GENOME-WIDE ANALYSIS OF BACTERIAL PROMOTER REGIONS

ELEAZAR ESKIN[1], URI KEICH[2], MIKHAIL S. GELFAND[3], PAVEL A. PEVZNER[2]

[1] *Department of Computer Science, Columbia University, New York, NY 10027,* eeskin@cs.columbia.edu

[2] *Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA 92093-0114,* {keich,ppevzner}@cs.ucsd.edu

[3] *IntegratedGenomics-Moscow, P.O.Box 348, Moscow, 117333, Russia,* gelfand@integratedgenomics.ru

Identifying prokaryotic promoter sequences is notoriously difficult and for most sequenced bacterial genomes the promoter sequences are still unknown. Since experimental analysis trails behind sequencing, genome-wide computational promoter discovery is often the only realistic way to discover these sequences in newly sequenced bacterial genomes. However, genome-wide samples for promoter discovery may be very large and corrupted complicating promoter discovery. We discuss three aspects of genome-wide promoter discovery: sample generation, signal finding algorithms, and scoring signals. We applied our new MITRA algorithm to analyze samples of divergent and convergent genes in 20 bacterial genomes and found strong putative dyad signals in 17 out of the 20 genomes. Moreover, in 12 out of 20 genomes the found signals are identical or similar to the known regulatory patterns (Pribnow-Gilbert boxes and CRP binding sites). Since many of putative signals correspond to previously known elements of bacterial transcriptional regulation, the remaining discovered signals are good candidates for unknown regulatory elements.

1 Introduction

A fundamental challenge of molecular biology is understanding the regulation of gene expression, in particular, on the level of transcription. In prokaryotes, genes encoding transcription factors may constitute up to 10% of the genome, as in *Pseudomonas aeruginosa* [1]. Thus an important component of genomic analysis is automated identification of transcription signals such as promoters. Since the experimental analysis trails behind the sequencing of new genomes, we are interested in discovery of regulatory sequences from complete genomes. This paper describes a method for discovering putative regulatory sites by fully automated genome-wide sequence analysis.

Discovering putative regulatory sites from complete genomes is a very difficult problem [2]. These difficulties are threefold and include (i) difficulties in sample generation, (ii) algorithmic difficulties in scaling to large and corrupted

samples, and (iii) statistical difficulties in assessing the significance of patterns that are discovered.

Regulatory signals can be modeled either as patterns or as profiles. This paper focuses on pattern approaches but the described algorithms are applicable to both patterns and profiles. Our algorithm will return a set of patterns which hopefully correspond to actual transcription signals in the genome. We use the term *monad* pattern to refer to a contiguous l-mer that occurs in the sample with up to d mismatches. Allowing for mismatches takes into account the fact that binding sites exactly matching the pattern are rare and often physiologically undesirable. We use the term $(l, d) - k$ pattern to denote a monad of length l that occurs in the sample k times (with up to d mismatches). Since bacterial regulatory signals are often build from two monad parts that occur with a fixed (or almost fixed) distance from each other, we model these binding sites by *dyad* patterns. A dyad pattern $(l_1 - (s_1, s_2) - l_2, d) - k$ are two monad patterns of length l_1 and l_2 respectively, that occur k times in the sample with a minimum separation distance of s_1 and a maximum separation distance of s_2. If $s_1 = s_2$ (a dyad with a fixed distance between the monads) we use the notation $(l_1 - s - l_2, d) - k$ patterns. This paper focuses on dyad signals that are common elements of bacterial transcriptional regulation.

There are many approaches to discovering monad signals [3,4,5,6,7,8] and dyad signals [9,10,11,12]. Recently there has been an emergence of powerful sample-driven approaches to monad pattern discovery [13,14,15] that are efficient enough to handle large genomic samples. For dyad signals, the sample-driven approaches include the algorithms presented in Marsan and Sagot 2000 [16] and MITRA [15]. For a given l_1, l_2, s_1, s_2, and k, these methods can find all $(l_1 - (s_1, s_2) - l_2, d) - k$ dyad patterns in the sample and are efficient enough to apply to samples of the total size exceeding $100,000$ nucleotides.

