Searching the Human Genome for Snail and Slug With DNA@Home

Kristopher Zarns and Travis Desell
Department of Computer Science
University of North Dakota
Grand Forks, North Dakota 58202-9015
Email: kzarns@gmail.com, tdesell@cs.und.edu

Sergei Nechaev and Archana Dhasarathy
Department of Basic Sciences
University of North Dakota
Grand Forks, North Dakota 58202-9061
Email: archana.dhasarathy@med.und.edu,
sergei.nechaev@med.und.edu

Abstract—**DNA@Home is a volunteer computing project that aims to use Gibbs Sampling for the identification and location of DNA control signals on full genome-scale datasets. A fault tolerant and asynchronous implementation of Gibbs sampling using the Berkeley Open Infrastructure for Network Computing (BOINC) was used to identify the location of binding sites of the SNAI1 (Snail) and SNAI2 (Slug) transcription factors across the human genome. A set of genes that are regulated by Slug but not Snail, and a set of genes that are regulated by Snail but not Slug were used to provide two datasets with known motifs. These datasets contained up to 994 DNA sequences, which to our knowledge is largest scale use of Gibbs sampling for discovery of binding sites. These genomic regions were analyzed using datasets containing various numbers of intergenomic regions. 1,000 parallel sampling walks were used to search for the presence of 1, 2 or 3 possible motifs. These runs were performed over a period of two months using over 1,500 volunteered computing hosts, and generated over 2.2 Terabytes of sampling data. High performance computing resources were used for post processing of the Gibbs Sampler output. This paper presents how intra- and interwalk analyses can aid in determining overall walk convergence. The results were validated against current biological knowledge of the Snail and Slug promoter regions, and present potential avenues for further biological study.**

I. Introduction

This paper presents an expansion on previous work done with DNA@Home [1] volunteer computing project. The DNA@Home project implements an asynchronous version of the Gibbs Sampling algorithm which performs parallel sampling walks using volunteer computing. DNA@home uses the Berkley Open Infrastructure for Network Computing (BOINC) [2] to provide massively scalable computing power to search for transcription factor binding sites (or *motifs*) in large datasets.

DNA@Home performed parallel Gibbs sampling runs with various parameters and data sets of varying sizes over a period of two months, aimed at identifying motifs related to the SNAI1 (Snail) and SNAI2 (Slug) genes. Each run had 1,000 parallel sampling walks, and the largest data sets contained 994 regions of DNA. This resulted in over 2.2 Terabytes of sampling data, which was analyzed using high performance computing resources. To our knowledge, this is the largest scale use of Gibbs sampling for de novo transcription factor binding site discovery.

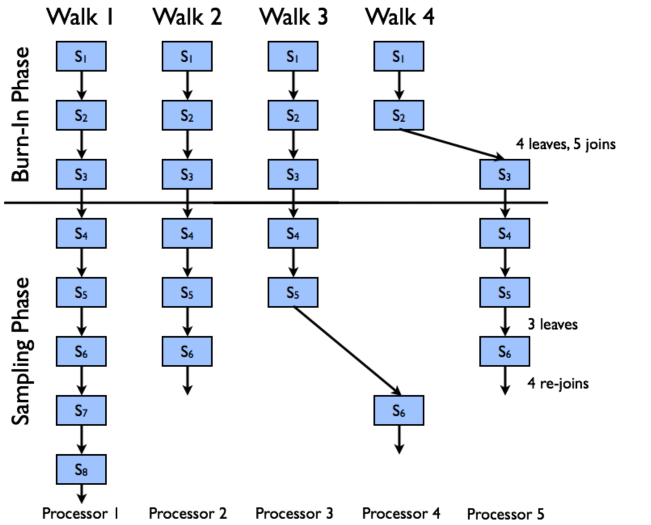


Fig. 1. This figure presents how DNA@Home performs parallel Gibbs sampling. Arrows represent *workunits*, or volunteer computing tasks, where hosts receive an initial state with depth x, Sx, and report a final state with depth y, Sy. Workunits have fixed walk lengths, in this case 1, however the runs described in this work had walk lengths of 10,000. When a walk completes its burn-in period, samples are taken. Processors can join and leave, restarting from walks of previously left processors.

A. The Gibbs Sampler

The Gibbs Sampler used by DNA@Home is set up to execute many walks in parallel. Each walk represents a run of the sampler with a different initial starting position. As a walk progresses it takes a number of steps. Each step is a move in the Markov chain. After a certain number of steps, a super-step, the resulting distribution is output, previous steps are forgotten and the current position is used to restart the walk (see Figure 1).

A major issue with Gibbs sampling is that the initial randomly chosen starting point will bias the result of the sampler. To overcome this bias, a certain number of steps should be discarded. The number of steps needed to overcome the initial bias is called the *burn-in*. The burn-in might be significantly larger than the size of each super-step, and is dependent on the dataset and parameters to the Gibbs sampler (which include how many motifs are being searched for, what type of motifs they are, and how many nucleotides long the motifs are). The burn in time needed is similar for each walk though there are some outlying walks which do not converge as quickly. Without determining a burn-in it is unclear how many steps need to be taken before a valid sample set is generated. The number of steps per super-step is called the step-size. A larger step-size provides a better sample and will make burn-in easier to detect. A smaller step size requires less computation.

It is possible to determine the required burn-in for an individual walk through convergence detection algorithms. Ideally the distribution generated at each super-step will approximate a stationary distribution, meaning that additional steps will not significantly alter the distribution. After the burn-in period is complete the sampler should converge. It is possible that some datasets and input parameters will not converge. If convergence is detected, the burn-in period is complete and the next result of the sampler is valid. No further steps are required and computation can cease.

While convergence is a useful tool for determining when a single walk has completed, it does not guarantee that the walk has a good distribution of samples that completely represents likely motifs, as it could have converged to a local optima. By using multiple walks, it is possible to discover additional motifs by settling into different local maxima, and provide a more global picture of the sampling space. Results show that analyzing distances between parallel walks can provide a good picture of whether or not the walks have converged, and if they have converged to local optima or global optima.

B. Biological Significance of Snail and Slug Motifs

The Snail family of Zinc-finger transcription factors, SNAI1, SNAI2, and SNAI3 are highly conserved across vertebrates [3]. The Snail1 (Snail) and Snail2 (Slug) transcription factors bind to the subset of E-box motifs (CAGGTG/CACCTG) present at gene promoters, and recruit co-factor complexes to alter gene expression [3]. By changing expression of genes such as E-cadherin, which helps cells adhere to one another, Snail and Slug trigger loss of cell-cell adhesion, and hence cellular movement. This caused cells to change their shape and migrate, a phenomenon called Epithelial-to-Mesenchymal Transition or EMT [4]. EMT is essential for proper embryonic development, but is also responsible for tumor invasion and metastasis [4]. While Snail and Slug have several functions in common, they yet appear to have distinct gene targets (Moreno-Bueno *et al.* , Dhasarathy *et al.*) [5] [6], which potentially has implications in their distinct roles at different stages of cancer metastasis. As yet, the molecular basis for the distinct regulation and binding affinity of downstream target genes by Snail and Slug are unknown.

