doi:10.1093/nar/gkm028

Identification of putative transcriptional regulatory networks in *Entamoeba histolytica* using Bayesian inference

Jason A. Hackney[1], Gretchen M. Ehrenkaufer[1] and Upinder Singh[2,*]

[1]Department of Microbiology and Immunology, Stanford University, Stanford, California, USA and [2]Department of Internal Medicine, Division of Infectious Diseases, and Department of Microbiology and Immunology, Stanford University, Stanford, California, USA

Received December 16, 2006; Revised January 3, 2007; Accepted January 4, 2007

ABSTRACT

Few transcriptional regulatory networks have been described in non-model organisms. In *Entamoeba histolytica* seminal aspects of pathogenesis are transcriptionally controlled, however, little is known about transcriptional regulatory networks that effect gene expression in this parasite. We used expression data from two microarray experiments, *cis*-regulatory motif elucidation, and a naïve Bayesian classifier to identify genome-wide transcriptional regulatory patterns in *E. histolytica*. Our algorithm identified promoter motifs that accurately predicted the gene expression level of 68% of genes under trophozoite conditions. We identified a promoter motif ($^{A}/_{T}$AAACCCT) associated with high gene expression, which is highly enriched in promoters of ribosomal protein genes and tRNA synthetases. Additionally, we identified three promoter motifs (GAATGATG, AACTATTTAAACAT$^{C}/_{T}$C and TGAACTTATAAACATC) associated with low gene expression. The promoters of a large gene family were highly enriched for these motifs, and in these genes the presence of ⩾2 motifs predicted low baseline gene expression and transcriptional activation by heat shock. We demonstrate that amebic nuclear protein(s) bind specifically to four of the motifs identified herein. Our analysis suggests that transcriptional regulatory networks can be identified using limited expression data. Thus, this approach is applicable to the multitude of systems for which microarray and genome sequence data are emerging.

INTRODUCTION

Identification of gene regulatory networks is one promise of the post-genomic era. Identification of *cis*-regulatory elements and the patterns of gene expression they control become increasingly possible as large-scale expression studies and high-throughput genome sequencing are carried out. A number of approaches can be used to identify *cis*-regulatory elements that control expression of large numbers of genes. Bioinformatic techniques can identify putative regulatory elements that are conserved in the promoters of co-expressed genes, conserved in promoters in evolutionarily distant species, or both in concert (1–3). Computational identification of motifs can be coupled with Bayesian statistics to allow the identification of potential interactions between transcription factors. In *Saccharomyces cerevisiae* and *Caenorhabditis elegans*, computational approaches and extremely large expression data sets, comprising upwards of 255 experiments, were used to identify regulatory modules controlling gene expression during many conditions, including progression through the cell cycle, sporulation and osmotic stress, and development (4,5). Additionally these studies accurately predicted gene expression based on promoter content. Unfortunately, however, in most systems such detailed microarray data are not currently available.

Entamoeba histolytica is a protozoan parasite and the etiologic agent of amebic colitis and amebic liver abscess causing invasive disease in 50 million people annually (6). The amebic life cycle has two stages, a trophozoite form that causes invasive disease, and an encysted form that is the transmissible agent. Changes in amebic transcription underlie developmental pathways, parasite response to host stresses, drug resistance and tissue invasion (7–10). A number of recent studies have utilized microarray

[*]To whom correspondence should be addressed. Email: usingh@stanford.edu

technology to characterize the amebic transcriptional profile associated with adhesion to collagen, parasite virulence, stage conversion and tissue invasion (8,11,12,13). Although global changes in gene expression were observed, the promoter elements controlling these transcriptional changes were not identified. Identifying regulatory pathways controlling transcriptional responses is key to understanding how and why amebae cause disease.

The basal transcriptional machinery in *E. histolytica* has been well characterized, including identification of a TATA box (TATTTAAA$^{G}/_{C}$) and an Initiator (Inr) element (AAAAATTCA) (14,15). In addition to the TATA box and Inr element, a third-core promoter element, the GAAC box (A$^{A}/_{T}$GAACT), is independently able to control the rate and site of transcription initiation (16,17). The presence of a third conserved core promoter element contributes to the unusual core promoter architecture in *E. histolytica* compared to other metazoan systems. A number of other regulatory elements and transcription factors have also been identified in *E. histolytica*. The most well characterized are the transcription factors upstream regulatory element 3-binding protein (URE3-BP), a calcium-sensitive EF hand protein and the *E. histolytica* enhancer binding proteins 1 and 2 (EhEBP1 and EhEBP2) (18–20). Additionally, the recent completion of the *E. histolytica* genome sequence indicates that canonical transcription factors are encoded in the genome (21). Thus, it appears that sequence-specific DNA-binding proteins control multiple aspects of basal and activated transcription in *E. histolytica*.

Although much work has been done in *E. histolytica* to characterize the regulation of a handful of genes, global transcriptional networks have not been identified. We have applied a gene regulatory network approach towards understanding coordinate control and regulation of gene expression in this parasite. Utilizing expression data from two microarray experiments, we identified *cis*-promoter motifs that correlated with the level of gene expression. In addition, we identified a set of three promoter motifs that when present in combination of $\geqslant$2 motifs were associated with increased gene expression in response to heat shock. Furthermore, by using electrophoretic mobility shift assays (EMSAs) we confirmed that a number of the motifs predicted by bioinformatic analyses specifically bind amebic nuclear protein(s).

MATERIALS AND METHODS

Ameba strains, culture and RNA isolation

Entamoeba histolytica (HM-1:IMSS) was grown axenically in trypticase-yeast extract-iron-serum (TYI-S-33) medium as previously described (22). Parasites were subjected to heat shock by exposure to 42°C for 1 h. Viability of the heat shock treated trophozoites was determined by Trypan blue exclusion. For RNA isolation, amebae were washed once with TYI-S-33 medium to remove dead cells, chilled on ice for 10 min, centrifuged for 10 min at 430 × *g*, resuspended in TRIZOL reagent (Invitrogen), lysed by a syringe needle and RNA isolated following the manufacturer's protocol. RNA used for microarray analysis was further purified using the RNeasy kit (Qiagen), according to the manufacturer's protocol.

Microarray design, hybridization and data transformation

Expression analysis was performed using a custom *E. histolytica* array from Affymetrix, Inc. (Santa Clara, CA, USA), as previously described (8). Probes were designed according to standard Affymetrix chip design protocols (http://www.affymetrix.com/support/technical/other/custom_design_manual.pdf); up to 16 paired oligonucleotides were designed per gene. A total of 9435 of the 9938 genes predicted in the *E. histolytica* genome are represented on the microarrays. Repetitive sequences from retrotransposon elements, tRNA genes and the ribosomal RNA episomal circle were not included on the array. Probes were also designed for intergenic regions, though these probes were not considered in this analysis. Labeled cRNA for hybridization was prepared from 4 μg of total RNA according to published Affymetrix protocol (http://www.affymetrix.com/support/technical/manual/expression_manual.affx). Hybridization and scanning were performed by the Stanford PAN facility according to Affymetrix protocols (http://cmgm.stanford.edu/pan/).

Two arrays from individual mid-log cultures of *E. histolytica* (HM-1:IMSS) trophozoites and two arrays from trophozoites subjected to heat shock as described above were included in this analysis. Raw data from the microarray scanner were loaded into the GCOS software (Affymetrix, Santa Clara, CA, USA). Data were scaled to have a mean value of 500. Data from probes designed to intergenic sequence were removed. The remaining scaled data were loaded into GeneSpring (Agilent Technologies, Palo Alto, CA, USA) and normalized per chip, to give a median expression value of 1, and yielding an approximately normal distribution. Normalized data were $\log_2$ transformed, giving a median value of 0, and the two replicates for each condition (untreated trophozoites and trophozoites subjected to heat shock) were each averaged.

Databases

The complete *E. histolytica* genome sequence was obtained from The Institute for Genome Research (TIGR, http://www.tigr.org/tdb/e2k1/eha1/). The amino acid sequence, nucleotide sequence and locations of all predicted open reading frames (ORFs) were retrieved from TIGR (download date February 21, 2006). This information was used to retrieve the region from −500 to −1 relative to the predicted translation start site for each ORF. We have sequence data for promoter regions of 7638 genes that are present on the microarray.

