Proceedings of the 2007 IEEE Symposium on Computational
Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Predicting Tumor Malignancies using Combined Computational Intelligence, Bioinformatics and Laboratory Molecular Biology Approaches

Jack Y. Yang
Department of Radiation Oncology, Massachusetts
General Hospital and Harvard Medical School
Boston, Massachusetts 02114 U.S.A
jyang@hadron.mgh.harvard.edu

Mary Qu Yang
National Human Genome Reserch Institute
National Institutes of Health, U.S. Department of Health
and Human Services, Bethesda, M.D. 20852 USA
yangma@mail.nih.gov

Andrzej Niemierko
Division of Biomathematics and Biostatistics and Dept. of
Radiation Oncology, Massachusetts General Hospital and
Harvard Medical School, Boston, MA 02114 U.S.A

Zuojie Luo and Jianling Li
Department of Endocrinology, First Affiliated Hospital,
Guangxi Medical University, Nanning, Guangxi, 500021,
China (Z. Luo also with Joslin Diabetes Ctr, Harvard Univ.)

Abstract—**Predicting tumor malignancies is an important but difficult task. For many tumors, especially neural and endocrine tumors, traditional pathological and histological analyses often can not effectively distinguish benign from malignant tumors. Developing synergistic bioinformatics and computational intelligence system is effective, because deterministic cancer markers do not always exist in individual patients. We proposed a parallel paradigm of cancer and use a number of ensemble methods including Boosting, Bagging and Consensus Networking, and have designed a novel classification scheme that advantageously combines several computational intelligence algorithms such as the variants of Self-Organizing Feature Map (SOFM) algorithms and the Maximum Contrast Tree (RMCT) algorithms. Boosting and Bagging have been advantageously combined. When all of the above are integrated into one synergistic intelligent medical decision system, the prediction power for the task has been significantly boosted. The system and features are validated by diagnosing new patients and by a number of laboratory molecular biology measurements. The outcomes of the research have improved cancer diagnosis and treatment planning, and may lead to diagnose microscopic diseases and better understanding of human genome mechanisms relating to malignant transformation.**

Keywords— Computational Intelligence, Bioinformatics, Parallel Paradigm of Cancr, Benign, Maliganant Transformation.

I. INTRODUCTION

Predicting malignancies plays essential roles not only in revealing human genome mechanisms of potential malignant transformation, but also in discovering effective prevention and treatment of cancers.

Recently, the National Human Genome Research Institute and National Cancer Institute, both part of NIH, U.S. Department of Health and Human Services, have launched The Cancer Genome Atlas (TCGA) with an overarching goal of understanding the molecular basis of cancer to improve our ability to diagnose, treat and prevent cancer [8].

The perspective of the TCGA project is that “cancer is not a single disease but a collection of diseases that arise from different combinations of genetic changes. Scientists must analyze the genetic material from different tumors and many patients to uncover the tell-tale genetic signatures of different cancer types” [8]. Based on the mission of TCGA, we have proposed a further **Parallel Paradigm of Cancer**: it is not only the genetic changes such as mutations of genes but also changes of gene expressions and regulatory networks that are eventually responsible for cancer development. Under this parallel paradigm, not only mutations of genes cause changes in gene regulatory networks; but also un-mutated genes with differential expressions and alternative splicing may also trigger the changes in the differential regulatory networks that are also responsible for cancers, especially when cells are in an abnormal environment. Because of the tiny differences between cancer and normal tissue in their same genotypes, however, their biological behaviors “phenotypes” are very different. Therefore, we investigate the differential expressions of genes among normal, benign and cancerous tissues in addition to the bioinformatics survey of human genome and cancer genetics. We need to solve current challenges in cancer diagnosis and prognosis:

- Ineffectiveness: Modern medical image technologies such as CAT scan, MRI scan, PET and X-rays all have their own problems and limitations. A detectable tumor tissue usually contains more than a hundred million tumor cells, which is 0.1 gram of a cluster of irregular tumor tissue in weight and a quarter inch in dimension. Those invisible tumor tissues are commonly referred to as microscopic diseases that are not diagnosable [1,9]. All these guarantee ineffectiveness.
- Inaccuracy: Highly characteristic (deterministic) cancer markers are not likely found for every individual patient. Accurate diagnosis and prognosis require verification of

1-4244-0710-9/07/$20.00 ©2007 IEEE

tumor biological behaviors which are largely unpredictable. Confirmations of cancer are typically made by pathological and histological analyses, which are not always effective, especially for neural and endocrine tumors. Therefore, many patients are actually diagnosed tumors of visible sizes but unknown malignancies.

- Inconsistency: The existence of multiple but inaccurate diagnostic methods guarantees inconsistency among differential diagnoses in many patients. We can not rely on any one of the multiple tumor associated antigens alone, because using another one independent may result a different diagnosis.
- Inefficiency: Because PET may detect much smaller tumors than CAT, MRI and X-rays and because blood and urine tests for screening cancer antigens are not always effective, from time to time, patients are identified of possible cancer that need further procedures for firm confirmations. But those procedures do not always work.

We need to solve above problems by developing synergistic bioinformatics and computational intelligence medical decision systems. This paper is a further development of our prototype system in [1] and advantageously combined the computational intelligence algorithms we developed before in [2-4]. We focus on the neural and endocrine tumors that are hard to identify malignancies [1,14,15].

