Using Genetic Programming to Classify Node Positive Patients in Bladder Cancer

Arpit A. Almal, MSc
Genetics Squared Inc
210 S. Fifth Ave
Suite A
01-734-929-9475
aalmal@genetics2.com

Anirban P. Mitra, MBBS
University of Southern California
Department of Pathology 2011 Zonal Ave HMR 312
Los Angeles, CA 90033
01-323-442-3477
amitra@usc.edu

Ram H. Datar, PhD
University of Southern California
Department of Pathology 2011 Zonal Ave HMR 312
Los Angeles, CA 90033
01-323-442-3477
datar@usc.edu

Peter F. Lenehan, MD, PhD
Genetics Squared Inc
210 S. Fifth Ave
Suite A
01-734-929-9475
plenehan@genetics2.com

David W. Fry, PhD
Genetics Squared Inc
210 S. Fifth Ave
Suite A
01-734-929-9475
dfry@genetics2.com

Richard J. Cote, MD
University of Southern California
Department of Pathology 2011 Zonal Ave HMR 312
Los Angeles, CA 90033
01-323-442-3477
cote_r@ccnt.hsc.usc.edu

William P. Worzel, Dip. CS
Genetics Squared Inc
210 S. Fifth Ave, Suite A
01-734-929-9475
bworzel@genetics2.com

ABSTRACT

Nodal staging has been identified as an independent indicator of prognosis. Quantitative RT-PCR data was taken for 70 genes associated with bladder cancer and genetic programming was used to develop classification rules associated with nodal stages of bladder cancer. This study suggests involvement of several key genes for discriminating between samples with and without nodal metastasis.

Categories and Subject Descriptors

I.5.3 [**Pattern Recognition**]: Clustering – *Algorithms and similarity measures* I.5.2 [**Pattern Recognition**]: Design Methology – *Classifier design and evaluation, feature design and evauation, Pattern Analysis* I.2.6 [**Artificial Intelligence**]: Learning – *Concept Learning and Induction* I.2.2 [**Artificial Intelligence**]: Automatic Programming – *Program Synthesis*

General Terms: Algorithms, Design, Experimentation.

GECCO'06, July 8–12, 2006, Seattle, Washington, USA.
Copyright 2006 ACM 1-59593-186-4/06/0007…$5.00.

Keywords

Bladder cancer, Nodal staging, Genetic Programming, Classification, Feature selection, Machine Learning

1. INTRODUCTION

Cancer of the urinary bladder is a major epidemiological problem that continues to grow each year. Bladder cancers encompass urothelial carcinomas (UCs, or transitional cell carcinomas or TCCs), squamous cell carcinomas (SCCs), adenocarcinomas and certain other infrequent tumor types. It is the fourth most common malignancy in males and the ninth most common malignancy in females in the United States. An average of 260,000 new cases of urinary bladder cancer are diagnosed worldwide every year with an estimated 63,210 cases in 2005 in the United States alone (about 47,010 in males and 16,200 in females) and 13,180 deaths (about 8,970 in men and 4,210 in women). The incidence of bladder cancer between 1997 and 2001 was estimated at 39 per 100,000 for males and 10 per 100,000 for females [1].

2. MOTIVATION

The current treatment for UC is based on the pathologic staging of the tumor. The staging, therefore, is highly important for clinical decision-making and exploring the various treatment options and the therapy thus chosen can result in significant morbidity and financial burden to the patient The traditional TNM classification, based on the location, depth and the metastasis of the tumor [2], or the World Health Organization (WHO) classification system for UC [3] relies on pattern recognition and nomenclature for reporting bladder cancer biopsies, interpretation of which can be highly subjective and can have a high frequency of inter- and intra-observer variations. One of the major concerns in UCs is a large percentage of recurrence after operation. Interpretations of biopsies can also be confounded by sampling problems like absence of the muscular layer in the specimen or the exclusion of the bladder wall in biopsies of large tumors that are growing exophytically that can affect the staging of the slide. Even between highly trained pathologists, there are no accepted definitions for microinvasion which is an important criterion to determine the risk for metastasis. Most significantly, the basic tools available to determine tumor behavior, malignant potential and chance for recurrence provided by the current pathologic staging modalities can be highly subjective. One thus needs to realize that while current histopathologic criteria can provide us with important morphological information about tumors in patient populations, they are unable to specify the risk for progression or response to treatment for an individual patient with bladder cancer. Previous studies by Esrig *et al* have shown the wide difference in recurrence and survival rates between patients of the same pathological stage with differences in their tumor p53 status [4]. Nodal involvement is considered to be an independent risk factor for recurrence and survival after cystectomy for organ-confined bladder cancer [5]. It can be safely assumed that metastasis is a phenomenon that has its roots in gene expression as their expression levels change much before phenotypic changes are observed. Hence, if we could identify staging criteria that are constituted using gene expression values, they could be used as more consistent prediction measures and can also provide appropriate biological insights. This study points to the need to incorporate objective methods of staging using molecular markers specific to bladder cancer to complement the morphologic approach and pattern a refined system of staging that focuses on the biologic behavior of the tumor and its predicted clinical outcome, thereby equipping the clinician with a better insight on the appropriate treatment regimen to be instituted.

