A Hybrid Ant Colony Optimization for the Prediction of Protein Secondary Structure

Chao CHEN, Yuan Xin TIAN, Xiao Yong ZOU*, Pei Xiang CAI, Jin Yuan MO

School of Chemistry and Chemical Engineering, Sun Yat-sen University, Guangzhou 510275

Abstract: Based on the concept of ant colony optimization and the idea of population in genetic algorithm, a novel global optimization algorithm, called the hybrid ant colony optimization (HACO), is proposed in this paper to tackle continuous-space optimization problems. It was compared with other well-known stochastic methods in the optimization of the benchmark functions and was also used to solve the problem of selecting appropriate dilation efficiently by optimizing the wavelet power spectrum of the hydrophobic sequence of protein, which is the key step on using continuous wavelet transform (CWT) to predict α-helices and connecting peptides.

Keywords: Ant colony algorithm, global optimization, wavelet power spectrum, protein structure prediction.

The prediction of secondary structure from amino acid sequence is the most familiar and well-defined problem, and is often regarded as the first step in understanding the protein folding problem[1]. In the past many different approaches have been developed for the secondary structure prediction: statistical, symbolic machine learning, neural network, *etc*. Qiu *et al.*[2] adopt continuous wavelet transform (CWT) to predict α-helices and short peptides connecting α-helices and β-strands, based on the theory of hydrophobic minima. The key step is to select the appropriate dilations, which has an enormous effect on the results of prediction.

The traditional estimators for the power spectrum of a signal have several disadvantages, especially when dealing with samples that are finite sized, have a complex geometry, are sampled irregularly, or in which the mean density is uncertain or varying. So, efforts on developing new estimators have never stopped. An estimator based on the wavelet transform was proposed by T. Christopher[3], wavelet spectral densities being additive contributions to the total energy of the signal.

Ant colony optimization (ACO), recently proposed by M. Dorigo[4], is inspired by adaptive natural real ants' behaviors and belongs to the class of biologically inspired heuristics. ACO algorithm was originally used for solving large-scale combinatorial discrete optimization problems including the traveling salesman problem (TSP), the quadratic assignment problem (QAP), the job-shop scheduling problem (JSP), *etc*. However, there are few studies on ACO for continuous optimization problems, which

[*] E-mail: ceszxy@zsu.edu.cn

have attracted more and more attention of the scientists.

In this article, a hybrid ant colony optimization (HACO), based on the concept of ant colony optimization and the idea of population in genetic algorithm (GA), was developed for the continuous function optimization. HACO consists of the following main steps.

1. Prepare parameters for the algorithm, including the numbers of artificial ants (*ant_num*) and sub-domains divided by the given domain (*N*), evaporation rate of pheromone (ρ), population size (*pop_size*) for evolution, the probabilities of crossover (*p_crossover*) and mutation (*p_mutation*), *etc*.
2. The population is randomly initialized and evaluated.
3. According to the strategy introduced in ref.[5], a procedure to check similarity between individuals is adopted, which can be helpful to keep diversity in the population and avoid converging to local minima.
4. Ant *k* move to region *j* and search with the probability shown below:

$$p_{ij}{}^{k}(t) = \tau_{ij}(t) \;/\; \sum_{r=1}^{N} \tau_{ir}(t)$$

5. Update the pheromone of the selected region *j*.
6. Generate new solutions by selecting different individuals and carrying out the operation in GA, such as crossover and mutation, according to the probability of *p_crossover* and *p_mutation*, and then evaluate them.
7. Globally update the pheromone.
8. Exchange the worse individuals with the better solutions searched out in step 6.
9. Before the maximum number of evaluations is reached, go to step 3.
10. Store and output the best solution.

To assess the HACO algorithm, we carried out benchmark problems using several standard multidimensional test functions with multiple minima. The following values of parameters are used: $20 \le ant\_num \le 50$, $0.5 \le \rho \le 0.7$, $50 \le N \le 100$, $pop\_size \le 8$, $p\_crossover = 0.9$, $p\_mutation = 0.08$. The stop criteria of HACO for each function are $|f - f_{\min}| < 10^{-6}$ or reaching 1,000 iterations. The performance of HACO is compared with other stochastic optimization methods, and the criterion for comparison is function

Table 1 Number of function evaluations required by HACO and other global methods to reach the optimal minima of standard test functions

Method	Test function					
	BR	CA	GP	H3	RA	SH
PRS	4850	—	5,125	5,280	—	6,700
SDE	2700	10,822	5,439	3,416	—	241,215
FAEA	394	303	490	488	2,762	446
HACO	320	405	384	530	2,670	454

Function abbreviations: BR, Branin; CA, Camelback; GP, Goldstein-Price; H3, Hartman; RA, Rastrigin and SH, Shubert. Methods abbreviations: PRS is the pure random search; SDE is the stochastic method; and FAEA refers to the fast annealing evolutionary algorithm. Results of PRS, FAEA are from ref.[5]. Results of SDE are from ref.[6]. Results of HACO are the average results over 100 continuous runs.

