Adaptive Filtering Techniques Combined with Natural Selection- Based Heuristic Algorithms in the Prediction of Protein-Protein Interactions

Christos M. Dimitrakopoulos[1], Konstantinos A. Theofilatos[1], Efstratios F. Georgopoulos[2], Spyridon D. Likothanassis[1], Athanasios K. Tsakalidis,[1] Seferina P. Mavroudi[1]

[1] Department of Computer Engineering & Informatics, University of Patras, Rio, GR-26500, Patras, Greece
{dimitrakop, theofilk, likothan, tsak, mavroudi}@ceid.upatras.gr

[2] Technological Educational Institute of Kalamata, 24100, Kalamata, Greece,
sfg@teikal.gr

Abstract. The analysis of protein-protein interactions (PPIs) is crucial to the understanding of cellular organizations, processes and functions. The reliability of the current experimental approaches interaction data is prone to error. Thus, a variety of computational methods have been developed to supplement the interactions that have been detected experimentally. The present paper's main objective is to present a novel classification framework for predicting PPIs combining the advantages of two algorithmic methods' categories (heuristic methods, adaptive filtering techniques) in order to produce high performance classifiers while maintaining their interpretability. Our goal is to find a simple mathematical equation that governs the best classifier enabling the extraction of biological knowledge. State-of-the-art adaptive filtering techniques were combined with the most contemporary heuristic methods which are based in the natural selection process. To the best of our knowledge, this is the first time that the proposed classification framework is applied and analyzed extensively for the problem of predicting PPIs. The proposed methodology was tested with a commonly used data set using all possible combinations of the selected adaptive filtering and heuristic techniques and comparisons were made. The best algorithmic combinations derived from these procedures were Genetic Algorithms with Extended Kalman Filters and Particle Swarm Optimization with Extended Kalman Filters. Using these algorithmic combinations high accuracy interpretable classifiers were produced.

Keywords: Protein-Protein Interactions, Adaptive Filtering, Genetic Algorithms. Particle Swarm Optimization, Least Squares Algorithm, Recursive Least Squares, Kalman Filtering

1 Introduction

In each living cell of the human organism, a variety of protein interactions take place. In recent years, researchers have tried to approach the problem of predicting all possible protein interactions in the human organism by implementing different computational techniques. At the beginning, most of them were based on the analysis of a sole feature, indicative of interaction between two proteins. Several examples of such features are features concerning the genomic sequence of the genes-generators of the reference proteins, features concerning the structure of the reference proteins, features concerning the sequences of the references proteins and many others [1,2]. The most recent computational approaches use various features as inputs for their classifiers in order to take advantage of all the available information [3,4].

In the present paper, we applied to the problem of PPI prediction several adaptive filtering techniques combined with the most contemporary heuristic methods which are based in the natural selection process. From this combination, a novel computational framework has been formed for the creation of a mathematical equation that gives the best classifier. We consider three classical parameter estimation algorithms (LMS, RLS and Kalman Filter) and one gain adaptation algorithm (IDBD). The gain adaptation algorithms have been shown to perform comparably to the best algorithms (Kalman and RLS), but they have a lower complexity [6].

Concerning the heuristic methods, each adaptive filtering technique was combined with a genetic algorithm and a Particle Swarm Optimization (PSO) heuristic technique. Genetic algorithms [7] and PSO [8] are the most contemporary heuristic methods and their process is based on the principle of natural selection. They are implemented in order to find the best subset of mathematical terms that constitutes the optimal mathematical model that can be used as the optimal classifier in the adaptive filtering techniques. In the RLS and LMS algorithms, the forgetting and convergence factors are been optimized respectively through the heuristic algorithms used.

In [5], genetic algorithms were firstly combined with Extended Kalman Filters for the problem of predicting PPIs. That method has been demonstrated to achieve higher classification performance than the PIPS naive Bayesian method [9] on the same dataset. In the terms of the current paper, the research has been extended by implementing different adaptive filtering techniques and heuristic algorithms in order to establish a novel classification framework. The proposed methodology was tested with a commonly used dataset and the advantages and disadvantages of each method selection are discussed. The mathematical equation that produced the best classifier in terms of classification performance and interpretability was presented and some first biological hypotheses about the classification model were derived.

