An Algorithmic Approach to Gene Regulatory Sequence Analysis

Eric C. Rouchka[1]
TR-ULBL-2008-01

March 24, 2008

[1]University of Louisville
Speed School of Engineering
Department of Computer Engineering and Computer Science
123 JB Speed Building
Louisville, Kentucky, USA 40292

eric.rouchka@louisville.edu

UNIVERSITY of LOUISVILLE dare to be great
Bioinformatics Laboratory Technical Report Series

Bioinformatics Research

An Algorithmic Approach to Gene Regulatory Sequence Analysis

Eric C. Rouchka[1,*]

[1]Department of Computer Engineering and Computer Science, University of Louisville, 123 JB Speed Building, Louisville, KY, USA

UNIVERSITY OF LOUISVILLE BIOINFORMATICS LABORATORY TECHNICAL REPORT SERIES REPORT NUMBER **TR-ULBL-2008-01**

ABSTRACT

Motivation: Genes code for proteins, the action molecules of life. The bulk of scientific effort has focused on the genes and their products. Yet only about 3% of the human genome codes for genes. The remainder, sometimes called "junk DNA", has received far less attention. Recent findings suggest that this junk may be far more important than previously believed. In particular this "junk" has been shown to contain important regulatory signals. A method for detecting common regulatory motifs using a modified Bernoulli approach to Gibbs Sampling is presented.

1 INTRODUCTION

DNA sequencing technology has outrun all other molecular biology methods leading to a vast data base of unexplored sequence data. The human genome project and other genome projects have greatly accelerated the growth of this data base. Since genes code for proteins, the action molecules of life, the bulk of scientific effort has focused on the genes and their products. Yet only about 3% of the human genome codes for genes. The remainder, sometimes called "junk DNA" has received far less attention. Recent findings suggest that this junk may be far more important than previously believed. In particular this "junk" has been shown to contain important regulatory signals.

1.1 Gene Regulation

Gene regulation is the fundamental process behind cellular alteration. It is used by single cell organisms to respond to changes in their environment and by multicellular organisms for cell differentiation. The most studied gene regulation involves the binding of regulator proteins to regulatory elements which are signal sequences that normally occur in an upstream fragment of the genome called the promoter. Far less studied are regulator elements that are more distant from the genes they regulate and regulation due to the binding of RNA gene regulators. While there are experimental methods to identify regulatory elements, they are time consuming and difficult and thus not particularly well suited to examine the vast database of sequences that is emerging from genome sequencing projects. Thus there is a need for computational and statistical methods to explore these databases to identify novel regulatory signals that may be hidden within the genomic "junk".

Gene regulation has the ability to cause short- and long-term effects. Short-term regulation generally leads to the production of induction or repression enzymes that allow cells to respond to changes in their environment. Long-term regulation is vital in determining the ontogeny of an individual. Regulatory regions differing from the norm may signal different mechanisms for gene regulation.

If the site of gene regulation can be determined, then the complex changes taking place in the expected motifs when things go wrong can begin to be analyzed better. Hopefully in the future gene therapy can be explored to help eliminate or lessen the effects of certain birth defects and cancers. A modified Gibbs Sampling approach using a Bernoulli method is presented for computational detection of novel motif patterns within similarly regulated nucleotide sequences.

2 METHODS

The goal is take a given set of amino acid or nucleotide sequences and determine common motif elements within them. The sequences composing this set can be derived from regions conserved among a number of different species, or from different regulatory regions within a single species. It is possible that more than one characteristic motif type exists in which case multiple gene regulation sites are probable. One approach known as site sampling assumes that each sequence contains exactly one motif element for each motif type. The alternative Bernoulli motif sampler assumes that each sequence can contain zero or more motif elements of each motif type.

*To whom correspondence should be addressed.

2.1 Residue Count Adjustments

When using Bayesian statistical methods, a distinction between items in the model and outside of the model needs to be made. In the case of motif sampling, occurrences of the motif will be considered part of the model, while those outside of motif areas are considered background noise.

