GemIdent: An Interactive Statistical Image Segmentation System

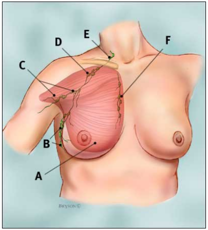
Adam Kapelner* & Susan Holmes**

*Department of Statistics
The Wharton School, University of Pennsylvania

**Department of Statistics
Stanford University

Breast Cancer and Lymphnodes

The breast drains lymph fluid to surrounding lymph nodes.

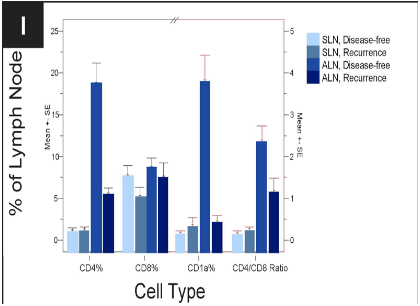


In breast cancer patients, cancer can metastasize in the lymph nodes.

This is a major prognostic indicator — most patients do not do well after lymphnode metastisis.

Criminals in Police Headquarters?

How is it possible that cancer can set up shop in lymph nodes, the immune cell centers?

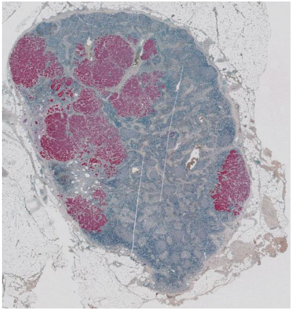


How is it possible that some patients can weather the storm and some cannot? Kohrt et al, 2005 showed that certain populations of immune cells is correlated with survival.

But what is happening biologically? Maybe the answer lies in how the cancer and the immune cells are *spatially distributed*.

A typical lymphnode

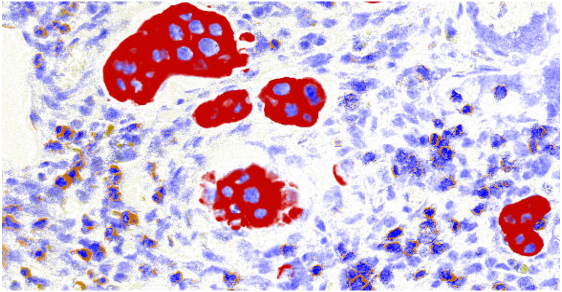
Here is a typical lymph node measuring about 1 cm in diameter. It was surgically removed and sliced into sections 3 μm thick. Cell populations of interest have been stained with different colors using immunohistochemistry.



In the above image, cancer is red, and CD4 T-cells (a type of immune cell) are brown, and all nuclei are "counterstained" blue.

Under the microscope

The entire node has been imaged using automated microscopy at 200X creating a few thousand subimages, for example:



If the goal is to find spatial distributions of cells, we must find the location of *every* cell nucleus in the lymphnode and determine the type of cell, *i.e.* its "phenotype".

The basic classification problem

Finding cells in images can be thought of as a classification problem for *each and every* pixel i, j:

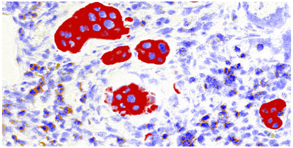
$$\hat{y}_{ij} = f(\mathbf{x}_{ij}), \qquad \hat{y}_{ij} \in \{0, 1, 2, \ldots, P\}$$

Where:

- $\mathbf{x}_{ij}$ is a vector of features / covariates
- "f" is a classification machine
- $\hat{y}_{ij}$ represents the machine's best guess for the pixel's phenotype. "0" represents not belonging to any phenotype.

Covariates?

Which features are important *i.e.* how do you build each $\mathbf{x}_{ij}$?



My naive thinking was:

I stain (color) information is obviously important

II rotational invariance

III the presence of stains at certain distances from the pixel of interest is key

Color separation via Mahalanobis distance

Images are in RGB format which means each pixel is has an 8-bit red value, 8-bit green value, and 8-bit blue value.

The image would be more useful if decomposed into $\{0, 1, \ldots, S\}$ relevant stains (where 0 represents the background, usually white). How to do this?

