A Support Vector Machine Classifier for Recognizing Mitotic Subphases Using High-Content Screening Data

Charles Y. Tao, Jonathan Hoyt and Yan Feng

J Biomol Screen 2007; 12; 490 originally published online Apr 13, 2007;
DOI: 10.1177/1087057107300707

A Support Vector Machine Classifier for Recognizing Mitotic Subphases Using High-Content Screening Data

CHARLES Y. TAO, JONATHAN HOYT, and YAN FENG

High-content screening studies of mitotic checkpoints are important for identifying cancer targets and developing novel cancer-specific therapies. A crucial step in such a study is to determine the stage of cell cycle. Due to the overwhelming number of cells assayed in a high-content screening experiment and the multiple factors that need to be taken into consideration for accurate determination of mitotic subphases, an automated classifier is necessary. In this article, the authors describe in detail a support vector machine (SVM) classifier that they have implemented to recognize various mitotic subphases. In contrast to previous studies to recognize subcellular patterns, they used only low-resolution cell images and a few parameters that can be calculated inexpensively with off-the-shelf image-processing software. The performance of the SVM was evaluated with a cross-validation method and was shown to be comparable to that of a human expert. (*Journal of Biomolecular Screening* 2007:490-496)

Key words: high-content screening, support vector machine, mitosis, cell cycle

INTRODUCTION

MAMMALIAN CELL CYCLE is a highly regulated cellular proliferation process. Complex cellular regulatory mechanisms called *cell cycle checkpoints* are in place to control the accurate timing of each phase and to ensure the fidelity of chromosome segregation.[1]

Deregulation of cell cycle control is a major hallmark of cancer development. Many existing cancer chemotherapies target DNA synthesis with nucleotide analogs. Severe toxicities are usually associated with this approach because normal cell proliferation is also inhibited by these nucleotide analogs.[2] Recently, many defects in mitotic checkpoint control have been found to be specific to cancer cells.[3,4] For this reason, efforts to develop new cancer-specific therapies are increasingly focused on mitosis.[5] Cancer therapies targeting mitotic checkpoint kinases are also emerging.[6] For example, Aurora A kinase was found to be overexpressed in a wide range of human tumors, and specific inhibitors for Aurora kinase activity have shown efficacy in animal models.[5] Interestingly, inhibition of Aurora kinase either by RNA interference (RNAi) or small molecules in cultured cells results in a specific cytokinesis defect. Detection of this phenotype cannot be accomplished with conventional cell death/cell proliferation assays, although it has been successfully achieved using microscopy.[7]

Mitosis is a morphologically distinctive process that usually lasts up to several hours. It is conventionally divided into 5 subphases—namely, prophase, prometaphase, metaphase, anaphase, and telophase. Complex cellular regulative mechanisms are in place to control the accurate timing of each subphase and to ensure the fidelity of chromosome separation.[1]

Genome-wide small interfering (siRNA) and compound screens targeting mitosis control using high-content microscopy are becoming the new technology of choice to identify novel cancer targets and lead compounds for a number of reasons: 1) the multifaceted information (intensity, location, and morphology) revealed in a high-content screen provides an accurate description of the state of a cell,[8,9] and 2) a large population of cells is assayed in each experiment, making it possible to calculate informative statistics on various stages of the cell cycle.

The ability to identify the different stages of mitosis accurately is essential to this approach. Traditionally, this task is accomplished by manual inspection of individual cell images. An expert determines the cell cycle phase based on both the intensity and morphological information contained in the image of a cell. For example, a smaller nucleus area due to chromosome condensation during metaphase can be used to differentiate prometaphase cells from metaphase cells.

However, this is not a feasible solution for high-content screening experiments, which often measure tens of millions of cells in a single experiment. As a result, automated classification is needed.

Previously, various machine learning approaches have been taken to automatically recognize protein subcellular location

Genome and Proteome Sciences, Novartis Institutes for Biomedical Research, Cambridge, MA.

Received Oct 19, 2006, and in revised form Jan 23, 2007. Accepted for publication Feb 7, 2007.

