaction, which is affected by regulatory mutations. If we fix the feature class, we do not see improvement for increasing k (a slight bit for $\mathcal{M}_3$). This means that 3 features are enough (for $N = 200$), the issue being to have mixed-moment features in the mix.

For the smaller sample size $N = 100$, the errors are greater but again the decrease from $\mathcal{M}_2$ to $\mathcal{M}_3$ remains significant for all values of k (albeit less than for $N = 200$). Much of the advantage of the added high-order moments is lost from $\mathcal{M}_1$ to $\mathcal{M}_2$ so that the errors remain essentially the same. We are observing the peaking phenomenon[23–25]: for fixed sample size, as the overall number of features grows, at first, the error decreases, but then it increases, the phenomenon being more prominent for small samples. Thus, the advantage of a superset of features is diminished and can actually be harmful. For a fixed feature class and increasing k, for $\mathcal{M}_1$ and $\mathcal{M}_2$, the errors get worse, and for $\mathcal{M}_3$, they remain the same. Feature selection mitigates to some extent the effect of peaking; however, as

Table 3. Average error rates $\hat{\varepsilon}_1^k, \hat{\varepsilon}_2^k, \hat{\varepsilon}_3^k$ for $k \in \{3,4,5\}$ and both networks PN1 and PN2.

			$k = 3$			$k = 4$			$k = 5$		
			M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3
			$\hat{\varepsilon}_1^3$	$\hat{\varepsilon}_2^3$	$\hat{\varepsilon}_3^3$	$\hat{\varepsilon}_1^4$	$\hat{\varepsilon}_2^4$	$\hat{\varepsilon}_3^4$	$\hat{\varepsilon}_1^5$	$\hat{\varepsilon}_2^5$	$\hat{\varepsilon}_3^5$
PN1	LDA	$N = 100$	0.2070	0.2073	0.1867	0.2092	0.2093	0.1867	0.2109	0.2106	0.1868
		$N = 200$	0.2008	0.1970	0.1629	0.2005	0.1973	0.1607	0.2006	0.1975	0.1590
	QDA	$N = 100$	0.2121	0.2129	0.1914	0.2189	0.2190	0.1948	0.2274	0.2270	0.2011
		$N = 200$	0.2029	0.1999	0.1684	0.2054	0.2030	0.1683	0.2084	0.2063	0.1696
	SVM	$N = 100$	0.2145	0.2152	0.1962	0.2163	0.2193	0.1992	0.2200	0.2226	0.1996
		$N = 200$	0.2040	0.2004	0.1699	0.2041	0.2014	0.1679	0.2045	0.2025	0.1692
	NNet	$N = 100$	0.2476	0.2412	0.2198	0.2479	0.2542	0.2234	0.2611	0.2563	0.2230
		$N = 200$	0.2219	0.2216	0.1868	0.2224	0.2196	0.1850	0.2268	0.2204	0.1820
PN2	LDA	$N = 100$	0.1019	0.1017	0.0995	0.1015	0.1018	0.0991	0.1002	0.1014	0.0998
		$N = 200$	0.0936	0.0907	0.0869	0.0923	0.0892	0.0847	0.0910	0.0882	0.0840
	QDA	$N = 100$	0.1048	0.1050	0.1035	0.1064	0.1069	0.1053	0.1076	0.1097	0.1079
		$N = 200$	0.0965	0.0935	0.0895	0.0962	0.0932	0.0893	0.0951	0.0938	0.0885
	SVM	$N = 100$	0.1081	0.1085	0.1092	0.1079	0.1119	0.1111	0.1085	0.1139	0.1147
		$N = 200$	0.0985	0.0953	0.0922	0.0981	0.0956	0.0917	0.0976	0.0956	0.0923
	NNet	$N = 100$	0.1358	0.1304	0.1277	0.1285	0.1319	0.1264	0.1349	0.1347	0.1266
		$N = 200$	0.1112	0.1059	0.1051	0.1095	0.1043	0.1030	0.1060	0.1046	0.1028

Abbreviations: LDA, linear discriminant analysis; NNet, neural network; PN1, Pathway Network 1; PN2, Pathway Network 2; QDA, quadratic discriminant analysis; SVM, support vector machine.

opposed to earlier studies in previous works,[23–25] in which there is no feature selection, the results in the work by Sima and Dougherty[26] demonstrate that peaking behavior is affected in peculiar ways by feature selection and is dependent on the classification rule.

