HHS Public Access

Author manuscript

Proc IEEE Int Conf Escience. Author manuscript; available in PMC 2016 August 01.

Published in final edited form as:

Proc IEEE Int Conf Escience. 2015 ; 2015: 429–438. doi:10.1109/eScience.2015.27.

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

Kristopher Zarns,
Department of Computer Science, University of North Dakota, Grand Forks, North Dakota 58202-9015

Travis Desell,
Department of Computer Science, University of North Dakota, Grand Forks, North Dakota 58202-9015

Sergei Nechaev, and
Department of Basic Sciences, University of North Dakota, Grand Forks, North Dakota 58202-9061

Archana Dhasarathy
Department of Basic Sciences, University of North Dakota, Grand Forks, North Dakota 58202-9061

Kristopher Zarns: kzarns@gmail.com; Travis Desell: tdesell@cs.und.edu; Sergei Nechaev: sergei.nechaev@med.und.edu; Archana Dhasarathy: archana.dhasarathy@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. Genes regulated by Slug but not Snail, and genes regulated by Snail but not Slug provided 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. 1000 parallel sampling walks were used to search for the presence of 1, 2 or 3 possible motifs using small, medium, and full size sets of these sequences. These runs were performed over a period of two months using over 1500 volunteered computing hosts and generated over 2.2 Terabytes of sampling data. High performance computing resources were used for post processing. This paper presents intra and inter walk analyses used to determine walk convergence. The results were validated against current biological knowledge of the Snail and Slug promoter regions and present avenues for further biological study.

I. Introduction

This paper presents new results from DNA@Home[1] [1], which uses BOINC [2] to provide massively scalable computing power to search for transcription factor binding sites (or

[1] http://csgrid.org/csg/dna

motifs) in large datasets. DNA@Home implements an asynchronous version of the Gibbs Sampling algorithm which performs parallel sampling walks using volunteer computing. DNA@Home performed parallel Gibbs sampling runs over a two month time period which varied in the number of motifs searched for and the dataset size used. The aim was to identify motifs related to the SNAI1 (Snail) and SNAI2 (Slug) genes. Each run had 1000 parallel sampling walks and the largest dataset contained 994 regions of DNA. This resulted in over 2.2 Terabytes of sampling data which was analyzed using high performance computing (HPC) resources. To our knowledge, this is the largest scale use of Gibbs sampling for de novo transcription factor binding site discovery.

A. The Gibbs Sampler

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

With Gibbs sampling the randomly chosen starting point biases the result of the sampler. To overcome this bias a certain number of steps, or a *burn-in*, should be discarded. The burn-in might be significantly larger than a 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 needed is similar for each walk though there are some outlier walks which do not converge as quickly. Burn-in must be completed before a valid sample set is generated. A larger super-step provides a better sample and will make burn-in easier to detect. A smaller super-step requires less computation per volunteer computing job.

Convergence detection algorithms can be used to determine burn-in for a walk. Ideally the distribution generated at each super-step will approximate a stationary distribution, meaning that additional steps will not significantly alter the distribution. After burn-in is complete the sampler should converge and the next result of the sampler should be valid. At this point, No further steps are required and computation can cease. It is worth noting that some datasets and input parameters will not converge.

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 may have converged to a local optima. By using multiple walks, it is possible to discover additional motifs by settling into different local optima. This provides 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 [5], [6], which potentially has implications in their distinct roles at different stages of cancer metastasis. The molecular basis for the distinct regulation and binding affinity of downstream target genes by Snail and Slug is still currently unknown.

II. Related Work

Lawrence *et al*. [7] discusses how to apply Gibbs Sampling to motif finding. The effect of differing random start sites on converge rate is described. Lawrence claims that larger datasets provide a better pattern model which improves the power of the Gibbs sampler.

A. Dataset Size

Table I relates the dataset analyzed by DNA@Home to other 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] an 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. ChIPMunk addresses the increased problem space created by ChIPSeq data. However, ChIPMunk is not a Gibbs Sampling algorithm. It is a greedy optimization using several heuristics which are specific to ChIPSeq.

Narlikar *et al*. [13] provides PRIORITY, a Gibbs Sampling algorithm which uses knowledge of transcription factor binding sites to use an informative prior probability. PRIORITY is shown as an improvement over AlignACE [14], MEME [10], MDscan [15] (a positional weight 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 Gibbs Sampler, BioProspector. Liu describes methods to validate that meaningful motifs were found. 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 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 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 Eigenvalue bounds. The Kolmogorov-Smirnov two sample statistic is a spectral method which can be used to test the null hypothesis of stationarity.

Jensen *et al*. [21] discuss finding motifs using Gibbs sampling when multiple motifs are present in the dataset. Jensen uses an annealing approach to shift the sampler to avoid being stuck in local maxima. The size of the shifts decreases over time according to a heat function.

