Classification and Clustering of Human Sperm Swimming Patterns

*Ji-won Choi, *Chizhong Wang, †Leonardo F. Urbano, ‡Puneet Masson, §Matthew VerMilyea, and *Moshe Kam
*Department of Electrical and Computer Engineering, New Jersey Institute of Technology,
Newark, New Jersey 07102 USA, Email: jc423@njit.edu, cw278@njit.edu, kam@njit.edu
†70 Beharrell St, Concord, MA 01742 USA, Email: leonardo.f.urbano@gmail.com
‡ Penn Fertility Care, Hospital of the University of Pennsylvania,
Philadelphia, PA 19104 USA, Email: Puneet.Masson@uphs.upenn.edu
§ Ovation Fertility, Austin, TX 78731 USA, Email: mvermilyea@ovationfertility.com

***Abstract*—The principal observed progressive swim types of sperm cells are *linear mean* and *circular swim*. Using motility characteristic parameters produced by CASA systems, we perform a parameter subset search to produce distinct clusters of the different swim types. For this task, the *artificial bee colony algorithm* (an iterative search algorithm modeled after the collective behavior of bees) and the well-studied k-means clustering algorithm were used on simulated and human sperm swim data. The result is distinct clusters with features of each types of swim. The clustering approach displays potential as a tool for automated sperm swim subpopulation analysis.**

***Index Terms*—Clustering, artificial bee colony algorithm, sperm imaging, particle tracking, sperm motility.**

I. INTRODUCTION

Semen analysis is an important assessment tool for infertility. A popular approach to semen analysis is computer assisted semen analysis (CASA), known to be fast, consistent and repeatable [1]. In CASA systems, semen sample video frames are analyzed to evaluate each specimen using image processing and particle tracking. Some of the major semen data that CASA system can provide are concentration (the number of sperm per milliliter of semen), percentage of motile sperm, and the motility characterstics of the sperm. Motility characterstics of the sperms are defined by various motility parameters, abbreviated as VCL, VSL, VAP, LIN, WOB, STR, ALH, and MAD [2].

The significance of sperm concentration and percent of motile sperm is clear in male fertility analysis. However, the significance of other sperm motility parameters is a subject of on-going research [3]. Several studies have used clustering methods to assist in this line of investigation. The clustering methodology proposed in this study is motivated by the review of clustering on motility parameters of CASA systems by Martinez-Pastor et al. [3]. The fully automated multi-sperm tracking algorithm introduced by Urbano et al. [4], [5] is used to explore the clustering task of human sperm swim subpopulation.

A. Background

Sperm cell movements can be divided into three major categories: progressive, non-progressive, and immotile [2]. Immotile movements define cells which show no movement at all, and non-progressive movements are defined by motility that lacks progression, where the sperm head show little displacement over time. Progressive movements characterize actively moving sperm cells. Two-dimensional progressive sperm swimming patterns observed by Babcock et al. [6] were categorized as *linear-mean* and *circular*.

The two progressive movements, linear mean and circular, differ by the presence and absence of head rolling. Sperm cells of different movement types have different values of motility parameters, therefore clustering by motility parameters may provide means of identifying the subpopulation of swimming patterns. Such clustering would allow better understanding of the cause and importance of these two swimming patterns.

B. Organization

This paper is organized as follows. In Section II, we describe the motility parameter calculation, sperm swim track simulation, sperm track data extraction, and sperm track clustering of simulated and real sperm swimming data. In Section III and IV, we provide results of clustering on simulated data and on two (2) human sperm samples, respectively.

II. METHODS

A. Motility Parameter Calculation

We use eight (8) standard motility parameters measured by many CASA systems [2]. These eight parameters are:

- VCL (curvilinear velocity)
- VSL (straight-line velocity)
- VAP (average path velocity)
- ALH (amplitude of lateral head displacement)
- LIN (linearity): VSL/VCL

*An earlier abstract in the general approach was presented at the Sixth IEEE International Conference on Healthcare Informatics (ICHI 2018).

978-1-5386-6396-7/18/$31.00 ©2018 IEEE

Authorized licensed use limited to: ANII. Downloaded on August 13,2020 at 03:22:29 UTC from IEEE Xplore. Restrictions apply.