Initially there are no motif elements for any of the motif types. Observed counts for each of the residues are calculated for the background based on the residues in the sequences. At the outset, observed motif counts are set to zero for each residue in each position of the motif. The observed counts are represented by the variable $q_{0,1}\ldots q_{i,n}$ where $q_{0,k}$ are the background counts, i is the number of motif types, and n is the alphabet length (4 for DNA sequences; 20 for amino acid sequences). In order to overcome issues with zero counts and overtraining of the data, pseudocounts are incorporated. The user can determine how much weight the pseudocounts hold by specifying a command line argument. The weight typically is in the range between 0 and 1 with a default of 0.1.

Motif pseudocounts are set using one of two methods. In the first method, there is no bias in the composition of the motif. In this uniformed priors method, pseudocounts for each residue in each position of the motif are calculated by setting them equal to the pseudocounts for that residue in the background. The second, known as the informed priors method, adds in prior knowledge weights specified by the user. Priors can be weighted for different positions in the motif according to confidence levels of the known motif structure. The observed counts and pseudocounts are important in calculations involving the Bayesian statistical formula.

Once the counts are calculated for the model without any motifs, the resulting null model is stored. An alignment of motif elements is randomly picked according to the motif length and initial number of motifs specified. Care is taken to ensure that motifs do not overlap the ends of the sequences or one another. At this point an alignment, albeit random, is given.

Observed background and motif counts are updated according to the initial alignment. This can be accomplished by adding in the residues occurring in the motif to the motif counts and deleting them from the background counts. The alignment is now held constant, and the sequence is traversed to see if the individual positions are motif starting positions according to the rules of either the site or Bernoulli sampler algorithm.

2.2 Site Sampler

The site sampler begins with a randomly assigned alignment of motifs, one per sequence. It then proceeds to follow the two steps of the Gibbs site sampler (Lawrence et al., 1993). The first set is the predictive update step in which one of the N sequences is chosen, beginning with the first sequence and proceeding to the last sequence. The motif element for each motif type in the current sequence is added to the background and the counts are updated accordingly. There is now a pattern description of the motif elements of the remaining sequences for each motif type.

The second step of the Gibbs site sampler is the sampling step. Each possible motif starting position is assigned a probability of being a motif starting position by following equations 1 - 4. Once a probability is assigned to each possible starting position, the probabilities are normalized so they sum to one. A single motif element is then sampled in for each motif type based on the probability weights. This is a random process, so the best fit may not be chosen, although the probabilities supply a weight to help ensure that descent elements are nearly always sampled when present.

$$P_{Motif}(t) = \prod_{j=1}^{J_t} \frac{c_{t,j,r} + b_{t,j,r}}{c_t + b_t} \tag{1}$$

$$P_{Element} = \frac{Odds}{1 - Odds} \tag{2}$$

$$P_{Background} = \prod_{j=1}^{J_t} \frac{c_r + q_r}{c + q} \tag{3}$$

$$Odds = \frac{P_{Motif}(t)}{P_{Background}} \tag{4}$$

For equations 1-4, t is the current motif type; J_t is the motif width, $c_{t,j,r}$ is the number of residues of type r at position j for motif type t; $b_{t,j,r}$ is the number of pseudocounts of type r at position j for motif type t; b_t is the total number of motif residue pseudocounts for motif type t; q_r is the number of background pseudocounts of type r, and q is the total number of background pseudocounts.

The site sampler will now proceed onto the next sequence and perform steps one and two until an alignment of elements has been chosen for each sequence. Once a correct pattern begins to appear in the first step, it will be improved upon by the weights associated with the sampling step until a local maximum alignment has not been obtained for a number of iterations through the site sampler. Such a lack of improvement suggests that the alignment is stuck in an "energy well" and will not improve. It is possible that this well is indeed the overall and not a local maximum.

2.3 Bernoulli Sampler

The Bernoulli sampler begins with an alignment of motifs randomly spread throughout the sequences. Since any given protein may have 0 or more specific motifs, all sequences are concatenated together to produce a single sequence. Inputs for the Bernoulli sampler include the initial sequences, the motif width(s) (in number of residues), and the initial estimate of the number of occurrences for each motif type.