II. Related Work

Lawrence *et al.* [7] provides an in depth discussion of how to apply the Gibbs Sampling algorithm to the motif finding problem. The differences in convergence rate between runs due to differing random start sites is described. Lawrence claims that the power of the Gibbs sampler increases when used with more sequences because the pattern model is improved by adding more data.

TABLE I
Project Dataset Comparison

Project	Gibbs	Sequences	Length	Runs	Motifs	Width
DNA@Home	Yes	994	1000	1000	1,2,3	6
ChIPMunk	No	10,000	1300	1	1	21
Bioprospector	Yes	60	800	250	1	8
PRIORITY	Yes	34	1300	5	0,1	8
W-AlignACE	Yes	176	800	5	2	10

A. Dataset Size

Table I relates the data set analyzed by DNA@Home to other recent work (note that ChIPMunk is not a Gibbs sampler). As described in Kulakovskiy *et al.* [8] many of the existing Gibbs Sampling motif discovery tools are not suited to processing the wealth of data provided by Next Generation Sequencing (NGS) data sources. Techniques like Chromatin Immunoprecipitation combined with sequencing (ChIP-Seq) determine where proteins bind on the genome, and can provide thousands of sequences with more than 1000 base pairs in each sequence. The size of the problem set and the efficiency of Gibbs sampling causes many approaches to reduce the dataset significantly so that it can be run on the available pool of hardware in a reasonable amount of time. DNA@Home overcomes these challenges through massive parallelism and volunteer computing. Kulakovskiy compares the efficiency of Weeder Pavesi *et al.* [9], Gibbs Sampler Lawrence *et al.* [7] and MEME Suite Bailey *et al.* [10]. Kulakovskiy also discusses the efficiency of cERMIT [11] another algorithm which takes advantage of the properties of ChIPSeq and HMS [12] which reduces the stochastic sampling set size and selects the alignment variable chauvinistically.

Kulakovskiy provides ChIPMunk which is suited for work on significantly larger scales than many of the previous Gibbs Sampling algorithms. However, ChIPMunk is not a Gibbs Sampling algorithm, it is a greedy optimization using several heuristics. ChIPMunk was meant to address the increased problem space created by the use of ChIPSeq data. ChipMunk takes advantage of several properties of ChIPSeq data to create hueristics which are specific to that type of dataset.

Narlikar *et al.* [13] provides PRIORITY, a Gibbs Sampling algorithm which takes advantage of knowledge of transcription factor binding sites to use an informative prior probability. PRIORITY is shown to be an improvement over AlignACE [14], MEME [10], MDscan [15] a weighted position matrix approach, and CONVERGE [16].

Similarly to Kulakovskiy and Narlikar, Che *et al.* [17] provides BEST which compares multiple motif finding programs: AlignACE [14], Biorprospector [18] and MEME [10].

Liu *et al.* [18] discusses the bioprospector a system for discovering motifs using Gibbs sampling. Liu implements the Gibbs sampler and describes methods used to validate that bioprospector found meaningful motifs. Bioprospector was run on 60 sequences of 800 base pairs. Liu discusses methods to improve on Lawrence's Gibbs Sampler by replacing the mixture model with a threshold sampler to account for relationships among input sequences. A third order Markov background model is used to take advantage of the larger dataset.

Chen *et al.* [19] provide W-AlignACE, a Gibbs Sampling method using an improved positional weight matrix. Chen compares W-AlignACE to the Gibbs Samplers AlignACE [14] and MDSCan [15].

B. Burn-In and Other Problems in Gibbs Sampling

There are many methods that can be used to determine the burn-in period and convergence rate of a Markov Chain Monte Carlo (MCMC) algorithm. Brooks *et al.* [20] has identified the following classes of methods for assessing convergence and determining burn-in in Gibbs Sampling: variance ratio, spectral, empirical kernel-based, regeneration and coupling, and semi-empirical methods that use Eigen Value bounds. The Kolmogorov Smirnov 2 sample statistic is a spectral method which can be used to test the null hypothesis of stationarity.

Jensen *et al.* [21] discusses methods for finding motifs using Gibbs sampling when multiple motifs are present in the data set. Jensen uses a annealing approach to shifting the sampler to avoid being stuck in local maxima. This shifting approach uses a heat function which decreases the size of shifts over time.

Woodward *et al.* [22] discusses problems with slow mixing and poor or nonexistent convergence of Gibbs sampling when used to detect motifs in genomic data if multiple motifs are present. In woodwards case, convergence rate decreased as the length of DNA sample increased. While DNA@Home used a average sequence length of 1000 base pairs, the number of sequences per dataset was varried. DNA@Home found that as predicted in Lawrence *et al.* a larger size dataset increases the rate of convergence [7].

III. Implementation

A. Generating the Dataset

To generate the dataset, genes from a list generated by global gene expression microarray analyses by Dhasarathy et al. [6] were used. In this experiment, Snail and Slug were independently expressed in human MCF-7 breast cancer cells, in a time course over 4 days. The genes that were uniquely regulated by Snail or Slug (both up or down) over the four days were compiled in a list, with overlaps being merged, thus generating two lists of genes unique to Snail or Slug regulation. The gene sequence of each of these genes was obtained at an interval of -500 to +500 base pairs from the transcription start site from the UCSC human genome browser (hg19) [23].

The initial sequence dataset used to generate the ranked list of genes for this experiment were taken from the Encode project at UCSC [23]. The track used was wgEncodeOpenChromChipMcf7Pol2SerumstimRawDataRep1. This track was chosen due to prior work in the workflow development which centered on Snail and Slug representation with regards to RNA Polymerase II (Pol II) binding. The entire Snail dataset contains 1,422 genes, while the entire Slug dataset only contains 412. Three different size FASTA files were generated for each dataset: small, medium, and large. To generate the FASTA files successively smaller sets of genes are used. The different data sets are illustrated in Table II.

TABLE II
DATASET CONFIGURATION AND BURN-IN INFORMATION

Motifs	Dataset	Size	Intervals	Genes	Burn In	Stable
1	Snail	Large	1442	994	<20,000	Yes
1	Snail	Medium	1442	100	<20,000	Yes
1	Snail	Small	1442	10	<20,000	No
1	Slug	Large	412	372	<20,000	Yes
1	Slug	Medium	412	99	<20,000	Yes
1	Slug	Small	412	10	<20,000	No
2	Snail	Large	1442	994	<20,000	Yes
2	Snail	Medium	1442	100	130,000	Yes
2	Snail	Small	1442	10	N/A	No
2	Slug	Large	412	372	60,000	Yes
2	Slug	Medium	412	99	200,000	Yes
2	Slug	Small	412	10	N/A	No
3	Snail	Large	1442	994	<20,000	Yes
3	Snail	Medium	1442	100	<20,000	Yes
3	Snail	Small	1442	10	N/A	No
3	Slug	Large	412	372	<20,000	Yes
3	Slug	Medium	412	99	<20,000	No
3	Slug	Small	412	10	N/A	No

The Gibbs Sampler takes a FASTA file containing the sequences that define the genes across the generated intervals. To generate the FASTA files a workflow was developed which takes sequenced data in FASTQ format and assigns each individual sequence a unique coordinate based on the sequence of human hg19 genome annotation using Bowtie [24], converts it to BedGraph format for display and verification, then associates the display data with gene intervals, filters that set for overlapping genes and ranks the genes based on the number of matching reads. The BowtieToBedGraph conversion software and CPPMatch ranking software was provided by Adam Burkholder of the National Institute of Health.