According to NHGRI-NIH, the cost to sequence genomes will be covered by major insurance policies. The era of affordable patient-specific medicine based on the full complement of genes is not too far away. However, deterministic cancer markers do not always exist in individual patients because even for the same type of cancer, the genetic mechanisms may be different. Human genome is abundant with alternative splicing – same gene but different protein products. To support our parallel paradigm of cancer, we developed a novel synergistic medical decision system to predict malignancies of human pheochromocytomas and paragangliomas and tumors of Cushing's syndrome of different adrenocortical diseases because identifying malignancies of all above tumors challenges all methods in pathology, histology and medical images [1,13-15].

Pheochromocytomas arising from the adrenal gland are closely related to paragangliomas arising from the paraganglionic system. The bioinformatics screening of human genome and statistics on patients have identified that only approximately 1/5 of pheochromocytomas are genetically inherited, resulted from mutations of the human genes SDHB, VHL, RET, NF1 and SDHD that cause the familial pheochromocytomas and extra-adrenal paragangliomas. Roughly 1/4 of paragangliomas are genetically inherited, resulted from mutations of human genes SDHD, PGL2 and SDHC. Most of those tumors are benign, however at least 8% of pheochromocytomas and 3% paragangliomas metastasize. Tumors of Cushing's syndrome are also related to the adrenal gland. Cushing's syndrome is called hypercortisolism or hyperadrenocorticism caused by excessive levels of the endogenous levels of corticosteroid hormone cortisol secreted by the adrenal glands that are related to the regulations by the pituitary gland and hypothalamus in the brain. Strictly, Cushing's syndrome refers to excess cortisol of any etiology while Cushing's disease which account to roughly 2/3 Cushing's syndrome refers only to hypercortisolism secondary to excess production of adrenocorticotropin (ACTH) from a pituitary gland adenoma. The rest of 1/3 Cushing's syndrome are tumors of a group of adrenocortical diseases that include adrenocortical carcinoma (ACC), adrenocortical adenoma (ACA) and adrenocortical hyperplasia (ACH). They all lead to hypercortisolism. Most of those adrenocrtical tumors are benign, however at least 1/6 of them metastasize. All of the above tumors are neuroendocrine tumors and are all difficult to identify malignancies based on clinical symptoms and pathological features. The above situations also suggest our parallel diagram that the aetiology of those neuroendocrine tumors are not limited to mutations of genes but also differential gene expressions that affect the gene regulatory networks. That is why we choose those tumors for developing a synergistic intelligent medical decision system.

TABLE I
EXPRESSION PROFILES OF FHIT, KI-67 AND PCNA IN ADRENOCORTICAL DISEASES WITH CLINICAL INFORMATION

Patients	Negative (%)			Total Positive (%)		
	FHIT	Ki-67	PCNA	FHIT	Ki-67	PCNA
Age<40	16.67	70.00	3.33	83.33	30.00	96.67
Age>40	11.05	26.32	10.53	78.95	73.68	89.47
Male	57.14	58.82	0.0	42.86	41.18	100
Female	15.62	78.12	9.37	84.38	21.88	90.63
Left side	11.43	68.86	7.14	78.57	31.14	92.86
Right side	14.29	76.19	4.76	85.71	23.81	95.24

TABLE II
EXPRESSION PROFILES S OF FHIT, KI-67 AND PCNA IN ADRENOCORTICAL DISEASES OF DIFFERENT MALIGNANCIES

Tissue	Negative (%)			Total Positive (%)		
	FHIT	Ki-67	PCNA	FHIT	Ki-67	PCNA
ACC	57.14	14.83	0.0	42.86	85.17	100
ACA	3.85	92.31	3.85	96.15	7.69	96.15
ACH	0.0	100.0	22.22	100.0	0.0	77.78

II. JOINT ROLE OF TUMOR ASSOCIATED GENE EXPRESSIONS

In many cases, pathological analyses and patient's symptoms are not sufficient to identify malignancies especially for neural and endocrine tumors. We use bioinformatics techniques to survey the human genome and tumor genetics and have identified several tumor associated markers such as expression profiles of hTERT, cyclin E, P27kip1, FHIT, Bax, Bcl-2, Fas, FasL, PCNA, and Ki-67 that are useful for predicting tumor malignancies of Cushing's syndrome and pheochromocytomas and paragangliomas. We investigated the role of an enzyme called telomerase in the process of tumor [1,12,21]. In human genome, telomerase is a protein complex composed of at least two sub-units that are coded by two different genes: hTERT (human Telomerase Reverse Transcriptase) and human Telomerase RNA (hTERC or hTR). It appears that inactivation of P53 [15] - a tumor suppressor and retinoblastoma proteins (pRb) are associated with telomeres shortening, thus affecting the integrity of human genome. At the activation of telomerase, cells may be "immortal" just like tumor cells. Cell arrest gene P27kip1 [16] belongs to a member of the universal cyclin-dependent kinase inhibitor family, which is able to arrest cell cycle [13] progression by complex cyclin-dependent kinase, therefore P27kip1 can be considered as a tumor suppressor gene along with FHIT (Fragile Histidine Triad). Fas is a tumor necrosis factor receptor and FasL is Fas ligand. Bax and Bcl-2 [15] are apoptosis related factors. PCNA, Ki-67 [15] and Cyclin E. [16] are all cell cycle [13-16] and cell proliferating related genes. We performed experiments to measure the expression levels of all above genes such as measuring those tumor-associated antigen levels by immunochemistry and measuring mRNA by *in situ* hybridization using cDNA probes. Our precise experiments showed likely high level of expressions of hTERT in malignant and borderline tumors (a separate group between benign and malignant tumors), but unlikely in benign tumors and no expressions in normal tissues [1]. Our experiments indicated clear tendencies that levels of expressions of cell proliferating related antigens such as PCNA, Ki-67, Cyclin E. and tumor necrosis related factors such as Fas, FasL increase while malignancies increase. The levels of expressions of cell arrest gene P27kip1 and tumor suppressor gene FHIT decrease while malignancies increase. Apoptosis related factors such as Bax and Bcl-2 are not highly characteristic because we do not know if necrosis - a distinct feature in malignancies is triggered by apoptosis or not. Tables 1-2 show the tendency patterns of FHIT, Ki-67 and p27Kip1 of ACC, ACA and ACH from our experiments. It appears that there are statistical significances of those gene expression levels, but none of them are deterministic. Tumor occurrence rate increases monotonically with age. The general pathway: Normal Tissue -> Benign -> Malignant Cancer cannot be reversed spontaneously. According to the multistage theory of cancer [10], cancer is originated from one or small number of specifically differentiated cells that usually take years to grow into visible size. If benign stage occurs as microscopic disease (invisible tumor [9]), then it is not detectable. A normal cell maintains a completely ordered gene expressions and regulatory networks while a tumor cell is not. Our parallel paradigm indicates that the degrees of malignancies are roughly proportional to the degrees of disorder in gene expressions and regulatory networks. This can be viewed by chaos theory [11] that an ordered system can "spontaneously evolve" to a disordered system. Table 1 shows that a malignant cancer marker Ki-67 [15] is evidently highly more likely expressed and tumor suppressor FHIT is slightly less likely expressed among patients above age 40. The overall degree of disorder of Ki-67, FHIT and PCNA expressions is higher for age over 40. Although PNCA (Proliferating Cell Nuclear Antigens) does not support the scenario of age, it is a highly characteristic malignant cancer marker as shown in table 2. It is possible that gene expressions of normal tissue surrounding malignant transformed tissue can be influenced by microscopic diseases. One clue coming from PCNA as shown independent of age in table 1 is actually detectable in certain normal tissues adjacent to some cancers. These phenomena concur with our parallel paradigm of cancer and may enable us to diagnose microscopic diseases. This motivated us to develop a synergistic bioinformatics and computational intelligent medical system to predict tumor malignancies using those tumor associated genes jointly.

