BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 10 2009, pages 1307–1313
doi:10.1093/bioinformatics/btn585

Data and text mining

Aneuploidy prediction and tumor classification with heterogeneous hidden conditional random fields

Zafer Barutcuoglu[1], Edoardo M. Airoldi[1,2], Vanessa Dumeaux[3], Robert E. Schapire[1] and Olga G. Troyanskaya[1,2,*]

[1]Department of Computer Science, Princeton University, 35 Olden Street, Princeton, NJ 08540, [2]Lewis-Sigler Institute for Integrative Genomics, Carl Icahn Laboratory, Princeton University, Princeton, NJ 08544, USA and [3]Institute of Community Medicine, Tromso University, Tromso, Norway

Received on July 29, 2008; revised and accepted on November 9, 2008
Advance Access publication December 3, 2008
Associate Editor: Jonathan Wren

ABSTRACT

Motivation: The heterogeneity of cancer cannot always be recognized by tumor morphology, but may be reflected by the underlying genetic aberrations. Array comparative genome hybridization (array-CGH) methods provide high-throughput data on genetic copy numbers, but determining the clinically relevant copy number changes remains a challenge. Conventional classification methods for linking recurrent alterations to clinical outcome ignore sequential correlations in selecting relevant features. Conversely, existing sequence classification methods can only model overall copy number instability, without regard to any particular position in the genome.

Results: Here, we present the heterogeneous hidden conditional random field, a new integrated array-CGH analysis method for jointly classifying tumors, inferring copy numbers and identifying clinically relevant positions in recurrent alteration regions. By capturing the sequentiality as well as the locality of changes, our integrated model provides better noise reduction, and achieves more relevant gene retrieval and more accurate classification than existing methods. We provide an efficient L_1-regularized discriminative training algorithm, which notably selects a small set of candidate genes most likely to be clinically relevant and driving the recurrent amplicons of importance. Our method thus provides unbiased starting points in deciding which genomic regions and which genes in particular to pursue for further examination. Our experiments on synthetic data and real genomic cancer prediction data show that our method is superior, both in prediction accuracy and relevant feature discovery, to existing methods. We also demonstrate that it can be used to generate novel biological hypotheses for breast cancer.

Contact: ogt@cs.princeton.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

One of the major challenges in the management of cancer is its heterogeneity: cancer patients with the same stage of disease can have markedly different treatment responses and survival outcomes. This heterogeneity cannot always be recognized by tumor morphology, but may reflect the complexity of underlying genetic aberrations.

Depending on the instability present in the tumor and the selection environment, tumor cells may acquire alterations, called *aneuploidies*, ranging from large segments with single copy number alterations to narrow homozygous deletions or high-level amplifications (Heim and Mitelman, 1989). Array comparative genomic hybridization (array-CGH) is a technique by which it is possible to detect and map genetic changes that involve gain or loss of segments of genomic DNA. Downstream analyses involve classifying the samples and finding copy number alterations that are associated with known biological markers. Finding regions of recurrent aneuploidy, called *amplicons*, for a tumor type can reveal candidate cancer genes that have undergone selection for altered expression associated with tumor growth (Albertson *et al.*, 2000; Brown *et al.*, 2006; Snijders *et al.*, 2005).

Although recent developments have enabled experiments to measure copy number on a genome scale with high genomic resolution, individual point measurements are still noisy, making the crucial separation of signal from noise difficult. A point deviation in array-CGH measurements can be due to a true difference in copy number, or a measurement artifact. A key factor for filtering out noise is to note the strong sequential correlation in copy numbers throughout the genome, and numerous methods have been successfully applied to sequentially detect regions of constant aneuploidy [see, Lai *et al.* (2005) for a survey].

Performing sequential aneuploidy detection on an individual genome, however, with no regard to recurrent patterns across different genomes, ignores correlations among similar tumor samples. In particular, if genomes in a sample set have been differentially labeled with a clinical target attribute (e.g. grade, subtype, recurrence, survival), then a *supervised* (label-aware) analysis can focus directly on the potentially clinically relevant patterns of aneuploidy, rather than relying solely on unsupervised sequential correlation. In addition to providing a direct predictive model for clinical diagnostic or prognostic applications, a supervised model can distinguish biomarker genes possibly relevant to tumor development from clinically irrelevant copy number changes.

Several studies have demonstrated the importance of supervised methods on CGH data for tumor classification, prognosis, and candidate gene search [see van Beers and Nederlof (2006) for a

*To whom correspondence should be addressed.

recent survey]. However, the all-purpose predictive models that have been used for analysis, such as naïve Bayes (Wessels *et al.*, 2002), support vector machines (SVMs) (Jonsson *et al.*, 2005) and various conventional statistics, all ignore the sequential information captured by unsupervised aneuploidy detection methods. This simplistic order-insensitive interpretation of array-CGH data is likely to cause the statistical bias of known correlations to be accounted for as variance, discarding clinically relevant signals as noise.

Only the recent hierarchical hidden Markov model (H-HMM) (Shah *et al.*, 2007) and fused SVM (Rapaport *et al.*, 2008) models demonstrate the benefits of supervised sequential array-CGH analysis over many tumor samples for identifying clinically important regions of aneuploidy, but identifying the causal genes within these amplicons remains an open challenge. Thus, no existing method can perform a supervised identification of the clinically relevant genes in the process of extracting copy number profiles for tumor classification.

