unpublished; available at ftp://ftp.isrec.isb-sib.ch/sib-isrec/pftools/); SAM version 1.3.3 (ref. 18;http://www.cse.ucsc.edu/research/compbio/sam.html); HMMER version 1.8.4 (S.R. Eddy, unpublished; available at http://hmmer.wustl.edu). Binding scores for oligonucleotide sequences were computed with the program pfsearch (pftools). Simulated SELEX data were generated by first converting the CTF/NFI profile (Fig. 1A) into an equivalent HMM with the program ptoh (pftools), and then by generating random instances from the HMM with the program hmme (HMMER). Exponential bases of 1.14, 1.19, and 1.36 were used to convert the same profile into different HMMs representing low-, medium-, and high-affinity binding sites, respectively. New HMMs were derived from the simulated SELEX data and from the Selex3 database with the program buildmodel (SAM). The details of the computational recipe can be found on our website (http://www.isrec.isb-sib.ch/selex_nf1/). The new profile (Fig. 2C) was computed from the new HMM with the program htop (pftools) using a logarithmic base of 1.26 for conversion (corresponding to the $10 \times \log_{10}$ scale used in the old profile). The profile weights were subsequently rescaled manually to conform to the conventions applied in the old profile (see Fig. 1A legend).

The covariance analysis was done on a set of 3,602 15-mer sites extracted from the Selex3 library with binding score ≥65 according to the new profile. Each sequence was presented in both orientations. The calculation of the χ^2 test variable and the mutual information value[2] was based on a 2×2 contingency table representation of the corresponding base frequencies: we considered only the presence or absence of the specific bases under consideration at each position.

Note: Supplementary information is available on the Nature Biotechnology website.

Acknowledgments
We thank Victor Jongeneel for support and suggestions, Roman Chrast and Stylianos Antonarakis for help with the SAGE procedure, Khalil Kadaoui for assistance, and Alan McNair for helpful comments on the manuscript. The financial support of the Ludwig Institute for Cancer Research, the Etat de Vaud, and the Swiss National Science Foundation (grants 31-63933.00 and 31-59370.99) are gratefully acknowledged.

Received 28 September 2001; accepted 10 May 2002

1. Bucher, P., Karplus, K., Moeri, N. & Hofmann, K. A flexible motif search technique based on generalized profiles. *Comput. Chem.* **20**, 3–29 (1996).
2. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids* (Cambridge Univ. Press, Cambridge, United Kingdom, 1998).
3. Ehret, G.B. *et al.* DNA binding specificity of different STAT proteins. Comparison of *in vitro* specificity with natural target sites. *J. Biol. Chem.* **276**, 6675–6688 (2001).
4. Klug, S.J. & Famulok, M. All you wanted to know about SELEX. *Mol. Biol. Rep.* **20**, 97–107 (1994).
5. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
6. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **3**, 175–185 (1998).
7. Roulet, E., Fisch, I., Junier, T., Bucher, P. & Mermod, N. Evaluation of computer tools for the prediction of transcription factor binding sites on genomic DNA. *In Silico Biol.* **1**, 21–28 (1998).
8. Roulet, E. *et al.* Experimental analysis and computer prediction of CTF/NF-I transcription factor DNA binding sites. *J. Mol. Biol.* **297**, 833–848 (2000).
9. Berg, O.G. & von Hippel, P.H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193**, 723–750 (1987).
10. Goodman, S.D., Velten, N.J., Gao, Q., Robinson, S. & Segall, A.M. *In vitro* selection of integration host factor binding sites. *J. Bacteriol.* **181**, 3246–3255 (1999).
11. Fields, D.S., He, Y.Y., Al-Uzri, A.Y. & Stormo, G.D. Quantitative specificity of the Mnt repressor. *J. Mol. Biol.* **271**, 178–194 (1997).
12. Vant-Hull, B., Payano-Baez, A., Davis, R.H. & Gold, L. The mathematics of SELEX against complex targets. *J. Mol. Biol.* **278**, 579–597 (1998).
13. Meisterernst, M., Gander, I., Rogge, L. & Winnacker, E.L. A quantitative analysis of nuclear factor I/DNA interactions. *Nucleic Acids Res.* **16**, 4419–4435 (1988).
14. Perier, R.C., Praz, V., Junier, T. & Bucher, P. The eukaryotic promoter database EPD. *Nucleic Acids Res.* **28**, 302–303 (2000).
15. Man, T.K. & Stormo, G.D. Non-independence of Mnt repressor-operator interactions determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.* **29**, 2471–2478 (2001).
16. Zhang, M.Q. & Marr, T.G. A weight array method for splicing signal analysis. *Comput. Appl. Biosci.* **9**, 499–509 (1993).
17. Burge, C.B. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
18. Hughey, R. & Krogh, A. Hidden Markov models for sequence analysis. Extension and analysis of the basic method. *Comput. Appl. Biosci.* **12**, 95–107 (1996).

