DNA Motif Detection Using Particle Swarm Optimization and Expectation-Maximization

C. T. Hardin

Department of Computer Science and Computer Engineering
University of Louisville
Louisville, KY 40292
cthard01@louisville.edu

Eric C. Rouchka

Department of Computer Engineering and Computer Science
University of Louisville
Louisville, KY 40292
ecrouc01@louisville.edu

ABSTRACT

Motif discovery, the process of discovering a meaningful pattern of nucleotides or amino acids that is shared by two or more molecules, is an important part of the study of gene function. In this paper, we propose a hybrid motif discovery approach based upon a combination of Particle Swarm Optimization (PSO) and the Expectation-Maximization (EM) algorithm. In the proposed algorithm, we use PSO to generate a seed for the EM algorithm.

1. INTRODUCTION

Many amino acid and nucleotide sequences with functional or structural similarities share short contiguous sequences on the order of 10-20 bases known as motifs. Motifs can have a wide variety of purposes including providing structural properties, ligand binding sites, or signaling sites. Depending on the purpose of the motif and the nature of the specificity required, these regions may be highly conserved (close to 100% identical), or may contain a more subtle signal. The detection of these common patterns in a set of similar biological sequences can give insight into the regulation mechanisms involved, whether it be functionally or structurally controlled.

Several motif discovery algorithms have been demonstrated. Expectation Maximization (EM) uses information theory to identify conserved patterns of nucleotides or amino acids that may exist in several unaligned genetic sequences [1;2]. This particular algorithm is likely to identify sequences with a local maximum score, and must be run many times to search for improved scores.

The Gibbs Sampler extends this concept by introducing a stochastic process to exit the local maximum and continue search for a better solution [3].

Meta-MEME uses hidden Markov models (HMMs) of protein families to predict the motif patterns [4]. These HMMs must be trained with a known set of conserved regions. However, once trained, they can be an effective tool for a searching a large database of genetic sequences.

Particle Swarm Optimization (PSO) is a socially inspired algorithm that has been applied to search in both continuous and discrete search spaces with multiple dimensions [5-8]. The concept is to have various particles "fly" through a search space in search of solutions based upon a given objective function.

2. MODEL DESCRIPTION

In this paper, we use PSO to search for a high value motif and then use that as a seed to begin the EM algorithm to further improve the motif locations.

The problem space consists of N genetic sequences. Sequence s_i has a length of k_i nucleotides. The object is to find a motif of length m, in each sequence, s_i, with the highest information content based upon EM scoring methods.

To apply PSO, we define a particle, $P = [p_i]$, as a vector containing the location of first character of the motif in sequence s_i. Further, we maintain a velocity vector $V = [v_i]$, where v_i is the current velocity of the particle within sequence s_i. Our current implementation requires the user to specify the length of the motif. (We are currently investigating methods to discover the optimal motif length.)

From this point forward, we use a standard implementation of PSO in which a number of particles are instantiated with random p_i and v_i values. The social factor, individual factor, and maximum velocity are specified, and the particles are allowed to "fly." After each step, the EM score of each particle is evaluated and the pBest and gBest values are updated. The particles are allowed to fly until they fail to improve the gBest score for *x* iterations.

Once the termination condition is met, the P matrix for the gBest solution is then used as a seed for the EM algorithm. EM is an iterative algorithm that enumerates all possible motif locations in an effort to find the best fit. It then repeats until no further improvement is found.

Even using PSO to seed the EM algorithm, we observe frequent cases of termination on a local solution. With the current implementation, we reinitialize the PSO and re-run the algorithm until no improvement is observed.

3. RESULTS

To test our methods, we used a data set previously considered by Stormo and Hartzell [8]. This data presents 18 sequences with each sequence being 105 characters long, and each sequence contains at least one CRP protein binding site. The CRP binding site is 22 characters in length.

The same dataset was analyzed using our PSO/EM algorithm and compared to the results of the Gibbs Sampler and MEME. The results are tabulated in the Appendix.

4. DISCUSSION

The PSO/EM algorithm correctly identified the region of the motif in 13 of 18 sequences. This compares favorably with the Gibbs Sampler (12 of 18) and MEME (14 of 18).