Bioinformatics

The MEME and MAST programs were downloaded from UCSD (http://meme.sdsc.edu). The MEME motif elucidation program was run with the command line

Table 1. Promoter motifs identified in genes with very high or very low expression in trophozoites

	Motif number	Motif consensus sequence	Motif occurrences in all promoters analyzed ($n = 7638$)	Number of occurrences in promoters of very highly expressed genes ($n = 477$)	Numbers of occurrences in promoters of very low expressed genes ($n = 630$)	P-value (high or low expression versus all other promoters analyzed)
Motifs identified in promoters of genes with very high expression ($\log_2$ normalized signal $\geqslant 4$)	**M27**	**CATCTCC$^{A}/_{T}$C T$^{C}/_{G}$**	**386**	**46**	**16**	**4.0e−3**
	M29	**G$^{A}/_{T}$AAT$^{A}/_{G}$GAAGAGAT$^{A}/_{T}$$^{A}/_{C}$**	**665**	**82**	**56**	**5.6e−5**
	M30	$^{C}/_{T}$$^{A}/_{T}$GTTG$^{A}/_{T}TG^{G}/_{T}$T	728	66	39	0.18
	M31	CTN$^{C}/_{T}$TTNTG$^{T}/_{C}$T	672	57	27	0.37
	M32	**$^{A}/_{C}$AAT$^{A}/_{T}$AAACAA$^{C}/_{A}$AA$^{G}/_{C}$A**	**925**	**95**	**55**	**8.5e−3**
	M33	**CCCAA$^{C}/_{T}$T$^{T}/_{A}$$^{A}/_{C}$TTAACA**	**454**	**55**	**20**	**1.3e−3**
	M35	TA$^{T}/_{G}$TTTTCTTTTTG$^{C}/_{T}$T	763	68	33	0.22
	M37**	**$^{A}/_{T}$AAACCCT**	**841**	**153**	**15**	**0.0**
	M38	TTT$^{C}/_{G}$TACGTTC	417	47	14	0.011
	M39	CTNCA$^{G}/_{C}$$^{C}/_{T}$$^{T}/_{A}N^{T}/_{C}T^{G}/_{C}C^{C}/_{G}G^{C}/_{G}$	1034	91	51	0.22
	M40**	**AAAAAGAACT$^{A}/_{T}$AAAAA**	**1819**	**221**	**57**	**0.0**
	M41**	**$^{A}/_{T}$TGTTATATATAACA**	**701**	**88**	**24**	**1.5e−5**
	M42	**$^{G}/_{T}$ACGTGG$^{A}/_{C}$A$^{A}/_{C}$CA$^{C}/_{A}$$^{G}/_{A}$**	**269**	**47**	**10**	**3.6e−7**
	M43	GGGTTT	214	21	6	0.17
	M44	**CCACGT**	**265**	**38**	**16**	**2.5e−4**
Motifs identified in promoters of genes with very low expression ($\log_2$ normalized signal $\leqslant -4$)	**M9****	**TGAACTTATAAACATC**	**314**	**16**	**50**	**0.0**
	M13	**A$^{G}/_{T}$AGGAGAAGG**	**794**	**65**	**67**	**3.8e–3**
	M15	**$^{C}/_{T}$$^{A}/_{T}$TTTCTTT$^{G}/_{T}$C**	**758**	**45**	**76**	**8.0e–6**
	M21	**ATGATANA$^{A}/_{C}$TTGTTG$^{A}/_{T}$**	**524**	**32**	**59**	**2.2e−6**
	M23**	**AACTATTTAAACAT$^{C}/_{T}$C**	**390**	**23**	**62**	**6.0e−8**
	M24**	**GAATGATG**	**404**	**33**	**67**	**1.2e−7**
	M25	$^{C}/_{T}$$^{G}/_{T}$GCTGCTC$^{C}/_{G}$$^{C}/_{T}$$^{A}/_{T}T^{A}/_{T}$GC	788	54	43	0.81

The motif number, consensus nucleotide sequence, occurrence in all promoters analyzed, and in each gene subset (very high expression, $\log_2$ normalized signal $\geqslant 4$ and very low gene expression, $\log_2$ normalized signal $\leqslant -4$) are shown. Fifteen motifs were significantly enriched ($P < 0.01$) in promoters of genes with either very low or very high expression relative to all other promoters in the genome (shown in bold). P-values were calculated using the hypergeometric distribution. Motifs identified with (**) are discussed in detail in the article.

arguments: *–dna –mod zoops –minw 6 –maxw 16 –minsites 35 –nmotifs 30*. This identifies 30 motifs found zero or one times in each promoter, with a width between 6 and 16 nucleotides. We created a custom background Markov model file with the nucleotide frequencies in the *E. histolytica* genome. This is necessary, as the mono- and dinucleotide frequencies of non-coding sequences in the *E. histolytica* genome are highly skewed (~80% A/T content), and the Markov chain created by MEME is sensitive to the background frequency of nucleotides. We used a custom Python program to determine the correlation coefficient between each pair of motifs. Motifs identified in both expression categories that had a correlation coefficient $\geqslant 0.7$ were eliminated from the analysis, leaving a total of 22 motifs (Table 1). We used the MAST program to identify all occurrences of each motif in the promoter sequence database. Command line option used for the MAST program was: *–ev 1000*. This allows sequences with an e-value of less than 1000. The relatively high e-value for motifs is required for finding some of the shorter motifs. We again made use of the background Markov model file previously generated. Custom parsers for MEME and MAST were written in the Python programming language, and are available from the Biopython project (http://biopython.org/).

To create a hidden Markov model (HMM) of the Ehssp gene family, the amino acid sequences of the C terminal domain for 48 representative Ehssp genes (shown in bold in Supplementary Table S1) were aligned using *clustalw* (http://clustalw.genome.jp/). The HMMER package was downloaded from Washington University (http://hmmer.wustl.edu). The *clustalw* alignment was used as input to the *hmmbuild* program, for generating a profile HMM. This model was run through *hmmcalibrate*, which calibrates the search statistics for the HMM. We used the *hmmsearch* program to search all *E. histolytica* predicted ORFs for significant similarity (e-value $< 1\mathrm{e}-7$) to our HMM.

Statistical analysis

We used the hypergeometric distribution to determine the significance of enrichment for each motif identified in the two expression categories. Briefly, the number of occurrences of each motif in genes with high expression and genes with low expression were identified, as well as the number of times the motif occurred in the promoter database as a whole. This allowed us to determine the relative enrichment of each motif for the expression data set in which it was identified. To confirm the motifs we identified as being overrepresented in the appropriate expression set, we randomized the nucleotide positions of the motif and measured the overrepresentation of the shuffled motif in the appropriate expression category. Randomization was performed 1000 times, and the number of times the shuffled motif was significantly enriched, with a P-value less than or equal to the original motif, in the appropriate expression category was counted.

Custom Bayesian classifier libraries were written in Python. Bayes' theorem, shown in the formula below, gives the conditional probability distribution of a random variable H given a second random variable E, given the marginal probability distribution of H, the conditional probability of E given H and the marginal probability distribution of E.

$$P(H \mid E) = P(E \mid H)P(H)/P(E)$$

In our case, we want to know the probability of a gene being expressed given the presence of each motif in its promoter. For each motif, the likelihood of it being present in genes with high or low expression was determined.

To create our Bayesian classifier, we used the motifs we previously identified using the MEME program. The promoter database for the entire *E. histolytica* genome was then queried using the MAST program against this set of motifs. For ease of calculation, only the most 3′ occurrence of each motif in a promoter was used in our Bayesian classifier. To determine the accuracy of the Bayesian classifier, we created training and cross-validation sets by randomly removing 25% of the genes. We then trained the classifier on the training set. The classifier was then evaluated using the cross-validation set. The random partitioning was repeated 1000 times, to give an estimate of the accuracy of the Bayesian classifier. For each correct prediction by the classifier, we determined which motifs were present in the promoter. By identifying the most frequent combinations of motifs in the correct predictions, we were able to determine if there was any evidence for significant co-occurrence of the motifs.