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 containing multiple motifs. In Woodwards case, convergence rate decreased as the length of DNA sample increased.

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 four days. The genes that were uniquely regulated by Snail or Slug (both up or down) over the four days were compiled into a list, with overlaps being merged. This generated two lists of genes unique to Snail or Slug regulation. The gene sequence were 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 wg En-code Open Chrom Chip Mcf7 Pol2 Serumstim Raw Data Rep1. This track was chosen due to prior

work which centered on Snail and Slug representation with regards to RNA Polymerase II (Pol II) binding. The entire Snail dataset contains 1422 genes, while the entire Slug dataset only contains 412. Three different size FASTA files were generated for each dataset: small, medium, and large. The datasets are illustrated in Table II.

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, associates the display data with gene intervals, filters for overlapping genes, and ranks the genes based on the number of matching reads. The Bowtie To Bed Graph conversion software and CPPMatch ranking software were provided by Adam Burkholder of the National Institute of Health [25].

B. Gibbs Sampling Configuration

The Gibbs Sampler was configured to search for 1, 2, or 3 motifs six nucleotides in length. This was done to examine how the number of motifs present in the dataset effects how quickly the parallel walks converge. There are six datasets, three for Slug and three 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, 1000 independent walks were created. Within a run each walk had the same dataset and started with a random initial starting sample and a different random seed. A super-step size of 10000 steps was used. 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 two sample statistic was used to test the null hypothesis of stationary distribution. Each time a walk was restarted, samples were reported for the last super-step of the walk. These samples represent the empirical distribution of that period. The test generated two values, a maximum distance between distributions and a probability that two distributions were generated from the same source distribution. The test sorts the sample to create the distribution function and then compares the distribution with the previous distribution.

The test is appropriate for testing Gibbs Sampling convergence for several reasons. The test is non-parametric, it does not 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, therefore 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[2], of which DNA@Home is a subproject. The number of simultaneously volunteered hosts participating in the project averaged around 1650 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 1500 users provide over 4100 compute hosts for the project. In total, 18 runs were made looking for 1 to 3 motifs using Snail and Slug datasets of small, medium, and large sizes (see Table II). These runs generated over 2.2 TB of sampling data. 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 runs that converged. As illustrated by Figure 2, the small dataset converged for the case of one motif for both datasets. However, the probability sample standard deviation (PSSD) 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 runs with a PSSD 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 20000 steps. The two 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 One Motif Runs*: Figure 2 shows a comparison of the one motif results for Snail and Slug. The small Slug datasets both show signs of convergence. However the high PSSD 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 20000 steps. The minimal PSSD and consistently high minimum probability suggest that all of the walks have converged.

2. *Analysis of the Two Motif Runs*: Figure 3 shows the results from searching for two motifs at once. This shows that using a larger dataset improves the rate at which the walks converge. In both the Slug and Snail two motif medium cases, the walks do not immediately converge. Instead of the convergence seen in the first 20000 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

[2] http://csgrid.org

deviation isn't reduced until 200000 steps. The Snail dataset sees the reduced standard deviation at 130000 steps. In both cases the large dataset performs better.

3. *Analysis of the Three Motif Runs*: Figure 4 shows that for Slug the medium size dataset converges quickly at around 40000 steps. However 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 were tested. Those tests proved to be extremely computationally expensive. The sampling datasets ranged from around 20 GB for the runs with one motif on the small number of sequences, to over 500 GB for the runs with three motifs on the large number of sequences. To compare the distance between each walk at every super-step a parallel analysis tool was developed using C++ and MPI which utilized HPC resources. A Beowulf HPC cluster with 32 dual quad-core compute nodes (for a total of 256 processing cores) was used. Each compute node had 64 GB of 1600-MHz RAM, two mirrored RAID 146-GB 15-K RPM SAS drives, two quad-core E5-2643 Intel processors which operate at 3.3 Ghz, and ran the Red Hat Enterprise Linux (RHEL) 6.2 operating system. All 32 nodes within the cluster were linked by a private 56 Gb InfiniBand FDR 1-to-1 network. The code was compiled and run using MVAPICH2-x [26] to allow highly optimized use of this network infrastructure.

Randomized sampling was required to calculate these results in a reasonable 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 1 or 3 motifs, which essentially converge within the first super-step. 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 average and median distances, 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. Generally runs with low average and median distances would generate enough parallel walks to get an appropriate sampling across all possible optima, while runs with high median and averages, would require either more sampling walks or motifs to be sure all regions of the search space are sampled correctly (*e.g.*, the large Slug and Snail datasets with one 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 III, IV, V, VI, VII, and VIII display the number of walks containing the motif, percentage of time the position was sampled, the gene symbol and chromosome(chr2 is chromosome2 for example), start of the region, end of the region, motif location offset from the start, five nucleotides before the motif, the motif in capitalized letters, five nucleotides after the motif, and if it contained the E-Box motifs. If multiple motifs were found for a sequence then multiple motifs were returned for that sequence. No motif was reported for a sequence if none of the motifs overcame the minimum percentage.

