An Evolutionary Approach for Analysing the Effect of Interaction Site Structural Features on Protein- Protein Complex Formation

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

[1] Department of ETCE,Jadavpur University, Kolkata, India
[2] Department of CSE,TJS Engg. College, Chennai, India
{chowdhuryarchana,pratyushar1}@gmail.com, konaramit@yahoo.co.in,
srmjana_73@yahoo.com

Abstract. Protein-protein complexes that dissociate and associate readily,often depending on the physiological condition or environment,play an important role in many biological processes.The impact of the features responsible for protein complex formation is not uniform. In this paper we have tried to rank the features required for stable protein-protein complex formation.We have employed Artificial Bee Colony with Temporal Difference Q learning algorithm to assign weights to the various atomic structure features. Experiments with data provide evidence that such an approach leads to improved clustering performance.

Keywords: protein-protein complex, interactome, Artificial Bee Colony algorithm, Temporal Difference Q learning, clustering.

1 Introduction

Most proteins form complexes to accomplish their biological functions [1]. Proteins perform and regulate many processes in the cell through interactions with other proteins. It has been estimated that 70% of proteins act through multi protein complexes in yeast [2].In order to predict an interaction site in situations where the binding partner is unknown, more emphasis is paid to properties which are observed in unbound structures which includes physicochemical parameters [3]and the evolutionary conservation of amino acid residues [4].Recent advances in highthroughput experimental methods for the identification of protein interactions have resulted in a large amount of diverse data that are some what incomplete and contradictory.Such experimental approaches studying protein interactomes have certain limitations that can be complemented by the computational methods for predicting protein interactions.

In this paper, we aim to analyze the effect of protein-protein interaction site structural features on clustering the protein-protein complex. For this task we have employed Artificial Bee Colony with Temporal Difference Q-Learning

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 656–661, 2013.

(ABC-TDQL)[5].It is shown that ABC-TDQL, inspired by global search capability of ABC [6]and from principles of reward and penalty of reinforcement learning,can give very promising results.

The rest of the paper is organized as follows: Section 2 give a brief idea about the structural features of protein-protein interaction site. Section 3 defines clustering problem in a formal language.Section 4 gives a brief description about the Artificial Bee Colony with Temporal Difference Q-Learning. Experiments and Results are provided in Section5.Section 6 concludes the paper.

2 The Structure of Protein-Protein Interaction Sites

The atomic structure of the recognition sites found in 63 protein-protein complexes of known three-dimensional structure is taken from [7].Sample dataset includes features such as: interface area of complex, number of hydrogen bond, number of water molecules,% interface area for non-polar atoms at interface,% interface area for polar atom,% interface area for charged atoms, number of interface atoms, packing density of buried atoms at interface with % interface area and packing density of atoms surrounded by solvent with % interface area.The dataset consists of protein protein complexs which belong to four groups of protease-inhibitor, antibody-antigen,enzyme-inhibitor and those that are involved in signal transduction.

3 Formulation of Problem

Problem Definition: Let $X_{N\times D} = \left\{\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_N\right\}$ be a set of N interaction sites of protein-protein complexes, each having D features. A partitional clustering algorithm tries to find out a partition of K clusters,such that the similarity of the interaction sites in the same cluster is maximum and interaction sites from different clusters differ as far as possible. Similarity is evaluated using the Euclidean distance $d(\vec{X}_i, \vec{X}_j)$.

Clustering Validity Index (CS Measure): Let $\vec{m}_i$ be the centroid of i-th cluster of protein-protein complexes. The CS measure [8]is then defined as

$$CS(K) = \frac{\frac{1}{k}\sum_{i=1}^{k}\left[\frac{1}{N_i}\sum_{x_i\in C_i} max_{x_q\in C_i}\left\{d(X_i, X_q)\right\}\right]}{\frac{1}{k}\sum_{i=1}^{k}\left[min_{j\in k, j\neq i}\left\{d(m_i, m_j)\right\}\right]} \tag{1}$$

CS measure is a function of ratio of sum of within-cluster-scatter to between cluster separation. It tries to find out clusters that have minimum within-cluster scatter (i.e. compact) and maximum between-cluster separation (i.e. well-separated)

Solution Representation and Fitness Function Evaluation: In the proposed method, for D dimensional interaction sites of N protein-protein comlexes,a solution is a vector of real numbers of dimension $D + D \times K$ where K is the number of clusters. The first D entries of $\vec{Z_i}$ represent the weights of the features which are randomly initialized and the remaining entries are reserved for K cluster centers, each D dimensional.Representation of the solution $\vec{Z_i}$ is shown in the following figure with its fitness function given in (2).

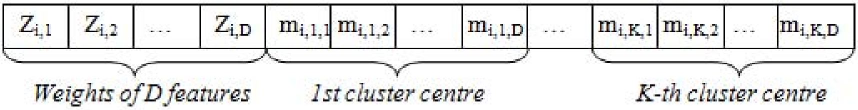

Fig. 1. Representation of a solution for ABC-TDQL-based clustering

$$f(\vec{Z_i}) = 1/CS_i(K) \tag{2}$$

4 Artificial Bee Colony with Temporal Difference Q-Learning (ABC-TDQL)

The ABC-TDQL [5] includes an ABC (with NP employed and NP onlooker bees) for global exploration and a TDQL for adaptive selection of scale factors for the individual members of the ABC using the Q-table. The row indices of the Q-table represent states $S_1, S_2, \cdots, S_{NP}$ of the population (based on fitness measure) obtained from the last update of the Q-table. The column indices correspond to uniformly quantized values of the scale factors $F_1, F_2, \cdots, F_{10}$ to be used in ABC. The steps of ABC-TDQL are given below.

1. Initialization: ABC-TDQL starts by initializing a population of NP, D-dimensional food sources (solutions)$\vec{Z_i}$ (G) at generation G=0 within the prescribed bounds and their fitness values (nectar amount)$fit(\vec{Z_i}\ (G))$ are evaluated for $i = [1, NP]$. The entries for the Q-table are initialized with small values.

2. Adaptive Selection of Parameters for Employed Bee Phase: The scale factor $F = F_j$ is randomly selected for individual food source from the meme pool satisfying(3)with r as a random number between (0,1).

$$\frac{\sum_{m=1}^{j-1} Q(S_i, 10F_m)}{\sum_{n=1}^{10} Q(S_i, 10F_n)} < r \leq \frac{\sum_{m=1}^{j} Q(S_i, 10F_m)}{\sum_{n=1}^{10} Q(S_i, 10F_n)} \tag{3}$$

3. Employed Bee Phase: An employed bee produces a modification $\vec{Z'_i}\ (G) = \left[Z_{i,1}(G), Z_{i,2}(G), \cdots, Z'_{i,j}(G), \cdots, Z_{i,D}(G)\right]$ on the position in her memory$\vec{Z_i}$ $(G) = [Z_{i,1}(G), Z_{i,2}(G), \cdots, Z_{i,j}(G), \cdots, Z_{i,D}(G)]$ and tests $fit(\vec{Z'_i}\ (G))$. $Z'_{i,j}(G)$ is computed using (4):

$$Z'_{i,j}(G) = Z_{i,j}(G) + 2 \times (F - 0.5) \times (Z_{i,j}(G) - Z_{k,j}(G)) \tag{4}$$

Here F is the scale factor in [0, 1] adaptively selected from the meme pool in the step 2, j and k are randomly selected such that $j \in [1, D]$, $k \in [1, NP]$, $k \neq i$. The bee replaces $Z_i(G)$ by $Z'_i(G)$ if $fit(\vec{Z'_i}\ (G))$>$fit(\vec{Z_i}\ (G))$.

4. Ranking of the Members: The food sources are sorted in descending order of probability of selection by onlooker bee as in (5).

$$prob(i) = fit(\overrightarrow{Z_i}\ (G))/\sum_{j=1}^{NP} fit(\overrightarrow{Z_j}\ (G)) \quad ,\forall i \tag{5}$$

5. Q-table Updating: Let a member at state S_i on selection of F_j moves to a new state S_k. Then $Q(S_i, 10F_j)$ will be updated following (7) with α and γ as the learning rate and discount factor respectively and

$$reward(S_i, 10F_j) = \left\{ \begin{matrix} fit(\overrightarrow{Z'_i}(G)) - fit(\overrightarrow{Z_i}(G)), \ \ if \ \ fit(\overrightarrow{Z'_i}(G)) > fit(\overrightarrow{Z_i}(G)) \\ -K\ (howeversmall), \quad otherwise \\ \end{matrix} \right\} \tag{6}$$

$$Q(S_i, 10F_j) = (1 - \alpha)Q(S_i, 10F_j) + \alpha(reward(S_i, 10F_j) + \gamma max_{F'} Q(S_k, 10F')) \tag{7}$$

6. Onlooker Bee Phase: Every onlooker bee probabilistically selects a food source depending on the probability value as stated in (5). Next steps 2 to 5 are repeated.
7. Scout Bee Phase: The abandoned food source is reinitialized randomly by the scout.
8. Convergence: After each evolution, steps are repeated until the condition for convergence is satisfied.

5 Experiment and Results

ABC-TDQL algorithm is run for 500 generations with population size of 50.The proposed algorithm is compared with other algorithms such as DE-TDQL, SaDE, DE/current-to-best/1, PSO and GA-based clustering method to evaluate the efficiency in assigning weights to the features.Table 1, represents the mean and the standard deviation (within parentheses) of final CS value, the intracluster distance, and the intercluster distance obtained for 50 independent runs of each of the algorithms.Table 2 represents the results comparing speed of various algorithms in terms of mean and standard deviation (within parenthesis) of the number of FEs, the CS cutoff value(0.10), the final intracluster distance, and the final intercluster distance over 50 independent runs for each algorithm. As the nominal partitions of the dataset is known, the mean number of misclassified data points for different features is calculated and presented in Table 3.

Table 4 represents various features arranged according to their decreasing weights obtained using the proposed algorithm.The entries in Table 4 are relevant and supported by the fact that the most important feature which helps to form stable protein- protein complexes include hydrophobic effect, which is gained from the surfaces buried in the recognition sites i.e. interface area.Then next is the contribution by non polar atoms, as buried atoms are more non-polar, 63% on average, in comparison to other interface atoms which remain partly accessible.

Table 1. Final solution

Algorithm	CS Measure	Intra-clu.Dist	Inter-clu.Dist
ABC-TDQL	0.133770 (0.01365)	1.034500 (0.01054)	7.881300 (0.01388)
DE-TDQL	0.263210 (0.02642)	1.197500 (0.01210)	7.918300 (0.01054)
SaDE	0.409070 (0.04092)	1.900700 (0.01908)	6.516400 (0.02642)
DE/current-to-best/1	0.609710 (0.06154)	4.248900 (0.04252)	4.906400 (0.04344)
PSO	0.814950 (0.08207)	5.620500 (0.05624)	3.721800 (0.05824)
GA	0.947700 (0.09491)	7.014900 (0.07139)	1.337700 (0.07374)

Table 2. Speed of algorithms

Algorithm	No. of FEs	Intra-clu. Dist	Inter-clu. Dist
ABC-TDQL	144321.29 (14.5320)	1.878600 (0.01932)	9.227900 (0.01711)
DE-TDQL	144321.29 (14.5320)	1.878600 (0.01932)	9.227900 (0.01711)
SaDE	268983.67 (27.5100)	3.663000 (0.03716)	7.638800 (0.03709)
DE/current-to-best/1	457415.26 (46.9980)	5.207600 (0.05220)	6.223300 (0.05159)
PSO	650836.54 (65.1550)	6.661200 (0.06700)	4.542400 (0.07000)
GA	843198.42 (84.4520)	8.233900 (0.08280)	2.417200 (0.08673)

Table 3. Mean classification error over nominal partition and standard deviation

ABC-TDQL	DE-TDQL	SaDE	DE/current-to-best/1	PSO	GA	Stat. Sig.
1.55(0.01)	1.61(0.01)	2.28(0.03)	4.98(0.05)	6.34(0.06)	8.06(0.08)	+

Table 4. Sorted Features

Index	Sorted Features
f01	Interface Area
f03	Number of Hydrogen Bonds
f04	% Interface Area of Polar Atoms
f05	V/V0
f06	% Interface Area for V/V0
f07	V'/V0
f08	% Interface Area for V'/V0
f09	Number of interface atoms
f10	Number of Water Molecules
f11	% Interface Area of Charged Atoms

Table 5. Statistical comparision

Classifier - algorithm	n_{01}	n_{10}	Zj	Comment
DE-TDQL	19	28	1.3617	Accepted
SaDE	18	30	2.5208	Accepted
DE/current-to-best/1	12	40	14.019	Rejected Rejected
PSO	9	45	22.685	Rejected
GA	5	47	32.326	Rejected

On average, neutral polar groups contribute 29% to the interface area. Thus, the fraction contributed by non polar groups is higher than the fraction contributed by neutral polar groups, whereas charged groups contribution is the least. Then comes electrostatic energy from the hydrogen bonds. For interface atoms that are buried in 63 complexes, which represents one-third of all interface atoms, the packing density derived from the Voronoi volume is within 7% of that of the protein interior. Since two-thirds of the interface atoms have non-zero solvent accessibility, their volume cannot be calculated in the absence of information of the structure of the solvent molecules with which they are in contact hence their contribution to stable complex formation is very less. McNemer's test is applied to determine the performance of two algorithms used for clustering of data points and is expressed as follows :

$$Z = (|n_{01} - n_{10}| - 1)^2/(n_{01} + n_{10}) \tag{8}$$

where n_{01} and n_{10} give the number of samples misclassified by one algorithm and not by the other with which comparision is done and vice versa.Table 5 represents the Z and the null hypothesis is rejected if $Z > \chi^2_{1,0.95} = 3.841459$, which indicates that the probability of the null hypothesis is correct only to a level of 5%. It is apparent from Table 5 that ABC-TDQL-based clustering technique has outperformed most of its competitors in inferring clusters of the data points except DE-TDQL and SaDE.

6 Conclusion

This paper proposes a novel technique for selecting features based on a clustering algorithm. The observations suggest that the structural features of the recognition sites, if found in other systems where proteins exist as stable independently, will interact to form specific complexes without major conformational changes, since the conformation of complexes considered for observation were almost same. Hence further research may involve investigating the recognition site features of complexes that involve large conformational change.

References

1. Gavin, A.C., et al.: Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415(6868), 141–147 (2002)
2. Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., et al.: Functional organization of the yeast proteome by systematic analysis of proteincomplexes. Nature 415, 141–147 (2002)
3. Janin, J., Chothia, C.: The structure of protein-protein recognition sites. J. Biol. Chem. 265, 16027–16030 (1990)
4. Lichtarge, O., Sowa, M.E.: Evolutionary predictions of binding surfaces and interactions. Curr. Opin. Struct. Biol. 12, 21–27 (2002)
5. Rakshit, P., Konar, A., Das, S., Nagar, A.K.: ABC-TDQL: an adap-tive memetic algorithm. In: 2013 IEEE Symposium Series on Computational Intelligence (accepted, to be published 2013)
6. Bhattacharjee, P., Rakshit, P., Goswami, I., Konar, A., Nagar, A.K.: Multi-robot path-planning using artificial bee colony optimization algorithm. In: NaBIC 2011, pp. 219–224 (2011)
7. Lo Conte, L., Chothia, C., Janin, J.: The Atomic Structure of Protein-Protein Recognition Sites. J. Mol. Biol. 285, 2177–2198 (1999)
8. Chou, C.H., Su, M.C., Lai, E.: A new cluster validity measure and its application to image compression. Pattern Anal. Appl. 7(2), 205–220 (2004)