Electrophoretic mobility shift assays (EMSAs)

Amebic nuclear proteins were obtained using the protocol described in (23). Briefly, 1×10^7 mid-to-late log-phase trophozoites were washed one time with ice-cold phosphate buffered saline (PBS) solution. Cells were harvested in ice-cold PBS and centrifuged for 5 min at 430 × *g*. Cells were then resuspended in hypotonic buffer (10 mM HEPES pH 7.9, 1.5 mM $MgCl_2$, 10 mM KCl, 6% NP-40) supplemented with protease inhibitors (500 µM AEBSF, 1 µM leupeptin, 1 µM E-64d), and incubated 20 min on ice. Nuclei were collected by centrifugation for 10 min at 1000 × *g* at 4°C, resuspended in hypertonic lysis buffer (20 mM HEPES pH 7.9, 420 mM NaCl, 1 mM EDTA, 1 mM EGTA, supplemented with the same protease inhibitor mix), and incubated on ice for 30 min. After lysis, nuclear and membrane fractions were removed by centrifugation at 18 000 × *g* for 20 min at 4°C. The soluble fraction, containing nuclear protein, was snap frozen on dry ice, and frozen at −80°C. Protein content was determined using Bradford reagent (24).

Double-stranded oligonucleotide probes (Supplementary Table S2) were designed for each motif, such that the most common nucleotide was used for each position. About 50 pmol of double-stranded oligonucleotide were labeled with α-^{32}P dATP using Klenow fragment (Invitrogen) in the supplied buffer according to manufacturer's protocol. Binding reactions occurred in binding buffer (10 mM TrisHCl pH 7.9, 50 mM NaCl, 1 mM EDTA, 3% glycerol, 1 mg/ml bovine serum albumin, 1 mg/ml salmon sperm DNA) with 5 µg of nuclear protein and 50 fmol of labeled probe. Binding occurred for 30 min at 30°C. For competition, 2- and 10-fold molar excess unlabeled oligonucleotide were added to the binding reactions prior to incubation at 30°C. Protein–DNA complexes were resolved on a 0.5 × TBE polyacrylamide gel, which was then vacuum-dried and exposed to a storage phosphor screen (Kodak). Phosphor screens were developed using software from Molecular Dynamics.

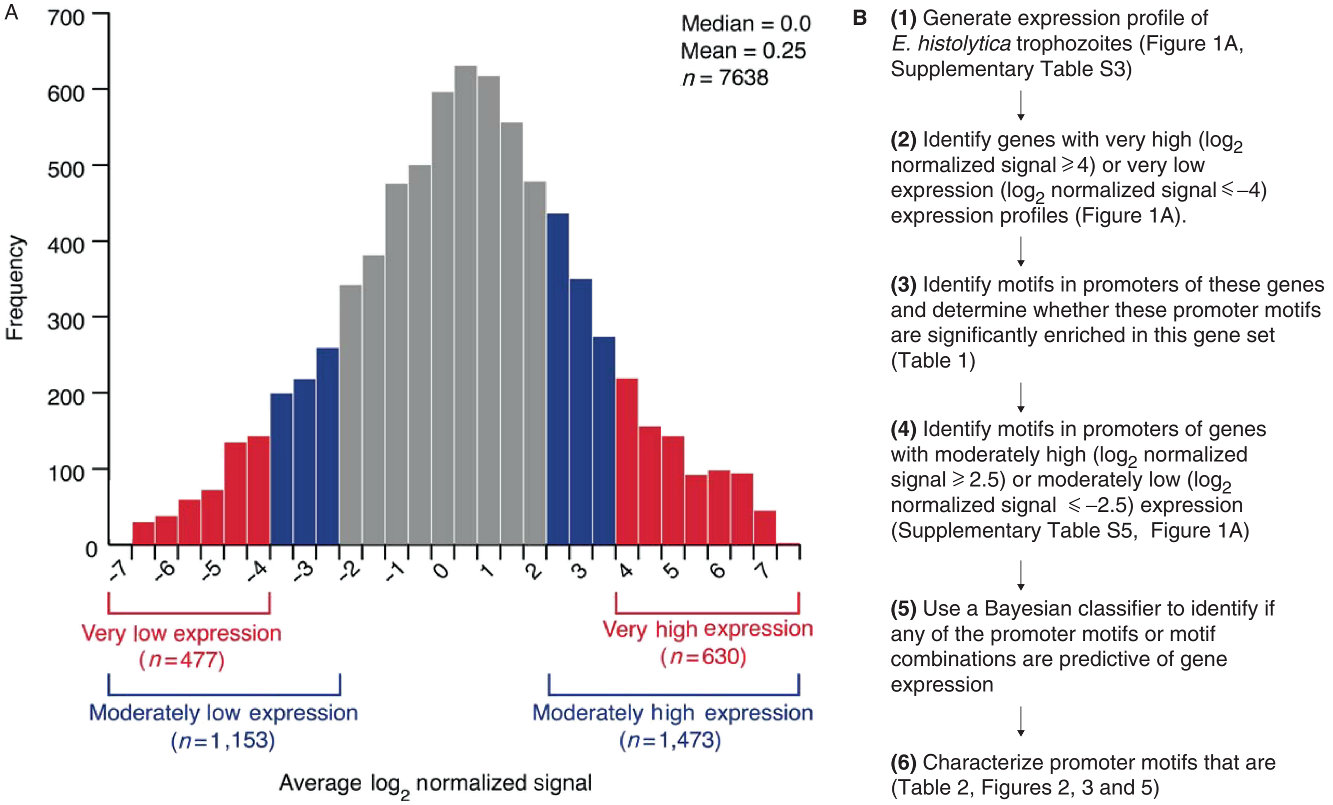
RESULTS

Expression profiling of *E. histolytica* trophozoites using a whole-genome oligonucleotide microarray

To identify the expression profile of mid-log phase *E. histolytica* (strain HM-1:IMSS) trophozoites, we used a custom short oligonucleotide microarray fabricated by Affymetrix, Inc (Santa Clara, CA, USA). These arrays were designed with probes targeting both predicted mRNAs (9435 of the 9938 predicted genes are represented on the array) and predicted intergenic sequences (8). Total RNA from two independent cultures was used to generate two expression profiles, which had a correlation coefficient of 0.96, indicating highly reproducible results. In addition, the hybridization intensity of the probes targeting coding sequences was significantly higher than those from intergenic sequences (Supplementary Table S3). Consistent with previous data using the same microarray, ~86–89% of probe sets showed hybridization above background levels and were considered 'Present' according to the Affymetrix GCOS software (Affymetrix, Inc., Santa Clara, CA, USA) (8). After removing probe sets targeting intergenic sequences and normalization, the data formed a unimodal approximately normal distribution (Figure 1A). There was excellent correlation between the microarray signal intensity and the amount of transcript as assessed by northern blot and semi-quantitative reverse transcriptase-polymerase chain reaction (RT-PCR) analysis ((8,12,25) and R. MacFarlane, G.M. Ehrenkaufer, and J.A. Hackney, unpublished data). Based on these results, we concluded that we could identify genes that differ in steady-state expression level in *E. histolytica* trophozoites.

Identification of motifs in promoters of genes with very high or very low gene expression profiles

Coordinate gene expression is often driven by conserved *cis*-promoter elements. To identify potential regulatory elements, we followed the bioinformatic procedure depicted in Figure 1B. The *E. histolytica* genome is highly compact (median intergenic size 326 bp), amebic genes have very short 5′-untranslated regions (median 21 bp), and almost all *E. histolytica* promoter elements identified to date are within 400 bp of the start codon ((14,27) and J.A. Hackney, unpublished data). The nucleotide sequences of the region from −500 to −1 relative to the predicted start codon were retrieved from the *E. histolytica* genome sequence data set and should contain the majority of promoter regulatory regions. Although this genomic region will often extend



(1) Generate expression profile of *E. histolytica* trophozoites (Figure 1A, Supplementary Table S3)

↓

(2) Identify genes with very high ($\log_2$ normalized signal ⩾ 4) or very low expression ($\log_2$ normalized signal ⩽ –4) expression profiles (Figure 1A).