In this work, we present a method that combines a sequential representation of copy numbers with outcome-related gene selection to build a supervised predictor for a clinical variable by selecting clinically relevant genes that 'drive' recurrent amplicons. Our method combines the sequential de-noising and the classification aspects in one integrated supervised architecture, so that they can cooperatively learn a better overall predictive model, without loss of relevant signal to either. We provide an efficient, regularized training algorithm that finds a sparse interpretable solution that directly identifies cancer-related genes. We extensively evaluate this method on both synthetic data and four biological datasets of breast, uveal melanoma and bladder tumors. We demonstrate that our method is substantially better than state-of-the-art methods and can be used to make new biological and clinically relevant hypotheses.

2 METHODS

2.1 Probabilistic model

Our model explicitly represents the discrete copy number at a probe location as a latent random variable. Each array-CGH measurement is an observed variable sampled exclusively from its underlying copy number's measurement-level mean with a random noise distribution. The sequentiality of copy numbers is represented by pairwise correlations between adjacent latent variables.

The entire sequence has a clinical label to be predicted, which in our model is affected directly by the discrete copy number profile. The real-valued observations relate to the sequence label only through the latent copy number's variables, making the sequence label conditionally independent of the observed measurements given the copy numbers. This decoupling reflects our explicit modeling of the observations as noisy representations of copy number levels: if we already knew the true copy numbers, the noisy observations would no longer be relevant to the prediction label. The model is illustrated in Figure 1.

Furthermore, we assume the sequence label to be directly affected by only a small subset of positions in the copy number profile. Part of the learning process is the selection of these positions, the cancer-related loci, by applying a sparse regularization on the $c_i - s$ edges in Figure 1.

The method first learns the model's parameters on a training dataset of array-CGH sequences with known sequence labels. A regularization parameter determines how many cancer-related positions are selected. Once the model is built, it can be used to predict the most likely sequence labels for new sequences. Discrete copy number profiles can also be queried as the most likely assignments of the latent copy number variables given observed data. For evaluations, a cross-validation or held-out samples protocol is used.

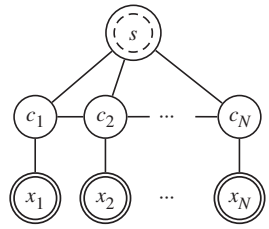

Fig. 1. The heterogeneous hidden conditional random field (HHCRF) model. The variables x_i are observed, c_i are hidden and the sequence label s is only observed during training. An exponential model for $p(s,\mathbf{c}|\mathbf{x})$ is tuned to maximize the class-conditional likelihood $p(s|\mathbf{x})$ of training data.

For a particular training example, let s be the clinical label of the whole sequence, let x_i denote the observation and c_i the latent variable at position $i \in \{1,\ldots,N\}$ whose value can be one of C different copy number states.

Given the observations $\mathbf{x}$ for an example, we use an exponential model for the conditional probability of the other variables:

$$p_{\boldsymbol{\theta}}(s,\mathbf{c}|\mathbf{x}) = \frac{1}{Z_{\boldsymbol{\theta}}(\mathbf{x})} \exp\left(\boldsymbol{\theta}\cdot\mathbf{f}(s,\mathbf{c},\mathbf{x})\right) \tag{1}$$

where $Z_{\boldsymbol{\theta}}(\mathbf{x})$ is a normalization factor, $\boldsymbol{\theta} = (\boldsymbol{\rho},\boldsymbol{\lambda},\boldsymbol{\omega})$ are the model parameters and $\mathbf{f}$ is a vector of features. In principle, the features could be any relevant real-valued functions of s, $\mathbf{c}$ and $\mathbf{x}$, but in our model, we consider features of only three types corresponding to the three edge types in Figure 1. Thus,

$$\begin{aligned}\boldsymbol{\theta}\cdot\mathbf{f}(s,\mathbf{c},\mathbf{x}) = \boldsymbol{\rho}\cdot\sum_{i=2}^{N}\mathbf{f}_{\text{pair}}(c_{i-1},c_i,s)\\ +\sum_{i=1}^{N}\boldsymbol{\lambda}_i\cdot\mathbf{f}_{\text{local}}(c_i,s)\\ +\boldsymbol{\omega}\cdot\sum_{i=1}^{N}\mathbf{f}_{\text{obs}}(c_i,x_i).\end{aligned} \tag{2}$$

The pairwise features $\mathbf{f}_{\text{pair}}$ and the corresponding parameters $\boldsymbol{\rho}$ model the sequence-wide correlation of adjacent nodes for each class. The local features $\mathbf{f}_{\text{local}}$ and their parameters $\boldsymbol{\lambda}_i$ model the correlation of latent variable c_i and the label s. And the observation features $\mathbf{f}_{\text{obs}}$ and their parameters $\boldsymbol{\omega}$ model the correlation of latent variable c_i and its noisy observation x_i.

For discrete latent variables and class label, the feature functions $\mathbf{f}_{\text{pair}}$ and $\mathbf{f}_{\text{local}}$ are typically defined to be 1 for a particular combination of arguments and 0 otherwise. The pairwise parameters $\boldsymbol{\rho}$ then correspond to (unnormalized) log-probabilities of a homogeneous hidden Markov model's (HMM) hidden state transitions. For real-valued observations, $\mathbf{f}_{\text{obs}}(c,x)$ can be defined as $(1,x,x^2)$ if $c = c'$ (and 0 otherwise) for each latent variable value c', the sufficient statistics for Gaussian distributions.

