International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 3 (2020), pp. 427-432

A Mathematical Methodology for Predicting the Primary Site of Metastatic Adenocarcinoma Cancer based on Rough Set Theory

Hossam A. Nabwey[1, 2]

[1]*Department of Mathematics, College of Science and Humanities in Al-Kharj, Prince Sattam bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia.*

[2]*Department of Basic Engineering Science, Faculty of Engineering, Menoufia University, Shebin El-Kom,32511, Egypt.*

https://orcid.org/0000-0002-7167-3822

Abstract

Patients with metastatic adenocarcinoma of unknown origin are a common clinical problem. Knowledge of the primary site is important for their management but histologically, such tumors appear similar. Better diagnostic markers are needed to enable the assignment of metastases to likely sites of origin on pathologic samples. The field of bioinformatics has transformed from being a discipline mainly associated with sequence databases and sequence analysis to a computational science that uses different types of data to describe biology. Thus, the ultimate goal of this research is to allow a computational simulation of cancer microarray data by propose a mathematical methodology based on Rough set theory to extracting decision rules which acts as classification scheme for prediction on biopsy material of the primary site in patients with metastatic adenocarcinoma of unknown origin. Finally a set of maximally classification rules are generated to predict origin for many tumors.

Keywords: Metastatic adenocarcinoma; tumor; bioinformatics; Rough set theory; feature selection.

1) INTRODUCTION

Molecular biology represents a very important application area for artificial intelligence and machine learning techniques in general and rough set-based methods in particular. Where the High-throughput experimental technologies have made it possible to obtain large scale data that can facilitate such research. So there is an urgent need for mathematical models, mathematical techniques and computer programs to analyze these resulting data. As well as, it is required to develop controlled vocabularies and data structures for representing knowledge in a readable form. Thus the field of bioinformatics has transformed from being a discipline mainly associated with sequence databases and sequence analysis to a computational science that uses different types of data to describe biology [1].

There is a particularly difficult challenges related to High-throughput experimental data which is inevitably obscured by a relatively large amount of noise and may be incomplete. These challenges make it especially important to choose good methods for validating the statistical and biological significance of the induced models. Furthermore, it demands a new method and theories which can deal with the uncertainty and vagueness in the data.

Rough set theory was used as a tool to reduce as well as dealing with uncertainty in datasets. Many heuristic algorithms are proposed based on rough set theory, also numerous approached based on rough set theory and other theories are investigated to extract decision rules and reduce the dimensionality of dataset [2-14].

The problem of reducing the dimensionality has been investigated for many numerous applications in different fields, since the irrelevant and redundant features in the dataset lead to low accuracy. There are two main approaches to reduce the input dimensionality, namely feature extraction and feature selection.

Standard medical classification systems for cancer tumors are based on clinical observations and the microscopical appearance of the tumor. These systems fail to recognize the molecular characteristics of the cancer that often corresponds to subtypes that need different treatment. Studying the expression levels of genes in tumor tissue may reveal such subtypes and may also diagnose the disease before it manifests itself on a clinical level [16]. Thus, the ultimate goal of this research is to allow a computational simulation of cancer microarray data to develop models for earlier detection and better understanding and treatment of cancer.

2) PROBLEM FORMULATION

The tumor is the main symptom that leads the patient to visit the doctor and from here the stage of discovery and testing of this tumor is begin. The most repeated question from the patient to his doctor is it a malignant tumor and leads to cancer or is it a secondary and benign tumor. Some 10% to 15% of cancers, however, present as metastases in solid organs, body cavities or lymph nodes. Most of these secondary tumors are adenocarcinomasو for which the seven commonest primary sites are breast, colon, lung, ovary, pancreas, prostate, and stomach [17-19]. The prognosis and therapy of patients with metastatic adenocarcinoma are linked to the site of origin, so these sites, and others, are investigated by clinical examination, radiology, and serum tumor markers.

If no primary cancer is found, then the metastatic deposit is usually biopsied, to confirm the diagnosis of malignancy and to subtype the tumor. Unfortunately, adenocarcinomas metastatic from different locations have similar microscopic appearances, which confound identification of the primary site. Patients with metastatic adenocarcinoma of unknown origin make up around 3% of all cancer patients and this category is among the 10 most common malignancies [20].