- WOB (wobble): VAP/VCL
- STR (straightness): VSL/VAP
- MAD (mean angular displacement)

The motility parameters are obtained by processing successive 1-second long track segments. The parameters of each particles are then averaged. We use the median of the parameter measurements as the average value. The median has shown better performance in clustering than the mean; the median is less impacted by outliers.

B. Approximate swim models of sperm

Observing the paths of progressive sperm movements in [6], we use the following models to describe linear mean and circular movements.

1) Linear Mean Swim Model:

$$\begin{bmatrix} x_n \\ y_n \end{bmatrix} = \begin{bmatrix} \cos(\theta_r) & -\sin(\theta_r) \\ \sin(\theta_r) & \cos(\theta_r) \end{bmatrix} \begin{bmatrix} r_v \sin(2\theta_{l_n}) \\ r_h \sin(\theta_{l_n}) \end{bmatrix} + \begin{bmatrix} x_{c_n} \\ y_{c_n} \end{bmatrix} \tag{1}$$

$$\begin{bmatrix} x_{c_{n+1}} \\ y_{c_{n+1}} \end{bmatrix} = \begin{bmatrix} x_{c_n} + V\cos(\theta_r) \\ y_{c_n} + V\sin(\theta_r) \end{bmatrix} \tag{2}$$

$$\theta_{l_{n+1}} = \theta_{l_n} + \Delta\theta_{l_n} \tag{3}$$

Where
x_n : horizontal position of the sperm head at step n
y_n : vertical position of the sperm head at step n
x_{c_n} : horizontal position of the average path at step n
y_{c_n} : vertical position of the average path at step n
$\Delta\theta_{l_n}$: rate of change in ribbon angle
r_h : width of ribbon
r_v : height of ribbon
θ_r : direction of forward movement
V : straight line path velocity
$\Delta\theta_{l_n}$: rate of change in ribbon angle

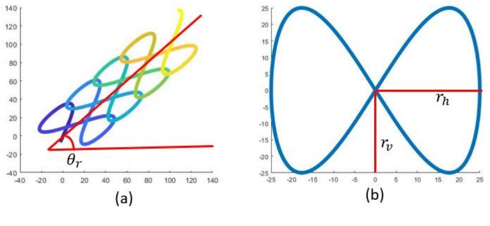

Fig. 1: Simulated Linear Mean Swim

Fig. 1a shows a simulated linear mean swim which has "ribbon-like" pattern, generated by using equations (1)-(3). Equation (1) calculates the horizontal and vertical position of the simulated sperm head by superimposing a ribbon on the center of the overall linear path. The ribbon consists of three variables, r_v, r_h, and θ_{l_n}. Fig. 1b, the graphical representation of the ribbon, shows r_v, the height of the ribbon, and r_h, width of the ribbon.

2) Circular Swim Model:

$$x_n = (r_c + a\sin(f_c\theta_{c_n}))\cos(\theta_{c_n}) \tag{4}$$

$$y_n = (r_c + a\sin(f_c\theta_{c_n}))\sin(\theta_{c_n}) \tag{5}$$

$$\theta_{c_{n+1}} = \theta_{c_n} + \Delta\theta_{c_n} \tag{6}$$

Where
r_c : radius of the circular path
f_c : frequency of sinusoid modulated on the circular path
a : amplitude of the sinusoid
$\Delta\theta_{c_n}$: rate of change in angle

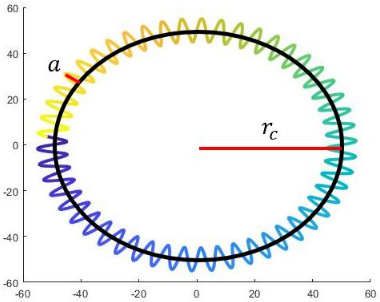

Fig. 2: Simulated Circular Swim

The circular swim is represented as a sinusoidal modulated circular path (see Fig. 2).

C. Specimen Video Collection and Track Data Extraction

Video clips of human semen specimens prepared and collected by the In-Vitro Fertilization laboratories at Penn Fertility Care were used in this study. The video clips are 200 magnified images of 640×480 pixel resolution (0.857 μm/pixel). The sperm swim track data have been obtained from the video clips by segmentation, localization, and track data association process.