III. The Molecular Biology experiments

Training instances are from tumor and tissue samples that are formalin fixed and paraffin wax embedded. All samples were inspected and reviewed dually by both pathological and molecular biology methods to ensure that all tumors have shown the classical histology and typical immunochemical patterns for neuroendocrine markers. Malignancy is defined as the presence of metastasis and/or extensive loan invasion. Tumors without metastases, but having some histological suspicious features are categorized as borderline tumors. The borderline tumors are more common in neural and endocrine tumors because it is so difficult to identify malignancies from benign tumors. The demographic data, clinical characteristics, and laboratory findings were all carefully annotated from the clinical records, laboratory finding, and follow-up data after treatments. We are authorized to use of those tumors and tissues for research purpose only.

To determine hTERT mRNA expressions, *in situ* hybridization experiments were performed using a standard clinical hTERT ISH Detection Kit [1]. We designed the biotin-labeled cDNA probes complementary to the hTERT mRNA using verified mRNA sequences. After removal of paraffin and dehydration, tissue slides were pretreated with proteinase K and then fixed with paraformaldehyde in phosphate-buffered saline (PBS) with standard procedures. The slides were hybridized and were then incubated with Avidin-Biotin-HRP Complex. Color reaction was detected by incubating slides in diaminobenzidine solution containing hydrogen peroxide, counterstained with hematoxylin. Darker colored nuclei and occasionally cytoplasmic stains are regarded as positive signals. Controls for specificity were performed by pre-treating tissue sections with RNAse and slides hybridization without probes. As positive controls, hybridization to known positive samples of bladder carcinoma tissue was performed. Traditionally only metastasized tumors are considered as malignant. We used immunochemical approaches to measure the levels of gene expressions.

Paraffin covers were removed from tissues and then placed on poly-1-lysine coated slides. Paraffin was cleared out in xylene, and samples were re-hydrated. For antigen unmasking, the tissues were treated by standard molecular biology laboratory procedures. Tissue sections were then washed in PBS. FHIT rabbit polyclonal antibodies were from Zhongshan

