Research Article

Restricted Boltzmann Machines for Classification of Hepatocellular Carcinoma

James A. Koziol,¹ Eng M. Tan,¹ Liping Dai,² Pengfei Ren,² and Jian-Ying Zhang²

¹ Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA 92037, USA

² Department of Biological Sciences, University of Texas at El Paso, El Paso, TX 79968, USA

Correspondence should be addressed to James A. Koziol; koziol@scripps.edu

Received 21 November 2013; Revised 14 March 2014; Accepted 28 March 2014; Published 14 April 2014

Academic Editor: Rosalia Maglietta

Copyright © 2014 James A. Koziol et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Multiple antigen miniarrays can provide accurate tools for cancer detection and diagnosis. These miniarrays can be validated by examining their operating characteristics in classifying individuals as either cancer patients or normal (non-cancer) subjects. We describe the use of restricted Boltzmann machines for this classification problem, relative to diagnosis of hepatocellular carcinoma. In this setting, we find that its operating characteristics are similar to a logistic regression standard and suggest that restricted Boltzmann machines merit further consideration for classification problems.

1. Introduction

We have previously investigated the utility of antibody profiles to seven tumor-associated antigens (TAAs) for discriminating between cancer patients and controls [1]. We found that these multiple antigen miniarrays could provide accurate tools for cancer detection and diagnosis and suggested that performance of the miniarrays might be enhanced by other combinations of TAAs appropriately selected for different cancer cohorts.

We return to this theme in the present paper, where we examine the utility of an expanded panel of 12 antibody profiles for cancer diagnosis of hepatocellular carcinoma (HCC), based on serum samples from newly diagnosed HCC patients and normal controls. To this end, we apply a fairly recent approach, restricted Boltzmann machines [2–5], to the classification problem at hand. For comparative purposes, we also utilize logistic regression, to provide a baseline of discriminative performance. Our aim here is to determine whether restricted Boltzmann machines provide a viable technique for classification of HCC, compared to the logistic regression standard.

2. Materials and Methods

2.1. Sera Samples. In all, sera samples from 175 HCC patients and 90 normal controls were amassed, as follows. Sera from 76 patients with HCC from Xiamen in China were obtained from the serum bank of the Cancer Autoimmunity Research Laboratory at the University of Texas (El Paso, Texas, USA), which were originally provided by a collaborator in Sun Yat-sen University (Guangzhou, China). 84 HCC patients' sera were collected from Korea and 15 from Japan. Ninety normal human sera (NHS) were originally obtained from the serum bank of the Autoimmune Disease Center at the Scripps Research Institute (La Jolla, CA, USA). All cancer patients were diagnosed according to established criteria; their serum samples were collected at the time of initial cancer diagnosis, when the patients had not received treatment with any chemotherapy or radiation therapy. Normal human sera were collected from adults during annual health examinations in people who had no obvious evidence of malignancy. This study was approved by the Institutional Review Board of the University of Texas at El Paso and collaborating institutions.