The position-dependent local parameters, which make the model heterogeneous, allow the model to interpolate between a homogeneous sequence-wide hypothesis and one that ignores correlations. If all local parameters are made zero, the model is a fully homogeneous random field, and classification only depends on sequence-wide stability of latent state. Conversely, if they are unconstrained and allowed to overpower the pairwise component, classification will depend almost fully on them, and the model will be akin to logistic regression (LR). In our model, we constrain the L_1 norm of the local parameters $\boldsymbol{\lambda}$ to adjust this tradeoff, which also encourages sparsity and results in an interpretable solution.

2.2 Training

The model is trained discriminatively, minimizing the conditional negative log-likelihood of labels over the empirical distribution $\tilde{p}(\mathbf{x},s)$ of the training

data:

$$\mathcal{L}_{\boldsymbol{\theta}} = -\sum_{\mathbf{x},s} \tilde{p}(\mathbf{x},s)\log p_{\boldsymbol{\theta}}(s|\mathbf{x}) \tag{3}$$

subject to the regularization constraint $||\boldsymbol{\lambda}||_1 \leq \beta$.

We use a gradient-based procedure to solve the optimization problem. The partial derivative of the objective loss with respect to any parameter θ_k is:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \theta_k} &= \sum_s \tilde{p}(s) \sum_{\mathbf{c}} p_{\boldsymbol{\theta}}(s,c|\mathbf{x}) f_k - \sum_{\mathbf{x},s} \tilde{p}(\mathbf{x},s) \sum_{\mathbf{c}} p_{\boldsymbol{\theta}}(c|s,\mathbf{x}) f_k \\ &= \mathbb{E}_{\tilde{p}(s)p_{\boldsymbol{\theta}}(s,c|\mathbf{x})}[f_k] - \mathbb{E}_{\tilde{p}(\mathbf{x},s)p_{\boldsymbol{\theta}}(c|s,\mathbf{x})}[f_k]. \end{aligned} \tag{4}$$

Although $p_{\boldsymbol{\theta}}(s|\mathbf{x})$ in (3) and the expectations in (4) call for marginalizing $p_{\boldsymbol{\theta}}(s,\mathbf{c}|\mathbf{x})$ as defined in (1) over the exponentially many value combinations of the latent variables $\mathbf{c}$, a dynamic programming solution exists, similar to the forward–backward procedure for HMMs, scaling linearly with sequence length.

2.3 Gradient LASSO

To satisfy the regularization constraint $||\boldsymbol{\lambda}||_1 \leq \beta$, we incorporate the Gradient LASSO algorithm (Kim and Kim, 2004), with a minor modification.

Gradient LASSO is an interior point method for optimizing a differentiable function subject to L_1 constraints. It maintains an explicitly sparse current solution, alternating between a coordinate-wise gradient step, which may add a new non-zero parameter, and a multivariate gradient step over the non-zero parameters, which may make one of them zero. The constraints are always kept satisfied, by starting inside the constraint simplex and bounding step sizes. When the current parameters satisfy the constraint by equality and local gradient descent is about to violate it, the gradient is projected onto the boundary, and linearity of L_1 constraint boundaries makes line search along the boundary possible.

Our version of Gradient LASSO (summarized in Algorithm 1) differs slightly from the original presented by Kim and Kim (2004): in the deletion step, if the current solution is not on the constraint boundary, we use a less conservative maximum step size Δ to accelerate learning, without affecting the final solution of the algorithm.

2.4 Unconstrained parameters

The unregularized parameters of our model $(\boldsymbol{\rho},\boldsymbol{\omega})$ are optimized after each two-step Gradient LASSO iteration, using the gradient-based L-BFGS algorithm (Nocedal, 1980), a limited-memory quasi-Newton method for unconstrained optimization, while the regularized parameters $\boldsymbol{\lambda}$ are kept unchanged.

Note that the unconstrained optimization step causes the constrained problem objective $L(\boldsymbol{\lambda})$ to change between iterations, and therefore the optimality of its current solution. The two-step Gradient LASSO algorithm, by adding newly relevant features and deleting obsolete features as necessary, is able to robustly cope with this concept drift without compromising sparsity, which would not have been possible with strictly growing or shrinking algorithms.

In our implementation, we constrained $\boldsymbol{\rho}$ to be diagonal to reduce model complexity, and used k-means clustering to initialize $\boldsymbol{\omega}$ to good starting values for faster convergence. Since our model is in exponential form, Gaussian parameters found by clustering observations can be multiplied out from $\exp\left[-(x-\mu)^2/(2\sigma^2)\right]$ to get feature coefficients for the form $\exp(\omega_1+\omega_1 x+\omega_1 x^2)$ for each copy number state.

2.5 Evaluation with synthetic data

To assess the performance of our method under different controlled conditions, we created synthetic datasets reflecting key properties of array-CGH microarrays using the following process.

In accordance with laboratory evidence suggesting that amplicons are selected based on certain underlying driver genes (Albertson, 2006), five 'oncogene' positions were randomly chosen for each dataset of fixed sequence length N. Then, amplicons of width $\sim\mathcal{N}(15,5)$ and uniform random offset were created to contain each oncogene position with probability $1-\varepsilon$ for positive examples and ε for negative examples (i.e. the inversion noise ε decreases the correlation of amplicon existence and positive label). Copy number levels were limited to normal (ratio $=1$) and amplified (ratio $=1.5$, reflecting tumor sample heterogeneity).