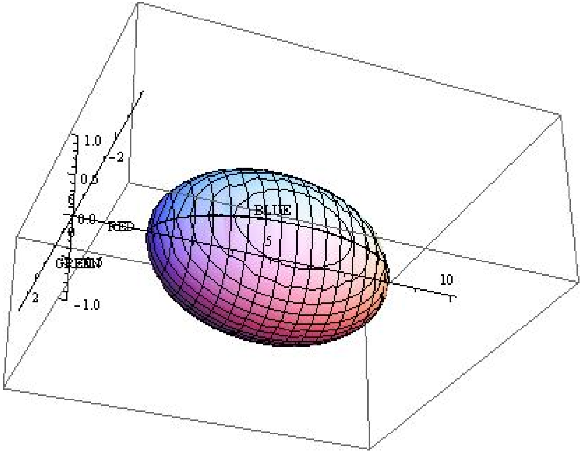
After getting a few examples of the color, use the Mahalanobis distance function:

$$d(\mathbf{t}_{ij}) = \sqrt{(\mathbf{t}_{ij} - \boldsymbol{\mu}_{\text{blue}})^T \ \hat{\Sigma}^{-1}_{\text{blue}} \ (\mathbf{t}_{ij} - \boldsymbol{\mu}_{\text{blue}})}$$

where $\mathbf{t}_{ij}$ is the 3×1 RGB-color vector at the pixel of interest, $\boldsymbol{\mu}_{\text{blue}}$ is the mean RGB-color vector for the "blue" stain, and $\hat{\Sigma}_{\text{blue}}$ is the 3×3 sample covariance matrix of all the examples of "blue" pixels.

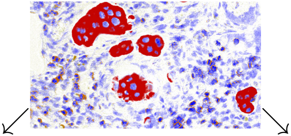
Color separation via Mahalanobis distance

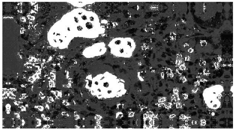
The distance function "*d*" assigns large values to pixels with colors far away from the "blue" stain, and small values to pixels with colors close to the "blue:"



Color Decomposition

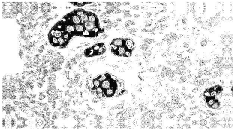
Thereby, our image can be decomposed into score matrices for each stain (as well as the background) which looks sort of like this:



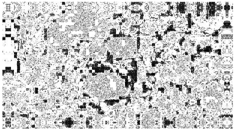


Stain: Blue

Stain: Brown



Stain: Red



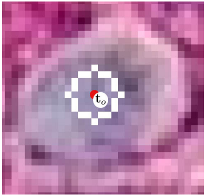
Stain: Background

We denote each of these score matrices F_s.

Ring Scores

For each pixel, we know that stain information away from the pixel itself is important, and we know that the images are rotationally invariant.

A natural choice would be to create "ring scores" which reflect "how much" of a certain stain exists a certain radius away from the pixel:



For each stain, rings $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_R\}$ (where R is the maximum radius) are considered important.

Feature computation and the feature vector

The score for stain s at radius r can be computed as follows by just adding up the values:

$$\ell_{s,r}(\mathbf{t}_o) = \sum_{\mathbf{t} \in \mathbf{c}_r} F_s(\mathbf{t}_o + \mathbf{t})$$

Now the feature vector for pixel i, j can be created by concatenating all the ring scores for each stain at each radius:

$$\mathbf{x}_{ij} = [\ell_{1,1}, \ell_{1,2}, \ldots, \ell_{1,R}, \ell_{2,1}, \ell_{2,2}, \ldots, \ell_{2,R}, \ldots\ldots, \ell_{S,1}, \ell_{S,2}, \ldots, \ell_{S,R}]$$

Note: somewhat naive — no interactions, no transformations

Supervised Statistical Learning

What would make sense would be to give a few examples of each phenotype, then create a machine that would make future predictions:

$$\underbrace{\begin{bmatrix} \text{---}\ \mathbf{x}_{ij}\ \text{---} \\ \text{---}\ \mathbf{x}_{ij}\ \text{---} \\ \text{---}\ \mathbf{x}_{ij}\ \text{---} \\ \text{---}\ \mathbf{x}_{ij}\ \text{---} \\ \vdots \\ \text{---}\ \mathbf{x}_{ij}\ \text{---} \\ \text{---}\ \mathbf{x}_{ij}\ \text{---} \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ \vdots \\ P \\ P \end{bmatrix}}_{\text{``Training Data''}} \xrightarrow{\text{Machine Creation}} \boxed{f}$$

What machines can you create that perform $\hat{y}_{ij} = f(\mathbf{x}_{ij})$?

