Anthony S. Guerrero (SCPD n4361176)
CS229 Class Project
12/15/2005

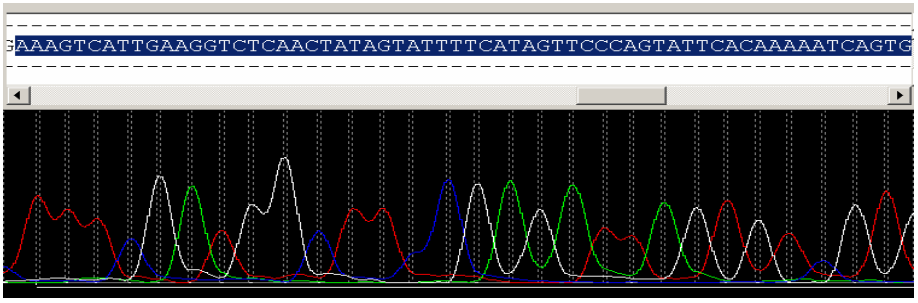
Identification of heterozygous point mutation events in DNA sequencing chromatograms.

Introduction.

The recent discovery of activating somatic mutations in cancer that correlate with phenotypes such as drug responsiveness, has generated renewed interest in the sequencing of genomes of tumor samples and cancer cancer cell lines with the goal of identifying the set of mutations that produce those phenotypes [1]. The two most popular strategies for discovering these events are array CGH [2], and direct sequencing of tumor samples and cells at specific loci of genes suspected a priori to be involved in tumor proliferation and survival. The latter technique involves using PCR amplification of the loci of interest and standard capillary electrophoresis DNA sequencing to generate chromatograms and sequences which are then compared to a reference normal sequence to reveal mutations. The detection of homozygous events is relatively straightforward, but the identification of heterozygous point events is problematic. The process of detecting heterozygous events involves detecting "peaks within peaks" of chromatogram waveforms and is plagued by a variety of artifacts in these signals which can potentially generate false positives. Proposed herein is a detection algorithm based on a classifier which distinguishes candidate "peaks within peaks" that are heterozygous point mutations from those that are false positives based on statistics about the candidate event and representation of these artifacts as interval-scale input variables to a machine learning algorithm.

Background.
Standard DNA sequencing technology using capillary eletrophoresis produces a chromatogram and a nucleotide sequence called by a base caller algorithm (Fig 1.).

Fig 1:



The output of the sequencer and basecaller is a binary file containing the four raw electrophoresis curves (one for each base G,C,A and T), a vector indicating the locations within the curves at which the bases were called (shown graphically in figure 1 as vertical dashed white lines), and a vector of the “call quality” [4] at each base (an integer from 0 to 100, equal to -10 * [log probability of error in the base call], according to the base caller algorithm).

Heterozygous point mutations manifest as peaks within peaks (Fig 2.) but are easily confused with a variety of artifacts including background noise (Fig 3.), cross-talk, dye-blobs (Fig 4.), end run noise (Fig 5.).

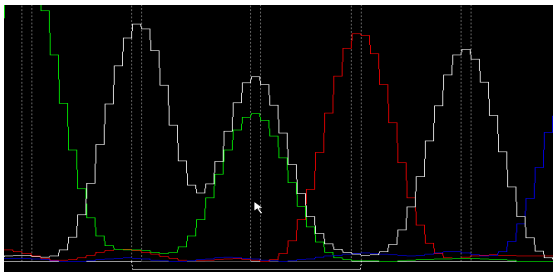


Fig 2. True Heterozygous Event.

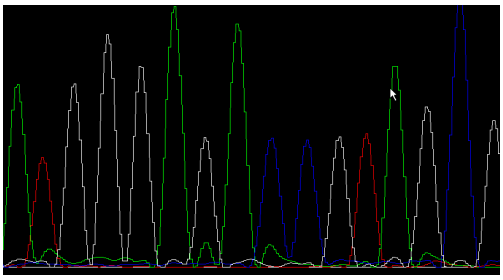


Fig 3. Background Noise, Non-events.

