KECERDASAN KOMPUTASIONAL A

REVIEW JURNAL

"A Robust Elicitation Algorithm for Discovering DNA Motifs Using Fuzzy Self-Organizing Maps"

Oleh :

I KOMANG ARYA GANDA WIGUNA

15/388479/PPA/04918

PROGRAM STUDI S2 ILMU KOMPUTER

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS GADJAH MADA

YOGYAKARTA

2016

A Robust Elicitation Algorithm for Discovering DNA Motifs Using Fuzzy Self-Organizing Maps

Dianhui Wang, Senior Member, IEEE, and Sarwar Tapan
IEEE Transactions on Neural Networks and Learning System, Vol. 24, no. 10, October 2013

Pendahuluan

Pada paper ini dilakukan penelitian terkait dengan mengidentifikasi motif DNA di daerah promoter untuk memahami mekanisme regulasi gen. Pendekatan komputasi untuk menemukan motif DNA diakui sebagai alat yang berguna untuk ahli biologi, yang sangat membantu dalam menghemat waktu eksperimental dan biaya di laboratorium.

Ide dasar di balik algoritma penemuan motif DNA berbasis clustering untuk mengekstrak satu set cluster, yang terdiri dari tipe yang sama sama panjang (*k*-length) subsequences (dikenal sebagai *k*-mer), diikuti oleh penggabungan dan mengoptimalkan beberapa potensi cluster dengan tingkat yang cukup untuk kehadiran sifat motif fungsional.

Self-organizing maps (SOMs) dapat digunakan untuk penemuan motif melalui mengelompokkan setara k-mer yang diekstrak dari urutan DNA. Namun, penemuan motif dengan algoritma berbasis SOM saat pengklusteran tidak adil memperlakukan sampel data yang tergeletak di sekitar batas-batas klaster dengan menetapkan mereka ke salah satu node, yang dapat mengakibatkan kinerja sistem tidak dapat diandalkan. Hal ini dapat berdampak negatif besar terhadap kinerja penemuan motif DNA, terutama dalam menemukan motif lemah. Ini berarti bahwa motif yang diambil dari pendekatan crisp partitioning sering gagal untuk melampirkan posisi motif lemah, yang akhirnya menurunkan kinerja dan keandalan alat motif temuan berbasis SOM.

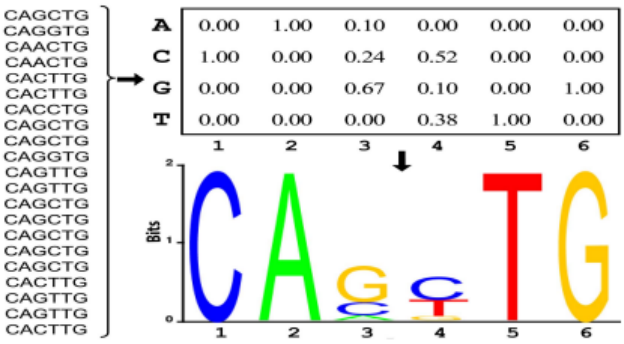
Untuk mengatasi masalah tersebut, pada paper ini menyajikan pendekatan berbasis Fuzzy SOMs (FSOMs) untuk menemukan motif, dimana FSOM melakukan fuzzy c-means (FCMs) soft clustering. Setiap cluster di FSOMs memiliki hubungan global untuk semua *k*-mer di dataset, yang meningkatkan kemungkinan melampirkan degenerasi contoh untuk motif fungsional. Termotivasi oleh algoritma pembelajaran tambahan dari fuzzy SOMs, algoritma batch learning untuk pelatihan jaringan FSOM dikembangkan dalam paper ini. Dalam paper ini, kerangka kerja yang kuat, bernama Robust Elicitation Algorithms for Discovering (READ) motif DNA, diusulkan dengan memanfaatkan beberapa FSOMs dan beberapa heuristik yang inovatif untuk optimasi calon motif DNA.

Paper ini bertujuan untuk mengembangkan kerangka kerja (framework) yang handal untuk menemukan motif DNA, di mana Fuzzy SOMs, dengan integrasi fuzzy c-means fungsi keanggotaan dan skema batch-learning standar, yang digunakan untuk mengekstrak motif diduga dengan panjang bervariasi secara rekursif.

