Electronic Supplementary Material (ESI) for ChemComm.

Simultaneous Single-Molecule Discrimination of cysteine and homocysteine with a protein nanopore

Yao Lu[a,†], Xue-Yuan Wu[a,†], Yi-Lun Ying*[a], Yi-Tao Long*[a b]

[a.] *School of Chemistry and Molecular Engineering, East China University of Science and Technology, Shanghai, 200237, P. R. China.*

[b.] *State Key Laboratory of Analytical Chemistry for Life Science, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing 210023, P. R. China.*

Email: yitaolong@nju.edu.cn; yilunying@ecust.edu.cn

Supporting Information

1. Experimental Section

Chemicals and Reagents. 1,2-Diphytanoyl-sn-glycero-3-phosphocholine (chloroform, ≥ 99%) was purchased from Avanti Polar Lipids Inc. (Alabaster, AL, USA). Trypsin EDTA and decane (anhydrous, ≥ 99%) were purchased from Sigma-Aldrich (St. Louis, MO, USA). The K238Q mutant proaerolysin was synthesized by our group and activated by digestion with trypsin for one hour at room temperature. The aldehyde-modified oligonucleotide (AHA4) was synthesized and HPLC-purified by Sangon Biotech Co., Ltd (Shanghai, China). All reagents and chemicals were of analytical grade and were used without further purification unless otherwise noted. All solutions were prepared with ultrapure water (18.2 MΩ cm at 25°C) using a Milli-Q System (EMD Millipore, Billerica, MA, USA).

Conjugation of the DNA probe with cysteine/homocystine**.** Cysteine/homocystine was saturated-solved in 10 mM Tris-HCl buffer (pH = 5). Then, 30 μL 100 μM AHA4 and 120 μL cysteine/homocystine solution were mixed together in a centrifuge tube and incubated under 70°C for 30 min. The reaction was characterized by Mass Spectra.

Nanopore Analysis. The single-channel recording was performed according to our previous studies.[1-2] All the nanopore experiments were carried out at 22 ± 2 °C. A planar lipid bilayer was formed by applying 30 mg/mL diphytanoyl-phosphatidyl-choline in decane to a 50 μm orifice in a Delrin bilayer cup (Warner Instruments, Hamden, CT, USA). Each chamber was filled with 1 mL of electrolyte (1 M KCl, 10 mM Tris and 1.0 mM EDTA, pH 8.0). Aerolysin (AeL, 0.5-2.0 μL, 1.5 $\mu g \cdot mL^{-1}$) was added to the *cis* compartment, which was defined as virtual ground. Once a stable single AeL nanopore was formed, the 20 μL reaction solution was added to the *cis* chamber. The current was detected with a pair of Ag/AgCl electrodes, recorded with a patch-clamp amplifier (Axopatch 200B, Axon Instruments, Forest City, CA, USA), filtered with a low-pass Bessel filter at 5 kHz, and then digitized with a Digidata 1440A A/D converter at a sampling rate of 250 kHz by running the Clampex 10.9 software (Molecular Devices, USA). The data analysis was performed by PyNano (https://decacent.github.io/PyNano/), Clampfit 10.9 (Axon Instruments, Forest City, CA, USA) and OriginLab 9.1 (OriginLab Corporation, Northampton, MA, USA). The scatter plots were set with I/I_0 on x axis and duration on the y axis. The I/I_0 histograms were fitted with Gaussian distribution, while duration histograms were fitted with a transformed

single -exponential function[1,2] as described in (1).

$$y = y_0 + \frac{Ae^{\frac{-2\left(logt - logt_c\right)^2}{w^2}}}{w\sqrt{\frac{\pi}{2}}} \quad (1)$$

$logt_c$ is the center of the gauss peak; w is the width of the gauss peak; A is the area of the gauss peak; y_0 is the offset of the gauss peak.

2. Comparation between WT and K238Q aerolysin nanopore.

Table S1 The I/I_0, full width at half maximum ratio ($FWHM/FWHM_{WT}$) and duration time for traversing the $poly(dA)_4$ through WT AeL and K238Q AeL at +100 mV, respectively

AeL Type*	I/I_0	$FWHM/FWHM_{WT}$†	Duration/ms
WT AeL ‖	0.50 ± 0.01	1	9.1 ± 1.4
K238Q AeL	0.52 ± 0.01	0.92	15.7 ± 2.2

‖ The data is obtained from Ref. 3.

