An Immune System Inspired Algorithm for Protein Function Prediction

Archana Chowdhury[1], Amit Konar[1],
Pratyusha Rakshit[1], and Janarthanan Ramadoss[2]

[1] Electronics and Telecommunication Engineering Department
Jadavpur University, Kolkata, India
[2] Department of Compuiter Science Engineering, TJS College of Engineering
Chennai, India
{chowdhuryarchana,pratyushar1}@gmail.com,
konaramit@yahoo.co.in, srmjana_73@yahoo.com

Abstract. An important problem in the field of bioinformatics research is assigning functions to proteins that have not been annotated. The extent, to which protein function is predicted accurately, depends largely on the Protein-Protein interaction network. It has been observed that bioinformatics applications are benefited by comparing proteins on the basis of biological role. Similarity based on Gene Ontology is a good way of exploring the above mentioned fact. In this paper we propose a novel approach for protein function prediction by utilizing the fact that most of the proteins which are connected in Protein-Protein Interaction network, tend to have similar functions. Our approach, an immune system-inspired meta-heuristic algorithm, known as Clonal Selection Algorithm (CSA), randomly associates functions to unannotated proteins and then optimizes the score function which incorporates the extent of similarity between the set of functions of unannotated protein and annotated protein. Experimental results reflect that our proposed method outperforms other state of the art algorithms in terms of precession, recall and F-value, when utilized to predict the protein function of Saccharomyces Cerevisiae.

Keywords: protein function prediction, protein–protein interaction network, gene ontology, annotated protein, clonal selection algorithm.

1 Introduction

The main problem in molecular biology is to understand the function of a protein, as the function of most of the proteins is unknown. It has been observed that even the most studied species, Saccharomyces cerevisiae, is reported to have more than 26 percent of its proteins with unknown molecular functions [1]. Huge amount of data continue to accumulate due to the application of high throughput technologies in various genome projects. Protein-Protein Interaction (PPI) is an important source of information among these databases. The introduction of high-throughput techniques have resulted in an amazing number of new proteins been identified. However, the function of a large number of these proteins still remains unknown.

M.K. Kundu et al. (eds.), *Advanced Computing, Networking and Informatics - Volume 1*,
Smart Innovation, Systems and Technologies 27,
DOI: 10.1007/978-3-319-07353-8_79, © Springer International Publishing Switzerland 2014

Several algorithms have been developed to predict protein functions, on the basic assumption that proteins with similar functions are more likely to interact. Among them, Deng proposed the Markov random field (MRF) model, which predicts protein functions based on the annotated proteins and the structure of the PPI network [2]andSchwikowski proposed neighbor counting approach [3]. In recent years, more and more research turned to predict protein functions semantically by combining the inter-relationships of function annotation terms with the topological structure information in the PPI network. To predict protein functions semantically, various methods were proposed to calculate functional similarities between annotation terms. Lord *et al.*[4] were the first to apply a measure of semantic similarity to GO annotations. Resink [5] used the concept of information content to calculate the semantic similarity between two GO terms.

In this paper, we aim to predict the function of an unannotated protein by using the topographical information of PPI network and function of annotated protein. The similarity of GO terms used to annotate proteins is measured using the information content of the respective terms as well as the terms which are common in the path from root to the GO terms. For this task we have employed the use of an immune system-inspired meta-heuristic algorithm, known as Clonal Selection Algorithm (CSA).The proposed method used a hypermutation strategy which provides the exploration capability to individual clone within the search space.

The rest of this paper is organized as follows: Section 2 give a brief idea about the definition and formulation of the problem as well as the scheme for solution representation. Section 3 provides an overview of the proposed. Experiments and Results are provided in Section 4.Section 5 concludes the paper.

2 Background of the Problem

2.1 Problem Definition

PPI network with N proteins is considered. The PPI network can be represented in the form of a binary data matrix $\mathbf{K}_{N\times N}$ where $k_{ij}=k_{ji}=1$and$k_{ij}=k_{ji}=0$ denotes the presence and absence of interaction between the proteins p_i and p_j respectively. The set of all functions for each protein p, is denoted as $F(p)$ thus, the set of all possible functions in the network is defined as $F= F(p_1)$U $F(p_2)$U… U $F(p_N)$ with number of all possible functions in the network $|F|=D$.