Biotechnology (Beijing, China). Ki-67 and PCNA mouse monoclonal antibody kits (ready-to-use) were the products from Maixin Biotechnology Development Co. (Fuzhou, China). Immunochemical procedures were mainly performed using the Maxvision™ HRP-Polymer anti-Mouse IHC Kits. For each antibody, negative control was performed using Immunol Staining Primary Antibody Dilution Buffer from Beyotime Institute of Biotechnology, instead of the primary antibody. We use the known positive samples of bladder carcinoma tissues, gastric adenocarcinoma tissues and lupus nephritis etc tissues as positive controls for hTERT protein, Ki-67 antigen and P27Kip1 and stomach tissue as the positive control for FHIT etc. All immunostained slides were inspected and analyzed using high-resolution Microscope Image Analyzer DMR+Q550 (Figures 1-3). Darker colored nuclear staining markers are regarded as positive signals, whereas cytoplasmic staining (non-specific background staining) markers are considered as negative. At least 10 (actually 20 or more) randomly chosen high-power (HPFs [x400]) areas were used to determine the antigen expression levels. As shown in Figure 3, left side have more concentrated stains than right sides, technically, randomly sampling more areas in a tissue is better. However, the immunochemical experiments are highly qualitative but least quantitative and also least expensive among all other methods. 10 randomly sampled areas are enough. We only use 4 levels of antigen expressions; therefore, the number of randomly sampling areas such as 10 areas or 20 areas would not likely matter at all as explain below: The expression level is measured as percentage of the number of positively stained tumor cells counted in all sampled areas over the total number of tumor cells counted in all sampled areas, the result is the average of those ratios over all sampled areas (at least 10 or 20 areas). If those average stained ratios are less than 5%, then they are considered as negative. 5%-25% positively stained cells are considered low levels of expressions (+), 25% -50% are considered as medium (++) whereas those with >50% positive tumor cells are considered high levels of expressions (+++). In general, we use above 4 levels to indicate a gene expression level; therefore, the numbers of randomly sampled areas such as at least 10 areas or 20 areas should not affect the results of above 4 levels. But just sample one or two areas may cause inaccuracy. As shown in fig. 3, apparently the left area has more percentage of cells stained than right area and if the percentage of cells stained is at the threshold values say 5%, 25% or 50% or around those percentages, expression levels may change from one grade to another. Statistically such situations are considered as rare events that can drastically affect the results, and that is the reason we sampled at least 10 or 20 area randomly to avoid classifying expression level one grade higher or lower. Never the less the immunochemical methods are not quantitative but are just qualitative. There are several other more quantitative methods that are more expensive but do not necessarily mean much better. For example, DNA microarray [17] has advantage of measuring gene expression levels of many genes simultaneously but usually also comes large noises and less specificity for individual genes. We consider whether an antigen is expressed or not is more important than the exact level of expression. All slides for FHIT and all other antigens for immunochemistry experiments have been dually evaluated by both pathological and cancer molecular biology analyses. Above are just brief description of our detailed experiments to qualitatively measure the gene expression levels of hTERT, PCNA, Ki-67, P27Kip1, FHIT, Cyclin E., Bax, Bcl-2, Fas and FasL. Note Fig 3 shows that tumor suppressor FHIT is highly expressed in ACH – aggregation of "normal" tissue, while Fig. 1 shows that FHIT is not expressed in malignant cancer ACC and Fig. 2 shows FHIT is medially expressed in benign tumor (ACA). Those results do support our parallel paradigm on cancers.

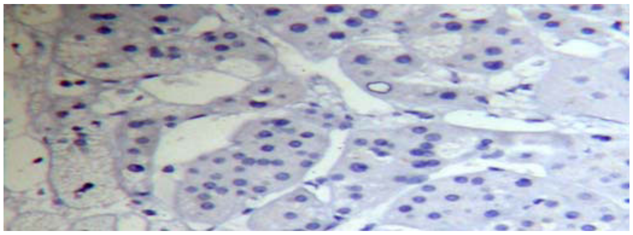

Fig. 1. Adrenocortical Carcinoma (ACC) FHIT Negative (-)

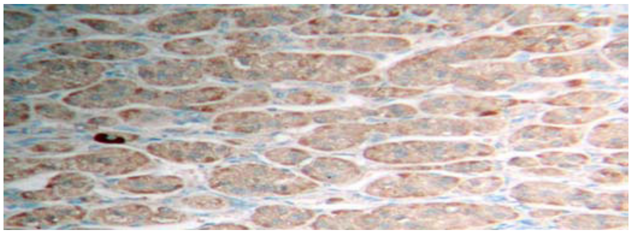

Fig. 2. Adrenocortical adenoma (ACA) FHIT Medium Positive (++).

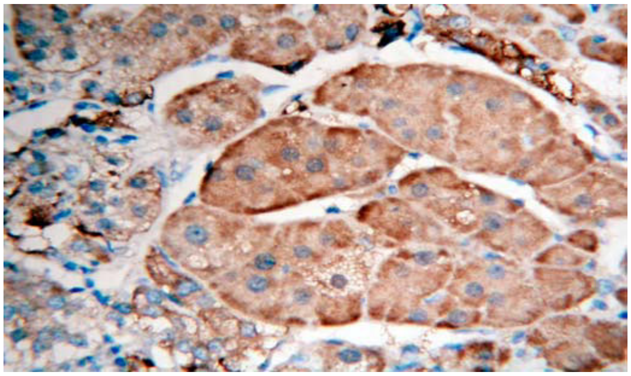

Fig. 3. Adrenocortical Hyperplasia (ACH) FHIT Strong Positive (+++).

IV. THE FEATURE SELECTION (FS) AND FILTERING

In classification problems, we are often interested in maximizing the true positive rate (also called the sensitivity), as this rate reflects the ability of the classifier to detect the "signal". Our system is to predict whether or not a given patient has malignant cancer (in this case the "signal" is "having malignant carcinoma"), then the cost of saying that the patient does not have malignant carcinoma when in fact the patient does (the false negative rate) is much higher than the cost of saying that the patient has malignant cancer when in fact the patient does not (the false positive rate). Thus it is more important to make the false negative rate smaller than it is to make the false positive rate small. Since true positive rate = 1 - false negative rate and true negative rate = 1 - false positive rate, it is desirable in many applications to make the true positive rate (i.e. the sensitivity) large at the expense of the true negative rate (i.e. the specificity). Sensitivity makes the *y-axis* and (1-specificity) makes the *x-axis* in Receiver Operating Characteristic (ROC) curve. A complete prefect random "classifier" gives a diagonal line with Youden Index = 0 (Youden index is the sensitivity + specificity – 1), while a perfect deterministic classifier always gives both sensitivity and accuracy equal to 1 with Youden Index = 1. A large ROC area and a large Youden Index indicate a good classifier. In our case, a true positive corresponds to the case of correctly classifying a malignant cancer patient. Malignant cancers tend to be less distinctive than benign compare to normal tissues [1]. Characteristic tumor associate gene expressions may turn on to have the desirable properties that they can be used to enhance sensitivity at the expense of specificity. To qualify for features measured by our experiments in the classifier, any two features must not be statistically correlated, and must give a satisfactory distance separation in the feature space (between classes) and must offer good generalization for the predictor [2]. Feature selection (FS) and filtering algorithms are divided into wrapper-based, or embedded, or filter-based FS.