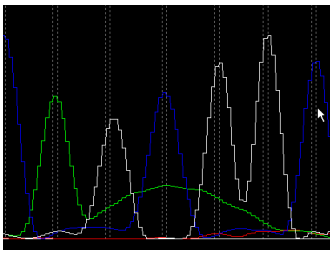


Fig 4. Dye Blob. Non-event.

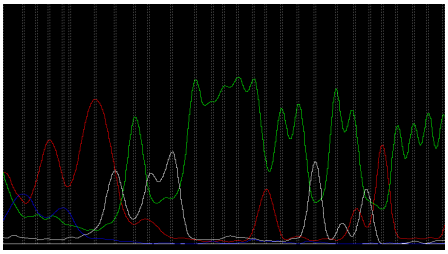


Fig 5. End Run Noise, Non-events.

Classification of Training Data:

One existing successful approach to identifying these events relies on the anti-correlation of the waveform of interest with respect to a waveform averaged from a pool of 'wild-type' or normal samples [3]. One embodiment of this approach is used by a program called "Mutation Surveyor™", the output of which was used to supply the training data for the classifier described here.

Attributes:

Proposed here is an algorithm which does not rely on such a pool of normal samples, and instead classifies events for a single chromatogram in isolation, solely by the local attributes of the chromatogram. The input to the classification algorithm is generated by locating in the chromatogram, in each channel, local-maxima that do not correspond to a called base at that position (non-call peaks). This produces four (possibly empty) sets of positions MA = {a1,..aA, A=number of non-call local maxima}, MG = {g1,..gG}, MC = {c1,..cC}, MT = {t1,..tT}. The peaks in these peak sets represent the candidates for classification. A classification of 'positive' constitutes the event detector.

The first and probably most important input variable for each peak is some measure of the magnitude of the local maxima with respect to its neighborhood. For each candidate peak, its full width half maximum, "FWHM" normalized by the base calling pitch and its height compared to the average call peaks of the same channel "peakratio" (Fig 6) in some neighborhood (+/- 20 bases) are calculated. Notice that the **FWHM** attribute was chosen as it appears that it may be useful for both rejecting noise peaks (with small width) and rejecting "dye-blobs" (with large peak widths).

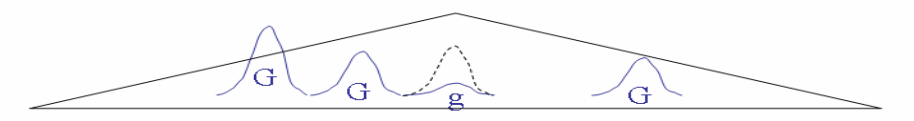


Fig 6. Peak Ratio = g/<G>

Also calculated for each peak is the proportional displacement of the peak with respect to its closest call peak normalized by the base calling pitch in the neighborhood, "posdisp" (Fig 7.). This attribute should help to distinguish noise peaks as peaks with a larger **posdisp** are increasingly likely to be noise as they do not spatially coincide with the most proximal base.

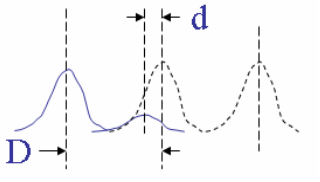


Fig 7. Proportional Displacement, 2d/D

For the purposes of eliminating potential false positives caused by regions of poor signal to noise ratio, the average local call quality in the neighborhood, not including the base over the peak of interest, is included as an attribute, "avequal", as is the specific call quality of the base over the peak of interest, "qual". We would expect the **avequal** attribute to be proportionally related to the likelihood of a positive (good surrounding base call quality implies that peak of interest is not likely to be noise). The specific call quality of the base over the peak of interest, **qual**, we would expect to relate inversely with the likelihood of a true positive (poor quality at the call over the peak implies a large significant underlying peak and likely positive).