Journal of Biomolecular Screening 12(4); 2007
DOI: 10.1177/1087057107300707

patterns.[10-14] However, most of these studies used manually acquired, high-resolution images, making it possible to extract rich, detailed quantitative information from the images for accurate classification. Furthermore, some image quantification parameters proposed in these studies are computationally expensive and cannot be calculated with off-the-shelf image-processing software. Finally, the patterns studied were typically distinct, easily separable subcellular structures.

Accurate, automated classification of mitotic subphases has been a challenge for the following reasons:

- The mitotic phase is significantly shorter in comparison with the other phases, resulting in a much smaller subpopulation (typically around 5%) of mitotic cells.
- The mitotic subphases represent the same cell in a succession of rapidly transitional states, and the boundaries between these states could be ambiguous.
- In most high-content screening experiments, image resolution is low due to considerations of image acquisition and processing time as well as data storage capacity.

Therefore, the classification method must be highly sensitive and able to use low-resolution images only.

In this study, we used a support vector machine (SVM) trained with a small number of cells that have been manually curated. SVM[15,16] recently has become the method of choice for many machine learning problems in computational biology.[17-20] We show in this article that for the purpose of recognizing mitotic subphases, high classification accuracy comparable to that of a human expert can be achieved with low-resolution images only. Furthermore, we used only a small number of parameters that are computationally inexpensive and can be calculated with most off-the-shelf image-processing software.

The benefit of being able to use only low-resolution images is obvious: they result in faster image acquisition and smaller file size and, therefore, are more amenable to high-throughput screening (HTS). As a result, more cells can be imaged, resulting in a larger sample size that is often statistically advantageous.

Using fewer parameters is also important for analyzing high-content screening data because any reduction in the time to calculate quantitative features on the single-cell level can be significant, considering the fact that tens of millions of cells are often scored in an experiment.

Neither the SVM implementation nor the image-processing methods used here are novel. Extensive discussions on the practical aspects of SVM, such as feature selection,[21,22] model selection,[23] and strategies for multiclass classification,[24,25] can be found in the literature. Many SVM implementations, such as LIBSVM,[23] BSVM,[26] Gist,[27] SvmFu3,[28] LS-SVMlab,[29] and SVM-light,[30] and accompanying tools are available, making it relatively easy to adopt the technology. For this article, we used LIBSVM, a widely available off-the-shelf implementation of the SVM.

The goal of this article is to show that SVM, when applied appropriately to even a simple set of parameters extracted from low-resolution images, can be a practical solution to the problem of automatically recognizing mitotic subphases.

METHODS

High-content cell cycle screening

Hela cells were seeded into black, clear-bottomed 384-well plates (Greiner) at a density of 2000 cells per well. The cells were then transfected with siRNA. After 48 h, the cells were pulse labeled with 1 μM Ethynyl-dU (Edu) for 40 min and then fixed and stained with 4 different fluorescent probes. Channel 1 measures the fluorescence of Hoescht 33342 (Invitrogen, Carlsbad, CA), a DNA-specific probe; channel 2 measures the fluorescence of rabbit antiphosphohistone H3 (PH3) antibody; channel 3 detects DNA synthesis that occurs during the S-phase by measuring the incorporation of rhodamine-azide-labeled deoxyuridine (courtesy of Timothy Mitchison); and channel 4 detects immunofluorescent labeling signals of alpha-tubulin (Sigma Aldrich, St. Louis, MO). The images were captured using Arrayscan Reader 3.5 (Cellomics, Pittsburg, PA), equipped with a 10× objective, and an ORCA-ER CCD camera (Hamamatsu, Shizuoka, Japan) running at 2 × 2 binning (512 × 512 pixels/image, 1.34 micron/pixel resolution).

Image processing and quantification

Image processing is a necessary step but not the focus of this article. Therefore, we simply used the program Cellomics Morphology Explorer that had been supplied by the manufacturer of the scanner. The software first recognizes each individual nucleus outlined by DNA stain using a standard watershed algorithm. Several hundred to a few thousand cells were identified for each condition. For each cell, the Cellomics image-processing software quantified the images by calculating more than a hundred parameters, which mainly fall into the following categories:

- Geometric properties: such as the area, perimeter, and shape of the cell nucleus; the location of a cell; average distance of a cell to its neighbors; and so on
- Intensity information: such as the content of a protein, as reflected by the intensity of the corresponding fluorescent dye, and the variance, skewness, kurtosis, and so on of the intensity distribution

These parameters represent various attributes of the physiological state of the cell, such as morphology, location, and content level of relevant biomolecules. Once the boundary of a cell is recognized, these parameters can easily be calculated.