A. Wrapper-based feature selection

Wrapper algorithms are interactive FS by using the inductive principles of learning principle in the FS steps. Wrapper algorithms usually outperform other FS but are extremely computationally expensive.

B. Embedded feature selection algorithms

Decision Trees and CART (classification and regression trees) exemplify embedded feature selections; the process of selecting a feature to split at each node of the tree is implicitly a feature-selection step.

C. Filter-based feature selection algorithms

Those preprocessing FS algorithms are independent to the learning algorithms and are usually computationally least expensive. We implemented a number of this type of FS [2-4,20] that including:

i. *T-test*
ii. *The Chi-Square Goodness-of-Fit test.*
iii. *The Bi-Normal Separation (BNS).*
iv. *Fisher's Permutation test*
v. *Distance Measures.*
vi. *Principle Component Analysis (PCA)*
vii. *Information Gain - how a decision tree selects a feature to split.*

We used a distance based FS. Let's consider a two-class, malignancy and benign. Given two features: *X* and *Y*, *D(X)* and *D(Y)* measure the separation of two classes subject to feature *X* and feature *Y*, respectively. If *D(X) > D(Y),* feature *X* is selected or if *D(X) < D(Y),* feature *Y* is selected. Hence, the decision rule is: $\boldsymbol{D(X)} \overset{>}{<} \boldsymbol{D(Y)}$

Distance measure can be in pair each time for every feature. After sorting these distances, we selected a number of most useful features for separating two classes. However, distance-measure based FS works in a pair-wise manner. We also used another feature selection method called Principle Component Analysis (PCA) to reduce the feature dimension. PCA can handle multiple features simultaneously. PCA is also called Karhunen-Loeve (K-L) transformation. K-L transformation is an orthonormal transformation of a vector ~X to same dimensional vector ~Y. In the transformation domain, the first principle component is the normalized linear combination with maximum variance; the second component has the next largest variance and so forth. Based on such ranking, only those with largest variance are preserved and the others are neglected. In fact the principle components are ranked by their ability to distinguish among classes. The implement procedure is as follows. Assume there are *N* instances in the training set and M features, Let ~ *X* represent a population of *N-dimensional* vectors, mean value of each feature m_x has been calculated. The mean value of each feature and KL transformation have been performed, the resulting covariance matrix of ~*Y* has been analyzed. Some features are correlated with each other. The correlation factor ρ_{xy} of two features *X* and *Y* can be obtained by the following equation:

$$\rho_{xy} = \frac{\mathrm{E[XY] - E[X]E[Y]}}{\sqrt{\mathrm{E[X^2]} - E\mathrm{[X]}^2}\sqrt{(\mathrm{E[Y^2]} - E\mathrm{[Y]}^2)}}$$

Our FS are made by decision rules based on both PCA and above distance measurements.

V. The Synergistic Medical Decision System

Training samples are 41 selected pheochromocytomas and 4 paragangliomas, 14 selected adrenocortical carcinoma, 26 selected adrenocortical adenoma, and 9 selected adrenocortical hyperplasia. There are also 9 selected normal adrenal glands from patients underwent nephrectomy of renal tumors. Both above FS methods ruled out apoptosis related factors Bax and Bcl-2. Fas and FasL correlated strongly. FS ranked the highly useful features in the following order: PCNA, Ki-67, hTERT, Cyclin E, FHIT, Fas and P27kip. The system uses the above 103 patients' samples with experimentally measured features of above 7 genes and further developments of our techniques in [2-4]. In principle we rely on the data and techniques that are generated and developed by us rather than others to ensure every scientific effort is reliable, accurate and rigorous to the best of our knowledge.

A. The Ensemble Method

Recently, there has been a surge of interest in using a machine learning technique called ensemble method to enhance the performance of smart engineering systems Ensemble method is a diverse class of methods that seek to combine the decisions of several computational intelligent classifiers in order to reduce misclassifications of a classifier. This class includes:

1. Consensus networking – In this approach, the test instances are fed into several computational intelligence classifiers and majority voting of the classification decisions of these classifiers are taken.
2. Boosting – This approach is a computational intelligence machine learning meta-algorithm. At each boosting round, a "*weak"* learner is trained with the data and output of the learner is feedback to the learned function, with some strength. Then, the data is re-weighted and boosting is focused on the data that are difficult to learn in the next boosting round, so that future *"weak"* learners will attempt to reduce the mis-classification errors.
3. Bootstrap Aggregation ("Bagging") – In this approach, the original data set is sampled (with replacement) to form *M* "bags" of data, each equal in size to the original dataset; a classifier is constructed based on each of the *M* bags. Then, given an instance to be classified, it can be fed it into each of the *M* classifiers to take the majority vote of these classifiers in forming the final classification decision.

Ensemble methods have been shown to be effective at reducing the generalization error. Several issues arise in the design of such a medical decision system:

- ➢ *What types of classifiers and ensemble methods should be combined?*
- ➢ *How should they be combined?*

As to the first question, our intelligent system combines the predictions of decisions from RMCT – Recursive Maximum Contrast Trees [4], PSHNN - Parallel Self-Organizing Hierarchical Neural Networks [6] and SOFM – the new variants of Self-Organizing Feature Map Algorithms [2]. As to the second question, we are investigating a multistage classification scheme in which each stage is composed of multiple classifiers whose decisions are combined by majority voting and consensus. Instances that are misclassified by the first stage are passed to the second stage. The idea being that by only focusing on the instances misclassified by the first stage, the second stage can concentrate on the more difficult parts of the feature space and so on. Our algorithm in the intelligent medical decision system is as follows:

• First step:
– Construct two very different computational intelligence classifiers, the SOFM [2] and RMCT [4].
– Pass the test instance to both classifiers:
- If both classifiers agree, then this is the consensus prediction.
- If they disagree, this may indicate the instance is difficult to predict reliably. Then we use the second step with additions of a third classifier and a more powerful computational intelligence algorithm namely the Boosting with Bagging to break the tie.