Our sample generation approach relies on comparative analysis of intergenic regions between divergently and convergently transcribed genes. We first take advantage of the relative location of genes in order to determine the regions where the binding sites most likely occur. Although regulatory elements are located upstream of genes, most upstream regions of bacterial genes do not contain promoters. Genes in bacterial genomes often form operons and only the intergenic region upstream of the first gene in the operon contain regulatory elements. Since the operon structure in bacterial genomes is rarely known, it is not clear how to automatically generate samples of regulatory regions. Our sample generation approach is based on the observation that intergenic regions between consecutive genes transcribed in divergent directions are guaranteed to be upstream regions of operons. By similar reasoning, intergenic regions between two convergently transcribed genes usually do not contain binding

sites. We use the intergenic regions between genes that are transcribed in convergent directions as our background sample. Since both divergent and convergent intergenic regions are selected from the same genome, the convergently transcribed intergenic regions allow us to estimate a background distribution for upstream regions with regulatory elements (the intergenic region between divergently transcribed genes).

A key ingredient for discovery of putative regulatory signals is a method to assess the statistical significance of a potential signal. The problem is nontrivial since our samples are large, biased, and contain low complexity regions. There exists a number of approaches to assessing the statistical significance of patterns. They include the shuffling approach [17], building statistical models to estimate the probability of a pattern [18] or profile [3,4], and using a background sample to assess the significance of a pattern depending on whether or not it is over represented in the sample [19]. Our ability to discover signals depends on the reliability of the method for determining the statistical significance of observed patterns in such large samples. For example consider the experimentally confirmed promoter represented by the Pribnow-Gilbert dyad $TTGACA-17-TATAAT$ in the *B. subtilis* genome. This dyad (with 2 allowed mismatches) occurs 143 times in our sample. However, it appears anywhere from 34 to 64 times at other separation distances from 3 to 23 nucleotides. Most of them are likely to be simply random events instead of having any biological meaning. It indicates a need for a new scoring approach that combines the traditional statistical analysis of monad patterns with the analysis of spacing and positional parameters.

Our scoring approach estimates the significance of patterns in the target sample against the patterns in the background sample. For each pattern, we compute the strength of the signal which measures the difference between the number of occurrences of the pattern and the expected number of occurrences based on the background distribution (strength score). However, even with scoring method that takes into account a background sample, it is still difficult to determine which patterns correspond to biologically meaningful signals. Our key idea is to incorporate two types of additional information to help make this determination. Firstly, we contrast every dyad pattern with a fixed separation distance against a dyad pattern with a "random" spacer (dyad score). Secondly, since some regulatory elements are positional (i.e., tend to occur at the same relative position) we also analyze the relative position of the signal to the start of the gene (positional score). Although we do not have reliable information about the transcription start position of the gene, we can still obtain a rough estimate of the relative positions of the signal using the translation start instead. Each type of information on its own is generally not sufficient to

make a determination on whether or not a signal is an actual binding site. In fact, some actual binding sites are not positional signals and some actual dyad binding sites have looser restrictions on their separation distance than other dyad binding sites. However, the combination of the three types of information helps us to decide whether or not a signal is a putative binding site.

We apply our new MITRA algorithm to analyze samples of divergent and convergent genes in 20 bacterial genomes and find particularly strong putative dyad signals in 9 out of the 20 genomes and signals that correspond to known binding sites in 12 of the 20 genomes. Details of the MITRA algorithm are presented in [15]. Detailed information about all of the signals reported in this paper is available at: `http://www.cs.columbia.edu/compbio/mitra/`.

2 Sample Generation

The main difficulty in the analysis of complete bacterial genomes is scarcity, or even lack, of the experimental data about location of regulatory regions and the operon structure. Thus, for any given gene, it is difficult to decide, whether it is the first gene in an operon (and thus the transcription factor binding sites are upstream of this gene), or it is preceded by other genes. Therefore, simply taking upstream regions for every gene in a bacterial genome would lead to extremely corrupted samples and failure of the motif finding algorithms.

A better approach to sample generation was first proposed by Washio et al., 1998 [20] and later explored by Sagot and colleagues in [2] and [21]. It is based on the observation that the divergently transcribed genes are guaranteed to be the most upstream genes of the respective operons. Thus, the target sample we used to search for the regulatory signals consists of genomic fragments between divergently transcribed genes. Similarly, a region between convergently transcribed genes cannot be an upstream region for any gene, and such regions formed a background sample (i.e., sample without binding sites).

