HHS Public Access

Author manuscript

IEEE Symp Comput Intell Bioinforma Comput Biol Proc. Author manuscript; available in PMC 2019 April 19.

Published in final edited form as:

IEEE Symp Comput Intell Bioinforma Comput Biol Proc. 2013 April ; 2013: 60–67. doi:10.1109/CIBCB.2013.6595389.

An Exploration Into Improving DNA Motif Inference by Looking for Highly Conserved Core Regions

Jeffrey A. Thompson and **Clare Bates Congdon**

Department of Computer Science, University of Southern Maine, Portland, Maine 04104

Abstract

Although most verified functional elements in noncoding DNA contain a highly conserved core region, this concept is not generally incorporated into *de novo* motif inference systems. In this work, we explore the utility of adding the notion of conserved core regions into a comparative genomics approach for the search for putative functional elements in noncoding DNA. By modifying the scoring function for GAMI, Genetic Algorithms for Motif Inference, we investigate tradeoffs between the strength of conservation of the full motif vs. the strength of conservation of a core region. This work illustrates that incorporating information about the structure of transcription factor binding sites can be helpful in identifying biologically functional elements.

I. Introduction

GAMI (Genetic Algorithms for Motif Inference) [1], [2] is designed to identify candidate functional elements in noncoding DNA, such as transcription factor binding sites (TFBSs). GAMI uses a genetic algorithms (GA) search to identify patterns (motifs) that recur in multiple sequences that are being compared. These sequences are often noncoding DNA from divergent species; conservation of nucleotide patterns between species with common ancestry is often used as an indicator of possible functional regions [3], [4], [5], [6], [7], [8]. GAMI is able to find hundreds of candidate functional elements in a single run, and is able to work with long sequence lengths, such as 1 million base pairs.

Transcriptional regulation is essential to the differential expression of genes in various tissues. Therefore, identifying TFBSs is an important step in understanding gene activity. Furthermore, mutations in TFBSs are responsible for a number of diseases [9], making it important to understand how genes are regulated under normal circumstances so that problems can be identified.

TFBSs are frequently chracterized by comparing known sites from multiple DNA sequences (often across species) and compiling the results into a position frequency matrix. Analysis of these matrices has shown that TFBSs often exhibit a conserved “core” of adjacent bases with a markedly higher conservation than surrounding bases [10]. This suggests that some parts of such TFBSs may be under tighter evolutionary constraint.

By default, GAMI evolves consensus motifs on the A, C, G, T alphabet and the default fitness function in GAMI is a measure of the conservation of the bases in the evolved motif. Although GAMI finds 100’s of candidate functional elements in a single run, there is no

component of the fitness function to encourage motifs with a highly conserved core region. In the work here, the fitness function is modified to score the full-length motif as well as a core region, so that both the general conservation and the core-region conservation can be incorporated.

In the remainder of this paper, Section II explains relevant background, including both characteristics of known TFBS and the GAMI system. Section III explains the modified scoring metric. Section IV explains our research methodology, including the the data curated for this work and the experimental design. Section V describes the results from these experiements; Section VI explores the interpretation of these results; and Section VII describes future work.

II. Background

A. Core Regions of TFBSs

A single transcription factor is frequently able to bind to a variety of nucleotide patterns. Futhermore, the binding sites for a transcription factor can vary between species. Therefore, in order to describe the binding sites for transcription factors, a number of reprsentations have been developed to account for site variability. One popular method is the position frequency matrix, which is used to characterize known TFBSs that are available in online databases such as TRANSFAC [11] and JASPAR [12]. This matrix records the frequency with which each base appears at each location in the binding sites for the same transcription factor across different sequences. Because the matrix is not specific to a single species or gene, it can be helpful in identifying binding sites for a transcription factor in other contexts. For example, Table I is a position frequency matrix constructed from 2000 binding sites for the RUNX1 transcription factor, which appears in the JASPAR database. The sequence logo in Figure 1 is a graphical way of showing the position frequency matrix, in which the height of each letter indicates the degree of conservation at each location. Each of these representations make it clear that there is a strongly conserved TGTGG motif at the core of the RUNX1 binding site. Thousands of these matrices now exist, characterizing many known transcription factor binding sites, many of which exhibit highly conserved core regions [10].