Realistic microarray measurement noise was then added, according to the exponential model proposed by Rocke and Durbin (2001) and using parameters estimated by Myers *et al.* (2004) from real human breast cancer array-CGH data. The 'clean' versions of all datasets, prior to microarray measurement noise addition, were also stored for comparison.

We generated 10 instances of 1000-sequence datasets for each combination of $N \in \{100, 1000\}$ and $\varepsilon \in \{0, 0.25\}$, with even positive/negative ratio. For each 1000-sequence instance, 50 examples were used for training and 950 for test.

Over the 10 instances for each setting, we ran our heterogeneous hidden conditional random fields model with $C=2$ states ('normal' and 'amplified') and $\beta \in \{5, 10, 20\}$ for 100 iterations, and compared it to a purely non-sequential LR model tuned by gradient descent with learning rate 0.1 and momentum 0.5 over 100 iterations.

As a sparsely regularized model for comparison, we used L_p-regularized LR(L_pLR) (Liu *et al.*, 2007) whose effectiveness has been demonstrated on expression microarray data. We used the parameters $p=0.1$ and $\gamma = 10^{-4}$ as suggested (though we did try other combinations with less success), and regularization weight $\beta \in \{0.1, 0.3, 0.5, 1, 3, 5\}$, gradient-optimized with learning rate 0.1 and momentum 0.5 over 500 iterations.

In addition, as a means of taking sequential correlations into account for noise reduction, we also ran L_pLR after preprocessing the data with a moving average of window size 50 ($L_p\text{LR}_{w50}$).

Algorithm 1 Gradient LASSO (modified)

Objective: min $L(\boldsymbol{\lambda})$ s.t. $||\boldsymbol{\lambda}|| \leq \beta$
repeat
  Addition step:
  Compute gradient $\nabla = (\partial L/\partial\lambda_1, \ldots, \partial L/\partial\lambda_d)$
  Choose coordinate $k = \arg\max_i |\nabla_i|$
  $h_k = -\beta\,\text{sign}(\nabla_k)$; $h_i = 0$ for all $i \neq k$
  $\hat{\alpha} = \arg\min_{\alpha\in[0,1]} L((1-\alpha)\boldsymbol{\lambda} + \alpha\mathbf{h})$
  $\boldsymbol{\lambda} \leftarrow (1-\hat{\alpha})\boldsymbol{\lambda} + \hat{\alpha}\mathbf{h}$

  Deletion step:
  Compute gradient $\nabla = (\partial L/\partial\lambda_1, \ldots, \partial L/\partial\lambda_d)$
  Let $\boldsymbol{\sigma} = \{i : \lambda_i \neq 0\}$
  Let $p = \nabla \cdot \mathbf{z}$ where $z_i = \text{sign}(\lambda_i)$

$$h_j = \begin{cases} 0 & \text{if } j \notin \boldsymbol{\sigma} \\ -\nabla_j + pz_j/|\lambda_j| & \text{if } j \in \boldsymbol{\sigma}, p < 0 \text{ and } ||\boldsymbol{\lambda}||_1 = \beta \\ -\nabla_j & \text{if } j \in \boldsymbol{\sigma}, \text{ otherwise} \end{cases}$$

$$\Delta = \begin{cases} \min_{j\in\boldsymbol{\sigma}}\{-\lambda_j/h_j : \lambda_j h_j < 0\} & \text{if } ||\boldsymbol{\lambda}||_1 = \beta \\ (\beta - ||\boldsymbol{\lambda}||_1)/||\mathbf{h}||_1 & \text{if } ||\boldsymbol{\lambda}||_1 < \beta \end{cases}$$

  $\hat{\alpha} = \arg\min_{\alpha\in[0,\Delta]} L(\boldsymbol{\lambda} + \alpha\mathbf{h})$
  $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \hat{\alpha}\mathbf{h}$
until converged

3 EXPERIMENTAL RESULTS

We evaluate our method on a range of synthetic datasets modeled after real cancer microarrays, and then on four biological datasets of breast, uveal melanoma and bladder tumors. The results demonstrate that our method performs substantially better than state-of-the-art classification methods, and is able to make new clinically relevant predictions for key amplicons and candidate marker genes.

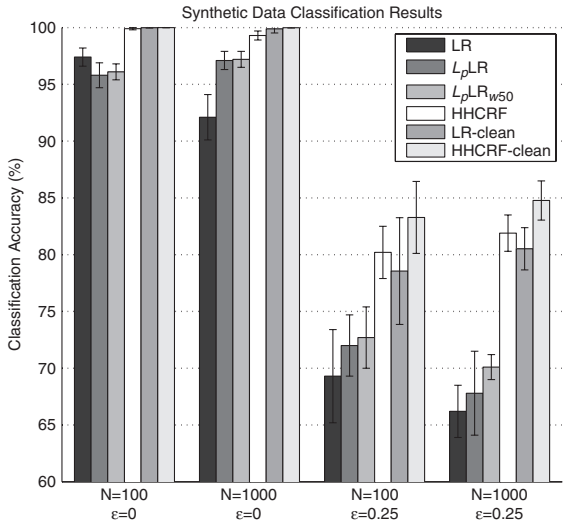

Fig. 2. Synthetic data classification accuracies (with SD error bars) for sequence length N and inversion noise ε over 10 instances of 50-training/950-test-example runs for the best cross-validated parameter settings of each model, for LR, with L_p regularization (L_pLR), preprocessed with a moving average of sequence window size 50 (L_pLR$_{w50}$), HHCRF, and results on data without simulated microarray measurement noise (LR-clean, HHCRF-clean).