Motif Modelling

Dalam komputasi, motif DNA dapat dimodelkan oleh satu set *k*-mers yang diekstraksi dari koleksi urutan promotor satu set gen yang diatur. Position Frequency Matrix (PFM)

digunakan untuk menggambarkan posisi signifikan dari nukleotida. Sebuah motif dapat divisualisasikan menggunakan logo seperti ditunjukkan pada Gambar 1.



Gambar 1 Representasi motif PFM dan visualisasi menggunakan entropy-based logo

Notasi berikut akan digunakan dalam makalah ini. Misalkan $D = \{s1, s2, ..., sn\}$ menjadi dataset dengan urutan yang diyakini mengandung motif peraturan. PFM dinotasikan sebagai M, digambarkan sebagai matriks, yaitu $M = [f(b_i, i)]_{4\times k}$, dimana $b_i \in \chi = \{A, C, G, T\}$ dan $f(bi,i)$ merupakan frekuensi relatif dari kemunculan b_i nukleotida di posisi ke-i. Demikian pula, k-mer, $K_s = q_1q_2 ... q_k$, dapat dikodekan sebagai matriks biner $K = [k(bi,i)]_{4\times k}$ dengan $k(q_i, i) = 1$ dan $k(b_i,i) = 0$ untuk $b_i = q_i$. Misalnya, k-mer $K_s = AGCGTGT$ dikodekan sebagai

$$K = encode(K_s) = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}_{4\times k}.$$

Untuk set binery encoded diberikan k-mer, $S=\{K_1, K_2, ..., K_P\}$, motif PFM M_S dapat dihitung dengan $M_S = \frac{1}{P}\sum_{i=1}^{P} K_i$. Dalam paper ini dijelaskan 2 cara untuk melakukan model dari motof DNA, yaitu :

- Motif modelling using MISCORE
  Mismatch-based matrix similarity score (MISCORE) adalah skema baru dan efektif untuk mengukur kesamaan antara k-mer dan model motif yang menggunakan karakterisasi gabungan fitur motif fungsional tanpa asumsi pada ketergantungan nukleotida. Menggunakan MISCORE, metrik untuk mengukur potensi dari motif kandidat untuk menjadi salah satu fungsional dapat didefinisikan. Mengingat satu set k-mer S, model PFM yang dapat dihitung sebagai $M_S = 1/|S| \sum_{\forall K\in S} e(K)$, dan MISCORE-based motif score (MMS), dinotasikan dengan $R(S)$, didefinisikan sebagai berikut :
  $$R(S) = \frac{1}{|S|}\sum_{\forall K\in S} r(K, M_s)$$
- Localized MMS for Model Assessment
  Posisi-spesifik metrik kesamaan menetapkan bobot yang sama untuk setiap posisi dalam model dan mengabaikan variabilitas antara blok lokal dari PFM motif fungsional. Karena PFM motif dapat dianggap sebagai descriptor dari preferensi yang mengikat, blok nukleotida yang mendasari diyakini membawa beberapa

informasi yang berguna yang merupakan karakterisasi keseluruhan pada motif. Oleh karena itu, PFM motif harus didekomposisi menjadi satu set blok lokal dengan bobot yang berarti sesuai dengan manfaatnya menjadi fungsional.

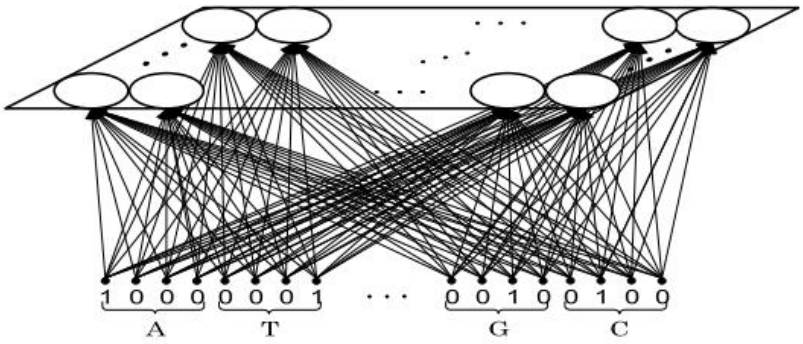
$$r_l(K, M_s) = \sum_{j=1}^{k-w+1} g_j \left(\frac{d(\beta_j(K), \beta_j(M_s))}{d(\beta_j(K), \beta_j(M_{ref}))} \right)$$

Perhatikan bahwa localized-MMS bertujuan untuk meningkatkan daya diskriminasi untuk motif yang lemah, sementara dalam tahapannya erat kaitannya dengan MMS untuk motif yang kuat.