Given the PPI data matrix $\mathbf{K}_{N\times N}$, a protein function prediction algorithm tries to find a set of possible functions $F(p)$ of an unannotated protein p based on the functions $F(p^{/})$ of all annotated proteins $p^{/}$ in the PPI network. Since the functions can be assigned to the unannotated protein p in a number of ways, a fitness function (measuring the accuracy of the function prediction) must be defined. The problem now turns out to be an optimization problem of finding a set of functions $F(p)$ of optimal adequacy as compared to all other feasible sets of functions for unannotated protein p.

2.2 Formulation of the Problem

The effectiveness of protein function prediction can be improved by taking the composite benefit of the topological configuration of the PPI network and the functional categories of annotated proteins through Gene Ontology (GO) [8], [9]. The protein functions are annotated using GO terms. GO is basically represented as a directed acyclic hierarchical structure in which a GO term may have multiple parents/ancestor GO terms. The probability, *prob*(*t*), for each GO term *t* in the GO tree is frequency of occurrence of the term and its children divided by the maximum number of terms in the GO tree. Thus the probability for each node/GO term will increase as we move up towards the root. The information content of a GO term in the GO tree is based on the *prob*(*t*) value and is given by

$$I(t) = -\log_{10} prob(t) \tag{1}$$

Here (1) shows that lesser is the probability of the GO term, more will be the information content associated with it. The similarity between two GO terms will be high if they share more information. The similarity between two terms, which is captured by the set of common ancestors, is the ratio of probability of the common terms between them to the probability of the individual terms.It can be represented as follows:

$$S(t_1,t_2) = \max_{t \in C(t_1,t_2)} \left(\frac{2 \log prob(t)}{\log prob(t_1) + \log prob(t_2)} \right) \tag{2}$$

Where $C(t_1, t_2)$ denotes the set of common ancestors of terms t_1 and t_2, $S(t_1,t_2)$ measures the similarity with respect to information content, in terms of the common ancestors of the terms t_1 and t_2.The value of the above similarity measure ranges between 0 and 1. With this representation scheme of protein functions, the similarity between a predicted function $f \in F(p)$ of unannotated protein *p* and a real function $f^{/}$ $\in F$ of the PPI network can be computed as follows.

$$sim(f, f') = S(t_1, t_2) \tag{3}$$

Where function *f*is annotated by t_1 and $f^{/}$ is annotated by t_2. Next, the score of the unannotated protein being annotated by the predicted function *f* is evaluated by (4).

$$score(p, f) = \sum_{\substack{j=1, \\ f' \in F(p_j)}}^{N} \frac{1}{dist(p, p_j)} \times sim(f, f') \tag{4}$$

Here $dist(p, p_j)$ represents the minimum number of edges between proteins *p* and p_j. Two important facts are included in (4), First is that, it assigns function *f* to protein *p* based on the similarity between *f* and all other protein functions available in the given network which conforms to the fact that theproteins with similar functions interact more frequently to construct the PPI network, secondly the term $1/dist(p, p_j)$ is included because proteins far away from *p* contribute less functional information than those having direct interaction with *p*. This is accomplished by assigning less weight to the proteins far away from *p* than its close neighbors. From (4), it is apparent that a high value of *score*(*p*, *f*) will indicate a higher adequacy in predicting *f* as a function of protein *p*.

2.3 Solution Representation and Cost Function Evaluation

In the proposed method a solution $\vec{X}_i$ is a vector of dimension *D* as *D*denotes maximum number of functions in the PPI network. The values of $\vec{X}_i$ belong to {0, 1} and the *j*-th parameter of $\vec{X}_i$ is interpreted as follows:

If $x_{i,j}$=1, then the *j*-th function f_j is predicted as a function of protein *p*. (5.a)

If $x_{i,j}$=0, then f_j is not predicted as a function of protein *p*. (5.b)

Let there be*D*=8 functions available in the network among which, the second, third, fifth and seventh have been predicted as assigned functions of unannotated protein *p*then the solution encoding scheme will be as shown in Fig.1.