Several of the genes that had the E-box in their promoter regions were known targets or predicted targets. 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*. [27]. The gene desmoplakin (DSP) as seen in table VII, which is a known target of Snail according to Ohkubo *et al*. [28], was also identified 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*. [29]. This implies that Snail or Slug could possibly bind to the ESRP2 sequence in certain contexts as seen in Tables VI, VII and VIII. Snail binds to E-box sequences at the ESRP1 gene promoter and represses it according to Reinke *et al*. [30]. While none of the Slug targets have been currently identified as direct targets, the data helps to pinpoint potential genes for validation 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, 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 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 1500 volunteer computing hosts for a period of two months, 18 different parallel Gibbs sampling runs were performed with varying parameters on datasets 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 analyzed using a HPC 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 there are plans to incorporate the various metrics utilized in the analysis of this sampling data into a web based user interface for project scientists. Further, there are plans to open DNA@Home up to external researchers, allowing them to submit their own FASTA files to perform their own Gibbs sampling runs.

Acknowledgments

Thank you to the many volunteers at DNA@Home for their time, compute cycles, and enthusiasm. Without them 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.]. Thank you to Adam Burkholder for providing the software to aid in generating these datasets [25].

References

1. Desell, T.; Newberg, LA.; Magdon-Ismail, M.; Szymanski, BK.; Thompson, W. Proceedings of the 2011 2nd International Congress on Computer Applications and Computational Science. Springer; 2012. Finding protein binding sites using volunteer computing grids; p. 385-393.
2. Anderson, DP.; Korpela, E.; Walton, R. e-Science. IEEE Computer Society; 2005. High-performance task distribution for volunteer computing; p. 196-203.
3. Villarejo A, Cortés-Cabrera Á, Molina-Ortíz P, Portillo F, Cano A. Differential role of snail1 and snail2 zinc fingers in e-cadherin repression and epithelial to mesenchymal transition. Journal of Biological Chemistry. 2014; 289(2):930–941. [PubMed: 24297167]
4. Nieto MA. The snail superfamily of zinc-finger transcription factors. Nature reviews Molecular cell biology. 2002; 3(3):155–166. [PubMed: 11994736]
5. Moreno-Bueno G, Cubillo E, Sarrió D, Peinado H, Rodríguez-Pinilla SM, Villa S, Bolós V, Jordá M, Fabra A, Portillo F, 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. 2006; 66(19):9543–9556. [PubMed: 17018611]
6. Dhasarathy A, Phadke D, Mav D, Shah RR, Wade PA. The transcription factors snail and slug activate the transforming growth factor-beta signaling pathway in breast cancer. PLoS One. 2011; 6(10):e26514. [PubMed: 22028892]
7. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. science. 1993; 262(5131):208–214. [PubMed: 8211139]
8. Kulakovskiy IV, Boeva V, Favorov AV, Makeev V. Deep and wide digging for binding motifs in chip-seq data. Bioinformatics. 2010; 26(20):2622–2623. [PubMed: 20736340]
9. Pavesi G, Mereghetti P, Mauri G, Pesole G. Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic acids research. 2004; 32(suppl 2):w199–w203. [PubMed: 15215380]
10. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. Meme suite: tools for motif discovery and searching. Nucleic acids research. 2009:gkp335.
11. Georgiev S, Boyle AP, Jayasurya K, Ding X, Mukherjee S, Ohler U, et al. Evidence-ranked motif identification. Genome Biol. 2010; 11(2):r19. [PubMed: 20156354]
12. Hu M, Yu J, Taylor JM, Chinnaiyan AM, Qin ZS. On the detection and refinement of transcription factor binding sites using chip-seq data. Nucleic acids research. 2010; 38(7):2154–2167. [PubMed: 20056654]
13. Narlikar L, Gordân R, Hartemink AJ. a nucleosome-guided map of transcription factor binding sites in yeast. PLoS computational biology. 2007; 3(11):e215. [PubMed: 17997593]

14. Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, Rouzé P, Moreau Y. A gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. Journal of Computational Biology. 2002; 9(2):447–464. [PubMed: 12015892]
15. Liu XS, Brutlag DL, Liu JS. An algorithm for finding protein–dna binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nature biotechnology. 2002; 20(8): 835–839.
16. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, et al. Transcriptional regulatory code of a eukaryotic genome. Nature. 2004; 431(7004):99–104. [PubMed: 15343339]
17. Che D, Jensen S, Cai L, Liu JS. Best: binding-site estimation suite of tools. Bioinformatics. 2005; 21(12):2909–2911. [PubMed: 15814553]
18. Liu X, Brutlag DL, Liu JS, et al. Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. Pacific symposium on biocomputing. 2001; 6(2001): 127–138. [PubMed: 11262934]
19. Chen X, Guo L, Fan Z, Jiang T. W-alignace: an improved gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/chip-chip data. Bioinformatics. 2008; 24(9):1121–1128. [PubMed: 18325926]
20. Brooks SP, Roberts GO. Convergence assessment techniques for markov chain monte carlo. Statistics and Computing. 1998; 8(4):319–335.
21. Jensen ST, Liu XS, Zhou Q, Liu JS. Computational discovery of gene regulatory binding motifs: a bayesian perspective. Statistical Science. 2004:188–204.
22. Woodard DB, Rosenthal JS, et al. Convergence rate of markov chain methods for genomic motif discovery. The Annals of Statistics. 2013; 41(1):91–124.
23. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG, et al. Encode data in the ucsc genome browser: year 5 update. Nucleic acids research. 2013; 41(D1):D56–D63. [PubMed: 23193274]
24. Langmead B, Trapnell C, Pop M, Salzberg S, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10(3):R25. [PubMed: 19261174]
25. Burkholder, A.; Zarns, K.; Scheidegger, A. “Chip-seq: Initial release”, Github Repository. Jun. 2015 [Online]. Available: http://dx.doi.org/10.5281/zenodo.18556
26. Huang, W.; Santhanaraman, G.; Jin, HW.; Gao, Q.; Panda, DK. Cluster Computing and the Grid, 2006 CCGRID 06 Sixth IEEE International Symposium on. Vol. 1. IEEE; 2006. Design of high performance MVAPICH2: MPI2 over InfiniBand; p. 43-48.
27. Ikenouchi J, Matsuda M, Furuse M, Tsukita S. Regulation of tight junctions during the epithelium-mesenchyme transition: direct repression of the gene expression of claudins/occludin by snail. Journal of cell science. 2003; 116(10):1959–1967. [PubMed: 12668723]
28. Ohkubo T, Ozawa M. The transcription factor snail downregulates the tight junction components independently of e-cadherin downregulation. Journal of cell science. 2004; 117(9):1675–1685. [PubMed: 15075229]
29. Gemmill RM, Roche J, Potiron VA, Nasarre P, Mitas M, Coldren CD, Helfrich BA, Garrett-Mayer E, Bunn PA, Drabkin HA. Zeb1-responsive genes in non-small cell lung cancer. Cancer letters. 2011; 300(1):66–78. [PubMed: 20980099]
30. Reinke LM, Xu Y, Cheng C. Snail represses the splicing regulator epithelial splicing regulatory protein 1 to promote epithelial-mesenchymal transition. Journal of Biological Chemistry. 2012; 287(43):36435–36442. [PubMed: 22961986]