3.1 Synthetic data

3.1.1 Classification In synthetic experiments with data generated to resemble real microarray data (Rocke and Durbin, 2001, Section 2), HHCRF consistently achieved significantly (by Student's paired t-test with $p < 10^{-4}$; i.e. confidence $> 99.99\%$) higher classification accuracy as compared with LR, L_p-regularized LR (L_pLR), and L_p-regularized LR preprocessed with a moving average of window size 50 (L_pLR$_{w50}$) (Fig. 2).

We also ran LR and HHCRF on the 'clean' versions (without microarray measurement noise) of the datasets (the other models were omitted since sparsity and smoothing became irrelevant in the absence of noise variance). Still, HHCRF performed better than LR, especially for the datasets with inversion noise ($\varepsilon = 0.25$), which suggests that HHCRF's pairwise parameters $\boldsymbol{\rho}$ capture sequence-wide stability properties and contribute to the classification task beyond simply filtering out observation noise.

Indeed, HHCRF accuracy on *noisy* data is comparable to the 'clean' data accuracy of LR, and indeed significantly better on the more difficult $\varepsilon = 0.25$ datasets (with 96% confidence for $N = 1000, \varepsilon = 0.25$), demonstrating the extent to which HHCRF is able to cope with experimental microarray noise.

3.1.2 Copy number inference The integral copy numbers for the classified sequences are the by-product of our model's classification task, obtainable by an efficient Viterbi-like max-product algorithm. Having the true underlying copy number states (normal versus amplified) for the synthetic data, we compared the states inferred by HHCRF to the true values. Note that the other models in the comparison cannot infer actual copy numbers at all. Table 1 summarizes the recovery of the true amplification states over all genes of all test sequences, where true positives are amplified genes inferred as amplified, and false positives are unamplified genes inferred as amplified. The high recall [TP/(TP+FN)] and comparatively lower precision [TP/(TP+FP)] reveal a tendency to avoid false negatives, not surprising considering that the discriminative loss is incurred only through selected oncogenes (non-zero local parameters) which are much more likely to be amplified than other genes, making false negatives more costly than false positives. In this situation, suggesting the biologist a more extensive candidate list is important, as additional information, such as known oncogene status can be used to filter candidates. Thus, our algorithm is effective in suggesting potential causative gene hypotheses that the user can examine for biologically interesting possibilities to follow up on.

Table 1. Synthetic data amplification results

N	ε_{inv}	Accuracy	Precision	Recall
100	0	76.1	52.9	98.2
1000	0	90.3	25.9	96.2
100	0.25	84.3	67.9	87.3
1000	0.25	88.6	21.7	86.2

Synthetic data amplification discovery statistics for the HHCRF models in Figure 2 over all genes in all test examples. True positives are amplified genes that were correctly inferred as amplified, and false positives are unamplified genes inferred by the model as amplified.

3.1.3 Oncogene discovery Comparing the sparse set of 'predicted oncogenes' selected by the model to the underlying true oncogenes requires a soft measure of overlap, both in set membership and also in terms of gene similarity, because Gradient LASSO reports only one in a group of genes that are always amplified together. For this purpose, we define a *co-amplification matrix* between the predicted oncogenes (rows) and the true oncogenes (columns), with entries denoting the correlation coefficients of the two genes' true copy numbers over test data. In practice, this copy number correlation provides a useful post-processing step to retrieve other candidate genes highly co-amplified with those selected by the model.

We then define *co-precision* as the mean of row maximums (average co-amplification of a predicted oncogene with the closest true oncogene) and *co-recall* as the mean of column maximums (average co-amplification of a true oncogene with the closest predicted oncogene). Thus, a model that returns only some of the true oncogenes, but no false predictions, will have high co-precision and low co-recall. Conversely, if all true oncogenes are found, but with many other spurious predictions, then co-recall will be high, and co-precision low. As desired, these measures are not affected much if several highly co-amplified genes are returned for one true oncogene.

These statistics, along with their harmonic mean (*co-F-measure*), are shown for the HHCRF models on the synthetic datasets in Figure 3.

The high co-recall values demonstrate successful recovery of most true oncogenes, decreasing with sequence length and ε difficulty, while the co-precision values indicate that the numbers of spurious predicted oncogenes were limited.

Also observable in Figure 3 is the effect of the regularization weight β on model complexity, directly increasing the number of predicted oncogenes.

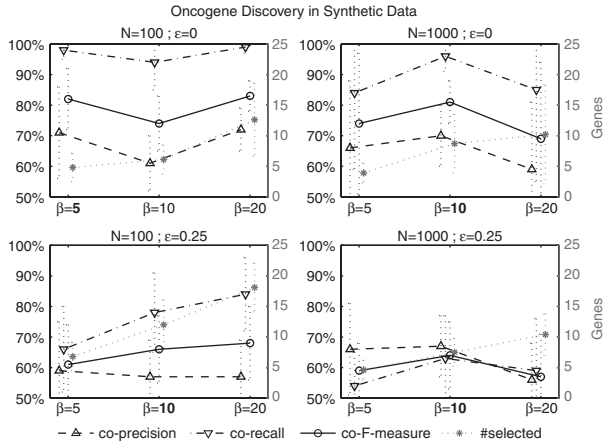

Fig. 3. Synthetic data oncogene discovery statistics (with SD error bars) for HHCRF with $\beta \in \{5, 10, 20\}$ over the 10 instances of each dataset. **Bold** labels indicate the β values with the highest classification accuracies in Figure 2. #selected (on the right y-axis) is the number of predicted oncogenes, compared with the five true oncogenes. Co-precision, co-recall and co-F-measure are percentages defined on the co-amplification matrix between the predicted and true oncogenes.