The algorithm starts at the very first position in the long, concatenated sequence and checks to see if the position is a possible motif start site, i.e. a motif placed there will not overlap two sequences or a previously set motif element of a different type. If it is determined to be a possible start site, checking is done to see if it overlaps a motif element of the current type or a motif type that has not yet been set. When this happens, the motif element it overlaps is taken out of the alignment and the counts are updated accordingly. Note that the element can get sampled back in later. The current position is then sampled in either as a motif element of the current type or as background based on the probabilities that are calculated via equations 1-4, in a similar fashion as the site sampler.

2.4 Calculating Alignment MAP Value

Once an alignment of motif elements has been found for all of the motif types (i.e. the motif elements for the last sequence have been set in the site sampler or the last position of the long sequence has been sampled in or out in the motif sampler), the current alignment is tested to see how well it describes the data. The value that is calculated is the maximum *a posteriori* (MAP) value (Liu et al., 1995). A maximal value of the MAP is desired. Rather than take the gamma functions (as shown in equation 10 of (Liu et al., 1995)), the natural log of the gammas is found. There are three parts to the MAP: the motif portion, the background portion and the beta distribution portion. The MAP for the site sampler and motif sampler is calculated in the same fashion with the exception that the beta distribution portion is dropped when using the site sampler since it is uninformative when the number of motifs is constant.

2.5 Fragmentation

It is possible that an alignment of motifs results in a descent map value that is not quite the maximal value. For instance, it could be possible, especially in the case of the motif sampler, that a good but not great alignment is found that is a shifted form of a better alignment.

Fragmentation allows the sampler to look at the current alignment and the surrounding positions to find which columns should be the ones used in the motif rather than just allowing for motifs that are continuous. The width of the field that can be checked is equal to 5 times the width of the current motif type. However, the field is generally greatly narrowed due to the limited flexibility of individual motif elements caused by overlapping ends of sequences or other motifs. Fragmentation goes through the columns that are currently used and picks out the worst column. Then a target column is chosen based on the possible column positions. The columns are associated with a weight to ensure that the closer a column is to the core of the motif element, the more likely it is to be chosen. The target column is sampled and the alignment and counts are reset accordingly. The process of fragmentation is examined every third time an alignment is found. It is possible that the worst column is still much better than the best available column, in which case the alignment will more than likely remain the same.

Fragmentation can alleviate the problem of shifting by picking the first column as the worst and replacing it with the last column or vice-versa. It will also deal with modeling more realistic regulatory sites where binding may or may not occur at contiguous sites. Fragmentation is covered in greater detail in (Liu et al., 1995).

2.6 Near Optimal Sampling

After the sampling steps have been iterated through a number of times (default is 10), each time running until a maximal map has not been found for the current run in the last 20 iterations or until the maximal number of iterations has been reached, the overall maximal map is stored. However, sampling of individual elements within the sequence is a random process, so it is possible to have some motif elements that have a low probability of being motif elements that are sampled in while others with a high probability of being motif elements are sampled out. Since such an alignment is suboptimal, a near optimal sampling routine is used to converge closer to the desired results.

First there is a preprocessing stage with near optimum sampling. In this stage, all of the positions that are motif elements in the maximal alignment are marked as "good" positions. All of the other possible motif starting positions are analyzed to see if their probability of being a motif element (according to equations 1, 2, 3 and 4) is above a certain threshold. If so, then they are marked as "good" positions, otherwise they are marked as "bad".

The sampling stage of near optimal sampling works exactly the same as in the site or motif sampler with one small difference. The exception is that only those positions which are marked "good" are possible starting positions. As a re-

sult, near optimal sampling is much faster than suboptimal sampling.

Near optimal sampling will run through for the maximum number of iterations (default is 500). A counter is used to indicate how many times each position gets sampled in as a motif element. This is used when the results are printed to find which positions occurred above a threshold percentage of the time.

2.7 Maximal MAP Calculation

Once near optimal sampling has completed, there are three different results that are available. The first is the maximum suboptimal alignment, the second is the maximum near optimal alignment and the third is the alignment of motif elements that occur over a threshold percentage of the time in near optimal sampling. If the flag is set on the command line to perform a maximization of the MAP, then the one alignment of the three that has the greatest map will be improved upon.

Maximal MAP calculation first sets the current alignment to the maximal alignment of the three found thus far. Then, in a single pass, beginning with the first position of the first sequence and ending with the last position of the last sequence, individual motif elements will be added in or taken out depending upon whether or not their inclusion improves the map value. The map must be calculated at each position in each sequence, a costly operation. What results is an alignment that has a map value that is as good, or better than the maximal MAP value found previously.