Dalam peper ini, motif modelling tersebut digunakan untuk menilai dan menentukan peringkat kandidat motif ekstrak dari FSOMs. Versi lokal bertujuan untuk meningkatkan kinerja sistem pada motif yang lemah.

Fuzzy Self-Organizing Maps

FSOMs dapat dianggap sebagai versi perbaikan dari SOMs klasik. Dalam penelitian ini digunakan algoritma batch learning untuk FSOMs. Algoritma ini akan diterapkan dalam paper READ ini untuk menemukan motif DNA.



Gambar 2 Binary encoded k-mer = "AT...GC" di FSOM

- SOM Batch Learning
  Dalam paper ini untuk penemuan motif DNA, datasetnya adalah seperangkat *k*-mer diekstrak dari input sequence dan node *i* dalam jaringan SOM menyimpan node-PFM (*Mi*). Menggunakan notasi ini, algoritma batch-learning standar dari jaringan SOM dapat digambarkan sebagai berikut:

$$M_i(t+1) = \frac{\sum_{j=1}^{|\Gamma|} h_{icj}(t) K_j}{\sum_{j=1}^{|\Gamma|} h_{icj}(t)}$$

  Dimana c_j adalah *best matching node* for K_j pada iterasi ke *i.*
- Fuzzy SOM Batch Learning
  Batch learning dapat dituliskan kembali dengan membership fungsi singleton sebagai berikut

$$M_i(t+1) = \frac{\sum_{j=1}^{|\Gamma|} \sum_{k=1}^{N} U_{kj}(t) h_{ik}(t) K_j}{\sum_{j=1}^{|\Gamma|} \sum_{k=1}^{N} U_{kj}(t) h_{ik}(t)}$$

Dimana U adalah binary membership matrix, yang mewakili berhubungan dengan K_j dan $M_k(t)$, dapat didefinisikan sebagai berikut:

$$U_{kj}(t)\begin{cases}1, k = \arg min_q\{r\left(K_j, M_q(t)\right)\}\\ 0, otherwise\end{cases}$$

Nilai keanggotaan matriks U dapat diganti dengan matriks keanggotaan kabur μ seperti pada algoritma FCM, di mana $\mu_{kj}(t)$ merupakan derajat keanggotaan dari K_j dengan memperhatikan model $M_k(t)$

$$\mu_{kj}(t+1) = \left[\sum_{l=1}^{N}\left(\frac{r(K_j, M_k(t))}{r(K_j, M_l(t))}\right)^{\frac{2}{m-1}}\right]^{-1}$$

Dimana m adalah fuzziness regulator.

Dalam paper ini diusulkan algoritma batch-learning umum untuk jaringan SOM dengan keanggotaan fuzzy yang terlibat dalam proses pembelajaran. Jaringan SOM dengan teknik pembelajaran seperti ini disebut sebagai fuzzy SOM. Sehingga, dari node-PFM (prototipe) *Mi* dapat diperbarui dengan menggunakan algoritma Fuzzy SOM batch learning berikut:

$$M_i(t+1) = \frac{\sum_{j=1}^{|\Gamma|}\sum_{k=1}^{N}\mu_{kj}^{m}(t)h_{ik}(t)K_j}{\sum_{j=1}^{|\Gamma|}\sum_{k=1}^{N}\mu_{kj}^{m}(t)h_{ik}(t)}$$

Robust Motif Discovery Framework

Framework READ dalam paper ini melakukan pelatihan beberapa FSOM untuk $k_{min} \leq k \leq k_{max}$. FSOMs dilatih oleh semua *k-mer* ekstraksi dari dataset yang diberikan dan peringkat atas motif calon *T* yang dihasilkan sebagai model awal. Kemudian dilakukan proses penggabungan untuk mengambil beberapa contoh motif hilang. Proses postprocessing ditingkatkan diterapkan untuk memperbaiki model calon motif. Akhirnya, peringkat teratas motif calon *T* dikembalikan sebagai motif akhir. Gambaran dari framework READ disajikan pada Gambar. 3, yang terdiri dari tiga komponen utama:

- **(C1) Pelatihan Sistem**
  Tahap pelatihan terdapat 3 proses yaitu
  1) Inisialisasi : struktur dari jaringan (node), *k-mer* dataset dan PFM model awal
  2) *Fuzzy SOM Learning* : fungsi dari membership untuk setiap node-PFM $M_i(t)$ dan update setiap node-PFM $M_i(t+1)$ di setiap akhir epoch
  3) *Termination* : tahap pelatihan selesai saat meksimum epoch (default 100 di READ). Cara lain dapat menggunakan *objective function* ($J_m(t)$) *FCM*, jika $J_m(t)$ bergerak secara terus sementara σ cukup menyusut, proses updating dapat dihentikan
- **(C2) Ekstraksi Motif**
  Hasil dari proses pelatihan akan membentuk fuzzy-PFM dari setiap node, dimana semua k-mer memiliki derajat tertentu. Untuk mendapatkan peringkat dari model fuzzy-PFM perlu disamakan dan menentukan metrik $(R_f(M_i))$ untuk mengevaluasi model motif fuzzy.

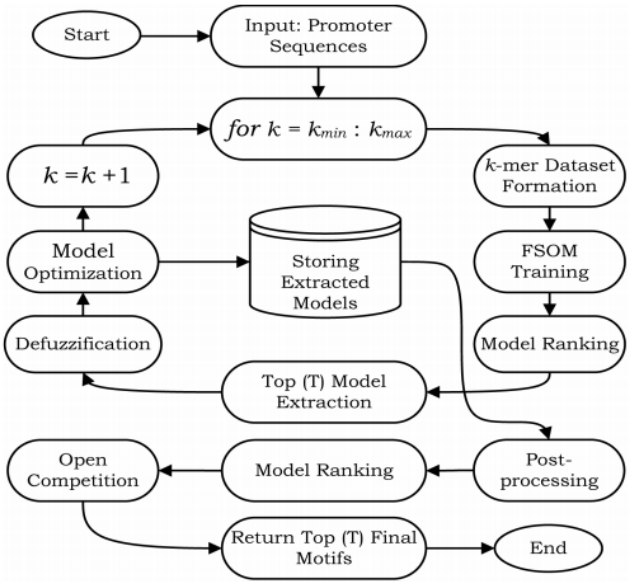
Nilai $R_f(M_i)$ yang kecil menunjukkan potensi untuk menjadi motif yang diduga. Candidate motif diurutkan berdasarkan Rf dan top-ranked T candidate akan dipilih. Setiap fuzzy candidate yang terpilih akan dikonversikan melalui defuzzifikasi, pilih nilai keanggotaan tertinggi untuk mepertahankan model.

Dalam paper ini diusulkan sebuah k-mer dari setiap node tetangga dari top-ranked candidate model. Hal ini dilakukan berdasarkan pengamatan bahwa model tetangga langsung (node) dari top-ranked model sering ditemukan untuk berbagai pola motif cukup mirip. Sehingga untuk menentukan sinyal motif lemah dari k-mer diambil dari k-mer yang terkandung dalam model tetangga melalui proses optimasi heuristik.

- **(C3) Post-processing**

  Post processing pada READ bertujuan untuk memperbaiki candidate ekstrak untuk meningkatkan sistem ketahanan dan kinerja motif penemuan. Pertama, ruang lingkup motif sasaran harus ditetapkan, yaitu maksimum (minimum) P_{max} (P_{min}) jumlah prediksi contoh motif per urutan promoter. Kedua, untuk beberapa kasus motif lemah dapat diambil (satu per satu) dari dataset k-mer melalui menambah atau mengganti prediksi yang ada dari setiap urutan promoter insi.

  Skema pengolahan post processing ini secara cepat dan mampu cepat mengubah menebak sub optimal dari motif ke dalam target satu. Oleh karena itu, skema perbaikan ini dapat digunakan untuk pemulihan efektif motif candidate noisy yang dihasilkan oleh ukuran peta pengaturan yang tidak tepat atau pelatihan yang tidak tepat dari FSOMs.