3.2 Breast cancer data

We applied our method to two breast cancer datasets for the task of identifying amplicons and potential causative genes predictive of high tumor grade. In both experiments, HHCRF successfully classified held-out examples significantly more accurately than a non-sequential SVM model, and made candidate gene predictions for relevance to tumor grade.

3.2.1 Pollack et al. (2002) breast tumor data On the 6691-gene human breast tumor array-CGH data from Pollack *et al.* (2002), we applied HHCRF with $C = 4$ copy number levels to classify tumors with histological grade 3-versus-all (17 positives out of 42). Over 5-fold cross-validation, held-out classification accuracies (mean±SD) for $\beta \in \{5, 10, 20\}$ were $76 \pm 7\%$, $67 \pm 10\%$, and $64 \pm 7\%$ respectively, compared with $60 \pm 20\%$ for a linear SVM. In addition to lower variance, HHCRF with $\beta = 5$ was statistically significantly more accurate (with 96% paired t-test confidence) than the SVM.

We then trained HHCRF with $\beta = 5$ on all 42 sequences, and examined the chosen genes. Table 2 shows the selected genes and their non-zero local weights. Among the selected genes, several have known connections to tumorigenesis. ARID1A has been identified as a presumptive tumor suppressor (Huang *et al.*, 2007), and VDUP1 is a known tumor suppressor (Han *et al.*, 2003). 'Homo sapiens clone 23596 mRNA sequence' has been observed to be highly expressed in breast cancer cell lines (Yi *et al.*, 2007), and downregulation of FLJ23403 (alias FAM38B) has been linked to human cancers (Beitzinger *et al.*, 2008).

Due to the non-grouping character of L_1 regularization, finding a relevant gene can suppress the subsequent detection of similar genes. In particular, Gradient LASSO picks only one gene out of a region that is always amplified together. To circumvent this effect, a correlation-based post-processing step can be applied after learning, to retrieve other relevant genes whose inferred copy numbers are highly correlated with the representative ones that were found by Gradient LASSO. For example, in Table 2, microarray feature 4505 does not match to a named gene, but its highest correlation (coefficient 0.69) in copy number is with the nearby feature 4515 (CUL4A), a known breast cancer-associated amplification (Nag *et al.*, 2004).

Table 2. Selected genes for Pollack *et al.* (2002) data

Index	Name	Weight	Evidence
98	ARID1A	$+0.61 \uparrow$	Huang *et al.* (2007)
353	VDUP1	$+1.30 \downarrow$	Han *et al.* (2003)
4505	co-amplified with CUL4A	$-0.69 \downarrow$	Nag *et al.* (2004)
5289	*H. sapiens* clone 23596	$+1.14 \uparrow$	Yi *et al.* (2007)
5634	FLJ23403	$-1.26 \downarrow$	Beitzinger *et al.* (2008)

Positive weights make a positive (high-grade) label more likely when amplified (↑) or deleted (↓), and negative weights make a negative label more likely. Microarray feature 4505 does not have a gene name, but it is highly co-amplified (corr.coeff. = 0.69) with nearby feature 4515 (CUL4A).

Table 3. Selected probes for Chin *et al.* (2006) data

Index	chr	Clone name	Weight
262	3	RP11-129P2	$-1.19 \uparrow$
566	5	CTD-2004C12	$+1.68 \downarrow$
657	6	RP11-47E20	$-1.25 \downarrow$
883	8	RP11-116F9	$+1.34 \downarrow$
953	8	RP11-44N11	$+1.25 \uparrow$
1725	16	RP11-52E21	$-0.90 \downarrow$
1738	16	RP11-140K16	$-0.38 \downarrow$
1780	17	DMPC-HFF#1-61H8	$+0.09 \uparrow$
2078	22	RP1-238C15	$-1.33 \downarrow$
2086	22	RP11-35I10	$-0.59 \downarrow$

Positive weights make a positive (high-grade) label more likely when amplified (↑) or deleted (↓), and negative weights make a negative label more likely.

3.2.2 Chin et al. *(2006) breast tumor data* The human breast tumor array-CGH data from Chin *et al.* (2006) has measurements for 2149 probe positions, not mapping directly to individual genes. Again, we ran HHCRF experiments with $C = 4$ for grade 3-versus-all (69 positives out of 141). The 5-fold cross-validation classification accuracies for $\beta \in \{5, 10, 20\}$ were $70 \pm 12\%$, $71 \pm 12\%$, and $67 \pm 07\%$, respectively, compared with $68 \pm 10\%$ for a linear SVM. HHCRF with $\beta = 10$ was more accurate than the SVM with 83% paired t-test confidence. As before, we then trained HHCRF with $\beta = 10$ on all 141 sequences for novel prediction, and Table 3 shows the selected probes.

Figure 4 shows part of a copy number profile extracted for high-grade breast tumor sequence `b0499`. In addition to determining the amplified and deleted regions, our model selected position 953 as a clinically relevant locus in determining tumor grade, predicted to correspond to the 'driver' gene for the 942..975 amplicon.