The sequences of 20 complete bacterial genomes (Table 1) were downloaded from the ERGO database [22]. The choice of the genomes was dictated by (i) availability of experimental information for some of these genomes and (ii) availability of several genomes from one taxonomic group. We used the limited set of experimentally confirmed promoters to verify that our predictions agree with the available data. We used genomes from the same taxonomic group (Gram-positive bacteria from the Bacillus/Clostridium group, mycoplasmas, chlamidiae, proteobacteria from the α, β and γ divisions, ϵ-proteobactera) to check whether our promoter predictions for these genomes produce similar putative patterns.

To create our samples, we extract the last 310 bases of the intergenic region

and the reverse complement of the first 310 bases. We remove the 10 bases closest to the gene to delete the strongly conserved Shine-Dalgarno signal that would dominate our results. In the cases of alternatives caused by overlapping genes the shortest intergenic fragments were selected. In cases where the intergenic region is longer than 620 nucleotides, a portion of the intergenic region is left out of the target sample. We perform the same procedure for creating the background sample that model the regions without the regulatory elements.

3 Finding Statistically Significant Signals

We use the MITRA algorithm to detect all "statistically significant" $(l_1 - (s_1, s_2) - l_2, d) - k$ patterns. MITRA is fully described in Eskin and Pevzner, 2002 [15] and is easily adapted to use the scoring method described below. Although MITRA was used for these experiments, any algorithm (such as Marsan and Sagot, 2000[16]) that can recover all $(l_1-(s_1,s_2)-l_2,d)-k$ patterns, properly modified to incorporate the scoring functions described below, would produce equivalent results.

We incorporate three types of information into assessing the significance of a signal: signal strength score, dyad separation score and positional score.

We use the background sample obtained from intergenic sequences between genes transcribed in convergent directions to estimate background distribution. We first describe our scoring method for monad patterns and then extend it to dyads patterns. For a pattern P we define $p_P = \frac{n_P}{n_B}$ as the number of l-mers in the background sample that are within d mismatches of the pattern, n_P divided by the total number of l-mers in the background sample, n_B. We smooth our estimates for p_P using Dirichlet priors [23] and adjust our estimates of p_P to take into account the differences in nucleotide composition between convergent and divergently transcribed intergenic regions.

Let n_T be the size of the target sample. Given that the pattern P occurs (with mismatches) o_P times in the target sample, we define the score of the pattern as $s_P = \frac{o_P - n_T p_P}{\sqrt{n_T p_P (1-p_P)}}$. For a single pattern, the score can be interpreted as the number of standard deviations from the mean if we assume a binomial distribution. The pattern score is simple and efficient enough to incorporate into MITRA for the exhaustive search to discover *all* top scoring patterns. Instead of returning all patterns that occur k times, we instead specify a minimal score threshold t. For a pattern P and minimum score threshold t, the minimum number of occurrences for the pattern k_P to make into the ranked list would be $k_P = n_T p_P + t\sqrt{n_T p_P (1 - p_P)}$.

We score the dyad patterns D composed of monad patterns P_1 and P_2 in a similar way. For each dyad pattern D, we estimate p_D from estimates

of the probabilities of the patterns P_1 and P_2. Since the mismatches of an instance of the dyad can be spread to both monads, we need to estimate a probability for each monad occurring with a certain number of mismatches. As above, we compute the counts for each occurrence of the pattern with i mismatches over the background sample and divide by the size of the sample. We use p_P^i to denote the probability of a pattern P occurring with i mismatches. We then estimate the probability for the dyad pattern D using $p_D = \sum_{i,j \text{ s.t. } i \geq 0, j \geq 0, i+j \leq d} p_{P_1}^i p_{P_2}^j$. If s is the number of allowable separation distances, the score for a dyad D, s_D is defined to be $s_D = \frac{o_D - sn_T p_D}{\sqrt{sn_T p_D(1-p_D)}}$. For a minimum score threshold t, we set minimum number of occurrences for a dyad pattern D as $k_D = sn_T p_D + t\sqrt{sn_T p_D(1-p_D)}$.

Two other types of information is the distribution of separation distances between the dyad signals and the distribution of the positions of instances of the signal. Many dyads which correspond to a binding site, have a peak in the histogram of separation distances at a certain separation distance such as in Figure 1(a) for *B.subtilis*. Similarly, many binding sites tend to occur in a similar position relative to the transcription start of the gene such as in Figure 1(c) *B.subtilis*. At the same time, for most regulatory signals the situation is more difficult, for example, the same histograms for *E.Coli* 1(b,d) show less pronounced peaks.