Similar effects are seen for network PN2 with LDA but the improvement is significantly less for PN2 as compared with PN1. In part this is because the overall error rates are much smaller and, perhaps, in part because there are less mixed moments to choose from or because there are no strong mixed-moment effects due to regulatory change (which can happen if the effects of the regulatory change are spread out in the steady-state distribution).

To get a better sense of the contribution of extra features resulting from single-cell classification, in Figure 3, we have plotted the distributions $\mathcal{D}_1^k, \mathcal{D}_2^k, \mathcal{D}_3^k$ for the LDA error rates $\varepsilon_1^{r,k}, \varepsilon_2^{r,k}, \varepsilon_3^{r,k}$, respectively (mean errors shown on the horizontal axis), for $N = 200$ and $k = 3, 5$. For both networks, classification improvement when moving from $\mathcal{M}_1$ to $\mathcal{M}_2$ to $\mathcal{M}_3$ is apparent as the error rate distributions shift to the left.

QDA, SVM, and NNet show similar behavior to that of LDA for network PN1 and $N = 200$. Most importantly, for all k, the errors decrease from $\mathcal{M}_1$ to $\mathcal{M}_2$ to $\mathcal{M}_3$, and the decrease from $\mathcal{M}_2$ to $\mathcal{M}_3$ is much greater than the decrease from $\mathcal{M}_1$ to $\mathcal{M}_2$. Once again, the differences are less pronounced for $N = 100$ on account of peaking. Analogous comments also apply to network PN2, but to a lesser degree, as is the case with LDA.

Real data

For the fluorescent imaging data, as we have profiled 1 gene for each cell, for each case, there are 3 features associated with that gene: mean, variance, and skewness. To test the potential of moment-based classification, we have tested all 7 feature combinations: mean, variance, skewness, mean plus variance, mean plus skewness, variance plus skewness, and all 3 features. Linear discriminant analysis is used for classification (QDA performs poorly on account of small sample size). The 10-fold cross-validation averaged over 10 repeats has been used to estimate errors. The simulation results are summarized in Table 4.

