Applied Soft Computing 9 (2009) 703–710

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

The genetic algorithm for breast tumor diagnosis—The case of DNA viruses

Ming-Hseng Tseng [a], Hung-Chang Liao [b,*]

[a] Department of Applied Information Sciences, Chung-Shan Medical University, Taiwan
[b] Department of Health Services Administration, Chung-Shan Medical University, No. 110, Sec. 1, Jian-Koa N. Road, Taichung 402, Taiwan

ARTICLE INFO

Article history:
Received 24 March 2007
Received in revised form 17 September 2008
Accepted 28 September 2008
Available online 25 October 2008

Keywords:
Genetic algorithm (GA)
Breast tumor
DNA viruses

ABSTRACT

The purpose of this study is to determine the bioinformatics between breast tumors and DNA viruses. To extract significant factors for breast tumors, two methods – analysis of variance (ANOVA) and information measurement – are used; then, the results show that human papillomavirus (HPV) is not significant factor. Further, genetic algorithm (GA)-based data mining was submitted to obtain the rule between other DNA viruses and breast cancer/fibroadenoma. The GA rule shows that the effect of breast cancer includes {herpes simplex virus (HSV)-1 (−), human herpesvirus (HHV)-8 (−,+)}, {HSV-1(−)}, or {HHV-8(−)}, and that the effect of fibroadenoma includes {HSV-1(+), HHV-8(+)}, {HSV-1(+)}, or {HHV-8(+)}. Furthermore, the Mantel–Haenszel test shows that GA rules make a distinction between breast cancer and fibroadenoma. Finally, regarding diagnosis cost and accuracy, this study suggests that considering HSV-1 alone or with HHV-8 is optimal, but considering only HHV-8 yields less accurate results.

1. Introduction

The incidence rate of breast cancer has substantially increased (from 9.2 to 23.5 per 100,000 women) from 1985 to 1995 in Taiwan [1]. Death from breast cancer has increased from 392 to 987 during 1984–1996 [2] as well. Until now, breast cancer has been the second-most diagnosed cancer in Taiwanese women [3]. Cheng et al. [4] have studied the unique features of breast cancer in Taiwan. Their results showed that 64.1% of breast cancers were diagnosed before the age of 50 in the patients investigated, and 29.3% were diagnosed before age 40. The breast cancer patients treated are likely to be younger on average than those in the United States. Reasons for the higher incidence of breast cancer in younger Taiwanese women are still unclear. Although there exist some recognized factors that increase the risk of breast cancer, the causes are still unknown, resulting in the lack of appropriate preventive treatment [5].

Yu et al. [6] investigated the risk factors for fibroadenoma in a case–control study involving 117 fibroadenoma cases in Australia. Results from their study revealed that fibroadenoma shared some risk factors with breast cancer. DNA viruses are part of the high-risk factors closely related to human cancers. It is estimated that DNA viruses contribute 20% to human cancers. Viruses, such as specific types of herpes simplex virus (HSV)-1, Epstein–Barr virus (EBV), cytomegalovirus (CMV), human papillomavirus (HPV), and human herpesvirus (HHV)-8, emerge as major causal factors of some human cancers [7]. Wu et al. [8] studied biological purging of breast cancer cells using an attenuated replication-competent HSV-1 in human hematopoietic stem cell transplantation. Hu et al. [9] developed a second generation of genetically modified HSV-1 with paclitaxel in the treatment of breast cancer in vitro. Teshigahara et al. [10] developed an oncolytic viral therapy for breast cancer with an HSV-1 mutant, HF10. They indicated that replication-competent HSV-1 mutants held significant potential as cancer therapeutic agents. A hybrid herpesvirus infectious vector, pH300, based on HSV-1 and EBV for gene transfer to human cells in vitro and in vivo has been studied by Wang and Vos [11].

Liu et al. [12] indicated that the EBV hybridoma technique offers several advantages over the other hybridoma systems for generating anti-breast cancer human monoclonal antibodies. Katano et al. [13] established a Breast-M and an EBV-infected B-cell line (Hairy-BM) from breast tumor tissue. Yip et al. [14] used the EBV-transformed peripheral blood mononuclear cells from individuals with breast cancer for the construction of human immunoglobulin gene libraries. Fina et al. [15] studied the frequency and genome load of EBV in 509 breast cancers from various geographical areas. The results showed that the EBV genome was not correlated with age, menopausal status, tumor, size, nodal status, or histological grade. Grinstein et al. [16] demonstrated EBV in carcinomas of the breast, lung, and other sites. Xue et al. [17] also studied the EBV gene in human breast cancer. Huang et al. [18] suggested that sporadic lytic EBV infection might contribute to polymerase chain reaction (PCR)-based detection of EBV in traditionally nonvirally associated epithelial malignancies. Ribeiro-Silva et al. [19] studied whether there is a

* Corresponding author. Tel.: +886 4 24730022x17171; fax: +886 4 26474352.
E-mail address: hcliao@csmu.edu.tw (H.-C. Liao).

1568-4946/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2008.09.013

relationship between latent infection with EBV and p53 and p63 expressions in breast carcinomas. A possible role for p63 in mammary tumorigenesis associated with EBV infection was submitted. Baeyens et al. [20] compared the radiation response in EBV cell lines derived from breast cancer patients with or without a BRCA1 mutation, revealing no significant difference.