We incorporate this information by assessing the statistical significance of the distribution of both the separation distance (between the two parts of the dyad) and of the position of the signal (relative to the estimated transcription start site). In the first case our null hypothesis is that every instance of the dyad is independently equally likely to fall in one of the $s_2 - s1 + 1$ possible bins: one for each possible separation distance. For the positional histogram, since there is often some flexibility in the position of the transcription factor relative to the transcription start site, we group the positions using bins of 30 bp.

We assess the statistical significance of the observed data using the statistic M which equals the maximal number of instances that fall in one bin. Under our null hypothesis M is distributed as the maximum multinomial bin which is given by[24]:

$$P(M \leq i) = \frac{N!}{N^N e^{-N}} \left[F_{\text{Pos}}[N/t](i)\right]^t P(W_i = N) \qquad i = 0, 1, \ldots,$$

where N is the total number of instances, t is the number of bins, $F_{\text{Pos}}[\lambda](i)$ is the cdf of a Poisson random variable with parameter λ evaluated at i, and W_i is a sum of t iid Poisson random variables ($\lambda = N/t$), each of which is subject to truncation at i.

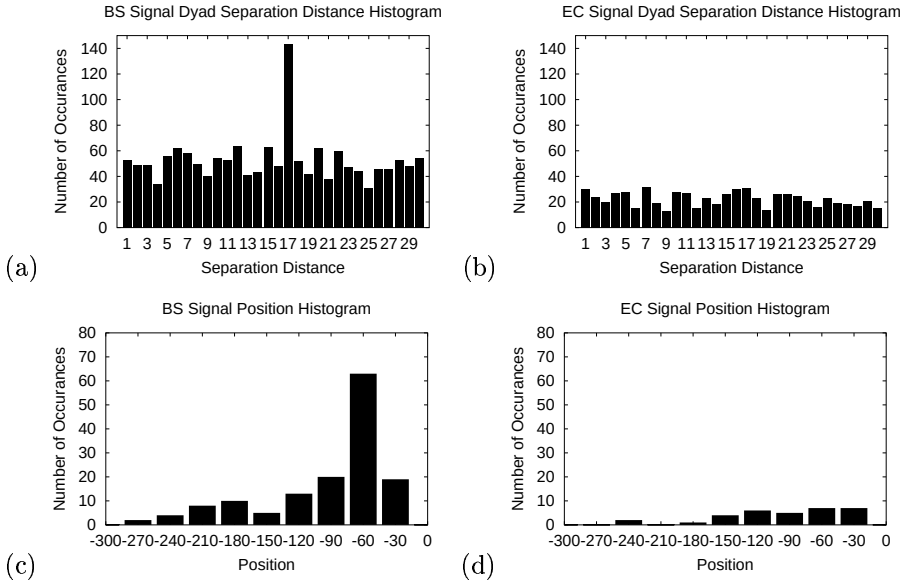

Figure 1: Histogram of Separation Distances and Positions for the Pribnow-Gilbert dyad signal TTGACA-17-TATAAT. Separation Distances in (a) BS genome and (b) EC genome. Positions in (c) BS genome (d) EC genome. We bin the positions of the signal instances into buckets (30 bp by default) since positions rarely exactly match and the exact transcription start position often unknown.

In practical terms $P(W_i = N)$ can be computed by a t-fold convolution of the truncated Poisson distribution. Note that the distribution of a Poisson (λ) random variable which is truncated at i is given by $P_{\mathrm{Pos}}[\lambda](j)/F_{\mathrm{Pos}}[\lambda](i)$ for $j = 0, 1, \ldots, i$, where $P_{\mathrm{Pos}}[\lambda](j) = e^{-\lambda}\lambda^j/j!$ is the standard Poisson probability mass function. Also, by Stirling's approximation, for a reasonably sized N we can approximate $\frac{N!}{N^N e^{-N}} \sim \sqrt{2\pi N}$. For both of these scores we compute the P-value of the statistic and report its negative log.

4 Finding putative regulatory elements in bacterial genomes

We performed blind experiments over 20 bacterial genomes (Table 1) and extracted the top dyad signals for each genome. We searched for dyads consisting of two conserved regions of length 6 with separation distances from 3 to 23 bases. To validate our results, we checked the found signals against the known regulatory elements in bacterial genomes. Table 2 shows the top signals

Table 1: Genome Intergenic Region Statistics. The first column is name of the genome. The ID is an abbreviation for the genome. The next columns list the number and the lengths of intergenic regions for divergent and convergent samples. The last column describe genome taxonomy.