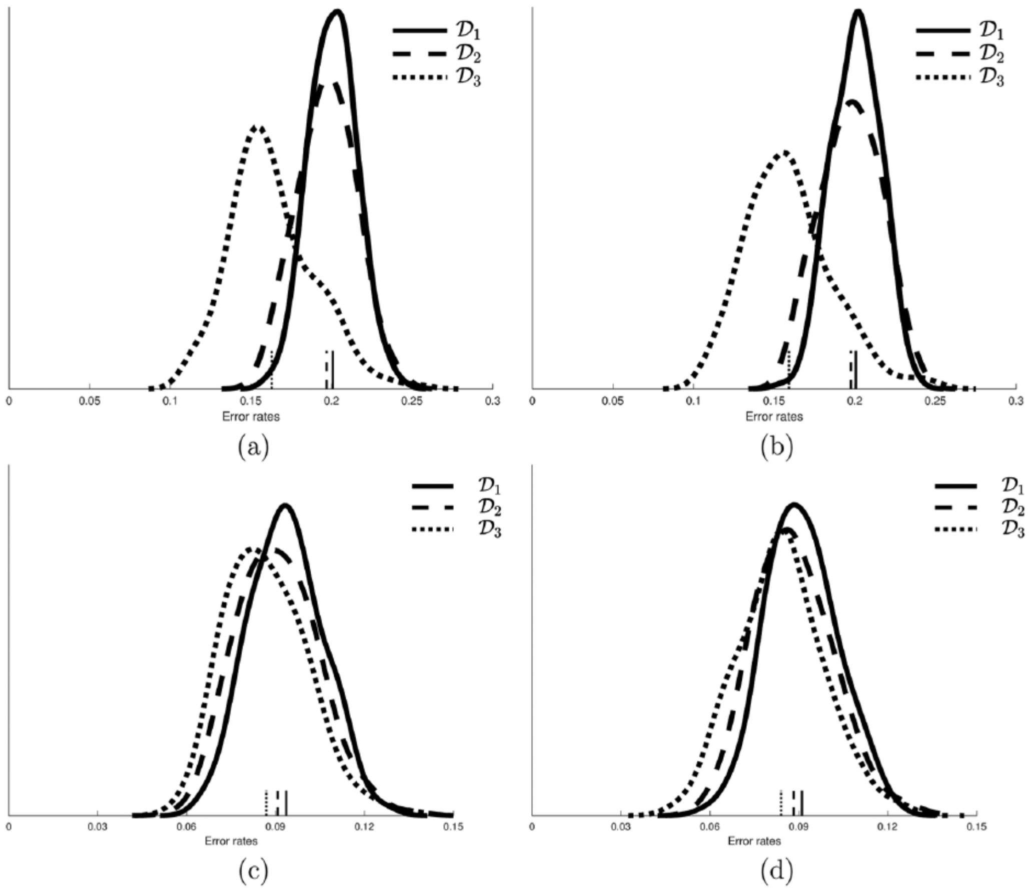

Figure 3. Distribution plots $\mathcal{D}_1^k, \mathcal{D}_2^k, \mathcal{D}_3^k$ (mean errors shown on the horizontal axis), for $N = 200$: (a) $k = 3$ for PN1, (b) $k = 5$ for PN1, (c) $k = 3$ for PN2, and (d) $k = 5$ for PN2.

Table 4. Classification error rates for linear discriminant analysis on all possible feature combinations for *Egfr* or Erbb2, based on 10-fold cross-validation repeated for 10 times.

	μ^1	μ^2	μ^3	$\mu^1 + \mu^2$	$\mu^1 + \mu^3$	$\mu^2 + \mu^3$	$\mu^1 + \mu^2 + \mu^3$
Egfr	0.597	0.434	0.440	0.516	0.416	0.376	0.406
Erbb2	0.083	0.246	0.579	0.038	0.075	0.248	0.038

μ^1—mean; μ^2—variance, μ^3—skewness.

With higher moments added to the feature pool, classification performance can improve significantly. For *Egfr*, it is hard to classify with only the mean. Using variance and skewness, classification performance improves from 0.597 to 0.376. For *Erbb2*, with the mean, the performance is already very good at 0.083. By adding variance, the error rate is cut by more than half to 0.038. Clearly, the higher moments improve performance.

Conclusions

The advent of single-cell expression measurement creates the potential for high-throughput expression-based classification with much greater accuracy than simply using mean expression over many cells. As we have demonstrated with both synthetic data generated from real networks and real single-cell data, higher order moments can improve moment-based classification, and the inclusion of mixed moments can make a more substantial improvement, not only because there are simply more features but also because mixed moments can capture gene-gene regulatory differences. Hopefully, these results will spur the development of more sophisticated single-cell technology so that practical sample sizes can be efficiently generated.

Author Contributions

CS, JH, and MLB generated the real data. CS and JH performed the computations. CS, JH, SK, and ERD drafted the manuscript. All authors discussed the results and commented on the final manuscript.