0	1	1	0	1	0	1	0

Fig. 1. Solution encoding scheme in the proposed method

In order to judge the quality of such a solution $\vec{X}_i$, for function prediction, the contribution by the entire set of predicted functions (denoted by set *F*(*p*)= {f_j| $x_{i,j}$=1 for *j*= [1, *D*]}) to annotate protein *p* is used for fitness function evaluation. Symbolically,

$$fit(\vec{X}_i) = \sum_{\forall f \in F(p)} score(p, f) \tag{6}$$

3 An Overview of Clonal Selection Algorithm (CSA)

CSA [6] is a population-based stochastic algorithm, which is inspired by the antigen driven affinity maturation process of B-cells in immune system. Overviews of the main steps of CSA are as follows:

A. Initialization

CSA starts with a population of *NP*,*D*-dimensional antibodies, $\vec{X}_i(t) = \{x_{i,1}(t), x_{i,2}(t), x_{i,3}(t), \ldots, x_{i,D}(t)\}$ for *i*= [1, *NP*] representing the candidate solutions, at the current generation *t* = 0 by randomly initializing in the range $[\vec{X}^{\min}, \vec{X}^{\max}]$ .Thus the *j*-th component of the *i*-th antibody at *t*=0 is given by

$$x_{i,j}(0) = x_j^{\min} + rand_{i,j}(0,1) \times (x_j^{\max} - x_j^{\min}) \tag{7}$$

Where $rand_{i,j}$(0, 1) is a uniformly distributed random number lying between 0 and 1. The affinity or the fitness $fit(\vec{X}_i(0))$ of the antibody $\vec{X}_i(0)$ is evaluated for *i*=[1, *NP*].

B. Selection of Antibodies for Cloning

The antibodies are sorted in descending order of fitness $fit(\vec{X}_i(t))$ for i= [1, NP] and the first n individuals of the population are selected for subsequent cloning phase.

C. Cloning

Each member $\vec{X}_k(t)$ of the selected subpopulation of n antibodies is allowed to produce clones for k= [1, n]. The number of clones c_k for the k-th individual is proportional to its affinity $fit(\vec{X}_k(t))$ and is calculated using (8). Here $fit_{\min}$= $fit(\vec{X}_n(t))$ and $fit_{\max}$= $fit(\vec{X}_1(t))$ represent the lowest and highest affinity of the sorted individuals in the subpopulation of n antibodies respectively. Similarly, $c_{\min}$ and $c_{\max}$ denote maximum and minimum number of clones.

$$c_k = \left\lfloor \frac{fit(\vec{X}_k(t)) - fit_{\min}}{fit_{\max} - fit_{\min}} \times (c_{\max} - c_{\min}) \right\rfloor + c_{\min} \tag{8}$$

D. Hypermutation

Each clone of $\vec{X}_k(t)$, denoted as $\vec{X}_k^l(t)$ for j= [1, D], k= [1, n] and l= [1, c_k] undergoes through the static hypermutation process using (9).

$$x_{k,j}^l(t+1) = x_{k,j}^l(t) + \alpha \times x_{k,j}^l(t) \times (x_j^{\max} - x_j^{\min}) \times G(0,\sigma) \tag{9}$$

Here α is a constant, however small and $G(0, \sigma)$ is a random Gaussian variable with zero mean and σ as the standard deviation. Usually, $\sigma$is taken as 1 [7]. The fitness $fit(\vec{X}_k^l(t))$ is evaluated for l= [1, c_k].

E. Clonal Selection

Let the set of matured clones (after hypermutation) corresponding to the k-th antibody, including itself, is denoted as S_k= $\{\vec{X}_k(t), \vec{X}_k^1(t+1), \vec{X}_k^2(t+1), ..., \vec{X}_k^{c_j}(t+1)\}$ for k= [1, n]. The best antibody in S_k with highest fitness is allowed to pass to the next generation. Symbolically,

$$\vec{X}_k(t+1) \leftarrow \arg\left(\max_{\forall \vec{X} \in S_k} (fit(\vec{X})) \right) \tag{10}$$

F. Replacement

The NP–n antibodies not selected for cloning operation are randomly re-initialized as in (7).