3. MATERIALS AND METHODS

3.1 Data Generation

Willey and colleagues have developed a modified quantitative method of standardized competitive reverse transcriptase PCR that allows simultaneous measurement of many genes using nanogram amounts of cDNA [6, 7]. The transcript levels are expressed as numerical values per million molecules of β-actin, a housekeeping gene chosen as a reference gene, thus affording intra- and inter-sample comparisons.

We have used this technique to obtain transcript profiles of 70 genes crucial in various cellular pathways that have been associated with tumor progression in various studies [8] (see Table 1).

3.2 Dataset & Study Design

The study involved data from 65 patients with 70 abovementioned genes being profiled for each one of them. We have divided the samples into a study and a validation set, wherein we learn from the data in the study set and test the robustness of our final solutions on the validation set. Nodal status for patients with UC was determined by histological examination of the lymph nodes obtained during bilateral pelvic lymphadenectomy during radical cystectomy.

Table 1: Genes Used In Study

Class	Genes
Angiogenesis	FGF5, FGFR4, VEGF, KDR
Anti-oxidation	GSTM3, GSTP1, GSTT1, SOD1
Apoptosis	ANXA5, BAD, BCL2, BCL2L1, CYPIA2, DAP, PTGS2, TGFBR2, TGIF, TNF, TNFAIP1, TNFSF10, TNFRSF1A, TRAF4
Apoptosis/Cell Cycle	CDKN1A, CDKN1B, CDKN2A, CDKN2C, GAPDH, RB1, RBL2, MXD1, TP53
Apoptosis/Cell Cycle/Gene Regulation	E2F1, E2F2, E2F4, E2F5
Cell Cycle	CCNA2, CCND3, CCNE1, CCNG1, CDC25C, CDC2, CDK7, CDK8, PCNA
Cell Growth Regulation	IGF1, IGF2R, PDGFB, PDGFRL
Invasion	CDH3, ICAM1, MMP16, TIMP2, BMP6
Signal Transduction	MAP2K6, MAPK12, MAPK9, MAPK8, STAT3, LYN, ERBB2
Gene Regulation	FOS, FOSL1, NFKB1, SP1, HSF1, MAP3K14, JUN, JUNB, MAX, MYC

The normal controls were cases that underwent radical prostatectomy without any lymph node dissection. For the purposes of this study we are classifying the normal samples as the node negative cases. The study set was composed of 11 Node positive (NP) cases and 23 Node negative (NN) cases, while the validation seet consisted of 10 NP cases and 21 NN cases. The sets were composed in such a way that the proportion of samples of each tumor stage remains the same in both the study and the validation set.

Using genetic programming on the study set, rules were developed for characterizing each stage molecularly using quantitative RT-PCR data. These rules were combined in a voting algorithm that was tested against a validation set. A subsequent analysis of the rules thus generated suggested some interesting features about tumor progression.

3.3 Developing Nodal Staging Rules

Genetic Programming (GP) [9, 10] is a method that lends itself naturally to the development of classifiers with the ability to automatically construct appropriate structure for the solution as well as select the variables. Importantly, the process does not require a large amount of the prior knowledge or effort in terms of structure selection or dimensionality reduction. There have been several initiatives where GP is used for analyzing medical/biological data [11,12] and for discrimination of cancers [13,14].

Overfitting is an important concern in any machine learning task, especially classification. Overfitting occurs when a classification algorithms draws strong inferences from the training samples and loses generality. In many biological or clinical datasets overfitting can be attributed to what is called the curse of dimensionality [15], i.e. when the number genes being studied are much larger than the number of samples in the training set. In this case, we have only 34 samples on the training set to learn from while there are more than 70 variables per sample, hence the danger of overfitting is a major concern in this study.

There are several approaches utilized to counter the overfitting problem – using simple rules, increasing the training samples, using a subset of test samples and integrating over different predictors. We have implemented all these methods in this study to prevent the GP system from overfitting the data, specifically because GP is considered to be a powerful search algorithm with a penchant for overfitting.

As mentioned earlier we have chosen a validation set from the samples in the study in order to identify how the predictors perform on the unseen data in order to gauge the amount of overfitting that has occurred. We were careful in maintaining the same proportion of samples with various attributes in the training and the validation sets For example, the proportion of the samples belonging to various T-stages for both the NP and NN cases were similar in the training and the validation sets.

