Scientific Review: A new insight on predicting tumour malignancies using synergistic computational intelligence and bioinformatics approaches

Jack Y. Yang* and Andrzej Niemierko*

Department of Radiation Oncology,
Massachusetts General Hospital and Harvard Medical School,
Harvard University,
Boston, Massachusetts 02114, USA
E-mail: jyang@bwh.Harvard.edu
E-mail: aniemierko@partners.org

Zuojie Luo*

Office of the University Provost and Dean of Academic Affairs,
Guangxi Medical University and the First Affiliated Hospital,
Nanning, Guangxi 530021, China
E-mail: zluo888@yahoo.com.cn

Mary Qu Yang*

National Human Genome Research Institute,
National Institutes of Health,
US Department of Health and Human Services,
Bethesda, MD 20852 USA and also Oak Ridge, D.O.E.
E-mail: yangma@mail.NIH.GOV
*Corresponding authors

Abstract: Recently, the National Human Genome Research Institute and National Cancer Institute, both part of NIH, US Department of Health and Human Services, have launched The Cancer Genome Atlas (TCGA). Based on the mission of TCGA, we have proposed a further parallel paradigm on cancer: it is not only the genetic changes (i.e., mutations of genes) but also changes of gene expressions and regulatory networks that are ultimately responsible for cancer development. Under this parallel paradigm, un-mutated genes with differential expressions and alternative splicing may also induce changes in the differential regulatory networks that also cause cancer when cells are subjected to unusual environments. We developed a novel synergistic computational intelligence and bioinformatics approach to predict malignancies of neuroendocrine tumours that are particularly important to discover the mechanisms of human genome mechanisms relating malignant transformation.

Keywords: computational intelligence; bioinformatics; parallel paradigm of cancer; benign tumour; malignant transformation.

Reference to this paper should be made as follows: Yang, J.Y., Niemierko, A., Luo, Z. and Yang, M.Q. (2009) 'Scientific Review: A new insight on predicting tumour malignancies using synergistic computational intelligence and bioinformatics approaches', *Int. J. Computational Intelligence in Bioinformatics and Systems Biology*, Vol. 1, No. 1, pp.4–14.

Biographical notes: Jack Y. Yang is a Harvard scientist, and chair of board of directors of *International Society of Intelligent Biological Medicine* (ISIBM). He is the Editor-in-Chief of *International Journal of Functional Informatics and Personalised Medicine*. He received his PhD and MS Degrees both from Purdue University, West Lafayette main campus and his post doctoral training was from Harvard Medical School and Indiana University School of Medicine. He also received training in biostatistics and bioinformatics from Johns Hopkins University, and in computer science from University of Illinois at Urbana-Champaign. He was trained as a combined experimental and computer scientist with more than 15 years of teaching, research and engineering practice experience in computer science and biomedical engineering. He has been an editor of more than a dozen journals and proceedings books and was the General Chair of the IEEE 7th International Conference on Bioinformatics and Bioengineering at Harvard Medical School and Co-PI of US National Science Foundation grant. He is also a consultant to IJCBS. He has published more than 100 papers. He specialises in cancer biology and artificial intelligence. http://en.wikipedia.org/wiki/Jack_Yang.

Andrzej Niemierko received his PhD from the University of Warsaw in Poland and post-doctoral training from Harvard Medical School. He was a recipient of NIH Fogarty Fellowship. He is currently Director of Division of Biostatistics and Biomathematics at Massachusetts General Hospital and Associate Professor of Radiation Oncology and Biophysics at Harvard Medical School.

Zuojie Luo received his MD from Guangxi Medical University in China. He was then a resident doctor, physician-in-charge and finally Chief Doctor and Director of Department of Endocrinology at Guangxi Medical University and First Affiliated Hospital. He spent a year as a Professor of Harvard Medical School before he took the provost position of Guangxi Medical University. Currently, he is the Provost and Dean of Academic Affairs of Guangxi Medical University, and Professor and Director of Department of Endocrinology of the First Affiliated Hospital.