The greater conservation of core regions may suggest that they have a more critical role in the function of a TFBS. Nevertheless, studies have shown that the highly conserved core is often not, in itself, sufficient for binding a transcription factor. For example, a study of the Nanog transcription factor [13] showed that despite identifying a highly conserved core region of 6 nt (nucleotides) in a Nanog binding site, 14 nt were required for Nanog to actually bind to the promoter. A study of the WRKY transcription factors in Arabidopsis [14] showed that despite being known for binding to a highly conserved Wbox element, flanking DNA sequences were partly responsible for specific binding of different WRKY family members. Other reasearchers have suggested that in some cases the less conserved flanking regions might be species-specific areas of the TFBS [15].

The biological importance of TFBS core regions was shown in an analysis of nucleotide conservation in bacterial protein- DNA complexes [19]. This work showed a significant correlation between evolutionary conservation and areas of contact between a transcription factor and its binding site. A later study suggested that the same result holds for yeast [20], a eukaryote. In addition, some X-ray crystallography has provided direct evidence of transcription factor contacts with DNA and the core region of a binding site [17], [13]. Figure 2 shows an example of a transcription factor making contact with its respective binding site. The computer generated image is based on X-ray crystallography of the RUNX1 transcription factor binding to DNA [17]. The points of contact between the transcription factor and the DNA fall within the core region of the TFBS [21].

The existence of core regions in TFBSs may be useful in comparative genomics approaches to studying transcriptional regulation, such as GAMI. By exploiting the presence of core regions, such approaches may be able to incorporate a more biologically accurate model when trying to identify candidate functional elements. This may help reduce the number of false positives, i.e. identified sites that are non-functional. Currently, gene regulation cannot reliably be worked out computationally. Instead, computational approaches are frequently used to narrow the range of tests that need to be performed in the wet lab. Therefore, reducing the number of false positives is important in decreasing the time and cost of these experiments. By extending GAMI to incorporate the notion of such conserved core sequences in transcription factor binding sites, we hope to increase the quality of the candidate sites that GAMI identifies.

B. The GAMI Algorithm

The target of GAMI's search is an N-mer that appears at least once in each input sequence. However, we allow imperfect matches, so a motif does not need to be fully represented in a sequence. Instead, N-mers that match more strongly are considered stronger motifs. The N-mer itself is a sequence of N bases from the set $\{A, C, G, T\}$. For example, if we are search for 8-mers, possible motifs identified include CATGCAAT, TAGGAACT, ACTTACGT, etc.

Aside from exhaustive search, there is not an algorithmic way to calculate the best motifs for a set of sequences, and depending on the number of sequences being examined, the sequence length, and the motif length, exhaustive search can be prohibitively computationally expensive. Therefore, most approaches to motif inference use some sort of heuristic search technique; GAMI uses a GA search.

III. Modifications to GAMI for this Work

Our extension to GAMI will focus on the fitness evaluation of candidate solutions. For this work, we weighted the fitness of candidate solutions to reward those containing the kinds of conserved core sequences talked about earlier. This should result in a higher probability that the final GAMI population will contain motifs exhibiting such cores.

The GAMI fitness evaluation uses a metric called "match percent" (MP). To evaluate the MP of a given motif, the best consecutive match for that motif is located within each sequence.

There may be more than one best match for a given motif and nucleotide sequence (but this does not alter the MP score). An example match for the motif AATTAGTG is shown in Table II. The (maximum) number of bases matched in each sequence is the score for that motif with that sequence. The score for the motif across all sequences in the data is the percent of the overall matches found out of the theoretical maximum possible (number of input sequences × motif length).