An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments

X. Shirley Liu[1], Douglas L. Brutlag[2], and Jun S. Liu[3]*

Published online: 8 July 2002, doi:10.1038/nbt717

Chromatin immunoprecipitation followed by cDNA microarray hybridization (ChIP–array) has become a popular procedure for studying genome-wide protein–DNA interactions and transcription regulation. However, it can only map the probable protein–DNA interaction loci within 1–2 kilobases resolution. To pinpoint interaction sites down to the base-pair level, we introduce a computational method, Motif Discovery scan (MDscan), that examines the ChIP–array-selected sequences and searches for DNA sequence motifs representing the protein–DNA interaction sites. MDscan combines the advantages of two widely adopted motif search strategies, word enumeration[1–4] and position-specific weight matrix updating[5–9], and incorporates the ChIP–array ranking information to accelerate searches and enhance their success rates. MDscan correctly identified all the experimentally verified motifs from published ChIP–array experiments in yeast[10–13] (STE12, GAL4, RAP1, SCB, MCB, MCM1, SFF, and SWI5), and predicted two motif patterns for the differential binding of Rap1 protein in telomere regions. In our studies, the method was faster and more accurate than several established motif-finding algorithms[5,8,9]. MDscan can be used to find DNA motifs not only in ChIP–array experiments but also in other experiments in which a subgroup of the sequences can be inferred to contain relatively abundant motif sites. The MDscan web server can be accessed at http://BioProspector.stanford.edu/MDscan/.

Although the 10 to 1,000 binding loci selected by ChIP–array experiments may contain false positives, those with high ChIP–array enrichment are more likely to represent true positives with multiple protein–DNA binding sites. MDscan takes advantage of this knowledge by first searching the highly ChIP–array-enriched fragments thoroughly, generating multiple candidate motif patterns, and then updating and refining the candidate motifs using other less likely sequences, guided by statistical scoring functions derived from Bayesian statistical formulation[7]. We applied MDscan to both simulated and biological data sets and compared its performance with BioProspector[9], CONSENSUS[5], and AlignACE[8].

In simulation studies, nine motif models were manually created (Table 1A), representing three different motif widths and three

[1]*Stanford Medical Informatics,* [2]*Department of Biochemistry, Stanford University, Stanford CA 94305.* [3]*Department of Statistics, Harvard University, 1 Oxford Street, Cambridge MA 02138.*
**Corresponding author (jliu@stat.harvard.edu).*

different degrees of conservation measured by information content. Each test data set contained 100 sequences of 600 base pairs each, generated from a third-order Markov model estimated from the yeast intergenic regions. Several motif segments were generated from the motif matrices and added to the data-set sequences, replacing some segments of the same width. The number of motif segments added follows four abundance schedules (Table 1B), imitating a feature of ChIP–array-selected sequences such that the top sequences contain more motif segments. For every combination of the three motif widths, three motif strengths, and four motif abundances, 100 test data sets were generated, giving rise to a total of 3 × 3 × 4 × 100 = 3,600 data sets. MDscan, BioProspector, CONSENSUS, and AlignACE were used to search for motifs in each data set, and the top five motifs reported from each program are summarized in Table 2. We are interested in the number of times a program successfully detects the motif inserted in the 100 tests for each motif width–strength–abundance combination, and the average rank of the correct motif if it comes up in the top five. When the expected number of motif sites in the top sequences was unknown, MDscan achieved a similar accuracy to that of BioProspector; both were more accurate than CONSENSUS and AlignACE. When a reasonable range of the expected bases per motif site (~300) in the top sequences was given, MDscan was further improved to give much better results than all the other three algorithms. Using parameters that give the best results for each program, MDscan was 3.3 times faster than CONSENSUS, 16 times faster than BioProspector, and 93 times faster than AlignACE.