B. Gibbs Sampling Configuration

The Gibbs Sampler was configured to search for 1, 2, or 3 motifs. Searching for more motifs per run was done to examine how the number of motifs present in the dataset effects on how quickly the parallel Gibbs sampling walks converge. There are six datasets to run, 3 for Slug and 3 for Snail. Those runs represent the different number of genes used in each dataset. The Snail and Slug datasets were run independently. For each run of the Gibbs Sampler 1000 independent walks were created. Each walk had the same dataset and started with a random initial starting samples and a different random seed. Each walk had a super-step size of 10,000 steps. After each super-step the resulting empirical distribution was stored for off-line calculation of the convergence rate and a new super-step was started from the current position with a new random seed.

C. *Checking for Convergence*

The Kolmogorov-Smirnov 2 sample statistic is used to test the null hypothesis of stationary distribution. Each time a walk is restarted a sample is reported for the last super-step of the walk. That sample represents the empirical distribution of that period. The Kolmogorov-Smirnov 2 sample test generates two values, a maximum distance between distributions and a probability that two distributions are generated from the same source distribution. Kolmogorov-Smirnov sorts the sample to create the distribution function and then compares the distribution with the previous distribution.

The Kolmogorov-Smirnov 2 sample test is appropriate for testing Gibbs Sampling convergence for several reasons. The test is non-parametric, it doesn't assume a particular known distribution. This is an advantage because the empirical motif distribution is unknown and unlikely to fit a common probability distribution. Transformations of the values being tested will not affect the result so using larger super-step sizes will make the results more accurate without distorting them.

IV. Results

A. *Gathering Results with DNA@Home*

The results for this work were gathered over a period of approximately 2 months using the University of North Dakota's Citizen Science Grid, of which DNA@Home is a subproject. The number of simultaneously volunteered hosts participating in the project averaged around 1,650 during this period. Near the end of this period, the BOINC *Charity Team* selected the project for an event, which resulted in a burst of an additional 400-500 compute hosts in February. As of March 2015, DNA@Home and the Citizen Science grid has had over 1,500 users provide over 4,100 compute hosts for the project. In total, 18 runs were made looking for one to three of motifs using Snail and Slug datasets of small, medium and large sizes (see Table II). These runs on DNA@Home generated over 2.2 Terabytes of sampling data with which the following analyses were generated. Convergence rates for individual walks are examined in Section IV-B, convergence of the entire parallel sampling walks is discussed in Section IV-C and a discussion and validation of the motifs found is presented in Section IV-D.

B. *Intrawalk Analysis and Burn-In Detection*

The burn-in period was well defined for all runs that converged. As illustrated by Figure 2, the small dataset converged for the case of 1 motif for both datasets. However, the probability sample standard deviation remained around 10% so those results were marked as unstable. Runs with 2 or 3 motifs did not converge for the small datasets. Their probability consistently hovers around 20% for all of the runs with a probability sample standard deviation of around 35%. Figures for the remaining small runs are not included. The number of motifs searched for affects the rate of convergence. For 1 and 3 motifs all of the medium and large runs converged by 20,000 steps. The 2 motif runs show that using more genes improves the rate of convergence. While this may seem counterintuitive, this is in agreement with the claims of Lawrence *et al.* [7], that convergence rates of Gibbs sampling increase with more sequences.

1) Analysis of the 1 Motif Runs: Figure 2 shows a comparison of the 1 motif results for Snail and Slug. The small Slug datasets both show signs of convergence. However the high probability sample standard deviation draws the quality of this data into question. The consistent presence of near zero probabilities also suggests that these results are not stable. The Medium dataset satisfies burn in and converges in under 20,000 steps. The minimal probability sample standard deviation and consistently high minimum probability suggest that all of the walks have converged.

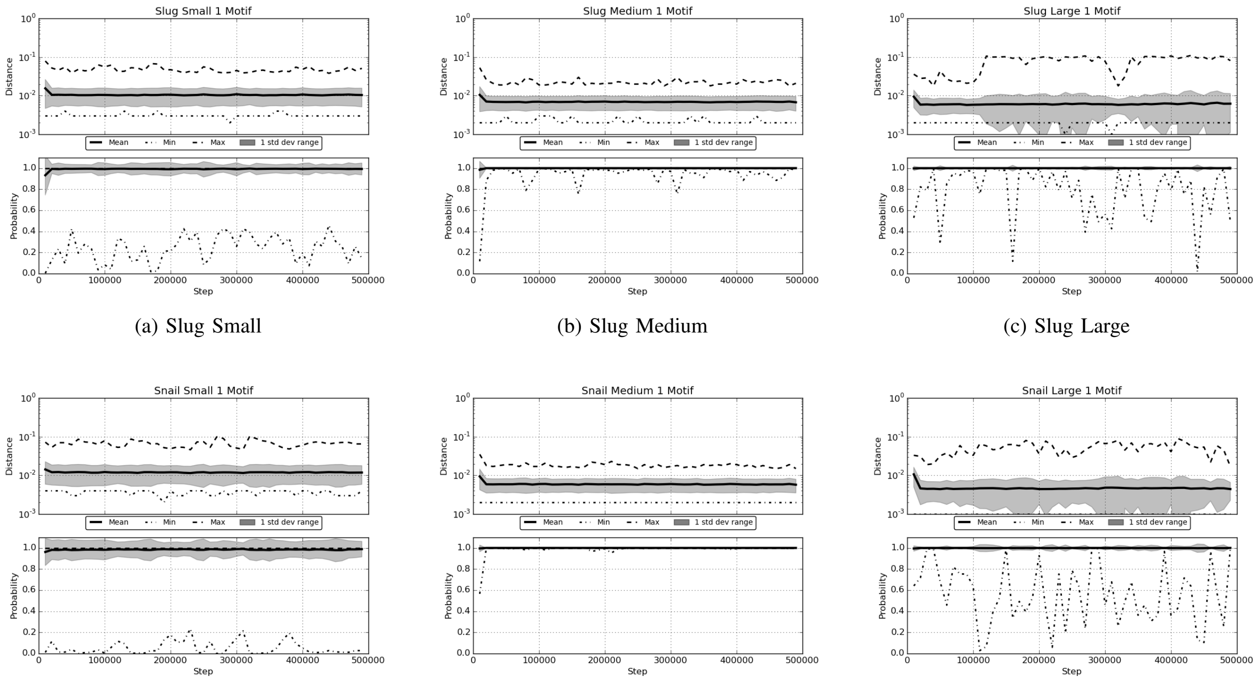
2) Analysis of the 2 Motifs Runs: Figure 3 shows the results from searching for 2 motifs at once. This shows that using a larger dataset improves the rate at which the walks converge. In both the Slug and Snail 2 motif medium cases, the walks do not immediately converge. Instead of the convergence seen in the first 20,000 steps in the other results for all numbers of motifs, this data shows that while the average probability of convergence is very high, the sample standard deviation is not reduced until much later. In the case of the Slug medium data the standard deviation isn't reduced until 200,000 steps. The Snail dataset sees the reduced standard deviation at 130,000 steps. In both cases, moving to the large dataset is an improvement.