Like the other two methods, the PSO/EM tends to have a consistent offset from the actual known motif location. PSO/EM consistently predicts a motif starting location 3 characters to the right of actual compared to Gibbs (2 characters left) and MEME (1 character left). This is believed to be a function of the information content of the specific motif versus the background and the scoring method applied by the algorithms.

5. MAIN CONTRIBUTIONS

This paper introduces the use of PSO into a new problem set, namely motif discovery. Even though motif discovery is primarily a problem in the domain of bioinformatics, it has potential application in pattern matching problems in other domains.

As far as we can ascertain, the technique presented here is the first hybrid utilization of PSO and EM in any problem domain.

6. SCOPE AND LIMITATIONS

This algorithm has only been tested on DNA sequences. The investigator must supply the expected length of the motif. We have implemented a scoring function for sequences of amino acids, but not yet tested it.

7. ACKNOWLEDGMENT

ER acknowledges support from the National Center for Research Resources (NCRR) grant 2P20RR016481-04 (Nigel G. F. Cooper, PI).

8. REFERENCES

[1] C. E. Lawrence and A. A. Reilly, "An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences," *Proteins*, vol. 7, no. 1, pp. 41-51, 1990.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1-38, 1977.

[3] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton, "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment," *Science*, vol. 262, no. 5131, pp. 208-214, Oct.1993.

[4] W. N. Grundy, T. L. Bailey, C. P. Elkan, and M. E. Baker, "Meta-MEME: motif-based hidden Markov models of protein families," *Comput. Appl. Biosci.*, vol. 13, no. 4, pp. 397-406, Aug.1997.

[5] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc.IEEE International Conference on Neural Networks*, 4 ed 1995, pp. 1942-1948.

[6] J. Kennedy, "The particle swarm: social adaptation of knowledge," in *Evolutionary Computation, 1997., IEEE International Conference on* 1997, pp. 303-308.

[7] J. Kennedy and R. C. Eberhart, "A discrete binary version of the particle swarm algorithm," in *Systems, Man, and Cybernetics, 1997.'Computational Cybernetics and Simulation'., 1997 IEEE International Conference on*, 5 ed 1997, pp. 4104-4108.

[8] G. D. Stormo and G. W. Hartzell, III, "Identifying protein-binding sites from unaligned DNA fragments," *Proc. Natl. Acad. Sci. U. S. A*, vol. 86, no. 4, pp. 1183-1187, Feb.1989.

APPENDIX

COMPARISON OF RESULTS FROM VARIOUS ALGORITHMS

Ref	LOCUS	Known Motif Locations	PSO/EM Location (Error)	Gibbs Location (Error)	MEME Location (Error)
1	cole1	17,61	64(3)	59(-2)	60(-1)
2	ecoarabop	17,55	58(3)	53(-2)	54(-1)
3	ecobglr1	76	79(3)	74(-2)	75(-1)
4	ecocrp	63	66(3)	61(-2)	62(-1)
5	ecocya	50	18(-32)	NONE	NONE
6	ecodeop	7,60	10(3)	5(-2)	6(-1)
7	ecogale	42	45(3)	40(-2)	41(-1)
8	ecoilvbpr	39	42(39)	NONE	38(-1)
9	ecolac	9,81	12(3)	7(-2)	8(-1)
10	ecomale	14	17(3)	12(-2)	13(-1)
11	ecomalk	29	64(-35)	59(-30)	34(-5)
12	ecomalt	41	44(3)	NONE	40(-1)
13	ecoompa	48	51(3)	46(-2)	47(-1)
14	ecotnaa	71	74(3)	69(-2)	70(-1)
15	ecouxu1	17	20(17)	15(-2)	74(57)
16	pbr-p4	53	56(3)	NONE	NONE
17	trn9cat	-1,84	36(37)	NONE	NONE
18	(tdc)	78	79(1)	74(-4)	76(-2)

This data set was obtained from GenBank, Release 55. (The *tdc* gene was not in that release; it was obtained from [8].

LOCUS is presented in alphabetical order.

Known motif locations were obtained from [8].