Restricting the complexity of the results is a measure that guarantees simpler rules in order to force the system to pick only those features of the system which contribute the most value to the predictor. GP provides several intuitive and simple ways of restricting complexity such as only using simple mathematical functions as the constituents of the programs created. Thus in this problem only simple mathematical operators like +,-,* and /, logical operators like 'and', 'not' and 'or', comparison operators like =,>,<,>=,<= are used. The one exception to the use of simple operators in the set of functions listed in Table 2 is 'exp' - an exponent function where exp(N) is equivalent to e^N. This was included because of our experience with gene expression data where the GP system will in essence normalize the data as needed when comparing expression levels that vary exponentially when compared to one another. Another way that complexity is controlled is by limiting the size and complexity of the programs produced. Motivation for this is derived from the Minimum Description Length (MDL) Principle [16] of risk minimization wherein the least complex solution is called most robust. Restricting the number of genes being used in any solution can also be used to alleviate the problem of the overfitting. This can be shown to relate to the degrees of freedom in the expression that is loosely related to the VC dimension [17, 18]. The results have therefore been restricted to have no more than 5-6 genes in a classifier program.

In order to select a robust classifier it is imperative to know the generalization rate of the classifier, especially in the case of small number of samples in the data set, where overfitting to the data can be relatively frequent. Cross validation [19] is a simple yet effective technique to evaluate how well the classifier generalizes to unseen data. By taking multiple subsets (folds) of the training data and using some of them to learn on and the remainder to internally test the generality of a result, the overall generality of the solution can be estimated. This approach allows a more complete use of the samples in the data set as it allows the system to train on potentially different datasets over several folds. In the case of relatively small number of samples leave one out cross validation (LOOCV) is often used. In LOOCV the number of folds are set equal to the number of samples in the training set and is advantageous in the sense that a larger number of training samples are available to learn on. LOOCV is approximately unbiased for true prediction error, but can have high variance because all the "training sets" are similar to each other. We are using a variation of cross validation that selects an optimum number of folds that give a good trade-off between the bias and variance in which the number of folds are the same as the number of target samples in the training set. As mentioned earlier there are 11 NPs and 23 NNs in the training set, hence instead of having 34 folds as would be the case if LOOCV was used, 11 folds were used which leads to a reduction in variance. This guarantees that there will be at least one target sample in each test fold.

Since GP is a stochastic process and often gives more than one solution with the same accuracy, the process is run for 20 times, so as to develop a reasonable sample of possible classifier sets. The set of classifiers belonging to a run with the best cross validation performance is selected. This set is used in a majority voting scheme to classify the samples in the validation set. Aggregate performance of these "meta-rules" on the test fold was taken as the predictor of the classification error, and the selected "meta-rule" was the one with the least test error. The majority voting scheme is a simple scheme to implement and increases the both the performance and consistency of the classifiers. It has been shown that for a binary classification scheme, if the individual rules are more than 50% accurate the performance of the classifier actually increases [20,21] and the resilience can be improved much more as estimation errors are reduced.

Table 2 summarizes the key parameters for this study including the mathematical operators used, the number of genes and operators allowed in a classifier and size of the population and other evolutionary drivers like crossover / mutation rate, etc.

In a clinical or biological setting, measures of accuracy like sensitivity and specificity are straightforward, easily comprehensible fitness measures.

Table 2: Genetic Programming Parameters Summary

Objective	Find a rule predicting nodal metastases in bladder cancer patients
Function Set	= , > , < , <= , >= , and, not, or, ?, +, -, *, /, exp
Input	StaRT-PCR gene expression values for selected genes
Fitness	Area under the curve (AUC)
Population Size/Deme	15000/16
Termination	Generation 100 or Success
Demes	16
Tournament Size	4
Elitism	Yes
Crossover/Mutation Rate	0.6/0.4
Initial tree depth/ Final	4 /5
Node count	7-12
Migration percentage/ frequency	5% every 5 generations

Positive predictive value and negative predictive value are similar and may be more appropriate fitness measures in clinical studies.

There is a problem in selecting just a single measure of accuracy out of sensitivity and specificity, in that both of them are inherently complementary and one can easily be increased while decreasing the other. The overall objective of maximizing both sensitivity and specificity is built into the receiver operating characteristic (ROC) evaluation of a test [22], and the search of the most informative test usually seeks to maximize the area under the curve (AUC) [23]. The AUC measure gives a direct indication of how distinctly are the samples separated into different classes. This fitness measure is also more robust than any other mathematical combination of the sensitivity and specificity, as there is no concept of boundary or threshold that might induce discontinuities into the system leading to strange behavior around them.

One of the other important concerns in the use of GP is the penchant for the system to be stuck in a local optima. This problem assumes a larger significance in the biological/clinical paradigm wherein the data is often very noisy as this can create a large number of local optima, while the unknown nature of the biological system makes it much more difficult to ascertain whether the optima discovered is local or global. This has been addressed by adjusting the parameters to perform a more explorative search that in GP terms would mean the maintenance of diversity. An increased amount of diversity allows for a more global search and helps avoid local optima [24,25]. The population size is also associated with allowing a better search of the space as a larger population can intuitively be thought of as performing a higher resolution sampling of the search [32]. The mutation rate is also kept significantly higher than is usual in order to provide the system with the capability of escaping the local optima by increasing the entropy in the evolutionary process. The crossover rate is moderate enough to prevent the system from spending a larger amount of time in a local neighborhood.