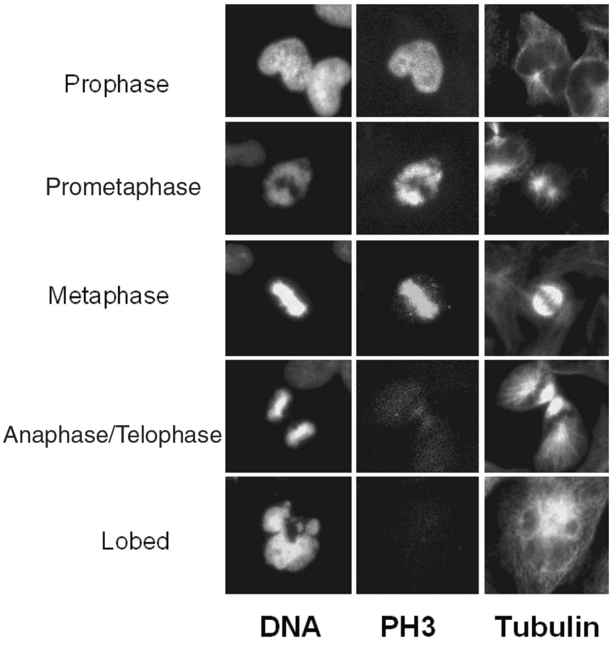

FIG. 1. Sample images of cells in various mitotic stages. The deoxyuridine channel is not shown here because it is a marker for DNA synthesis during the S-phase only.

Training data

A list of 369 cells with a high mitotic index (PH3 level) and confirmed visually by the expert to be in mitosis was manually classified. Some examples are shown in **Figure 1**. The cells were then manually assigned to each of the following classes: prophase, prometaphase, metaphase, anaphase/telophase, and multilobed.

The following criteria were used by the expert to annotate the training set: cells that appeared as normal interphase cells by Hoescht staining with increased levels of histone H3 phosphorylation (PH3) were categorized as prophase. Cells with elevated PH3 and condensed chromosomes that had not completely aligned or achieved bipolar attachment to the mitotic spindle were classified as prometaphase cells. Cells were identified as metaphase based on elevated PH3 with the complete bipolar attachment and alignment of all chromosomes between the spindle poles. Anaphase/telophase cells were identified as pairs of rod-shaped chromatin that were smaller in size than metaphase cells and exhibited markedly decreased levels of PH3 staining. The multilobed phenotype, common in cells with compromised mitotic checkpoint control, is characteristic of cells that exit mitosis prior to proper alignment and segregation of sister chromatids. In these cells, the nuclei often appear irregularly shaped with protruding lobes. Histone H3 staining is absent in these cells.

Support vector machine

To train an SVM with a training data set (x_i, y_i), i = 1, . . . , n, where $x_i \in R^n$ and $y_i \in \{-1, 1\}$ is a binary class label, the following optimization problem is solved[16]:

$$\min_{w,\beta,\xi_i} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i,$$

subject to

$$y_i(w^T \phi(x_i) + \beta) \geq 1 - \xi_i, \qquad (1)$$

$$\xi_i \geq 0, \sum_i \xi_i \leq Constant.$$

$C > 0$ is the penalty for each erroneous classification, and ξ_i is a slack variable to allow for margin violations. The total budget for margin violations is given by the constant. The training vectors x_i are mapped into a higher dimensional space by the function ϕ. In practice, the problem is simplified because only the so-called kernel function $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$, instead of ϕ itself, needs to be specified. The preferred 1st choice of kernel function is called the radial basis function (RBF), which can handle the nonlinear relationship between class labels and attributes. It has the following form:

$$K(x_i, x_j) = \exp(-\gamma \,\|\, x_i - x_j \|^2), \qquad (2)$$

where $\gamma > 0$ determines the effective range of the kernel.

The SVM was trained with the LIBSVM implementation.[23] To avoid large numerical values dominating smaller ones, each parameter is scaled so that its mean is zero and its standard deviation is 1.