2 Materials and Methods

2.1 Dataset

The dataset used in this paper was created by retrieving material from the PIPS database [9] which contains information for about 17.643.506 human protein interactions. The positive samples in the dataset were the 16536 known human protein interactions extracted from Human Protein Reference Database (HPRD) [10]. The negative dataset was built using randomly chosen protein pairs provided by the PIPS Database. A 1:1 ratio of positive and negative samples was selected in our original dataset, in order to force our classifiers to achieve higher sensitivity and discover a large number of protein-protein interactions. The dataset was split into two equal subsets, training and testing, keeping the 1:1 ratio between the positive and negative samples.

For every protein pair the following features as in [9] were used as inputs:

- Expression: Gene expression profiles from 79 physiologically normal tissues (Pearson correlation of co-expression over all conditions).
- Orthology: Interactions of homologous protein pairs from yeast, fly, worm and human. The similarity function used is the InParanoid score function [11].
- Fully Bayesian combination of the following features:
    - Localization: Here PLST predictions are used [12].
    - Domain co-occurrence: Chi-square score of co-occurrence of domain pairs.
    - Post-translational modifications (PTM) co-occurrence: PTM pair enrichment score has been calculated as the probability of co-occurrence of two specific PTMs in all pairs of interacting protein pairs divided by the probability of occurrence of both of these PTMs separately.
- Transitive: This is a module that considers local topology of the underlying network predicted using combinations of above features and it works on the premise that a pair of proteins is more likely to interact if it shares interacting partners.

All feature values were normalized in a range from 0 to 1.

2.2 Adaptive Filtering Techniques

The adaptive filtering techniques used in this survey are the least mean squares (LMS) [13], the recursive least squares (RLS) [14], the Extended Kalman filtering method (EKF) [15] and the incremental delta-bar-delta algorithm (IDBD) [16].

The LMS algorithm is by far the most widely used algorithm in adaptive filtering because of its low computational complexity, proof of convergence in stationary environment and stable behavior when implemented with finite-precision arithmetic.

To guarantee the convergence of LMS it is good practice to set the learning rate in the range,

$$0 < \mu < \frac{1}{\lambda max} \quad (1)$$

, where λ_{max} is the largest eigenvalue of $R = E[x(k)x^T(k)]$, where x(k) is the vector of the input variables.

Least-squares algorithms aim at the minimization of the sum of the squares of the difference between the desired signal and the model filter output. When new samples of the incoming signals are received at every iteration, the solution for the least-squares problem can be computed in recursive form resulting in the recursive least-squares (RLS) algorithms.

Kalman filter is a set of mathematical equations which constitutes an efficient means to estimate the state of a process by minimizing the mean of the square error. The Extended Kalman Filter (EKF) is the nonlinear version of the Kalman Filter, and its goal is to approach the situation where the process to be estimated or the measurement relationship to the process is nonlinear. The specific equations for the time and measurement updates are divided into two groups. The time update equations and the measurement update equations.

Gain adaptation algorithms [16] implement a sort of meta-learning in that the learning rate is adapted based on the current inputs and on the trace of previous modifications [17]. The incremental delta-bar-delta (IDBD) uses a different adaptive learning rate for each input. This action can lead to the improvement of the system, when some of the inputs are irrelevant. In IDBD, each element of the $k_i(t)$ of the gain vector k(t) is computed separately.

2.3 Heuristic Methods

Heuristic methods are implemented to fasten the process of finding a "good enough" solution to a problem, where the usage of an exhaustive search is impractical. Because of the large search space of our problem (2^{113} possible solutions), 2 classical heuristic methods were used, genetic algorithms and particle swarm optimization.

Genetic Algorithms (GAs) [7] are search algorithms inspired by the principle of natural selection. They are useful and efficient when the search space is big and complicated or when there is not any available mathematical analysis of the problem. A population of candidate solutions, called *chromosomes*, is optimized through a number of evolutionary cycles and genetic operations, such as *crossovers* or *mutations*.