↓

(3) Identify motifs in promoters of these genes and determine whether these promoter motifs are significantly enriched in this gene set (Table 1)

↓

(4) Identify motifs in promoters of genes with moderately high ($\log_2$ normalized signal ⩾ 2.5) or moderately low ($\log_2$ normalized signal ⩽ –2.5) expression (Supplementary Table S5, Figure 1A)

↓

(5) Use a Bayesian classifier to identify if any of the promoter motifs or motif combinations are predictive of gene expression

↓

(6) Characterize promoter motifs that are (Table 2, Figures 2, 3 and 5)

Figure 1. (**A**) Histogram of expression data from *E. histolytica* trophozoites. The average $\log_2$-transformed normalized signal for the two microarrays from *E. histolytica* trophozoites was used to generate a histogram of expression values. The expression categories used in this article are indicated on the histogram. (**B**) Schematic of bioinformatic analyses utilized in the study. To identify promoter motifs predictive of gene expression, we used the procedure outlined here. After determining the expression profile of *E. histolytica* trophozoites (**1**), we retrieved the promoter regions from all the genes represented on the microarray. The genes with highest ($\log_2$ normalized signal ⩾4) and lowest ($\log_2$ normalized signal ⩽−4) expression were identified (**2**). Significant promoter motifs were identified in the promoters of each of these sets of genes (**3**). Occurrences of these motifs were identified in genes with more moderate expression values ($\log_2$ normalized signal ⩾2.5 or ⩽−2.5) (**4**). We then created a Bayesian classifier using the most 3′ occurrence of each motif as a predictive variable. The classifier was trained using a subset of the microarray data and tested using a cross-validation set (**5**). The motifs most useful in predicting gene expression, either singly or in groups, were identified, and genes potentially controlled by these elements were characterized (**6**).

into the adjacent gene, we do not believe this should pose a substantial problem with our algorithm, as elements identified in neighboring genes will not likely be significantly enriched in either expression category. Our analysis will not identify motifs that are present >500 bp upstream from the start codon or those present downstream of the stop codon (in the 3′-untranslated region or adjacent genomic regions). In addition, as we are only examining steady-state levels of gene expression, our approach cannot address differences in mRNA stability or other post-transcriptional regulatory processes. Promoter regions that were >98% identical to another promoter and those for which <500 bp of sequence was available (at the end of a contig for example) were not analyzed. Following these criteria, a total of 7638 unique promoter sequences were obtained.

In order to identify motifs which correlated with gene expression level in *E. histolytica* trophozoites, we used the MEME program (27) to identify conserved motifs in the promoter regions of genes with very high (defined as a $\log_2$ normalized signal ⩾4; $n = 630$) and very low (defined as $\log_2$ normalized signal ⩽−4; $n = 477$) expression profiles (Figure 1A). Thirty motifs were identified in each set of promoters, with several motifs identified in both sets. We removed motifs identified in the promoters of both gene sets that had a correlation coefficient of ⩾0.7, leaving a total of 22 motifs (Table 1). We then used the MAST program (28) to identify all occurrences of these 22 motifs in the promoters of all *E. histolytica* genes for which we have promoter sequences. Since we hypothesized that motifs found in each of these sets would be predictive of gene expression level, we would expect, for example, that motifs identified in promoters of highly expressed genes would be enriched in the promoters of this set of genes. Using the hypergeometric distribution, we found that 15 of the 22 promoter motifs were significantly over-represented in the promoters of genes of the appropriate expression data set ($P < 0.01$), compared to the rest of genes for which we have promoter sequences (Table 1). That is, the promoter motif was more frequently found in the appropriate data set than would be expected by random chance. However, the relative enrichment of a motif in a given gene expression data set was lower than data reported elsewhere (which is often as much as 100-fold enrichment (1)) most likely due to our analysis of only one expression condition instead of the complex data sets typically analyzed. Of the fifteen motifs identified as being significantly enriched in the gene expression data set of interest, seven were biased to the 3′ end of promoter regions (towards the start codon of the gene), seven were equally distributed along the 500 bp promoter region, and only one was biased towards the 5′ end of the putative promoter region (Supplementary Figure S1). This suggests the motifs we identified likely represent motifs in the promoter regions of genes of interest, and not 5′ or 3′ elements from the neighboring gene. To determine whether the motifs were overrepresented because of the motif sequence, or merely because of nucleotide bias, we randomly permuted each motif 1000 times and determined whether the shuffled motif showed similar enrichment in the appropriate expression category (Supplementary Table S4). In many cases, no significant motifs were generated by the randomization. In a few cases, however, we did find randomly shuffled motifs that had *P*-values equal to or lower than the original motifs we identified, likely due to the low nucleotide variability of a motif (e.g. the $^{A}/_{T}$AAACCCT motif has low nucleotide variability and thus at least some of the 1000 randomly shuffled motifs are likely to substantially resemble the original motif).

Use of a Bayesian classifier to predict gene expression in *E. histolytica*

We wished to further characterize the motifs described above in order to identify potential cooperative action. For this analysis, we used a Bayesian classifier to identify groups of motifs that were significantly predictive of gene expression levels. The concept of Bayesian statistics is that previously observed frequencies of event occurrences give an indication of the likelihood of those events occurring again. In our case, we are trying to determine the likelihood that a given gene will have a specific expression level based on whether or not it has one or more motifs in its promoter region. Bayesian statistics have been previously coupled with identification of promoter elements to identify patterns of gene regulation dependent on multiple transcription factor binding sites (4,5,29). Because transcription factors often coordinately influence gene expression, this type of analysis is better suited to identification of combinations of motifs than identification of the single motif alone. We hypothesized that we could use the motifs identified in promoters of genes with very high and very low gene expression levels to classify a larger set of genes with more moderate gene expression levels (Figure 1A and B). As expected, many of the motifs that were significantly enriched in promoters of genes with very high gene expression were also enriched in the promoters of more moderately expressed genes (defined as $\log_2$ normalized expression value $\geqslant 2.5$; 1473 genes, 19% of genes for which we have promoter data; Supplementary Table S5). Likewise, most of the motifs significantly enriched in promoters of genes with very low gene expression were also enriched in the promoters of genes with low gene expression (defined as $\log_2$ normalized expression value of $\leqslant -2.5$; 1153 genes, 15% of genes for which we have promoter data) (Supplementary Table S5). We can make no predictions about genes with median gene expression levels ($\log_2$ normalized expression values between -2.5 and 2.5) as the motifs we previously identified should only be indicative of low or high gene expression; thus these genes were not considered. We also removed genes whose promoters contain none of the motifs, as we cannot make any predictions about their expression levels. A final data set of 1584 genes was used in the Bayesian analysis. To assess the accuracy of our Bayesian classifier, we used a cross-validation strategy, randomly removing one quarter (396) of genes for later testing. We then trained the classifier on the remaining data set of 1188 genes, which includes genes with $\log_2$ normalized expression values $\geqslant 2.5$ or $\leqslant -2.5$ and at least one previously identified motif in their upstream regions. After training the Bayesian classifier, the expression levels of 396 genes in the cross-validation set were predicted using the motifs present in their promoters. We repeated this test 1000 times, each time removing a random set of 396 genes and training the classifier on the remaining 1188 genes. Overall, our prediction accuracy was $68 \pm 2\%$. The *P*-value for correctly predicting this fraction of genes is 7.4×10^{-7}, according to the binomial distribution. Highly expressed genes were more likely to be correctly predicted at $72 \pm 2\%$, while low expressed genes were correctly predicted at $62 \pm 2\%$. The difference in predictive power between the two expression categories is likely due to the greater number of genes with high gene expression in our data set. As Bayes' theorem incorporates the background frequency of each class occurring and weights the predictions accordingly, our predictions are necessarily weighted toward the genes with high expression, which are the more common set. We believe our results compare favorably with other reports using Bayesian statistics to infer regulatory patterns considering the relatively small size of our gene expression data set (2 arrays versus ~255 arrays used in other studies) (4,5).

Motifs predictive of high gene expression in trophozoites

We wished to further characterize the promoter motifs that were used to correctly predict gene expression profiles in *E. histolytica*. For this analysis, we identified all occurrences of a given motif in the promoters of all genes for which we have sequence data, regardless of expression level. We identified a single motif (consensus sequence $^{A}/_{T}$AAACCCT) that was predictive of high gene expression in trophozoites (282 of 1473 highly expressed genes have at least one copy of this motif in their promoters, $P < 0.01$) (Table 2). This motif is very similar to an enhancer element in *Schizosaccharomyces pombe* (AAACCCT), which is found upstream of all *S. pombe* histone genes (30,31). In *S. pombe* this element is in direct repeats, whereas in *E. histolytica* gene promoters it is generally present in only a single copy. Although we did find this motif in the promoters of 4 of the 13 histone genes in *E. histolytica*, the prevalence was not statistically significant. However, in *E. histolytica*, this motif was highly enriched in the promoters of ribosomal proteins (72 of 134, $P < 0.01$), and tRNA synthetase genes

Table 2. A motif highly enriched in *E. histolytica* genes with high expression levels

	Total number of genes in data set	Number of genes with $^{A}/_{T}$AAACCCT in promoter region	*P*-value (overrepresentation of $^{A}/_{T}$AAACCCT versus all promoters analyzed)
All promoters analyzed	7638	841	NA
Highly expressed genes ($\log_2$ normalized signal $\geqslant 2.5$)	1473	282	3.0e−7
Low expressed genes ($\log_2$ normalized signal $\leqslant 2.5$)	1153	68	1
Histone genes	13	4	0.02
tRNA synthetase genes	28	15	7.0e−7
Ribosomal protein genes	134	72	1.2e−7

We identified occurrences of the $^{A}/_{T}$AAACCCT motif in the promoters of the 7638 genes on the microarray for which we have promoter sequence. The $^{A}/_{T}$AAACCCT motif is over-represented in the promoters of genes with moderately high expression ($\log_2$ normalized signal $\geqslant 2.5$), but not in promoters of genes with moderately low expression ($\log_2$ normalized signal $\leqslant -2.5$). This promoter motif is also highly over-represented in the promoters of ribosomal proteins and tRNA synthetases, each of which show high expression in log-phase trophozoites. We did not find significant enrichment of this motif in *E. histolytica* histone promoters. The *P*-values were calculated using the hypergeometric distribution. $P < 0.01$ considered statistically significant.