The SVM described above is for the case when class labels are binary (i.e., there are only 2 classes). However, the current classification problem is a multiclass one. An extensive comparison of the various multiclass methods was made by Hsu and Lin,[24] who concluded that the "one-against-one" and direct acyclic graph (DAG) methods are more suitable for practical use than the other methods. Although this conclusion remains controversial,[25] we decided to take a practical approach here: we will simply use the one-against-one method as long as the classification results are satisfactory.

Using the one-against-one method with k classes, $k(k-1)/2$ SVMs were trained, 1 for each pair of classes. The predicted class of a particular data point was decided with a simple majority vote of the results of the k $(k-1)/2$ SVMs.

For best performance, the SVM was optimized by tuning the key parameters. The key parameters of the SVM are the penalty parameter C and γ in the RBF function, which determines the effective range of the kernel. The optimal combination of parameters was searched on a simple grid. C was chosen from the geometric series (2^{-10}, 2^{-9}, . . . , 2^{10}), and γ was chosen from the

Table 1. Feature Set 1 (FS1), Consisting of 59 Parameters

XCentroid	YCentroid	AreaCh1
PerimCh1	ShapeP2ACh1	ShapeLWRCh1
ShapeBFRCh1	LengthCh1	WidthCh1
AngleCh1	FiberLengthCh1	FiberWidthCh1
ConvexHullAreaRatioCh1	ConvexHullPerimRatioCh1	EqCircDiamCh1
EqSphereVolCh1	EqSphereAreaCh1	EqEllipseLWRCh1
EqEllipseProlateVolCh1	EqEllipseOblateVolCh1	NeighborMinDistCh1
NeighborAvgDistCh1	NeighborVarDistCh1	TotalIntenCh1
AvgIntenCh1	VarIntenCh1	SkewIntenCh1
KurtIntenCh1	EntropyIntenCh1	DiffIntenDensityCh1
TotalIntenCh2	AvgIntenCh2	SpotFiberCountCh3
TotalIntenCh3	AvgIntenCh3	VarIntenCh3
SkewIntenCh3	KurtIntenCh3	EntropyIntenCh3
DiffIntenDensityCh3	IntenCoocMaxCh3	IntenCoocContrastCh3
IntenCoocEntropyCh3	IntenCoocASMCh3	TotalIntenCh4
AvgIntenCh4	VarIntenCh4	SkewIntenCh4
KurtIntenCh4	EntropyIntenCh4	DiffIntenDensityCh4
IntenCoocMaxCh4	IntenCoocContrastCh4	IntenCoocEntropyCh4
IntenCoocASMCh4	AvgRadialIntenCh4	VarRadialIntenCh4
SkewRadialIntenCh4	KurtRadialIntenCh4	

geometric series ($2^{-5}, 2^{-4}, \ldots, 2^{5}$). For each pair of ($C$, γ), the overall error rate was estimated with the leave-one-out cross-validation method, and the pair that gave the lowest overall error rate was selected.

RESULTS AND DISCUSSION

Feature selection

Selecting a subset of parameters to train the SVM is important for 2 reasons: first, a smaller set of attributes reduces both the computational cost and the time needed to export data, which could be significant in a high-content screen due to the huge number of cells assayed in an experiment; second, irrelevant, distracting parameters could potentially cause performance deterioration of a machine learning algorithm.[31]

Among the more than 100 parameters that can be exported from the image-processing software, we first removed those that are obviously irrelevant. The resulting 59-parameter feature set, FS1, is shown in **Table 1**.

To further reduce the number of parameters, we performed feature selection on FS1. Several feature selection approaches were applied, resulting in feature sets of different sizes. The performance of each feature set was evaluated using cross-validation, as discussed later.