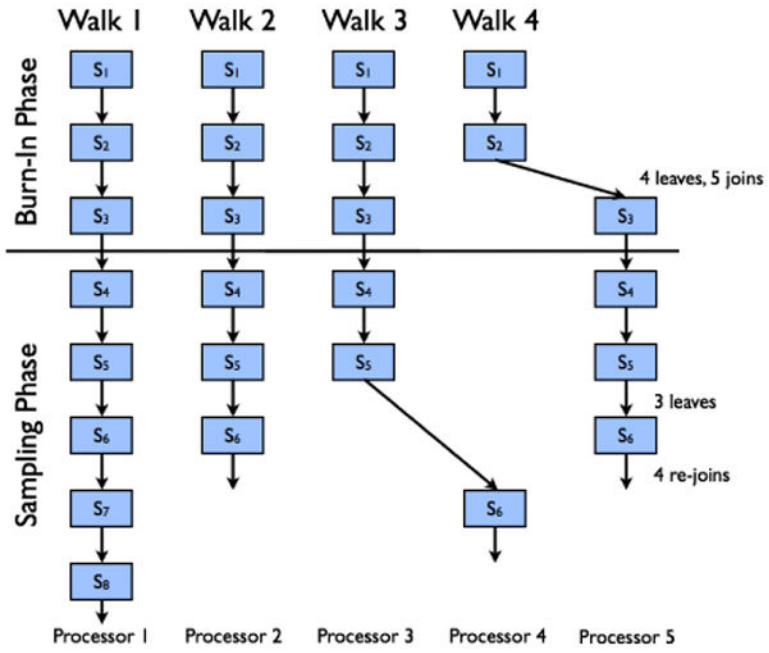
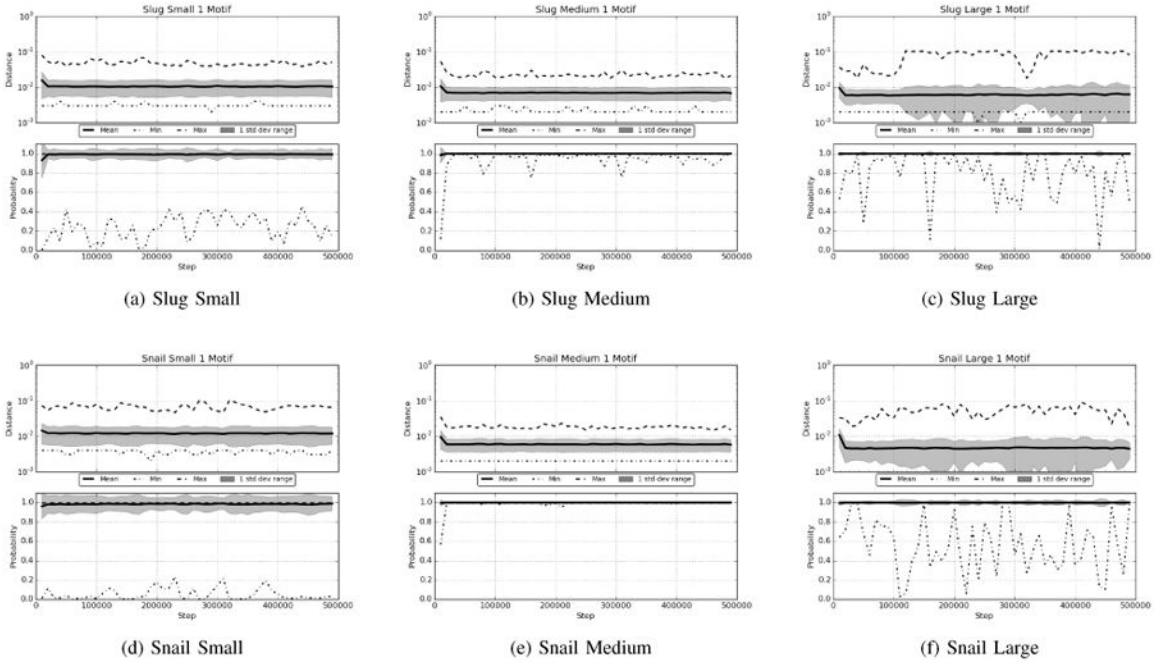


Fig. 1.

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 Figure, 1), however the runs described in this work had walk lengths of 10000. When a walk completes its burn-in period, samples are taken. Processors can join and leave, restarting from walks of previously left processors.



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

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

Fig. 2.

One 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 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

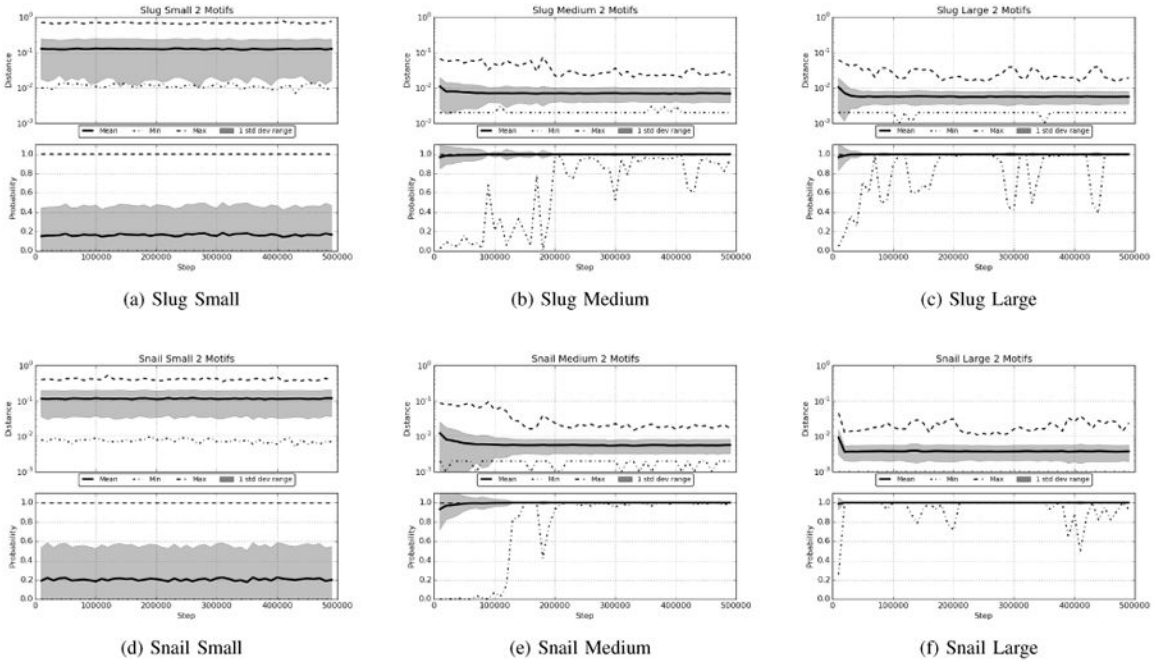


Fig. 3.