(15 of 28, $P < 0.01$). Importantly, the occurrences of this motif were not restricted to a single type of tRNA synthetase or ribosomal protein subunit, suggesting the presence of the motif in these promoters is not simply through evolutionary conservation of duplicated promoter sequences (Supplementary Table S6).

Amebic nuclear proteins bind in a sequence-specific manner to motifs M37, M40 and M41

In order to confirm that the motifs identified on the basis of bioinformatic analyses bind amebic nuclear proteins in a sequence-specific manner, we used EMSAs. We chose motifs such as M37, M40 and M41 that had significant enrichment within the appropriate expression category, were biased toward the 3′ end of the promoter, and had strong conservation at each position in the motif. We identified that *E. histolytica* trophozoite nuclear extract bound motifs M37 and M40, with a single band identified in the binding reaction, which was competed by 2- or 10-fold molar excess of cold oligonucleotide, but not by a similar excess of shuffled cold oligonucleotide designed to retain similar nucleotide content as the original motif (Figure 2A and B; Supplementary Table S2). For motif M41, three bands were observed upon interaction of the oligonucleotide containing the motif and amebic nuclear extract, however only one band (marked with an arrow) appears to be binding in a sequence-specific manner (Figure 2C). This confirms that our computational approach successfully identified promoter motifs that are recognized by specific DNA-binding proteins in *E. histolytica*.

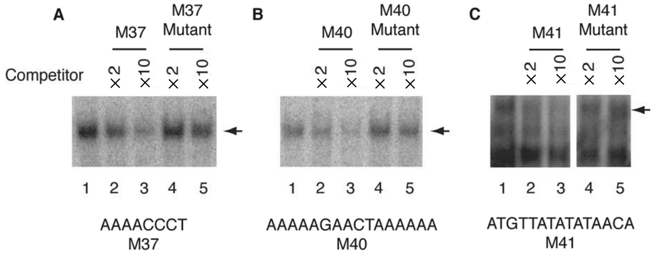


Figure 2. Sequence-specific binding of *E. histolytica* nuclear proteins to M37, M40 and M41 motifs identified in promoters of genes with high expression levels. (**A**–**C**) EMSAs with an oligonucleotide conforming to the M37, M40 or M41 consensus sequence were performed using nuclear extracts prepared from mid-to-late log-phase trophozoites. Lanes for each panel: (1) nuclear extract plus labeled oligonucleotide without competitor, (2) 2-fold molar excess unlabeled oligonucleotide, (3) 10-fold molar excess unlabeled oligonucleotide, (4) 2-fold molar excess unlabeled mutant oligonucleotide, (5) 10-fold molar excess unlabeled mutant oligonucleotide. Arrows indicate DNA–protein complexes we believe to be specific.

Motifs correlated with low gene expression in trophozoites

We also identified a set of three motifs (M24, M23 and M9) that were overrepresented in promoters of genes with very low expression (Table 1, Figure 3A). Examination of the motif combinations most frequently found by our Bayesian classifier indicated that these motifs were likely to show significant co-occurrence in promoters of genes with low expression. Thus, we analyzed gene expression of all genes that have these motifs (either single or multiple motif combinations) in their promoter regions. We found that genes whose promoters contained only one of the three motifs, but not the other two, had expression levels close to median (Figure 3B). However, genes whose promoters contained $\geqslant 2$ motifs had extremely low gene expression signal. Additionally, we found a strong propensity toward co-occurrence of these three motifs in the promoter of a given gene ($P < 0.01$, Supplementary Figure S2A). Genes that had $\geqslant 2$ of these motifs were also more likely to be in the set of genes with moderately low gene expression than genes with only a single motif.

The motif M24 (GAATGATG) is novel, not previously described in *E. histolytica* or other organisms. The motif M23 (AACTATTTAAACAT$^{C}/_{T}$C) is very similar to the amebic TATA box, but has additional flanking sequences. The M9 motif (TGAACTTATAAACATC) is composed of two motifs: M9A (TGAACT), similar to the amebic GAAC element and M9B, a novel motif (TATAAACATC) (14). The two parts of M9 appear to be inseparable in the context of the other two motifs, M23 and M24. M9B by itself is not significantly co-associated with M24 or M23 and also does not predict low gene expression levels (Supplementary Figure S2B).

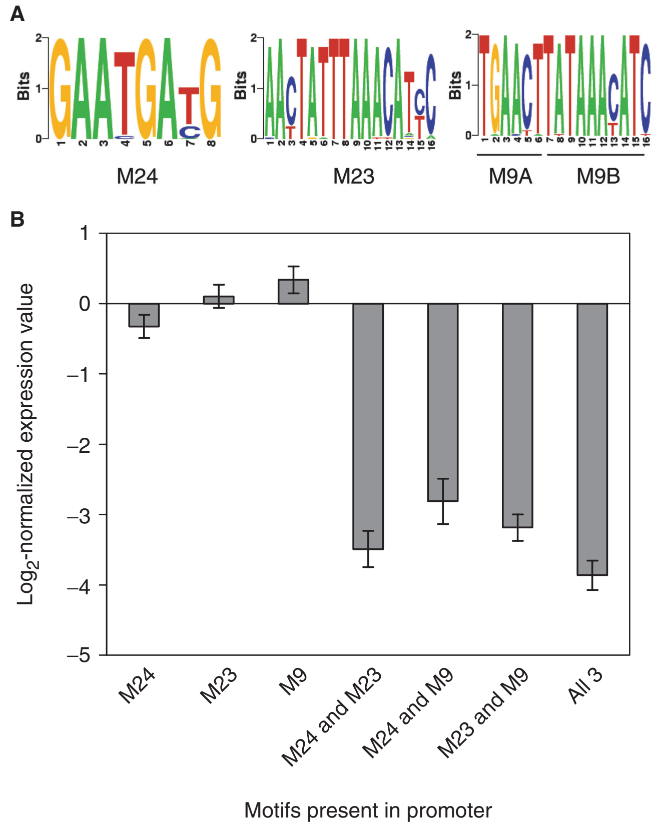


Figure 3. A combination of three promoter motifs that predict low gene expression. (**A**) Three promoter motifs, M24 (GAATGATG), M23 (AACTATTTAAACAT^C/$_T$C) and M9 (TGAACTTATAAACATC) are predictive of low gene expression in *E. histolytica* trophozoites. Motif M9 is comprised of two separate motifs: M9A, TGAACT and M9B, TATAACATC. Motifs are depicted as sequence logos (http://weblogo.berkeley.edu), in which the height of each nucleotide represents the information content of that base at that position. (**B**) The average expression of all genes from the microarray data with each of the single promoter motifs and combinations of motifs are shown. The average expression of genes with any of the individual motifs was close to median, while genes with any combination of two or all three motifs had a strong negative bias in gene expression level.

The majority of genes that have M9, M23 and M24 promoter motifs belong to the Ehssp gene family

A total of 58 genes have all three motifs (M9, M23 and M24) in their promoter regions (Supplementary Figure S3A) and 182 genes contain at least two of the three motifs (Supplementary Figure S3B). The majority of genes that contain ⩾2 of these motifs are annotated as hypothetical proteins by The Institute for Genomic Research, but a number of those have significant BLASTP matches to a polymorphic, charged antigen, *E. histolytica* stress-sensitive protein 1 (Ehssp1) (32). Three subfamilies of the Ehssp gene family were previously identified, varying in domain composition, primarily of the charged medial domain (32). We created an HMM of the C-terminal domain shared by all three subfamilies. Searching the *E. histolytica* genome sequence using this HMM, we identified 253 members of the Ehssp gene family (Supplementary Table S1).

