UNIVERSITY OF M'sila

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Department of Computer Science

Thesis Graduation Diploma for obtaining

Master in Academic Computing

Domain: Computer Mathematics

Sector: Computer basic

Option: Advanced Information System

TOPIC

Motif Finding Using Ant Colony Optimization

Supervised by:

Mr. Salim BOUAMAMA

Ms. Ouarda ASSAS

Prepared by:

Leyla SALMI

Promotion: 2011 /2012

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

CHAPTER 2: RELATED WORK

CHAPTER 3: ANT COLONY OPTIMIZATION METAHEURISTIC (ACO)

CHAPTER 4: THE PROPOSED APPROACH MOTIF FINDING USING ANT COLONY OPTIMIZATION (MFACO)

CHAPTER 5: IMPLEMENTATION DETAILS AND EXPERIMENTAL RESULTS

CHAPTER 6: CONCLUDING REMARKS AND FUTURE WORK

ABSTRACT

A challenging problem in molecular biology is to identify of the specific binding sites of transcription factors in the promoter regions of genes referred to as motifs. This thesis presents an Ant Colony Optimization approach and how it can be used to provide the motif finding problem with promising solutions. The proposed approach incorporates a modified form of the Gibbs sampling technique as a local heuristic optimization search step. Further, it searches both in the space of starting positions as well as in the space of motif patterns so that it has more chances to discover potential motifs. The approach has been implemented and tested on some datasets including the *Escherichia coli* CRP protein dataset. Its performance was compared with other recent proposed algorithms for finding motifs such as MEME, MotifSampler, BioProspector, and in particular Genetic Algorithms. Experimental results show that our approach could achieve comparable or better performance in terms of motif accuracy within a reasonable computational time.

Keywords: Bioinformatics, Ant Colony Optimization, Motif Finding, Meta-heuristics

REFERENCES

[1]. Liu X, Brutlag D, Liu J. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pacific Symposium on Biocomputing 2001; 6: 127-138. [online software] [accessed February 2012]. Available from URL http://bioprospector.stanford.edu/cgi-bin/BPsearch.pl

[2]. Human Genome Project Information Web Site. [online] [accessed January 2012]. Available from URL http://www.ornl.gov/sci/techresources/Human_Genome/

[3]. Krane DE, Raymer ML. Fundamentals Concepts of Bioinformatics. San Francisco: Benjamin Cummings; 2003: 314.

[4]. Jones NC, Pevzner PA. An introduction to Bioinformatics Algorithms. Cambridge: MIT Press; 2004: 435.

[5]. Campbell NA, Reece JB. Biology, 7th ed. Benjamin Cummings; 2005. [on CD-ROM]

[6]. Attwood T, Parry-Smith D. Introduction to Bioinformatics. London: Addison Wesley; 1999.

[7]. Man KF, Tang KS, Kwong S. Genetic Algorithms: concepts and designs. London: Springer-Verlag; 1999.

[8]. Lesk A, Introduction to Bioinformatics. New York: Oxford University Press; 2002.

[9]. Bulyk M. Computational prediction of transcription-factor binding site locations. Genome Biology 2003; 5(201).

[10]. Keith JM et al. A simulated annealing algorithm for finding consensus sequences. Bioinformatics 2002; 18(11): 1494-1499.

[11]. Liu FM, Tsai JP, Chen RM, Chen SN, Shih SH. FMGA: Finding motifs by genetic algorithm. IEEE 4th Symposium on Bioinformatics and Bioengineering (BIBE'04); 2004; 459-466.

[12]. Thijs G et al. Higher-order background model improve the detection of promoter regularity elements by Gibbs sampling. Bioinformatics 2001; 17(12): 1113-1122

[13]. Lawrence C, Altschul S, Boguski M, Liu J, Neuwald A, Wootton J. Detecting Subtle Sequence Signals: A gibbs sampling strategy for multiple alignments. Science 1993 Oct 8; 262(5131): 208-214.