Mary Qu Yang is an American Computer Scientist and Biologist, she is Editor-in-Chief of *International Journal of Computational Biology and Drug Design*. She received her PhD, MSECE and MS Degrees, all from Purdue University main campus in West Lafayette and her post-doctoral training from NIH main campus in Maryland. She also completed the research specialist training from NIH, US Department of Health and Human Services and Oak Ridge, DOE and received training in biostatistics and bioinformatics from Johns Hopkins University. She was a visiting scholar of Dr. Jun S. Liu's statistical and computational genomics laboratory of Harvard University in Cambridge. She was a recipient of the Outstanding Interdisciplinary Bilsland Dissertation Fellow for Computer Engineering (Advisor: Dr. Okan K. Ersoy) and Biophysics (Advisor: Dr. Albert W. Overhauser) Dual Degrees and NIH Fellow for the National Human Genome Research. She works in both engineering practice and translational medicine and was trained as a combined experimental and computer scientist with more than 15 years of teaching, research and engineering practice experience. She was the scientific review, advisory and steering committee chair of IEEE Bioinformatics and Bioengineering at Harvard Medical School in 2007 and advisory committee chair of Biocomp and IJCBS, as well as an honorary consulting editor of *International Journal of Functional Informatics and Personalised Medicine*, an official journal of *International Society of Intelligent Biological Medicine*. She has been an editor of a number of journals and proceedings books including *Journal of Supercomputing* (Springer Science), *International Journal of Pattern Recognition and Artificial Intelligence* (World Scientific), *IEEE Bioinformatics and Bioengineering* (IEEE), *International Journal of Bioinformatics Research and Applications* (InderScience) and *International Conference on Bioinformatics and Computational Biology*. She developed hybrid intelligent systems and contributed to the research in bidirectional promoters, cancer genomics, transmembrane proteins and their disease relevancies. She is known for Intelligent Computing Research and Education and was a recipient of a number of outstanding achievement and best paper awards including IEEE Computer Society Bioinformatics and Biomedicine Distinguished Workshop Keynote Lecturer, IEEE Bioinformatics and Bioengineering Outstanding Achievement Award, Artificial Neural Networks in Engineering Best Paper Award, World Congress on Computer Science, Computer Engineering and Applied Computing Outstanding Achievement Award. She has published more than 100 peer-reviewed scientific papers and book chapters and has delivered many invited talks including a number of keynote lectures to promote the emerging field of translational bioinformatics and personalised medicine. She specialises in genomics and machine learning.

1 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. Many cancers are resulted from genomic ‘instabilities’ that involve multiple abnormal gene expressions and regulations that are hard to identify by traditional pathological and histological analyses. Recently, the National Human Genome Research Institute and National Cancer Institute, both part of NIH, US 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. 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 analyse the genetic material from different tumours and many patients to uncover the tell-tale genetic signatures of different cancer types.” (http://cancergenome.nih.gov)

Based on the mission of TCGA, we have proposed a further parallel paradigm on cancer: it is not only the genetic changes (i.e., mutations of genes) but also changes of gene expressions and regulatory networks that are ultimately responsible for cancer development (Yang et al., 2007a; Yang and Yang, 2007). 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 induce changes in the differential regulatory networks that also cause cancer when cells are subjected to unusual environments (Yang et al., 2007; Yang and Yang, 2007). We consider that the differences between cancer and normal tissue are small in terms of their genotype, yet their biological behaviour 'phenotypes' are very different. Therefore, our approach focuses on the investigation of differential expressions of genes among normal, benign and cancerous tissues in addition to the genome-wide survey of human genome and cancer genetics.