Inspection of typical chromatograms demonstrates the ubiquitous presence of unusually high levels of noise at both the beginning and end of the waveform, referred to as 'end run noise' above. In anticipation of the fact that **avequal**, which will sample the left and right neighborhood of a peak, might not be an accurate estimate of whether or not the peak of interest is near the extremes of the waveform, two additional attributes, "maxqual" and "minqual", which represent the maximum and minimum of the two averages produced by only sampling from right and left of the peak. The attribute **minqual** might be expected to serve as a good indicator of whether a particular candidate peak is near the end of the waveform as this would produce a low left or right flanking average base call quality.

Finally, a measure of how the peak in question compares to the distribution of other background peaks in the same channel (which should be mostly noise) is calculated the Z-score of the peak in question with respect to the average and standard deviation of all the background peaks in the same channel. This attribute is called "lzscore".

In summary, the attributes are thus: **peakratio, lzscore, FWHM, posdisp, qual, avequal, maxqual and minqual**.

Data Collection:
A program was written to parse chromatograms to yield the attributes mentioned above. The set of attributes were collected for 255 heterozygous point mutations detected in 202 chromatograms by Mutation Surveyor™. These 256 data points represent the 'positives' of the training data set. The set of 'negatives' were gathered by processing 11 chromatograms classified by Mutation Surveyor™ as 'wild-type' (meaning they should have no events) and collecting attributes for all non-call peaks in the 11 wild-type chromatograms. This yielded 21,101 'negative' data points as the number of non-event background peaks vastly outnumbers the number of heterozygous mutation event peaks.

Training:
Visual inspection of the dimension pair-wise plot of a hypercube including a 10% margin around the hypercube containing all the 'positive' data (Fig. 8) shows that the data is not trivially separable by any pair of dimensions, although peakratio, "lzscore" and the neighborhood quality metrics show promise. A linear and radial kernel SVM [6], logistic regression and a regression tree model (pruned by 10 fold cross validation and a "1-SE" rule for the optimal tree [5]) were trained on randomly sampled subsets of the training data of increasing size for ten iterations. Their total error, false positive and false negative rates averaged for the ten iterations on each training data set size were tabulated (Fig. 9).