[14]. Pevzner PA, Sze S. Combinatorial approaches to finding subtle signals in DNA sequences. Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB-00); AAAI Press; 2000; San Diego, California; 269-278.

[15]. Keich U, Pevzner PA. Finding motifs in the twilight zone. Bioinformatics 2002; 18(10): 1374-1381.

[16]. Price A, Ramabhadran S, Pevzner PA. Finding subtle motifs by branching from sample strings. Bioinformatics 2003; 19(Suppl. 2): 149-155.

[17]. Hertz G, Stormo G. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 1999; 15(7): 563-577.

[18]. Stormo GD. DNA binding sites: representation and discovery. Bioinformatics 2000, 16(1): 16-23.

[19]. Che D, Song Y, Rasheed K. MDGA: Motif discovery using a genetic algorithm. Proceedings of the 2005 Conference on Genetic and Evolutionary Computation (GECCO '05); ACM Press; 2005 June 25-29; Washington, USA; 447-452.

[20]. Stormo GD, Hartzell GW. Identifying protein-binding sites from unaligned DNA fragments. Proc. Natl. Acad. Sci. USA; February 1989; 86(4): 1183-1187.

[21]. Smyth B. Computing Patterns in Strings. Harlow, England: Pearson Education; 2003: 423-43.

[22]. Dorigo M, Gambardella LM. Ant colonies for traveling salesman problem. BioSystems 1997; 43: 73-81.

[23]. Dorigo M, Maniezzo V, Colorni A. The ant system: optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics 1996; Part B, 26(1): 29-42.

[24]. Stützle T, Dorigo M. ACO algorithms for the traveling salesman problem. In: Miettinen K, Mäkelä M, Neittaanmäki P, Periaux J, editors. Evolutionary Algorithms in Engineering and Computer Science. Chichester, UK: John Wiley & Sons; 1999: 163- 183.

[25]. Gambardella LM, Dorigo M. An ant colony system hybridized with a new local search for the sequential ordering problem. INFORMS Journal on Computing 2000; 12(3): 237-255.

[26]. Chu D, Till M, Zomaya A. Parallel ant colony optimization for 3D protein structure prediction using the HP lattice model. IEEE Proceedings of the 19th International Parallel and Distributed Processing Symposium (IPDPS'05); 2005.

[27]. Wang G, Gong W, Kastner R. Instruction scheduling using MAX-MIN ant colony optimization. Great Lakes Symposium on Very Large Scale Integration (GLSVLSI'2005); ACM Press; 2005 April 17-19; Chicago, Illinois, USA.

[28]. Bailey L, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology; AAAI Press; Aug 1994; Menlo Park, California; 28 36. [online software] [accessed February 2012]. Available from URL http://meme.sdsc.edu/meme/meme.html.

[29]. Thijs G. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. Journal of Computational Biology 2002; 9(2): 447-464. [online software] [accessed February 2012]. Available from URL http://homes.esat.kuleuven.be/~thijs/Work/MotifSampler.html.

[30]. Zheng W. Relation between weight matrix and substitution matrix: motif search by similarity. Bioinformatics 2005; 21(7): 938-934.

[31]. W. Thompson, E. Rouchka, and C. Lawrence. Gibbs Recursive Sampler: finding transcription factor binding sites. Nucleic Acids Research Journal 2003; 31(13): 3580- 3585.

[32]. Notredame C, Higgins D. SAGA: Sequence alignment by genetic algorithm. Nucleic Acids Research Journal 1996; 24(8): 1515-1524.

[33]. Seehuus R, Tveit A, Edsberg O. Discovering biological motifs with genetic programming. Proceedings of the 2005 Conference on Genetic and Evolutionary Computation (GECCO '05); ACM Press; 2005 June 25-29; Washington, USA; 401-408.

[34]. Buhler J, Tompa M. Finding motifs using random projections. Journal of Computational Biology 2002; 9(2): 225-242.