Following [4], in the segmentation phase, the image was convolved with 11×11 pixel Gaussian filters 5 times to reduce image noise. Then, the image was convolved with 9×9 Laplacian of Gaussian (LoG) filter to produce spot-enhance image. Using Otsu's method [7], the intensity threshold was calculated for the image, and the image was binarized. The binarized image then went through morphological enhancements, erosion and dilation, to reduce spurious detection due to noise. Finally, the sperm cells were "localized" by labeling any group with more than or equal to 5 pixels as a detected sperm head.

As the final step of track data extraction, using the joint probability density association filter (JPDAF) [8], the particle location data were used to track the particles throughout the video clip. By using JPDAF, misdetection probability of particles during of head collisions in sperm cells were

Authorized licensed use limited to: ANII. Downloaded on August 13,2020 at 03:22:29 UTC from IEEE Xplore. Restrictions apply.

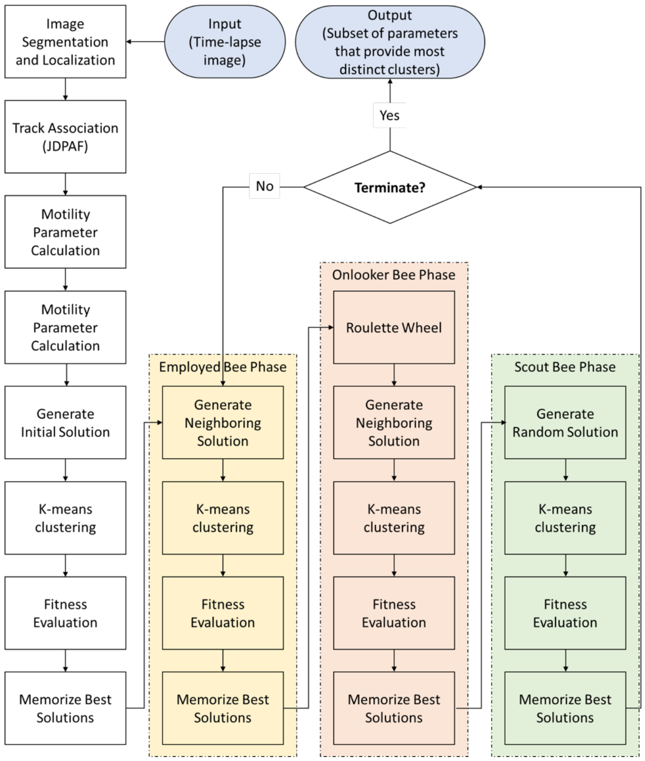

Fig. 3: Process Flowchart (Input: Time-lapse Image, Output: Subset of parameters that provide most distinct clusters)

reduced, giving more accurate tracking results compared to other popular track association methods such as the Nearest Neighbor algorithm [4], [5]. For detailed information regarding the procedures we have used (segmentation, localization, and track data association), see [4].

D. Sperm Track Clustering and Artificial Bee Colony Algorithm Design

In attempt to find the subset of motility parameters for k-means clustering that will provide distinct clusters of each swimming types, we have used the artificial bee colony (ABC) algorithm. ABC algorithm is an optimization algorithm based on foraging behavior of honeybee swarm, proposed by Karaboga [9]. The main goal of the ABC algorithm is to find the optimal "food source" or optimal solution for a given problem. ABC algorithm accomplishes this task by using three types of bees: employed bees, onlooker bees, and scout bees.

After gathering food at a food source, bees return to a "dance floor" within the hive. On the "dance floor," each of bees performs a waggle dance. The dance provides information to the other bees, such as the direction, distance, and "richness" of the food source [10]. The duration of the dance is longer for richer food source. In the ABC algorithm, the richness of a food source is referred as the *fitness* of a solution.

For our design of the ABC algorithm, each of the possible subset of motility parameters is considered a food source, which will be evaluated by bees. The aim of the algorithm would be to find a subset of motility parameters that will produce most distinct clusters by k-means clustering. Distinctiveness of clusters is measured by the *silhouette values* [11].

A "silhouette value" is a measure of similarity between different data points within the same cluster. The silhouette value for particle i is

$$S_i = \frac{b_i - a_i}{max(a_i, b_i)}, \tag{7}$$

where a_i stands for average distance from data point i to other data points within the same cluster, and b_i is the minimum distance from data point i to other data points in other clusters.

