Supplementary data for: Large-scale motif discovery using DNA Gray code and equiprobable oligomers

Natsuhiro Ichinose [1], Tetsushi Yada [1] and Osamu Gotoh [1]

[1]Graduate School of Informatics, Kyoto University

S.1 Properties of equiprobable oligomers

Table S1: Average and standard deviation of information contents of equiprobable oligomers as functions of the order of Markov model and the threshold θ.

Order		$\theta = 5$	$\theta = 10$	$\theta = 15$	$\theta = 20$
1	average	6.1	11	16	21
	standard deviation	0.48	0.6	0.65	0.62
2	average	6.1	11	16	21
	standard deviation	0.55	0.65	0.66	0.66
3	average	6.1	11	16	21
	standard deviation	0.55	0.68	0.68	0.68
4	average	8.2	11	16	21
	standard deviation	0.73	0.69	0.69	0.69

Table S1 shows the averages and standard deviations of the information contents of equiprobable oligomers under various settings of background Markov model and threshold value. The background Markov model is derived from the human promoter sequences in cisRED [Robertson *et al.*, 2006]. The results indicate that the standard deviations are almost constant whereas the averages increase in proportion to the threshold values. Since we use a large threshold value for large-scale sequences (ex., $\theta = 24$ for the cisRED sequences), the variation in the information contents is small relative to the average, indicating that the approximation of the equiprobable oligomers is effective enough for our target size of sequences.

S.2 Validation of threshold adjustment

Table S2: Correlation coefficient nCC for the threshold difference γ.

		nCC						
L [Mbp]	% MCS	$\gamma = -3$	$\gamma = -2$	$\gamma = -1$	$\gamma = 0$	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$
0.6	5	0.014	0.026	0.12	0.18	0.22*	0.22*	0.22*
	10	0.016	0.13	0.25	0.33	0.34*	0.33	0.3
	50	0.087	0.25	0.35	0.4*	0.4*	0.38	0.35
	80	0.11	0.22	0.32	0.38*	0.38*	0.37	0.35
1.2	5	-0.013	0.076	0.19	0.27	0.28*	0.26	0.26
	10	0.069	0.17	0.3	0.36*	0.36*	0.34	0.31
	50	0.17	0.24	0.37	0.42*	0.42*	0.4	0.37
	80	0.15	0.23	0.35	0.39	0.4*	0.38	0.35
12	5	0.08	0.2	0.33	0.39*	0.39*	0.36	0.33
	10	0.12	0.25	0.39	0.43	0.44*	0.41	0.38
	50	0.14	0.27	0.4	0.44*	0.44*	0.42	0.39
	80	0.12	0.24	0.35	0.4*	0.4*	0.38	0.36

(*) best nCC, (L) length of sequences, (% MCS) percentage of motif-containing sequences

In our method, a cluster is defined as a contiguous region in the DNA Gray code within which the counts of oligomers are greater than 0, so that the threshold θ is critical for good balance between sensitivity and specificity. For the best balance, the expectation value of the count should be approximately 1 in the background model. Let p be the probability of occurrence of a particular equiprobable oligomer, then the expectation value of its count is pL, where L is the total length of sequences. Accordingly, $pL \approx 1$. From Table S1, the probability p can be approximated by $-\log_2(p) = \theta + \epsilon$ and $\epsilon \approx 1$. Therefore, the balanced condition of the threshold is estimated to be $\theta' = \log_2(L) - 1$, which is used as the default value in the application.

We validate the above estimate by the simulation with synthetic data that consist of background sequences and motif models. The background sequences are random sequences generated by the third-order Markov model derived from the human promoter sequences. The motif models are 97 vertebrate motifs in the JASPAR database (JASPAR CORE 2008) [Sandelin *et al.*, 2004]. The instances of the motif models are randomly generated from their PWMs and embedded in the background sequences.

Based on the simulation, we compute the correlation coefficient at the nucleotide level as a function of the threshold θ expressed by the difference γ from the default value, *i.e.*, $\theta = \theta' + \gamma$. The results shown in Table S2 indicate that nCC has the best value at $\gamma = 0$ or 1 under various conditions of L and the percentage of motif-containing sequences, indicating the adequacy of our estimates of the default value of the threshold. Note that the default value is tuned so that the sensitivity is slightly dominant over the specificity on purpose, because a high sensitivity is usually desired in the practical cases. The value for γ can be adjusted by the option (`-p` or `--add-threshold`) at the runtime of Hegma.