3) Analysis of the 3 motifs Runs: Figure 4, shows that for Slug, while the medium size dataset converges quickly at around 40,000 steps, the stability of that convergence is brought into question by the fluctuating standard deviation. Again, using the large dataset for Slug improves the quality of the result.

C. *Interwalk Analysis*

Convergence rates for the parallel sampling walks as a whole was tested, which proved to be extremely computationally expensive. The sampling data sets ranged from around 20 GB for the runs with 1 motif on the small number of sequences, to over 500 GB for the runs with 3 motifs on the large number of sequences. In order to compare the distance between each walk at every super step, a parallel analysis tool was developed using C++ and the Message Passing Interface (MPI) to utilize high performance computing resources. These results were gathered using a Beowulf HPC cluster with 32 dual quad-core compute nodes (for a total of 256 processing cores). Each compute node has 64GBs of 1600MHz RAM, two mirrored RAID 146GB 15K RPM SAS drives, two quad-core E5-2643 Intel processors which operate at 3.3Ghz, and run the Red Hat Enterprise Linux (RHEL) 6.2 operating system. All 32 nodes within the cluster are linked by a private 56 gigabit (Gb) InfiniBand (IB) FDR 1-to-1 network. The code was compiled and run using MVAPICH2-x [25], to allow highly optimized use of this network infrastructure.

Even utilizing a HPC cluster and MPI, randomized sampling was required in order to calculate these results in a reasonable



(a) Slug Small

(b) Slug Medium

(c) Slug Large

(d) Snail Small

(e) Snail Medium

(f) Snail Large

Fig. 2. 1 Motif Kolmogorov-Smirnov Analysis after Burn-In. Top row: Slug shows improved convergence rate as the dataset size increases. Bottom row: Snail similarly converges sooner for larger datasets. In all of the Kolmogorov-Smirnov graphs, the top subgraph represents a y-log view of the average largest difference between super steps. The solid line is the mean, the dash-dot line is the minimum, the dashed line is the maximum, and the shaded region is the 1st standard deviation. The lower subplot shares the same legend however y values now range between 0 and 1. The lower subplot represents the probability that the current super step was generated from the same distribution as the previous super step

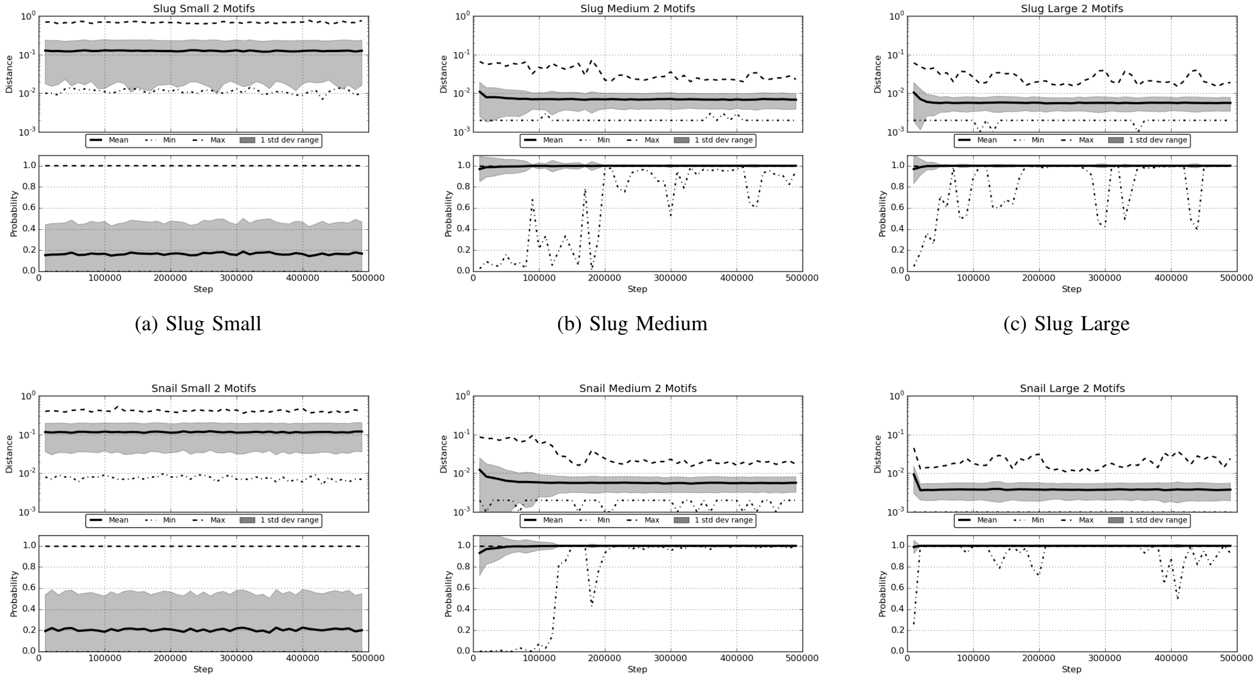
amount of time. Figures 5, 6 and 7 display the minimum, average, median and maximum distance between each walk in a random sample of 100 walks at each super step and were generated over a period of two days using the HPC cluster. The distance between any two walks was calculated as the average difference in the number of samples at each position within the sequences for each motif.

Similar to the interwalk comparison, runs with two motifs take significantly longer to converge than those with one or three motifs, which essentially converge within the first subwalk. Also, comparing the interwalk distance of the parallel sampling walk provides another strong measure with which to determine if the individual walks have converged to different local optima or if there is a consistent global optimum across all walks. For runs with high maximum distances and low averages and median distance, groups of walks would have converged to different regions. For runs with high maximum, median and average distances, walks would have converged to many different regions without grouping together. For runs with low maximum, average and median distances, all the walks grouped to a similar region. In general, for runs with low average and median distances, we would have generated enough parallel walks to get an appropriate sampling across all possible optima, while for those with high median and averages, either more sampling walks or motifs would be required to make sure the all regions of the search space are being sampled correctly (*e.g.*, the large Slug and Snail data sets with 1 motif).

D. Motif Validation and Analysis

The top ten motifs for Snail and Slug from each walk in the large datasets that were represented in greater than 10% of walks, and that occurred with greater than 10% frequency and which contained the Snail or Slug binding site sequence (also known as the 'E-Box' sequence, CAGGTG or CACCTG) within the combination of the reported motif and its left and right neighbors were examined. Tables VI, VII, VIII, III, IV, and V display the walk number, percentage of time the position was sampled, the gene name, starting nucleotide location, 5 nucleotides before the motif, the motif in capitalized letters, 5 nucleotides after the motif, ending nucleotide location and if it contained the E-Box. The gene name also includes the chromosome it was found in (chr2 is chromosome 2, for example), and the start and end position of the gene sequence in that chromosome. If multiple motifs were found for the sequence within that distribution at sample percentages over the minimum then multiple lines are returned for that sequence. If none of the motifs for a sequence overcame the minimum percentage then no motif was reported for the sequence.