Traditionally, cancer classification has been based on clinical criteria and histopathology, which is itself based on tissue and cell morphology. For diagnostic pathology, ideally we want high coverage and high accuracy i.e., by using markers as less as possible we can correctly predict origin for many tumors and misclassify few tumors.

In this work the main goal is to represent a mathematical methodology to be a scheme for the immunohistochemical evaluation of adenocarcinomas by profiling selected candidate site specific markers to improve the prediction of primary sites.

Through bioinformatic analysis, the data were taken from literature [17] where expression profiling of 10 tumor markers including (CA125, CDX2, cytokeratins 7(CK7) , cytokeratins 20(CK20), estrogen receptor (ER), gross cystic disease fluid protein15 (GCDFP-15), lysozyme, mesothelin, prostate-specific antigen(PSA), and thyroid transcription factor1(TTF1)) are generated using tissue microarrays and immunohistochemistry. The sample size equal 352 primary tumors in four tissue microarrays, with 261 adenocarcinomas from seven main primary sites (breast, colon, lung, ovary, pancreas, prostate and stomach). Other tumors which arise at the same or nearby sites or enter the differential diagnosis were also included as shown in Table 1.

Table 1. Decision table for immunohistochemistry results for 352 primary tumors by using 10 markers

U	frequency	% Positivity of the marker										Primary site
		PSA	TTF1	GCDFP-15	CDX2	CK20	CK7	ER	mesothelin	CA125	lysozyme	
X1	35	0	0	54	0	0	83	77	3	6	14	Breast
X2	18	0	0	6	0	0	89	83	94	89	0	Ovary serous
X3	10	10	0	10	20	30	40	40	30	20	20	Ovary Mucinous
X4	10	10	0	10	0	0	60	30	50	90	10	Endometrial adenoca
X5	53	0	2	2	0	19	69	0	47	53	51	Pancreas
X6	6	0	0	17	17	83	83	0	83	67	67	Ampullary adenoca
X7	10	0	0	0	10	30	80	0	70	50	90	Cholangiocarcinoma
X8	34	3	3	0	18	18	35	0	21	9	85	Stomach
X9	21	0	0	0	10	33	48	0	19	10	57	Esophageal adenoca
X10	47	0	0	9	83	68	4	2	2	0	53	Colon
X11	18	100	11	0	0	0	0	11	0	0	6	Prostate
X12	46	0	91	4	2	2	91	9	39	39	43	Lung
X13	3	0	100	0	0	0	0	0	0	0	0	Lung small cell ca
X14	7	0	57	0	0	0	43	0	14	14	14	Lung squamous ca
X15	6	0	0	0	0	0	67	0	50	50	17	Mesotheliome
X16	7	0	0	0	0	0	0	0	0	0	0	Esophageal squ ca
X17	15	0	0	0	0	0	0	0	0	0	7	Renal cell ca
X18	6	0	0	0	0	0	0	0	0	0	0	Hepatocellular ca

3) ANALYSIS

Rough set theory is the core of the proposed methodology for extracting decision rules which acts as classification scheme for prediction on biopsy material of the primary site in patients with metastatic adenocarcinoma of unknown origin. Rough set theory has been a methodology of database mining or knowledge discovery in relational databases. In its abstract form, it is a new area of uncertainty mathematics closely related to fuzzy theory. It is a formal approximation of a crisp set defined by its two approximations – Upper approximation and Lower approximation as shown in fig. 1.

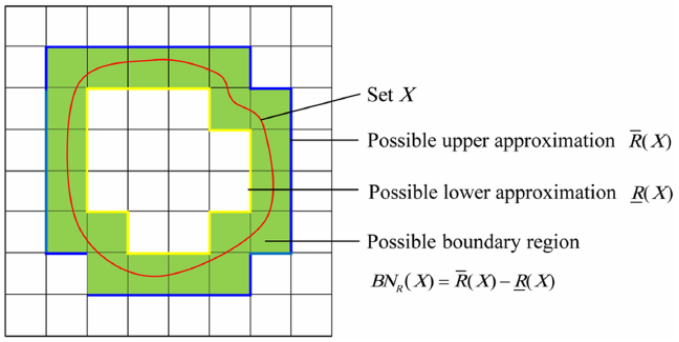

Fig. 1: representation of Upper approximation and Lower approximation

By applying the proposed methodology we will begin with discretizing the decision table given in table 1. Discretization is to set up some discretization delimiting points (also called breakpoints) in attribute values interval. In this work the rules for discretization all the attributes in decision table 1 by replacing the values of the attributes as:

- If the Positivity of the marker equal 0% then we put 0
- If the Positivity of the marker greater than zero and less than 50% then we put 1
- If the Positivity of the marker greater than or equal 50% then we put 2

As shown in Table 2.