As mentioned above, our approach is to modify the way that GAMI calculates fitness. MP finds the number of matches for a motif in each sequence to calculate the total percentage of matches, which is used as the motif's fitness, as shown in Table II. Our new scoring method locates the best match for a motif in each sequence and counts the number of matches per column in the motif. Then, it slides a "window" through the column scores to identify a cluster of the highest values (III), which we will consider the core region of the motif. In both Table II and Table III, the number of matches is the same, but once the core is identified (the bases with highlighted scores), the score is weighted so the fitness incorporates both the MP of the entire motif and the MP of just the core region (the weight is user-defined).

Therefore, we added a new parameter to GAMI called *Core Score Weight* that represents the balance between the importance of the core region score and the overall motif score to a motif's fitness. Core Score Weight is specfied as a real value from 0 to 1. A value of 0 means that GAMI should run as normal. A value of .5 means the core region contributes as much to fitness as the overall motif score. A value of 1 would mean that the core region is the only part of the motif considered for fitness; this is disallowed, as it would be better then to run GAMI to look for motifs of the shorter (core region) length directly. The length of the core region is also a parameter to the system.

This approach was implemented to expediently get a sense of the value of incorporating conserved core regions into the search. One difficulty with this design is the fact that GAMI often identifies more than one equally strong match in a sequence for any given motif. These locations are all reported as part of GAMI's final results. However, the implementation just described depends on using a single best match in each sequence. Attempting to consider the strongest window among all best matching locations in each sequence is a combinatoric problem that could greatly slow GAMI's search. However, we are not performing an exhaustive search. Genetic algorithms such as GAMI attempt to find more optimal solutions to complex problems without guaranteeing solutions that are the most optimal. Therefore, for the initial design of this work, when equally good locations for a motif are encountered, one is selected at random for the core region evaluation. The evaluation is done 10 times, and the best score of the 10 is used. In the future, this approach will be modified.

IV. Methodology

In prior work, it was established that GAMI is an effective approach to DNA motif inference. Therefore, for the initial phase of this project, we will focus on comparing the performance of our extensions to GAMI to its original form. To that end, we will consider the following pertinent questions:

1. Are both versions of GAMI able to identify known transcription factor binding sites in our datasets?
2. Are there differences in the types of candidate functional elements identified by each version of GAMI?
3. Does the modified scoring function improve the recovery of known transcription factor binding sites in our data?

A. Data

For the first phase of this project we identified a biological data set containing as many known transcription factor binding sites as possible, without examining those sites in advance to see if they had particular characteristics (such as highly conserved cores). Nevertheless, the structure of TFBSs varies considerably and we will need to test our design on a number of different datasets in the future.

We identified the sequence upstream from the G6PC gene as a good candidate for our work here, based on the number of known TFBSs for this gene listed in the PAZAR database of regulatory sequence annotation [22]. We searched PAZAR for a gene with greater than five TFBSs annotated in humans that were present in any noncoding region, excluding the UTRs (untranslated regions), and which include a variety of different sites. The UTRs were excluded because they are not annotated for all available genomes, making the dataset harder to assemble. Known binding sites for the upstream region of G6PC are listed in Table V.

We then curated the noncoding DNA upstream of G6PC from the human genome, as well as from orthologous sequences in other species, looking in either NCBI's Gene database and the Ensembl database. This gave us sequences from 38 different species. We excluded sequences containing any N's, leaving us with 20 different sequences. The fish species contained more than one copy of the G6PC gene and so only the upstream of the first one annotated was used. The sequences are described in Table IV.

B. Parameter Settings

In this initial work we are comparing our modifications of GAMI to its unmodified form. Therefore, we did not focus on tuning the parameter settings. We simply used settings that have worked well on other projects, since any differences between the orignal GAMI and our modified version should still be apparent.

For all experiments reported here, we used a population size of 1,000, a crossover rate of 0.8, and a mutation rate of 0.02. The number of trials was set at 50,000 (which refers to the number of fitness function evaluations; due to elitism and the ability to recognize when a reproduction operator has no effect, there is not a clean mapping between the number of trials and the number of generations). Fifty percent elitism was used to preserve the best 500 distinct motifs in the population every generation. Therefore, at most 500 new motifs are created every generation, and the result of a run can be considered the 500 best solutions in the final population. The 80 percent crossover rate means that 80 percent of the remaining motifs are candidates for crossover (a total of 400). The 2 percent mutation rate means that a

nucleotide in a solution has a 2 percent chance of being set to a random value (possibly the same as it was before). Rank-based selection was used.