2.2. Expression and Purification of CAPER α Recombinant Protein. Twelve antigens, IMP1, p62, Koc, p53, survivin, p16, HCC1, p90, RalA, NPM1, MDM2, and 14-3-3 ζ , were expressed as recombinant proteins. Recombinant p62 was expressed from a clone isolated from a cDNA expression library by immunoscreening with antibody from a patient with HCC [6]. p62 cDNA was subcloned into pET28a vector producing a fusion protein with NH-terminal 6x histidine and T7 epitope tags. The recombinant protein expressed in *Escherichia coli* BL21 (DE3) was purified using nickel column chromatography. Koc cDNA cloned into the pcDNA3 vector [7] was similarly subcloned into pET28a vector and recombinant protein expressed as above. Imp1 construct pCMV5-Imp1 was kindly provided by Nielsen et al. [8] and p53 clone (p53SN3) by Yuxin Yin of Columbia University, New York, NY, and subcloned into pET28a for protein expression. Survivin cDNA was amplified from human survivin EST clone (BG258433) before subcloning in pET28a vector. Plasmid pET-HCC1 carrying HCC1 cDNA was derived from the previous study [9]. cDNA encoding RalA amplified by PCR from a human expressed sequence tag (EST) clone (#BM560822) was subcloned into pET28a vector. NPM1 construct GFP-NPM WT (plasmid ID: 17578), MDM2 construct pGEX-4T MDM2 WT (plasmid ID: 16237), and 14-3-3 ζ construct GST-14-3-3 WT (plasmid ID: 1944) purchased from Addgene were subcloned into pET28a vector. p16 cDNA was amplified by RT-PCR from human HeLa cells and was subcloned into the pGEX vector expressing p16 with glutathione S transferase (GST) fusion partner. The GST gene fusion system was used for the expression and purification of p16 recombinant protein [10]. p90 has been cloned from a cDNA expression library by immunoscreening with antibody from a patient with gastric cancer and then subcloned into pGEX [11]. Expression of adequate amounts of recombinant protein was examined in SDS-PAGE and Coomassie blue staining was used to determine that expression products of expected molecular sizes were produced. In addition, western blot analysis was used to confirm that the bands seen in SDS-PAGE were reactive with reference antibodies.

2.3. Enzyme-Linked Immunosorbent Assay (ELISA). Purified recombinant proteins were diluted in phosphate-buffered saline (PBS) to a final concentration of 0.5 $\mu\text{g/mL}$ for coating 96-well Immunolon2 microtiter plates (Fisher Scientific, Huston, TX, USA) overnight at 4°C. The human serum samples diluted 1:200 were incubated in the antigen-coated wells for 90 min. Horseradish peroxidase- (HRP-) conjugated goat anti-human IgG (Santa Cruz Biotechnology, Inc., Santa Cruz, CA, USA) was used as the secondary antibody at 1:4,000 dilution and, after reaction for 90 min, the wells were washed with PBS containing 0.05% Tween 20. The substrate 2,2'-azino-bis(3-ethyl-benzothiazoline-6-sulfonic acid) diammonium salt (ABTS) (Sigma, St. Louis, MO, USA) was used as the detecting agent. The optical density (OD) value of each well was read at 405 nm, and the cutoff value for determining a positive reaction was designated as the mean absorbance of the 90 normal sera plus 3 standard deviations (mean + 3 SDs). Each sample was tested in duplicate. Each run of ELISA

included 8 NHS representing a range of absorbance above and below the mean of 89 normal human sera, and the average OD value of 8 NHS was used to normalize all absorbance values to the standard mean of the entire 90 normal samples. All positive sera were further confirmed by western blotting.

2.3.1. Data Preparation. We prepared two datasets for further consideration, as follows. ELISA results are typically dichotomized by selection of a threshold or cutoff absorbance value: an observed absorbance value above the threshold connotes a positive reaction and below the threshold a negative reaction. We chose cutoffs of normal means + 3 standard deviations (SDs) on all of the antibody assays and derived the corresponding dataset wherein input for each individual consisted of dichotomous variables "positive" or "negative" on each antibody assay, along with group membership (cancer or control).

The normal mean + 3 SDs cutoff has conventionally been used to distinguish abnormal or positive from normal or negative in antibody assays, so this dichotomized dataset should lead to benchmarks of standard performance. Nevertheless, choosing cutoffs in terms of means plus multiples of SDs does discard some information, and it is possible that operating characteristics of our classifiers might be improved by utilizing actual absorbance values on each of the antibody assays for each individual rather than the dichotomous positive or negative determinations. Since absorbance scales differ for the various assays, we first normalized the data by transforming each assay's absorbance values into percentiles of the Gaussian distribution fitted to the absorbance values of the controls. With this normalization, we derived our second dataset, in which input for each individual consisted of normalized values between 0 and 1 on each antibody assay, along with group membership.