Table S2 The I/I_0, full width at half maximum ratio ($FWHM/FWHM_{WT}$) and duration time for traversing the $poly(dA)_4$ through WT AeL and K238Q AeL at +160 mV, respectively

AeL Type*	I/I_0	$FWHM/FWHM_{WT}$†	Duration/ms
WT AeL ‖	0.54 ± 0.01	1	1.9 ± 0.1
K238Q AeL	0.57 ± 0.01	0.92	223.5 ± 1.9

*All the experiments were performed in 1.0 M KCl, 10 mM Tris, 1.0 mM EDTA at pH 8.0 at +160 mV.

† $FWHM/FWHM_{WT}$ is the ratio between the FWHM value for specific type of AeL to that for WT AeL.

‖ The data is obtained from Ref. 3.

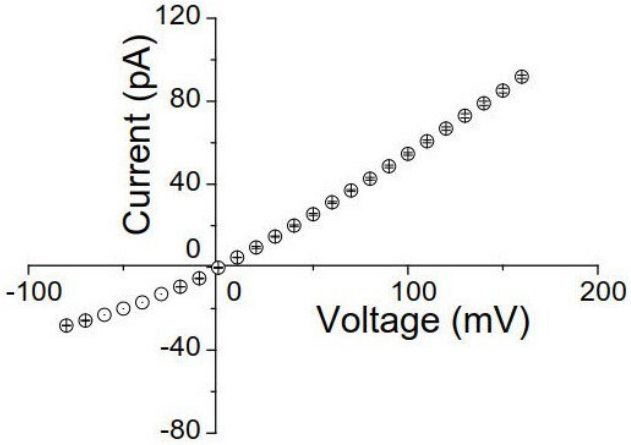

Fig.S1 Current-voltage curves of K238Q aerolysin nanopore. The data was acquired 1.0 M KCl, 10 mM Tris, and 1.0 mM EDTA, pH=8.0, 22 ± 2℃. The error bars are based on three independent nanopore experiments.

3. Voltage-dependent experiments on AHA4 with the K238Q aerolysin.

The signals distribution of AHA4 is composed of two populations, which are labelled PI and PII. According to previous studies of α-hemolysin and aerolysin,[5,6] PI is ascribed to the entering of oligonucleotide into the aerolysin, whereas PII resulted from collision of the oligonucleotides with the cis opening of the aerolysin.

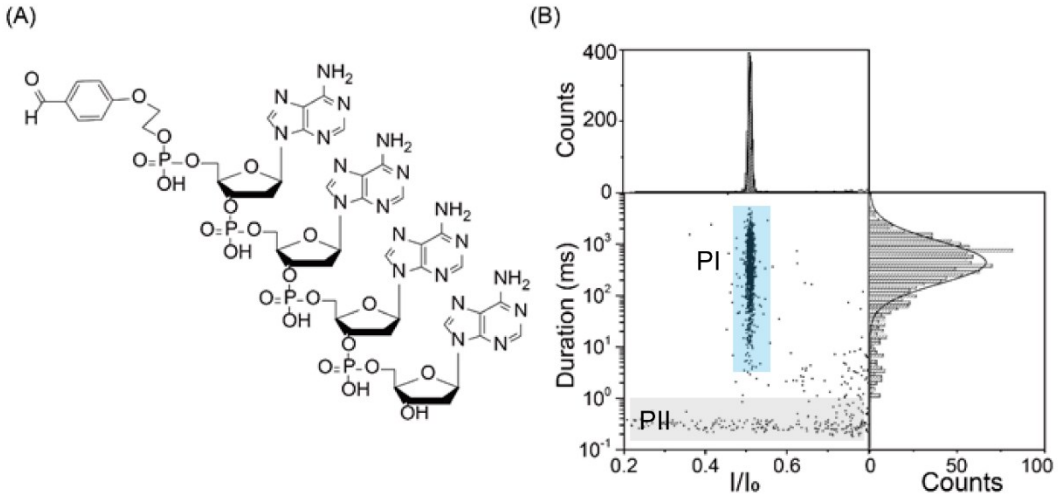

Fig. S2 (A) Chemical structure of AHA4. (B) The scatter plot of PI (blue) of AHA4 with mean I/I_0 of 0.53 and duration of 446.7 ms. The I/I_0 histograms (top) were fitted with Gaussian distribution, while each duration histogram (right) was fitted with a transformed single-exponential function. All data was obtained in 1.0 M KCl, 10 mM Tris, and 1.0 mM EDTA, pH=8.0, 22 ± 2℃ at the bias potential of +160 mV.