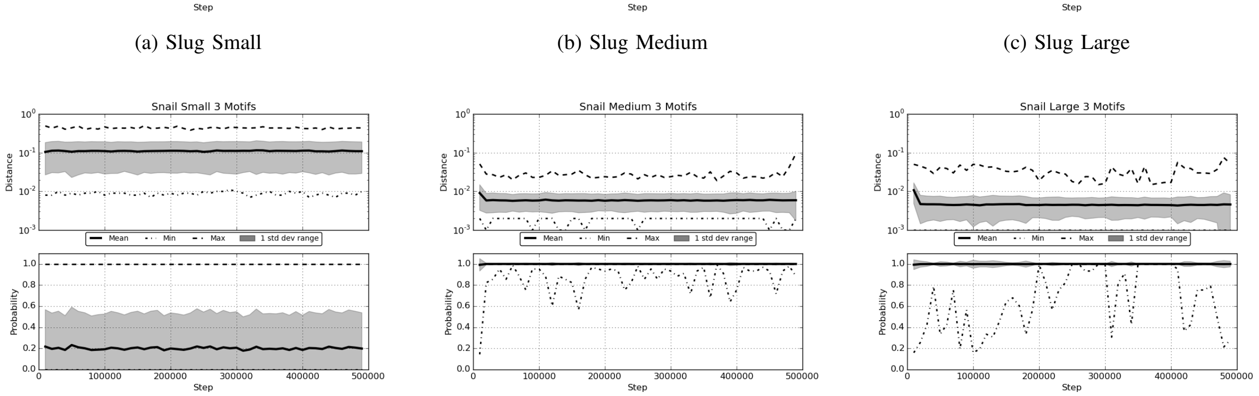
Several of the genes that had the E-box in their promoter regions were previously known targets or predicted targets.



(a) Slug Small (b) Slug Medium (c) Slug Large

(d) Snail Small (e) Snail Medium (f) Snail Large

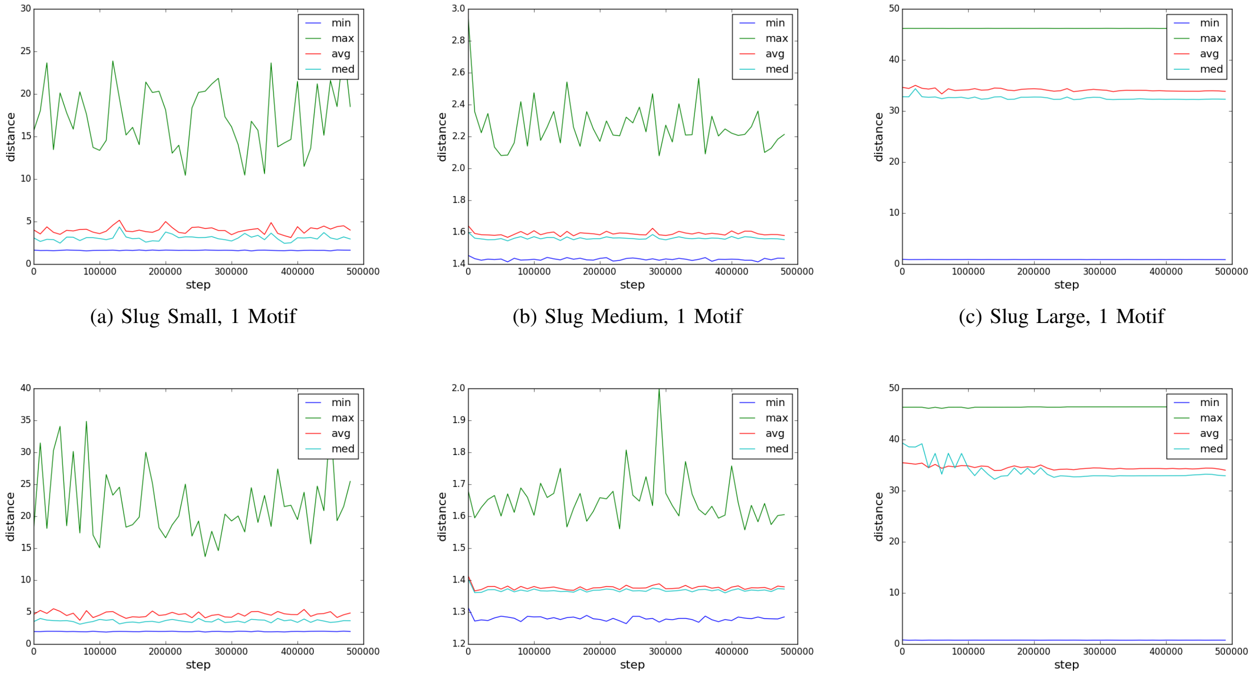
Fig. 3. 2 Motifs Kolmogrov-Smirnov Analysis. Top row: Slug shows instability for the small dataset and slower convergence of the large dataset vs the 1 motif runs. Bottom row: Snail also shows instability for the small dataset however Snail converges more quickly than Slug. The 2 motif runs do not converge as quickly as the 1 or 3 motif runs. However once converged the 2 motif runs on the large Snail dataset do not show the repeated low minimums in the probability section that the other motif groupings show.



(a) Slug Small (b) Slug Medium (c) Slug Large

(d) Snail Small (e) Snail Medium (f) Snail Large

Fig. 4. 3 Motifs Komogorov-Smirnov Analysis. Top row: Slug is unstable for the small dataset. The rate of convergence is similar to the 1 motif results for the medium and large datasets. Bottom row: Snail is unstable for the small dataset and converges sooner than Slug for the Medium and Large datasets.