2.3.2. Statistical Methods. We used two methods for the classification problem, logistic regression and restricted Boltzmann machines (RBMs). Logistic regression is an established technique, whereas restricted Boltzmann machines are somewhat novel. Briefly, a restricted Boltzmann machine is a machine learning technique popularized by Hinton [2]. An RBM has the same generic form as a single hidden layer neural network. It consists of one layer of visible units and one layer of hidden units, with symmetrically weighted connections between the visible and the hidden units, but with no connections within each layer of units. Originally, RBMs were proposed with both visible and hidden units representing binary states, but generalizations to continuous visible and hidden units are available [3]. We utilized a discriminative training algorithm [12, 13] for fitting RBMs to training data and assessed performance on validation data as described next.

For both RBMs and logistic regression, we used 10-fold cross-validation for determination of operating characteristics of the classifiers. The entire data set was randomly divided into 10 equally sized subsets, and the classifiers were trained on 9 subsets and tested with the remaining subset (the validation subset). This cross-validation was repeated 9

additional times (so that each of the 10 subsets served as the validation subset once only), and the results on the test sets combined to calculate the predictive accuracy and error rates of each classifier. We remark that the initial random division was stratified, to preserve the ratio of cancer cases to controls (normals) in each subset.

All computations were undertaken in MATLAB R2012b (The Mathworks Inc., Natick, Massachusetts, USA). Cross-validation samples were generated with the crossvalind procedure in MATLAB. Restricted Boltzmann machines were fit using MATLAB `matrbm` code available at <https://code.google.com/p/matrbm/>, with parameter adjustments as suggested by Hinton [14]. Logistic regressions were implemented with the `glmfit` procedure in MATLAB.

3. Results

We begin with the dichotomized data for all 12 antibodies (normal mean + 3 SDs cutoffs as described in Statistical Methods). With these data, logistic regression achieved an overall sensitivity of .697 (range .588 to .824) and specificity of .811 (range .556 to 1.0), these statistics being obtained from test samples based on 10-fold cross-validation. In comparison, restricted Boltzmann machines with 12 hidden nodes achieved an overall sensitivity of .720 (range .611 to .824) and specificity of .800 (range .556 to 1.0) on these same test samples.

Next, we investigated whether operating characteristics could be improved by replacing the dichotomies with continuous responses, in particular, normalization, as described in Materials and Methods. With the normalized data for all 12 antibodies, logistic regression achieved an overall sensitivity of .909 (range .824 to 1.0) and specificity of .789 (range .667 to 1.0), these statistics being obtained on the test samples from 10-fold cross-validation as before. In comparison, restricted Boltzmann machines with 12 hidden nodes achieved an overall sensitivity of .926 (range .824 to 1.0) and specificity of .700 (range .333 to .889). Sensitivities of both classifiers improve dramatically with the normalized data. There is some attenuation of specificity with logistic regression on the normalized data and a much greater drop-off in specificity with RBMs. Summary values are given in Table 1.

Note that there is some flexibility with restricted Boltzmann machines, namely, the number of hidden nodes. We examined whether the performance of RBMs is dependent on the choice of number of hidden nodes, by repeatedly fitting RBMs to the dichotomized and normalized datasets, while systematically varying the number of hidden nodes from 2 to 16. For each fit, we obtained summary estimates of sensitivity and specificity from 10-fold cross-validation. We plot these estimates versus number of hidden nodes in Figure 1(a) sensitivity estimates, Figure 1(b) specificity estimates. It turns out that the performance of RBMs in this setting seems rather insensitive to the number of hidden nodes.

We then addressed the question of whether the performance of logistic regression or restricted Boltzmann machines could be improved by judicious selection of input

TABLE 1: Operating characteristics of logistic regression (LR) and restricted Boltzmann machine (RBM) classifiers for hepatocellular carcinoma (HCC), using individual profiles of 12 autoantibodies. Sensitivities and specificities were determined from 10-fold cross-validation of the sample of 175 HCC patients and 90 normal controls.