An 8-parameter feature set FS2 (**Table 2**), selected using a classification tree approach with the Gini index as a measure of the impurity,[32] was chosen for the following reasons: 1) its compact size, 2) the accuracy of classification results (discussed later), and 3) the fact that it is dominated by channel 1 (DNA) features, which the expert also had relied on heavily for manual annotation. The classification tree was constructed using the R package rpart.[33] The 8 parameters are those that were actually used in the resulting classification tree to split the nodes.[31]

Table 2. Selected Feature Sets

Feature Set	*Selected Parameters*		
FS2			
AreaCh1	WidthCh1	SkewIntenCh1	VarIntenCh1
TotalIntenCh2	AvgIntenCh2	KurtIntenCh3	KurtIntenCh4
FS3			
TotalIntenCh1	AreaCh1	TotalIntenCh2	TotalIntenCh3

Feature set 2 (FS2) was selected using a decision tree approach. Feature set 3 (FS3) was manually selected based on biology of the cell cycle.

The composition of automatically selected feature sets such as FS2 is often hard to interpret. Therefore, we also manually selected a very small feature set FS3, consisting of only 4 parameters—namely, the total intensity of channels 1 to 3 (DNA, PH3, and EdU) and the area of the nucleus, based solely on our understanding of the biology of the cell cycle. FS3 is also shown in **Table 2**.

Classification results

Comparison of the classification predicted by the SVM with the actual classification by the expert was made. Two different cross-validation methods were used to evaluate the performance of the SVM: the "leave-one-out" and the 10-fold cross-validation.

With the leave-one-out cross-validation method, the SVM trained with all but 1 of the 369 cells in the data set was used

Table 3. Cross-Validation Results, Using Leave-One-Out Cross-Validation

	Expert Annotated					
SVM Predicted	*Pro*	*Prometa*	*Meta*	*Ana/Telo*	*Lobed*	*Accuracy*
FS1						
Pro	**35**	1	0	0	3	
Prometa	11	**92**	5	0	2	
Meta	5	6	**99**	0	0	89.2%
Ana/Telo	1	0	0	**78**	1	
Lobed	2	2	0	1	**25**	
FS2						
Pro	**31**	3	1	0	0	
Prometa	15	**91**	6	0	0	
Meta	5	7	**96**	0	0	88.3%
Ana/Telo	1	0	1	**79**	2	
Lobed	2	0	0	0	**29**	
FS3						
Pro	**30**	1	0	0	1	
Prometa	12	**72**	17	0	2	
Meta	8	27	**86**	0	0	79.4%
Ana/Telo	3	0	1	**79**	2	
Lobed	1	1	0	0	**26**	

SVM, support vector machine; FS1, feature set 1; FS2, feature set 2; FS3, feature set 3.

to predict the category of the excluded cell. This was repeated 369 times so that each cell was excluded exactly once.

Another method is the 10-fold cross-validation. With this method, the data set was randomly divided into 10 almost equally sized subsets. The sampling was stratified across all classes to ensure that each class was represented proportionally in the training and test sets. Nine of the 10 subsets were combined and used to train the SVM, and the remaining subset was used for testing. This was repeated 10 times by selecting a different subset for testing each time.

We found that these 2 methods consistently produced very similar results. Therefore, only the results of the leave-one-out method are shown in **Table 3**.

Sensitivity and specificity were calculated as another measure of the performance of the SVM, as shown in **Table 4**. Again, only the leave-one-out numbers are shown because the 10-fold cross-validation gave very similar results. For each class, the sensitivity and specificity of prediction were calculated as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \tag{4}$$

where *TP* is the number of true positives, *FN* is the number of false negatives, and so on.

Table 4. Sensitivity and Specificity, Using Leave-One-Out Cross-Validation

	Pro	*Prometa*	*Meta*	*Ana/Telo*	*Lobed*
Sensitivity (%)					
FS1	64.8	91.1	95.2	98.7	80.6
FS2	57.4	90.1	92.3	100.0	93.5
FS3	55.6	71.3	82.7	100.0	83.9
Specificity (%)					
FS1	98.7	93.3	95.8	99.3	98.5
FS2	98.7	92.2	95.5	98.6	99.4
FS3	99.4	88.4	86.8	97.9	99.4

FS1, feature set 1; FS2, feature set 2; FS3, feature set 3.

For example, 99 cells were classified as in metaphase by both the SVM and the expert, and 5 cells were classified as in metaphase by the expert but not by the SVM. Therefore, *TN = 99*, *FN = 5*, and the sensitivity is 95.2%. In addition, the SVM contradicted the expert by classifying 11 other cells as in metaphase. Therefore, *TN = 369 – 95 – 5 – 11 = 254*, *FP = 11*, and the specificity is 95.8%.