Gambar 3 READ framework

Performance Evaluation

- Comparison Using Real DNA Datasets

  Evaluasi akan dilakukan menggunakan real DNA dataset dan akan dibandingkan dengan menggunakan algoritma lainya seperti SOMEA, SOMBRERO, MEME AlignACE dan WEEDER. Hasil pada gambar 4, dimana hasil yang diambil rata-rata

dari 10 kali percobaan untuk mencari nilai precision (P), recall (R) dan F-measure (F). Dari tabel dapat dilihat READ unggul dari SOM-based SOMBRERO (state-of-the-art). Peningkatan mencapai rata-rata recal 10.39%, precision 30.99% dan F-measure 24.66%. Hal ini menunjukkan bahwa pendekatan berbasis Fuzzy SOM pada paper ini lebih baik dalam mengambil sinyal motif lemah (weak motif) daripada pendekatan berbasis SOM-based. READ juga memiliki rata-rata F-measure sebesar 0.73, hal ini menunjukkan yang terbaik diantara tools lainnya, WEEDER (0.72), MEME (0.65) dan AlignACE (0.69).

PERFORMANCE COMPARISON USING EIGHT REAL DNA DATASETS

	The average of F-measure (F), recall (R) and precision (P) rates over 10 runs																	
	READ			SOMEA			SOMBRERO			MEME			AlignACE			WEEDER		
TF	*R*	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*
CRP	0.76	0.84	0.80	0.91	0.89	**0.90**	0.83	0.43	0.56	0.59	0.88	0.69	0.83	0.98	**0.90**	0.75	0.83	0.79
GCN4	0.48	0.70	0.55	0.69	0.45	0.54	0.80	0.41	0.53	0.52	0.52	0.52	0.61	0.62	0.60	0.64	0.87	**0.73**
ERE	0.92	0.59	0.71	0.74	0.58	0.65	0.80	0.59	0.67	0.72	0.82	**0.77**	0.75	0.77	0.76	0.76	0.54	0.63
MEF2	0.96	0.87	**0.91**	0.81	0.99	0.89	0.35	0.22	0.27	0.92	0.80	0.85	0.86	0.87	0.86	0.88	0.88	0.88
SRF	0.91	0.77	**0.83**	0.84	0.74	0.79	0.67	0.83	0.74	0.87	0.72	0.79	0.83	0.71	0.77	0.83	0.71	0.76
CREB	0.82	0.78	**0.81**	0.89	0.67	0.77	0.83	0.43	0.56	0.59	0.88	0.69	0.52	0.66	0.57	0.79	0.71	0.75
E2F	0.69	0.74	0.71	0.82	0.64	0.71	0.76	0.67	0.71	0.68	0.64	0.65	0.75	0.68	0.71	0.89	0.67	**0.76**
MyoD	0.65	0.42	**0.51**	0.66	0.39	0.49	0.50	0.32	0.39	0.23	0.38	0.27	0.34	0.31	0.32	0.43	0.50	0.46
avg	0.77	0.71	**0.73**	0.80	0.67	0.72	0.69	0.49	0.55	0.64	0.71	0.65	0.69	0.70	0.69	0.75	0.71	0.72

Gambar 4 Hasil perbandingan menggunakan real DNA datasets

- Comparison Between SOMs, FSOMs and FCMs

  Pada pengujian ini dilakukan evaluasi FSOM dengan membandingkan READ dengan pendekatan SOM-based ($READ_s$) dan pendekatan FCM ($READ_f$) dalam penentuan motif DNA. Masing-masing pendekatan dilakukan evaluasi terhadap 8 dataset DNA dengan inisialisasi nilai parameter yang sama dengan dalam READ.

  Hasil menunjukkan READ dengan algoritma FSOM menunjukkan hasil yang lebih baik dari pada READ dengan algoritma SOM-based. Nilai rata-rata recal READ (0.77) dibandingkan dengan $READ_s$ (0.66). Dalam hasil ini menunjukkan bahwa pendekatan berbasis FSOM lebih masuk akal dalam kaitannya dengan fuzziness dalam dataset DNA.

  Hasil dengan menggunakan FCM juga menunjukkan bahwa kinerja yang masih rendah jika dibandingkan dengan READ menggunakan FSOM dalam penentuan motif DNA. Titik centroids dari neightborhood function dalam tahap awal pelatihan FSOM yang membedakan dengan FCM dimana hal ini yang menjadi salah satu faktor yang berkontribusi terhadap peningkatan kerja READ daripada $READ_f$.