We have implemented a GP system using coarse grained distributed GP wherein the subpopulations are separated across several demes [islands] which has also been shown to increase the performance of the algorithm by protecting against premature convergence within a single population. In order to allow for the propagation of good quality genetic material across demes we allow migration of well performing individuals from one deme to another. In the current scheme we are using a ring topology-based migration strategy. The settings have been indicated in Table 2. The island multi-population model has been shown to improve the performance in terms of searching a larger space when there is no particular benefit achieved in increasing the population for the search space [26]. Tournament size has been fixed so as to allow a proper balance between chance and selectivity of the individuals.

3.4 Results

The final "meta-rule" that we generated was constituted of 11 rules used in a majority voting scheme as mentioned above. Our prediction is based on agreement of more than 5 rules out of 11. The constituent rules are listed below.

1) exp(exp(HSF1)) - exp(MAD)/(KDR-MAP2k6) > 2.718

2) (MAP2K6/KDR) * (exp(TGIF) - MAP2K6/ICAM1) > 0.709

3) (ICAM1 - CDK8) / (exp(JUNB) * (JUNB - exp(TGFBR2))) > 1.32

4) ANXA5 * MAP2K6 / (KDR * (ICAM1-CDK8)) > 1.701

5) (ICAM1 - MAP2K6) * exp(MAP2K6 - KDR) > 3653.813

6) (ICAM1 - CDK8) * TP53 / (exp(TGFBR2) * PTGS2) > 21941.453

7) (CCND3 / MAP2K6) * (exp(BMP6) - (KDR/MAP2K6)) > .201

8) MAP2K6 / (CDKN1A * exp(MAPK12) *(CDC25C – KDR)) > 7.703

9) (ANXA5 - exp(PDGFRL)) / (CDKN1A * (KDR - exp(TGFBR2))) > .044

10) ANXA5 / (CDKN1A * (exp(PTGS2) - (CDK8/ICAM1))) > 79.002

11) MAP2K6 / (KDR * (ICAM1 - (TNFAIP1/exp(PDGFB)))) > 1.182

Table 3: Result Metrics for the "Meta-Rule" performance on the validation set

True Positive (TP)	6
True Negative (TN)	19
False Positive (FP)	2
False Negative (FN)	4
Accuracy	81%
Sensitivity	60%
Specificity	90%
Positive Predictive Value (PPV)	75%
Negative Predictive Value (NPV)	83%

Table 3 shows the results of the meta-rule used in a majority voting scheme on the validation set. 6 out of 10 node positive samples were correctly identified and 19 out of 21 node negative samples were identified. Various metrics are also presented that are calculated from the results. An accuracy of 81% is calculated by summing the True Positive and True Negative results and dividing by the total number of samples (i.e., Acc=(TP+TN)/Total Samples). Similarly the Sensitivity is 60% as calculated by dividing the True Positive value by the sum of the True Positive and the False Negative values (i.e., Sens=TP/(TP+FN) or the percentage of total targeted samples correctly identified divided by the total number of targeted samples). Specificity is the complementary measure of the True Negative value divided by the sum of the True Negative and the False Positive values (i.e., Spec=TN/(TN+FP) or the percentage of samples correctly identified as not belonging to the target class. In clinical studies, the Positive Predictive Value and Negative Predictive Value are also used and so we present these values as well. The Positive Predictive Value is calculated by dividing the number of True Positive samples by the sum of the True Positive and False Positive values (i.e., PPV=TP/(TP+FP) or the percentage of samples correctly identified as belonging to the target class divided by the number predicted to belong to the target class). The Negative Predictive Value is complementary metric calculated by dividing the True Negative by the sum of the of the True Negative value and the False Negative value (i.e., NPV=TN/(TN+FN) or the number of samples correctly identified as not belonging to the target class divided by the number of samples predicted to not belong to the target class.

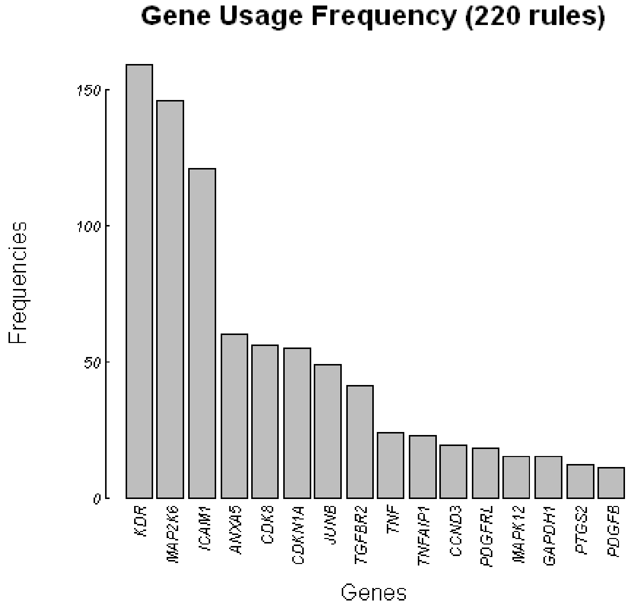

Fig. 1 Gene Frequency across 20 runs (220 rules)