Following these successful results on simulation data, we applied MDscan to 12 published ChIP–array experiments[10–13], all performed on yeast (Table 3). The proteins of interest in these experiments are Ste12, Gal4, Rap1, and all the cell cycle–related proteins: Mbp1, Swi4, Swi6, Fkh1, Fkh2, Ndd1, Mcm1, Ace2, and Swi5. Ste12 responds to mating pheromones in haploid yeast cells and transcriptionally activates >200 genes[10]. Eight of the top ten motifs found by MDscan agreed with the published STE12 motif[14] (all-capital type is used to represent protein complexes such as MBF or binding motifs such as STE12; initial capitalization represents proteins such as Ste12). Gal4, one of the best-characterized transcriptional activators, activates genes involved in galactose metabolism[10].

Table 1A. Nine motif models for three motif widths and three motif strengths

Motif width (consensus)	Motif information content		
	S1	S2	S3
W8 (GACTACCA)	1.114	0.962	0.795
W12 (GACTACCATGGA)	0.944	0.840	0.753
W16 (AGGATCTAATGATCCT)	0.832	0.750	0.665

Table 1B. Four motif abundances

Expected copies of motif segments	Motif abundance			
	A1	A2	A3	A4
Among top 5 sequences	3	2.5	2	1.5
Among middle 35 sequences	1.4	1.1	0.8	0.5
Among last 60 sequences	0.4	0.3	0.2	0.1
Total expected motif segments	88	69	50	31

Motif information content is defined as

$$\frac{1}{w} \times \sum_{i=1}^{w}\sum_{j=A}^{T} p_{ij} \log_2(4 \times p_{ij}),$$

where p_{ij} is the frequency of base j at motif position i. Information content can range from 0 to 2, reflecting the weakest to the strongest motifs.

The motifs discovered by MDscan not only agreed with the published GAL4 motif[10], but also revealed a stronger palindromic pattern: CGGASSASTSTSSTCCG. Rap1 is a repressor–activator protein. In the telomere regions, it represses telomeric transcription and regulates telomere length; in other genomic regions, it represses transcription at silent mating–type loci, and regulates the transcription of genes encoding ribosomal proteins and glycolytic enzymes[12]. For the 727 Rap1-selected sequences, MDscan identified all top ten motifs as CACACACACAC, with sites coming from eight sequences containing extremely high numbers of $(CA)_n$ repeats. After these eight sequences were removed, MDscan found three different motifs in the remaining 719 sequences. As many of the top ten ChIP–array-enriched sequences were in telomere regions, we divided the data set into two groups containing 575 nontelomere and 142 telomere sequences, respectively, to perform MDscan. Among the three motifs identified from the 719 sequences, one motif with consensus CACCCATACAT appeared in all top ten results from the nontelomere group, agreeing perfectly with all the biologically verified RAP1 consensus sequences[15–17]; the other two were found in the telomere group, and could be the potential differential telomere binding sites of Rap1.

There are about 800 cell cycle–regulated genes, the transcription of which is regulated by transcription factors in a serial fashion[13]. Swi4 and Swi6 form SBF, and Mbp1 and Swi6 form MBF; both heterodimers are active during G1/S phase and regulate Ndd1 (ref. 13). MDscan correctly found the MCB motif from Mbp1 targets, the SCB motif from Swi4 targets, and both motifs from Swi6 targets[11,18]. Fkh1 activates S/G2-state genes[13], and MDscan identified the SFF motif from its targets[18]. The Mcm1–Fkh2–Ndd1 complex activates G2/M genes and regulates Swi5 and Ace2 (ref. 13). Targets of each component in the Fkh2–Mcm1–Ndd1 complex were found to contain the MCM1 motif[18]. In addition, MDscan detected a shorter motif in Fkh2, which resembled the SCB motif. Swi5, Ace2, and Mcm1 activate M/G1 genes, and they all regulate Cln3, which in turn activates SBF and MBF[13]. MDscan correctly identified the SWI5 motif from both Swi5 and Ace2 targets[18].

