Amino Acids (2015) 47:217–221
DOI 10.1007/s00726-014-1871-3

SHORT COMMUNICATION

HBC-Evo: predicting human breast cancer by exploiting amino acid sequence-based feature spaces and evolutionary ensemble system

Abdul Majid · Safdar Ali

Received: 15 May 2014 / Accepted: 4 November 2014 / Published online: 10 December 2014

Abstract We developed genetic programming (GP)-based evolutionary ensemble system for the early diagnosis, prognosis and prediction of human breast cancer. This system has effectively exploited the diversity in feature and decision spaces. First, individual learners are trained in different feature spaces using physicochemical properties of protein amino acids. Their predictions are then stacked to develop the best solution during GP evolution process. Finally, results for HBC-Evo system are obtained with optimal threshold, which is computed using particle swarm optimization. Our novel approach has demonstrated promising results compared to state of the art approaches.

Keywords Breast cancer diagnosis · Protein sequences · Amino acids · Physicochemical properties · Genetic programming · Evolutionary ensemble system

Introduction

Human breast cancer is the second major cause of worldwide cancer-related deaths. Both men and women are afflicted by breast cancer though it is common in women (Good et al. 2014). Similar to other cancers, it could be effectively treated if diagnosed in early stages. Molecular signatures of breast cancer are predicted using basic knowledge from biosciences-related fields of proteomics, genomics, etc. Generally, most of the mutations that affect genes result in protein sequences with truncations, insertions or deletions. This change in protein sequences might be useful for drug discovery and cancer prediction. In our view, descriptors generated using physicochemical properties of amino acids of protein related to cancer could be very beneficial for the diagnosis of breast cancer. These properties are utilized in the study of protein interactions, structures, and sequence-order effects (Li et al. 2009; Jiang et al. 2008; Maqsood et al. 2012; Khan et al. 2011).

In the literature, researchers have proposed ensemble approaches for the early diagnosis, prognosis and prediction for human breast cancer. Xin et al. (2012) have used random forest (RF) algorithm for prediction of DNA-binding residues in protein sequences. Aminzadeh et al. (2011) have applied RotBoost ensemble using microarray genes and achieved 94.39 % accuracy. Ruxandra and Stoean (2013) have employed evolutionary algorithms (EAs) with the combination of support vector machines (SVM) and reported accuracy up to 97.07 %. Lavanya, and Rani (2012) have used bagging and boosting ensemble approaches for ensemble-based decision-making system. They have achieved accuracies up to 97.85 and 95.56 %, respectively. Mostly, ensemble approaches have limited performance due to small number of biological samples and class imbalance. These approaches do not effectively combine diverse classifiers that are individually trained in different feature spaces. On the contrary, our proposed GP-based (HBC-Evo) system has exploited diversity in feature and decision spaces to yield more accurate results.

Electronic supplementary material The online version of this article (doi:10.1007/s00726-014-1871-3) contains supplementary material, which is available to authorized users.

A. Majid (✉) · S. Ali
Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad 45650, Nilore, Pakistan
e-mail: abdulmajiid@pieas.edu.pk

S. Ali
e-mail: safdarali@pieas.edu.pk

Proposed HBC-EVO system

The proposed HBC-Evo system is evaluated using two datasets of human protein primary sequences for cancer/non-cancer (C/NC) and breast/non-breast cancer (B/NBC). These datasets are retrieved from Munteanu et al. (2009). They have composed 1,056 protein primary sequences. The C/NC dataset is comprised of 865 non-cancer and 191 cancer-related proteins sequences, whereas the B/NBC dataset is comprised 865 non-cancer and 122 breast cancer-related protein sequences.