[35]. Cordan O, Herrera F, Stützle T. A review on the ant colony optimization metaheuristic: basics, models and new trends. Mathware & Soft Computing 2002; 9.

[36]. Maniezzo V, Gambardella L.M, Luigi F.D. Ant Colony Optimization.

[37]. Dorigo M, Stützle T. The ant colony optimization metaheuristic: algorithms, applications, and advances. In: Glover F, Kochenberger G editors. Handbook of Metaheuristics. Kluwer academic Publishers; 2001.

[38]. Mullen R.J, Monekosso D, Barman S, Remagnino P. A review of ant algorithms. Expert Systems with Applications 2009; 36: 9608–9617.

[39]. Blum C. Ant colony optimization: introduction and recent trends. Physics of Life Reviews 2005; 2: 353-373.

[40]. Dorigo M, Stützle T. Ant Colony Optimization. Cambridge: MIT Press, 2004: 1-24.

[41]. Dorigo M, Di-Caro G, Gambardella LM. Ant algorithms for discrete optimization. Artificial Life 1999; 5(2): 137-172.

[42]. Stützle T, Hoos H. MAX-MIN ant system. Future Generation Computer Systems 2000; 16(8): 889-914.

[43]. Liao, Y.J., Yang, C.B., Shiau, S.H. Motif finding in biological sequences. In: Proc. Of 2003 Symposium on Digital Life and Internet Technologies, Tainan, Taiwan, pp. 89–98 (2003).

[44]. Blum C, Dorigo M. The Hyper Cube Framework for Ant Colony Optimization. IEEE Transactions on Systems, Man, and Cybernetics April 2004; Part B, 34(2): 1161- 1172.

ملخص

يعتبر تحديد المواقع الخاصة بارتباط عوامل النسخ على مستوى الجينات مشكلة صعبة في مجال البيولوجيا الجزيئية و التي تدعى بالأنماط. عملنا يقدم نهج التحسين من قبل مستعمرات النمل التي يمكن استخدامها لإيجاد حلول واعدة لمشكلة التعرف على الأنماط. النهج المقترح يتضمن صيغة معدلة من أسلوب عينات جيبس كخطوة للبحث المحلي الأمثل. وتم تنفيذ هذا النهج واختباره على قاعدة بيانات لبكتريا الايشيريشيا كولي لإظهار أنه يمكن من تحقيق أداء مماثل أوأفضل من حيث دقة النمط خلال مدة زمنية معقولة.

الكلمات المفتاح: المعلوماتية الحيوية، مستعمرات النمل، التعرف على الأنماط، التعريف الاستدلالي.

ABSTRACT

A challenging problem in molecular biology is to identify of the specific binding sites of transcription factors in the promoter regions of genes referred to as motifs. Our work presents an Ant Colony Optimization approach and how it can be used to provide the motif finding problem with promising solutions. The proposed approach incorporates a modified form of the Gibbs sampling technique as a local heuristic optimization search step. The approach has been implemented and tested on the *Escherichia coli* CRP dataset to show that it could achieve comparable or better performance in terms of motif accuracy within a reasonable computational time.

Keywords: Bioinformatics, Ant Colony Optimization, Motif Finding, Meta-heuristics.

RESUME

L'identification des sites spécifiques de liaison de facteurs de transcription dans les régions promotrices de gènes appelés motifs est un problème difficile de la biologie moléculaire. Notre travail présente une approche d'optimisation de colonie de fourmis qui peut être utilisé pour fournir des solutions prometteuses au problème de la reconnaissance du motif. L'approche proposée intègre une forme modifiée de la technique Gibbs Sampling comme une étape de recherche heuristique d'optimisation locale. L'approche a été implémentée et testée sur la base de données *Escherichia coli CRP* pour montrer qu'il pouvait atteindre des performances comparables ou meilleures en termes de la précision du motif dans un temps de calcul raisonnable.

Mots clés: Bioinformatique, Optimisation de Colonie de Fourmis, Reconnaissance du Motif, Méta-heuristiques.