Table 4 shows consistently high specificity in all classes. This indicates that the SVM generates very few false positives. The sensitivity is also high in all but the prophase population. Therefore, the number of false negatives is also low, except for prophase.

Misclassified cells

To investigate the reasons for the misclassification, the misclassified cells using FS1 were reexamined by the expert, sometimes using other information, such as tubulin structure, which had not been used during the initial classification. The expert reaffirmed the initial classification in about 60% of the cases. Twenty percent of the cases seemed to be borderline because the cells seemed to be in a transitional state between the phases. In such cases, either the expert or the SVM could be right. The expert reversed the initial classification and agreed with the SVM in about 10% of the cases, after using other information such as tubulin channel images that had not been used originally. The remaining 10% were determined to be cells not in mitosis (such as apoptotic cells) or cells whose images are blurry.

In the reaffirmed cases, there are often some indications from the original images why the SVM did not make a correct classification. In most cases, the state of a cell is atypical and not represented sufficiently in the training set. A larger training set, incorporating more diverse situations, is likely to help improve the performance in such cases.

Separation of prophase and prometaphase

We observed the largest error rate in prophase and prometaphase separation. The reasons are multiple. Unlike metaphase, when the chromosomes are tightly aligned along the spindle midzone, prophase and prometaphase are both much more dynamic compared to metaphase. The transition from prophase to prometaphase is a continuous and dynamic process. The boundary between these 2 subphases of mitosis is not distinct. An expert separates the 2 almost solely based on the appearance of chromosome condensation. In fact, due to the ambiguity, 2 experts inspecting the same images could have come to different conclusions in some cases. A training set that includes more prophase and prometaphase cells, as well as more descriptors of image textures, might lead to improvement in the separation of these 2 phases.

CONCLUSIONS

In conclusion, we have shown that SVMs are a valuable tool to classify the overwhelming number of cells assayed in a high-content screening into different phases of cell cycle.

It is interesting to evaluate the performance of FS3, which had been manually selected solely based on prior knowledge of cell biology.

Not surprisingly, as shown in **Tables 3** and **4**, there is some deterioration in performance because significantly fewer parameters had been used. However, the performance is still acceptable. This observation is important under certain circumstances where the cost of time and storage is an issue, but the requirement for accuracy is lower. For such applications, the benefit of using a smaller set probably outweighs the deterioration in performance.

Due to the varying experimental conditions from screen to screen, it might be necessary to generate a training set for each screen. In fact, we have seen in practice that the SVM trained with one screen usually cannot be applied directly to another because of the differences resulting from different cell lines and experimental conditions. Even so, the workload can be dramatically reduced. For example, assuming that 5% of the cell population is mitotic, there will be about a half-million mitotic cells in a whole genome screen, which makes manual annotation almost impossible. With the SVM approach, only a small training set of a few hundred cells needs to be annotated manually.

If the screens are very similar in nature, however, the SVM trained on the data of one screen should be applicable to the data of another. Systematic plate-to-plate and well-to-well variations, typical in most high-throughput screens, can be minimized with data normalization procedures such as median polishing. Such procedures, in addition to scaling of data, are necessary for optimal performance of the SVM.

The tasks of differentiating various stages in interphase (G1, S, G2) and separating interphase cells from mitotic cells are much easier because these interphase stages can often be determined by distinct, reliable markers such as DNA content, deoxyuridine level, and PH3 level.