2.8 Expectation-Maximization

The final step when dealing with the site sampler is to use the maximal motif alignment as input into the Expectation Maximization (EM) Algorithm (Lawrence and Reilly, 1990). There are two steps to the EM Algorithm: the expectation step and the maximization step.

The expectation step calculates the expected number of residues of each type in each of the positions in the characteristic motif. This is accomplished by traversing through the sequences and finding the probability *P* that the current position is a motif element start site. Now assume that the motif does indeed start there. It will have a width of *J*. Assuming that the current position is position *i* of sequence *S*, the expected values can be updated using equation 5.

$$\varepsilon_{b,j} += P \tag{5}$$

where

$$j = 0..J; b = S_{i+j-1}$$

The maximization step is used to calculate the population frequencies according to equations 6 and 7.

$$P_{b,j} = \frac{\varepsilon_{b,j}}{N_m} \tag{6}$$

$$P_{b,0} = \frac{\varepsilon_{b,0}}{N_b} \tag{7}$$

For the above equations, N_m is the number of motifs and N_b is the number of background residues. The first time through, the probability for each of the positions has been set according to the maximum alignment found previously in the expectation maximization algorithm. Convergence of the algorithm is achieved once the population frequencies stabilize from iteration to iteration.

By using the expectation maximization algorithm, hypothesis testing can be used to find how informative a model is in describing a possible motif alignment. This is accomplished by comparing the log likelihood ratio estimates calculated for two different motif models using equation 8.

$$\log L = N_m \sum_{j=1}^{J} \sum_{b=A}^{T} P_{b,j} \log_e (P_{b,j}) + N_b \sum_{b=A}^{T} P_{b,0} \log_e P_{b,0} \tag{8}$$

Once expectation maximization if finished, a near optimal alignment results according to the user specified inputs. Tinkering with the inputs according to how subtle the alignment is will allow the sampler to detect an alignment closer to an optimal solution. In the same respect, if the alignment is not too subtle, the sampler will converge faster and will not need to try as many iterations or seed values.

Table I: Porin Data Results

S,	M	BEG	pre	MOTIF	post	END	PROB	T	DESCRIPTION	STRAND
1,	1	149	ffglv	DGLNFAVQYLG	knerd	159	0.28	F	OMPF_ECOLI OUTER	B7
1,	2	170	tarrs	NGDGVGGSISY	eyegf	180	0.96	F	OMPF_ECOLI OUTER	B8
1,	3	212	ngkka	EQWATGLKYDA	nniyl	222	1.00	F	OMPF_ECOLI OUTER	B10
1,	4	255	fankt	QDVLLVAQYQF	dfglr	265	0.99	F	OMPF_ECOLI OUTER	B12
1,	5	293	dvdlv	NYFEVGATYYF	nknms	303	0.98	F	OMPF_ECOLI OUTER	B14
1,	6	328	klgvg	SDDTVAVGIVY	qf	338	0.98	F	OMPF_ECOLI OUTER	B16
2,	1	37	sgttd	SGLEFGASFKA	hesvg	47	0.86	F	PORI_RHOCA PORIN.	B3
2,	2	57	gaetg	EDGTVFLSGAF	gkiem	67	0.99	F	PORI_RHOCA PORIN.	B4
2,	3	99	lddrg	GNDIPYLTGDE	rltae	109	0.12	F	PORI_RHOCA PORIN.	--
2,	4	159	aaytf	GNYTVGLGYEK	idspd	169	0.98	F	PORI_RHOCA PORIN.	B9
2,	5	182	lmadm	EQLELAAIAKF	gatnv	192	0.97	F	PORI_RHOCA PORIN.	B10
2,	6	261	tiddv	TYYGLGASYDL	gggas	271	0.18	F	PORI_RHOCA PORIN.	B14
3,	1	6	eisln	GYGRFGLQYVE	drgvg	16	0.31	F	porin - R. blastica	B1
3,	2	43	ttetd	QGVTFGAKLRM	qwddg	53	0.99	F	porin - R. blastica	B3
3,	3	88	gnvdt	AFDSVALTYDS	emgye	98	0.90	F	porin - R. blastica	--
3,	4	171	iaadw	SNDMISLAAAY	ttdag	181	0.99	F	porin - R. blastica	B9
3,	5	222	lstag	DQVTLYGNYAF	gattv	232	1.00	F	porin - R. blastica	B12
3,	6	263	dyqfa	EGVKVSGSVQS	gfane	273	0.97	F	porin - R. blastica	B15
3,	7	277	qsgfa	NETVADVGVRF	df	287	0.84	F	porin - R. blastica	B16