	The average of R, P and F over 10 runs											
	READ				READ_s				READ_f			
TF	R	P	F	$\delta(F)$	R	P	F	$\delta(F)$	R	P	F	$\delta(F)$
CRP	0.76	0.84	**0.80**	(0.03)	0.68	0.73	0.70	(0.11)	0.61	0.77	0.68	(0.05)
GCN4	0.48	0.70	**0.55**	(0.10)	0.48	0.57	0.51	(0.08)	0.36	0.55	0.43	(0.05)
ERE	0.92	0.59	**0.71**	(0.06)	0.76	0.52	0.62	(0.14)	0.70	0.53	0.60	(0.08)
MEF2	0.96	0.87	**0.91**	(0.08)	0.70	0.63	0.66	(0.25)	0.68	0.60	0.64	(0.32)
SRF	0.91	0.77	**0.83**	(0.02)	0.79	0.78	0.78	(0.07)	0.58	0.67	0.62	(0.06)
CREB	0.82	0.78	**0.81**	(0.02)	0.79	0.76	0.78	(0.03)	0.77	0.71	0.74	(0.07)
E2F	0.69	0.74	**0.71**	(0.02)	0.56	0.60	0.58	(0.22)	0.63	0.50	0.56	(0.21)
MyoD	0.65	0.42	**0.51**	(0.09)	0.54	0.37	0.44	(0.10)	0.48	0.36	0.41	(0.15)
avg	**0.77**	**0.71**	**0.73**	(**0.05**)	0.66	0.62	0.64	(0.12)	0.60	0.59	0.58	(0.12)

Note: $\delta(F)$ denotes a standard deviation over 10 runs.

Gambar 5 Hasil perbandingan FSOM, SOM dan FCM

- Comparison Using Multiple Motif Datasets

  Pengujian ini menggunakan data dari Annoated Regulatory Binding Web (ABS, v1.0) database, pengujian ini dilakukan untuk melihat bagaimana tools ini melakukan secara bersamaan menemukan beberapa motif. Dari hasil pengujian pada gambar 6, menunjukan READ mencapai tingkat recall lebih baik dari SOMBRERO pada $\text{dataset}_{1,2,4,5}$. READ juga menghasilkan tingkat recall lebih baik dari MEME dan WEEDER pada 11 dari 15 motif. READ (0.45) memperoleh peningkatan sebesar 20% lebih dari SOMBRERO (0.36) dan peningkatan 9.6% dadi SOMEA (0.42) dalam hal F-measure dihitung dengan menggunakan semua dataset. Karena tingkat precision yang lebih baik, MEME mendapatkan rata-rata terbaik pada F-measure. Namun READ menunjukan keunggulan dalam recall daripada MEME yang memperoleh peningkatan sebesar 21,2% yang menguntungkan dalam beberapa motif penemuan.

  Tools berdasarkan SOM menghadapi dua kendala utama. Pertama, SOM-based menggunakan variasi panjang konsensus dari beberapa motif. Dalam pelatihan menggunakan panjang k-mer yang berbeda adalah solusi yang layak untuk mengatasi masalah ini dan telah diadopsi dalam READ untuk menemukan motif dengan variabel panjang k-mer. Kedua, sulit untuk menemukan map size yang tepat yang dapat melayani secara bersamaan tugas beberapa motif penemuan. Seperti dapat dilihat, map size yang lebih kecil akan menghasilkan tingkat presisi miskin tetapi tingkat recall meningkat, sementara map size yang lebih besar akan menghasilkan tingkat recall miskin tetapi tingkat presisi ditingkatkan.

  Pelatihan dengan map size yang berbeda dan berbagai panjang k-mer, baik dari kisaran yang cukup besar, dapat dilakukan dengan menggunakan tools motif penemuan berbasis SOM ini di banyak tugas motif penemuan. Namun, beban komputasi menjadi postprocessing besar dan lebih lama akan harus dibuat. Framework READ yang diusulkan dalam paper ini dapat secara otomatis menemukan solusi sub optimal melalui pelatihan berganda dengan berbagai panjang k-mer meskipun prototipe PFM secara acak diinisialisasi. READ melakukan dengan robust karena penggunaan jaringan FSOM, strategi penggabungan, dan skema postprocessing membaik, yang memberikan

kontribusi untuk penanganan ketidakpastian model yang disebabkan oleh ukuran peta pengaturan dan / atau random Model inisialisasi