In the previous studies of CMV, Strender et al. [21] studied a group of 17 patients who had undergone modified radical mastectomy for breast cancer. Lee et al. [22] found that the caveolin expression was significantly reduced in human breast cancer cells provided that the caveolin cDNA linked to the CMV promoter was transfected into human mammary cancer cells. Still et al. [23] indicated that the constitutive expression of the gene under the control of CMV promoter in mouse fibroblasts results in cellular transformation and anchorage-independent growth. They further suggested that the inappropriate expression could impart a proliferated advantage. Svane et al. [24] analyzed the impact of high-dose chemotherapy on antigen-specific T cells responsive to CMV immunity in breast cancer patients. Their results indicated that the introduction of a preventive cancer vaccination in combination with intensive chemotherapy might be a realistic treatment option. Akbulut et al. [25] studied the cytotoxic effect of replication-competent adenoviral vectors carrying L-plastin promoter-regulated E1A and cytosine deaminase genes in cancers of breast, ovary, and colon. A similar vector driven by the CMV promoter has also been constructed as a control. This treatment resulted in decreased tumor size and decreased tumor cell growth rate. Ma et al. [26] demonstrated that the inhibition of the PKB-dependent survival pathway could promote apoptosis and thermosensitization in malignant breast cancer cells, with relative sparing of their normal counterpart. Zhu et al. [27] showed that CXCR4 had a low expression of luciferase (0.32%) compared to that of the CMV promoter in mice live in vivo. The CXCR4 was proven to be a good candidate tissue-specific promoter for cancer gene therapy for melanoma and breast cancers.

Recent studies have revealed a possible role for HPV in the pathogenesis of breast cancer. No interaction was observed between types of oral contraceptives or with any recognized risk factor for breast cancer. Oral contraceptives may act as a promoter for HPV-induced carcinogenesis [28]. Chang et al. [29] demonstrated a high frequency of abnormalities of this gene in human breast cancer. They found that there was no genomic deletion or rearrangement in spite of the presence of abnormal transcripts and no definite relationship between the abnormal transcripts and HPV infection. Liu et al. [30] showed that 6 out of 17 (35%) of the breast cancers were identified as being HPV positive in the PCR/dot blot analysis with both the HPV E6-E7 and L1 primer sets. Widschwendter et al. [31] suggested that HPV DNA might be transported from the original site of infection to the breast tissue by the bloodstream, and that it possibly existed in the carcinogenesis of breast neoplasia in some patients.

Allan et al. [32] detected two women who developed Kaposi's sarcoma in the lymphedematous arm many years after surgery for breast cancer. Kaposi's sarcoma-associated herpesvirus (KSHV, HHV-8) was suggested to be associated with breast cancer [33]. Klein and Klein [34] studied the surveillance against tumor. HHV-8 is a relevant viral agent in this context. Andres [35] found that Kaposi's sarcoma had a high incidence in the renal transplanted population, and it was related with HHV-8.

In order to determine the relationship between DNA viruses and breast tumors, the data mining technique is used in this study. Clinical databases belong to a domain where the process of data mining has become a requirement because of the rapid increase of clinical data [36]. Data mining can be used as an intelligent diagnostic tool in healthcare. Pendharkar et al. [37] demonstrated the association between statistical, mathematical, and neural approaches for mining breast cancer patterns. Wang et al. [38] investigated a simple Bayesian belief network for the diagnosis of breast cancer. They both found that neural networks (NNs) and rule induction-based models performed better overall classification than did traditional statistical methods. However, most past studies using data mining techniques on breast cancer have focused on describing individual patient characteristics, including demographic and environmental factors, but have not considered the risk of DNA viruses.

In this paper, the authors use the data mining technique of genetic algorithms (GAs) to find the bioinformatics. Applications of the GAs to numerous problems across a variety of fields have demonstrated that it is a robust and effective approach in searching very complex spaces [39,40]. GAs are mainly suitable for multi-parameter optimization problems with an objective function, depending on various hard and soft constraints [41]. Moreover, GAs generate rules that have higher accuracy and larger coverage than do inductive learning methods (e.g., CHAID algorithm) and NNs [42]. In this study, a GA method is proposed as the data mining technique for extracting knowledge rules of breast tumor diagnosis prediction. Furthermore, the extracted knowledge can be used in the control and management of the disease.

2. GA techniques

GAs are stochastic search methods for solving optimal solutions within large and complicated search spaces. The techniques are based on ideas from Darwin's theory of natural selection and "survival of the fittest" [43,44]. According to the mechanism of struggle for existence, natural selection, and the exchange of genetic information, the species with optimal fitness will govern the world. Rather than starting from a single guess within the search space, GAs are initialized with the population of guesses. These are usually random and will be spread throughout the search space. A typical algorithm uses three main operators – selection, crossover, and mutation – to improve the fitness of a population of guesses toward convergence at the global optimum [44,45].

The point of GAs is to search a proper combination of multiple parameters to achieve the greatest level of satisfaction, either minimum or maximum, depending on the requirement of the problem [40]. There are some differences between GAs and other convectional search methods [40,44]:

(1) GAs use the principles of probabilistic rules to find optimal solutions.
(2) GAs work in strings of characters representing the parameters, not the parameters themselves.
(3) GAs employ searching in a population to handle multiple parameters that search for combinations of different variables, reducing the probability of obtaining local optimum.
(4) GAs use objective function information only and are without auxiliary mathematical knowledge.

These distinct features provide the efficiency and robustness for GAs, thus resulting in advantages over convectional searching methods.

3. The discovery of knowledge rules using a GA

Although numerous experimental studies reported the usefulness of NNs in classification studies, there is a major drawback in building and using a model in which the user cannot readily comprehend the final rules that NN models acquire. An advantage of presenting an approach using GAs is that it is capable of