3.5 Using Genetic Programming for Feature Selection

From a pure mathematical view, feature selection may be regarded as a pure combinatorial problem [27]. The difficulty with feature selection when there are many features is that it is a NP-hard problem and unless some information loss is acceptable, we generally have to spend a huge amount of computational effort to discover the most significant combination of features. GP can leverage its population-based method to serve as a powerful feature selection tool with the computational burden alleviated by parallelization. Because GP can find structure in large data sets, the performance can be improved since we can include genes that individually do not show a large amount of correlation but can be an important player when used with other genes in a predictor.

Since it facilitates automatic variable selection as well as structure generation of the solution, it could find relationships that in general are out of the reach of a classical algorithm [28].

To select the features using GP we used the statistics that can be extracted from the multiple runs with different parameters. For dimensionality reduction we calculate 'gene frequency' by tabulating the frequency of gene usage in the best performing rules for each fold across many runs. The motivation is due to cross validation consistency [29] wherein genes that are important for the classification solution would be repeated more often than others. Taking a look at the gene frequency results (Fig. 1) we can see that there are 3 genes that show a very high frequency of usage compared to the other genes and so can be thought to account for most of the variance in the solution. The p-value for the corresponding genes frequencies against a null hypothesis of a uniform random sampling are KDR – 9.69E-130, MAP2K6 1.13E-110 and ICAM1 4.10E-78 which suggests that frequencies to be of large statistical significance. These p-values are calculated by assuming a uniform distribution of gene selection and randomly assembling a rule with the same number of genes on average as is discovered in this analysis. Given the number of rules produced in the analysis, a binomial distribution is calculated to predict the expected number of selections of a gene compared with the actual number of selections.

We then ran the entire GP process using only the above 3 genes for the same classification problem. The rules generated (shown in Figure 2) showed a marked increase in the robustness of the results, along with a slight increase in the prediction accuracy. The result can be understood to improve in robustness due to previously mentioned reasons for decreasing complexity in order to prevent overfiting. The results of these runs are shown in Table 4.

Fig. 2 List of 3-Gene Rules

1) (MAP2K6 / KDR) * (1.0 - (MAP2K6 / ICAM1)) > .71
2) MAP2K6 * ((ICAM1 – KDR) / (ICAM1 * KDR)) > .705
3) MAP2K6 * ((ICAM1 – KDR) / (ICAM1 * KDR)) > .705
4) (MAP2K6 / KDR) – exp(KDR) – (MAP2K6 / ICAM1) > -.294
5) (MAP2K6 / KDR) * (1.0 - (MAP2K6 / ICAM1)) > .71
6) ICAM1 / (MAP2K6 – KDR * exp(ICAM1)) > .4.092
7) (MAP2K6 / KDR) * (1.0 - (MAP2K6 / ICAM1)) > .71
8) (MAP2K6 / KDR) – (MAP2K6 / (ICAM1 + exp(ICAM1))) > .705
9) MAP2K6 * ((ICAM1 – KDR) / (ICAM1 * KDR)) > .69
10) (MAP2K6 / KDR) – (MAP2K6 / (ICAM1 * exp(KDR))) > .705
11) (MAP2K6 / KDR) – (MAP2K6 / (ICAM1 + exp(ICAM1))) > .886

Table 4. Results & Metrics for the GP run using 3 genes

TP	7
TN	18
FP	3
FN	3
Accuracy	81%
Sensitivity	70%
Specificity	86%
PPV	70%
NPV	86%

It is interesting to note that one of the rules has a PPV of 100% and most of the rules are similarly constituted in terms of the usage of genes. The prediction accuracy is also higher for this rule. Performance of that rule on the validation set is shown in Table 5.

Table 5. Results & Metrics for the performance "(MAP2K6 / KDR) * (1.0 - (MAP2K6 / ICAM1)) > .71" rule on the Validation set

TP	7
TN	21
FP	0
FN	3
Accuracy	90%
Sensitivity	70%
Specificity	100%
PPV	100%
NPV	88%

4. DISCUSSION

Perusing the various classifiers generated in the study we observed several 'gene motifs' i.e. recurring mathematical genetic constructs in the discriminant solutions generated. These can provide us with an insight into gene regulation and the various pathways that might be similarly regulated or are correlated in terms of the tumor progression. Examples of a few motifs are (MAP2K6/KDR), (ICAM1/MAP2K6), (ICAM1-KDR), (ICAM1-CDK8), etc. It is interesting to observe a few of these in the rules presented here. Since all of the above are used in the rules of the form 'if [expression] > constant then target', we can identify the logical relationships between the genes. These relationships, though simplistic, give some insight into how their respective pathways are being affected.