PERFORMANCE COMPARISON USING ARTIFICIAL DATASETS WITH PLANTED MULTIPLE MOTIFS

		READ			SOMEA			SOMBRERO			MEME			WEEDER			READ (SOM)		
	3 TFs	*R*	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*
$Dataset_1$	CREB	0.39	0.29	0.33	0.43	0.26	0.33	0.44	0.26	0.33	0.20	1.00	0.33	0.00	0.00	0.00	0.33	0.25	0.29
	MyoD	0.27	0.19	0.23	0.48	0.23	0.31	0.20	0.08	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.12	0.14
	TBP	0.28	0.20	0.23	0.36	0.21	0.26	0.20	0.12	0.15	0.07	0.50	0.12	0.00	0.00	0.00	0.19	0.14	0.16
	avg	0.31	0.23	0.26	0.42	0.23	**0.30**	0.28	0.15	0.20	0.09	0.50	0.15	0.00	0.00	0.00	0.23	0.17	0.20
$Dataset_2$	NFAT	0.36	0.29	0.32	0.39	0.27	0.31	0.36	0.21	0.26	0.44	0.78	0.56	0.00	0.00	0.00	0.30	0.24	0.26
	HNF4	0.85	0.64	0.73	0.57	0.40	0.47	0.63	0.39	0.48	0.60	0.82	0.69	0.40	1.00	0.57	0.73	0.55	0.63
	SP1	0.53	0.43	0.47	0.50	0.53	0.50	0.53	0.35	0.42	0.38	0.54	0.44	0.00	0.00	0.00	0.54	0.43	0.48
	avg	0.58	0.45	0.51	0.49	0.40	0.43	0.51	0.32	0.39	0.47	0.71	**0.56**	0.13	0.33	0.19	0.52	0.41	0.46
$Dataset_3$	CAAT	0.28	0.20	0.23	0.43	0.21	0.25	0.32	0.17	0.22	0.29	0.80	0.42	0.00	0.00	0.00	0.30	0.21	0.25
	SRF	0.49	0.35	0.40	0.70	0.40	0.50	0.59	0.28	0.38	0.29	0.57	0.38	0.00	0.00	0.00	0.36	0.25	0.30
	MEF2	0.61	0.46	0.52	0.79	0.45	0.57	0.65	0.31	0.27	0.80	0.57	0.67	0.27	1.00	0.42	0.53	0.40	0.45
	avg	0.46	0.33	0.38	0.64	0.35	0.44	0.52	0.25	0.29	0.46	0.65	**0.49**	0.09	0.33	0.14	0.39	0.29	0.33
$Dataset_4$	USF	0.61	0.52	0.56	0.68	0.39	0.48	0.73	0.48	0.57	0.41	0.88	0.56	0.00	0.00	0.00	0.35	0.30	0.32
	HNF3B	0.21	0.14	0.17	0.47	0.25	0.31	0.26	0.13	0.17	0.15	1.00	0.27	0.00	0.00	0.00	0.16	0.11	0.13
	NFKB	0.89	0.81	0.85	0.71	0.47	0.56	0.66	0.46	0.54	0.80	0.57	0.67	0.33	1.00	0.50	0.76	0.69	0.72
	avg	0.57	0.49	**0.53**	0.62	0.37	0.45	0.55	0.36	0.43	0.45	0.82	0.50	0.11	0.33	0.17	0.42	0.36	0.39
$Dataset_5$	GATA3	0.48	0.36	0.41	0.61	0.37	0.46	0.49	0.33	0.36	0.40	0.75	0.52	0.40	1.00	0.57	0.45	0.34	0.38
	CMYC	0.94	0.75	0.83	0.74	0.47	0.57	0.89	0.70	0.84	0.75	1.00	0.86	0.19	0.75	0.30	0.58	0.47	0.52
	EGR1	0.61	0.43	0.51	0.66	0.36	0.47	0.47	0.26	0.33	0.64	0.81	0.72	0.00	0.00	0.00	0.58	0.41	0.48
	avg	0.68	0.51	0.58	0.67	0.40	0.50	0.62	0.43	0.51	0.60	0.85	**0.70**	0.20	0.58	0.29	0.54	0.40	0.46
avg{5 *datasets*}		0.52	0.40	0.45	0.57	0.35	0.42	0.49	0.30	0.36	0.41	0.71	**0.48**	0.11	0.32	0.16	0.42	0.32	0.37

Gambar 6 Hasil pengujian dataset

- Robustness Analysis

  Pengujian ini dilakukan untuk mengetahui ketahanan (robustness) dari tools dimana jumlah cluster sangat mempengaruhi kualitas clustering sehingga dilakukan pengujian dari penggunaan map size yang berbeda.