The sum of silhouette values is high for "good" clustering, and low for "poor" clustering. For our ABC algorithm, fitness value (the "richness" of the food source) was defined as the sum of silhouette value for all the data used in clustering (Eq. (8)). The fitness function for "food source" s for c clusters, with cluster j containing n_j data points is defined as

$$F(s) = \sum_{j=1}^{c} (\sum_{i=1}^{n_j} S_i). \tag{8}$$

The first type of bees, employed bees, exploit the known food sources. These employed bees return to their hive, and "dance" on the dance floor. Onlooker bees observe the dance, and decide on their destination. The duration of the dance is proportional to the richness of the food source; the bees that has been to a rich food source will dance longer than the bees who has been to a poor food source. Therefore, it is more likely for the onlooker bees sight the bee that has been to a rich food source. As a result, the onlooker bees has more chance of going to a rich food source. Scout bees perform a random search in order to find even better food sources. These three behavior types drive the ABC algorithm.

First, the bees are sent to random locations to search for food. Each bee evaluates each food source, and returns to the hive with a fitness value. Few of the "best" sources are chosen as candidate locations for employed bees and onlooker bees to exploit. Next, the algorithm goes through three phases (employed, onlooker, and scout bees), and repeats until a termination requirement has been met. For our design, ABC algorithm is terminated at the user-defined iteration.

As the first step, the employed bees are sent to the best food sources that were chosen during the initialization phase. The employed bees look at the neighboring site to see if any of the neighboring sites contain a better food source than the food sources listed as the best food sources. If any of the bees finds a food source that is better, the new site replaces the worst food source listed in the best food source list. Next, the onlooker bees choose their destination by observing dances of the employed bees. For the onlooker bees, a roulette wheel with the selections proportional to the fitness values in the best food source list is used to select a destination. The probability

Authorized licensed use limited to: ANII. Downloaded on August 13,2020 at 03:22:29 UTC from IEEE Xplore. Restrictions apply.

of choosing food source s_m from total number of best food sources s_o is defined to as

$$p_t(s_m) = \frac{F(s_m)}{\sum_{k=1}^{s_0} F(s_k)}. \tag{9}$$

Afterwards, the onlooker bees observe the neighboring sites, and update the best food source list if any better food source was found. Lastly, during the scout bee phase, the scout bees are sent to random locations to evaluate food sources, and the best food source list is updated if any better food source was found. The scout bees enable a global search, creating a possibility of avoiding convergence to local maximum. The process is represented as a flowchart in Fig. 3.

In the context of our problem, every possible subset of motility parameters is possible food source for the bees to exploit. In our design, we create an 8-bit binary word to indicate which motility parameters are used. Only the motility parameters corresponding to a "1" are selected for the subset used later in k-means clustering. Initial random food sources are selected by choosing a random binary number between 00000000 and 11111111. The neighboring "food source" is determined by flipping one of the bits in the food source word randomly.

Clustering of the subset of parameters was performed through the unsupervised-learning based k-means clustering method. Specifically, we used the built-in function for k-means clustering in MATLAB, which uses k-means++ algorithm by Arthur [12]. It was assumed that there were three (3) distinct clusters in the processed data, with the three (3) clusters representing each of the sperm swim categories, namely the linear mean, circular, and non-progressive (including immotile). Each k-means clustering has been repeated 5 times to reduce the probability of finding a local maximum.

III. K-means Clustering of Synthetic Sperm Swim Tracks with the ABC Algorithm

Using the mathematical models (1)-(6), 300 instances each of linear mean, circular, and immotile (or dead) swim type data were generated. The parameters used in linear mean and circular swim for each simulated sperm was distributed normally; the mean and the variance of each parameter are listed in table I. The mean and the variance of the simulation parameters were chosen so that the VCL and VSL values of the two progressive swimming types would overlap. In the simulation, it was assumed that all the particles are disturbed by small additive Brownian motion (to model the influence of movement of the fluid).

For the ABC algorithm, ten (10) employed bees, five (5) onlooker bees, and five (5) scout bees were used for 10 iterations. The ABC algorithm selected the motility parameters ALH, WOB, and STR for k-means clustering. The ALH, WOB, and STR values of the three types of simulated swimming data are shown in Fig. 4. The data points are colored according to their swimming pattern; red for dead, green for linear mean, and blue for circular. Here, We can see that in spite of the overlap in the VCL and VSL values, the two progressive swimming types were correctly clustered.