Table 2. The discrtized Decision table of Table1.

U	% Positivity of the marker										Primary site
	PSA	TTF1	GCDFP-15	CDX2	CK20	CK7	ER	mesothelin	CA125	lysozyme	
X1	0	0	2	0	0	2	2	1	1	1	Breast
X2	0	0	1	0	0	2	2	2	2	0	Ovary serous
X3	1	0	1	1	1	1	1	1	1	1	Ovary Mucinous
X4	1	0	1	0	0	2	1	2	2	1	Endometrial adenoca
X5	0	1	1	0	1	2	0	1	2	2	Pancreas
X6	0	0	1	1	2	2	0	2	2	2	Ampullary adenoca
X7	0	0	0	1	1	2	0	2	2	2	Cholangiocarcinoma
X8	3	1	0	1	1	1	0	1	1	2	Stomach
X9	0	0	0	1	1	1	0	1	1	2	Esophageal adenoca
X10	0	0	1	2	2	1	1	1	0	2	Colon
X11	2	1	0	0	0	0	1	0	0	1	Prostate
X12	0	2	1	1	1	2	1	1	1	1	Lung
X13	0	2	0	0	0	0	0	0	0	0	Lung small cell ca
X14	0	2	0	0	0	1	0	1	1	1	Lung squamous ca
X15	0	0	0	0	0	2	0	2	2	1	Mesotheliome
X16	0	0	0	0	0	0	0	0	0	0	Esophageal squ ca
X17	0	0	0	0	0	0	0	0	0	1	Renal cell ca
X18	0	0	0	0	0	0	0	0	0	0	Hepatocellular ca

After that we compute the reduct of the discretized decision table. With the aid of software called ROSETTA which is an RST analysis toolkit we get the reducts of table 2 as shown in table 3

Table 3. Reducts of Table 2.

	Reduct	Support	Length
1	{TTF1, GCDFP-15, CK7, lysozyme}	100	4
2	{TTF1, GCDFP-15, CA125, lysozyme}	100	4
3	{TTF1, GCDFP-15, mesothelin, lysozyme}	100	4
4	{PSA, TTF1, CK20, mesothelin, lysozyme}	100	5
5	{TTF1, CK20, CK7, ER, lysozyme}	100	5
6	{TTF1, CK20, ER, CA125, lysozyme}	100	5
7	{PSA, TTF1, CK20, CA125, lysozyme}	100	5
8	{TTF1, CDX2, CK20, ER, mesothelin, lysozym	100	6

So we can bring a nearly 136 rule which can be used for classifying new cases and determine the primary site. Part of the Generated Rule Set are shown below :

Part of the Generated Rule Set

IF TTF1(0) AND GCDFP-15(2) AND CK7(2) AND lysozyme(1) => decision(Breast)

IF TTF1(0) AND GCDFP-15(1) AND CK7(2) AND lysozyme(0) => decision(Ovary serous)

IF TTF1(0) AND GCDFP-15(1) AND CK7(1) AND lysozyme(1) => decision(Ovary Mucinous)

IF TTF1(0) AND GCDFP-15(1) AND CK7(2) AND lysozyme(1) => decision(Endometrial adenoca)

IF TTF1(1) AND GCDFP-15(1) AND CK7(2) AND lysozyme(2) => decision(Pancreas)

IF TTF1(0) AND GCDFP-15(1) AND CK7(2) AND lysozyme(2) => decision(Ampullary adenoca)

IF TTF1(0) AND GCDFP-15(0) AND CK7(2) AND lysozyme(2) => decision(Cholangiocarcinoma)

IF TTF1(1) AND GCDFP-15(0) AND CK7(1) AND lysozyme(2) => decision(Stomach)

IF TTF1(0) AND GCDFP-15(0) AND CK7(1) AND lysozyme(2) => decision(Esophageal adenoca)