We also applied the other three algorithms to these ChIP–array-selected data sets. BioProspector failed to find the SCB motif from Swi4 and Swi6 targets, the SFF motif from Fkh1 targets, and the MCM1 motif from Fkh2 targets. AlignACE failed to find the MBF motif from Swi6 targets, the SFF motif from Fkh1 targets, the MCM1 motif from Ndd1 targets, and the SWI5 motif from Ace2 and Swi5 targets. With large data sets such as the Rap1 nontelomere sequences, MDscan was ~35 times faster than BioProspector and ~400 times faster than AlignACE. CONSENSUS only found seven correct motifs in the 12 experiments; among these motifs only four are partially correct (with mismatched or shifted bases from the correct consensus).

MDscan first uses a word-enumeration strategy to look for oligomers of width *w* (*w*-mers) that are abundant in the top sequences. Because MDscan enumerates only existing *w*-mers in the top sequences, its search time increases only quadratically with respect to the total number of bases in the top sequences for all motif sizes, and linearly with respect to the bases in the remaining sequences. By using *m*-match criterion (see Experimental Protocol) for initialization and adopting a statistically derived scoring function for subsequent evaluation and refinement, MDscan overcomes the limitation in existing word-enumeration algorithms of inflexible base substitutions. As shown by the results for the Rap1 and Swi6 data, MDscan can also find multiple motifs in one run.

Existing statistically based algorithms such as AlignACE[8], BioProspector[9], Gibbs motif sampler[7,19], and MEME[6] rely on iterative procedures (either expectation maximization or Gibbs sam-

Table 2. Simulation results of MDscan, BioProspector, CONSENSUS, and AlignACE

Tests	MDscan Expect 300 bases/site		MDscan Expect unknown		BioProspector		CONSENSUS		AlignACE Expect 6,000 bases/site	
	Times found	Avg. rank	Times found	Avg. rank	Times found	Avg. rank	Times found	Avg. rank	Times found	Avg. rank
W8S1A1	100	1.00	100	1.00	100	1.00	85	1.12	100	1.70
W8S1A2	97	1.01	96	1.01	100	1.00	51	1.22	91	2.49
W8S1A3	92	1.09	93	1.19	92	1.08	36	1.61	52	3.38
W8S1A4	68	1.57	54	1.87	46	1.65	11	2.09	4	3.25
W8S2A1	100	1.10	99	1.03	99	1.02	49	1.51	70	2.97
W8S2A2	91	1.23	80	1.15	97	1.10	27	1.30	36	3.56
W8S2A3	78	1.28	59	1.53	71	1.38	7	1.86	9	4.22
W8S2A4	37	2.68	20	2.05	19	1.89	3	3.67	0	0.00
W8S3A1	82	1.41	70	1.31	81	1.20	6	1.50	11	4.18
W8S3A2	60	1.50	48	1.94	63	1.37	3	2.33	0	0.00
W8S3A3	39	1.90	33	2.24	36	1.56	4	1.50	1	5.00
W8S3A4	20	2.45	8	2.50	4	3.25	2	1.00	0	0.00
W12S1A1	100	1.00	100	1.00	100	1.00	100	1.00	100	1.00
W12S1A2	100	1.00	100	1.00	100	1.00	98	1.00	100	1.01
W12S1A3	99	1.00	99	1.01	100	1.01	81	1.02	96	1.49
W12S1A4	91	1.11	81	1.10	74	1.22	39	1.54	50	2.80
W12S2A1	100	1.00	100	1.00	100	1.00	94	1.00	99	1.12
W12S2A2	99	1.01	96	1.00	100	1.03	86	1.12	95	1.68
W12S2A3	95	1.00	93	1.08	86	1.10	47	1.45	63	2.83
W12S2A4	72	1.19	58	1.33	43	1.44	14	1.57	14	3.36
W12S3A1	98	1.00	98	1.01	97	1.00	76	1.14	95	1.83
W12S3A2	93	1.03	87	1.06	93	1.03	49	1.35	71	2.62
W12S3A3	81	1.25	68	1.28	68	1.21	18	1.83	23	3.00
W12S3A4	51	1.69	33	1.76	16	1.69	7	1.43	1	3.00
W16S1A1	100	1.00	100	1.00	100	1.00	100	1.00	100	1.00
W16S1A2	100	1.00	100	1.00	100	1.00	100	1.00	100	1.01
W16S1A3	100	1.00	100	1.00	100	1.00	93	1.02	100	1.12
W16S1A4	97	1.00	95	1.01	87	1.10	71	1.45	63	1.97
W16S2A1	100	1.00	100	1.01	100	1.00	99	1.00	100	1.01
W16S2A2	100	1.00	100	1.00	100	1.00	95	1.00	100	1.05
W16S2A3	99	1.00	99	1.00	97	1.02	73	1.22	83	1.53
W16S2A4	79	1.05	79	1.11	61	1.20	31	1.55	35	2.69
W16S3A1	99	1.00	99	1.00	100	1.00	94	1.06	100	1.11
W16S3A2	100	1.00	100	1.00	99	1.00	84	1.21	95	1.63
W16S3A3	94	1.03	91	1.01	86	1.09	41	1.51	49	2.61
W16S3A4	69	1.17	62	1.10	26	1.31	11	1.55	7	3.86