TABLE I: Simulation Parameters

Parameter	Mean	Variance
r_h	25	5
r_v	25	5
V	0.5	0.05
$\Delta\theta_{l_n}$	3	0.3
r_c	50	10
f_c	50	5
a	2	0.2
$\Delta\theta_{c_n}$	2	0.2

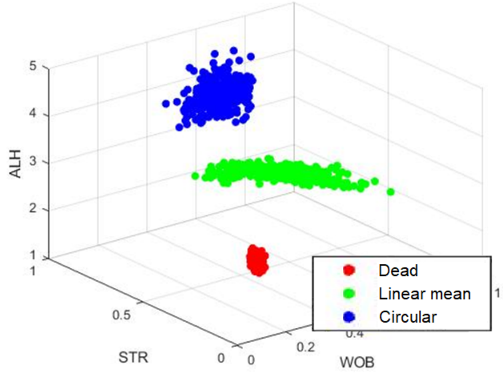

Fig. 4: ALH, WOB, and STR values of simulated swim data

IV. K-means Clustering of Human Sperm Swim Tracks with ABC Algorithm

Ten (10) employed bees, five (5) onlooker bees, and five (5) scout bees were used in the ABC algorithm. For real sperm data, 100 iterations are used to reduce the likelihood of converging to a local maximum. We present clustering results for two (2) different samples, one at 15 FPS (frames per second) and the other at 30 FPS (both observed for 10 seconds).

A. Clustering Outcomes - Sample A (shown in Fig. 5)

In sample A, 21 sperm swim tracks were observed. Motility parameters used for this clustering were LIN, WOB, and STR.

- Cluster 1: 3 tracks with very low values of LIN, WOB, and STR.
- Cluster 2 (circular): 9 tracks with high values of LIN, WOB, and STR. Tracks show circular swimming path, with small lateral displacement.

Authorized licensed use limited to: ANII. Downloaded on August 13,2020 at 03:22:29 UTC from IEEE Xplore. Restrictions apply.

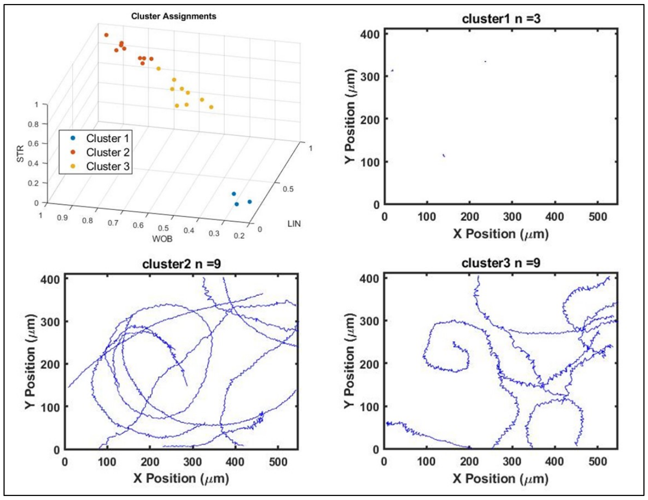

Fig. 5: Sample A Clustering Outcomes (30 FPS, 10 seconds)

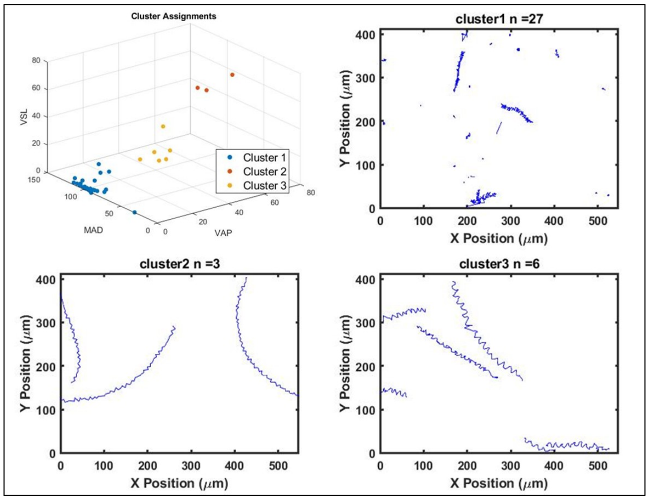

Fig. 6: Sample B Clustering Outcomes (15 FPS, 10 seconds)

- Cluster 3 (linear-mean): 9 tracks with high values of STR and smaller LIN and WOB values compared to cluster 2. Tracks show approximate linear swimming path with jagged-like patterns.