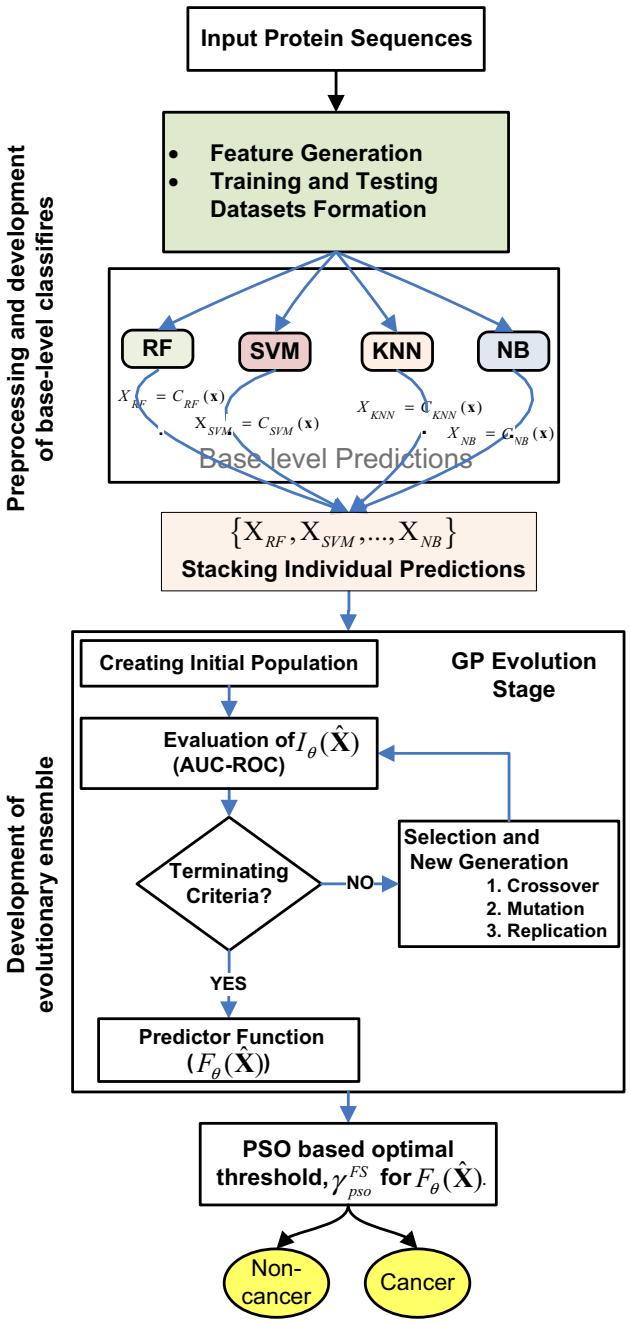

Fig. 1 Basic architecture of the HBC-Evo system

The basic architecture of the proposed HBC-Evo system is shown in Fig. 1. The objective of this work is the early diagnosis/prediction of human breast cancer using protein amino acid molecules associated with cancer. Protein sequences are essential to almost all bodily functions. The chief function of amino acids is to act as the building blocks of protein. Amino acids could be connected jointly in varying form of linear chains, i.e., in sequences, to form enormous types of protein molecules. Proteins of a tissue would reflect the initial changes caused by the successive genetic mutations, which lead to cancer. Such changes/mutations would be exploited for the early diagnosis of breast cancer. Supplementary Fig. S1 illustrates the change in the composition of the amino acid molecules of the cancerous proteins with respect to non-cancerous proteins. It is observed that more disturbances occurred in P, S, Y, C, R, and N amino acid molecules. These molecules would offer high discrimination power between cancerous and non-cancerous proteins. To incorporate this discriminant feature, the molecular descriptors of amino acid sequences are formed using numerical values of their physiochemical properties of hydrophobicity and hydrophilicity (see Supplementary Table S1). These descriptors information are then transformed into well-known feature spaces of amino acid composition (AAC), split amino acid composition (SAAC), pseudo-amino acid composition-series (PseAAC-S), and pseudo-amino acid composition-parallel (PseAAC-P) using statistical and/or mathematical methods. Detailed information of these feature spaces is available in Majid et al. (2014) and Safdar et al. (2014). For data balancing, we have employed mega-trend diffusion technique that oversamples the minority class in feature space. The predicted information of base learners is extracted using naïve bayes (NB), k-nearest neighbor (KNN), SVM, and RF algorithms. In the second phase, to exploit decision space, the predictions of individual classifiers are stacked to develop HBC-Evo during GP evolution process. In training process, GP algorithm has automatically extracted useful discriminant features (attributes). Particle swarm optimization (PSO) algorithm is applied to find optimal threshold for the best individual developed at the end of GP run.