In our approach, a simple GA was used where each chromosome comprises *term genes* that encode the best term subset and *parameter genes* that encode the best choice of parameters for each adaptive method used. The term genes are binary genes with the value of 1 indicating that this specific term should be included in the final classification model and the value of 0 indicating the opposite. The parameters which

are optimized using GA are the convergence factor μ for the LMS adaptive filtering method and the forgetting factor λ for the RLS adaptive filtering method.

For the genetic algorithm used in our hybrid methodology, the one-point crossover and the mutation operators were used. Crossover and mutation probabilities for the GA where set to 0.9 and 0.01 respectively. Crossover is used in hope that new chromosomes will have good parts of old chromosomes and maybe the new chromosomes will be better. However it is good to leave some part of population survive to next generation. This is the reason a high (but not equal to one) crossover probability was used. As already mentioned, mutation is made to prevent falling GA into local extreme, but it should not occur very often, because then GA will in fact change to random search. That is the main reason why a small mutation probability was applied. Furthermore, roulette selection was used for the selection step and elitism to raise the evolutionary pressure in better solutions and to accelerate the evolution.

The size of the initial population was set to 30 chromosomes after experimentation in the training dataset. The termination criterion is the maximum number of 100 generations to be reached combined with a termination method that stops the evolution when the population is deemed as converged. The population is deemed as converged when the average fitness across the current population is less than 5% away from the best fitness of the current population. Specifically, when the average fitness across the current population is less than 5% away from the best fitness of the population, the diversity of the population is very low and evolving it for more generations is unlikely to produce different and better individuals than the existing ones or the ones already examined by the algorithm in previous generations.

The particle swarm optimization (PSO) [8] algorithm is another population based heuristic search algorithm based on the simulation of the social behavior of birds within a flock.

In our approach PSO searches the best mathematical term's subset for building the optimal classifier and the best parameters for each adaptive filtering method. Thus its particle is consisted of 113 variables indicating whether to include a term in the subset or not, plus one variable for the parameter selection of the adaptive filtering method. The values of the variables for the term selection range from 0 to 2. Values bigger than 1 indicate that this specific term should be included in the final classification model and the values less than 1 indicate the opposite.

In order to make a fair comparison with the GA search method, the initial population size of the swarm was set to 30 particles and the total number of iterations was set to 100. The termination criteria of convergence was not applied here because of the adaptive search behavior used in our PSO implementation which needs the total number of iterations to be used in order to perform in its full strength and effectiveness.

As mentioned above, the convergence factor μ for the LMS adaptive filtering method and the forgetting factor λ for the RLS adaptive filtering method are been optimized through the GA and the PSO algorithms. According to the theory, the convergence factor μ for LMS algorithm has been selected in the range $[0, 1/\lambda_{max}]$, in order to guarantee convergence of the mean square error, where λ_{max} is the largest

eigenvalue of the autocorrelation matrix of the input x(k). Concerning RLS algorithm, the forgetting factor λ is chosen such that $2\mu = 1 - \lambda$, where μ is the corresponding convergence factor of the LMS algorithm. Hence, the values of the parameter λ are close to 1 and the values of the parameter μ are close to 0.

2.4 Evaluation in the Hybrid Methodology

The main idea of our proposed classification method is to use a heuristic method to find a "good enough" subset of terms in order to build the mathematical model for our predictor and then apply an adaptive filtering technique to find its optimal parameters. The search space of our problem consists of 2^{113} possible solutions and hence is extremely large. GAs and PSOs are heuristic methods that can deal with large search spaces and do not get trapped in local optimal solutions. The 113 mathematical terms that were used in our method were taken from 3 known nonlinear mathematical models, which are the Volterra Series, the exponential and the polynomial model, as described in [5].

In our approach, a simple Genetic Algorithm and PSO were used as heuristic methods. Each chromosome in these methods comprises of genes that encode the best subset of mathematical terms to be used in our classifier. Roulette selection was used for the selection step including elitism to accelerate the evolution of the population in the Genetic algorithm.