For each method, the first column indicates the number of times the correct motif was within the top five motifs reported in the 100 test data sets; the second column indicates the average rank of the correct motif when it was in the top five. MDscan results for 300 bases/site and unknown motif frequency, CONSENSUS results for seeding with first and proceeding linearly, and AlignACE results for 6,000 bases/site (or ten expected motif sites for the whole data set) shown here are the best results among their different parameter specifications.

pling) with nonquantifiable convergence properties[20]. In practice, they often encounter serious local-maximum problems when dealing with large data sets and require multiple runs to ensure meaningful findings. As the data set becomes larger (as shown with Rap1 nontelomere sequences), the speed and sensitivity advantages of MDscan become more substantial.

Although it incorporates sequence-ranking information to accelerate the search and enhance sensitivity, MDscan is tolerant of moderate ranking errors—its only requirement is that the top sequences have stronger motif signals in general. If a reasonable range of the total number of motif elements in the top sequences can be estimated (300 base pairs per motif site in the simulation, for example), MDscan can use a scoring function derived from a Bayes formulation[7] to achieve a higher overall accuracy.

MDscan can also be applied to search the upstream sequences of all the induced or repressed genes in a genome from a single microarray experiment by using the most induced or most repressed sequences as the top group. In a preliminary test, MDscan finished searching through all motif widths between 5 and 15 in one yeast cell-cycle expression data set[18] in an hour and successfully identified the MBF, STRE, and REB1 motifs and another motif known to be involved in RNA processing (E. M. Conlon, X. S. Liu, J. D. Lieb, D. L. Brutlag, and J. S. Liu, unpublished data).

MDscan equips biologists with a computational tool for discovering transcription-factor binding motifs in data generated by gene expression and ChIP-array experiments. The combination of a heuristic word-enumeration approach and rigorous statistical modeling underlying MDscan also offers a promising strategy for tackling gene regulation problems in higher eukaryotes.