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

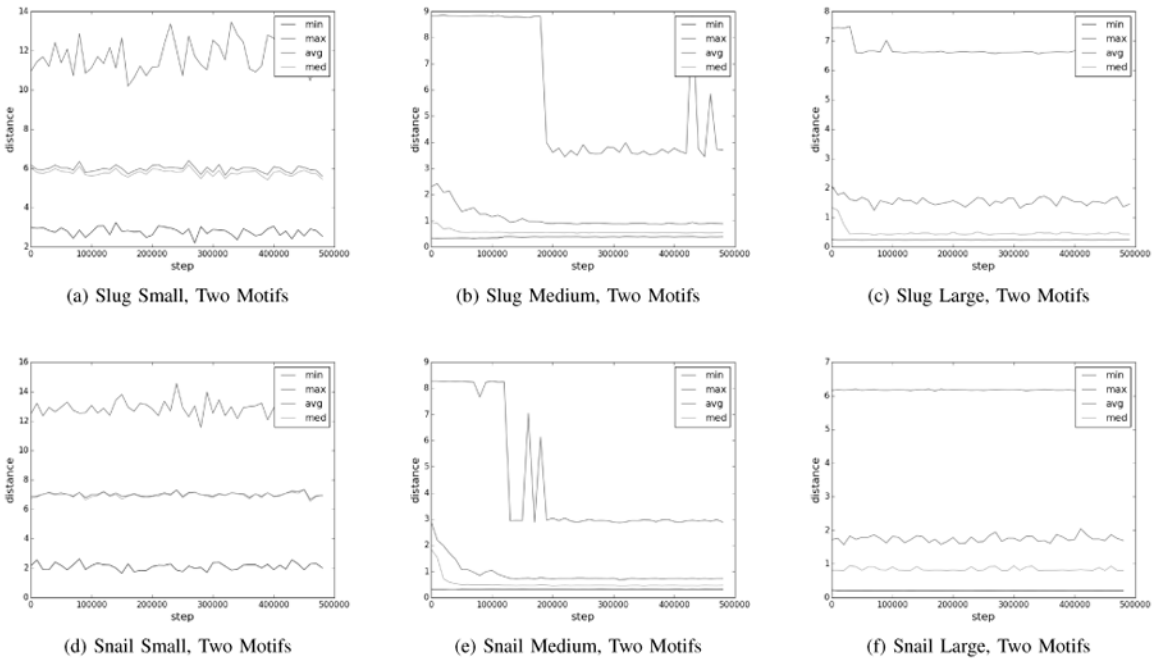
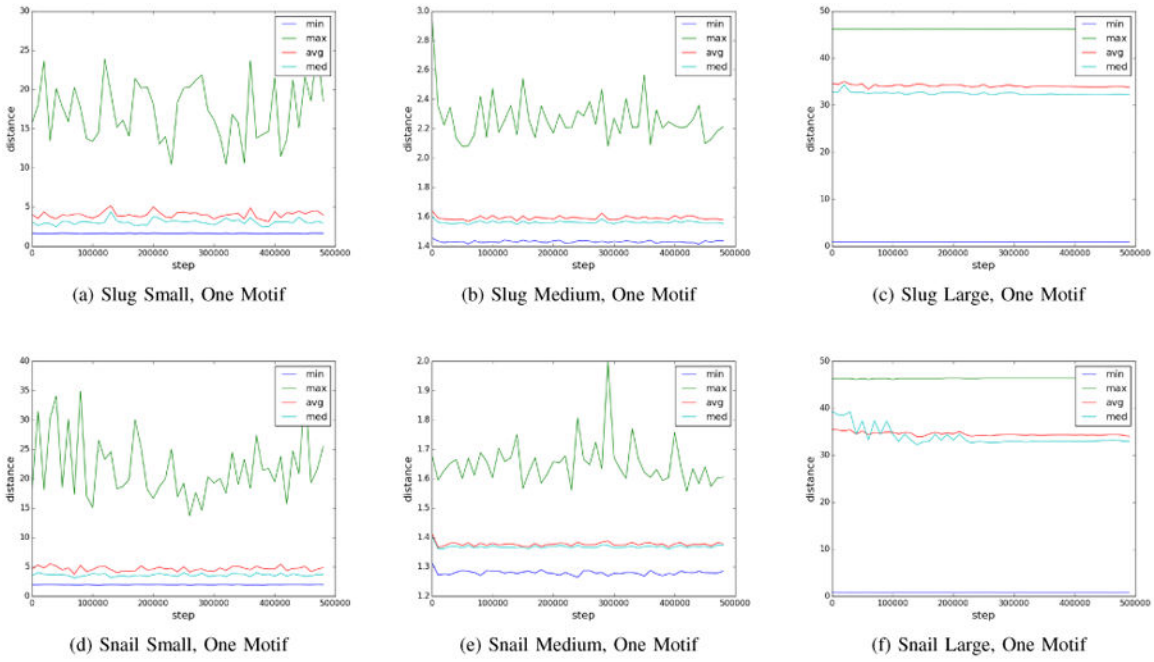