Based on our experience, we are aimed to solve some of the limitations and problems that challenge today's cancer diagnosis and prognosis:

- *Inaccuracy***.** Deterministic cancer markers are not likely found for every individual patient. Accurate diagnosis requires verification of tumour biological behaviours and prognosis is largely unpredictable as of today. Confirmations of cancer are typically made by pathological and histological analyses, which are not always effective especially for neural and endocrine tumours. Therefore, many patients are actually diagnosed tumours of visible sizes but unknown malignancies.
- *Inefficiency.* Cancer diagnoses often involve CAT scan, MRI scan, PET, X-rays, and multiple blood and urine tests. Many patients are identified suspicious of cancer but can not be confirmed.
- *Inconsistency.* The existence of multiple but inaccurate diagnostic methods guarantees inconsistency among differential diagnoses in many cases. Without the help of a real intelligent diagnosis machine, using any of the multiple tumour associated antigens alone may conflict the diagnosis by using another one independent. Degrees of malignancies are largely unpredictable as of today.
- *Ineffectiveness.* Modern medical image technologies such as CAT, MRI, PET and X-rays all have their own problems and limitations. A detectable tumour tissue actually contains more than a 100 million tumour cells. That is about 0.1 grams of tissue in weight and a quarter inch in dimension. Those invisible tissues are commonly referred to as microscopic diseases, which can not be diagnosed. This guarantees ineffectiveness.

We are aimed to solve the above problems by developing a synergistic experimental molecular biology, bioinformatics and computational intelligent medical system. This paper is a summary and further development of the conference reports in Yang et al. (2006a, 2007a) and advantageously combines a number of computational intelligence and bioinformatics techniques we developed in Yang et al. (2006a, 2006b, 2007a, 2007b, 2007c), Yang and Laura (2007a, 2007b), Yang, (2005) and Yang and Yang (2007).

Tumour occurrence rate increases monotonically with age. The general pathway: Normal Tissue → Benign → Malignant Cancer cannot be reversed spontaneously (Note that benign may not necessarily be a mid-stage for every tumour, however we consider that a benign stage may often occur in microscopic disease, and thus is not always detectable) (Yang et al., 2006a; Okunieff et al, 2005; Yang and Yang, 2007). Cancer is not caused by a single factor. According to the "multi-stage theory of cancer" (Douglas and Weinberg, 2000), tumours can be mono- or polyclonal, especially if there's a predisposition. However, patients are not born with cancer (new born benign tumours are rare), cancerous tissue originates from one or very small number of specifically differentiated or mutated cell(s), often originated from normal cell(s). With a few exceptions, it takes years before a visible tumour can be detected by modern medical imaging technology, but malignancy often can not be confirmed by pathological and histological analyses. Malignant transformation is actually often encoded in the patient's human genome and is often associated with the expressions of a few cell proliferating associated genes and inactivation of tumour suppressor genes. A normal cell maintains a completely ordered gene expressions and regulatory networks while a tumour cell is not. We have proposed the degree of malignancy is roughly proportional to the degree of disorder in gene expressions and regulatory networks of the cell (Yang et al., 2007a; Yang and Yang, 2007). This appears plausible because an ordered system can 'spontaneously' go to a disordered system and less disordered can 'spontaneous' go a higher disordered system according to statistical physics theory called the theory of chaos (Ian, 1989). It appears that gene expressions of surrounding normal tissues may also be influenced by microscopic diseases. For example, a malignant cancer marker called Proliferating Cell Nuclear Antigens (PCNA) is detectable in certain normal tissues adjacent to some cancers. These phenomena may enable us to detect cancers of microscopic diseases.

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. Deterministic cancer markers do not likely 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. Here, we developed a synergistic computational and molecular biology and bioinformatics approach to predict potential malignancies. We have combined a number of computational intelligence ensemble methods that including Boosting, Bagging and Consensus Networking, and have designed a novel classification scheme that advantageously combines several computational intelligence algorithms, namely the variants of Self-Organising Feature Map (SOFM) algorithm (Yang, 2005; Yang et al., 2006b) and the Maximum Contrast Trees (RMCT) algorithm (Yang et al., 2007c; Yang, 2005). New computational intelligence methods such as Boosting and Bagging (Yang et al., 2007c; Yang, 2005) have been advantageously combined. When all of the above are combined into one integrated intelligent medical decision system, the prediction power of the system has been significantly enhanced (Table 1).