Table S3 The I/I_0, full width at half maximum ratio ($FWHM/FWHM_{WT}$) and duration time for AHA4 traversing through WT AeL and K238Q AeL, respectively

AeL Type*	I/I_0	$FWHM/FWHM_{WT}$	Duration/ms
WT AeL	0.47 ± 0.01	1	6.0 ± 1.0
K238Q AeL	0.53 ± 0.01	0.88	446.7 ± 1.0

*All the experiments were performed in 1.0 M KCl, 10 mM Tris, 1.0 mM EDTA at pH 8.0 at +160 mV.

† $FWHM/FWHM_{WT}$ is the ratio between the FWHM value for specific AeL type to that of WT AeL.

4. HPLC-MS characterizations of probe AHA4, AHA4-C and AHA4-H**.**

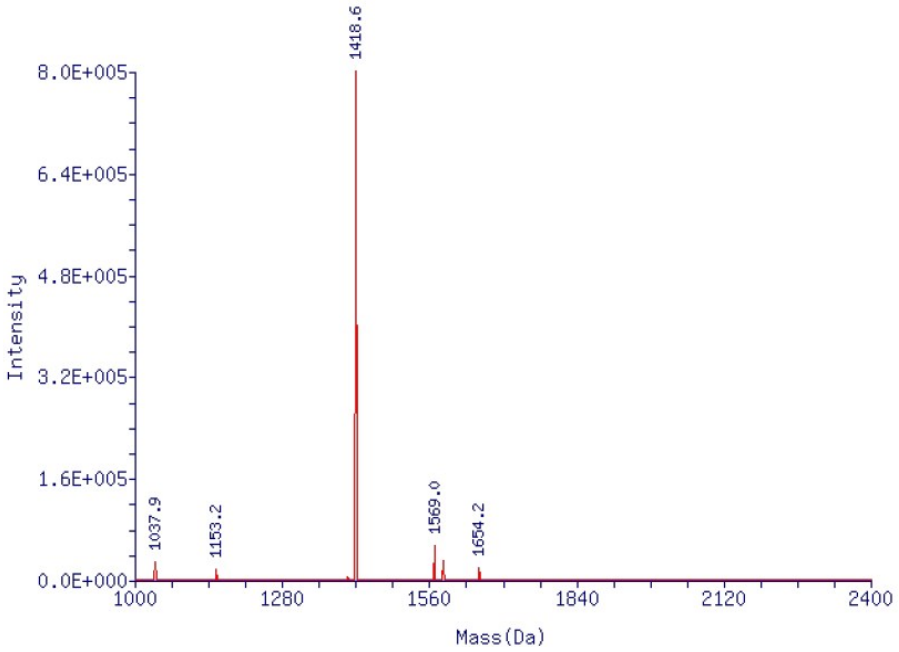

Fig. S3 Mass spectrometry characterization of probe AHA4. Calculated: m/z = 1419.0.