In the Ehssp gene family, members that have ⩾2 of the M9, M23 and M24 promoter motifs, the motif position and organization were highly conserved (Supplementary Figure S3C). This stereotypical spacing was retained regardless of the relative divergence of the rest of the promoter sequences for each given gene (data not shown). However, Ehssp family genes that do not retain ⩾2 motifs appear to be more divergent from each other than those that have retained the motifs (data not shown). This suggests that either the Ehssp genes that have ⩾2 motifs are more recent gene duplication events than the Ehssp genes with ⩽1 motif, or that Ehssp genes with ⩾2 motifs were subjected to stronger evolutionary pressure than Ehssp genes with ⩽1 motif. At present, we cannot distinguish between these two possibilities.

There are 43 genes that contain ⩾2 of the promoter motifs M9, M23 and M24 but which do not belong to the Ehssp gene family (Supplementary Table S7). Unlike the genes in the Ehssp gene family, these genes do not tend to have low expression in trophozoites (median $\log_2$ normalized signal 0.0 ± 0.70). There were no apparent functional categories enriched within this set of genes (data not shown). The spacing of motifs M9, M23 and M24 in these promoters differs from the Ehssp genes: while there is a trend toward motifs M9 and M23 being found at the 3′ end of the promoter, the distribution throughout the promoters is more broad than in the Ehssp genes (Supplementary Figure S3D), and the relative ordering of the motifs is not as conserved as in the Ehssp genes (data not shown).

Ehssp-family genes with ⩾2 of the motifs M9, M23 and M24 are transcriptionally up-regulated during heat shock

When the Ehssp gene family was first identified, it was shown that some members of this gene family were responsive to heat and oxidative stresses (32). However, the authors did not determine how many members of the Ehssp gene family were responsive to stress, or if there were other members of the Ehssp gene family that were not stress responsive. To determine this, we subjected two independent cultures of *E. histolytica* (HM-1:IMSS) trophozoites to heat shock for 1 h at 42°C on different days, isolated RNA and hybridized it to two Affymetrix microarrays. The results from the two microarrays were comparable to each other (correlation coefficient = 0.9) and the data gave a unimodal approximately normal distribution (Supplementary Figure S4). Our data were comparable to previous analyses of the heat shock response in *E. histolytica* and other systems, with ∼10% of the genes on the microarray up-regulated by ⩾2-fold (33,34). Additionally, many genes previously identified as up-regulated under heat shock (including the heat shock proteins and Ehssp genes) were identified as such in our analysis (21,32,35). A scatter plot showing microarray expression data for all Ehssp genes (those with ⩾2 motifs and ⩽1 motif) and non-Ehssp genes with ⩾2 motifs under standard culture and heat shock conditions is depicted in Figure 4. The Ehssp genes with ⩾2 motifs were significantly more likely to be up-regulated and to a much higher degree by heat shock than either the Ehssp

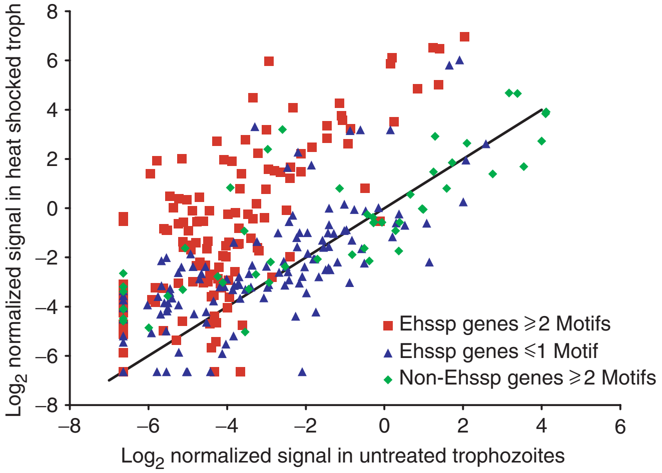


Figure 4. Presence of the M9, M23 and M24 motifs in the Ehssp gene family promoters is predictive of strong up-regulation under heat shock conditions. We performed microarray analysis to determine the heat shock responsiveness of the Ehssp gene family and non-Ehssp genes that have the M9, M23 and M24 motifs. Averaged $\log_2$ transformed normalized signal for each Ehssp gene in untreated trophozoites is plotted against the averaged $\log_2$ transformed normalized signal in heat-shock-treated trophozoites. Ehssp family genes with ≥2 motifs are shown as red squares, while Ehssp genes with ≤1 motif are shown as blue triangles. Non-Ehssp family genes with ≥2 motifs are shown as green diamonds. The average fold increase in heat shock conditions compared to untreated trophozoites for the Ehssp genes with ≥2 motifs was significantly higher than in the Ehssp genes with ≤1 motif, or the non-Ehssp family genes with ≥2 motifs. Note that the graph is plotted on a $\log_2$–$\log_2$ scale. The black line indicates equivalent signal between untreated trophozoites and heat-shock-treated trophozoites.

genes with ⩽1 motif or the non-Ehssp family genes with ⩾2 motifs (Figure 4 and Supplementary Table S7). Overall, the Ehssp family genes with ⩾2 motifs had a 25.5 (±4.73)-fold increase in expression (± standard error) after heat shock, while the Ehssp family genes with ⩽1 motif had a 3.77 (±0.82)-fold increase in expression (± standard error) after heat shock ($P = 1.2 \times 10^{-5}$ compared to the Ehssp genes with ⩾2 motifs, Supplementary Table S7). The non-Ehssp family genes with ⩾2 motifs had a 4.84 (±1.43)-fold increase in expression (± standard error) after heat shock ($P = 5.1 \times 10^{-5}$ compared to Ehssp genes with ⩾ 2 motifs, $P = 0.51$ compared to Ehssp genes with ⩽1 motif; Supplementary Table S7). Although some probes for the Ehssp family and non-Ehssp family genes with ⩽1 motif are up-regulated under heat shock conditions, they are a significant minority of the data set. Whether this represents signal cross-hybridization between members of this gene family or that other motif(s) are responsible for the up-regulation of these genes is not clear at present.

M9-specific DNA-binding protein(s) in *E. histolytica* nuclear extract from heat-shock-treated parasites

The transcriptional up-regulation of the Ehssp genes with ⩾2 motifs M9, M23 and M24 under heat shock conditions could be due to a number of scenarios. One possibility is that amebic nuclear proteins that function as repressors bind to the motifs under standard culture conditions, but do not bind under heat shock conditions. Thus a loss of repression and resultant induction of gene expression would occur under heat shock conditions. Alternatively, activators may bind to the motifs only under heat shock conditions, facilitating induction of gene expression. We decided to identify whether specific nuclear proteins bind to the M9 motif differentially under heat shock conditions to determine which of these models was correct. We chose the M9 motif because it (i) is highly overrepresented in heat-shock-responsive Ehssp genes, (ii) is novel and (iii) has a highly conserved distribution in the Ehssp gene promoters. We performed EMSA analysis with the M9 motif, both with *E. histolytica* nuclear extract from untreated trophozoites, and with nuclear extract from parasites exposed to heat shock at 42°C for 1 h (Figure 5). A faint band was seen with the M9 motif and standard nuclear extract, but that band does not appear to represent a specific interaction. In contrast, under heat shock conditions, four bands were identified by EMSA analysis, two of which appear to be specific interactions to M9. Both 2- and 10-fold molar excess unlabeled oligonucleotide with M9 substantially competed the binding interaction, whereas competition with M9A mutant and M9B mutant were equally ineffective at competing the two specific bands, although two other bands were substantially competed (Figure 5B). This suggests that some amebic nuclear proteins bind M9 in a much stronger complex when bound to both sites of the motif, than to either part alone. Overall, the data suggest that protein complex(es) bind to M9 specifically under heat shock conditions and are likely involved in transcriptional activation. Further kinetic studies will be needed to determine the exact nature of the binding between the protein(s) and M9 motif and the potential roles of the M23 and M24 motifs in this binding and gene activation.