Table 1 Performance comparisons among ensemble method, SOFM, SOM, decision tree and SVM on test data set for malignancy prediction

Performance	*Ensemble method* (%)	*SOFM* (%)	*SOM* (%)	*Decision trees* (%)	*SVM* (%)
Average accuracy	95.2	94.6	88.9	88.6	87.8
Standard deviation	2.3	2.9	4.3	4.8	4.9

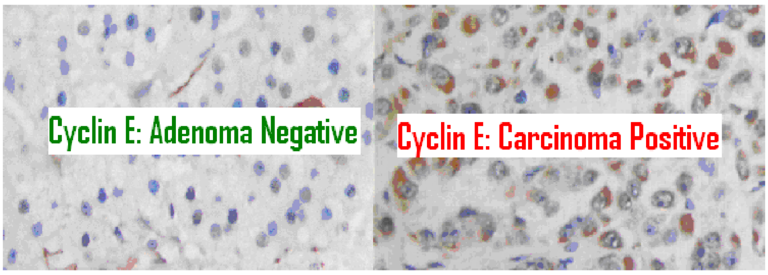
2 Joint role of tumour associated gene expressions

In general, it is very difficult to identify potential malignancies of neural and endocrine tumours based on clinical symptoms and pathological features. To conquer such difficulties, we conducted a survey of human genome and tumour genetics and have identified several tumour 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 malignancies of tumours of Cushing's syndrome and pheochromocytomas and paragangliomas (Yang et al., 2006a, 2007a). We obtained a considerable of samples of pheochromocytomas, paragangliomas, Adrenocortical Carcinoma (ACC), Adrenocortical adenoma (Aca), Adrenocortical Hyperplasia (ACH) and normal adrenal glands for measuring their tumour associated gene expression levels and use them as training samples. We are authorised to use those samples for research purpose only. Malignancy is defined as the presence of metastasis and/or extensive loan invasion. We explain roles of a few tumour markers below:

Recently, the Fragile Histidine Triad (FHIT) has been identified at chromosomal region 3p14.2. The biological function of the FHIT in the human genome has not been fully characterised yet. But FHIT has been known that deletion and differential gene expression levels of FHIT are closely associated to the malignancies and prognoses of a variety of human tumours. Therefore, FHIT can be considered as a tumour suppressor. Also recently, the Fas gene has been identified at chromosome 10q24.1. Fas is a tumour necrosis factor receptor. FasL is Fas ligand. Fas is known as TNFRSF6 and has also confusingly been known as APO-1, APT1, and CD95.

Furthermore, we investigated the role of an enzyme called telomerase in the process of tumour (Yang et al., 2006a, 2007a; Kim et al., 1994). 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). Because of the '*Hayflick Limit*' of cell culture, it appears that inactivation of P53 – a tumour suppressor and retinoblastoma proteins (pRb) eventually make cell entering into 'para' – apoptosis called 'crisis' and telomeres are often damaged, thus affecting the integrity of human genome (i.e., exposed chromosome ends may show double stranded breaks as a result of telomere shortening). Particularly, during cell divisions, many mutations and genomic instability may occur because that the fused chromosomes are randomly ripped apart. At the activation of telomerase, cells may Be 'immortal' just like tumour cells because telomerase re-activation allows cell proliferation 'forever' (technically, there are other means of reaching senescence than telomere shortening, here 'forever' means indefinitely). Our research is to investigate the telomerase activations in human genome, and to reveal essential roles in cancer development by 'immortalities' caused by telomerase. To determine hTERT mRNA expression levels, we performed in situ hybridisation using a standard clinical hTERT ISH Detection Kit (Yang et al., 2006a, 2007a). We designed the biotin-labeled cDNA probes complementary to the hTERT mRNA using verified mRNA sequences (Nakamura et al., 1997). Our precise experiments from the cDNA probes showed the high expressions of hTERT mRNA in most malignant and borderline neuroendocrine tumours. By comparison, there is no signal ever detected for hTERT mRNA expression in normal tissues. In addition, only very few benign tumours have produced any detectable signals. The results are dually confirmed by immunochemical experiments using antibodies that can indicate the expression levels of hTERT antigens. Those experimental measurements support our scenario that at re-activation of human telomerase, cells become 'immortal' – a sign of malignant transformation. Similar experiments have been conducted to determine the expression levels of cyclin E (see Figure 1), P27kip1, FHIT, Bcl-2, Bax, Fas, FasL (Fas ligand), PCNA, and Ki-67 among malignant, benign and normal tissues. Results showed clear tendencies that levels of expressions of cell proliferating related antigens such as PCNA, Ki-67, Cyclin E. and tumour related factors such as Fas, FasL increase while malignancies increase. The levels of expressions of cell arrest genes and tumour suppressor genes such as P27Kip1, FHIT decrease while malignancies increases. Bax and Bcl-2 are apoptosis related genes, it is unknown if malignant necrosis is triggered by apoptosis or not, therefore, apoptosis related factors such as Bax and Bcl-2 are not highly characteristic in our immunochemical experiments, yet they are not completely useless. Our experiments showed statistically significant tendencies in population analyses; however, individual gene expression level is not deterministic on each individual patient.