Table 2 Precision of each variable (x_1 to x_5) for HACO and other methods

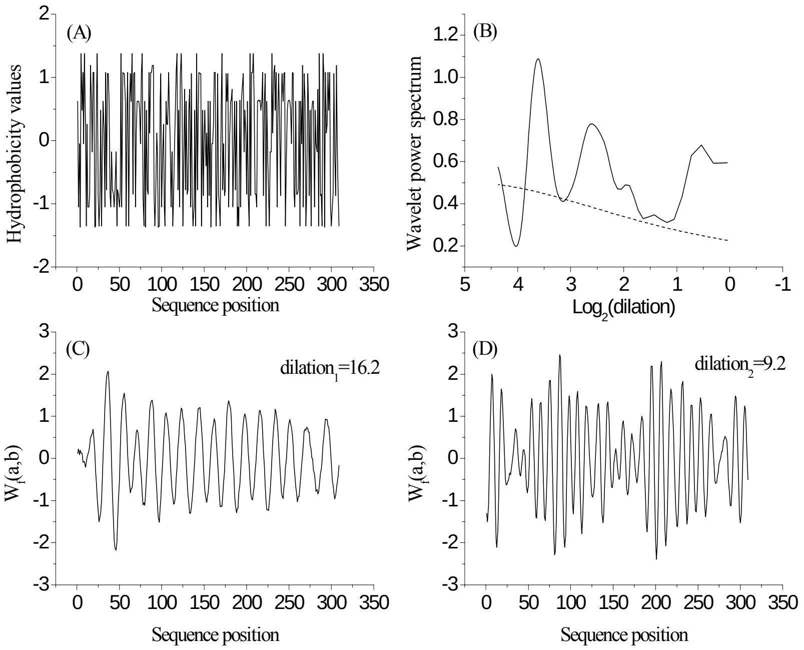
Variable	FSA (100,000)	SAS (3,710)	TRUST (89)	FAEA (1,413)	HACO (1,130)	Exact
x_1	-2.702844	-2.903506	-2.90353	-2.9035340	-2.90375	-2.903534
x_2	-3.148829	-2.903527	-2.90353	-2.9035340	-2.90338	-2.903534
x_3	1.099552	1.000241	1.00004	0.9999996	0.99964	1
x_4	1.355916	0.999855	0.99997	0.9999996	1.00012	1
x_5	1.485936	1.000194	0.99997	1.0000004	1.00080	1

The amount given in parentheses is the number of function evaluations. Methods abbreviations: FSA, the fast simulated annealing; SAS, the stochastic approximation paradigm; TRUST, the terminal repeller unconstrained subenergy tunneling; and FAEA, the fast annealing evolutionary algorithm. Results of FAEA are from ref.[5]. Results of FSA, SAS and TRUST are from ref.[6]. Results of HACO are the average results over 100 continuous runs.

evaluations. The results in **Table 1** and **2** demonstrate that HACO is a fast and global method and produces consistent and accurate results.

In this section HACO is utilized to solve the problem of selecting appropriate dilations by optimizing the wavelet power spectrum of the hydrophobic sequence of protein. There are three main steps. Firstly, transform the amino acids of protein into the sequence of Fauchereand hydrophobic free energies per residue. Secondly, estimate the power spectrum of the hydrophobic sequence with the method of wavelet transform, adopting the Morlet as the wavelet basis. Thirdly, optimize the power spectrum with HACO and search out the dilations corresponding to its minima, under which the α-helices and connecting peptides can be well predicted. As an example, 1gca was chosen from the Brookhaven Protein Databank (PDB) to describe the whole procedure. **Figure 1** shows the predicted results, of which **(A)** is the hydrophobic values of plot and **(B)** is the corresponding wavelet power spectrum plot at dilation range of 1-20. In **Figure 1(B)** the solid line is the wavelet power spectrum plot and the dashed line is the 95% confidence level, which definition refers to ref.[3]. **Figure 1(B)** shows that there are only two local minima below the confidence line, where the corresponding dilations can be optimized by HACO. The two dilations are 9.2 and 16.2, under which connecting peptides and α-helices can be predicted accurately. **(C)** and **(D)** shows the CWT plots of hydrophobic value sequences with the two dilations, in which the number and the relevant positions of the α-helices and connecting peptides of protein 1gca can be extracted according to the minima. In order to investigate the feasibility of our method, we randomly select forty proteins from PDBsum database as the test objects, including ten all-α proteins, ten all-β, ten α+β and ten α/β. The results indicate that the appropriate dilations for all the selected proteins are approximately equal to 9 and 16. But, the average prediction accuracies of α-helices and connecting peptides, being 79.0% and 85.4%, are prior to the ones, being 77.4% and 80.8% of ref.[2], in which the dilations were set equal to 9 and 16 exactly, but arbitrarily.

Figure 1 The predicted results of protein 1gca secondary structure



Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.20475068) and the Guangdong Provincial Natural Science Foundation (No.031577).

References

1. H. Wako, T. L. Blundell, *J. Mol. Biol.*, **1994**, *238*(5), 682.
2. J. D. Qiu, R. P. Liang, X. Y. Zou, J. Y. Mo, *Talanta,* **2003**, *61*(3), 285.
3. T. Christopher, P. C. Gilbert, *Bull. Amer. Meteor. Soc.,* **1998**, *79*(1), 61.
4. M. Dorigo, V. Maniezzo, A. Colorni, *IEEE Syst. B,* **1996**, *26*(1), 29.
5. W. S. Cai, X. G. Shao, *J. Comput. Chem.,* **2002**, *23*(4), 427.
6. J. Barhen, V. Protopopescu, D. Reister, *Science*, **1997**, *276*(16), 1094.

Received 17 January, 2005