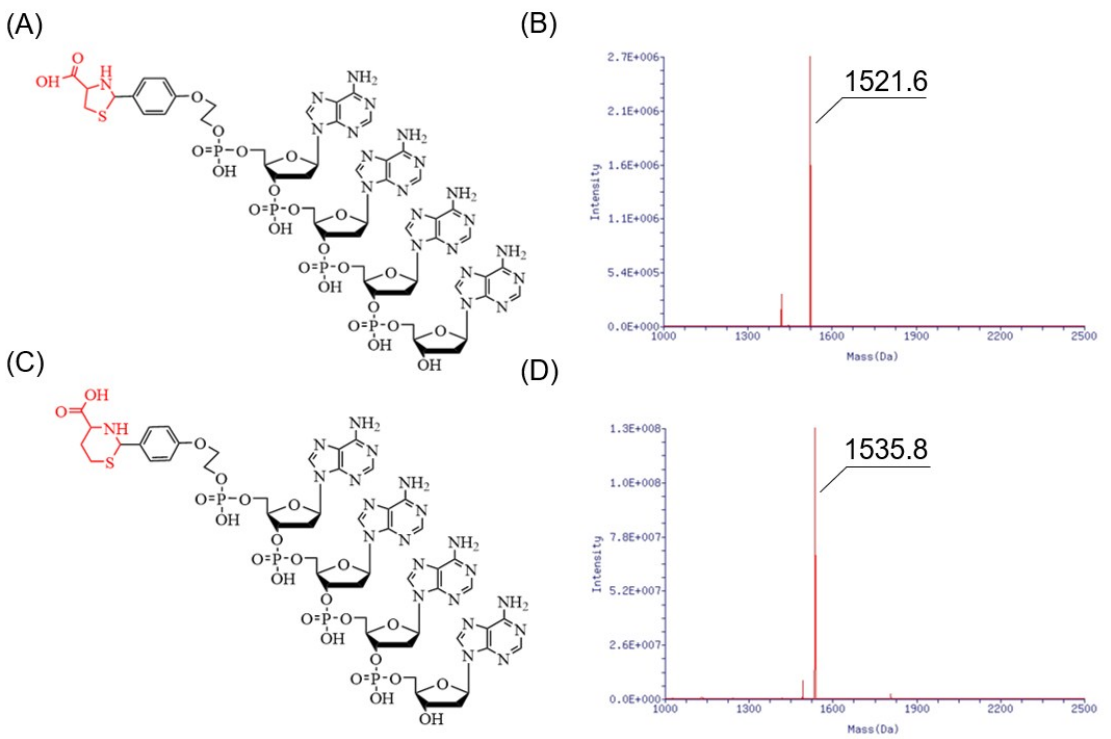

Fig. S4 Chemical structure (A) and mass spectrometry characterization (B) of AHA4-C. Calculated: m/z = 1522.2. Chemical structure (C) and mass spectrometry characterization

(D) of AHA4-H. Calculated: m/z = 1536.2.

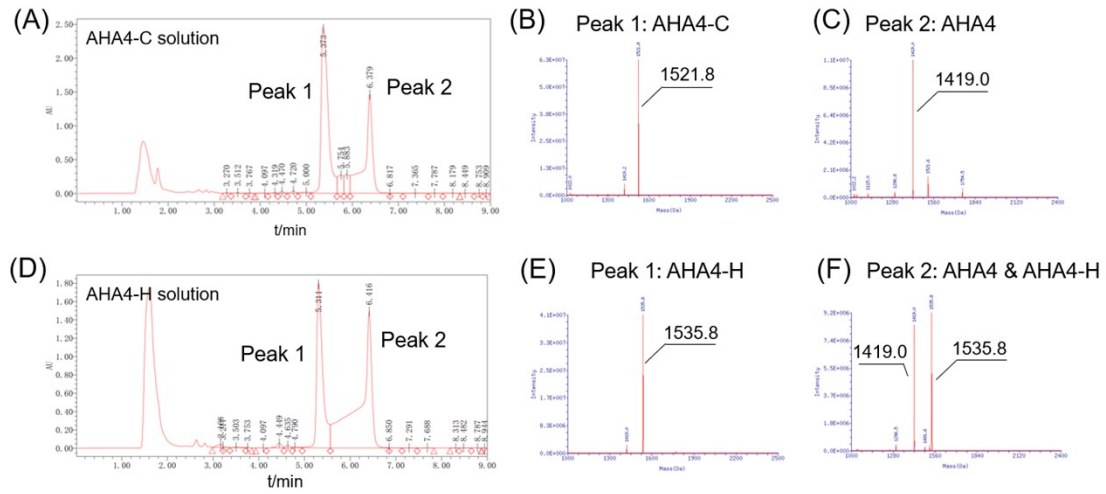

Fig. S5 The HPLC spectrum (A) of AHA4-C product solution with the corresponding mass spectra of peak 1 (B) and peak 2 (C); The HPLC spectrum (D) of AHA4-H product solution with the corresponding mass spectra of peak 1 (E) and peak 2 (F). The chromatography was performed on Xbridge C18 column from Waters, USA, having the following specifications: internal diameter 4.6 mm, height 150 mm, and particle size 3.5 µm. The mobile phase consisting of A: water and B: acetonitrile with the gradient set up as mentioned below which delivered good baseline separation of the targeted peaks. The gradient used was as follows: zero time condition was 84.5% A and it was decreased to 79.5% A in 7 minutes. The column conditioning and equilibration were performed in 5 minutes attaining the initial conditions. The chromatographic column maintained at 35°C and the flow rate used was 2 mL/min. The peaks of target AHA4-C and AHA4-H eluted at retention times 5.373 min and 5.311 min, respectively, in the sample extract.