Experimental protocol

Consider a set of n DNA sequences selected from ChIP–array experiments, ranked according to their ChIP–array enhancement scores, from highest to lowest. MDscan first scrutinizes the top t (~3–20) sequences in the ranking to form a set of promising candidates. Assuming the protein-binding motif to be of width w, MDscan enumerates each nonredundant w-mer (seed) that appears in both strands of the top t sequences and searches for all w-mers in the top t sequences with at least m base pairs matching the seed (called m-matches). The m is determined so that the chance that a pair of randomly generated w-mers are m-matches of each other is <0.15%. For each seed, we find all the m-matches in the top t sequences and use them to form a motif weight matrix. If the expected number of bases per motif site in the top sequences can be estimated, we use the following approximate maximum a *posteriori* (MAP) scoring function of Liu *et al.*[7] to evaluate a matrix:

Table 3. MDscan results for sequences selected by ChIP–array experiments

Biological test (number of sequences (seq), data size (kb), running time on 1GHz PC)	Published binding motif consensus		MDscan results and motif rankings	
Ste12 (26 seq, 18 kb, 4 s)	STE12	TGAAACA[14]	* TGMAACA	1,2,5,8,9
			AAACMAA	3,10
			* GAAACAA	4
			* YTGAAAC	6,7
Gal4 (23 seq, 17 kb, 23 s)	GAL4	$CGGN_{11}CCG$[10]	* CGGASCASTSTSSTCCG	1–8,10
			* GGAGCACTGTTGACCGA	9
Rap1 (727 seq, 309 kb, 10 s)	RAP1	Same as below	* CACACACACACAC	1–10
Rap1 (719 seq, excluding the 8 seq with CA repeats, 305 kb, 10 s)	RAP1	RTRCACCCANNCMCC[15]	~ GGCACTTGCATCA	1,3
		RMAYCCRMNCAYY[16]	* ACCCATAYCTCAC	5,6
		RMACCCANNCAYY[16]	^ ACCCTTACACTAC	2
		ACACCCAYACAYYY[17]	^ CACTTACCCTACC	4,10
			^ ACTTACCCTACCA	7
			^ CTTACCCTACCMCY	8,9
Rap1 (577 nontelomere seq, excluding the 3 seq with CA repeats, 254 kb, 15 s)	RAP1	Same as above	* ACACCCATACATC	1–3,7
			* K ACACCCATACAT	4–6,8
			* CACCCATACATCT	9,10
Rap1 (142 telomere seq, excluding the 5 seq with CA repeats, 51 kb, 6 s)	RAP1	Same as above	~ KGCACTTGCMTCA	1,2
			~ GCACTTGCCTCAG	4
			^ MACTTACCCTACC	3,5,10
			^ ACTTACCCTACCA	6,8
			^ CTTACCCTACCAT	7,9
Mbp1 (137 seq, 92 kb, 3 s)	MCB	ACGCGT[18]	* ACGCGT	1–3,6,7
			* RACGCG	4,8–10
			* CGCGTC	5
Swi4 (210 seq, 148 kb, 6 s)	SCB	CGCGAAAA[11]	* ACGCGAA	1,4
	SCB	CACGAAA[18]	AACGCGA	2,3,5
			AAACGCG	6
			* CGCGAAA	7–10
Swi6 (222 seq, 143 kb, 7 s)	MCB	ACGCGT[18]	# ACGCGTM	1–3,6
			# GACGCGT	4
	SCB	CGCGAAAA[11]	# TAACGCG	5
	SCB	CACGAAA[18]	# AAACGCG	7,8
			@ ACGCGAA	9
			@ CGCGAAA	10
Fkh1 (184 seq, 120 kb, 10 s)	SFF	GTMAACAA[18]	* GTAAACAA	1,3,5–8
			SCGSGKSG	2,4
			* TAAACAAA	9,10
Fkh2 (197 seq, 135 kb, 7–12 s)	SCB	CGCGAAAA[11]	% ACGCSAAA	1,4
	SCB	CACGAAA[18]	% AAACGCGA	5
			% CGCCAAAA	6
			% AACGCGAA	7
	MCM1	TTACCNAATTNGGTAA[18]	& TTTCCTAATTAGGAA	1,2,4
			& TCCTAATTAGGAAAT	3,5
			& CCBAATTAGGAAATA	6–8,10
			& TTCCTAATTAGGAAA	9
Ndd1 (123 seq, 83 kb, 20 s)	MCM1	TTACCNAATTNGGTAA[18]	* TTTCCTAATTAGGAAA	1–4
			* TTCCTAATTAGGAAAT	5–6
			TCCTAATTAGGAAATA	7–8,10
			CCTAATCAGGAAATAT	9
Mcm1 (116 seq, 79 kb, 15 s)	MCM1	TTACCNAATTNGGTAA[18]	* TTTCCTAATTAGGAAA	1–6
			* ATTTCCTAATTAGGAA	7–10
Ace2 (77 seq, 56 kb, 3 s)	SWI5	ACCAGC[18]	* DCCAGC	1,4,8
			* CCAGCR	2,3,5,6,9
			CCSGSC	7,10
Swi5 (90 seq, 64 kb, 2 s)	SWI5	ACCAGC[18]	* RCCAGC	1,2,10
			* CCWGSM	3–9