(a) Slug Small, 1 Motif

(b) Slug Medium, 1 Motif

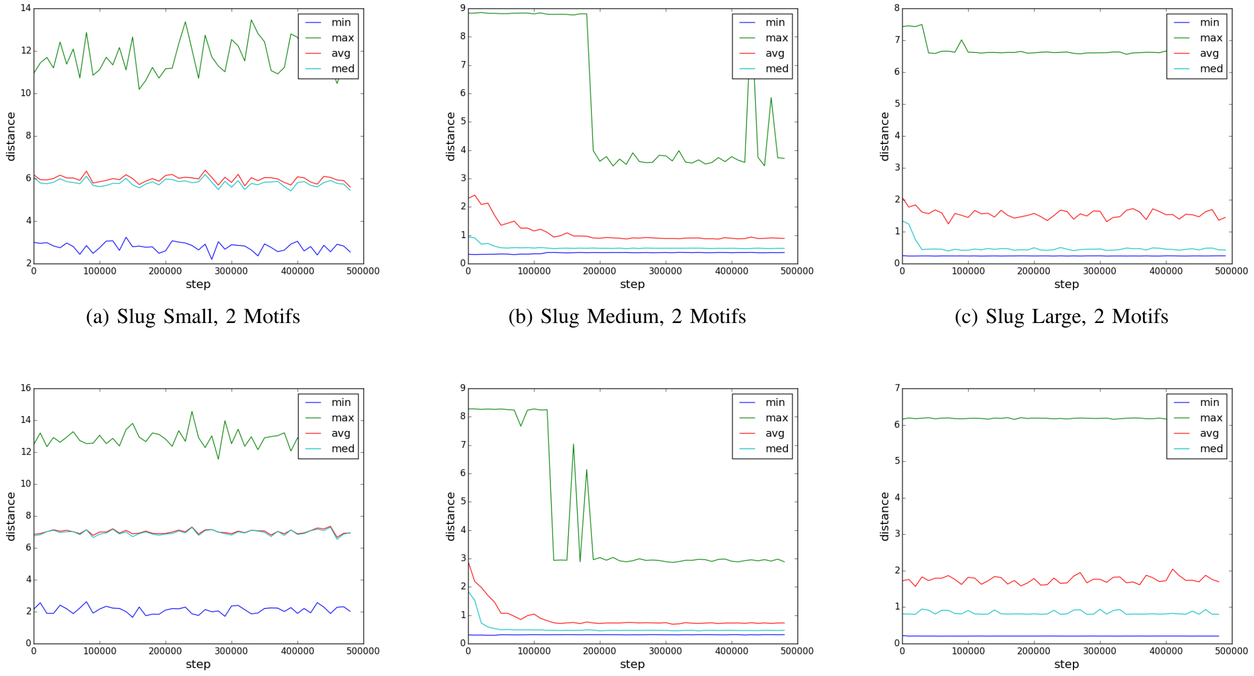
(c) Slug Large, 1 Motif

(d) Snail Small, 1 Motif

(e) Snail Medium, 1 Motif

(f) Snail Large, 1 Motif

Fig. 5. This figure presents the min, average, median and max distances between a random sample of 100 walks after every 10,000 steps in the walk for the 1 motif runs. Interestingly, for a medium number of intergenomic regions, the distances between the walks are the smallest. For the small set, the average and median distances stay low, but the high maximum distances suggest some instability. For the large set, it becomes obvious that one motif is not sufficient, given the consistently high average and maximum distance between walks.



(a) Slug Small, 2 Motifs

(b) Slug Medium, 2 Motifs

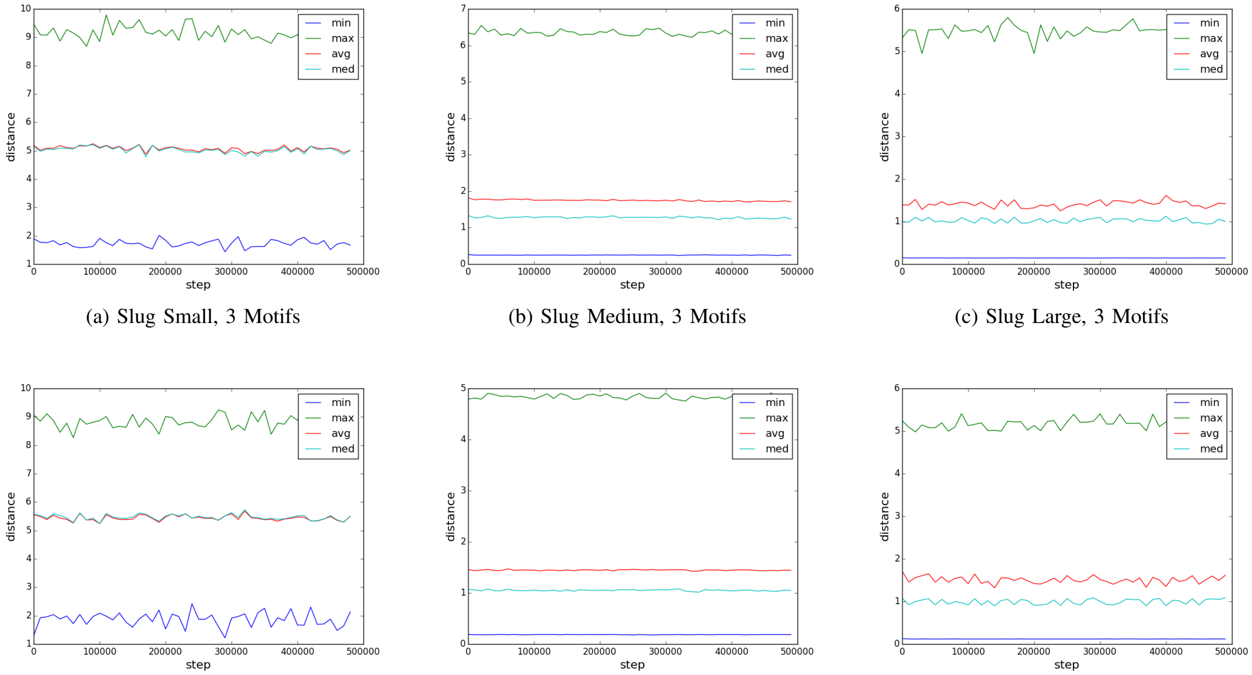
(c) Slug Large, 2 Motifs

(d) Snail Small, 2 Motifs

(e) Snail Medium, 2 Motifs

(f) Snail Large, 2 Motifs

Fig. 6. This figure presents the min, average, median and max distances between a random sample of 100 walks after every 10,000 steps in the walk for the 2 motif runs. Some of the medium and large intergenomic region have a noticably longer time to convergence. The distances between walks stays similar for all size data sets.



(a) Slug Small, 3 Motifs (b) Slug Medium, 3 Motifs (c) Slug Large, 3 Motifs

(d) Snail Small, 3 Motifs (e) Snail Medium, 3 Motifs (f) Snail Large, 3 Motifs

Fig. 7. This figure presents the min, average, median and max distances between a random sample of 100 walks after every 10,000 steps in the walk for the 2 motif runs. In contrast to the 1 motif runs, the disitance between walks on the three motif runs showed a marked decrease in distance with the larger datasets, suggesting that there were better matches for more motifs in the alrger data set sizes.

For example, Claudin-7 (CLDN7) as seen in table VII, a cell membrane protein, was shown to be regulated by Snail binding to its promoter E-box sequences by Ikenouchi et al. [26]. The gene desmoplakin (DSP) as seen in table VII, which is another known target of Snail according to Ohkubo et al. [27], was also identified in our walks as possessing E-box sequences. Another gene, ESRP2, while not identified as a direct target of Snail or Slug, does contain E-box sequences that can be bound by a protein called Zeb1, which performs similar functions to Snail and Slug according to Gemmill et al. [28]. This implies that Snail or Slug could possibly bind to the ESRP2 sequence in certain contexts as seen in Tables VI, VII and VIII. Indeed, Snail binds to E-box sequences at the ESRP1 gene promoter and represses it according to Reinke et al. [29]. While none of the Slug targets have been currently identified as direct targets, the data does help to pinpoint potential genes that can be validated by experimental approaches to discover novel ways of gene regulation. Overall, using gene regulation data from the microarray lists and then searching for E-box sequence motifs in those gene promoters, the data can be used to predict which of these are regulated by direct binding of Snail and Slug. Once validated, these genes could serve as future therapeutic targets for drug delivery, and/or biomarkers for cancer metastasis.

V. Conclusions and Future Work

This paper presents the use of the DNA@Home volunteer computing project to search for transcription factor binding sides around genes related to the Snail and Slug family of Zinc-finger transcription factors. Utilizing over 1,500 volunteer computing hosts for a period of two months, 18 different parallel Gibbs sampling runs were performed with varying parameters on data sets with up to 994 DNA sequence regions. To our knowledge, these present the largest scale use of Gibbs sampling for de novo detection of transcription factor binding sites.