5. Discrimination of AHA4 and AHA4-H**.**

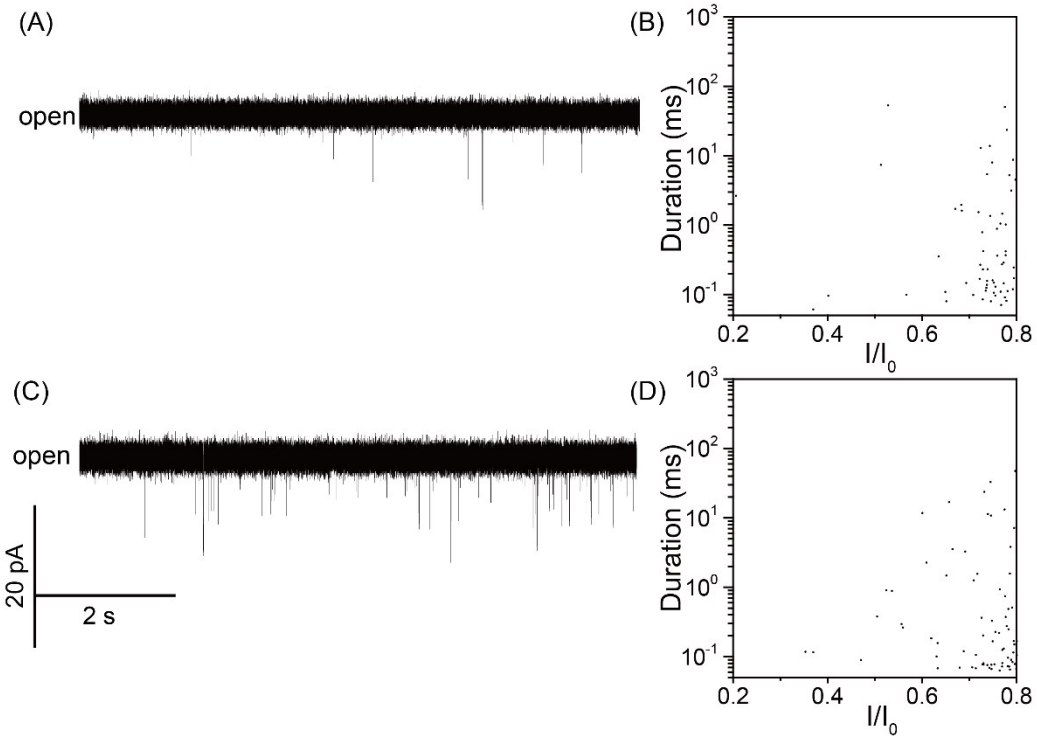

Fig. S6 K238Q AeL analysis of 10 mM Cys/Hcy. (A) The current trace of 10 mM Cys; (B) The scatter plots for 20 min current recording of Cys . (C) The current trace of 10 mM Hcy; (D) The scatter plots for 20 min current recording of Hcy. All data was obtained in 1.0 M KCl, 10 mM Tris, and 1.0 mM EDTA, pH=8.0, 22 ± 2℃ at the bias potential of +160 mV.

Spectra Clustering Algorithm

The spectral clustering is an unsupervised clustering method based on graph theory. The points in dataset are regarded as the vertices of the graph, and the edges between the vertices are weighted based on similarity criterion.[7,8,9] Then, the weighted undirected graph is cut into k subgraphs and the sum of the weights between the subgraphs should be as small as possible, while the sum of weights within the subgraphs should be as large as possible. Compared with other traditional clustering methods[10,11,12], the dataset with non-overlapping distributions can be classified optimally by spectral clustering.