• Second step:
– Construct an additional classifier, PSHNN [6].
– Pass the test instance to all 3 classifiers (SOFM, RMCT and PSHNN), but each classifier is also trained by Boosting with Bagging; the consensus prediction is obtained by taking the majority vote of all 3 classifiers.

Our development of new variants of Self-organizing Feature Map algorithms (SOFM) [2] is inspired by Kohonen's SOM (Self-Organizing Maps) [7] and Ersoy's PSHNN (Parallel Self-organizing Hierarchical Neural Networks) [6] algorithms but differs from the neural networks SOM algorithm by dropping the topological neighborhood and replacing it with the concept of a global neighborhood generated by ranking with significant variants denoted as variants of Self-Organizing Feature Map algorithm (SOFM). The new algorithm solves two common severe problems of SOM and many other Neural Networks (NN) algorithms.

- ➢ *Results of many NN and SOM are affected by the order in which instances are presented to the network [5-7]. For a medical decision system, we need solid robustness and accuracy in diagnosis, especially when we deal with fatal diseases.*
- ➢ *The trajectories of the neurons can oscillate wildly as a result from many NNT and SOM algorithms [5-7]. Our new SOFM solves the problems using a stepwise procedure to minimize the objective function [2].*

$$\mathrm{Q}\,(\mathrm{t}) = \frac{1}{2n} \sum_{instance\ i} \sum_{neuron\ j} m_{ij} \,||\, \vec{x}_i - \vec{W}_j(t) ||^2$$

where gradient of Q with respect to the weight vector $\vec{W}_k$ of neuron *k* is: $\nabla_{\vec{w}_k} Q = \frac{1}{n} \sum_{ins\tan ce\ i} m_{i,k} (\vec{W}_k(t) - \vec{x}_i)$

The novel batch update rule of the SOFM is thus given as: $\vec{W}_k(t+1) = \vec{W}_k(t) - \eta_t(\frac{1}{n}) \sum_{instance\ i} m_{i,k} (\vec{W}_k(t) - \vec{x}_i)$

Because the novel batch update of SOFM [2] performs gradient descent on an "averaged" error surface, the trajectories of neurons is much less variable that gradient descent rule in SOM. When used jointly with our fixed initial neuron assigned and fixed small learning rate, the SOFM algorithm is not affected by the order in which instances are presented to the network. However, whenever the SOFM and RMCT in the Consensus Networking machines give conflicting decisions, we need additional computational intelligence algorithms to break the tie. This motivated us to develop a new computational intelligence algorithm called the Boosting with Bagging that is applied to SOFM, RMCT and PSHNN for the final majority voting decision.

B. *The Boosting with Bagging*

Boosting is a computational intelligence method that can be combined with Bagging to improve the performance of a classifier. We demonstrated that when combined appropriately, Boosting with Bagging is resistant to over-fitting and the variance of the overall estimator is reduced, while the bias remains roughly the same [2-4]. Boosting with Bagging has been applied to SOFM, RMCT and PSHNN for the final majority voting decision. We are interested in incorporating useful confidence information into the intelligent system. We combine bagging with a generalization of traditional boosting algorithm that allows confidence information to be incorporated. The combined Boosting with Bagging algorithm emphasizes on *weaker* learner for each boosting run. Assuming we have *N* training instances, then we construct a classifying function $f(\vec{x}_i)$. Class label y_i is either *0* or *1*. The square error of the classifier $f(\vec{x}_i)$ is given by: $error(i) = \{f(\vec{x}_i) - y_i\}^2$. The procedure of Boosting with Bagging is described as following: Initialization: $\alpha_0 = 1; t = 1; W_i = P_i = 1/N$ where $i = 1, 2, 3, ..., N$. for $t = 1$ to T. Take *n* subsamples, choose one of subsamples that gives smallest error.

$$\varepsilon_t = \sum_{i=1}^{N} P_i^t (1 - h_{ty_i}(x_i))$$

Update coefficient α_t, weight W_i of training instance and probability P_i of instance at *t* boosting round.

$$\alpha_t = \ln(-\frac{\varepsilon_t}{1-\varepsilon_t})$$

$$W_i^{t+1} = W_i^t e^{-a_t h_{y_i}^t(\vec{x}_i)}$$

$$P_i^{t+1} = \frac{W_i^{t+1}}{\sum_{i=1}^{N} W_i^{t+1}} \qquad t = t + 1; \text{ End}$$

The confidence instance $\vec{x}$ belongs to class *k* is determined by the following equation: $\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_2, \widehat{\theta}_3, ...\widehat{\theta}_n$

The Boosting with Bagging will reduce variance error but will not affect bias error; it can be verified as following:

Assume that we want to form an estimator of a quantity based on observations. We can express the error of this estimate as the sum of a variance component and a bias component. Let us assume observations $x_1, x_2, x_3 .... x_n$ Estimator $\hat{\theta}(x_1, x_2, x_3 .... x_n)$ and corresponding true $\theta(x_1, x_2, x_3 .... x_n)$. Thus

$$Error = E[(\hat{\theta} - \theta)^2] = E[\{\hat{\theta} - E(\hat{\theta}) + (E[\hat{\theta}] - \theta)\}^2]$$
$$= E[(E[\hat{\theta}] - \hat{\theta})^2 + 2(\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta) + (E[\hat{\theta}] - \theta)^2]$$
$$= E[(E[\hat{\theta}] - \hat{\theta})^2] + E[2(\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta) + E[(E[\hat{\theta}] - \theta)^2]$$
$$= Var(\hat{\theta}) \ + \{Bias(\hat{\theta}, \theta)\}^2$$

And there are *m* observed estimators: $\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_2, \widehat{\theta}_3, ...\widehat{\theta}_n$, average their predictions to obtain an overall estimation. The variance of the overall estimate is:

$$Var(\hat{\theta}) = Var(\frac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i) = \frac{1}{m^2}(Var\sum_{i=1}^{m} \hat{\theta}_i)$$
$$= \frac{1}{m^2}\sum_{i=1}^{m} Var(\hat{\theta}_i) \cong \frac{1}{m^2} m Var(\hat{\theta}_i) = \frac{1}{m} Var(\hat{\theta}_i)$$

while the bias of the overall estimate is:

$$Bias(\overline{\hat{\theta}}, \theta) = Bias\{\frac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i - \theta\}$$
$$= Bias\{\frac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i - \frac{1}{m}\sum_{i=1}^{m} \theta\} = Bias\{\frac{1}{m}\sum_{i=1}^{m} (\hat{\theta}_i - \theta)\}$$
$$= \frac{1}{m}\sum_{i=1}^{m} Bias\{\hat{\theta}_i, \theta\} \cong \frac{1}{m} m Bias(\hat{\theta}_i - \theta)$$
$$= Bias(\hat{\theta}_i - \theta)$$

Therefore, we can see that variance of the overall estimator is reduced, while the bias remains roughly the same. The Boosting with Bagging improves the performance of the intelligent system.

TABLE III
PERFORMANCE COMPARISONS OF THE ENSEMBLE METHOD (E.M.) AGAINST DIFFERENT COMPUTATIONAL INTELLIGENCE ALGORITHMS: SOFM, SOM, RMCT, PARALLEL SELF-ORGANIZING HIERARCHICAL NEURAL NETWORKS (PSH), DECISION TREES (DT) AND SUPPORT VECTOR MACHINES (%)

	E.M	SOFM	SOM	RMCT	PSH	DT	SVM
Accuracy	95.1	94.5	88.7	93.8	89.6	88.3	87.9
Standard Deviation	2.1	2.8	4.6	2.3	3.2	5.1	4.8

VI. RESULTS, DISCUSSION AND CONCLUSION

Based on our experience, currently there is no universal effective therapy for malignant neural and endocrine tumors. There is no precise histological or pathological method to distinguish between benign and malignancies among those tumors. The treatments and prognoses are not only just quite different, but also very often determined inappropriately. Malignancies are largely unpredictable. We are part of the international efforts to search for deterministic malignant cancer markers. As our research proceeds, we found that developing synergistic bioinformatics and computational intelligence system is effective, because deterministic cancer markers do not always exist in individual patients. We are responsible to launch bioinformatics study of human genome and cancer genetics in identifying useful features in the development of the synergistic bioinformatics and computational intelligence system for the task. We will continuously using some of bioinformatics techniques we developed [1-4] to screen the human genome and other sequence data to identify more features for the task. We will improve above synergistic bioinformatics and computational intelligence system to predict malignancies of more types of tumors with following purposes:

1. To discover and identify useful features for predicting tumor biological behaviors.
2. To experimentally verify these features, and to jointly utilize them in designing medical decision systems.
3. To discover mechanisms of human genome relating malignant transformation.
4. To exploit the synergy between bioinformatics and computational intelligence.
5. To diagnose microscopic diseases and to treat cancer.

For many types of tumors, malignant transformation from mortal, normal cells to “immortal” cancer cells is often associated with the activation of telomerase and subsequent telomere maintenance. However, malignant transformation is also often associated with differential gene expressions and inactivation of one or a few tumor suppressor genes. A normal cell must maintain a completely ordered gene expressions and regulatory networks while tumor cell must not. Our parallel paradigm indicates that degree of malignancies is roughly proportional to the degree of disorder in gene expressions and regulatory networks. This concurs with theory of chaos [11] that an ordered or less disordered system can "spontaneously" go to a disordered or higher disordered system. This scenario is generally valid for "spontaneous" solid tumors, excluding the leukemia and virtual infected cancers which are not "spontaneous" tumors that benign tumors do not exist as mid-steps between normal and cancerous tissues. However, most cancers do take several steps and "long" time before the normal tissues transform into malignancies. If joint cancer associated antigens are detectable in a benign or even normal tissue, this may indicate a sign of malignant transformation. But deterministic cancer markers are not likely found in individual patients, we therefore developed this synergistic bioinformatics and computational intelligence medical decision system utilizing multiple tumor associated markers jointly in combination with the machine learning techniques we developed before [1-4]. Results showed when advantageously combine those techniques into one integrated synergistic intelligent system; the prediction power has been significantly improved to an overall accuracy of 95.1% +/- 2.1%. Benchmarks of the synergistic Ensemble Method (EM) system with a component of SOFM against other popular algorithms such as Support Vector Machines (SVM-light [19], 6.01), decision trees [18], SOM [7] and the Parallel Self-organizing Hierarchical Neural Networks (PSH [5,6]) are reported in Table 3 using a "leave one out" validation test on the 103 patients. The intelligent system use the variants of SOFM alone reached an overall 94.5% accuracy (Table 3). Because of random seeds and different order of input instances, the results may slightly different in the overall performance from one run to another even using same data of 103 patients.