After each evolution, we repeat from step B until one of the following conditions for convergence is satisfied. The conditions include restraining the number of iterations, maintaining error limits, or the both, whichever occurs earlier.

4 Experiments and Results

The GO terms [8] and GO annotation dataset [9] used in the experiments were downloaded from Saccharomyces Genome Database (SGD). We filtered out all regulatory relationships, and maintain only the relationships resulting in 15 main functional categories for Saccharomyces cerevisiae as given in [10]. Protein-Protein interaction data of Saccharomyces cerevisiae were obtained from BIOGRID [11] database (http://thebiogrid.org/). To reduce the effect of noise, the duplicated interactions and self-interactions were removed. The final dataset consists of 69,331 interaction protein pairs involving 5386 annotated proteins. Let $\{f_{r1}, f_{r2}, \ldots, f_{rn}\}$ be the set of n real functions of protein p and $\{f_{p1}, f_{p2}, \ldots, f_{pm}\}$ denotes the set of m functions predicted by protein function assignment scheme. It is obvious that $1 \leq m, n \leq D$. The three performance metrics used to evaluate the effectiveness of our proposed method are:

$$Precision = \frac{\sum_{j=1}^{m} \max_{i=1}^{n}(sim(f_{r_i}, f_{p_j}))}{m} \tag{11}$$

$$Recall = \frac{\sum_{i=1}^{n} \max_{j=1}^{m}(sim(f_{r_i}, f_{p_j}))}{n} \tag{12}$$

$$F - value = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{13}$$

An algorithm having higher values for the above metrics supersedes others. The evaluation of these metrics were conducted on test datasets by varying the number of proteins in the network N= [10, 200] for a particular unannotated protein. We have used only biological process for our experiment as assigning biological process to unannotated protein includes biological experiments which are very costly.In our study, we have compared the relative performance of the proposed scheme with Firefly Algorithm (FA) [12], Particle Swarm Optimization (PSO) [13], and also with the existing methods like Indirect Neighbor Method (INM) [14] and Neighbor Counting (NC) [3] in Table 1 and Fig. 2, Fig. 3, Fig. 4, for predicting functions of protein YBL068W.We report here results for only the above mentioned protein in order to save space.The omitted results for different proteins follow a similar trend as those stated above. The proposed approach was applied on annotated proteins of the network as the real functions for the same will be known to us. For CSA, $c_{\min}$ and $c_{\max}$ are set to 2 and 10 respectively.For all the evolutionary/swarm algorithm-based prediction schemes, the population size is kept at 50 and the maximum function evaluations (FEs) is set as 300000 and also best parametric set-up for already existing method is used. It is evident from Table1 and Fig. 2-4 that our algorithm outperforms others with respect to the aforementioned performance metrics irrespective of number of proteins in the network.

Table 1. Comparative analysis for predicting functions of YBL068W with N=80

Protein	Real Function	Real Functions Predicted by Different Algorithms				
		CSA	FA	PSO	INM	NC
YBL068W	GO:0009116	GO:0009116	GO:0009116	GO:0009116	x	x
	GO:0006015	GO:0006015	X	x	x	x
	GO:0009165	x	GO:0009165	GO:0009165	GO:0009165	x
	GO:0044249	GO:0044249	X	GO:0044249	GO:0044249	GO:0044249
	GO:0031505	GO:0031505	GO:0031505	x	GO:0031505	x
	GO:0009156	x	X	x	x	GO:0009156
	GO:0016310	GO:0016310	GO:0016310	GO:0016310	x	x