These runs generated over 2.2 Terabytes of sampling data, which was in part analyzed using a high performance computing cluster to determine statistics about the distances between the parallel sampling walks. This information provides insight as to how well these runs were performing sampling, in terms of convergence regions of local optima or a singular region of a global optima. This is valuable information for determining how many motifs to search for, if the burn-in period has completed, and if the runs have generated enough samples to provide a reliable distribution of likely transcription factor binding sites. The use of parallel sampling walks allows Gibbs sampling to be performed at much larger scales and to more quickly gather samples.

This work provided a large scale example of the capabilities of DNA@Home, and we wish to incorporate the various metrics utilized in the analysis of this sampling data into a web based user interface for project scientists. Further, we plan to open DNA@Home up to external researchers, allowing them to submit their own FASTA files to perform their own Gibbs

TABLE III
E-box Slug Large 1 Motif Step 30,000

Walk Count	% Hits	Gene	Offset	Motif	cacctg	caggtg
426	13.27	FAM136A chr2 70528720 70529720	83	cacctGCCGCCgagga	True	False
426	13.69	WWOX chr16 78132826 78133826	833	caggtGCCTCCacagt	False	True
426	14.31	TMEM116 ERP29 chr12 112450523 112451651	921	caggtGCCGCCggggc	False	True
426	15.33	GNS chr12 65152726 65153726	278	caggtGGCGGGggctg	False	True
426	16.41	MYD88 chr3 38179468 38180468	851	caggtGGCGGCcgact	False	True
426	19.34	BST2 chr19 17515884 17516884	233	caggtGGCGGCctggg	False	True
426	19.97	COQ9 CIAPIN1 chr16 57480836 57481869	419	cacctGCCGCCtgggc	True	False
663	10.88	ZNF57 chr19 2906605 2907605	737	cacctGGAAAGttctg	True	False
966	13.68	PIN1 chr19 9945382 9946382	25	caggtGGGAAGaggga	False	True

TABLE IV
E-box Slug Large 2 Motifs Step 70,000

Walk Count	% Hits	Gene	Offset	Motif	cacctg	caggtg
213	10.36	TMEM41A chr3 185216345 185217345	287	cacctGCCTCCagcct	True	False
296	10.89	BST2 chr19 17515884 17516884	233	caggtGGCGGCctggg	False	True
299	10.89	WWOX chr16 78132826 78133826	833	caggtGCCTCCacagt	False	True
696	14.16	RORC chr1 151803848 151804848	11	cacctGGGAGGgcctg	True	False
696	17.86	LCLAT1 chr2 30669636 30670636	615	caggtGGGAGGctgga	False	True
697	13.79	PIN1 chr19 9945382 9946382	25	caggtGGGAAGaggga	False	True

TABLE V
E-box Slug Large 3 Motifs Step 30,000

Walk Count	% Hits	Gene	Offset	Motif	cacctg	caggtg
423	13.27	FAM136A chr2 70528720 70529720	83	cacctGCCGCCgagga	True	False
423	13.69	WWOX chr16 78132826 78133826	833	caggtGCCTCCacagt	False	True
423	14.33	TMEM116 ERP29 chr12 112450523 112451651	921	caggtGCCGCCggggc	False	True
423	15.29	GNS chr12 65152726 65153726	278	caggtGGCGGGggctg	False	True
423	16.40	MYD88 chr3 38179468 38180468	851	caggtGGCGGCcgact	False	True
423	19.38	BST2 chr19 17515884 17516884	233	caggtGGCGGCctggg	False	True
423	19.94	COQ9 CIAPIN1 chr16 57480836 57481869	419	cacctGCCGCCtgggc	True	False
695	10.86	ZNF57 chr19 2906605 2907605	737	cacctGGAAAGttctg	True	False
962	13.67	PIN1 chr19 9945382 9946382	25	caggtGGGAAGaggga	False	True

TABLE VI
E-box Snail Large 1 Motif Step 30,000

Walk Count	% Hits	Gene	Offset	Motif	cacctg	caggtg
392	14.63	TPD52 chr8 81082845 81083845	412	cacctGGAGGGacgag	True	False
396	22.80	RABAC1 chr19 42463028 42464028	841	cacctGGAGGGcttgc	True	False
429	27.44	FAM195A chr16 691619 692619	519	caggtGGAGGGccggc	False	True
469	13.31	RALGAPA2 chr20 20508402 20509402	127	caggtGGAAAGataag	False	True
491	11.01	MYL7 chr7 44180416 44181416	447	cacctGGGAGAccgct	True	False
499	20.34	SLC22A17 chr14 23821160 23822160	572	caggtGGGAGGgaggg	False	True
900	19.24	ESRP2 chr16 68269636 68270636	912	cacctGGGAAAgggga	True	False
946	15.43	STX3 chr11 59522031 59523031	979	cacctGGGAAGcgctc	True	False
1224	26.68	TXNRD2 chr22 19928859 19929859	174	cacctGGGAAGggggc	True	False

TABLE VII
E-box Snail Large 2 Motifs Step 30,000

Walk Count	% Hits	Gene	Offset	Motif	cacctg	caggtg
298	11.78	IVD chr15 40697185 40698185	989	caggtGAGGAGactga	False	True
298	14.18	CLDN7 chr17 7165764 7166764	206	caggtGAGGAGgaaga	False	True
298	18.90	DSP chr6 7541369 7542369	361	caggtGGGGAGgggcg	False	True
298	22.82	MKL2 chr16 14164695 14165695	257	caggtGAGAAGgaggc	False	True
430	10.48	C10orf35 chr10 71389502 71390502	504	caggtGGGAGGaaacc	False	True
471	14.52	ESRP2 chr16 68269636 68270636	912	cacctGGGAAAgggga	True	False
471	17.27	SLC22A17 chr14 23821160 23822160	572	caggtGGGAGGgaggg	False	True
769	16.82	STX3 chr11 59522031 59523031	979	cacctGGGAAGcgctc	True	False
771	37.80	TXNRD2 chr22 19928859 19929859	174	cacctGGGAAGggggc	True	False

sampling runs.

VI. Acknowledgements

We would like to thank the many volunteers at DNA@Home for their time, compute cycles and enthusiasm. Without them

TABLE VIII
E-BOX SNAIL LARGE 3 MOTIFS STEP 30,000

Walk Count	% Hits	Gene	Offset	Motif	cacctg	caggtg
383	15.59	SGK3 chr8 67686915 67687915	713	caggtGGAGGGacccc	False	True
385	22.89	RABAC1 chr19 42463028 42464028	841	cacctGGAGGGcttgc	True	False
418	27.45	FAM195A chr16 691619 692619	519	caggtGGAGGGccggc	False	True
459	13.39	RALGAPA2 chr20 20508402 20509402	127	caggtGGAAAGataag	False	True
470	11.01	MYL7 chr7 44180416 44181416	447	cacctGGGAGAccgct	True	False
501	20.36	SLC22A17 chr14 23821160 23822160	572	caggtGGGAGGgaggg	False	True
891	19.14	ESRP2 chr16 68269636 68270636	912	cacctGGGAAAgggga	True	False
930	15.40	STX3 chr11 59522031 59523031	979	cacctGGGAAGcgctc	True	False
1208	26.60	TXNRD2 chr22 19928859 19929859	174	cacctGGGAAGggggc	True	False

this work would not be possible. This work was supported by the National Institutes of Health [5P20GM104360 to N.S. and A.D] and University of North Dakota School of Medicine Faculty Seed Grant [to K.Z., T.D., S.N. and A.D.]. We would also like to thank Adam Burkholder for providing the software to aid in generating these data sets.