A clinician could easily utilize our HBC-Evo system for the diagnosis of cancer using protein sequences of the affected tissue. The protein sequences of amino acids can be obtained from DNA sequence, mass spectroscopy, Edman degradation, etc. These protein sequences could simply be fed to HBC-Evo system. If the protein is related to cancer, i.e., pattern of amino acids in protein is changed, it will diagnose cancer patient, otherwise non-cancer patient. Once the cancer is diagnosed, the clinician may

proceed further by recommending other standard tests to know the severity of the cancer.

The input dataset of protein sequences is randomly divided into two different parts for training (Trn) and testing (Tst). On the average, two-third data are used for training and one-third data are remaining for model testing. Base classifiers are trained for Trn data and new meta-data (Trn_meta) are obtained for the development of HBC-Evo. These base predictions are then stacked for HBC-Evo. The Tst data are employed to evaluate the base classifiers and thus new meta-data (Tst_meta) are obtained for the evaluation of the system.

For m base-level classifiers $\{C_1, C_2, \ldots, C_m\}$, consider a set of N training data samples (Trn) of protein sequences, $S_t = \left\{\left(\mathbf{x}^{(n)}, t^{(n)}\right)\right\}_{n=1}^{N}$, where $\mathbf{x}^{(n)}$ indicates the nth feature vector correspond to target $t^{(n)}$. Using these base-level classifiers, for each ith examples, we obtained a set of m predictions $\left\{X_1^i, X_2^i, \ldots, X_m^i\right\}$. Therefore, at meta-level training, we obtained m-dimensional feature vector $\hat{\mathbf{X}}_{\mathrm{m}} = (X_1, X_2, \ldots, X_{\mathrm{m}})$. Consequently, for GP training, a Trn_meta data of N samples is formed, i.e., $S_d = \left\{\left(\hat{\mathbf{X}}^{(n)}, t^{(n)}\right)\right\}_{n=1}^{N}$. GP algorithm combines the preliminarily predicted information $\hat{\mathbf{X}}^{(n)}$ of the base-level classifiers to their target labels $t^{(n)}$.

GP is a powerful optimization evolutionary algorithm that searches for possible candidate solutions in the defined problem space. During GP evolution, predictor function $\varphi = I_\theta(\hat{\mathbf{X}})$ is formed, where $\hat{\mathbf{X}} \in \Re^m$, $\varphi \in \Re$, and θ represents suitable set of GP parameters. To find proper structure of the fitness function, we provided suitable set of functions (like plus/sin/log/etc.), variables (like x/y/etc.), and random numbers. In GP, candidate solutions $I_\theta(\hat{\mathbf{X}})$ are represented in the form of tree structure. GP tree are created using Ramped Half-and-Half method to produce initial population of 100 individuals. We used area under the curve (AUC) of receiver operating characteristic (ROC) curve (AUC-ROC) for GP fitness. Maximum fitness score shows how successfully $I_\theta(\hat{\mathbf{X}})$ move towards the optimal solution. The fitness probability of individual candidates $I_\theta(\hat{\mathbf{X}})$, in the population P_s, is computed as follows:

$$\Pr(I_\theta) = \frac{I_\theta}{\sum_{P_s} I_\theta}. \tag{1}$$

By employing crossover operator, new offspring is produced by randomly choosing parts of selected individual parents. We opted for a variable crossover/mutation ratio that automatically corrects this ratio to converge global minima. Finally, GP simulation cease to proceed when the best individual function, $I_\theta(\hat{\mathbf{X}}) \rightarrow F_\theta(\hat{\mathbf{X}})$, is developed.