*, Motif discovered by MDscan agreed with the published motif consensus; ~, a second and ^, a third motif identified as potential binding to Rap1 at telomere regions; #, SCB motif and @, MCB motif both found in Swi6 targets; %, MCB motif and &, MCM1 motif found in Fkh2 targets when different motif widths were used for the search.

$$\frac{x_m}{w} \times \left[\sum_{i=1}^{w}\sum_{j=A}^{T} p_{ij}\log p_{ij} - \frac{1}{x_m}\sum_{\text{all segments}} \log(p_0(s)) - \log(\text{expected bases/site})\right]$$

where x_m is the number of m-matches aligned in the motif, p_{ij} is the frequency of nucleotide j at position i of the motif matrix and $p_0(s)$ is the probability of generating the m-match s from the background model. When the expected number of sites in the top sequences is unknown, we evaluate the motif matrix with:

$$\frac{\log(x_m)}{w}\left[\sum_{i=1}^{w}\sum_{j=A}^{T} p_{ij}\log p_{ij} - \frac{1}{x_m}\sum_{\text{all segments}}\log(p_0(s))\right]$$

We use a Markov background model estimated from all the intergenic regions of a genome. For example, the probability of generating ATGTA (supposing the three bases preceding this segment are CTT) from this background model is

$$\begin{aligned} p_0(ATGTA) = {} & p(A|\text{previous 3 bases } CTT) \times p(T|\text{previous 3 bases } TTA) \times \\ & p(G|\text{previous 3 bases } TAT) \times p(T|\text{previous 3 bases } ATG) \times \\ & p(A|\text{previous 3 bases } TGT) \end{aligned}$$

After computing the scores for all the w-mer motifs established in this step, we save the highest 10–50 "seed" candidate motifs for updating in the next step.

In the motif updating process, every retained candidate motif matrix is used to scan all the w-mers in the remaining sequences. A new w-mer is added into a candidate weight matrix if and only if the motif score of that matrix is increased. We further refine each candidate motif by re-examining all the segments that are already included in the motif matrix during the updating step. A segment is removed from the matrix if doing so increases the motif score. The aligned segments for each motif usually stabilize within ten refinement iterations. MDscan reports the highest-scoring candidate motifs as the protein–DNA interaction motif.

Acknowledgments

The authors thank the Brown lab at Stanford (especially Jason D. Lieb) and the Young lab at MIT (especially Bing Ren) for their valuable data and scientific insight. This work is supported by National Human Genome Research Institute grants R01 HGF02235 and R01 HG02518-01, and National Science Foundation grant DMS-0094613.