3 RESULTS

The results produced by the Bernoulli sampler were compared with known motif positions and results obtained using other samplers. The three main data sets used were a bacterial porins data set with known motif positions (Neuwald et al., 1995), dinucleotide-binding proteins whose motifs have been detected by a previous implementation of a Gibbs sampling algorithm (Liu et al., 1995) and a CRP binding data set with CRP binding sites that have been found using an implementation of the expectation maximization algorithm (Lawrence and Reilly, 1990).

3.1 Porin Dataset

The porin data consists of a set of 16 distantly related bacterial integral outer membrane proteins. A blast score of 185 has been performed to preprocess the data to remove similar sequences. The location of the predicted porins is given in Table I, while 16 known beta strands for three of these proteins is given in Table II.

For Table I, the first column represents the sequence number (S) followed by the motif number (M). The second and sixth columns are the motif beginning (BEG) and ending (END) start sites. The third (pre) and fifth (post) columns are the residues that flank the motifs (MOTIF) in the fourth column. The seventh column represents the probability of the motif belonging to the alignment (PROB). The eighth column indicates whether it is a forward (F) or

Table II: Known Alignment for Porin Beta Strands

Strand	OmpF	Rhodobactor Capsulatus	R. blastica
B1	8 - 23	1 - 15	1 - 16
B2	38 - 51	18 - 35	24 - 39
B3	54 - 65	39 - 47	46 - 57
B4	83 - 91	59 - 65	69 - 75
B5	94 - 100	68 - 74	78 - 84
B6	135 - 141	118 - 125	131 - 139
B7	150 - 161	127 - 135	141 - 150
B8	170 - 182	148 - 158	163 - 172
B9	184 - 195	160 - 171	174 - 183
B10	211 - 222	181 - 192	193 - 201
B11	224 - 235	194 - 206	205 - 214
B12	253 - 265	227 - 240	222 - 232
B13	268 - 281	242 - 255	234 - 245
B14	294 - 303	258 - 271	252 - 261
B15	306 - 315	274 - 285	264 - 274
B16	331 - 340 1 - 7	292 - 301	278 - 289

Table III: Known Dinluceotide Motif Positions

Motif (Seq, #)	Begin	End
1-1	195	220
2-1	23	48
4-1	6	31
5-1	6	31
1-1	8	33
8-1	41	66
8-2	179	204
11-1	468	493
12-1	3	28
14-1	5	30
21-1	298	323
21-2	348	373
22-1	11	36
22-2	222	247
24-1	6	31
27-1	215	240
28-1	24	49
29-1	148	173
41-1	162	187
41-2	204	229
42-1	7	32
42-2	383	408
44-1	6	31

Motif (Seq, #)	Begin	End
46-1	6	31
47-1	756	781
48-1	7	32
49-1	6	31
50-1	260	285
51-1	184	209
55-1	297	322
56-1	5	30
57-1	80	105
63-1	286	311
70-1	4	29
70-2	38	63
71-1	3	28
76-1	5	30
77-1	119	144
78-1	5	30
79-1	167	192
82-1	13	38
83-1	290	315
85-1	9	34
86-1	218	243
89-1	16	41

reverse complement (R) element. Column nine contains a description of the sequence and column 10 indicates which strand the element is contained within.

Of all the motifs found for the first three proteins, only two do not fit into one of the beta strands (see Table II). While only 17 of 48 motif elements have been found, the results are impressive considering that the known motif alignment has elements of varying length. Thus, it would be best to consider multiple motifs and run this data set again to sample more motif elements.

3.2 Dinucleotide Dataset

The dinucleotide is a set of 91 distantly related proteins that bind to one of three cofactors used by several different enzymes. The maximal alignment has been identified by an implementation of the Gibbs sampler using fragmentation (Liu et al., 1995). These sites are provided in Table III.