Generally, the computational complexity of the GP algorithm is related to the complexity of the problem. The computational time consumed during each GP simulation is highly dependent on parameters of population size, number of generations, and input data size. All simulations are carried out using Intel(R) dual-core processor of 2.93 GHz, 2 GB RAM with Windows 7 operating system, in MATLAB R2013a environment. On the average, the GP runs have provided temporal cost of 1,235.77 and 570.35 s for the highest performing feature spaces of PseAAC-S (60 dimensions) and PseAAC-P (40 dimensions), respectively. Therefore, our approach in PseAAC-P feature space is nearly two times more efficient than PseAAC-S space. The detail of parameters setting is given in the supplementary Table S2. Supplementary Fig. S2 shows improvement in the best GP individuals in each generation. This figure indicates increase in complexity of the best individuals with respect to nodes and tree depth. The best numerical function for PseAAC-P in prefix form is given in Eq. 2 and its tree structure is shown in supplementary Fig. S3.

PSO algorithm is applied to find the optimal threshold for the best function $F_\theta(\hat{\mathbf{X}})$. This algorithm is used to select the optimal threshold values in each feature space (FS), i.e., $\gamma_{\mathrm{pso}}^{\mathrm{PseAAC}-S} = 0.4141$ for C/NC and 0.6391 for B/NBC. The best predictions of the HBC-Evo ($\hat{E}_{\mathrm{Ens}}^{\mathrm{FS}}$) are computed using $\gamma_{\mathrm{pso}}^{\mathrm{FS}}$ for cancer and non-cancer patients as follows:

$$\hat{E}_{\mathrm{Ens}}^{\mathrm{FS}} = \begin{cases} C & \text{if } (F_\theta^{\mathrm{FS}}(\hat{\mathbf{X}}) \geq \gamma_{\mathrm{pso}}^{\mathrm{FS}}) \\ \mathrm{NC} & \text{otherwise} \end{cases} \tag{3}$$

$$F_\theta^{\mathrm{PseAAC-P}}(\hat{\mathbf{X}}) = \mathrm{plus}(\mathrm{plus}(\cos(X_2), \mathrm{times}(\mathrm{times}(\mathrm{minus}(\mathrm{minus}(\mathrm{minus}(\cos(\cos(\mathrm{minus}(\mathrm{minus}(\mathrm{minus}(X_4, X_3), X_3), X_3))), \cos(\sin(X_1))), X_1), X_3), \mathrm{abs}(\sin(\cos(\mathrm{minus}(\mathrm{minus}(\cos(X_2), X_2), \cos(\mathrm{minus}(\mathrm{abs}(X_3), X_3)))))), X_3)), \cos(\mathrm{minus}(\mathrm{minus}(\cos(X_2), X_2), \cos(\mathrm{minus}(\mathrm{abs}(X_3), X_3))))). \tag{2}$$

Results and discussion

The improved performance of the proposed HBC-Evo system is revealed compared to state of the art approaches. Table 1 demonstrates that $\mathrm{RF}_{\mathrm{PseAAC\text{-}S}}$ and $\mathrm{RF}_{\mathrm{PseAAC\text{-}P}}$ models have achieved an overall accuracy of 97.93 and 97.10 % for C/NC and B/NBC datasets, respectively. Thus, an improvement of 7.93 and 7.10 % is found compared to QPDR model (Munteanu et al. 2009). The performance of $\mathrm{RF}_{\mathrm{PseAAC\text{-}S}}$ model is highest among the individual classifiers. Detailed results of the individual classifiers are provided in Supplementary Table S3.

Table 1 Performance comparison of state of the art approaches and HBC-Evo

Approaches	C/NC			B/NBC		
	AUC	Acc	Q-avg. statistic	AUC	Acc	Q-avg. statistic
Fuzzy-GA (Pena-Reyes and Sipper 1999)	NA	N/A	NA	NA	97.36	NA
QPDR (Munteanu et al. 2009)	NA	90.00	NA	NA	91.80	NA
Fuzzy-SVM (Li et al. 2011)	NA	N/A	NA	NA	96.35	NA
KNN (Majid et al. 2014)	NA	96.01	NA	NA	94.54	NA
SVM (Majid et al. 2014)	NA	96.71	NA	90.00	95.18	NA
Individual*						
$RF_{PseAAC-S}$	99.56	97.93	NA	–	–	NA
$RF_{PseAAC-P}$	–	–	–	99.20	97.10	NA
Proposed HBC-Evo						
Ens_{AAC}	99.91	98.83	0.397	99.82	97.80	0.399
Ens_{SAAC}	99.93	98.89	0.398	99.70	97.69	0.391
$Ens_{PseAAC-S}$	99.95	99.01	0.397	99.89	98.40	0.393
$Ens_{PseAAC-P}$	99.87	98.40	0.398	99.86	98.30	0.399