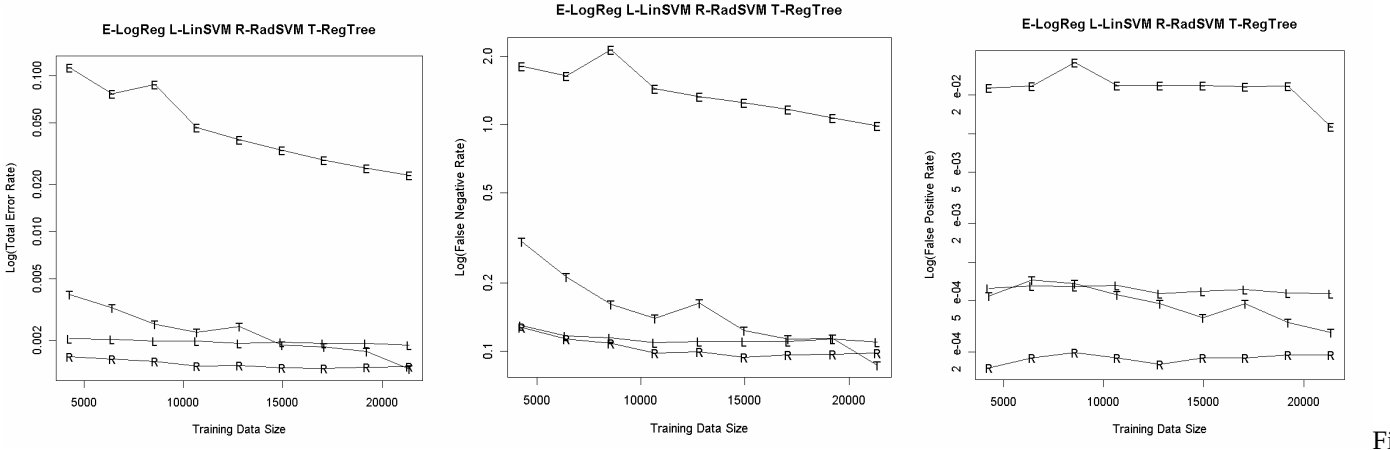


Fig 9

The SVM and logistic regression models seem to not improve substantially with larger training data set sizes, suggesting an inherent bias problem and possible poor lack of fit to the classification problem. The regression tree model seemed to improve steadily on all error rates, suggesting that the model may converge further for larger training data sets and seems to converge to a lower general error rate and, more importantly, a lower false negative rate when trained on all the data. One might imagine a radial kernel SVM to be capable of recapitulating quite accurately the decision boundary of a regression tree, so the fact that the regression tree exceeds the performance of the radial kernel SVM for a large enough data set size is a bit surprising. This is suggestive of some possible 'sharp angles' or high-dimensional corners in the optimal decision boundary. These could possibly be mirroring fixed thresholds inherent in the Mutation Surveyor™ software used to supply the training data.

The optimal pruned tree (Fig. 10) shows that a majority of the positives classify by the lzscore, minqual and peakratio and a majority of the negatives classify by lzscore and peakratio.

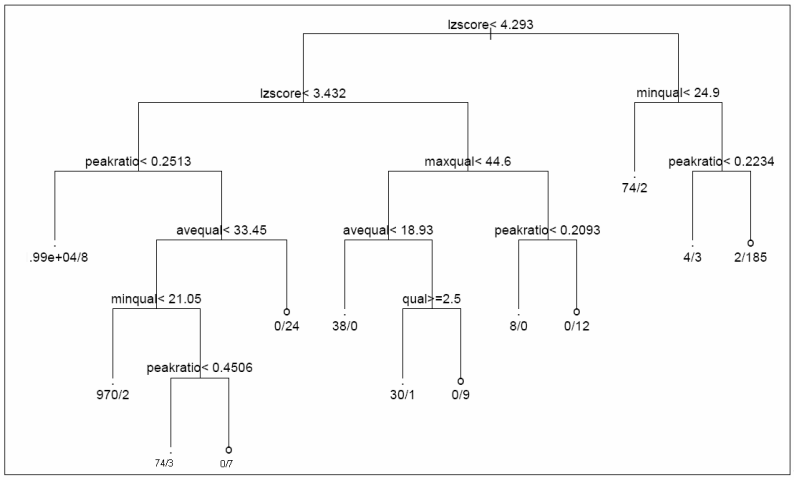


Fig 10

Improvements:
Manual inspection of some of the most frequently misclassified true positives shows that even when the background peak might be weak, often the foreground peak is measurably reduced in amplitude in the presence of a heterozygous event. This suggests that one might add an additional parameter to measure how attenuated the foreground peak is compared to the neighboring call-peaks of the same channel. Usually, although not always, chromatogram readouts come in pairs, one chromatogram produced from a PCR reaction on the forward strand of DNA and one independently produced from a PCR reaction on the reverse strand. Exploiting the corroboration expected between forward and reverse strand readouts would allow one to train a machine learning algorithm with an asymmetric loss structure favoring false positives over false negatives and use corroboration to further eliminate false positives.