IF TTF1(0) AND GCDFP-15(1) AND CK7(1) AND lysozyme(2) => decision(Colon)

IF TTF1(1) AND GCDFP-15(0) AND CK7(0) AND lysozyme(1) => decision(Prostate)

IF TTF1(2) AND GCDFP-15(1) AND CK7(2) AND lysozyme(1) => decision(Lung)

IF TTF1(2) AND GCDFP-15(0) AND CK7(0) AND lysozyme(0) => decision(Lung small cell ca)

IF TTF1(2) AND GCDFP-15(0) AND CK7(1) AND lysozyme(1) => decision(Lung squamous ca)

IF TTF1(0) AND GCDFP-15(0) AND CK7(2) AND lysozyme(1) => decision(Mesotheliome)

IF TTF1(0) AND GCDFP-15(0) AND CK7(0) AND lysozyme(1) => decision(Renal cell ca)

IF TTF1(0) AND GCDFP-15(2) AND mesothelin(1) AND lysozyme(1) => decision(Breast)

IF TTF1(0) AND GCDFP-15(1) AND mesothelin(2) AND lysozyme(0) => decision(Ovary serous)

IF TTF1(0) AND GCDFP-15(1) AND mesothelin(1) AND lysozyme(1) => decision(Ovary Mucinous)

IF TTF1(0) AND GCDFP-15(1) AND mesothelin(2) AND lysozyme(1) => decision(Endometrial adenoca)

IF TTF1(1) AND GCDFP-15(1) AND mesothelin(1) AND lysozyme(2) => decision(Pancreas)

IF TTF1(0) AND GCDFP-15(1) AND mesothelin(2) AND lysozyme(2) => decision(Ampullary adenoca)

IF TTF1(0) AND GCDFP-15(0) AND mesothelin(2) AND lysozyme(2) => decision(Cholangiocarcinoma)

IF TTF1(1) AND GCDFP-15(0) AND mesothelin(1) AND lysozyme(2) => decision(Stomach)

IF TTF1(0) AND GCDFP-15(0) AND mesothelin(1) AND lysozyme(2) => decision(Esophageal adenoca)

IF TTF1(0) AND GCDFP-15(1) AND mesothelin(1) AND lysozyme(2) => decision(Colon)

IF TTF1(1) AND GCDFP-15(0) AND mesothelin(0) AND lysozyme(1) => decision(Prostate)

IF TTF1(2) AND GCDFP-15(1) AND mesothelin(1) AND lysozyme(1) => decision(Lung)

CONCLUSION

This paper suggested a methodology for predicting the primary site of metastatic adenocarcinoma cancer based on rough set theory. Data analysis led to a simplified diagnostic panel and set of decision rules are extracted which can classify a set of primary and metastatic tumors correctly. Also the obtained results are in good agreement with previous studies. The technique has been simplified logic-based rules, reduces the time and resources required to building knowledge for prediction on biopsy material of the primary site in patients with metastatic adenocarcinoma of unknown origin, leading to improved management and therapy.

ACKNOWLEDGMENTS

The author thank Prince Sattam bin Abdulaziz University, Deanship of Scientific Research at Prince Sattam bin Abdulaziz University for their continuous support and encouragement.

REFERENCES

[1] Kanehisa, Minoru, and Peer Bork. "Bioinformatics in the post-sequence era." Nature genetics 33, no. 3 (2003): 305-310.

[2] Nabwey, Hossam A. "A Hybrid Approach for Extracting Classification Rules Based on Rough Set Methodology and Fuzzy Inference System and Its Application in Groundwater Quality Assessment." In Advances in Fuzzy Logic and Technology 2017, pp. 611-625. Springer, Cham, 2017.

[3] Nabwey, Hossam A., M. Modather, and M. Abdou. "Rough set theory based method for building knowledge for the rate of heat transfer on free convection over a vertical flat plate embedded in a porous medium." In 2015 International Conference on Computing, Communication and Security (ICCCS), pp. 1-8. IEEE, 2015.