Fragmentation was used to obtain the maximal alignment given in Table IV. The columns that have an '*' beneath them are the chosen columns for the motif. Comparing this alignment with the previously calculated motif sites (Liu et al., 1995), it can be observed that 40 of the 45 motif elements have been chosen while only one additional motif element (79, 2) was found. An explanation for the discrepancy could be that the alignment calculated using the Bernoulli motif sampler is close to but not equal to the maximal alignment. Four of the five sites that differ are sampled in at frequencies just above the threshold with the previous sampler (Liu et al., 1995).

3.3 CRP Dataset

The CRP data contains a set of 18 proteins that have a CRP binding site. CRP is a dimeric DNA binding protein. Footprint sites have been reported (Stormo and Hartzell, III, 1989), as shown in Table V. The results using the site sampler with a width of 22, expectation maximization and no fragmentation are produced as shown in Table VI.

Comparing these results to the footprint sites in Table VI, it can be seen that 20 of 24 footprint sites are found. Three of the sites (those marked with an *) are sites that have been previously found using an expectation maximization approach (Lawrence and Reilly, 1990). The site marked by + is the only site that does not appear in the footprint sites or those sites previously sampled.

Table VI: Footprint Sites for CRP Data

Sequence	Footprint Sites
cole 1	17, 61
eco arabop	17, 55
eco bglrl	76
eco crp	63
eco cya	50
eco deop	7, 60
eco gale	42
eco ilvbpr	39

Sequence	Footprint Sites
eco lac	9, 80
eco male	14
eco malk	29, 61
eco ompa	48
eco tnaa	71
eco uxul	17
pbr-p4	53
trn9cat	1, 84