Genome Name	Genome ID	Div. Regions	Div. Nucleotides	Conv. Regions	Conv. Nucleotides	Genome Taxonomy
Bacillus subtilis	BS	552	132145	244	47087	*Bacillus* group
Campylobacter jejuni	CJ	168	25837	36	4638	ϵ-proteobacteria
Chlamydia muridarum	CMU	120	29176	55	5751	*Clamidiales*
Chlamydophila pneumoniae	CPX	136	33274	67	8805	*Chlamidiales*
Chlamydia pneumoniae	CQ	141	36973	68	11460	*Chlamidiales*
Synechocystis sp.	CY	525	138494	330	50311	*Cyanobacteria*
Escherichia coli	EC	589	148936	374	61262	γ-proteobacteria
Haemophilus influenzae	HI	228	48346	181	23944	γ-proteobacteria
Helicobacter pylori	HP	169	33123	109	26091	ϵ-proteobacteria
Lactococcus lactis	LLX	217	53342	139	23081	*Bacillus* group
Mycoplasma genitalium	MG	30	5013	7	1404	*Mycoplasmales*
Mycoplasma pneumoniae	MP	51	9535	41	10235	*Mycoplasmales*
Mycobacterium tuberculosis	MT	530	93471	290	43742	*Mycobacteria*
Neisseria meningitidis ser. B	NX	287	55651	266	45793	β-proteobacteria
Pseudomonas aeruginosa	PA	790	194787	450	80089	γ-proteobacteria
Rickettsia prowazekii	RP	104	42597	99	62895	α-proteobacteria
Streptococcus pyogenes	ST	169	47300	159	33165	*Bacillus* group
Thermotoga maritima	TM	163	25332	29	4542	*Thermotogales*
Ureaplasma urealyticum	UU	55	12728	40	10623	*Mycoplasmales*
Pasteurella multocida	VK	267	58281	238	30640	γ-proteobacteria

from each of the 20 genomes with respect to strength score and provides the strength score, the dyad score and the positional score for each signal. We do not report a signal if (i) it is a slight variation of a higher scoring signal, (ii) if it is a shifted variant of a higher scoring signal, or (iii) if it is a reverse complement of a higher scoring signal.

The set of putative signals identified by our algorithm contains a number of known signals and several promising candidates for more detailed analysis. Among the known signals are the classical promoter signals consisting of the standard Gilbert and Pribnow boxes. They have been found in all Gram-positive bacteria from the *Bacillus* group: *B. subtilis*, *S. pyogenes*, and *L. lactis*, as well as in alpha-proteobacterium *R. prowazekii*.

For nine genomes (BS, CY, EC, HI, LLX, NX, PA, ST, and VK) we discovered the particularly strong signals with high dyad and strength scores. For six of these genomes, these signals correspond to known biological signals or to variants of known signals.

In *B. subtilis. (BS)*, one of the strongest dyad signals that we recovered was the classical Pribnow-Gilbert promoter consensus $TTGACA-17-TATAAT$. This signal is over-represented in the divergent intergenic regions relative to the convergent intergenic regions, has a very strong distance peak at the distance 17 as well as a strong positional peak at distance in the range -90 to -60. The distribution of separation distances is shown in Figure 1(a) and the positional distribution is shown in 1(c). In *E. coli. (EC)*, the found signal in *E. coli* perfectly matches the binding signal of the transcription factor CRP (TGTGAT-4-ATCACA). In *H. influenzae. (HI)*, all three signals found in *H.*

influenzae are interesting. The first found dyad TGCGGT-12-CGTTTT signal has a strongly conserved region around the dyad represented by the longer dyad AAAAGTGCGGTNA-10-CGTTTT. The second found signal is the binding signal of the transcription factor CRP. In addition, the third signal has an additional interesting feature. Although it looks like an *AT* rich signal it tends to occur in non-*AT* rich regions which suggests that it is a real binding site. In *L. lactis. (LLX)*, the found dyad corresponds to the canonical Pribnow-Gilbert promoter consensus. In *S. pyogenes. (ST)*, the found dyad corresponds to the canonical Pribnow-Gilbert promoter consensus. In *P. multocida. (VK)*, the found dyad is a slightly shifted form of the binding signal of the transcription factor CRP. In addition, there are two lower-scoring but still interesting palindromes AATGTG-10-CACATT and AATTTG-12-CAAATT that may be the binding signals for yet unknown transcription factors. We plan to do detailed analysis of these signals in order to determine the corresponding regulons.