The GP analysis in this study clearly shows an unequivocal preference to use ICAM1, MAP2K6 and KDR in a specific relationship to define node positive bladder cancer specimens. The association of metastatic disease with the expression levels of these proteins is not unreasonable considering their function and involvement in tumor biology. Several reports, for example, indicate that the expression levels of ICAM1 correlate with metastatic potential, migration, and infiltration ability [34-38]. In addition, expression of ICAM-1 by tumor cells has often been associated with tumor progression with highest levels usually occurring in metastatic tumors [39-48]. Ligation of ICAM-1 induces activation of MAP2K6, which in turn activates p38 [49-51], a protein kinase that has been shown to be closely associated with an invasive phenotype for several tumor types including bladder, pancreatic and prostate[52-54] as well as poor prognosis in node-positive breast cancer [55]. Other studies have shown a direct correlation of MAP2K6 activity with metastatic potential [56-58]. Finally, the more robust rules in this study also imply that the expression level of tumor KDR is consistently lower in relation to ICAM-1 and MAP2K6 when there is nodal involvement in bladder cancer. Although a precise reason for why this relationship should exist is unknown, some studies have established a more aggressive phenotype in cancers that have lower expression of KDR [59].

We tried an interesting approach of only using Boolean expressions for generating classifiers, in order to explore whether Boolean rules would identify the same logical conclusions that we drew from the motif analysis. The GP system found almost the same motifs in the Boolean search though it was constrained to construct rules that only used Boolean operators. A few examples that show the constructs are 'if (ICAM1 > MAP2K6) then Node positive' or 'if (MAP2K6 > KDR) then Node positive'. These rules have a reasonably good performance on the training set and thus lend some credence to our results.

A burning problem in GP setup is to develop a technique where a single rule can be automatically selected, as there is often an embarrassment of riches in the form of too many results with the same fitness. We believe this can be addressed using the approach of dimensionality reduction described above as this reduces the search space to a reasonably small area thus increasing the probability of finding the correct solution is increased across all the runs. The single rule highlighted in Table 5 was discovered in the analysis employing the three most frequently used genes from our study and it proved to be much more robust in terms of predicting the nodal stage in the validation set.

5. CONCLUSIONS

Genetic programming can be seen as an appropriate learning and hypothesis generation tool in a biological/clinical setting, with the additional value of the human readability of the results. This allows better insights into the mechanisms of the biological processes. The small number of samples in this study limits the statistical power of the conclusions. However this study can be regarded as a strong hypothesis generating process regarding the importance of the three genes identified and their use in determining nodal status, and can motivate further similar studies with larger sample size to validate these results.

6. REFERENCES

[1] American Cancer Society. Cancer Facts and Figures 2005. Atlanta: American Cancer Society; 2005

[2] Hermanek P, Hutter RVP, Sobin LH, Wagner G, Wittekind CH, eds. TNM Atlas, 4th edn. XX. Springer, Berlin, 1997

[3] Epstein JI, Amin MB, Reuter VR, Mostofi FK, The World Health Organization/International Society of Urological Pathology Consensus Classification of urothelial (transitional cell) neoplasms of the urinary bladder. Am J Surg Pathol 22, (1998), 1435-48

[4] Esrig D, Elmajian D, Groshen S, et al. Accumulation of nuclear p53 and tumor progression in bladder cancer. N Engl J Med. 331, (1994), 1259-64

[5] Lotan Y, A Gupta, SF Shariat, GS Palapattu, A Vazina, PI Karakiewicz, PJ Bastian, CG Rogers, G Amiel, P Perotte, MP Schoenberg, SP Lerner, AI Sagalowsky. Lymphovascular invasion is independently associated with overall survival, cause-specific survival, and local and distant recurrence in patients with negative lymph nodes at radical cystectomy. J Clin Oncol, 23,27 (2005),6533-6539

[6] Apostolakos MJ, Schuermann WH, Frampton MW, Utell MJ, Willey JC. Measurement of gene expression by multiplex competitive polymerase chain reaction. Anal Biochem. 213 , (1993), 277-84

[7] Crawford EL, Warner KA, Khuder SA, Zahorchak RJ, Willey JC. Multiplex standardized RT-PCR for expression analysis of many genes in small samples. Biochem Biophys Res Commun. 293 , (2002), 509-16

[8] RJ Cote and RH Datar, Therapeutic approaches to bladder cancer: identifying targets and mechanisms. Critical Reviews in Oncology-Hematology (Supplemental), 46, (2000), S67-S83.