As described previously, two new parameters were added for this work: *Core Region Length and Core Score Weight*. Core Region Length specifies the length of window used to look for the region of highest conservation within a TFBS. Core Score Weight specifies the percentage of the total score that the core region should account for. In the experiments here, we used Core Region Length and Core Score Weights of 0/0, 4/0.1, 4/0.5, and 4/0.9.

We ran GAMI using motif lengths of 6, 7, 8, 9, 12, 13, and 15. These lengths correlate to the lengths of the known binding sites in Table V. Normally, the lengths of candidate functional elements will not be known in advance and we would run GAMI using a variety of motif lengths. Since we have the lengths for known functional elements in this dataset, we ran GAMI with these lengths in order to make the results easier to interpret. However, the motif length does not need to be exact. For example, a motif length of 15 should find the locations of candidate functional elements that are slightly shorter or longer.

C. Experimental Design

We limited the input sequences to the 1000 bases immediately upstream of G6PC in order to make our evaluation more clear. If we included the entire noncoding region upstream of G6PC, then GAMI would identify many candidate solutions but would not be as focused on the area containing the currently known TFBSs. Since we would not be able to tell which method finds more of these unknown TFBSs, this would make it harder to see if the core region extension is beneficial.

As described in the parameters section above, we ran GAMI using four different Core Score Weight settings doing twenty experiments at each setting. (Multiple experiments are called for due to the stochastic nature of genetic algorithms.) This allows us to compare the efficacy of different Core Score Weight settings with this data set. We then checked the results of each set of experiments for the recovery of the known transcription factor binding sites listed in Table V. A successful recovery is defined as a match that is located in the same location in the human genome as one of the binding sites from Table V, rather than matching the sequence of nucleotides given. This is for two reasons:

1. GAMI finds *consensus motifs*. This means that each base in the motif represents the most highly conserved base at that location across all of the input sequences, which may not match the sequence given in Table V.
2. GAMI may find a sequence that matches the known transcription factor binding site at another location in the input sequence. Although it may be functional, the sequence alone does not determine functionality. True functionality is a result of complex interactions including the sequence, the conformation of the strand, interaction of the transcription factor with other proteins or RNA, and many other elements.

Furthermore, we will not demand that the location is a precise match but that it is within one base of the known binding site. Frequently, the areas surrounding a TFBS are also well

conserved, which can lead GAMI to find a motif that is located near the functional site. However, allowing matches to be any further off could confuse which site was found in the case of overlapping TFBSs. For example, the binding sites with accession numbers RS0001888 and RS0001889 (Table V) would be ambiguous if matches were any further off.

V. Results

The results of the experiments are shown in Table VI. In general, the TFBSs with lengths 6-9 were recovered at all settings, although the the placement was slightly less accurate at Core Score Weight (CSW) settings of 0.1 and 0.5 for these TFBSs. Recovery of the longer motifs was less consistent. One of the TFBSs of length 12 (RS0001886) was recovered by the unmodified GAMI and using a CSW of 0.9. The other TFBS of length 12 was not recovered at any setting. Out of the two TFBSs of length 13, RS0144884 was recovered only using a Core Score Weight (CSW) of 0.9, and RS0144885 was only recovered using a CSW of 0.1. Finally, the only setting that recovered the length 15 TFBS, RS0001884, was the CSW of 0.1.

The candidate solutions identified by all experiments were also analyzed for conservation and complexity.