The evaluation process in both heuristic techniques is described in the following steps:

- For every individual of the population, use the Adaptive Filtering method (using the training dataset) to compute the best parameters for the subset of terms that arises from the individual's genes.
- Use the mathematical model found at the previous step to compute the fitness of the classifier in the validation set.
- For every individual compute the output of the following function:
  $F = a^* G_m + b^*sensitivity + c^*specificity - d^*terms$ (2),
  where G_m is the geometric mean of sensitivity and specificity. Hence, the proposed method tries to produce one mathematical model that optimizes the combination of sensitivity and specificity measures, sensitivity and specificity alone and simultaneously minimize the number of mathematical terms to be used in the mathematical model.
- Rank the individuals from the one having the lower fitness value to the one having the bigger fitness value. Using this ranking in the evaluation procedure, better scaled fitness scores can be achieved.
- Give as fitness value for every individual their ranking number. For example, the worst individual will take fitness equal to 1, the next fitness equal to 2 etc.

During the training we used 5-fold cross validation.

3 Experimental Results

In order to evaluate the performance of each combination of algorithms we applied it to the dataset described in Section 2.1. In order to make a fair comparison between all combinations, their classification thresholds were optimized. For all methods we experimented running them 100 times and in Table 1, the average values for each combination of algorithms are presented.

Table 1. Classification performance of all possible combinations between adaptive filtering and heuristic algorithms. Gm is the geometric mean of sensitivity and specificity.

Algorithm Combination	Gm	Sensitivity	Specificity	# of terms
Kalman with GA	0.7911	0.7540	0.8302	12.3
Kalman with PSO	0.7927	0.7568	0.8304	39.3
RLS with GA	0.7826	0.7291	0.8400	14.6
RLS with PSO	0.7797	0.7308	0.8319	26.1
LMS with GA	0.7012	0.6388	0.7698	15.6
LMS with PSO	0.6845	0.6418	0.7300	38.6
IDBD with GA	0.6704	0.4976	0.9033	26.6
IDBD with PSO	0.6694	0.5516	0.8124	37.3

As it is clear from Table 1, Kalman Filtering algorithm when combined with the PSO heuristic method achieves the highest geometric mean, which is the combination of sensitivity and specificity. Hence, we can assume that Kalman with PSO achieves the highest classification performance. However, at the same time it is observed that the optimal mathematical models generated by Kalman with GA are the simplest ones concerning complexity. Particularly, classifiers generated from Kalman with GA algorithm have an average of 12.3 mathematical terms. While IDBD and RLS succeed to achieve higher specificity than Kalman algorithm, they cannot achieve a higher overall classification performance through the geometric mean measure, because of their low sensitivity values.

In general, GAs lead to less complexity classifiers than the classifiers produced by PSO for all possible adaptive filtering algorithms. According to the classification performance, adaptive filtering techniques can be ranked from the best one to the worse as Kalman, RLS, LMS and IDBD. This fact can also be observed at figure 1, where the performance of each algorithm has been depicted based on the evaluation measure, as described in equation (2). The evaluation measure is a multi-objective measure, as it considers more than one metrics simultaneously. Kalman and RLS perform almost the same, but Kalman achieves a slightly higher accuracy. LMS algorithm is known for its computational simplicity, but RLS algorithm and Kalman filter have higher performance as they focus on the minimization of the mean square error between the desired signal and the model filter output. IDBD, which is a gain adaptation algorithm, fails to compete with the other algorithms. Maybe this happens because the learning rate of the IDBD algorithm is not optimized through the heuristic

algorithms, but within the algorithm itself (meta-learning parameter). When dealing with the PPI prediction problem our first objectives are classification performance and interpretability and not computational complexity because of the off-line nature of the problem. Thus, among the local search algorithms explored in the present paper, Extended Kalman filters are the best solution.

Another observation is that the results of the adaptive filtering techniques combined with PSO have larger variance than the corresponding algorithms combined with GA. This means that PSO is more unstable than GA algorithm, and hence, GA has a smoother convergence. Moreover, we can observe that RLS, LMS and IDBD have -in general- larger variance than the Kalman algorithm which is more stable. This fact can be attributed to the parameters that RLS, LMS and IDBD try to optimize through the heuristic algorithm. As a result, they have a larger search space to explore and the complexity of the heuristic problem rises making them difficult to converge using the same iterations. In the Kalman case, both GA and PSO have low variance in the measures' results and PSO achieves to overcome the classification accuracy of the GA.