REFERENCES

1. Lake B, Chen S, Sos BC, et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotech.* 2018;36:70–80.

2. Fiers MWEJ, Minnoye L, Aibar S, GonzÃąlez-Blas B, Atak ZK, Aerts S. Mapping gene regulatory networks from single-cell omics data. *Brief Func Genom*. 2018:elx046.
3. Tang F, Lao K, Surani MA. Development and applications of single-cell transcriptome analysis. *Nat Meth*. 2011;8:S6–S11.
4. Bartfai T, Buckley PT, Eberwine J. Drug targets: single-cell transcriptomics hastens unbiased discovery. *Trends Pharm Sci*. 2012;33:9–16.
5. Karbalayghareh A, Braga-Neto UM, Dougherty E. Classification of single-cell gene expression trajectories from incomplete and noisy data. *IEEE/ACM Trans Comput Biol Bioinform*. 2017;15:99.
6. Kauffman S. Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theoret Biol*. 1969;22:437–467.
7. Kauffman SA. *The Origins of Order: Self-organization and Selection in Evolution* (Oxford Statistical Science Series). Oxford, UK: Oxford University Press; 2003.
8. Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*. 2002;18:261–274.
9. Faryabi B, Chamberland J-F, Vahedi G, Datta A, Dougherty E. Optimal intervention in asynchronous genetic regulatory networks. *IEEE J Select Topics Signal Proc*. 2008;2:412–423.
10. Wang Z, Jin S, Liu G, et al. DTWscore: differential expression and cell clustering analysis for time-series single-cell RNA-seq data. *BMC Bioinformatics*. 2017;18:270.
11. Esfahani MS, Dougherty ER. Effect of separate sampling on classification accuracy. *Bioinformatics*. 2014;30:242–250.
12. Karbalayghareh A, Braga-Neto U, Dougherty E. Classification of state trajectories in gene regulatory networks. *IEEE/ACM Trans Comput Biol Bioinformatics*. 2018;15:68–82.
13. Lauritzen S. *Graphical Models* (Oxford Statistical Science Series). Wotton-under-Edge, UK: Clarendon Press; 1996.
14. Qian X, Dougherty ER. Effect of function perturbation on the steady-state distribution of genetic regulatory networks: optimal structural intervention. *IEEE Trans Signal Proc*. 2008;56:4966–4976.
15. Hua J, Sima C, Cypert M, et al. Dynamical analysis of drug efficacy and mechanism of action using GFP reporters. *J Biol Syst*. 2012;20:403–422.
16. Hua J, Sima C, Cypert M, et al. Tracking transcriptional activities with high-content epifluorescent imaging. *J Biomed Optics*. 2012;17:046008.
17. Blake WJ, KAErn M, Cantor CR, Collins JJ. Noise in eukaryotic gene expression. *Nature*. 2003;422:633–637.
18. Elowitz MB, Leibler S. A synthetic oscillatory network of transcriptional regulators. *Nature*. 2000;403:335–338.
19. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Meth*. 2008;5:877–879.
20. Taniguchi K, Kajiyama T, Kambara H. Quantitative analysis of gene expression in a single cell by qPCR. *Nat Meth*. 2009;6:503–506.
21. Pudil P, Novovicova J, Kittler J. Floating search methods in feature selection. *Patt Recogn Lett*. 1994;15:1119–1125.
22. Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*. 1964;29:1–27.
23. Hughes G. On the mean accuracy of statistical pattern recognizers. *IEEE Trans Informat Theory*. 1968;14:55–63.
24. Jain A, Waller W. On the optimal number of features in the classification of multivariate Gaussian data. *Patt Recogn*. 1978;10:365–374.
25. Hua J, Xiong Z, Lowey J, Suh E, Dougherty ER. Optimal number of features as a function of sample size for various classification rules. *Bioinformatics*. 2005;21:1509–1515.
26. Sima C, Dougherty ER. The peaking phenomenon in the presence of feature selection. *Patt Recogn Lett*. 2008;29:1667–1674.