Received 21 August 2001; accepted 10 May 2002

1. van Helden, J., Andre, B. & Collado-Vides, J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* **281**, 827–842 (1998).
2. Bussemaker, H.J., Li, H. & Siggia, E.D. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl. Acad. Sci. USA* **97**, 10096–10100 (2000).
3. Sinha, S. & Tompa, M. A statistical method for finding transcription factor binding sites. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 344–354 (2000).
4. Vilo, J., Brazma, A., Jonassen, I., Robinson, A. & Ukkonen, E. Mining for putative regulatory elements in the yeast genome using gene expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 384–394 (2000).
5. Hertz, G.Z., Hartzell, G.W. & Stormo, G.D. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.* **6**, 81–92 (1990).
6. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).
7. Liu, J.S., Neuwald, A.F. & Lawrence, C.E. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.* **90**, 1156–1170 (1995).
8. Roth, F.P., Hughes, J.D., Estep, P.W. & Church, G.M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16**, 939–945 (1998).
9. Liu, X., Brutlag, D.L. & Liu, J.S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.* 127–138 (2001).
10. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
11. Iyer, V.R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).
12. Lieb, J.D., Liu, X., Botstein, D. & Brown, P.O. Promoter-specific binding of Rap1p revealed by genome-wide maps of protein-DNA association. *Nat. Genet.* **28**, 327–334 (2001).
13. Simon, I. *et al.* Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **106**, 697–708 (2001).
14. Dolan, J.W., Kirkman, C. & Fields, S. The yeast STE12 protein binds to the DNA sequence mediating pheromone induction. *Proc. Natl. Acad. Sci. USA* **86**, 5703–5707 (1989).
15. Graham, I.R. & Chambers, A. Use of a selection technique to identify the diversity of binding sites for the yeast RAP1 transcription factor. *Nucleic Acids Res.* **22**, 124–130 (1994).
16. Buchman, A.R., Kimmerly, W.J., Rine, J. & Kornberg, R.D. Two DNA-binding factors recognize specific sequences at silencers, upstream activating sequences, autonomously replicating sequences, and telomeres in *Saccharomyces cerevisiae. Mol. Cell Biol.* **8**, 210–225 (1988).
17. Idrissi, F.Z. & Pina, B. Functional divergence between the half-sites of the DNA-binding sequence for the yeast transcriptional regulator Rap1p. *Biochem. J.* **341**, 477–482 (1999).
18. Spellman, P.T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).
19. Lawrence, C.E. *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214 (1993).
20. Liu, J.S. *Monte Carlo Strategies in Scientific Computing* (Springer, New York, 2001).

Genome-wide internal tagging of bacterial exported proteins

Jeannie Bailey and Colin Manoil*

Published online: 1 July 2002, doi:10.1038/nbt715

As a result of the explosive growth of bacterial genomic and postgenomic information, there is a pressing need for efficient, inexpensive strategies for characterizing the *in vivo* behavior and function of newly identified gene products. We describe here an internal tagging procedure, based on transposon technology[1,2], to facilitate the analysis of membrane-bound and secreted proteins in Gram-negative bacteria. The technique is based on a broad–host range transposon (IS*phoA*/hah), which may be used to generate both alkaline phosphatase (AP) gene fusions and 63-codon in-frame insertions in the genome. The 63-codon insertion encodes an influenza hemagglutinin epitope and a hexahistidine sequence, permitting sensitive detection and metal affinity purification of tagged proteins. For each gene targeted, it is thus possible to monitor the disruption of phenotype (using the transposon insertion), the gene's transcription and translation (using the AP reporter activity), and the behavior of the unfused protein (using the internal tag). Studies on a sequence-defined collection of *Escherichia coli* strains generated using the transposon showed that the synthesis and subcellular localization of tagged proteins could be readily monitored. The use of IS*phoA*/hah should provide a cost-effective approach for genome-wide *in vivo* studies of the behavior of exported proteins in a number of bacterial species.

Transposons that generate fusions to the *E. coli* AP gene (*phoA*, have been previously shown to allow the selective identification of genes encoding exported proteins in bacteria[3,4]. However, further analysis of the proteins identified may be limited, since the hybrid polypeptides produced lack the C-terminal sequences of the target protein, and are thereby susceptible to cellular proteolysis. Also, they frequently exhibit altered subcellular localization, and usually lack detectable function of the target protein. In-frame insertions derived from *phoA* fusions are less prone to these difficulties, but are generally constructed with plasmid-borne genes using *in vitro* manipulations that are not well suited for large-scale genomic studies[5–7].

*Department of Genome Sciences, 357730, University of Washington, Seattle, WA 98195. *Corresponding author (manoil@u.washington.edu).*