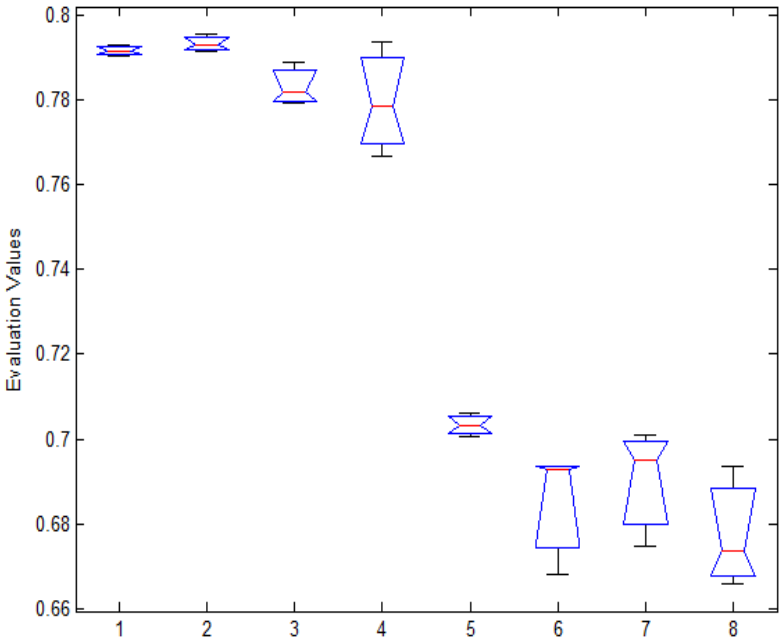

Fig. 1. Evaluation value is the average of the three main measures (geometric mean, sensitivity, specificity), for each combination of algorithms. 1: Kalman GA, 2: Kalman PSO, 3: RLS GA, 4: RLS PSO, 5: LMS GA, 6: LMS PSO, 7: IDBD GA, 8: IDBD PSO.

Next, we present the classifier with the higher evaluation value (equation 2), that resulted from the Kalman with GA combination. We observe that there are a lot of nonlinear mathematical terms in the model, like exponential or polynomial, revealing the complexity of the real model and indicating that this classification model cannot

be derived with simple classification algorithms. The model attributes higher importance to terms including the third feature x_3, which is a combination of three distinct features. Furthermore, the third feature dominates the model, as 4 terms out of 9 consist only of the third feature.

$$out = -4.0961{x_3}^3 - 0.5481x_2{x_4}^3 - 0.4992{x_4}^4 + 1.1192x_2e^{-{x_3}^2} + 4.2501x_3e^{-{x_3}^2} + 0.7043x_4e^{-{x_2}^2} + 0.3599e^{-{x_3}^2} + 3.5403{x_3}^7 - 0.6968{x_2}^9 \quad (3)$$

4 Conclusions

We have studied the influence of different combinations between adaptive filtering techniques and heuristic algorithms on PPI prediction problem in order to produce the optimal hybrid method for this problem. A search space of possible mathematical models is been searched through the heuristic algorithms (GA, PSO) and the optimal parameters for each specific model are been calculated by using an adaptive filtering technique (Kalman, RLS, LMS, IDBD). The mathematical model resulted from the above process constitutes our optimal classifier. All methods were tested in a human protein interaction dataset extracted from the PIPS database.

Our research concluded that the combination of an Extended Kalman Filter combined with PSO – as described in Section 2.3 – gives us the highest classification performance. Moreover, the Extended Kalman Filter when combined with GA produces the less complicated classifiers. In general, PSO heuristic has a larger variance than GA technique, except for the Kalman algorithm where both are stable. In the Kalman algorithm, PSO achieves to overcome the classification accuracy of GA.

In this paper, we presented a new classification framework for predicting PPIs combining heuristic algorithms with adaptive filtering techniques. According to the need of the researchers for high classification accuracy or low complex classifiers, the most convenient combination of algorithms can be used. Finally, we intent to use our method for the prediction of novel human PPIs.