Regression & Machine Learning

Regression *e.g.* OLS (a la Stat 102)

- Effect Tests
- p-values
- Confidence Intervals
- Prediction

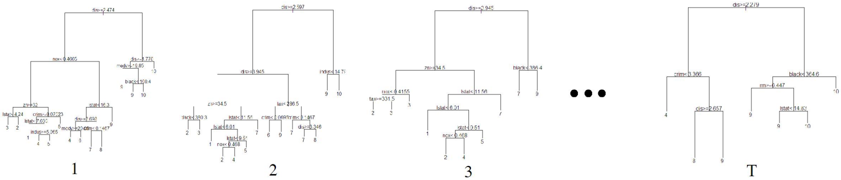
Machine Learning

- ~~Effect Tests~~
- ~~p-values~~
- ~~Confidence Intervals~~
- Prediction

As long as we only care about prediction. . .

Random Forests

For pixel classification, we chose the "machine" with the lowest error rate. At the time of conception (2006), Random Forests beat out the contenders for our data.



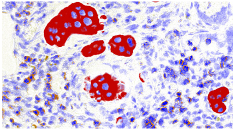
Random Forests Algorithm - Decision trees (i.e. CARTs) with a few twists:

1. Bootstrap training data
2. Select a subset of features to evaluate at each node
3. Repeat 1-2 to build T trees

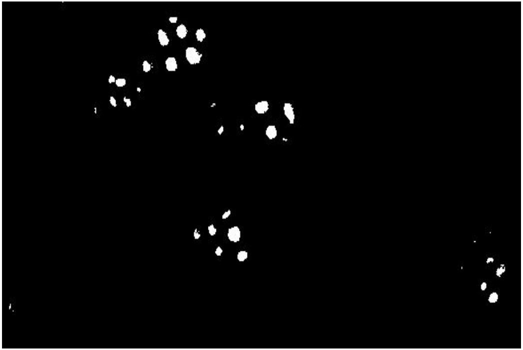
To predict, let trees vote democratically

Classification

With 5 training points for cancer, and 50 NON points, we build a random forest, then proceed to classify *each and every* pixel of the example image:



$\xrightarrow{f}$



These blobs are then post-processed using a simple erosion algorithm to yield centroids *i.e.* coordinates of each cell.

Introduction to the software

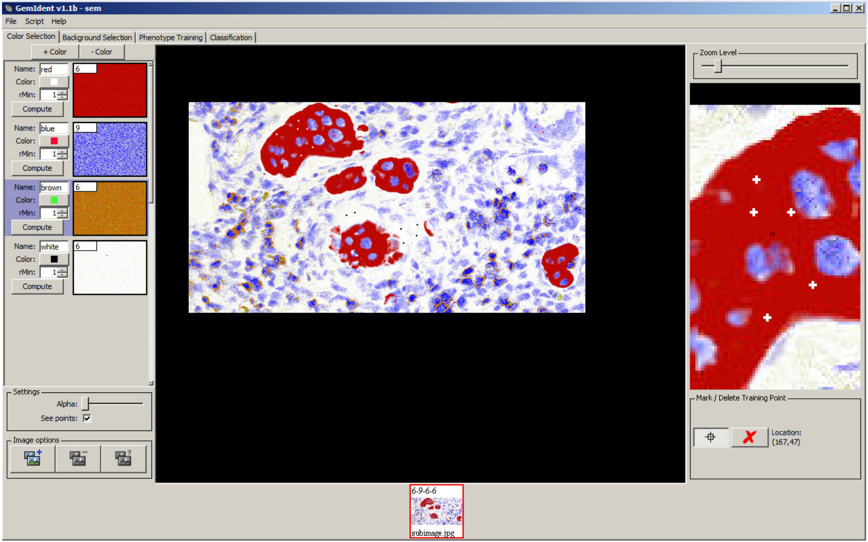
This algorithm, coupled with a graphical user interface, was developed in Java into a program called GemIdent and open-sourced under GPL.