Many of the other signals detected also correspond to known binding sites. In *N. meningitidis (NX)* and *R. prowazekii (RP)*, we detect canonical Pribnow-Gilbert boxes. Among other identified signals, there are modified forms of the Pribnow-Gilbert promoter consensuses: TTGACA-19-ATAATT in *C. pneumoniae (CPX)* (the Pribnow box is shifted by 1 bp to the right; the spacer length is longer than in other species), TTAATC-21-TATAAT in *H. pylori (HP)*, identified earlier in [21] (unusual Gilbert box and longer spacer), TTGACC-17-TAGAAT in *P. aeruginosa (PA)* (modified boxes), and TTGCCA-17-TACAAT in *P. multocida (VK)* (modified boxes). Many of the other signals that we find are palindromic signals (Table 2).

The known promoter signal of *E. coli* was too weak to be discernible in this analysis as well as previous genome-wide analyses[21]. The promoter signal in *Mycoplasmas* is also very weak; even given a sample of mapped promoters it is not possible to derive a good consensus [25]. The signals identified in *M. tuberculosis* do not resemble the promoter consensus of a closely related bacterium *M. paratuberculosis* [26]. We also did not find any signal corresponding to the suggested consensus TTTAAGT-(15-19)-TATAAT of *C. jejuni* [27].

Some signals may still be artifacts. In particular, the AT-rich signals of *Mycoplasmas* represent neither promoters identified in experimental study [25], nor can they be binding signals of transcription factor HrcA (the CIRCE box TTAGCACTC-9-GAGTGCTAA) identified in [28]. Despite the method's inability to find these signals, the fact that we were able to identify many of the promoters and transcription factor binding signals demonstrates the power of the method and indicates that at least some of the identified candidates deserve closer look.

Table 2: Top scoring dyad signals in 20 bacterial genomes. Underlined signals are particularly strong (strength scores greater than 10 and either a dyad score or position score greater than 7). The Signal Class column labels if the signal falls into a known biological signal. Classes are defined as (PB) Pribnow-Gilbert signal, (PB*) variant Pribnow-Gilbert signal, (CRP) CRP signal, (CRP*) variant CRP signal, (PU) palindromic signal for a possibly unknown factor.