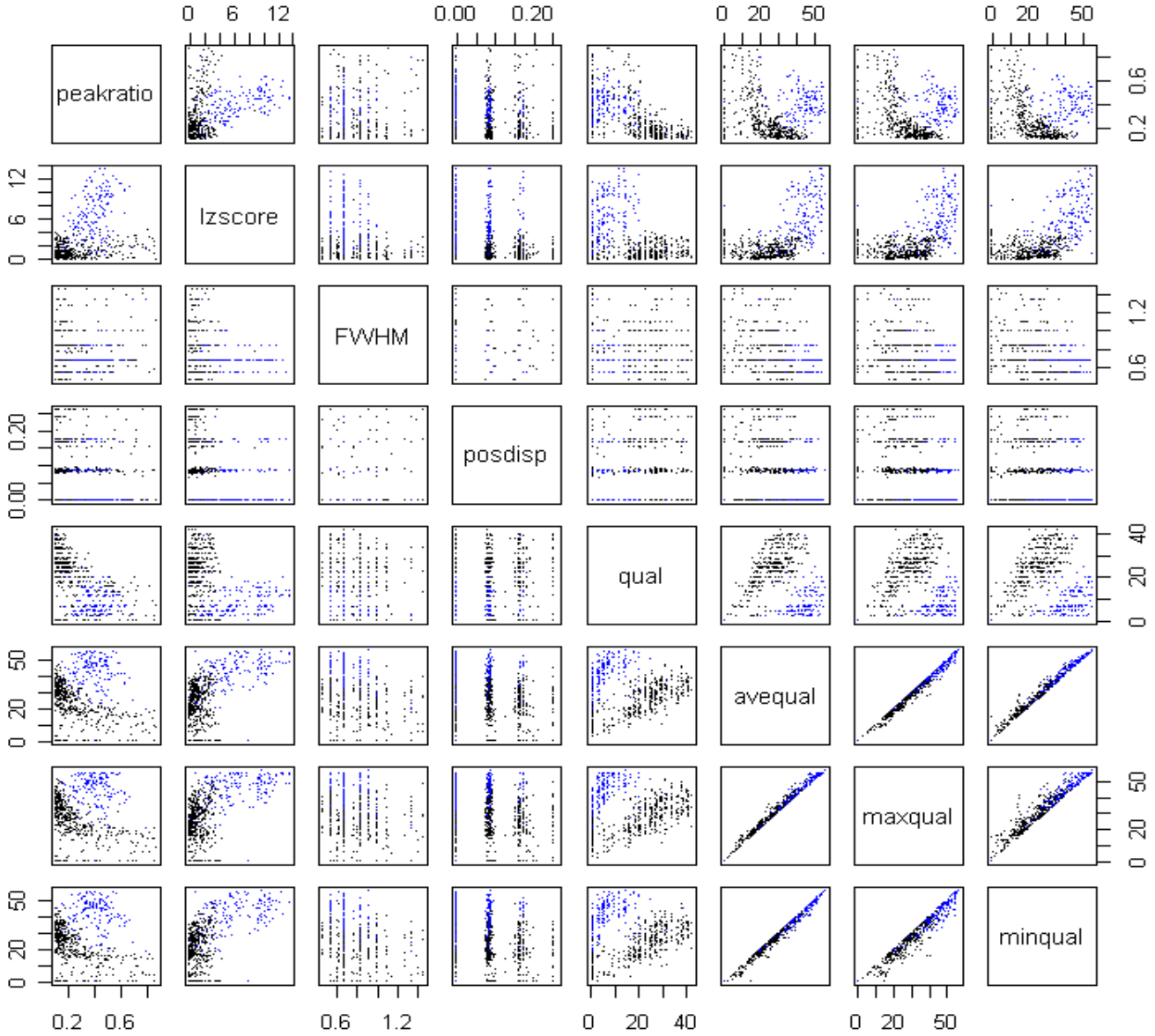


Figure 8, Hypercube containing all positives plus 10% margin. Positives are blue.

References:
[1] "Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib", T. J. Lynch et al., N. Engl. J. Med. 350, 2129 (2004).
[2] Nakao, K., Mehta, K.R., Fridlyand, J., Moore, D.H., Jain, A.N., Lafuente, A., Wiencke, J.W., Terdiman, J.P. and Waldman, F.M. (2004) High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization, Carcinogenesis, 25, 1345-1357.
[3] http://www.softgenetics.com/ms/index.htm
[4] http://www.phrap.com/phred/
[5] Therneau TM, Atkinson EJ: An introduction to recursive partitioning using the RPART routines. Tech Rep Mayo Foundation 1997. p 13.
[6] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm