Predicting Multicomponent Protein Assemblies Using an Ant Colony Approach

Vishwesh Venkatraman and David W Ritchie

INRIA Nancy Grand Est,
615 Avenue du Jardin Botanique,
54506, Vandoeuvre-lès-Nancy, France
vishwesh.venkatraman@inria.fr, dave.ritchie@inria.fr

Abstract

Biological processes are often governed by functional modules of large protein assemblies such as the proteasomes and the nuclear pore complex, for example. However, atomic structures can be determined experimentally only for a small fraction of these multicomponent assemblies. In this article, we present an ant colony optimization based approach to predict the structure of large multicomponent complexes. Starting with pair-wise docking predictions, a multigraph consisting of vertices representing the component proteins and edges representing scored transformations is constructed. Thus, the assembly problem corresponds to identifying minimum weighted spanning trees that yield arrangements of components with few atomic clashes. The utility of the approach is demonstrated using protein complexes taken from the Protein Data Bank. Our algorithm was able to identify near-native solutions for 5 of the 6 cases tested, including one 6-component complex. This demonstrates that the ant colony model provides a useful way to deal with highly combinatorial problems such as assembling multicomponent protein complexes.

Key words

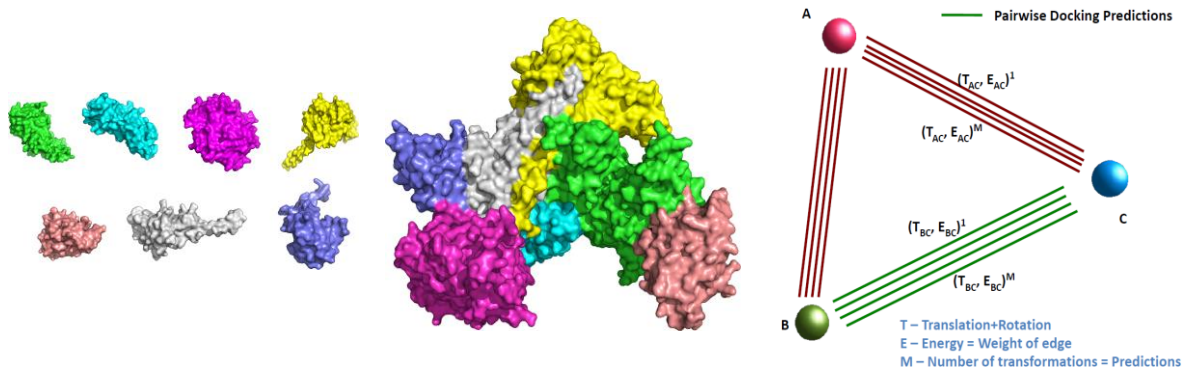
Macromolecular assembly, multiple protein docking, *Hex*, ant colony optimization

1 Introduction

Biological cells contain many large macromolecular units that govern complex and important biological processes. Elucidating the structural characteristics of these multicomponent protein complexes can contribute towards a better understanding of the interactions between these macromolecules. However, due a number of practical reasons, atomic structures can be determined experimentally only for a small fraction of these multimolecular assemblies [1]. Efforts to model such structures computationally have therefore become increasingly valuable. While pair-wise docking has been used to predict binary interactions, only very few groups have attempted to model multicomponent complexes [2]. Here, we present an ant colony optimization approach to predicting the structure of multicomponent complexes. The efficacy of the algorithm is demonstrated using three examples taken from the Protein Data Bank (PDB) [3].

2 Assembling Macromolecules

Starting with a set of N component structures, as illustrated in Figure 1, the goal is to predict the native complex formed by the interactions between the proteins. In this case, combinatorial complexity is a major problem because for N subunits that form the complex, there are $N^{N-2}K^{N-1}$ possible solutions where K is the number of predictions for each pair of subunits. Hence both N and K have to be relatively small for the calculation to be tractable.



A **B** **C**

Figure 1: The combinatorial assembly problem. (A) The component proteins. (B) The target complex to be predicted. (C) The structure prediction problem can be formulated as a graph with proteins at the vertices and with parallel edges representing possible docking solutions.

2.1 The Prediction Algorithm

Ant colony optimization (ACO) is a stochastic global optimization technique inspired by the foraging behaviour of ants which use pheromone trail information to guide the search for promising solutions. In this work, we use an ACO approach to address the multicomponent protein assembly prediction problem. The first step consists of using the *Hex* docking algorithm [4] to generate pair-wise docking predictions for the N input protein structures yielding $N(N-1)/2$ sets of predictions. In this study, we have chosen to retain the best K=100 orientations for each pair. A complete multigraph composed of vertices representing the constituent proteins and parallel edges representing the high scoring transformations is then constructed (Figure 1C). A connected subgraph, or spanning tree, of this large graph represents a solution, *i.e.* a predicted complex. The problem therefore corresponds to identifying minimum weighted spanning trees (MSTs). More formally, the problem can be expressed as an edge-weighted k-cardinality problem [5] although the exact weights of the edges are not known *a priori*. This problem has been shown to be NP-hard [6]. The ACO algorithm starts with a population of ants, each representing a randomly generated MST (set of graph edges). Each edge is associated with a pheromone value (initially set to a random value) that is updated according to the lowest weighted tree found during each iteration and the best solution found so far. A special feature of this algorithm is that at each instance, the ant must calculate atomic clashes which will contribute to the final score. Each ant (solution) is thus assigned a total score based on the calculated *Hex* energy for each edge of the MST and the number of atomic clashes between the components. During each iteration, the algorithm outputs the solution with the lowest weight (best score). The solutions are assessed using the root mean square deviation (RMSD) distance (measured in Angstrom units) between the C-α atoms of the predicted and the experimental structures and are ranked according to the calculated total scores.

3 Results and Discussion

We have implemented our own version of Blum's algorithm [5] in C++. The approach has been tested using three target complexes (in their bound form) taken from the PDB. These include (i) the three-component complex VHL-elonginC-elonginB, PDB code 1VCB, (ii) the three-component IκBα/NF-

κB complex, PDB code 1IKN, and (iii) the seven-component Arp2/3 complex, PDB code 1K8K. For this third target, additional test cases were constructed by selecting 4, 5 and 6 components of the complex. The ACO algorithm was used with a population size of 10 and was iterated 1000 times. The results are summarised in Table 1. A RMSD of less than 10Å is considered to be close to a near-native solution. All calculations were performed on a Linux machine with Quad Intel Core2 2.83GHz CPUs. Table 1 shows that the algorithm is able to find near-native solutions for 5 of the 6 test cases. We believe these good results could be further improved by using a better hydrophobicity-based scoring function to eliminate more infeasible orientations.

Complex	Chains	Time (seconds)	Rank	RMSD (Å)	Best RMSD (Å)
1VCB	A,B,C	2627	1	0.58	0.58
1IKN	A,C,D	4637	1	9.17	0.88
1K8K	A,B,D,E	7408	1	4.96	2.19
1K8K	A,B,D,E,F	10118	2	9.48	2.99
1K8K	A,B,D,E,F,G	11646	15	4.63	3.53
1K8K	A,B,C,D,E,F,G	22016	-	-	10.21

Table 1: Results of the ant colony protein assembly algorithm. The Rank and RMSD columns show the rank and RMSD of the first near-native solution found. The Best RMSD column indicates the lowest RMSD solution obtained by the ACO algorithm which does not always correspond to the first near-native solution.

References

[1] ALBER F., DOKUDOVSKAYA S., VEENHOFF L.M., ZHANG W., KIPPER J., DEVOS D., SUPRAPTO A., KARNI-SCHMIDT O., WILLIAMS R., CHAIT B.T., ROUT M.P., and SALI A, *Determining the architectures of macromolecular assemblies*. Nature, 450(7170), pp 683-694, 2007.

[2] INBAR Y., BENYAMINI H., NUSSINOV R. and WOLFSON H.J., *Prediction of multimolecular assemblies by multiple docking*, Journal of Molecular Biology, 349, pp 435-447, 2005.

[3] BERNSTEIN F.C., KOETZLE T.F., WILLIAMS G.J., MEYER JR. E.E., BRICE M.D., RODGERS J.R., KENNARD O., SHIMANOUCHI T. and TASUMI M., *The Protein Data Bank: A Computer-based Archival File For Macromolecular Structures.* Journal of Molecular Biology, 112, pp 535-542, 1977.

[4] RITCHIE D.W. and VENKATRAMAN V., *Ultra-Fast FFT Protein Docking On Graphics Processors.* Bioinformatics, 26, pp 2398-2405, 2010.

[5] BLUM C., *Ant Colony Optimization For The Edge-weighted k-cardinality Tree Problem*. Proceedings of the Genetic and Evolutionary Computation Conference, San Francisco, USA, pp. 27-34, 2002.

[6] FISCHETTI M., HAMACHER H.W., JØRNSTEN K. and MAFFIOLI F., *Weighted k-cardinality trees: Complexity and polyhedral structure.* Networks, 24, pp 11–21, 1994.