[4] Nabwey, H.A.. An approach based on Rough Sets Theory and Grey System for Implementation of Rule-Based Control for Sustainability of Rotary Clinker Kiln. International Journal of Engineering Research and Technology, Volume 12, Number 12 (2019), pp. 2604-2610

[5] Shaaban, Shaaban M., and H. Nabwey. "A decision tree approach for steam turbine-generator fault diagnosis." International Journal of Advanced Science and Technology 51 (2013): 59-66.

[6] Shaaban, Shaaban M., and Hossam A. Nabwey. "A probabilistic rough set approach for water reservoirs site location decision making." In International Conference on Computational Science and Its Applications, pp. 358-372. Springer, Berlin, Heidelberg, 2012.

[7] Shaaban, Shaaban M., and Hossam A. Nabwey. "Rehabilitation and reconstruction of asphalts pavement decision making based on rough set theory." In International Conference on Computational Science and Its Applications, pp. 316-330. Springer, Berlin, Heidelberg, 2012.

[8] Shaaban, M., and A. Nabwey. "Transformer fault diagnosis method based on rough set and generalized distribution table." Int J Intell Eng Syst 5 (2012): 17-24.

[9] Mohamed, Hossam Abd Elmaksoud. "An Algorithm for Mining Decision Rules Based on Decision Network and Rough Set Theory." In International Conference on Ubiquitous Computing and Multimedia Applications, pp. 44-54. Springer, Berlin, Heidelberg, 2011.

[10] Zhao, Hong, Ping Wang, Qinghua Hu, and Pengfei Zhu. "Fuzzy Rough Set Based Feature Selection for Large-Scale Hierarchical Classification." IEEE Transactions on Fuzzy Systems 27, no. 10 (2019): 1891-1903.

[11] Nabwey, Hossam A., and Mahdy S. El-Paoumy. "An integrated methodology of rough set theory and grey system for extracting decision rules." International Journal of Hybrid Information Technology 6, no. 1 (2013): 57-65.

[12] Pathak, H.K., George, R., Nabwey, H.A., El-Paoumy, M.S. and Reshma, K.P., 2015. Some generalized fixed point results in ab-metric space and application to matrix equations. Fixed Point Theory and Applications, 2015(1), pp.1-17.

[13] George, R., Nabwey, H.A., Reshma, K.P. and Rajagopalan, R., 2015. Generalized cone b-metric spaces and contraction principles. Mat. Vesn, 67(4), pp.246-257.

[14] Nabwey, H.A., Boumazgour, M. and Rashad, A.M., 2017. Group method analysis of mixed convection stagnation-point flow of non-Newtonian nanofluid over a vertical stretching surface. Indian Journal of Physics, 91(7), pp.731-742.

[15] Nabwey, H.A. Rough Set Approach for Analyzing the Effect of Viscoelastic and Micropolar Parameters on Hiemenz Flow in Hydromagnetics. International Journal of Engineering Research and Technology, Volume 13, Number 1 (2020), pp. 170-180

[16] Hvidsten, T.R. and Komorowski, J., 2007. Rough sets in bioinformatics. In Transactions on rough sets VII (pp. 225-243). Springer, Berlin, Heidelberg.

[17] Dennis, J.L., Hvidsten, T.R., Wit, E.C., Komorowski, J., Bell, A.K., Downie, I., Mooney, J., Verbeke, C., Bellamy, C., Keith, W.N. and Oien, K.A., 2005. Markers of adenocarcinoma characteristic of the site of origin: development of a diagnostic algorithm. Clinical Cancer Research, 11(10), pp.3766-3772.

[18] Varadhachary, G.R., Abbruzzese, J.L. and Lenzi, R., 2004. Diagnostic strategies for unknown primary cancer. Cancer: Interdisciplinary International Journal of the American Cancer Society, 100(9), pp.1776-1785.

[19] Swierczynski, S.L., Maitra, A., Abraham, S.C., Iacobuzio-Donahue, C.A., Ashfaq, R., Cameron, J.L., Schulick, R.D., Yeo, C.J., Rahman, A., Hinkle, D.A.

and Hruban, R.H., 2004. Analysis of novel tumor markers in pancreatic and biliary carcinomas using tissue microarrays. Human pathology, 35(3), pp.357-366.

[20] Pavlidis, N., Briasoulis, E., Hainsworth, J. and Greco, F.A., 2003. Diagnostic and therapeutic management of cancer of an unknown primary. European journal of cancer, 39(14), pp.1990-2005.