B. Clustering Outcomes - Sample B (shown in Fig. 6)

In sample B, 36 sperm swim tracks were observed. Motility parameters used for this clustering were VSL, VAP, and MAD.

- Cluster 1: 27 tracks with low straight-line velocity and average path velocity. Linear-mean progressive tracks can be observed in this cluster. Few tracks show small linear progression. Due to the very small speed, they have been clustered with the dead or immotile cells.
- Cluster 2 (circular): 3 tracks with high value of straight-line velocity and average path velocity. Tracks show circular swimming path, with small lateral displacement.
- Cluster 3 (linear-mean): 6 tracks with smaller straight-line and average path velocity compared to cluster 2. Tracks show approximate linear swimming path with jagged-like patterns.

C. Discussion

For both sperm samples, we observed that using more parameters than the selected parameters chosen by the ABC algorithm did not provide better clusters (as judged by the silhouette values). Also, the parameters used for clustering were not the same for the two samples. The use of ABC algorithm allows sample-specific clustering, which leads to clusters whose membership can be described by swimming characteristics. We were able to separate cells based on the types of the two progressive movements from each other. In all the examples, three (3) clusters were specified for the k-means clustering. Attempts to increase the number of clusters to four (4) resulted in poor performance.

V. Conclusions

K-means clustering guided by the ABC algorithm (for parameter selection) provided distinct clusters of the different sperm swimming types. The algorithm produced clusters of moving sperm swim types characterized by presence or absence of sperm head roll. The ability to categorize sperm swim movement could provide a tool that will lead to a better understanding of sperm swimming patterns.

References

[1] A. M. Crespilho, "Sensitivity evaluation of the computer-assisted sperm analysis (CASA) in the determination of frozen-thawed bull semen concentration," Braz. J. Vet. Res. Anim. Sci., São Paulo, v. 54, n. 3, pp. 247-252, 2017

[2] *Laboratory Manual for the Examination and Processing of Human Semen*, 5th ed., World Health Org., Geneva, Switzerland, 2010

[3] F. Martinez-Pastor, E. Jorge Tizado, J. J. Garde, L. Anel, P. de Paz, "Statistical Series: opportunities and challenges of sperm motility subpopulation analysis," *Theriogenology*, vol.75, issue 5, pp.783-795 Mar 2011

[4] L. F. Urbano. "Automatic Tracking and Motility Analysis of Human Sperm in Time-Lapse Images," *IEEE Transactions on Medical Imaging*, vol.36, No.3, Mar 2017.

[5] L. F. Urbano. "Roboust Automatic Multi-Sperm Tracking in Time-Lapse Images," Ph.D thesis, Drexel University, Philadelphia, 2014-15.

[6] Babcock et al., "Episodic rolling and transient attachments create diversity in sperm swimming behavior," *BMC Biology*, 2014 12:67

[7] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern*, vol.9, no.1, pp.62-66, Jan 1979.

[8] Y. Bar-Shalom, F. Daum, and J. Huang "The probabilistic data association filter," *IEEE Control Syst.*, vol.29, No.6, pp. 82-100, Dec. 2009.

[9] D. Karaboga, "An idea based on honey bee swarm for numerical optimization," Technical report TR06. Kayseri, Turkey: Computer Engineering Department, Engineering Faculty, Erciyes University. 2005

[10] V. Tereshko, A. Loengarov, "Collective Decision-Making in Honey Bee Foraging Dynamics," Computing and Information Systems Journal, ISSN 1352-9404, vol.9, no.3, October 2005.

[11] P. J. Rousseeuw. "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis", *Computational and Applied Mathematics*, vol.20, pp. 53–65, 1987

[12] D. Arthur, S. Vassilvitskii, "K-means++: The Advantages of Careful Seeding," *In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007, pp. 1027–1035

Authorized licensed use limited to: ANII. Downloaded on August 13,2020 at 03:22:29 UTC from IEEE Xplore. Restrictions apply.