Figure 1 Immunochemical stains to measure gene expression levels of Cyclin E between benign and malignant tissues (Image Analyser for Pathology DMR+Q550) (see online version for colours)



Based on our experimental results, all of the above tumour associated markers are useful but none of the above markers are highly characteristic and none of them can be d independently to diagnose malignancies of tumours (Yang et al., 2006a, 2007a), we need to develop a novel synergistic method utilising them jointly. The advantage of developing a intelligent diagnostic tool by synergistic bioinformatics and computational intelligence methods make precise prediction of tumour biological behaviours for neuronendocrine tumours for the first time (Yang et al., 2006a, 2007a). Without such synergistic efforts, based on our experiences, physicians will continuously experience difficulties in diagnosing malignancies and prognoses of neural and endocrine tumours. Therefore we developed the world first of its kind synergistic bioinformatics and computational intelligent medical decision system to predict malignancies of tumours of Cushing's syndrome and pheochromocytomas and paragangliomas that combined the use of those tumour associated markers jointly (Yang et al., 2006a, 2007a; Yang and Yang, 2007).

3 Synergistic experimental design

Given the above analyses, the experimental design here is to develop more powerful synergistic bioinformatics and computational intelligence approaches for an intelligent diagnosis system based on our previous research in Yang et al. (2006b, 2007a, 2007c) and Yang (2005). 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 (Yang et al., 2007a; Yang, 2005; Codrington, 1997). Ensemble method is a diverse class of methods that seek to combine the decisions of several computational intelligent classifiers in order to improve the performance of the classification task. This class includes:

- *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.

- *Boosting*. This approach is a new applied mathematical technique. 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 PCNA ill attempt to reduce the classification errors.
- *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 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.

Although there are perfect mathematical proofs that ensemble methods can effective reducing the generalisation error, however, several issues arise in the design of such an intelligent system utilising ensemble method:

- *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 (Yang et al., 2007c; Yang, 2005), Parallel Self-Organising Hierarchical Neural Networks (PSHNN) Choe et al. (2000) and SOFM – the variants of Self-Organising Feature Map Algorithms (Yang, 2005; Yang et al., 2006b). 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 based on intelligent medical decision system in Yang et al. (2006a, 2007a) and Yang and Yang (2007) is as follows:

First step:

- Construct two very different computational intelligence classifiers, the SOFM (Yang, 2005; Yang et al., 2006b) and RMCT (Yang et al., 2007c; Yang, 2005).
- 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 named Boosting with Bagging to break the tie.

Second step:

- Construct an additional classifier, PSHNN (Choe et al., 2000).
- Pass the test instance to all three 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 three classifiers.