3.3 Institut Curie melanoma and bladder data

We also obtained successful results by applying our model on uveal melanoma and bladder tumor data from Institut Curie, used in the evaluation of the fused SVM algorithm in Rapaport *et al.* (2008).

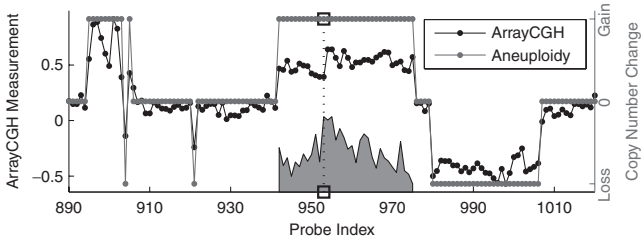

Fig. 4. Aneuploidies detected in a high-grade breast tumor from Chin *et al.* (2006). Our method detects the amplified and deleted regions, and also pinpoints probe 953 (Open Square) in the 942–975 amplicon as one of the 10 clinically important positions selected for relevance to high tumor grade, by analyzing across all tumor profiles in the dataset. The shaded area shows copy number correlations with the selected probe.

HHCRF classification performance exceeded (uveal melanoma tumors) or was comparable to (bladder tumors) that of fused SVM in these results. HHCRF produces a more interpretable model, outputting a specific set of outcome-related 'amplicon-driving' genes.

It should be noted that fused SVM does not limit the amplitudes of altered regions to a shared set of copy number levels; this may provide a better fit in the presence of high variance in tumor heterogeneity across many samples. Alternatively, if the effects of tumor heterogeneity and normal cell contamination have already been normalized out by the increasingly popular flow cytometric sorting techniques, HHCRF assumptions will hold stronger. In practice, model selection should ultimately be guided by application objectives and the particular data at hand.

3.3.1 Uveal melanoma tumors The uveal melanoma tumor dataset has array-CGH profiles with 3649 probes on non-sex chromosomes. Classifying by whether liver metastasis occurred within 24 months versus not (35 positives out of 78 tumors), HHCRF with $C=5$ states made a total of 10 test errors (87% accuracy) over 10 cross-validation folds for $\beta=5$, 11 errors (86% accuracy) for $\beta=10$, and 8 errors (90% accuracy) for $\beta=20$, compared with the best 10-fold cross-validation results from fused SVM at 17 errors (78% accuracy).

3.3.2 Bladder tumors The bladder carcinoma dataset contains array-CGH profiles with 2143 probes on non-sex chromosomes. On classification by tumor stage Ta-versus-T2+ (16 'stage Ta' positives out of 48 tumors with stage labels), HHCRF with $C=5$ states made a total of seven test errors (85% accuracy) over 10 cross-validation folds for $\beta=5$ and $\beta=10$, and 8 test errors (83% accuracy) for $\beta=20$. The best HHCRF error is on par with the best leave-one-out estimate of fused SVM (seven errors) reported in Rapaport *et al.* (2008).

Classifying by tumor grade 1-versus-higher (12 'grade 1' positives out of 57 tumors), HHCRF with $C=5$ states made a total of 10 test errors (82% accuracy) over 10 cross-validation folds for $\beta=5$, 9 errors (84% accuracy) for $\beta=10$, and 11 errors (81% accuracy) for $\beta=20$, compared with the best leave-one-out estimate reported by fused SVM (seven errors).

4 EXTENSIONS

The correlation-based post-processing step, retrieving similar genes from the selected oncogenes, can be necessary because of the L_1 loss minimized by Gradient LASSO: if two or more genes are equally important, picking only one of them is L_1-optimal for the algorithm. The desired grouping effect can be provided by a hybrid L_1+L_2 extension of Gradient LASSO, analogous to the Elastic Net (Zou and Hastie, 2005) extension of LASSO (Tibshirani, 1996), which will select all similarly important genes simultaneously due to the L_2 component.

Several other extensions are possible. If both array-CGH and expression microarray data are available for a dataset, HHCRF can use them together, by simply adding a new set of observed variables stemming from the same latent copy numbers. If information is available on the varying physical spacing of individual probes along the genome, it can be directly encoded into the pairwise features of HHCRF, as in the HMM model by Rueda and Diaz-Uriarte (2007). Although using a finite set of possible copy number levels may be sufficient in practice, incorporating *hierarchical Dirichlet processes* can allow copy numbers to grow arbitrarily (Teh *et al.*, 2006). Then array-CGH measurements can also be modeled to have an explicitly linear dependency on copy number, further reducing model complexity. Replacing maximum likelihood training with a Bayesian treatment, working with posterior distributions of model parameters (similar to Qi *et al.*, 2005) can reduce overfitting during training. Maximizing the classification margin (similar to Taskar *et al.*, 2004) may also improve generalization.

We also implemented a generative version of our model, explicitly assuming $p(x_i|c_i)$ to be Gaussian (as in Shah *et al.*, 2007) and modeling the joint probability $p_{\boldsymbol{\theta}}(s,\mathbf{c},\mathbf{x})$ to maximize the joint likelihood. The observation parameters, updated relatively slowly in the discriminative model, are tuned more directly by the joint gradient, and are expected to be less sensitive to initial values. However, the generative updates proved to be too aggressive in our experiments, overpowering the effect of the supervision label on loss. A discriminative training scheme, optimizing the conditional likelihood on the generative model, remains to be explored.

5 CONCLUSION

We presented the HHCRF, an array-CGH analysis method for jointly classifying tumors by clinical label, extracting copy number profiles, and identifying clinically relevant genes. We demonstrated its effectiveness on synthetic and real datasets, and described a generative variation and other extensions.