The software has the following main features organized into panels:

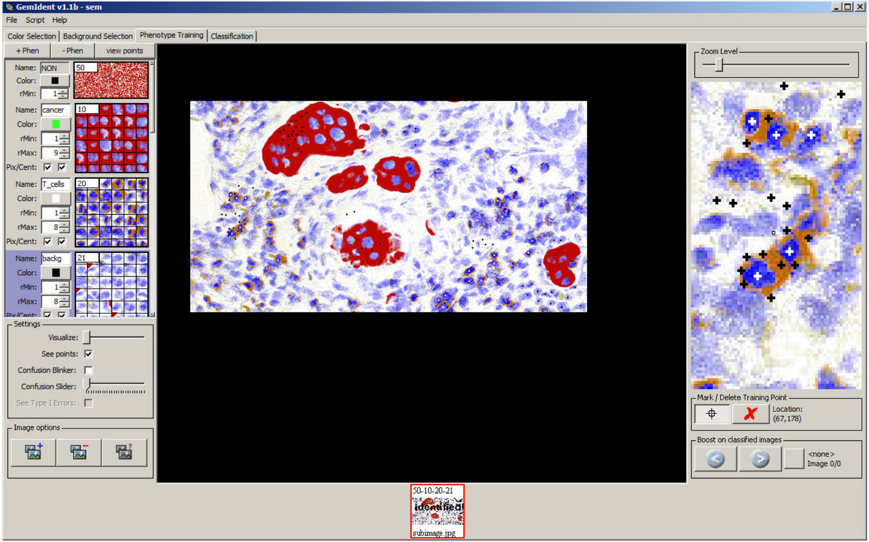
- Color Selection
- Phenotype Training
- Classification
- Data Analysis (not shown)