  Dari hasil yang diperoleh pada gambar 7, framework READ menunjukkan nilai standar deviasi yang kecil daripada SOMBRERO dan SOMEA yang menunjukkan ketahanan yang lebih baik dalam menangani perubahan map size.

ROBUSTNESS ANALYSIS OF SOM AND FSOMS BASED TOOLS USING DIFFERENT MAP SIZES

	The average of F-measure over 10 runs with different map sizes											
	map size $= 10 \times 10$			*map size* $= 15 \times 15$			*map size* $= 20 \times 20$			*standard deviation*		
TF	READ	SOMEA	SOMBRERO	READ	SOMEA	SOMBRERO	READ	SOMEA	SOMBRERO	READ	SOMEA	SOMBRERO
CREB	0.80	0.70	0.41	0.79	0.76	0.67	0.78	0.72	0.67	**0.008**	0.031	0.150
CRP	0.79	0.81	0.71	0.79	0.66	0.71	0.69	0.58	0.52	**0.060**	0.117	0.110
E2F	0.70	0.58	0.73	0.70	0.69	0.63	0.71	0.72	0.67	**0.004**	0.074	0.050
ERE	0.72	0.53	0.42	0.76	0.66	0.60	0.71	0.61	0.74	**0.028**	0.066	0.160
GCN4	0.53	0.41	0.44	0.50	0.51	0.52	0.49	0.58	0.60	**0.018**	0.085	0.080
MEF2	0.92	0.68	0.92	0.85	0.91	0.80	0.75	0.82	0.44	**0.087**	0.116	0.250
MyoD	0.50	0.32	0.23	0.52	0.49	0.42	0.44	0.47	0.49	**0.043**	0.093	0.135
SRF	0.82	0.70	0.67	0.81	0.77	0.72	0.72	0.71	0.71	0.055	0.038	**0.026**

Gambar 7 Hasil pengujian ketahanan dengan map size berbeda pada tiap dataset

Kesimpulan

Penemuan komputasi motif DNA adalah tugas yang bermakna dan menantang di bioinformatika. Dari perspektif pembelajaran-sistem, menemukan candidate motif melalui pembelajaran adalah untuk menambang satu set pola halus dengan beberapa signifikansi statistik. Hal ini tidak dapat dilakukan tanpa pemahaman yang baik tentang motif DNA, dan, karena itu, karakterisasi pada motif dan kesamaan yang tepat metrik yang digunakan dalam pembelajaran sangat penting untuk lebih meningkatkan tools yang ada.

Paper ini memberikan kontribusi untuk pengembangan algoritma elisitasi yang kuat untuk penemuan motif DNA menggunakan jaringan Fuzzy SOM. Sebuah algoritma pembelajaran batch yang baru untuk FSOMs diusulkan dengan mengintegrasikan fungsi keanggotaan FCM dalam algoritma batch-belajar dari jaringan SOM standar. Untuk mencapai kinerja yang lebih baik dan dapat diandalkan, berdasarkan pekerjaan peneliti sebelumnya dan skema pembelajaran yang diusulkan disarankan dalam paper ini, beberapa jaringan FSOM digunakan untuk mengekstrak candidate motif dengan berbagai panjang k-mer dari urutan masukan dengan penggabungan efektif model dan pengolahan optimasi. Dua pendekatan berbasis SOM, SOMBRERO dan SOMEA, dan tools terkemuka lainnya termasuk MEME terkenal, AlignACE, dan WEEDER yang diuji dengan menggunakan delapan dataset nyata dan lima dataset buatan. Hasil penelitian menunjukkan bahwa algoritma yang diusulkan dalam paper ini memiliki potensi yang baik untuk menguntungkan dan kokoh melakukan motif penemuan DNA.