S.3 Performance evaluation with ChIP-seq data

Table S3: Prediction statistics and calculation time for ChIP-seq data.

	nSn	nPPV	nCC	sSn	sPPV	sASP	time [s]
Hegma	0.057^{+}	0.61^{+}	0.052	0.088^{+}	0.84^{+}	0.47^{+}	0.28*
Weeder	0.21*	0.55	0.058^{+}	0.32*	0.33	0.33	1100
MEME	0.042	0.69*	0.069*	0.050	1.0*	0.52*	640
MDscan	0.057^{+}	0.53	0.017	0.088^{+}	0.40	0.24	1.7^{+}
BioProspector	0.036	0.56	0.023	0.058	0.62	0.34	55

(*) best statistics, ($^{+}$) second best

Table S3 shows the prediction statistics and the calculation time tested on the set of human ChIP-seq data in the cisRED database (Human Stat1 ChIP-seq peaks 1) [Robertson *et al.*, 2006]. Although the size of the data set is much smaller than our target size (kbp vs. Mbp), the sites in this dataset are more strongly supported by experiments than those of the main cisRED database. The number of sequences is 226, and the total number of nucleotides is approximately 136 kbp, of which valid (unmasked) nucleotides amount to approximately 122 kbp. The number of ChIP-seq peaks is 5,951. Because of the small data size, we can test five representative motif finding methods, Hegma (this work), Weeder (ver. 1.4.2) [Pavesi *et al.*, 2004], MEME (ver. 4.7.0) [Bailey and Elkan, 1994], MDscan (ver. 2004) [Liu *et al.*, 2002], and BioProspector (ver. 2004) [Liu *et al.*, 2001]. We ran these programs in the following settings:

Hegma	hegma datafile.fa
Weeder	weederlauncher.out datafile.fa HS medium
MEME	meme datafile.fa -maxsize 150000 -dna -V
MDscan	MDscan -i datafile.fa
BioProspector	BioProspector -i datafile.fa -W 8

The results indicate that MEME performs best among the five methods as indicated by the highest nCC and $sASP$ values. Since the statistics of Hegma are mostly the second best, its performance is proved to be stable. Furthermore, Hegma is three or four orders of magnitude faster than MEME. Together with the results shown in the text, we can conclude that Hegma may discover biologically relevant motifs in a wide range of sizes and types of sequence data.

S.4 Details of the results for ten most frequent motifs in cisRED

Tables S4 shows the details of the comparison between Hegma and Weeder with respect to their performance for the ten most frequent motifs in cisRED. nCC and $sASP$ of Weeder are superior to those of Hegma only for two examples of *AhR* and *HNF-1α* when the percentage of motif-containing sequences is 100% (indicated by asterisks). The significant superiority of Hegma over Weeder is supported by the small p-values of Wilcoxon signed rank tests: 9.3×10^{-7} for nCC and 5.0×10^{-7} for $sASP$.

Table S4: Prediction statistics for ten most frequent motifs.

% MCS	Motif	Method	nSn	nPPV	nCC	sSn	sPPV	sASP
40	AhR	Hegma	0.36	0.039	0.11	0.49	0.049	0.27
		Weeder	0	0	-0.0005	0	0	0
	aMEF-2	Hegma	0.2	0.019	0.053	0.25	0.026	0.14
		Weeder	0	0	0	0	0	0
	POU2F1	Hegma	0.061	0.0076	0.011	0.078	0.0097	0.044
		Weeder	0	0	0	0	0	0
	Pax-5	Hegma	0.26	0.026	0.072	0.36	0.039	0.2
		Weeder	0	0	0	0	0	0
	DEAF-1	Hegma	0.24	0.023	0.065	0.31	0.033	0.17
		Weeder	0	0	0	0	0	0
	CREB	Hegma	0.21	0.025	0.064	0.34	0.035	0.19
		Weeder	0	0	0	0	0	0
	HNF-1α	Hegma	0.13	0.015	0.033	0.17	0.021	0.094
		Weeder	0	0	0	0	0	0
	DP-1	Hegma	0.43	0.039	0.12	0.59	0.044	0.32
		Weeder	0	0	0	0	0	0
	RSRFC4	Hegma	0.15	0.014	0.037	0.18	0.018	0.1
		Weeder	0	0	0	0	0	0
	POU3F2	Hegma	0.14	0.014	0.033	0.18	0.019	0.1
		Weeder	0	0	0	0	0	0
60	AhR	Hegma	0.37	0.043	0.11	0.52	0.06	0.29
		Weeder	0	0	0	0	0	0
	aMEF-2	Hegma	0.19	0.017	0.04	0.23	0.023	0.13
		Weeder	0	0	0	0	0	0
	POU2F1	Hegma	0.077	0.0093	0.011	0.095	0.012	0.053
		Weeder	0	0	0	0	0	0
	Pax-5	Hegma	0.28	0.028	0.074	0.37	0.039	0.2
		Weeder	0	0	0	0	0	0
	DEAF-1	Hegma	0.23	0.025	0.061	0.31	0.035	0.17
		Weeder	0	0	0	0	0	0
	CREB	Hegma	0.23	0.026	0.062	0.34	0.034	0.18
		Weeder	0	0	0	0	0	0
	HNF-1α	Hegma	0.079	0.0078	0.009	0.12	0.012	0.065
		Weeder	0	0	0	0	0	0
	DP-1	Hegma	0.33	0.033	0.09	0.44	0.038	0.24
		Weeder	0	0	0	0	0	0
	RSRFC4	Hegma	0.17	0.015	0.036	0.22	0.022	0.12
		Weeder	0	0	0	0	0	0
	POU3F2	Hegma	0.081	0.0095	0.012	0.1	0.012	0.058
		Weeder	0	0	0	0	0	0