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 method called Boosting with Bagging that is applied to SOFM, RMCT and PSHNN for the final majority voting decision. Boosting can be combined with Bagging to improve the performance of a classifier. We demonstrate 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 (Yang et al., 2006b, 2007c; Yang, 2005). The intelligent medical decision system uses the developments of ensemble methods to increase accuracy of the predicting malignancies. The RMCT (Yang et al., 2007c; Yang, 2005) and PSHNN (Choe et al., 2000) are well developed methods, we recently developed SOFM (Yang et al., 2007c; Yang, 2005). Therefore, we can see that variance of the overall estimator is reduced, while the bias remains roughly the same. Boosting with Bagging significantly improve the performance of the intelligent system in predicting malignancies of tumours and diagnosing microscopic diseases. Our results showed that when all of those combined ensemble methods are integrated into one synergistic intelligent system, prediction power has been significantly improved. We benchmarked our system against other popular algorithms such as Support Vector Machines (SVM-light (Joachims, 2002) version 6.01), Decision Trees (Quinlan, 1997) and Self-Organised Maps (Kohonen, 1982) using 3-fold cross validations. Results are shown in Table 1. Because of the random seeds and different order of input instances, the performance may slightly different from one run to another, however the intelligence system that advantageous combined a number of computational intelligence techniques showed significant boost of prediction power.

4 Discussion and conclusion

As for any bioinformatics based intelligent decision system, additional features are generally beneficial to improve performance. It is thus beneficial to discover and identify more cancer markers. We will continuously search for potential markers throughout the human genome and other available data using the computational intelligence techniques we developed (Yang et al., 2006a, 2006b, 2007a; Yang and Laura, 2007a, 2007b; Yang, 2005; Yang and Yang, 2007). We will further validate those markers by experimental measurements such as using cDNA probes via quantitative RT-PCR (quantitative reverse transcriptase polymerase chain reaction) and immunochemical techniques (Yang et al., 2006a, 2007a; Yang and Yang, 2007). It is quite clear that accurate determination of malignancies of Adrenocortical Adenomas (ACA) and Carcinomas (ACC) in Cushing's syndrome and the human pheochromocytomas and paragangliomas are not effective by traditional pathological and histological analyses, we have conducted genome-wide survey of tumour associated gene expressions and we have experimentally determined the differences among normal, benign and malignant tissues using differential expression profiles of cyclin E, P27kip1, FHIT, Bax, Bcl-2, Fas, FasL, PCNA, hTERT and Ki-67. We have validated those markers by both experimental immunochemical and histopathological analyses (Yang et al., 2006a, 2007a; Yang and Yang, 2007). We developed a synergistic intelligent medical decision system to predict malignancies of ACA and ACC in Cushing's syndrome and the pheochromocytomas and paragangliomas. Our approach focus on the joint use of tumour associated gene expressions by synergistic laboratory findings, bioinformatics and computation intelligence. The exciting results we obtained and the intelligent system we designed mark the beginning of further systematic research on developing more reliable and more accurate diagnostic tools. This also motivated our great interest in revealing human genome mechanisms relating to potential malignant transformation and diagnosing cancers of microscopic diseases from genotypes to phenotypes.

Acknowledgements

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

References

Choe, W., Bina, M. and Ersoy, O.K. (2000) 'Neural network schemes for detecting rare events in human genomic DNA', *Bioinformatic*s, Vol. 16, pp.1060–1072.

Codrington, C.W. (1997) *Image Segmentation: A Competitive Approach*, PhD Thesis, University West Lafayette, Purdue.

Douglas, H. and Weinberg, R.A. (2000) 'The hallmarks of cancer', *Cell*, Vol. 100, pp.57–70.

Ian, S. (1989) *Does God Play Dice? The Mathematics of Chaos*, Penguin Books Ltd., Harmondsworth, Middlesex.

Joachims, T. (2002) *Learning to Classify Text using Support Vector Machines*, Kluwer Academic Publishers.

Kim, N.W.P., Piatyszek, M.A., Prowse, K.R., Shay, J.W. *et al.* (1994) 'Specific association of human telomerase activity with immortal cells and cancer', *Science*, Vol. 266, pp.2011–2015.

Kohonen, T. (1982) 'Self-organizing formation of topologically correct feature maps', *Biological Cybernetics*, Vol. 43, No. 1, pp.59–69.

Nakamura, T., Morin, G., Chapman, K., Weinrich, S., Andrews, W., Lingner, J. *et al.* (1997) 'Telomerase catalytic subunit homologs from fission yeast and human', *Science*, Vol. 277, pp.955–959.

