# Ant Colony Optimization for Prediction of Compound-Protein Interactions

**Akhmad Rezki Purnajaya** [1]*
* Teknik Perangkat Lunak, Universitas Universal
rezkipurnajaya@gmail.com [1]

**Article Info**

***Article history:***

Received 2019-10-11
Revised 2019-10-18
Accepted 2019-10-29

***Keyword:***

*Ant colony optimization,*
*Compound-protein interaction,*
*Drug-target analysis,*
*Link prediction.*

## ABSTRACT

The prediction of Compound-Protein Interactions (CPI) is an essential step in drug-target analysis for developing new drugs. Therefore, it needs a good incentive to develop a faster and more effective method to predicting the interaction between compound and protein. Predicting the unobserved link of CPI can be done with Ant Colony Optimization for Link Prediction (ACO_LP) algorithms. Each ant selects its path according to the pheromone value and the heuristic information in the link. The path passed by the ant is evaluated and the pheromone information on each link is updated according to the quality of the path. The pheromones on each link are used as the final value of similarity between nodes. The ACO_LP are tested on benchmark CPI data: Nuclear Receptor, G-Protein Coupled Receptor (GPCR), Ion Channel, and Enzyme. Result show that the accuracy values for Nuclear Receptor, GPCR, Ion Channel, and Enzyme dataset are 0.62, 0.62, 0.74, and 0.79 respectively. The results indicate that ACO_LP has good accuracy for prediction of CPI.

## I. Introduction

The identification of Compund-Protein Interactions (CPI) is a key in the development of drugs, especially herbs. At the same time, many attempts have been made to molecular databank constitution to explore a relation possessed by compounds, synthetic compounds or natural compounds. They are extracted from animals, plants, or microorganisms. However, there is a little knowledge about the interaction between compund and protein. For example, US PubChem database stores 30 million chemical compunds, but information of some compounds interaction to protein target is very limited [1].

Predictive link is a method to predict a relationship between networks by concluding new link to known set of nodes and links. Determination of predictive link is done if a similarity between compound and protein is obtained. The use of binary data to determine similarity of compound and protein is expected to improve efficiency of computation process and accelerate the acquisition of similarity result and predictive link [1-3]. The computational method for predicting the CPI is thus essential in drug or herbal medicine studies. The method can reduce time, cost, and failure rate for discovering new drugs or herbal medicines [4].

Ant-Colony Optimization (ACO) is the first algorithm to find the optimal path in a graph, based on the behavior of ants that find their way between their colonies and food sources [5]. An artificial ant algorithm was created in predicting links by utilizing swarm intelligence. Each ant selects its path according to the pheromone value and the heuristic information (weight) in the link. The path passed by the ant is evaluated and the pheromone information on each link is updated according to the quality of the path. The final pheromones on each link are used as the interaction score between the nodes and the threshold is performed against whether the link is potentially interacting or not-interacting. The results of the experiments in study of Chen and Ling [6] on a number of real networks show that the predicted links with based ACO algorithms have higher prediction accuracy than other prediction methods that also carried out the same network experiments. In addition, in terms of ACO performance on predicted links or ACO_LP can maintain a low time complexity.

In this study used four Yamanishi datasets (i.e., Nuclear Receptor, G-Protein Coupled Receptor (GPCR), Ion Channel, and Enzyme) which are common benchmark dataset on CPI prediction. The CPI prediction result using BLM method will then be evaluated by using Area Under Curve (AUC) and accuracy [7]. AUC is a numerical measure to differentiate model performance and can be employed to show how successful the model rankings are by separating positive and negative observations. AUC is known to have proven to be a reliable performance measure for class imbalance problem [8].

## II. Methods

If the problem of CPI is applied to machine learning, then we can set the notation that i,j is the number of compound pairs $(C_1,C_1),...,(C_i,C_j)$ and the compound similarity $S(C_i,C_j)$ where compund Ci is connected to compound $C_j$. Protein similarity calculation is denoted that k,l is the number of compound pairs $(P_1, P_1),..., (P_k,P_l)$ and the protein similarity $S(P_k,P_l)$ where protein $P_k$ is connected to protein $P_l$. Actual of CPI is denoted that i is the number of compounds and k is the number of proteins $(C_1,P_1),...,(C_i,P_k)$.

### A. Materials

In this study used four Yamanishi datasets [9], Nuclear Receptor, GPCR, Ion channel, and Enzyme which are benchmark datasets on CPI prediction. These datasets were downloaded from http://web.kuicr.kyotou.ac.jp/supp/yoshi/drugtarget/. The datasets consist of adjacency matrix of the CPI data class, compound structure similarity matrix, and protein sequence similarity matrix. Yamanishi datasets reduced by selecting compound-protein interaction networks of Yamanishi datasets and filter into matrix of size 10x10.

Adjacency matrix of Nuclear Receptor, GPCR, Ion Channel, and Enzyme dataset consist of 93 non-interacting and 7 interacting pairs, 95 non-interacting and 5 interacting pairs, 93 non-interacting and 7 interacting pairs, and 98 non-interacting and 2 interacting pairs respectively.

### B. Ant Colony Optimization for Link Prediction

The basic idea is that a graph or network has a link that indicates a connection between each node whether it is a known or unknown link. Then the set of pheromones and heuristic information of each node, the greater the value of pheromones and heuristics, the higher the chance that the node has an interaction. Each ant passes through the graph until it reaches the n node and forms the path. In the first iteration the ants will form their paths, some nodes may be selected multiple times and some nodes may not be selected at all. The ants will move to paths that have pheromones and high heuristic information based on probability. After the ants to the last node or first iteration is completed then evaluated the quality of each path. Paths with higher total number of vertices (one node) will have higher node quality scores. Then the pheromones are updated by adding the quality score information. The renewed pheromone will affect the journey of the ants in the next iteration just as in the first iteration. The intensity of pheromone information at each node may increase or decrease due to the evaporation process in each iteration. Communication between the ants expressed in pheromones will find potential nodes and not. Finally, $T_{ij}$ pheromones on all vertices $(v_i, v_j)$ are used for interaction scores between nodes and their outputs are expressed in terms of final matrix scores [6].

The framework of our ant colony optimization based algorithm for link prediction ACO LP is as follows:

```
Input :  τ  : Adjacency pheromone matrix (small random)
         η  : Matrix heuristic information
         MaxIt: Maximum  number of iterations
         nAnt: Number of ants
         ε : Threshold for error feromon information
Output : Score : The Final pheromone matrix
Begin
1. t = 1;
2. Parameter initialization;
         λ : Pheromone parameter
         γ : Weight parameter
         C : Fitnesss path rate
         ρ : Evaporation rate
   Set the initial values of pheromone matrix τ and
   heuristic matrix η according to (2) dan (3)
         $\tau_{ij} = \lambda * (a_{ij} + \varepsilon)$        (2)
         $\eta_{ij} = \gamma * |\Gamma(i, j)|$        (3)
3. Repeat
4.       For k=1 to m do /* for the m ants*/
5.             Ant k randomly selects a node $s_1$
6.          for i = 1 to n-1 do
7.                Ant k selects the next node according to (4);
         $p_{ij}^{k} = \frac{\tau_{ij}^{\alpha} \cdot \eta_{ij}^{\beta}}{\sum_{k=1}^{n} \tau_{ik}^{\alpha} \cdot \eta_{ik}^{\beta}}$        (4)
8.          End for i
9.          Calculate the fitness of the path formed by ant k
            according to  (5)
         $Q(S) = C * \frac{1}{n} \sum_{i=1}^{n} d(s_i),$        (5)
10.      End for k
11.      Update the pheromone values according to (6), (7)
         and (8)
         $\Delta\tau_{ij}^{k}(t) = Q(S)$        (6)
         $\Delta\tau_{ij}(t) = \sum_{k=1}^{m} \Delta\tau_{ij}^{k}(t)$        (7)
         $\tau_{ij}(t+1) = \rho \cdot \tau_{ij}(t) + \Delta\tau_{ij}(t)$        (8)
12. Until max1≤i,j≤n | τij (t + 1)  - τij(t) | ≤ ε or t > Nc ;
13. Score = τ;
14. Output the score matrix Score;
End
```

The next step is to initialize some parameters of the ant colony optimization algorithm. The initial parameters of the ant colony optimization algorithm can be shown in table below:

TABLE I
INITIAL PARAMETERS OF ACO_LP

| Parameter | Value |
|---|---|
| Maximum number of iterations (MaxIt) | 7 |
| Number of ants (nAnt) | 4 |
| Phromone parameter (λ) | 0.1 |
| Threshold for error of phromone (ε) | 0.65 |
| Weight (γ) | 0.95 |
| Fitness path rate (C) | 0.04 |
| Evaporation rate (ρ) | 0.4 |

### C. *Receiver Operating Charateristic (ROC)*

To find out the accuracy ACO_LP algorithm, we can compare between the predicted links CPI results using of testing data with actual CPI of training data. The sensitivity, specificity, and accuracy are calculated by equation (9-12), TP, FP, TN, FN are the number of true positives, false positives, true negatives and false negatives, respectively [7].

$$Sensitivity = \frac{True\ Positive}{(True\ Positive + False\ Negative)} \quad (9)$$

$$Specificity = \frac{True\ Negative}{(True\ Negative + False\ Positive)} \quad (10)$$

$$AUC = \frac{Sensitivity + Specificity}{2} \quad (11)$$

$$Accuracy = \frac{True\ Positive + True\ Negative}{(Length\ Positive + Length\ Negative)} \quad (12)$$

### D. *Ratio of Positive Data (RPD)*

After the performance prediction is obtained, the ratio of positive data (interacting data class) can be calculated by shown in equation below:

$$Positive\ Data\ Ratio = \frac{n_1}{n_s \times n_p} \quad (13)$$

where $n_l$ is the number of interacting class, $n_s$ is the number of compounds, and $n_p$ is the number of protein [10].

## III. RESULT

In this study, interactions made by pharmaceutically useful compound-protein class of Yamanishi datasets with matrix of size 10x10 to find link prediction of ten compounds with ten other proteins. The compound-protein interaction network for each protein class using a bipartite graph representation. In the bipartite graph, the heterogeneous nodes correspond to either compounds or target proteins, and edges correspond to interactions between them. The edge is placed between a drug node and a target node if the protein is a known target of the compound.

In the ACO_LP algorithm, the ant agent is used to move randomly on a CPI network. Each ant selects its path according to the pheromone value and the heuristic information at its vertex. The initial value of pheromones on each node in the graph is set to random. The initial value of the heuristic information in each vertex is arranged according to the corresponding nodes based on the similarities between the compounds, the similarity between the proteins and the weight of the unobserved link and observed link interaction.

The path which the ant has passed is evaluated and the pheromone information in each node is updated according to the quality of the path it contains. Finally, the pheromones on each node are used as the value of the final score of the pair of nodes. Each dataset is predicted 10 times to get an average performance value. In this experiments, performance of ACO_LP algorithm can be shown in table and figure below:

TABLE II
VALUE OF THE PREDICTION PERFORMANCE

| Dataset | Sensitivity | Specificity | AUC | Accuracy |
|---|---|---|---|---|
| Nuclear Receptor | 0.399 | 0.638 | 0.519 | 0.621 |
| GPCR | 0.380 | 0.629 | 0.505 | 0.617 |
| Ion Channel | 0.271 | 0.771 | 0.512 | 0.736 |
| Enzyme | 0.200 | 0.802 | 0.501 | 0.790 |

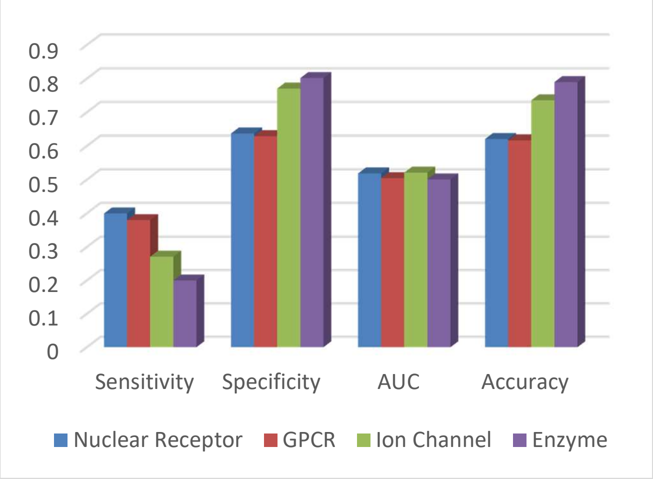

Fig. 1 Statistics of The Prediction Performance

From AUC and accuracy values, we conclude that the ACO_LP algorithm has good performance for prediction CPI. However, sensitivity values are much lower because adjacency matrix of CPI has more non-interacting data class than interacting data class. Prediction of CPI while ignoring data balance is that the prediction might be biased towards the majority class while ignoring the minority class. Futhermore, ACO_LP can accurately predict CPI although has imbalanced class in this case.

The ACO_LP algorithm is able to find new interacting pairs in CPI. This is evidenced by the increase in the percentage increase in ratio of positive data by 29% in the Nuclear Receptor dataset, 34% in the GPCR dataset, 13% in the Ion channel, and finally 16% in the Enzyme dataset, as shown in Table III.

TABLE III
PERCENTAGE INCREASE IN RATIO OF POSITIVE DATA

| | Dataset | | | |
|---|---|---|---|---|
| | **Nuclear receptor** | **GPCR** | **Ion channel** | **Enzyme** |
| **Training data** | | | | |
| Interacting | 7 | 5 | 7 | 2 |
| Non-interacting | 93 | 95 | 93 | 98 |
| Ratio of positive data | 7% | 5% | 7% | 2% |
| **Prediction data** | | | | |
| Interacting | 36 | 39 | 20 | 18 |
| Non-interacting | 64 | 61 | 80 | 82 |
| Ratio of positive data | 36% | 39% | 20% | 18% |
| **Percentage increase in Ratio of Positive Data** | **29%** | **34%** | **13%** | **16%** |

## IV. CONCLUSION

The empirical results show that the ACO_LP algorithm can achieve link prediction results by using fast computing time. So in this study, ACO_LP algorithm can be an alternative way to applied in Drug Target Interaction field. One of the reasons for the ACO_LP algorithm achieving high-quality results is because it uses pheromone and heuristic information that reflects local and global network structures so that it can take into consideration the attributes and structural information of networks from four of Yamanishi Dataset (Nuclear receptor, GPCR, Ion channel, Enzyme).

In the future, there is a potential that ACO_LP algorithm can not only be applied for CPI prediction, but it can also be used for drug-target analysis prediction which also has a class imbalance problem. This algorithm can provide more information about detect new CPI for drug repositioning.

## REFERENCES

[1] N. Ajay, “Morphological Similarity: A 3D Molecular Similarity Method Correlated With Protein-Ligand Recognition,’’ *Journal of Computer-Aided Molecular Design,* Page: 199–213, 2000.

[2] T. Yasuo, and Y. Yoshiro, “Scalable Prediction Of Compound-Protein Interactions Using Minwise Hashing,” *International Conference on Genome Informatics*,” doi:10.1186/1752-0509-7-S6-S3, 2013.

[3] H. Woong, Z. Xiaolei, G. Mark, and K, Daisuke, “3D Compound Comparison Methods and Their Application in Drug Discovery,” *Molecules,* Page:12841-12862; doi:10.3390/molecules200712841, 2004.

[4] S. Kim, D. Jin, and H. Lee, “Predicting drug-target interactions using drug-drug interactions”. *PLoS ONE* 8, http://dx.doi.org/10.1371/journal.pone.0080129, 2013.

[5] Engelbrecht and P. Andries, “Computational intelligence : an Introduction”, John Wiley & Sons Ltd. England, 2007.

[6] C. Bolun, and C. Ling. ”A Link Prediction Algorithm Based on Ant Colony Optimization”, *Applied Intelligence*; DOI: 10.1007/s10489-014-0558-5, 2014.

[7] P. Sonego, A. Kocsor and S. Pongor, “ROC analysis: Applications to the classification of biological sequences and 3D structures”. *Briefings in Bioinformatics* 9, 198-209, 2007.

[8] T. Fawcett, “ROC Graphs: Notes and practical considerations for data mining researchers”. *Patter n Recognition Letters* 31, 1-38, 2003.

[9] Y. Yamanishi, A. Michihiro, G. Alex, H. Wataru and K. Minoru, “Prediction of drug–target interaction networks from the integration of chemical and genomic spaces”, *Bioinformatics*, Vol. 24 ISMB, pages i232–i240, doi:10.1093/bioinformatics/btn162, 2008.

[10] CW. Harris, “Problems in measuring change”. Madison: University of Wisconsin Press. pp. 167–198, 1967.