DISCUSSION

We have identified patterns of promoter motifs in genes with similar expression levels in the protozoan parasite *E. histolytica* using Bayesian algorithms. Despite the small size of our microarray data set, we identified a number of interesting motifs that were predictive of gene expression in *E. histolytica* trophozoites. Additionally, we identified a set of three motifs, which in combination of ⩾2 motifs predicted low baseline gene expression and transcriptional up-regulation under heat shock. Finally, we have validated by EMSA analysis that a number of motifs we identified represent specific binding sites for amebic nuclear protein(s).

Microarray expression analysis is a powerful method to characterize transcriptional changes that regulate multiple aspects of pathogenesis. The large data sets produced by such experiments create the opportunity to find global regulatory patterns. Previous studies in *S. cerevisiae* and *C. elegans* generated statistical models of gene expression, predicted gene expression levels based on promoter motifs and identified regulatory modules

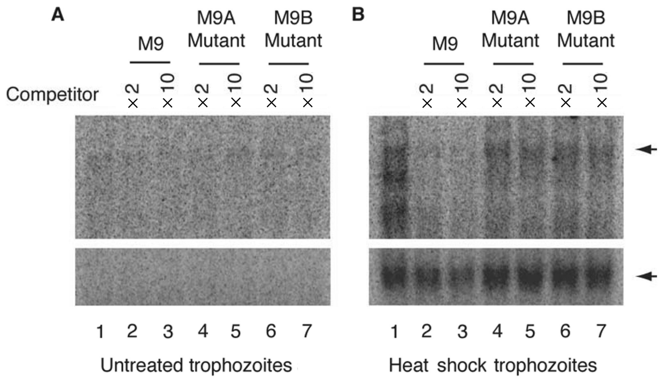


Figure 5. An amebic nuclear protein(s) binds in a sequence-specific manner to M9 under heat shock conditions, but not under standard culture conditions. (**A**) EMSAs with the M9 oligonucleotide (TGAACTTATAAACATC) were performed using nuclear extracts from trophozoites under standard culture conditions. No significant binding proteins were identified. Lanes: (1) nuclear extract plus labeled oligonucleotide without competitor, (2) 2-fold excess specific competitor, (3) 10-fold excess specific competitor, (4) 2-fold excess competitor with M9A mutated, (5) 10-fold excess competitor with M9A mutated, (6) 2-fold excess with M9B mutated and (7) 10-fold excess competitor with M9B mutated. (**B**) EMSAs with the M9 oligonucleotide were performed using nuclear extract from trophozoites subjected to heat shock. Several bands were found in the lane with nuclear extract alone, all of which were competed using the specific oligonucleotide competitor, but two of which were not competed with mutated oligonucleotides. Lanes: (1) nuclear extract plus labeled oligonucleotide without competitor, (2) 2-fold excess specific competitor, (3) 10-fold excess specific competitor, (4) 2-fold excess competitor with M9A mutated, (5) 10-fold excess competitor with M9A mutated, (6) 2-fold excess with M9B mutated and (7) 10-fold excess competitor with M9B mutated. Arrows indicate complexes we believe to be specific. Top and bottom panels were separated to allow clearer visualization of the different bands.

controlling gene expression during the cell cycle, sporulation and osmotic stress, and developmental changes (4,5). These studies relied on exquisitely detailed expression data from multiple conditions and time points of interest with data from up to 255 microarrays in *S. cerevisiae*. In these situations, the vast amount of transcriptional profiling proved invaluable. However, such detailed information is unlikely to be available for most other systems of interest.

We have shown that substantially less expression data can still be successfully applied to the identification of global transcriptional networks. Using data from two microarrays hybridized with RNA from log-phase *E. histolytica* trophozoites we were able to identify promoter motifs that correlate with gene expression level. Although none of these motifs were restricted to promoters of genes within a given expression profile, several motifs were enriched in the promoters of genes with similar functions or gene family members, likely indicating biologically significant enrichment. Three of these motifs bound to specific protein(s) in nuclear extracts from *E. histolytica* trophozoites. The promoter motifs identified by this method should not be core promoter elements as these would be found in all genes, and several of the motifs identified are enriched in groups of co-regulated genes (the ribosomal protein genes, tRNA synthetases and the Ehssp gene family). Thus we would expect transcriptional regulators identified by this approach to be either general transcription factors, or transcriptional enhancers or repressors. One interesting motif identified in this analysis ($^{A}/_{T}$AAACCCT) strongly correlated with high gene expression under trophozoite conditions. In *E. histolytica* this motif is highly enriched in the promoters of several groups of genes, including ribosomal proteins and tRNA synthetases. Ribosomal protein genes often show co-regulated expression, as has been extensively detailed in *S. cerevisiae* (36). Additionally, two promoter motifs were identified in 95% of ribosomal protein genes in *Toxoplasma gondii*, though neither of these motifs is similar to the $^{A}/_{T}$AAACCCT motif identified in *E. histolytica* (37). An analogous promoter motif (AAACCCT) has been described in *S. pombe*, where it is present in tandem repeats and is found in the promoter of all histone genes (30).

We identified three promoter motifs (M24, GAATGAGT; M23, AACTATTTAAACAT$^{C}/_{T}$C; and M9, TGAACTTATAAACATC) that in any combination of $\geqslant$2 motifs were highly predictive of low baseline gene expression in *E. histolytica* trophozoites. The M9, M23 and M24 motifs are overrepresented in the promoters of a large gene family homologous to Ehssp1, a stress-sensitive antigen (32). A majority (55%) of the 253 predicted genes in this family contain $\geqslant$2 of these promoter motifs. In promoters of the Ehssp gene family that contain $\geqslant$2 motifs, the position and spacing of these motifs is conserved. Interestingly, it appears that the Ehssp genes with $\geqslant$2 motifs appear to be more similar to each other than Ehssp genes with $\leqslant$1 motif, suggesting that they either have a more recent evolutionary origin, or that the Ehssp genes with $\leqslant$1 motif may no longer be subjected to the same degree of evolutionary pressure as the Ehssp genes with $\geqslant$2 motifs.

Ehssp family genes with $\geqslant$2 motifs were highly up-regulated by heat shock. In contrast, the Ehssp family genes with $\leqslant$1 motif and the non-Ehssp family members with $\geqslant$2 motifs showed more modest up-regulation of gene expression after heat shock. We have demonstrated that the M9 motif binds amebic nuclear protein(s) in a specific manner in nuclear extracts prepared from heat-shocked trophozoites, whereas no substantial specific interaction was identified with nuclear extracts prepared from untreated trophozoites. Our model is that transcriptional activator(s) bind to the M9 motif under heat shock conditions, up-regulating transcription of the Ehssp genes. How binding of the potential activator(s) relates to the presence of other motif(s), and to the spacing between motif M9 and the transcription start site are yet to be determined.

Few complex transcriptional regulatory networks have been identified to date in other parasitic systems. Recent bioinformatic analysis of *P. falciparum* promoters relied on large-scale expression data combined with evolutionary conservation to identify 12 putative regulatory elements, two of which had been previously described (2). This work also identified an overrepresented group of promoter motifs, many of which had potentially opposing activities.

The most complex transcriptional network characterized in a protozoan parasite to date was found in analysis of promoters from heat shock protein (hsp) genes in *P. falciparum* (3). This work identified a novel G-box motif that was conserved in the promoters of the related *Plasmodium* species, *P. yoelii*, *P. berghei* and *P. vivax*. Further *in vitro* analysis identified a second control element within the *hsp86* promoter.

We have demonstrated a method for identification of promoter motifs in *E. histolytica* promoters relying upon small expression data sets. As more amebic expression profiles are determined, this analysis can be extended to identify novel genetic regulatory pathways involved in pathogenesis and developmental control. This work represents a major step forward in identifying transcriptional regulatory networks in protozoan parasites and indicates that this statistical method can be broadly applied in other organisms.

SUPPLEMENTARY DATA

Supplementary Data is available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank all the members of the Singh lab, Swathi Rao, Jon Boyle and Jeroen Saeij for their comments and helpful discussion. J.A.H. was supported by NIH grants T32-AI07502 and T32-AI07328. G.M.E. was supported by NIH grants T32-AI07502 and AI-068899. U.S. was supported in part by NIH grants AI-53724 and AI-37941. U.S. is a member of the Stanford University Digestive Disease Center funded by the NIH grant, P30 DK56339. The authors have no competing interests. Funding to pay the Open Access publication charge was provided by NIAID.