The system has been put in test and validation for new patients. So far the system has reliably predicted 6 malignancies that are later confirmed by the presentations of metastasis and or extensive loan invasions. There is no conflicting report such as predicted benign but found metastasis. We still need longer follow up time to record all tumor recurrences for further validation of the system. The statistical test based on the limited available patients indicated the system has achieved a high confidence level on predicting malignancies and is generally reliable to predict malignancies of the types of neuroendocrine tumors. We will continuously search for potential markers for different types of tumors and validate those markers by using cDNA probes via quantitative RT-PCR - quantitative reverse transcriptase polymerase chain reaction, DNA microarray, FISH - Fluorescent *In situ* Hybridization and immunochemistry (if all affordable). The system is expected to be further extended to predict malignancies of different types of tumors.

The exciting results we obtained mark the beginning of further systematic research on developing more reliable and more accurate diagnostic tools utilizing the synergistic laboratory molecular biology, bioinformatics and computational intelligence. This also motivated our great interest in revealing the human genome mechanism relating to potential of cancer development from normal tissues. The research has many potential applications, not only provides a viable alternative diagnostic tool and better understanding of human genome mechanisms, but also provides useful information for better treatment planning and cancer prevention. We will further explore those applications and fulfill the task of predicting tumor malignancies.

ACKNOWLEDGMENT

We thank Drs. Laura Elintiski (NIH/NHGRI), Okan K. Ersoy, Craig W. Codrington (Purdue University), Yinfen Qin, Minyi Wei, Yan Ma, Xinghuan Liang, Decheng Lu, Jing Xian (Guangxi Medical University) and Zhiheng He (Joslin Diabetes Center, Harvard University) for valuable discussions and or experimental supports. The research was partially supported by NIH/NCI R01 CA50628-15 and NIH/NHGRI and Havard University /HMS/MGH Post Doctoral Research Specialist Fellowship Awards.

REFERENCES

[1] Zuojie Luo, Mary Qu Yang, Yan Ma, Jian;ing Li, Yinfen Qin, Minyi Wei, Xinghuan Liang, Decheng Lu, Jing Xian, Zhiheng He, Okan K. Ersoy and Jack Y. Yang "Developing Intelligent Systems for Distinguishing Benign and Malignant Tumors" *ANNIE*, pp 191-8, 2006

[2] Mary Qu Yang "Predicting Protein Structure and Function using Machine Learning Ph.D. thesis. *Purdue University*, West Lafayette, 2005

[3] Mary Qu Yang and Jack Y. Yang "IUP: Intrinsically Unstructured Proteins – A Software Tool to Analysis Polypeptide Structures" *IEEE BIBE.* pp. 3-13. 2006

[4] Mary Qu Yang, Jack Y. Yang and Okan K. Ersoy "Classifying Protein Single Labeled, Multiple Labeled with Protein Functional Classes" *International Journal General Systems*. Vol. 36 issue 1. pp.91-107, 2007

[5] Okan K. Ersoy "Parallel Self-Organzing Hierarchical Neural Networks", *IEEE Trans. Neural Networks* Vol 1, No.2 1990.

[6] W. Choe, Okan K. Erosy et.al. "Neural Network Schemes for Detecting Rare Events in Human Genomic DNA" *Bioinformatics*16, 1060-72

[7] T. Kohonen,. "Self-organizing formation of topologically correct feature maps". *Biolog. Cybernetics*,43(1):59–69, 1982.

[8] Statement of TCGA, NIH http://cancergenome.nih.gov.

[9] P. Okunieff, D. Morgan, A. Niemierko, H. Suit "Radiation dose response of human tumors". *Int J. Radiat.Oncol. Biol. Phys*. 32(4):1227-37. 1995;

[10] Douglas Hanahan, and Robert A. Weinberg "The Hallmarks of Cancer" *Cell,* Vol. 100, p.p. 57–70, Cell Press, 2000.

[11] Ian Stewart "Does God Play Dice? The Mathematics of Chaos", Penguin Books, Harmondsworth, Middlesex 1989,

[12] N.W. Kim, M. Piatyszek et.al."Specific association of human telomerase activity with immortal cells and cancer". *Science* 266: 2011-15. 1994.

[13] R. J. Michalides "Cell cycle regulators: mechanisms, role in aetiology, prognosis, treatment of cancer". *J. Clinic. Pathology* 52(8):555-68, 1999

[14] R. Lloyd et. al. "Aberrant p27kipl in endocrine and other tumors". *Am J Pathol* 150: 401-7. 1997

[15] T. Hadar et.al. "Express p53,Ki-67,Bcl-2 Parathyroid adenoma & residual normal tissue"*Path. Oncol. Res.* 11(1):45-9. 2005

[16] P. Porter et.al. Expression of cell-cycle regulators p27Kip1and cyclin E, alone & in combination, correlate with survival in young breast cancer patients. *Nature Med* 3:222-5 1997

[17] U. Alon, N. Barkai et .al. Data pertaining to the article 'Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*. 96(12): 6745-50, 1999

[18] R. Quinlan "Data Mining Tools C5.0" © *RULEQUEST RES.* 1997

[19] T. Joachims "Learning to Classify Text using Support Vector Machines" *Kluwer Academic Publishers* 2002

[20] Craig W. Codrington "Image Segmentation: A Competitive Approach". Ph.D. Thesis *Purdue Uniersity West Lafayette.*1997

[21] T. M. Nakamura, G. B. Morin et. al. "Telomerase catalytic subunit homolog from fission yeast and human". *Science* 277: 955-9. 1997