A particularly important feature of our method is to estimate the clinical significance of detected copy number changes. When the genome-wide profile is scanned for potentially new regions of interest, quantitative statistics about the aberrations are critical in order to decide which region to pursue for further examination. Our model highlights the most clinically relevant aneuploidy regions as those containing the predictive genes it has selected. The method also allows prioritization of genes harbored within the chromosomal regions of interest, starting with the explicitly selected genes and extending to others in similarity by co-amplification. In previous studies, prior biological knowledge was heavily used to infer causal genes in amplified regions, and thus, many known or putative oncogenes were credited as the driver genes, while some potentially

novel cancer-driving genes may have been overlooked. In contrast, in addition to detecting aneuploidies, our method explicitly identifies both amplicons and individual genes whose copy numbers are the most discriminative of the clinical label, suggesting specific targets for further biological investigation.

Funding: National Science Foundation (IIS-0513552); National Institute of Health (R01 GM071966); NSF CAREER award DBI-0546275; NIGMS Center of Excellence (P50 GM071508).

Conflict of Interest: none declared.

REFERENCES

Albertson,D. *et al*. (2000) Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nat. Genet.*, **25**, 144–146.

Albertson,D.G. (2006) Gene amplification in cancer. *Trends Genet.*, **22**, 447–455.

Beitzinger,M. *et al*. (2008) p73 poses a barrier to malignant transformation by limiting anchorage-independent growth. *EMBO J.*, **27**, 792–803.

Brown,L.A. *et al*. (2006) Amplification of EMSY, a novel oncogene on 11q13, in high grade ovarian surface epithelial carcinomas. *Gynecol. Oncol.*, **100**, 264–270.

Chin,K. *et al*. (2006) Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, **10**, 529–541.

Han,S.H. *et al*. (2003) VDUP1 upregulated by TGF-beta1 and 1,25-dihydorxyvitamin D3 inhibits tumor cell growth by blocking cell-cycle progression. *Oncogene*, **22**, 4035–4046.

Heim,S. and Mitelman,F. (1989) Primary chromosome abnormalities in human neoplasia. *Adv. Cancer Res.*, **52**, 1–43.

Huang, J. *et al*. (2007) Genomic and functional evidence for an ARID1A tumor suppressor role. *Genes Chromosomes Cancer*, **46**, 745–750.

Jonsson,G. *et al*. (2005) Distinct genomic profiles in hereditary breast tumors identified by array-based comparative genomic hybridization. *Cancer Res.*, **65**, 7612–7621.

Kim,Y. and Kim,J. (2004) Gradient LASSO for feature selection. In *ICML '04: Proceedings of the 21st International Conference on Machine Learning*, ACM, New York, NY, USA, p. 60.

Lai,W.R. *et al*. (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**, 3763–3770.

Liu,Z. *et al*. (2007) Sparse logistic regression with Lp penalty for biomarker identification. *Stat. Appl. Genet. Mol. Biol.*, **6**, Article 6.

Myers,C.L. *et al*. (2004) Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics*, **20**, 3533–3543.

Nag,A. *et al*. (2004) Cul4A physically associates with MDM2 and participates in the proteolysis of p53. *Cancer Res.*, **64**, 8152–8155.

Nocedal, J. (1980) Updating quasi-Newton matrices with limited storage. *Math. Comput.*, **35**, 773–782.

Pollack, J. R. *et al*. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci. USA*, **99**, 12963–12968.

Qi,Y. *et al*. (2005) Bayesian conditional random fields. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, Jan 6–8, 2005*, Cowell,R.G. and Ghahramani,Z. ed., Savannah Hotel, Barbados, pp. 269–276.

Rapaport,F. *et al*. (2008) Classification of arrayCGH data using a fused SVM. *Bioinformatics*. **24**, i375–i382.

Rocke,D.M. and Durbin,B. (2001) A model for measurement error for gene expression arrays. *J. Comput. Biol.*, **8**, 557–569.

Rueda,O.M. and Diaz-Uriarte,R. (2007) Flexible and accurate detection of genomic copy-number changes from aCGH. *PLoS. Comput. Biol.*, **3**, e122.

Shah,S.P. *et al*. (2007) Modeling recurrent DNA copy number alterations in array CGH data. *Bioinformatics*, **23**, i450–i458.

Snijders,A.M. *et al*. (2005) Rare amplicons implicate frequent deregulation of cell fate specification pathways in oral squamous cell carcinoma. *Oncogene*, **24**, 4232–4242.

Taskar,B. *et al*. (2004) Max-margin Markov networks. *Adv. Neu. Infor. Proc. Sys.*. **16**, 51.

Teh,Y.W. *et al*. (2006) Hierarchical Dirichlet processes. *J. Am. Stat. Assoc.*, **101**, 1566–1581.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, **58**, 267–288.

van Beers,E.H. and Nederlof,P.M. (2006) Array-CGH and breast cancer. *Breast Cancer Res.*, **8**, 210.

Wessels,L.F.A. *et al*. (2002) Molecular classification of breast carcinomas by comparative genomic hybridization: a specific somatic genetic profile for BRCA1 tumors. *Cancer Res.*, **62**, 7110–7117.

Yi,C.-H. *et al*. (2007) Loss of fibulin-2 expression is associated with breast cancer progression. *Am. J. Pathol.*, **170**, 1535–1545.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B*, **67**, 301–320.