REFERENCES

[1] T. Desell, L. A. Newberg, M. Magdon-Ismail, B. K. Szymanski, and W. Thompson, "Finding protein binding sites using volunteer computing grids," in *Proceedings of the 2011 2nd International Congress on Computer Applications and Computational Science*. Springer, 2012, pp. 385–393.

[2] D. P. Anderson, E. Korpela, and R. Walton, "High-performance task distribution for volunteer computing." in *e-Science*. IEEE Computer Society, 2005, pp. 196–203.

[3] A. Villarejo, Á. Cortés-Cabrera, P. Molina-Ortíz, F. Portillo, and A. Cano, "Differential role of snail1 and snail2 zinc fingers in e-cadherin repression and epithelial to mesenchymal transition," *Journal of Biological Chemistry*, vol. 289, no. 2, pp. 930–941, 2014.

[4] M. A. Nieto, "The snail superfamily of zinc-finger transcription factors," *Nature reviews Molecular cell biology*, vol. 3, no. 3, pp. 155–166, 2002.

[5] G. Moreno-Bueno, E. Cubillo, D. Sarrió, H. Peinado, S. M. Rodríguez-Pinilla, S. Villa, V. Bolós, M. Jordá, A. Fabra, F. Portillo *et al.*, "Genetic profiling of epithelial cells expressing e-cadherin repressors reveals a distinct role for snail, slug, and e47 factors in epithelial-mesenchymal transition," *Cancer research*, vol. 66, no. 19, pp. 9543–9556, 2006.

[6] A. Dhasarathy, D. Phadke, D. Mav, R. R. Shah, and P. A. Wade, "The transcription factors snail and slug activate the transforming growth factor-beta signaling pathway in breast cancer," *PLoS One*, vol. 6, no. 10, p. e26514, 2011.

[7] 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*, vol. 262, no. 5131, pp. 208–214, 1993.

[8] I. V. Kulakovskiy, V. Boeva, A. V. Favorov, and V. Makeev, "Deep and wide digging for binding motifs in chip-seq data," *Bioinformatics*, vol. 26, no. 20, pp. 2622–2623, 2010.

[9] G. Pavesi, P. Mereghetti, G. Mauri, and G. Pesole, "Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes," *Nucleic acids research*, vol. 32, no. suppl 2, pp. W199–W203, 2004.

[10] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, "Meme suite: tools for motif discovery and searching," *Nucleic acids research*, p. gkp335, 2009.

[11] S. Georgiev, A. P. Boyle, K. Jayasurya, X. Ding, S. Mukherjee, U. Ohler *et al.*, "Evidence-ranked motif identification," *Genome Biol*, vol. 11, no. 2, p. R19, 2010.

[12] M. Hu, J. Yu, J. M. Taylor, A. M. Chinnaiyan, and Z. S. Qin, "On the detection and refinement of transcription factor binding sites using chip-seq data," *Nucleic acids research*, vol. 38, no. 7, pp. 2154–2167, 2010.

[13] L. Narlikar, R. Gordân, and A. J. Hartemink, "A nucleosome-guided map of transcription factor binding sites in yeast," *PLoS computational biology*, vol. 3, no. 11, p. e215, 2007.

[14] G. Thijs, K. Marchal, M. Lescot, S. Rombauts, B. De Moor, P. Rouzé, and Y. Moreau, "A gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes," *Journal of Computational Biology*, vol. 9, no. 2, pp. 447–464, 2002.

[15] X. S. Liu, D. L. Brutlag, and J. S. Liu, "An algorithm for finding protein–dna binding sites with applications to chromatin-immunoprecipitation microarray experiments," *Nature biotechnology*, vol. 20, no. 8, pp. 835–839, 2002.

[16] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo *et al.*, "Transcriptional regulatory code of a eukaryotic genome," *Nature*, vol. 431, no. 7004, pp. 99–104, 2004.

[17] D. Che, S. Jensen, L. Cai, and J. S. Liu, "Best: binding-site estimation suite of tools," *Bioinformatics*, vol. 21, no. 12, pp. 2909–2911, 2005.

[18] X. Liu, D. L. Brutlag, J. S. Liu *et al.*, "Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes." in *Pacific symposium on biocomputing*, vol. 6, no. 2001, 2001, pp. 127–138.

[19] X. Chen, L. Guo, Z. Fan, and T. Jiang, "W-alignace: an improved gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/chip-chip data," *Bioinformatics*, vol. 24, no. 9, pp. 1121–1128, 2008.

[20] S. P. Brooks and G. O. Roberts, "Convergence assessment techniques for markov chain monte carlo," *Statistics and Computing*, vol. 8, no. 4, pp. 319–335, 1998.

[21] S. T. Jensen, X. S. Liu, Q. Zhou, and J. S. Liu, "Computational discovery of gene regulatory binding motifs: a bayesian perspective," *Statistical Science*, pp. 188–204, 2004.

[22] D. B. Woodard, J. S. Rosenthal *et al.*, "Convergence rate of markov chain methods for genomic motif discovery," *The Annals of Statistics*, vol. 41, no. 1, pp. 91–124, 2013.

[23] K. R. Rosenbloom, C. A. Sloan, V. S. Malladi, T. R. Dreszer, K. Learned, V. M. Kirkup, M. C. Wong, M. Maddren, R. Fang, S. G. Heitner *et al.*, "Encode data in the ucsc genome browser: year 5 update," *Nucleic acids research*, vol. 41, no. D1, pp. D56–D63, 2013.

[24] B. Langmead, C. Trapnell, M. Pop, S. Salzberg *et al.*, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol*, vol. 10, no. 3, p. R25, 2009.

[25] W. Huang, G. Santhanaraman, H.-W. Jin, Q. Gao, and D. K. Panda, "Design of high performance MVAPICH2: MPI2 over InfiniBand," in *Cluster Computing and the Grid, 2006. CCGRID 06. Sixth IEEE International Symposium on*, vol. 1. IEEE, 2006, pp. 43–48.

[26] J. Ikenouchi, M. Matsuda, M. Furuse, and S. Tsukita, "Regulation of tight junctions during the epithelium-mesenchyme transition: direct repression of the gene expression of claudins/occludin by snail," *Journal of cell science*, vol. 116, no. 10, pp. 1959–1967, 2003.

[27] T. Ohkubo and M. Ozawa, "The transcription factor snail downregulates the tight junction components independently of e-cadherin downregulation," *Journal of cell science*, vol. 117, no. 9, pp. 1675–1685, 2004.

[28] R. M. Gemmill, J. Roche, V. A. Potiron, P. Nasarre, M. Mitas, C. D. Coldren, B. A. Helfrich, E. Garrett-Mayer, P. A. Bunn, and H. A. Drabkin, "Zeb1-responsive genes in non-small cell lung cancer," *Cancer letters*, vol. 300, no. 1, pp. 66–78, 2011.

[29] L. M. Reinke, Y. Xu, and C. Cheng, "Snail represses the splicing regulator epithelial splicing regulatory protein 1 to promote epithelial-mesenchymal transition," *Journal of Biological Chemistry*, vol. 287, no. 43, pp. 36 435–36 442, 2012.