Fig. 4.

Three Motif Kolmogorov-Smirnov Analysis. Top row: Slug is unstable for the small dataset. The rate of convergence is similar to the one 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, One Motif (b) Slug Medium, One Motif (c) Slug Large, One Motif

(d) Snail Small, One Motif (e) Snail Medium, One Motif (f) Snail Large, One Motif

Fig. 5.

This figure presents the minimum, average, median, and maximum distances between a random sample of 100 walks after every super-step for the one 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.

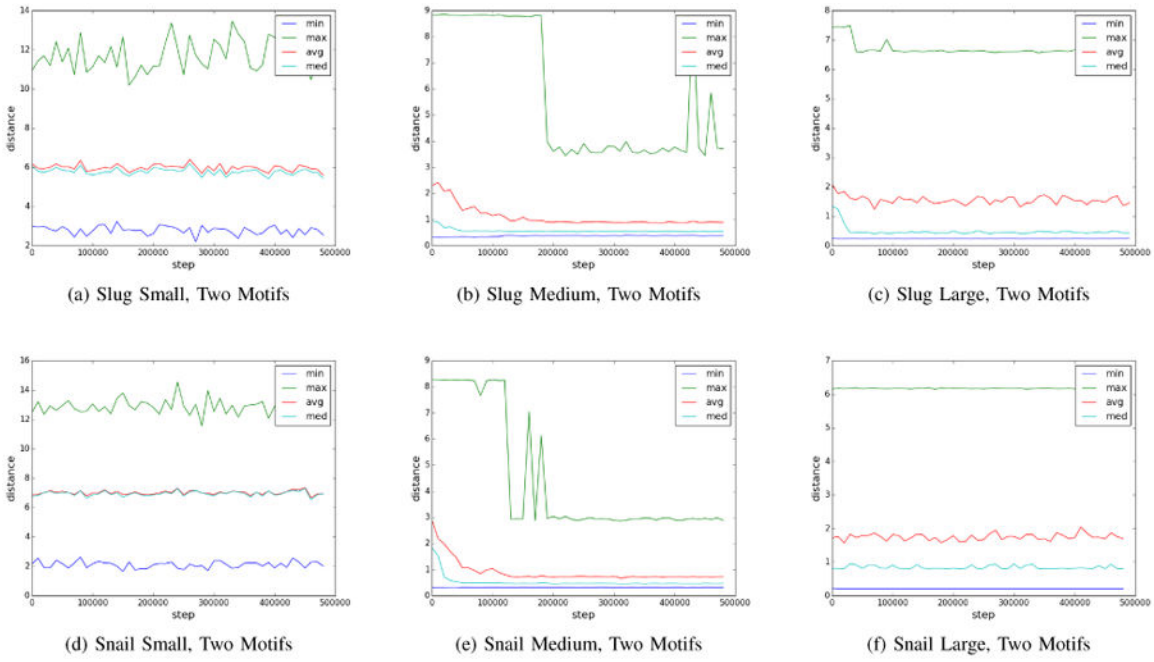


Fig. 6.

This figure presents the minimum, average, median, and maximum distances between a random sample of 100 walks after every super-step for the two motif runs. Some of the medium and large intergenomic region have a noticeably longer time to convergence. The distances between walks stays similar for all size datasets.