% MCS	Motif	Method	nSn	nPPV	nCC	sSn	sPPV	sASP
80	AhR	Hegma	0.33	0.038	0.092	0.46	0.053	0.26
		Weeder	0.031	0.029	0.022	0.049	0.027	0.038
	aMEF-2	Hegma	0.26	0.024	0.059	0.34	0.033	0.19
		Weeder	0	0	0	0	0	0
	POU2F1	Hegma	0.079	0.0097	0.0062	0.1	0.013	0.058
		Weeder	0	0	0	0	0	0
	Pax-5	Hegma	0.26	0.028	0.063	0.36	0.04	0.2
		Weeder	0	0	0	0	0	0
	DEAF-1	Hegma	0.21	0.021	0.047	0.29	0.032	0.16
		Weeder	0.002	0.047	0.0083	0.003	0.082	0.042
	CREB	Hegma	0.23	0.026	0.057	0.38	0.039	0.21
		Weeder	0	0	0	0	0	0
	HNF-1α	Hegma	0.072	0.0076	0.0017	0.097	0.011	0.054
		Weeder	0	0	0	0	0	0
	DP-1	Hegma	0.36	0.039	0.098	0.46	0.044	0.25
		Weeder	0.031	0.039	0.028	0.057	0.039	0.048
	RSRFC4	Hegma	0.16	0.014	0.028	0.21	0.022	0.12
		Weeder	0	0	0	0	0	0
	POU3F2	Hegma	0.13	0.014	0.021	0.17	0.017	0.091
		Weeder	0	0	0	0	0	0
100	AhR	Hegma	0.32	0.039	0.086	0.45	0.054	0.25
		Weeder	0.42	0.04	0.1*	0.68	0.033	0.36*
	aMEF-2	Hegma	0.25	0.022	0.049	0.31	0.032	0.17
		Weeder	0	0	0	0	0	0
	POU2F1	Hegma	0.091	0.011	0.0033	0.13	0.014	0.07
		Weeder	0	0	0	0	0	0
	Pax-5	Hegma	0.23	0.026	0.051	0.32	0.038	0.18
		Weeder	0	0	-0.0013	0	0	0
	DEAF-1	Hegma	0.21	0.022	0.042	0.29	0.033	0.16
		Weeder	0.0052	0.022	0.0062	0.018	0.05	0.034
	CREB	Hegma	0.22	0.028	0.053	0.34	0.039	0.19
		Weeder	0.011	0.043	0.017	0.026	0.059	0.043
	HNF-1α	Hegma	0.069	0.0073	-0.0056	0.097	0.011	0.054
		Weeder	0.19	0.052	0.084*	0.33	0.056	0.2*
	DP-1	Hegma	0.35	0.038	0.089	0.46	0.045	0.25
		Weeder	0.21	0.043	0.074	0.36	0.038	0.2
	RSRFC4	Hegma	0.17	0.014	0.022	0.22	0.02	0.12
		Weeder	0	0	0	0	0	0
	POU3F2	Hegma	0.11	0.011	0.0064	0.15	0.015	0.084
		Weeder	0	0	0	0	0	0

References

[Bailey and Elkan, 1994] Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proc. of the 2nd Int. Conf. on Intelligent Systems for Molecular Biology*, 28–36

[Liu *et al.*, 2001] Liu, X., Brutlag, D.L. and Liu, J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes, *Pac. Symp. Biocomp.*, 127–138

[Liu *et al.*, 2002] Liu, X.S., Brutlag, D.L. and Liu, J.S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments, *Nat. Biotech.*, **20**(8), 835–839

[Pavesi *et al.*, 2004] Pavesi, G., Mereghetti, P., Mauri, G. and Pesole, G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes., *Nucleic Acids Res.*, **32**, W199–W203

[Robertson *et al.*, 2006] Robertson, A.G. *et al.* (2006) cisRED: A database system for genome scale computational discovery of regulatory elements, *Nucleic Acids Res.*, **34** (Database issue), D68–73

[Sandelin *et al.*, 2004] Sandelin, A., Alkema, W., Engström, P., Wasserman, W.W. and Lenhard, B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles, *Nucleic Acids Res.*, **32** (suppl. 1), D91–D94