	Dichotomized data		Normalized data	
	Sensitivity	Specificity	Sensitivity	Specificity
LR	0.697	0.811	0.909	0.789
RBM	0.720	0.800	0.926	0.700

TABLE 2: Operating characteristics of logistic regression (LR) and restricted Boltzmann machine (RBM) classifiers for hepatocellular carcinoma (HCC), using individual profiles of either 5 (normalized data) or 6 (dichotomized data) autoantibodies. Sensitivities and specificities were determined from 10-fold cross-validation of the sample of 175 HCC patients and 90 normal controls.

	Dichotomized data		Normalized data	
	Sensitivity	Specificity	Sensitivity	Specificity
LR	0.646	0.889	0.903	0.789
RBM	0.731	0.722	0.949	0.578

variables. We implemented the selection procedure via step-wise logistic regressions, to arrive at models in which variables were included if and only if they were jointly statistically significant at the 0.10 level. From the dichotomized data, 6 antibodies were thus selected: Koc, NPM1, p53, p62, p90, and RalA. From the normalized data, 5 antibodies were so selected: HCC1, Koc, p16, p90, and survivin. We then reduced each dataset to these respective subsets of antibodies and refit logistic regressions and RBMs. With the reduced dichotomized data, logistic regression achieved an overall sensitivity of .646 (range .530 to .765) and specificity of .889 (range .778 to 1.0), from the test samples in 10-fold cross-validation as before. In comparison, RBMs with 12 hidden nodes achieved an overall sensitivity of .731 (range .556 to 1.0) and specificity of .722 (range .556 to 1.0). With the reduced normalized data, logistic regression achieved an overall sensitivity of .903 (range .778 to 1.0) and specificity of .789 (range .556 to 1.0). In comparison, RBMs with 12 hidden nodes achieved an overall sensitivity of .949 (range .882 to 1.0) and specificity of .578 (range .222 to .778). Summary values are given in Table 2.

4. Discussion

In a previous study [1], we reported that multiple antigen miniarrays could serve as useful tools for cancer detection and diagnosis. The utility of autoantibodies in cancer diagnosis was demonstrated there, because of the typical absence of elevated or depressed levels of particular autoantibodies in normal individuals. In addition, we had previously proposed [15, 16] that autoantibodies might successfully be used as indicators of aberrant cellular mechanisms in tumorigenesis. In the oncologic setting, we further suggest that autoantibody panels might be adopted as predictive markers; that is, they

FIGURE 1: Operating characteristics of restricted Boltzmann machines (RBMs) for classification of hepatocellular carcinoma. (a) Sensitivity. (b) Specificity. Serum samples were obtained from 90 normal subjects and 175 newly diagnosed HCC patients. ELISA levels of 12 cancer-associated antibodies were obtained for each individual; as described in the Materials and Methods, the data were then either normalized or dichotomized. Numbers of hidden nodes in the RBMs were systematically varied from two through 16, and associated sensitivities and specificities were determined from 10-fold cross-validation of the normalized or dichotomized data. Sensitivities of the RBMs remain unchanged as the number of hidden nodes increases; and dichotomization of the data results in a loss of sensitivity compared to the normalized (continuous) data. Specificities of the RBMs show more variability than sensitivities; in addition, dichotomization of the data seems to improve the specificities of the RBMs compared to the normalized data.

might provide diagnostics for identifying treatment of choice in subgroups of patients based on phenotypic characteristics.

In this current study, we examined the utility of an expanded panel of 12 antibody profiles for cancer diagnosis of hepatocellular carcinoma, using restricted Boltzmann machines as our classifier. Restricted Boltzmann machines are a recent machine learning technique popularized by Hinton and Salakhutdinov at the University of Toronto [2, 3]. Fischer and Igel [4, 5] provide readable introductions to RBMs for the uninitiated. Use of RBMs for classification and discrimination was proposed by Larochelle and Bengio [12], and we utilize one of their suggested methods in our setting, relating to accurate diagnosis of hepatocellular carcinoma on the basis of a panel of 12 antibody profiles.