Before classification, the raw current signals caused by AHA4, AHA4-C and AHA4-H traversing K238Q are preprocessed by home designed Pynano software (https://decacent.github.io/PyNano/). The features of the current blockades, including residual current and duration time, are extracted with threshold method and regarded as

the input dataset of spectral clustering. Here, the current blockages fall in the region of I/I_0 of 0.2 - 0.8 and duration > 0.1 ms. Then, the graph $G = (V,E)$ is constructed based on the dataset. The residual current and duration time of blockades represent a set of vertices $V = [v_i]_{n \times 2}$, while the edge between v_i and v_j is weighted by Gaussian kernel function:

$$w_{ij} = w_{ji} = \exp\left(-\frac{\| v_i - v_j \|_2^2}{2\sigma^2}\right) \quad (1)$$

where the parameter σ denotes the width of neighborhoods and $W = [w_{ij}]_{n \times n}$ represents the weighted adjacency matrix. Then, the degree matrix D and the normalized Laplacian matrix L can be calculated by Equation 2 and Equation 3.

$$D = \sum_{j=1}^{n} \omega_{ij} \qquad (i = 1,2,\ldots,n) \quad (2)$$

$$L = D^{-1/2}(D - W)D^{-1/2} \quad (3)$$

After the graph is constructed, the essence of the cut of graph is the eigen-decomposition of Laplacian matrix L. The first k eigenvectors labeled as $F_{n \times k}$ are computed, while k denotes the number of clusters. Here, we set the cluster number into 3. For $i = 1,2,\ldots,n$, f_i is denoted as the new feature vector and clustered into two or three clusters with k-means or other algorithms. The spectral clustering is implemented in sklearn, which is a Python package. It provides two clustering methods for f_i: k-means is a popular method but it is sensitive to initialization, while discretization is more robust and suitable for the clustering of multiple experimental data under the same experimental conditions.

Detailed process:

Input:

- Number of clusters, K
- Dataset $V = [v_i]_{N \times 2}$

Algorithm:

- Compute the weighted adjacency matrix $W = [w_{ij}]_{N \times N}$ and construct a fully connected similarity graph.

$$w_{ij} = w_{ji} = \exp\left(-\frac{\| v_i - v_j \|_2^2}{2\sigma^2}\right)$$

- Compute the degree matrix D, which is defined as the diagonal matrix with the degrees d_i on the diagonal.

$$d_i = \sum_{j=1}^{n} \omega_{ij}$$

- Compute the normalized Laplacian matrix L.

$$L = D^{-1/2}(D-W)D^{-1/2} = I - D^{-1/2}WD^{-1/2}$$

- Compute the first K eigenvectors labeled as $F_{N\times K}$ by eigen-decomposition of Laplacian matrix L, and for $i = 1,2,...,N$, the row vector f_i is denoted as the new feature vector.
- Cluster each row of $F_{n\times k}$ into K clusters using k-means or any algorithm.
- Assign the label of v_i to the cluster k according to the label of f_i.

Output:

- Clusters $C_1, C_2, ..., C_K$

Reference

1 Z. L. Hu, Z. Y. Li, Y. L. Ying, J. J. Zhang, C. Cao, Y. T. Long, H. Tian, Anal. Chem., 2018, 90, 4268-4272.

2 Z. L. Hu, M. Y. Li, S. C. Liu, Y. L. Ying, Y. T. Long, Chem. Sci., 2019, 10, 354-358.

3 C. Cao, Y.-L. Ying, Z.-L. Hu, D.-F. Liao, H. Tian and Y.-T. Long, *Nat. Nanotechnol.*, 2016, **11**, 713–718.

4 Y. Q. Wang, M. Y. Li, H. Qiu, C. Cao, M. B. Wang, X. Y. Wu, J. Huang, Y. L. Ying and Y. T. Long, *Anal. Chem.*, 2018, **90**, 7790–7794.

5 M. P. Gallego, M. F. Breton, F. Discala, L. Auvray, J. M. Betton, J. Pelta, *ACS Nano*, 2014, **8**, 11350–11360.

6 A. Meller, L. Nivon, D. Branton, *Phys. Rev. Lett.*, 2001, **86**, 3435–3438.

7 S. Park and H. Zhao, *Bioinformatics*, 2018, **34**, 2069–2076.

8 C. Ding, *Energy*, 2004, **CI**, 1–20.

9 T. Caliñski and J. Harabasz, *Commun. Stat.*, 1974, **3**, 1–27.

10 R. Kannan, S. Vempala and A. Veta, 2002, 367–377.

11 W. S. K. Fernando, P. H. Perera, H. M. S. P. B. Herath, M. P. B. Ekanayake, G. M. R. I. Godaliyadda and J. V. Wijayakulasooriya, *15th Int. Conf. Adv. ICT Emerg. Reg. ICTer 2015 - Conf. Proc.*, 2016, 21–24.

12 A. Paccanaro, J. A. Casbon and M. A. S. Saqi, *Nucleic Acids Res.*, 2006, **34**, 1571–1580.