www.gemident.com

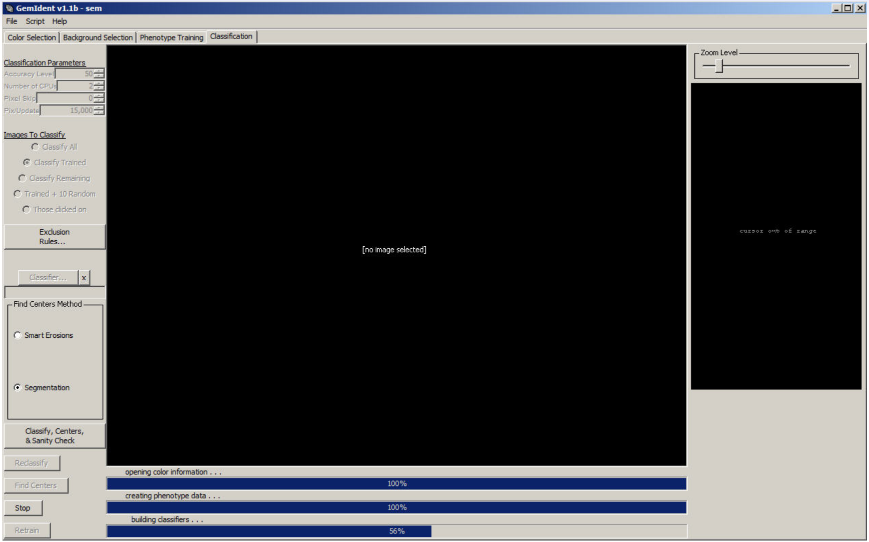
Color Selection



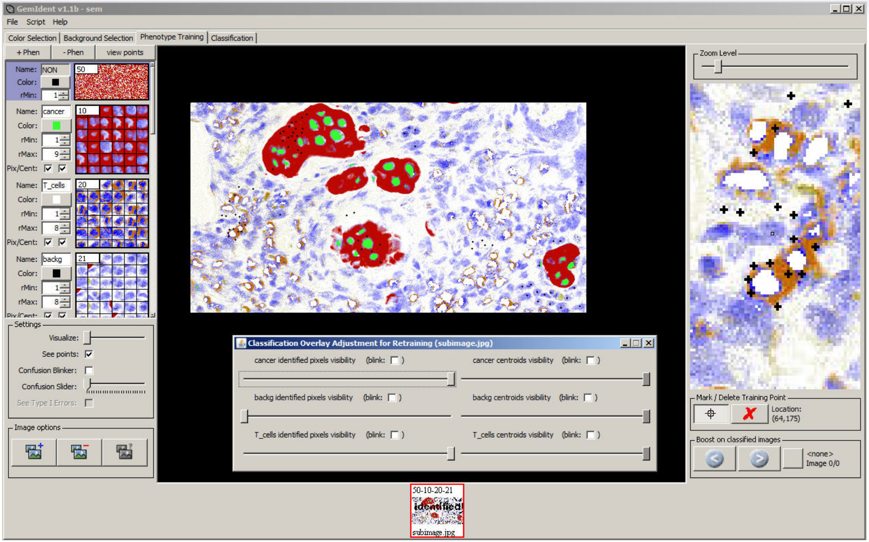
Phenotype Training



Classification

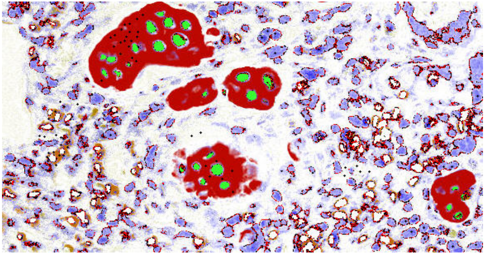


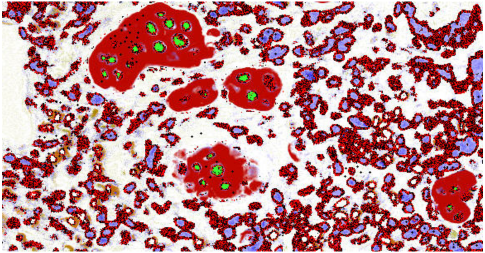
Phenotype Retraining



"Confusion" Measure

The ultimate in interactive boosting. . . when trees fight?





The Output

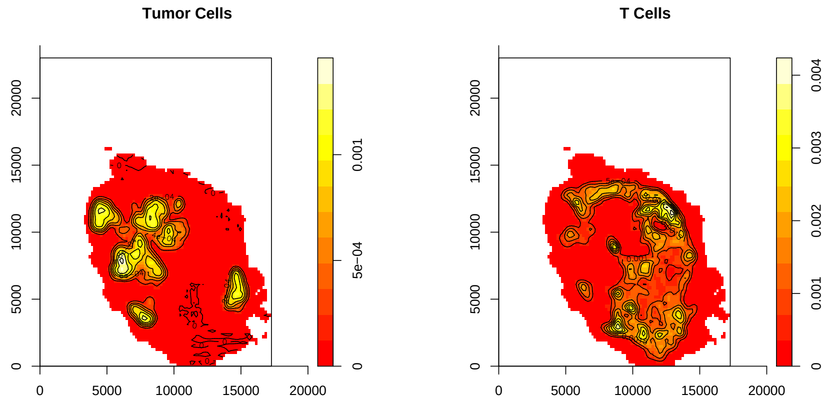
A big CSV file with every coordinate for every cell type across every subimage:

```
filename,locX,locY,globalX,globalY
stage_0147,63,49,6416,10263
stage_0147,92,122,6445,10336
stage_0147,95,243,6448,10457
stage_0147,109,243,6462,10457
stage_0147,247,719,6600,10933
stage_0147,276,757,6629,10971
stage_0147,389,830,6742,11044
stage_0147,397,394,6750,10608
stage_0147,407,820,6760,11034
```

There are 10^5 to 10^6 cells per lymphnode.

Spatial Analysis done in `R` with `spatstat`, `spdep`, `DCluster` packages.