Table IV: Predicted Dinucleotide Locations

S,	M	BEG	pre	MOTIF	post	END	PROB	T	DESCRIPTION
1,	1	195	tqgst	CAVFGLGGVGLSVIMGCKAAGAARII	gvdin	220	1.00	F	ADHE_HORSE
2,	1	23	rsynk	ITVVGVGAVGMACAISILMKDLADEV	alvdv	48	1.00	F	LDHM_SQUAC
4,	1	6	agwsc	LVTGGGGFLGQRIICLLVEEKDLQEI	rvldk	31	1.00	F	3BHS_BOVIN
5,	1	6	malqq	FGLIGLAVMGENLALNIERNGFSLTV	ynrta	31	1.00	F	6PGD_SYNP7
6,	1	8	ankni	IFVAGLGGIGFDTSREIVKSGPKNLV	ildri	33	1.00	F	ADH1_DRONA
8,	1	41	ektpq	ICVVGSGPAGFYTAQHLLKHPQAHVD	iyekq	66	1.00	F	ADRO_HUMAN
8,	2	179	lscdt	AVILGQGNVALDVARILLTPPEHLEA	lllcq	204	1.00	F	ADRO_HUMAN
11,	1	468	dqvie	VFVIGVGGVGGALIEQIYRQQPWLKQ	khidl	493	1.00	F	AK1H_SERMA
14,	1	5	mqfd	YIIIGAGSAGNVLATRLTEDPNTSVL	lleag	30	1.00	F	BETA_ECOLI
21,	1	348	lvqrl	IGIEARGRIGRAVAVALHRASRALDV	adhrq	373	1.00	F	CRTI_APHSP
22,	1	11	egmgr	AVVIGAGLGGLAAAMRLGAKGYKVTV	vdrld	36	1.00	F	CRTI_RHOCA
22,	2	222	kfgvh	YAIGGVQAIADAMAKVITDQGGEMRL	ntevd	247	1.00	F	CRTI_RHOCA
24,	1	6	mtnir	VAIVGYGNLGRSVEKLIAKQPDMDLV	gifsr	31	1.00	F	DDH_CORGL
27,	1	215	mkkak	IAVQGIGNVGSYTVLNCEKLGGTVVA	maewc	240	1.00	F	DHE3_CLODI
28,	1	24	plplt	VGVLGSGHAGTALAAWFASRHVPTAL	wapad	49	1.00	F	DHNO_AGRT7
29,	1	148	fkgkp	ALIVGGGGTARTAIYVLRKWLGVSKI	yivnr	173	0.99	F	DHQA_ASPNI
41,	1	162	fglna	VVIGASNIVGRPMSMELLLAGCTTTV	thrft	187	1.00	F	FOLD_ECOLI
41,	2	204	lenad	LLIVAVGKPGFIPGDWIKEGAIVIDV	ginrl	229	0.98	F	FOLD_ECOLI
42,	1	7	tfqad	LAIVGAGGAGLRAAIAAAQANPNAKI	alisk	32	1.00	F	FRDA_ECOLI
44,	1	6	msrak	VGINGFGRIGRLVLRAAFLKNTVDVV	svndp	31	1.00	F	G3P_SCHMA
46,	1	6	mqpir	LGLVGYGKIAQDQHVPAINANPAFTL	vsvat	31	0.99	F	GAL_PSEFL
49,	1	6	mnqva	VVIGGGQTLGAFLCHGLAAEGYRVAV	vdiqs	31	1.00	F	GUTD_ECOLI
50,	1	260	svtvc	VGHLGGLDIAERDIARLRGLGRTVSD	siavr	285	1.00	F	HDNO_ARTOX
51,	1	184	lssqk	TVIIGAGKMACLLVKHLLAKGATDIT	ivnrs	209	1.00	F	HEM1_SYNY3
55,	1	297	dvqsd	IVAQGFGSLGLMTSILVTPDGKTFES	eaahg	322	0.99	F	IDHP_YEAST
56,	1	5	mslr	IGVIGTGAIGKEHINRITNKLSGAEI	vavtd	30	1.00	F	IDH_BACSU
57,	1	80	fkndt	FALIGYGSQGYGQGLNLRDNGLNVII	gvrkd	105	1.00	F	ILV5_YEAST
63,	1	286	lsdht	VLFQGAGEAALGIANLIVMAMEKEGV	skeaa	311	1.00	F	MAOX_ANAPL
70,	1	4	mtd	IAFLGLGNMGGPMAANLLKAGHRVNV	fdlqp	29	1.00	F	MMSB_PSEAE
71,	1	3	mk	ALHFGAGNIGRGFIGKLLADAGIQLT	fadvn	28	1.00	F	MTLD_ECOLI
76,	1	5	msnt	IVVVGAGVIGLTSALLLSKNKGNKIT	vvakh	30	1.00	F	OXDA_FUSSO
77,	1	119	lydrt	VGIVGVGNVGRRLQARLEALGIKTLL	cdppr	144	1.00	F	PDXB_ECOLI
78,	1	5	mktq	VAIIGAGPSGLLLGQLLHKAGIDNVI	lerqt	30	1.00	F	PHHY_PSEFL
79,	1	167	vppak	VMVIGAGVAGLAAIGAANSLGAIVRA	fdtrp	192	1.00	F	PNTA_ECOLI
79,	2	463	aisgi	IVVGALLQIGQGGWVSFLSFIAVLIA	sinif	488	0.96	F	PNTA_ECOLI
82,	1	13	aesyt	LGFIGAGKMAESIARGAVRSGVLPPS	rirta	38	1.00	F	PROC_SOYBN
85,	1	9	lkeyv	IVSGASGFIGKHLLEALKKSGISVVA	itrdv	34	1.00	F	RFBJ_SALSP
86,	1	218	lagkv	AVVAGYGDVGKGSAASLKAFGSRVIV	teidp	243	1.00	F	SAHH_CAEEL
89,	1	16	katkv	IGIIGLGDMGLLYANKFTDAGWSVIC	cdree	41	1.00	F	TYR1_YEAST
				******* ** * * * *					