Genome ID	Signal Pattern	Number Occurrences	Strength Score	Dyad Score	Position Score	Signal Class
BS	TTGACA-17-TATAAT	143	20.10	8.86	7.42	PB
BS	CCTCCT-16-CATTAT	62	13.15	4.07	3.53	
BS	TATAAT-5-TATAAT	151	11.25	5.81	7.48	
CJ	TTCCCT-10-AAATTT	54	4.79	3.86	2.76	
CJ	TACCAT-8-TAAAAT	58	2.71	8.19	6.32	
CJ	TTTAAC-11-TAGAAT	71	4.19	8.30	6.74	
CMU	AATTAT-6-ATAATT	39	2.61	7.67	0.37	PU
CMU	AATATA-18-TATATT	37	1.21	7.51	2.37	PU
CMU	CATTGT-12-TCTTCT	21	1.81	6.75	0.09	
CMU	GCAACA-21-AAATAA	22	0.86	3.49	0.62	
CPX	TTGACA-19-ATAATT	27	3.43	7.16	1.92	PB*
CPX	GTGCAA-11-TTTTTC	25	3.81	6.95	0.30	
CPX	ATTAAT-12-ATTAAT	42	2.78	2.17	0.74	PU
CPX	ATTATT-6-ATTAAT	57	3.21	3.35	4.11	
CQ	AAAATT-6-ATAATG	49	3.17	7.76	1.57	
CQ	ATTATT-6-ATTAAT	57	3.91	3.25	5.98	
CQ	ATTAAT-14-ATTATT	54	4.42	4.48	2.76	
CY	ATTGTA-11-AATTTT	76	7.67	3.75	3.43	
CY	TGTTAA-4-TGTTAC	57	12.77	7.83	3.30	
EC	TGTGAT-4-ATCACA	108	24.04	8.01	2.99	CRP
HI	TGCGGT-12-CGTTTT	58	5.84	7.80	3.14	
HI	TGTGAT-4-ATCACA	62	17.64	7.11	6.44	CRP
HI	AAAATT-6-AATTTT	342	13.21	9.85	8.26	PU
HP	ATTATA-10-TATAAT	89	7.17	7.81	6.96	PU
HP	ATTTTA-18-TATGCT	49	6.04	7.97	2.51	
HP	TTAAGC-21-TATAAT	51	6.62	7.79	6.41	PB*
HP	GTATAA-7-ATTATA	56	8.78	8.09	6.49	
LLX	TTGACA-17-TATAAT	125	19.87	8.62	7.29	PB
LLX	TTATAA-5-TTATAA	203	8.41	9.54	7.76	PU
MG	AGTAAA-10-TTTACT	20	0.11	5.99	2.16	PU
MG	AATCAA-11-ACTTTT	21	0.78	6.70	0.09	
MP	TCCAAA-14-TTTTTA	23	2.91	6.14	0.49	
MP	TTGTAA-15-TAATTA	20	4.33	4.18	0.92	
MP	TTAAAA-17-TTAGTA	23	2.88	6.46	0.49	
MP	TTATTA-18-CAAACA	20	3.91	6.08	3.58	
MT	CGGCCC-10-CGGGCC	80	6.47	8.53	1.37	
MT	CGATAC-12-CGCCGC	51	4.84	7.57	2.22	
MT	GGCCCG-8-CCAGGC	66	7.37	8.31	4.86	
NX	CCGCCG-12-CGGCGG	33	10.86	7.40	3.13	PU
NX	TTGACA-17-TATAAT	43	12.87	7.71	6.24	PB
NX	CTTCAG-3-GCATAG	50	19.61	6.68	4.45	
NX	TATAGT-6-ACTATA	30	7.78	7.17	0.50	PU
PA	TTGACC-17-TAGAAT	31	9.99	7.18	5.91	PB*
PA	TGTCAC-5-TGTCAC	34	9.51	7.23	5.18	
PA	TATAAT-6-CAATTT	30	11.77	7.21	5.88	
PA	TATAAT-3-CGGCCT	50	8.63	7.50	6.39	
RP	TTGACA-17-TATAAT	42	6.55	7.56	4.80	PB
RP	CTTTAA-21-TTAAAG	44	6.99	7.32	1.32	PU
RP	AATTAT-22-TTTCCC	20	6.56	6.56	0.92	
RP	TAATTA-9-AACCAT	39	5.46	7.52	1.07	
ST	TTGACA-17-TATAAT	92	18.56	8.24	6.99	PB
ST	AATTAT-3-ATAATA	123	10.80	8.85	5.97	
TM	TTGACA-17-TATAAT	33	4.34	7.43	5.98	PB
UU	TATAAT-13-TACAAT	39	1.12	0.37	3.34	
VK	AATGTG-10-CACATT	58	12.36	7.76	3.14	PU
VK	AATTTG-12-CAAATT	79	8.97	8.57	3.94	PU
VK	ATTGTA-12-AAATTT	78	6.72	7.02	6.67	
VK	TTGCCA-17-TACAAT	30	7.85	7.13	5.88	PB*
VK	GTGATC-4-TCACAA	70	27.03	7.33	6.72	CRP*

5 Conclusion

We presented an approach for fully automatic discovery of putative regulatory signals in bacterial genomes. The approach emphasizes the interplay of three processes: sample generation, signal finding, and scoring.

We applied our MITRA algorithm to 20 bacterial genomes and detected signals that correspond to known binding sites in 12 of the 20 genomes. The majority of the strong signals detected by MITRA, do in fact correspond to known biological binding sites. Of the 14 particularly strong signals detected by MITRA, 4 correspond to a Pribnow-Gilbert signal or one of its variants and 3 correspond to a CRP signal. A very promising direction is to further examine the remaining 7 strong signals to determine whether or not they correspond to actual binding sites.

6 Acknowledgments

We are grateful to Alexei Dolgopolov for assistance with generating the data samples from genomic sequences. MG was partially supported by grants from INTAS (99-1476), HHMI (55000309) and RFBR (00-15-99362).