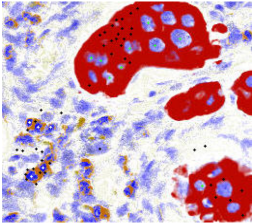
Density plots in R of a typical lymphnode



Supervised learning

Supervised learning is . . . supervised.

This means someone has to sit and click on examples of cells (as well as NON-phenotypes).



This is actually quite expensive. In fact, this is what users of GemIdent complain about the most.

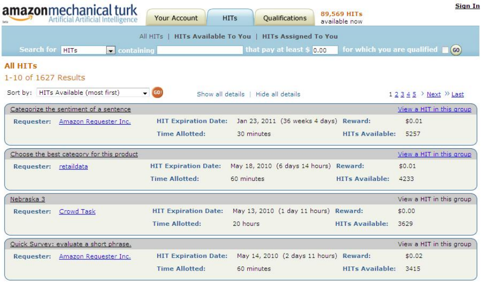
Imagine we have 100's of lymphnodes requiring careful training. Could there be an easier way?

Imagine a world. . .

where inside of a box, you can place any task, and have anyone in the world do it for a price you set. . .

MTurk

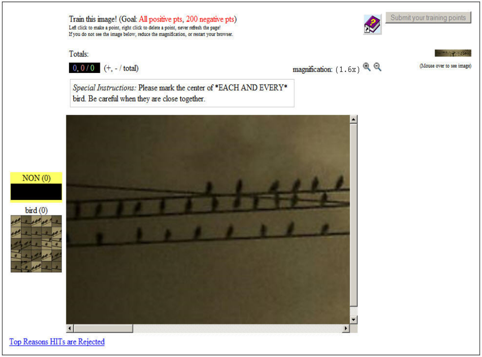
The most popular crowdsourcing marketplace is MTurk where you can freely post tasks.



Why not ask the world to click on cells for us for 10 cents an image?

DistributeEyes

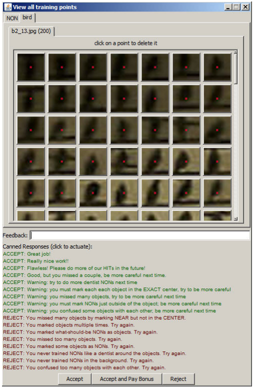
We created an add-on to GemIdent that takes images, and crowdsources the phenotype training step. Below is a DistributeEyes task that involves clicking on pigeons.



We can ensure the quality of the work by making workers watch an instructional video and passing a quiz.

Checking the training data

From within GemIdent, we can see the worker's points and choose to pay them or not. If they did a good job, we can even give them a bonus.



Verifying the usefulness

We ran an experiment with 600 image-labeling tasks (of a variety of images) to check accuracy and find patterns in workers that do the best job.

The hope

The results were very positive. Crowdsourcing can accelerate data collection.

We hope that this concept will be adopted by the medical research community as well as the statistical community because the price of supervised machine learning applications will drop.

References

L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: citeseer.ist.psu.edu/breiman01random.html

Ihaka, R., and R. Gentleman (1996): "R: A Language for Data Analysis and Graphics," Journal of Computational and Graphical Statistics, 5(3), 299–314.
cran.r-project.org/

S. Holmes ,A. Kapelner and P.P. Lee, An Interactive Java Statistical Image Segmentation System: GEMIDENT. Journal of Statistical Software, vol 30.

A. Kapelner, S. Holmes and P.P. Lee, "An Interactive Statistical Image Segmentation and Visualization System", medivis, pp. 81-86, International Conference on Medical Information Visualisation - BioMedical Visualisation (MediViz 2007), 2007.

H. E. Kohrt, N. Nouri, K. Nowels, D. Johnson, S. Holmes, and P. P. Lee, "Profile of immune cells in axillary lymph nodes predicts disease-free survival in breast cancer," PLoS Med, vol. 2(9), p. e284, 2005.

R. Tibshirani, T. Hastie and J. Friedman, The Elements of Statistical Learning. NY.: Springer, 2001.

Acknowledgements

Peter P. Lee
Holbrook Kohrt
Francesca Setiadi
Kyle Woodward

Funding from NIH/ NIGMS R01 and NSF-DMS.