We found that overall operating characteristics of RBMs were comparable to logistic regression: RBMs typically had greater sensitivities, but smaller specificities, than logistic regression with the same input variables. There were dramatic increases in the sensitivities of the logistic regression and RBM classifiers with the normalized data from the panel of 12 antibody profiles compared with the dichotomized data, to levels exceeding 0.9. The tradeoff here was diminution in specificities. Dichotomization at a smaller cutoff level than mean + 3 SDs (e.g., mean + 2 SDs) would similarly achieve increased sensitivity, but decreased specificity, relative to the mean + 3 SDs cutoff reported here.

We have reported operating characteristics of the logistic regression and RBM classifiers solely with test data from our cross-validation procedure. Operating characteristics

were typically about .05 higher when the classifiers were assessed on the training data, though this represents an overly optimistic assessment of classifier performance. As with other classifiers, logistic regression has its limitations. Our use of logistic regression as a comparator is motivated by its ease of interpretation and relatively good performance in our previous studies; as well, both cases and controls are numerous and not totally unbalanced, and our explanatory variables are not highly correlated.

We remark that our interest in fixed cutoffs is motivated primarily by historical precedent, that is, by the long history of their usage in immunoassay assessment to demark normality from abnormality. In immunological bioassay systems, it has been repeatedly observed that there are a few “normal” controls who are high responders. There are many explanations for this, including nonspecific binding reactions with antigen and the possibility that, among a “normal” population, there are preclinical cases of individuals who have specific antibody. The long accepted approach to this problem is to establish a cutoff at +2 or +3 SDs of the mean of “normals” as the differential marker, with the former showing higher sensitivity but lower specificity and the latter lower sensitivity but higher specificity. We chose +3 SDs of the mean as the cutoff in our studies and have been using this in all our previous studies. Theoretically, at least, dichotomization of continuous variables engenders an information loss and is to be deplored (e.g., [17, 18]). In the present setting, dichotomization resulted in a drop in sensitivities, but an increase in specificities, of the classifiers

relative to the continuous data, representing a tradeoff rather than a clear win for the continuous data.

Reduction in the number of input variables was affected by means of a logistic regression variable selection technique. The operating characteristics of the logistic regression classifier with continuous data were unaffected by this variable reduction; with the dichotomized data, sensitivity decreased slightly, but specificity improved. Under variable reduction, sensitivities of the RBM classifier improved slightly, but specificities declined. As with support vector classifiers, variable selection is not inherent with restricted Boltzmann machines; and, judging from our investigation of the number of hidden units, overfitting of the RBMs does not seem to be a cause of concern in our setting.

We find it encouraging that, in this particular setting, restricted Boltzmann machines are competitive with logistic regression classifiers. A priori, we would expect discrimination between cancer cases and controls on the basis of threshold values of tumor-associated antigens to be quite reasonable; such monotone patterns are easily modeled with logistic regression. In addition, restricted Boltzmann machines might also capture interaction patterns automatically, which might be especially advantageous were we to expand the predictors to include demographic or other information. Note, however, that our datasets are rather sparse compared to the large datasets typically studied with restricted Boltzmann machines. Their relatively good performance in our classification problem is promising, and RBMs merit further consideration in similar investigations. We encourage others to explore further the utility of the RBM approach in similar settings.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors thank the reviewers for their perceptive comments. This study was supported in part by National Institute of Health Grant AG007996 (JAK).