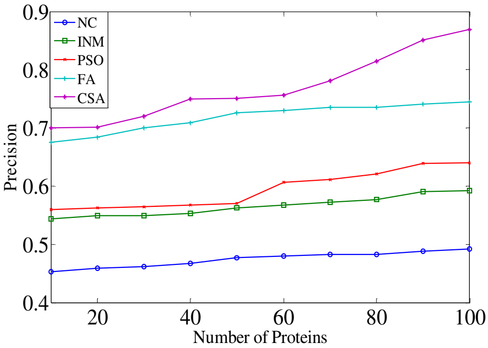

Fig. 2. Comparative analysis of precision plot

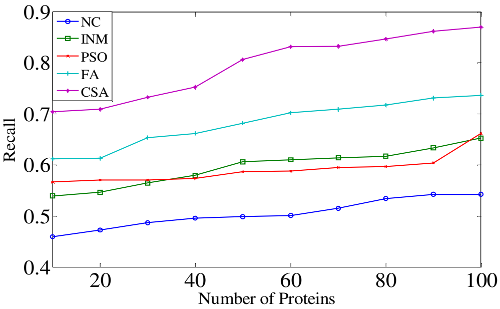

Fig. 3. Comparative analysis of recall plot

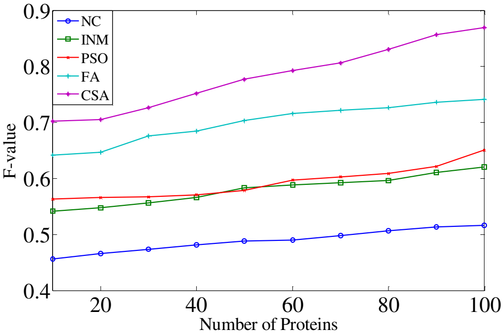

Fig. 4. Comparative analysis of F-value plot

5 Conclusion

Protein-Protein Interaction network play an important role in predicting the function of unannotated protein. In this paper we proposed a novel technique to predict the function of the unannotated protein based on the topological information as well as the functions of annotatedproteins of the PPI network of Saccharomyces Cerevisiae. Semantic similarity between proteins based on information content of the Go term is utilized to associate a function to an unannotated protein. Our approach does not entirely depend on the assumption that two interacting proteins are likely to have the same function or share functions. The simulation results reflect that our approach outperforms other state-of-the-art approaches.

References

1. Breitkreutz, B.J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D.H., Bähler, J., Wood, V., Dolinski, K., Tyers, M.: The BioGRID Interaction Database: 2008 Update. Nucleic Acids Research 36, D637– D640(2008)
2. Deng, M.H., Zhang, K., Mehta, S., Chen, T., Sun, F.Z.: Prediction of protein function using protein-protein interaction data. Journal of Computational Biology 10(6), 947–960 (2003)
3. Schwikowski, B., Uetz, P., Field, S.: A network of protein protein interactions in yeast. Nature Biotechnology 18, 1257–1261 (2000)
4. Lord, P.W., Stevens, R.D., Brass, A., Goble, C.A.: Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19(10), 1275–1283 (2003)
5. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of International Joint Conference for Artificial Intelligence, pp. 448–453 (1995)
6. Castro, D., Nunes, L., Zuben, F.J.V.: The clonal selection algorithm with engineering appli-cations. In: Proceedings of GECCO, pp. 36–39 (2000)

7. Felipe, C., Guimarães, F.G., Igarashi, H., Ramírez, J.A.: A clonal selection algorithm for optimization in electromagnetics. IEEE Transactions on Magnetics 41(5), 1736–1739 (2005)
8. Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J.: Gene ontology: tool for the unification of biology. Nature Genetics 25, 25–29 (2000)
9. Dwight, S., Harris, M., Dolinski, K., Ball, C., Binkley, G., Christie, K., Fisk, D., Issel Tarv-er, L., Schroeder, M., Sherlock, G.: Saccharomyces Genome Database (SGD) provides sec-ondary gene annotation using the Gene Ontology (GO). Nucleic Acids Research 30, 69–72 (2012)
10. Chowdhury, A., Konar, A., Rakshit, P., Janarthanan, R.: Protein Function Prediction Using Adaptive Swarm Based Algorithm. SEMCCO 2, 55–68 (2013)
11. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 34, D535–D539 (2006)
12. Yang, X.-S.: Firefly algorithms for multimodal optimization. In: Watanabe, O., Zeugmann, T. (eds.) SAGA 2009. LNCS, vol. 5792, pp. 169–178. Springer, Heidelberg (2009)
13. Kennedy, J.: Particle swarm optimization. In: Encyclopedia of Machine Learning, pp. 760–766 (2010)
14. Chua, H.N., Sung, W.-K., Wong, L.: Exploiting indirect neighbors antopological weight to predict protein function from protein protein interactions. Bioinformatics 22(13), 1623–1630 (2006)