Acknowledgments. This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

References

1. J.R. Bock and D.A. Gough, "Predicting protein-protein interactions from primary structure," Bioinformatics, vol. 17, May. 2001, pp. 455-460.
2. H.N., Chua, W.K. Sung, and L., Wong, Exploiting indirect neighbors and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22(13):1623-1630,2006.

3. X. Chen and M. Liu, "Prediction of protein-protein interactions using random decision forest framework," Bioinformatics, vol. 21, Dec. 2005, pp. 4394-4400.
4. P. Fariselli, F. Pazos, A Valencia, and R. Casadio, "Prediction of protein—protein interaction sites in heterocomplexes with neural networks," European Journal of Biochemistry 1 FEBS, vol. 269, Mar. 2002, pp. 1356-1361.
5. Theofilatos, K.A.; Dimitrakopoulos, C.M.; Tsakalidis, A.K.; Likothanassis, S.D.; Papadimitriou, S.T.; Mavroudi, S.P.; , "A new hybrid method for predicting protein interactions using Genetic Algorithms and Extended Kalman Filters", *2010, 10th IEEE International Conference on Information Technology and Applications in Biomedicine (ITAB)*, 3-5 Nov. 2010
6. P.S. Diniz, Adaptive Filtering: Algorithms and Practical Implementation" Springer, 2002.
7. Holland J., Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence, Cambridge, Mass: MIT Press, 1995.
8. Kennedy J. and Eberhart R.C., Particle Swarm Optimization, In Proceedings of the IEEE International Joint Conference on Neural Networks, pages 1942-1948,1995.
9. M. Scott and G. Barton, "Probabilistic prediction and ranking of human protein-protein interactions," BMC Bioinformatics, vol. 8, 2007, p. 239.
10. G.R. Mishra, M. Suresh, K. Kumaran, N. Kannabiran, S. Suresh, P. Bala, K. Shivakumar, N. Anuradha, R. Reddy, T.M. Raghavan, S. Menon, G. Hanumanthu, M. Gupta, S. Upendran, S. Gupta, M. Mahesh, B. Jacob, P. Mathew, P. Chatterjee, K.S. Arnn, S. Sharma, K.N. Chandrika, N. Deshpande, K. Palvankar, R. Raghavnath, R. Krishnakanth, H. Karathia, B. Rekba, R. Nayak, G. Vishnupriya, H. G.M. Kumar, M. Nagini, G.S.S. Kumar, R. Jose, P. Deepthi, S. S. Mohan, T.K. B. Gandhi, H.C. Harsha, K.S. Deshpande, M. Sarker, T. S. K. Prasad, and A Pandey, "Human protein reference database 2006 update," Nucleic Acids Research, vol. 34, Jan. 2006, pp. 0411-414.
11. K.P. O'Brien, M. Remm, E.L. Sonnhammer: Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleid Acids Res 2005, 33(Database issue):D476-80.
12. M. S. Scott, D. Y. Thomas, and M. T. Hallett, "Predicting subcellular localization via protein motif co-occurrence," Genome Research, vol. 14, Oct. 2004, pp. 1957-1966.
13. B. Widrow and M. E. Hoff, "Adaptive switching circuits," *WESCOM Conv. Rec.*, pt. 4, pp. 96-140, 1960.
14. G. C. Goodwin and R. L. Payne, *Dynamic System Identification: Experiment Design and Data Analysis,* Academic Press, NewYork, NY, 1977.
15. G. Welch and G. Bishop, "An Introduction to the Kalman Filter," UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL, 1995.
16. R. S. Sutton. Adapting bias by gradient descent: An incremental version of delta-bar-delta. In Proceedings of the Tenth National Conference on Artificial Intelligence, pages 171 –176. MIT Press, 1992.
17. Prediction Update Algorithms for XCSF: RLS, Kalman Filter, and Gain Adaptation
18. Ratnaweera A., Halgamuge S. and Watson H., Particle Swarm Optimization with Self -Adaptive Acceleration Coefficients, In Proceedings of the First International Conference on Fuzzy Systems and Knowledge Discovery, pages 264-268, 2003