[9] Andre, D. and Koza, J.R, A parallel implementation of genetic programming that achieves super-linear performance. In Hamid R. Arabnia, ed., Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications, volume III, (Sunnyvale, 9-11 August 1996) 1163--1174

[10] Koza, J.R, Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press, Cambridge, MA, 1992

[11] Brameier M., Haan J., Krings A., MacCallum R, Automatic discovery of cross-family sequence features associated with protein function
BMC Bioinformatics (2006), 7:16

[12] Driscoll JA, Worzel WP, MacLean CD., Classification of Gene Expression Data with Genetic Programming. In Genetic Programming Theory and Practice. R Riolo, WP Worzel (eds.), Springer Science+Business Media, Inc., NY, pp: 25-42, 2004

[13] J.H. Hong, S.B. Cho. Cancer Prediction Using Diversity-Based Ensemble Genetic Programming. Lecture Notes in Computer Science 3558 (2005) 294-304

[14] Langdon W., Buxton B., Genetic Programming for Mining DNA Chip Data from Cancer Patients. Genetic Programming and Evolvable Machines 5(3), (2004),251-257

[15] Hastie T., Tibshirani R., and Friedman J., The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer series in statistics. Springer, Berlin, 2001.

[16] Rissanen, J., Modeling by shortest data description, Automatica 14, (1978) 465-471.

[17] Vapnik, V. N. and Chervonenkis, A. Y, On the uniform convergence of relative frequencies of events to their probabilities. Theory Probab. Apl., (1971) 16, 264-280.

[18] Vapnik, V.N, The Nature of Statistical Learning Theory, Springer-Verlag, Berlin, 1995.

[19] Schaffer, C., Selecting a classification method by cross-validation, Machine Learning 13(1), (1993), 135-143.

[20] Narasimhamurthy A, A framework for the analysis of majority voting. Lectures notes in Computer Science 2749, (2003), 268-274,

[21] Narasimhamurthy A, Theoretical bounds for majority voting performance for a binary classification problem. PAMI, in press

[22] Zou KH, Hall WJ, Shapiro DE. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. Stat Med. 15, 16, (Oct, 1997) 2143-56.

[23] Handley JA and McNeil BJ, The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), (1982 Apr), 29-36.

[24] S. Gustafson., An Analysis of Diversity in Genetic Programming. Ph.D. Dissertation, School of Computer Science and Information Technology, University of Nottingham, Nottingham, U.K, 2004

[25] E. Burke, S. Gustafson, and G. Kendall, A survey and analysis of diversity measures in genetic programming. In GECCO 2002: Proceedings of the Genetics and Evolutionary Computation Conference, (New York, 9-13 July 2002). Morgan Kaufmann,716–723

[26] Folino, C. Pizzuti, G. Spezzano L. Vanneschi, M. Tomassini, Diversity Analysis in Cellular and Multipopulation Genetic Programming in Proceedings of Congress on Evolutionary Computation 2003, (CEC'03) (Canberra 8-12 Dec 2003), IEEE Press, Piscataway, NJ, 305-311

[27] Keinosuke Fukunaga, Introduction to Statistical Pattern Recognition, AcademicPress, London, 1990.

[28] Moore, J.H., Parker, J.S., Olsen, N.J., and Aune, T., Symbolic discriminant analysis of microarray data in autoimmune disease. Genetic Epidemiology 23, (2002), 57-69.

[29] Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Plummer, W.D., Parl, F.F. and Moore, J.H., Multifactor Dimensionality Reduction Reveals High-Order Interactions among Estrogen Metabolism Genes in Sporadic Breast Cancer. American Journal of Human Genetics, 69, (2001), 138-147.

[30] MacLean CD, Wollesen EA, Worzel WP., Listening to Data: Tuning a Genetic Programming System. In: Genetic Programming Theory and Practice II. U-M O'Reilly, T Yu, R Riolo, WP Worzel (eds.), Springer Science+Business Media, Inc., NY, (2005), 245-62

[31] Daida, J.M., What Makes a Problem GP-Hard? A Look at How Structure Affects Content. Genetic Programming Theory and Practice, eds. Riolo R.L. & Worzel W.P., Kluwer, Boston, MA, 2003, 99-118

[32] Bellman, R. Adaptive Control Processes: A Guided Tour, Princeton University Press, NJ, (1961).

[33] K. Sastry, U.-M. O'Reilly, D. E. Goldberg, Population Sizing for Genetic Programming Based Upon Decision Making in O'Reilly, U.M., et al. Genetic Programming Theory and Practice II. Kluwer Academic Publishers, Boston, MA (2004), 49-66

[34] Roche Y, Pasquier D and Rambeaud JJ., Fibrinogen mediates bladder cancer cell migration in an ICAM-1-dependent pathway. *Thromb Haemost,* 89, (2003), 1089-97.

[35] Ozer G, Altinel M, and Kocak B., Potential value of soluble intercellular adhesion molecule-1 in the serum of patients with bladder cancer. *Urol Int,* 70, (2003), 167-71.

[36] Palumbo JS, Potter JM, and Kaplan LS., Spontaneous hematogenous and lymphatic metastasis, but not primary tumor growth or angiogenesis, is diminished in fibrinogen-deficient mice. *Cancer Res,* 62, (2002), 6966-72.

[37] Rosette C, Roth RB, and Oeth P., Role of ICAM1 in invasion of human breast cancer cells. *Carcinogenesis,* 26, (2005), 943-50.

[38] Jackson AM, Alexandrov AB, and Prescott S., Expression of adhesion molecules by bladder cancer cells: modulation by

interferon-gamma and tumour necrosis factor-alpha. *J Urol,* 148, (1992), 1583-6.