REFERENCES

1. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P: *Molecular Biology of the Cell*. New York: Garland Science, 2002.
2. Hardman JG, Limbird LE, Gilman AG: *Goodman & Gilman's The Pharmacological Basis of Therapeutics*. 10th ed. New York: McGraw-Hill Professional, 2001.
3. Weaver BA, Silk AD, Montagna C, Verdier-Pinard P, Cleveland DW: Aneuploidy acts both oncogenically and as a tumor suppressor. *Cancer Cell* 2007;11:25-36.
4. Baker DJ, Chen J, van Deursen JM: The mitotic checkpoint in cancer and aging: what have mice taught us? *Curr Opin Cell Biol* 2005;17: 583-589.
5. Harrington EA, Bebbington D, Moore J, Rasmussen RK, Ajose-Adeogun AO, Nakayama T, et al: VX-680, a potent and selective small-molecule inhibitor of the Aurora kinases, suppresses tumor growth in vivo. *Nat Med* 2004;10:262-267.
6. Sausville EA: Aurora kinases dawn as cancer drug targets. *Nat Med* 2004;10:234-235.
7. Eggert US, Kiger AA, Richter C, Perlman ZE, Perrimon N, Mitchison TJ, et al: Parallel chemical genetic and genome-wide RNAi screens identify cytokinesis inhibitors and targets. *PLoS Biol* 2004;2:e379.
8. Mitchison TJ: Small-molecule screening and profiling by using automated microscopy. *Chembiochem* 2005;6:33-39.
9. Perlman ZE, Slack MD, Feng Y, Mitchison TJ, Wu LF, Altschuler SJ: Multidimensional drug profiling by automated microscopy. *Science* 2004;306:1194-1198.
10. Zhou XB, Wong STC: High content cellular imaging for drug development. *IEEE Signal Processing Magazine* 2006;23:170-174.
11. Boland MV, Murphy RF: A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of Hela cells. *Bioinformatics* 2001;17:1213-1223.
12. Long X, Cleveland WL, Yao YL: Automatic detection of unstained viable cells in bright field images using a support vector machine with an improved training procedure. *Comput Biol Med* 2006;36:339-362.
13. Huang K, Murphy RF: Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *BMC Bioinformatics* 2004;5:78.
14. Murphy RF: Cytomics and location proteomics: automated interpretation of subcelluar patterns in fluorescence microscope images. *Cytometry A* 2005;67:1-3.
15. Boser BE, Guyon IM, Vapnik VN: A training algorithm for optimal margin classifiers. In Haussler D (ed): *5th Annual ACM Workshop on COLT*. Pittsburgh: ACM Press, 1992:144-152.
16. Cortes C, Vapnik V: Support-vector networks. *Machine Learning* 1995; 20:273-297.
17. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al: Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA* 2000;97:262-267.
18. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D: Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 2000;16:906-914.
19. Ding CH, Dubchak I: Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* 2001;17:349-358.
20. Mao Y, Zhou X, Pi D, Sun Y, Wong ST: Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection. *J Biomed Biotechnol* 2005;2:160-171.

21. Chen Y-W, Lin C-J: Combining SVMs with various feature selection strategies. In Guyon I, Gunn S, Nikravesh M, Zadeh L (eds): *Feature Extraction, Foundations and Applications*. New York: Springer-Verlag, 2006.
22. Zhou XB, Liu KY, Sabatini B, Wong STC: Mutual information based feature selection in studying perturbation of dendritic structure caused by TSC2 inactivation. *Neuroinformatics* 2006;4:81-94.
23. Chang CC, Lin CJ: LIBSVM: a library for support vector machines [Online]. Retrieved from http://www.csie.ntu.edu.tw/cjlin/libsvm
24. Hsu CW, Lin CJ: A comparison of methods for multi-class support vector machines. *IEEE Trans Neural Networks* 2002;13:415-425.
25. Rifkin R, Klautau A: In defense of one-vs-all classification. *J Machine Learning Res* 2004;5:101-141.
26. http://www.csie.ntu.edu.tw/~cjlin/bsvm/
27. http://microarray.cpmc.columbia.edu/gist/
28. http://five-percent-nation.mit.edu/SvmFu/
29. http://www.esat.kuleuven.ac.be/sista/lssvmlab/
30. Joachims T: Making large-scale SVM learning practical. In Schölkopf B, Burges C, Smola A (eds): *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA: MIT Press, 1999.
31. Witten IH, Frank E: *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. San Francisco: Morgan Kaufmann, 2005.
32. Venables WN, Ripley BD: *Modern Applied Statistics with S*. 4th ed. New York: Springer-Verlag, 2002.
33. http://cran.r-project.org/src/contrib/Descriptions/rpart.html

Address reprint requests to:
Charles Y. Tao
Genome and Proteome Sciences
Novartis Institutes for Biomedical Research
250 Massachusetts Avenue
Cambridge, MA 02139

E-mail: charles.tao@novartis.com