NA not available

* In addition to RF, we also used NB, KNN, and SVM as base classifiers (Supplementary Table S3)

For C/NC dataset, $Ens_{PseAAC-S}$ has attained 99.95 % AUC of ROC and achieved an overall accuracy of 99.01 % (Table 1). Performance of the HBC-Evo system is higher than the highest performing individual classifier, i.e., $RF_{PseAAC-S}$ 99.56 % in terms of AUC and 97.93 % in terms of Acc. Therefore, combining different individual approaches through the proposed evolutionary approach has effectively exploited diversity of learning algorithms in different decision spaces. For B/NBC dataset, $Ens_{PseAAC-S}$ has achieved 99.89 % AUC of ROC and an overall accuracy of 98.40 %. The average diversity of ensemble models is described in terms of Q-statistics (Table 1). The lower value (<1) of Q-statistics highlights the higher diversity that is useful for the proposed approach. Supplementary Table S4 shows other performance measures of sensitivity, specificity, *F* score, etc.

Conclusion

The developed HBC-Evo system has combined the preliminary information of individual classifiers for the diagnosis of breast cancer. This system has effectively exploited the change that occurred in amino acids compounds related to cancer proteins using their physicochemical properties of hydrophobicity and hydrophilicity. Our $Ens_{PseAAC-S}$ model has demonstrated excellent improvement over other state of the art approaches for cancer prediction. The proposed HBC-Evo system could be used for academia, practitioners, and technicians for the early diagnosis of breast cancer using protein amino acid sequences.

Conflict of interest It is stated that we authors do not have any type of "conflict of interest" in the submission of this paper.

References

Aminzadeh F, Shadgar B, Osareh A (2011) A robust model for gene analysis and classification. Int J Multimed Appl 3:11–20

Good BM, Loguercio S, Griffith OL, Nanis M, Wu C, Su AI (2014) The cure: making a game of gene selection for breast cancer survival prediction. arXiv preprint arXiv 1402:3632

Jiang X, Wei R, Zhao Y, Zhang T (2008) Using Chou's pseudo amino acid composition based on approximate entropy and an ensemble of AdaBoost classifiers to predict protein subnuclear location. Amino Acids 34(4):669–675

Khan A, Majid A, Hayat M (2011) CE-PLoc: an ensemble classifier for predicting protein subcellular locations by fusing different modes of pseudo amino acid composition. Comput Biol Chem 35(4):218–229

Lavanya D, Rani KU (2012) Ensemble decision making system for breast cancer data. Int J Comput Appl 51(17):0975–8887

Li ZC, Zhou XB, Dai Z, Zou XY (2009) Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis. Amino Acids 37(2):415–425

Li DC, Liu CW, Hu SC (2011) A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets. Artif Intell Med 52:45–52. doi:10.1016/j.artmed.2011.02.001

Majid A, Safdar A, Mubashar I, Nabeela K (2014) Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines. Comput Methods Progr Biomed 113(3):792–808

Maqsood H, Khan A, Yeasin M (2012) Prediction of membrane proteins using split amino acid and ensemble classification. Amino Acids 42(6):2447–2460

Munteanu CR, Magalhães AL, Uriarte E, González-Díaz H (2009) Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. J Theor Biol 257(2):303–311

Pena-Reyes CA, Sipper M (1999) A fuzzy-genetic approach to breast cancer diagnosis. Artif Intell Med 17:131–155

Ruxandra S, Stoean C (2013) Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection. Expert Syst Appl 40:2677–2686

Safdar A, Majid A, Khan A (2014) IDM-PhyChm-Ens: intelligent decision-making ensemble methodology for classification of human breast cancer using physicochemical properties of amino acids. Amino Acids 46(4):977–993

Xin M, Guo J, Liu H, Xie J, Sun X (2012) Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information. IEEE ACM Trans Comput Biol Bioinf 9(6):1766–1775