Conflict of interest statement. None declared.

REFERENCES

1. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.
2. van Noort,V. and Huynen,M.A. (2006) Combinatorial gene regulation in *Plasmodium falciparum*. *Trends Genet.*, **22**, 73–78.
3. Militello,K.T., Dodge,M., Bethke,L. and Wirth,D.F. (2004) Identification of regulatory elements in the *Plasmodium falciparum* genome. *Mol. Biochem. Parasitol.*, **134**, 75–88.
4. Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.
5. Segal,E., Shapira,M., Regev,A., Pe'er,D., Botstein,D., Koller,D. and Friedman,N. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.
6. WHO (1997) Amoebiasis. *Wkly. Epidemiol. Rec.*, **72**, 97–99.
7. Sanchez,L., Enea,V. and Eichinger,D. (1994) Identification of a developmentally regulated transcript expressed during encystation of *Entamoeba invadens*. *Mol. Biochem. Parasitol.*, **67**, 125–135.
8. Gilchrist,C.A., Houpt,E., Trapaidze,N., Fei,Z., Crasta,O., Asgharpour,A., Evans,C., Martino-Catt,S., Baba,D.J. *et al.* (2006) Impact of intestinal colonization and invasion on the *Entamoeba histolytica* transcriptome. *Mol. Biochem. Parasitol.*, **147**, 163–176.
9. Akbar,M.A., Chatterjee,N.S., Sen,P., Debnath,A., Pal,A., Bera,T. and Das,P. (2004) Genes induced by a high-oxygen environment in *Entamoeba histolytica*. *Mol. Biochem. Parasitol.*, **133**, 187–196.
10. Marchat,L.A., Gomez,C., Perez,D.G., Paz,F., Mendoza,L. and Orozco,E. (2002) Two CCAAT/enhancer binding protein sites are cis-activator elements of the *Entamoeba histolytica* EhPgp1 (mdr-like) gene expression. *Cell Microbiol.*, **4**, 725–737.
11. Debnath,A., Das,P., Sajid,M. and McKerrow,J.H. (2004) Identification of genomic responses to collagen binding by trophozoites of *Entamoeba histolytica*. *J. Infect. Dis.*, **190**, 448–457.
12. MacFarlane,R.C. and Singh,U. (2006) Identification of differentially expressed genes in virulent and nonvirulent *Entamoeba* species: potential implications for amebic pathogenesis. *Infect. Immun.*, **74**, 340–351.
13. Ehrenkaufer,G.M., Haque,R., Hackney,J.A., Eichinger,D.M., and Singh, U. (2007) Identfication of developmentally regulated genes in Entamoeba histolytica: insights into mechanisms of stage conversion in a protozoan parasite. *Cellular Microbiology*, doi:10.1111/j.1462-5822.2006.00882.
14. Singh,U., Rogers,J.B., Mann,B.J. and Petri,W.A. Jr (1997) Transcription initiation is controlled by three core promoter elements in the hgl5 gene of the protozoan parasite *Entamoeba histolytica*. *Proc. Natl. Acad. Sci. USA.*, **94**, 8812–8817.
15. Purdy,J.E., Pho,L.T., Mann,B.J. and Petri,W.A. Jr (1996) Upstream regulatory elements controlling expression of the *Entamoeba histolytica* lectin. *Mol. Biochem. Parasitol.*, **78**, 91–103.
16. Singh,U., Gilchrist,C.A., Schaenman,J.M., Rogers,J.B., Hockensmith,J.W., Mann,B.J. and Petri,W.A. (2002) Context-dependent roles of the *Entamoeba histolytica* core promoter element GAAC in transcriptional activation and protein complex assembly. *Mol. Biochem. Parasitol.*, **120**, 107–116.
17. Singh,U. and Rogers,J.B. (1998) The novel core promoter element GAAC in the hgl5 gene of *Entamoeba histolytica* is able to direct a transcription start site independent of TATA or initiator regions. *J. Biol. Chem.*, **273**, 21663–21668.
18. Schaenman,J.M., Gilchrist,C.A., Mann,B.J. and Petri,W.A. Jr (2001) Identification of two *Entamoeba histolytica* sequence-specific URE4 enhancer-binding proteins with homology to the RNA-binding motif RRM. *J. Biol. Chem.*, **276**, 1602–1609.
19. Gilchrist,C.A., Holm,C.F., Hughes,M.A., Schaenman,J.M., Mann,B.J. and Petri,W.A. Jr (2001) Identification and characterization of an *Entamoeba histolytica* upstream regulatory element 3 sequence-specific DNA-binding protein containing EF-hand motifs. *J. Biol. Chem.*, **276**, 11838–11843.
20. Gilchrist,C.A., Leo,M., Line,C.G., Mann,B.J. and Petri,W.A. Jr (2003) Calcium modulates promoter occupancy by the *Entamoeba histolytica* Ca^{2+}-binding transcription factor URE3-BP. *J. Biol. Chem.*, **278**, 4646–4653.
21. Loftus,B., Anderson,I., Davies,R., Alsmark,U.C., Samuelson,J., Amedeo,P., Roncaglia,P., Berriman,M., Hirt,R.P. *et al.* (2005) The genome of the protist parasite *Entamoeba histolytica*. *Nature*, **433**, 865–868.
22. MacFarlane,R.C., Shah,P.H. and Singh,U. (2005) Transcriptional profiling of *Entamoeba histolytica* trophozoites. *Int. J. Parasitol.*, **35**, 533–542.
23. Gilchrist,C.A., Mann,B.J. and Petri,W.A. Jr (1998) Control of ferredoxin and Gal/GalNAc lectin gene expression in *Entamoeba histolytica* by a *cis*-acting DNA sequence. *Infect. Immun.*, **66**, 2383–2386.
24. Bradford,M.M. (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.*, **72**, 248–254.
25. Bruchhaus,I., Loftus,B.J., Hall,N. and Tannich,E. (2003) The intestinal protozoan parasite *Entamoeba histolytica* contains 20 cysteine protease genes, of which only a small subset is expressed during in vitro cultivation. *Eukaryotic Cell*, **2**, 501–509.
26. Bruchhaus,I., Leippe,M., Lioutas,C. and Tannich,E. (1993) Unusual gene organization in the protozoan parasite *Entamoeba histolytica*. *DNA Cell Biol.*, **12**, 925–933.
27. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
28. Bailey,T.L. and Gribskov,M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.

29. Tamada,Y., Kim,S., Bannai,H., Imoto,S., Tashiro,K., Kuhara,S. and Miyano,S. (2003) Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, **19** (**Suppl 2**), II227–II236.
30. Matsumoto,S. and Yanagida,M. (1985) Histone gene organization of fission yeast: a common upstream sequence. *EMBO J.*, **4**, 3531–3538.
31. Oliva,A., Rosebrock,A., Ferrezuelo,F., Pyne,S., Chen,H., Skiena,S., Futcher,B. and Leatherwood,J. (2005) The cell cycle-regulated genes of *Schizosaccharomyces pombe*. *PLoS Biol.*, **3**, e225.
32. Satish,S., Bakre,A.A., Bhattacharya,S. and Bhattacharya,A. (2003) Stress-dependent expression of a polymorphic, charged antigen in the protozoan parasite *Entamoeba histolytica*. *Infect. Immun.*, **71**, 4472–4486.
33. Causton,H.C., Ren,B., Koh,S.S., Harbison,C.T., Kanin,E., Jennings,E.G., Lee,T.I., True,H.L., Lander,E.S. *et al.* (2001) Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell*, **12**, 323–337.
34. Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.
35. Weber,C., Guigon,G., Bouchier,C., Frangeul,L., Moreira,S., Sismeiro,O., Gouyette,C., Mirelman,D., Coppee,J.Y. *et al.* (2006) Stress by heat shock induces massive down regulation of genes and allows differential allelic expression of the Gal/GalNAc lectin in *Entamoeba histolytica*. *Eukaryot. Cell*, **5**, 871–875.
36. Zhao,Y., McIntosh,K.B., Rudra,D., Schawalder,S., Shore,D. and Warner,J.R. (2006) Fine-structure analysis of ribosomal protein gene transcription. *Mol. Cell. Biol.*, **26**, 4853–4862.
37. van Poppel,N.F., Welagen,J., Vermeulen,A.N. and Schaap,D. (2006) The complete set of *Toxoplasma gondii* ribosomal protein genes contains two conserved promoter elements. *Parasitology*, **133**, 19–31.