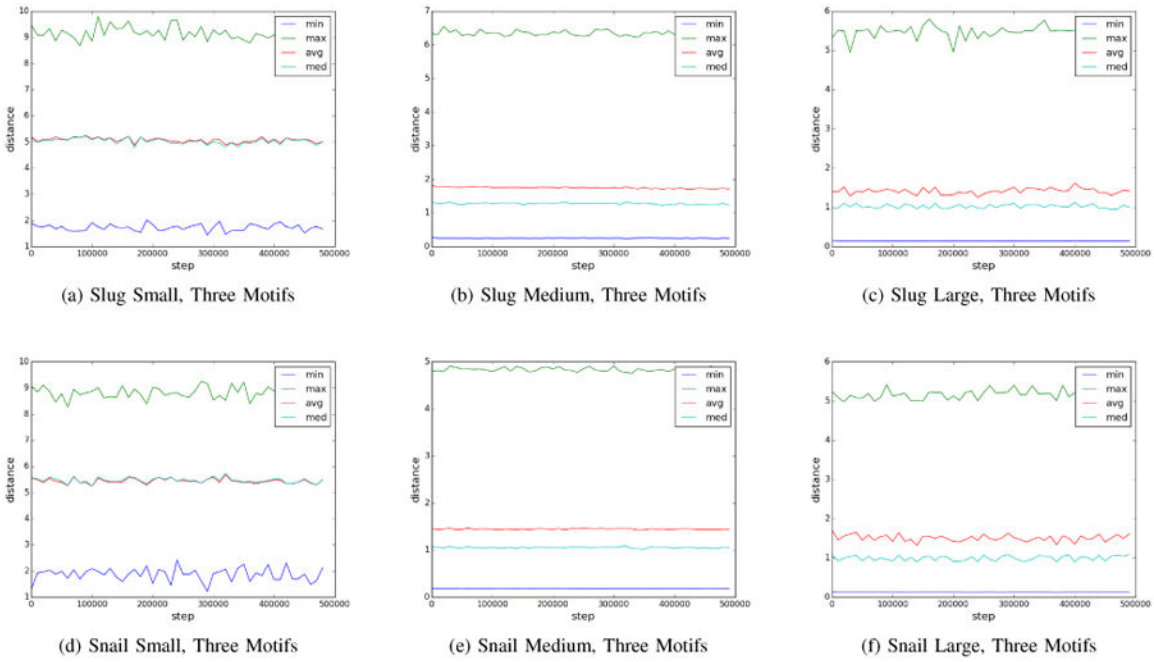


Fig. 7.

This figure presents the minimum, average, median, and maximum distances between a random sample of 100 walks after every super-step for the three motif runs. In contrast to the one motif runs, the distance between walks showed a decrease in distance with the larger datasets, suggesting that there were better matches for more motifs in the larger dataset sizes.

Table I

Project Dataset Comparison

Project	Gibbs	Sequences	Length	Runs	Motifs	Width
DNA@Home	Yes	994	1000	1000	1,2,3	6
ChIPMunk	No	10000	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

Table II

Dataset Configuration and Burn-In Information

Motifs	Dataset	Size	Intervals	Genes	Burn In	Stable
1	Snail	Large	1442	994	<20000	Yes
1	Snail	Medium	1442	100	<20000	Yes
1	Snail	Small	1442	10	<20000	No
1	Slug	Large	412	372	<20000	Yes
1	Slug	Medium	412	99	<20000	Yes
1	Slug	Small	412	10	<20000	No
2	Snail	Large	1442	994	<20000	Yes
2	Snail	Medium	1442	100	130000	Yes
2	Snail	Small	1442	10	N/A	No
2	Slug	Large	412	372	60000	Yes
2	Slug	Medium	412	99	200000	Yes
2	Slug	Small	412	10	N/A	No
3	Snail	Large	1442	994	<20000	Yes
3	Snail	Medium	1442	100	<20000	Yes
3	Snail	Small	1442	10	N/A	No
3	Slug	Large	412	372	<20000	Yes
3	Slug	Medium	412	99	<20000	No
3	Slug	Small	412	10	N/A	No

Table III

E-box Slug Large One Motif Step 30000

Walk Count	% Hits	Gene	Start	End	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 Two Motifs Step 70000

Walk Count	% Hits	Gene	Start	End	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

Proc IEEE Int Conf Escience. Author manuscript; available in PMC 2016 August 01.

Table V

E-box Slug Large Three Motifs Step 30000

Walk Count	% Hits	Gene	Start	End	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

Proc IEEE Int Conf Escience. Author manuscript; available in PMC 2016 August 01.

Table VI

E-box Snail Large One Motif Step 30000

Walk Count	% Hits	Gene	Start	End	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

Proc IEEE Int Conf Escience. Author manuscript; available in PMC 2016 August 01.

Table VII

E-box Snail Large Two Motifs Step 30000

Walk Count	% Hits	Gene	Start	End	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

Proc IEEE Int Conf Escience. Author manuscript; available in PMC 2016 August 01.

Table VIII

E-box Snail Large Three Motifs Step 30000

Walk Count	% Hits	Gene	Start	End	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

Proc IEEE Int Conf Escience. Author manuscript; available in PMC 2016 August 01.