References

- [1] J. A. Koziol, J.-Y. Zhang, C. A. Casiano et al., "Recursive partitioning as an approach to selection of immune markers for tumor diagnosis," *Clinical Cancer Research*, vol. 9, no. 14, pp. 5120–5126, 2003.
- [2] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.
- [3] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [4] A. Fischer and C. Igel, "An introduction to restricted Boltzmann machines," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, L. Alvarez, Ed., vol. 7441 of *Lecture Notes in Computer Science*, pp. 14–36, Springer, Berlin, Germany, 2012.
- [5] A. Fischer and C. Igel, "Training restricted Boltzmann machines: an introduction," *Pattern Recognition*, vol. 47, pp. 25–39, 2014.
- [6] J.-Y. Zhang, E. K. L. Chan, X.-X. Peng, and E. M. Tan, "A novel cytoplasmic protein with RNA-binding motifs is an autoantigen in human hepatocellular carcinoma," *The Journal of Experimental Medicine*, vol. 189, no. 7, pp. 1101–1110, 1999.
- [7] F. Müller-Pillasch, U. Lacher, C. Wallrapp et al., "Cloning of a gene highly overexpressed in cancer coding for a novel KH-domain containing protein," *Oncogene*, vol. 14, no. 22, pp. 2729–2733, 1997.
- [8] J. Nielsen, J. Christiansen, J. Lykke-Andersen, A. H. Johnsen, U. M. Wewer, and F. C. Nielsen, "A family of insulin-like growth factor II mRNA-binding proteins represses translation in late development," *Molecular and Cellular Biology*, vol. 19, no. 2, pp. 1262–1270, 1999.
- [9] H. Imai, E. K. L. Chan, K. Kiyosawa, X.-D. Fu, and E. M. Tan, "Novel nuclear autoantigen with splicing factor motifs identified with antibody from hepatocellular carcinoma," *The Journal of Clinical Investigation*, vol. 92, no. 5, pp. 2419–2426, 1993.
- [10] K. Looi, R. Meghiorino, F.-D. Shi, X.-X. Peng, Y. Chen, and J.-Y. Zhang, "Humoral immune response to p16, a cyclin-dependent kinase inhibitor in human malignancies," *Oncology Reports*, vol. 16, no. 5, pp. 1105–1110, 2006.
- [11] L. S. Hoo, J. Y. Zhang, and E. K. L. Chan, "Cloning and characterization of a novel 90 kDa "companion" auto-antigen of p62 overexpressed in cancer," *Oncogene*, vol. 21, no. 32, pp. 5006–5015, 2002.
- [12] H. Larochelle and Y. Bengio, "Classification using discriminative restricted boltzmann machines," in *Proceedings of the 25th International Conference on Machine Learning*, pp. 536–543, ACM, New York, NY, USA, July 2008.
- [13] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: a comparison of logistic regression and naïve Bayes," in *Advances in Neural Information Processing Systems*, T. G. Diettrich, S. Becker, and Z. Ghahramani, Eds., pp. 841–848, MIT Press, Cambridge, Mass, USA, 2001.
- [14] G. E. Hinton, "A practical guide to training restricted Boltzmann machines," Tech. Rep. UTML TR 2010-003, Department of Computer Science, University of Toronto, 2010, <http://www.csri.utoronto.ca/~hinton/absps/guideTR.pdf>.
- [15] E. M. Tan, "Autoantibodies as reporters identifying aberrant cellular mechanisms in tumorigenesis," *The Journal of Clinical Investigation*, vol. 108, no. 10, pp. 1411–1415, 2001.
- [16] E. M. Tan and J. Y. Zhang, "Autoantibodies to tumor-associated antigens: reporters from the immune system," *Immunological Reviews*, vol. 222, no. 1, pp. 328–340, 2008.
- [17] O. Naggara, J. Raymond, F. Guilbert, D. Roy, A. Weill, and D. G. Altman, "Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms," *American Journal of Neuroradiology*, vol. 32, no. 3, pp. 437–440, 2011.
- [18] C. Bennette and A. Vickers, "Against Quantiles: categorization of continuous variables in epidemiologic research, and its discontents," *BMC Medical Research Methodology*, vol. 12, article 21, 2012.