[39] Miele ME, Bennett CF, and Miller BE., Enhanced metastatic ability of TNF-alpha-treated malignant melanoma cells is reduced by intercellular adhesion molecule-1 (ICAM-1, CD54) antisense oligonucleotides. *Exp Cell Res,* 214, (1994), 231-41.

[40] Koyama S, Ebihara T, and Fukao K, Expression of intercellular adhesion molecule 1 (ICAM-1) during the development of invasion and/or metastasis of gastric carcinoma. *J Cancer Res Clin Oncol,* 118, (1992), 609-14.

[41] Maruo Y, Gochi A, and Kaihara A., ICAM-1 expression and the soluble ICAM-1 level for evaluating the metastatic potential of gastric cancer. *Int J Cancer,* 100, (2002), 486-90.

[42] O'Hanlon DM, Fitzsimons H, and Lynch J., Soluble adhesion molecules (E-selectin, ICAM-1 and VCAM-1) in breast carcinoma. *Eur J Cancer,* 38, (2002), 2252-7.

[43] Regidor PA, Callies R, and Regidor M, Expression of the cell adhesion molecules ICAM-1 and VCAM-1 in the cytosol of breast cancer tissue, benign breast tissue and corresponding sera. *Eur J Gynaecol Oncol,* 19, (1998), 377-83.

[44] Santarosa M, Favaro D, Quaia M, et al., Expression and release of intercellular adhesion molecule-1 in renal-cancer patients. *Int J Cancer,* 62, (1995), 271-5.

[45] Sun JJ, Zhou XD, Liu YK, et al, Invasion and metastasis of liver cancer: expression of intercellular adhesion molecule 1. *J Cancer Res Clin Oncol,* 125, (1999), 28-34.

[46] Tempia-Caliera AA, Horvath LZ, Zimmermann A, et al., Adhesion molecules in human pancreatic cancer. *J Surg Oncol,* 79, (2002), 93-100.

[47] Yano S, Nokihara H, Yamamoto A, et al., Multifunctional interleukin-1beta promotes metastasis of human lung cancer cells in SCID mice via enhanced expression of adhesion-, invasion- and angiogenesis-related molecules. *Cancer Sci,* 94, (2003), 244-52.

[48] Roux PP and Blenis J, ERK and p38 MAPK-activated protein kinases: a family of protein kinases with diverse biological functions. *Microbiol Mol Biol Rev,* 68, (2004), 320-44.

[49] Wang Q and Doerschuk CM, The p38 mitogen-activated protein kinase mediates cytoskeletal remodeling in pulmonary microvascular endothelial cells upon intracellular adhesion molecule-1 ligation. *J Immunol,* 166, (2001), 6877-84.

[50] Zarubin THan J., Activation and signaling of the p38 MAP kinase pathway. *Cell Res,* 15, (2005), 11-8.

[51] Huang X, Chen S, Xu L, et al., Genistein inhibits p38 map kinase activation, matrix metalloproteinase type 2, and cell invasion in human prostate epithelial cells. *Cancer Res,* 65, (2005), 3470-8.

[52] Lin M, DiVito MM, Merajver SD, et al., Regulation of pancreatic cancer cell migration and invasion by RhoC GTPase and caveolin-1. *Mol Cancer,* 2004, 4 :21.

[53] Ott I, Weigand B, Michl R, et al., Tissue factor cytoplasmic domain stimulates migration by activation of the GTPase Rac1 and the mitogen-activated protein kinase p38. *Circulation,* 111, (2005), 349-55.

[54] Esteva FJ, Sahin AA, Smith TL, et al., Prognostic significance of phosphorylated P38 mitogen-activated protein kinase and HER-2 expression in lymph node-positive breast carcinoma. *Cancer,* 100, (2004), 499-506.

[55] Kim MS, Lee EJ, Kim HR, et al., p38 kinase is a key signaling molecule for H-Ras-induced cell motility and invasive phenotype in human breast epithelial cells. *Cancer Res,* 63, (2003), 5454-61.

[56] Shin I, Kim S, Song H, et al., H-Ras-specific activation of Rac-MKK3/6-p38 pathway: its critical role in invasion and migration of breast epithelial cells. *J Biol Chem,* 280, (2005), 14675-83.

[57] Behren A, Binder K, Vucelic G, et al., The p38 SAPK pathway is required for Ha-ras induced in vitro invasion of NIH3T3 cells. *Exp Cell Res,* 303, (2005), 321-30.

[58] Gakiopoulou-Givalou H, Nakopoulou L, Panayotopoulou EG, et al., Non-endothelial KDR/flk-1 expression is associated with increased survival of patients with urothelial bladder carcinomas. *Histopathology,* 43, (2003), 272-9.

[59] Ferrer FA, Miller LJ, Lindquist R, et al., Expression of vascular endothelial growth factor receptors in human prostate cancer. *Urology,* 54, (1999), 567-72.