Okunieff, D. Morgan, A. and Suit, N.H. (2005) 'Radiation dose response of human tumors', *International Journal of Radiation Oncology. Biological Physics*, Vol. 32, No. 4, pp.1227–1237.

Quinlan, J. (1997) 'Data Mining Tools C5.0' © *RULEQUEST RES.*

Yang, J.Y., Yang, M.Q. Ersoy, O.K., Luo, Z., Ma, Y.Y., Li, J., Qin, Y., Wei, M., Liang, X., Lu, D. Xian, J. and He, Z. (2006a) 'Developing intelligent systems for distinguishing benign and malignant tumors', *Bio-informatics and Computational Biology and Evolutionary Computation*, pp.191–198, ASME Press (ISBN 0-7918-0256-6): Based on the Best Paper Award and Smart Engineering Systems Design Award at Artificial Neural Networks in Engineering International Conference. http://web.umr.edu/~annie/annie06_Final/bpa06.htm

Yang, J.Y. and Yang, M.Q. (2007) 'Using bioinformatics and machine learning to predict potential malignancies', in Arabnia, H.R., Yang, M.Q. and Yang, J.Y. (Eds.): *Proceeding of Biocomp'07*, pp.1–6. World congress in computer science, computer engineering, and applied computing invited banquet, *Keynote Lecture: Impact and Significance of Developing Synergistic Computational Intelligence and Bioinformatics Approaches in Human Genome and Comparative Cancer Genome Research*, CSERA Press, http://www.world-academy-of-science.org/worldcomp07/ws/keynotes/keynote_banquet

Yang, J.Y., Yang, M.Q., Niemierko, A., Luo, Z. and Li, J. (2007a) 'Predicting tumor malignancies using combined computational intelligence, bioinformatics and laboratory molecular biology approaches', *IEEE CIBCB*, pp.46–53.

Yang, M.Q. (2005) *Predicting Protein Structure and Function using Machine Learning PhD Thesis with Outstanding Interdiciplinary Blisland Dissertation Fellow Award for Biological Physics and Computer Engineering Dual Degree*, Purdue University, West Lafayette

Yang, M.Q. and Laura, L.E. (2007a) 'A comprehensive study of bidirectional promoters in the human genome', *Invited Keynote Lecture with Outstanding Research Achievement Award and Educational Service Award at International Symposium of Bioinformatics Research and Applications (ISBRA 2007)*, http://www.cs.gsu.edu/ISBRA Lecture Notes in Bioinformatics, Springer, LNBI 4463 pp.361–371.

Yang, M.Q. and Laura L.E. (2007b) 'Orthology and multiple class prediction of functional elements in the human genome', *BMC Genomics*, Suppl. Issue for Biocomp'07 (Best Paper Award), http://www.world-academy-of-science.org/worldcomp07/ws/BIOCOMP07

Yang, M.Q., Laura, M.K. and Laura, L.E. (2007b) 'Comprehensive annotation of bidirectional promoters identifies co-regulation among breast and ovarian cancer genes', *PLoS Computational Biology*, Vol. 3, No. 4, doi:10.1371/journal.pcbi.0030072.

Yang, M.Q., Yang, J.Y. and Ersoy, O.K. (2006b) 'Developing new variants of the self-organizing feature map algorithms', *Infra-structure Systems Engineering, Bio-informatics and Computational Biology and Evolutionary Computation*, ASME Press, pp.183–190 (ISBN 0-7918-0256-6): based on the Second Runner Up Best Paper Award in Theoretical Developments in Computational Intelligence at Artificial Neural Networks in Engineering. http://web.umr.edu/~annie/annie06_Final/bpa06.htm

Yang, M.Q., Yang, J.Y. and Ersoy, O.K. (2007c) 'Classifying protein single labeled, multiple labeled with protein functional classes', *International Journal of General System*, Vol. 36, No. 1, pp.91–107 *Based on Novel Smart Engineering Systems Design Award at 2003 Artificial Neural Networks in Engineering International Conference*, http://web.umr.edu/~annie/bpa03.htm

Website

Statement from TCGA of NCI-NHGRI, NIH, http://cancergenome.nih.gov