Seed = 838822581 Difference of Logs of Maps = 100.53212

Table V: Predicted CRP Data Set

```
 S,  M   BEG  pre          MOTIF         post    END   PROB T DESCRIPTION
 1,  1    38 gcgct ATTCTCGCCCGATGCCACAAAA accag    17   0.31 R cole1
 1,  2    61 actgt TTTTTTGATCGTTTTCACAAAA atgga    82   0.34 F cole1
 2,  1    38 ttttc TGCCGTGATTATAGACACTTTT gttac    17   0.05 R ecoarabop
 2,  2    55 attga TTATTTGCACGGCGTCACACTT tgcta    76   0.45 F ecoarabop
 3,  1    76 ttaat AACTGTGAGCATGGTCATATTT ttatc    97   0.49 F ecobgirl
 4,  1    63 tgcat GTATGCAAAGGACGTCACATTA ccgtg    84   0.46 F ecocrp
 5,  1    71 gtcta AAACGTGATCAATTTAACACCT tgctg    50   0.49 R ecocya
 6,  1    28 cactg TAATGCGATCTGGTTCAAATAA ttcac     7   0.26 R ecodaop
 6,  2    81 caaca CACTTCGATACACATCACAATT aagga    60   0.25 R ecodaop
*7,  1    24 aattc TTGTGTAAACGATTCCACTAAT ttatt    45   0.33 F ecogale
+8,  1    15 gggtt TTTTGTTATCTGCAATTCAGTA caaaa    36   0.10 F ecoilvbpr
 8,  2    39 gtaca AAACGTGATCAACCCCTCAATT ttccc    60   0.37 F ecoilvbpr
 9,  1    30 gccta ATGAGTGAGCTAACTCACATTA attgc     9   0.36 R ecolac
*9,  2    94 ttgtt ATCCGCTCACAATTCCACACAA catac    73   0.28 R ecolac
10,  1    35 tcgct TTGTGTGATCTCTGTTACAGAA ttggc    14   0.50 R ecomale
11,  1    29 acggc TTCTGTGAACTAAACCGAGGTC atgta    50   0.08 F ecomalk
11,  2    82 acgat TTTTGCAAGCAACATCACGAAA ttcct    61   0.47 R ecomalk
12,  1    62 tgtct GAATTTGCACTGTGTCACAATT ccaaa    41   0.35 R ecomalt
13,  1    48 ttcat ATGCCTGACGGAGTTCACACTT gtaag    69   0.49 F ecoompa
14,  1    71 cgaac GATTGTGATTCGATTCACATTT aaaca    92   0.48 F ecotnaa
15,  1    38 tctaa TTGGGTTAACCACATCACAACA atttc    17   0.49 R ecouxu1
16,  1    53 atatg CGGTGTGAAATACCGCACAGAT gcgta    74   0.48 F pbr322
17,  1   105       CGTGCCGATCAACGTCTCATTT tcgcc    84   0.99 R trn9cat
*18,  1    97 aggat ATGTGCGACCACTCACAAATTA acttt    76   0.26 R (tdr)
                    **********************
```

```
Seed = 838740876
```

4 DISCUSSION

Results for the porin, dinucleotide, and CRP datasets indicate that the Bernoulli sampler approach to regulatory site prediction produces viable results. The concepts outlined in this technical report have been expanded into a Gibbs Recursive Sampler (Thompson et al., 2003) that implements multiple motif detection and model specific information such as reverse complementary regions. Detection of motifs within sequences will continue to be a topic of interest as more and more disparate genome sequences become available.

ACKNOWLEDGEMENTS

The work reported here was originally reported as ER's Master of Science Thesis at Rennselaer Polytechnic Institute in May, 1996. It is published here in slightly modified form as a technical report in order to provide a widespread public dissemination of the work. ER is currently supported by NIH-NCRR grant P20RR16481 and NIH-NIEHS grant P30ES014443. The contents of this manuscript are solely the responsibility of the authors and do not represent the official views of NCRR, NIEHS, or NIH.

REFERENCES

Lawrence,C.E. et al. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262, 208-214.

Lawrence,C.E. and Reilly,A.A. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7, 41-51.

Liu,J.S., Neuwald,A.F. and Lawrence,C.E. (1995) Bayesian Models for Multiple Sequence Alignment and Gibbs Sampling Strategies. *Journal of the American Statistical Association*, 90, 1156-1171.

Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.*, 4, 1618-1632.

Stormo,G.D. and Hartzell,G.W., III. (1989) Identifying protein-binding sites from unaligned DNA fragments 2. *Proc. Natl. Acad. Sci. U. S. A*, 86, 1183-1187.

Thompson,W., Rouchka,E.C. and Lawrence,C.E. (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res.*, 31, 3580-3585.