1. C.K. Stover et al. Complete genome sequence of pseudomonas aeruginosa pa01, an opportunistic pathogen. *Nature*, 406:959–964, 2000.
2. A. Vanet, L. Marsan, and M. Sagot. Promoter sequences and algorithmical methods for identifying them. *Research in Microbiology*, 150:779–799, 1999.
3. T. L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21:51, 1995.
4. G. Hertz and G. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15:563–577, 1999.
5. C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.
6. A. Neuwald, J. Liu, and C. Lawrence. Gibbs motif sampling: Detection of bacterial outer membrane repeats. *Protein Science*, 4:1618–1632, 1995.
7. J. Buhler and M. Tompa. Finding motifs using random projections. In *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology (RECOMB01)*, pages 69–76, 2001.
8. P. A. Pevzner and S. Sze. Combinatorial approaches to finding subtle signals in DNA sequences. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 269–278, 2000.
9. D. GuhaThakurta and G. D. Stormo. Identifying target sites for cooperatively binding factors. *Bioinformatics*, 17:608–621, 2001.
10. M.S. Gelfand, E.V. Koonin, and A.A. Mironov. Prediction of transcription regulatory sites in archaea by a comparative genomic approach. *Nucleic Acids Research*, 28:695–705, 2000.
11. J. van Helden, A. F. Rios, and J. Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Research*, 28:1808–1818, 2000.

12. X. Liu, D.L. Brutlag, and J.S. Liu. Bioprospector: Discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. In *Proceedings of the 2001 Pacific Symposium on Biocomputing*, volume 6, pages 127–138, 2001.
13. M. Sagot. Spelling approximate or repeated motifs using a suffix tree. *Lecture Notes in Computer Science*, 1380:111–127, 1998.
14. G. Pavesi, G. Mauri, and G. Pesole. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, 17:S207–S214, July 2001. Proceedings of the Ninth International Conference on Intelligent Systems for Molecular Biology.
15. E. Eskin and P. Pevzner. Finding composite regulatory patterns in dna sequences. *Bioinformatics*, Supplement 1:S354–63, 2002. Proceedings of the Tenth International Conference on Intelligent Systems for Molecular Biology (ISMB-2002).
16. L. Marsan and M. Sagot. Algorithms for extracting structured motifs using a suffix tree with applications to promoter and regulatory site consensus identification. *Journal of Computational Biology*, 7:345–360, 2000.
17. S. Karlin, F. Ost, and B. E. Blaisdell. Patterns in dna and amino acid sequence and their statistical significance. In M. S. Waterman, editor, *Mathematical Methods for DNA Sequences*, pages 133–158, Boca Raton, France, 1989. CRC Press.
18. A. Denise, M. Regnier, and M. Vandenbogaert. Assessing the statistical significance of overrepresented oligonucleotides. In *Proceedings of the 1st Workshop on Algorithms in BioInformatics*, Denmark, 2001.
19. Y. Barash, G. Bejerano, and N. Friedman. A simple hypergeometric approach for discovering putative transcription factor binding sites. In *Proceedings of the 1st Workshop on Algorithms in BioInformatics*, 2001.
20. T. Washio, J. Sasayama, and M. Tomita. Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination. *Nucleic Acids Res.*, 26:5456–5463, 1998.
21. A. Vanet, L. Marsan, A. Labigne, and M. Sagot. Inferring regulatory elements from a whole genome. an analysis of Helicobacter pylori σ^{80} family of promoter signals. *Journal of Molecular Biology*, 297:335–353, 2000.
22. R. Overbeek, N. Larsen, G.D. Pusch, M. D'Souza, E.Jr. Selkov, N. Kyrpides, M. Fonstein, N. Maltsev, and E.T. Selkov. Wit: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.*, 28:123–125, 2000.
23. M. H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.
24. B. Levin. A representation for multinomial cumulative distribution functions. *Ann. Statist.*, 9(5):1123–1126, 1981.
25. J.Weiner, R. Herrmann, and G.F. Browning. Transcription in mycoplasma pneumoniae. *Nucleic Acids Res.*, 28:4488–4496, 2000.
26. J.P. Bannantine, R.G. Barletta C.O. Thoen, and R.E. Andrews. Identification of mycobacterium paratuberculosis gene expression signals. *Microbiology.*, 143:921–928, 1997.
27. M.M. Wosten, M. Boeve, M.G. Koot, A.C. van Nuene, and B.A. van der Zeijst. Identification of campylobacter jejuni promoter sequences. *J Bacteriol.*, 180:594–599, 1998.
28. R. Segal and E.Z. Ron. Regulation and organization of the groe and dnak operons in eubacteria. *FEMS Microbiol Lett*, 138:1–10, 1996.