- *Conservation*. The mean conservation of all motifs using their best locations in each sequence. In other words, conservation is a measure of how well the motif matches the input sequences. The results of this analysis are illustrated in Figure 3.
- *Complexity*. The mean complexity of motifs was measured as in Fogel, et al. [10],

$$K = \frac{1}{w} log_N \left(\frac{w!}{\prod n_i!} \right) \quad (1)$$

where K is complexity, w is motif length, $N = 4$ (the number of types of bases), and n_i is the number of nucleotides of type i, where i is an element of $\{A, T, G, C\}$. This is a measure of how well represented each base is in the motif, giving motifs with a more balanced representation of each base a higher score. The complexity score was normalized [1..100] so that it could be compared between motifs of different lengths. The results of this analysis are illustrated in Figure 4.

In Figure 3 there are clear differences in how well motifs are conserved between the unmodified GAMI and runs using various Core Score Weight settings. At all motif lengths, a CSW of 0.1 results in motifs with a higher average conservation across the input sequences. Conversely, a CSW of 0.9 results in motifs with a lower average conservation for motif lengths greater than 7.

Similarly, there are differences in the complexity of motifs recovered at various Core Score Weights (Figure 4). For motif lengths 6-7, the unmodified GAMI recovered more complex

motifs than any of the CSW settings. For motif lengths 9-15 CSW settings of 0.5 and 0.9 recovered more complex motifs than either the unmodified GAMI or a CSW of 0.1 (which were about the same).

VI. Discussion

The results in Table VI show that shorter motifs are more easily recovered than longer motifs in this data set at all Core ScoreWeight Settings. This is not surprising given two factors:

1. The longer the TFBS, the more likely that mutations will occur [23]. If one considers that each base has some probabiliity of mutation, then the longer the sequence considered, the greater the probability that at least one base in that sequence will mutate in any given period of time. Therefore, it is more likely for a longer TFBS to have less conservation than a shorter one in a given set of input sequences, which may make it harder to recover.
2. The length of the core region is not proportional to the length of the TFBS. Many observed core regions are 4-5 bases long. These more highly conserved areas therefore occupy a greater proportion of the length of short TFBSs, possibly making them easier to recover. That being said, as one study showed [10], core regions as annotated in the TRANSFAC database or by tools such as MatInspector [21] use an arbitrary definition of core length, so the relationship of core length to TFBS length is currently hard to assess with certainty.

Table VI reveals that Core Score Weight settings of 0.1 and 0.9 were the most effective at recovering known TFBSs in this data set, each recovering one more site than the unmodified GAMI. When the CSW was set to 0.5, the fewest sites were recovered, and no sites longer than 9 bases were recovered. However, the unmodified GAMI tended to recover the exact location of the known TFBSs more frequently than when the CSW was set.

It is difficult to know exactly why an individual solution occurred, but we can consider the likely impact of the parameters. A CSW of 0.1 means that the core region should comprise 10% of the candidate solution’s fitness score. In some cases this will not be enough to make individuals with strong cores but more weakly conserved flanking areas score better than motifs that have no strong clusters of conservation. However, in other cases, such a setting will provide a small boost to the score of an individual that would normally score a little too low to be preserved. A CSW of 0.5 balances the importance of the core region equally against the overall conservation. A CSW of 0.9 means that nearly all of the motif’s fitness score would be based on its core region. This is akin to searching for short motifs and therefore should have little effect on the recovery of short TFBSs, as in the results we present here. Although the setting would ignore most of the impact of flanking areas in longer motifs, possibly making them harder to identify, it would also be the setting least impacted by noise in the data.

Conservation of motifs across the input sequences is GAMI’s primary filter for identifying candidate functional elements. However, we were not surprised to see that the CSW setting

of 0.9 reduced the overall conservation of most recovered motifs (Figure 3). This setting strongly emphasizes conservation in a relatively short core. Therefore, the overall conservation has only a small impact on a motif's score, and we would expect to see a greater number of motifs with lower overall conservation. However, that does not necessarily make a CSW of 0.9 less useful than other settings. Although conservation can be a good indicator of possible functional areas in the noncoding DNA, conservation does not equal functionality. Furthermore, there are functional elements of noncoding DNA that are not directly involved in transcriptional regulation [24]. Therefore, our goal was to filter out some of the highly conserved but less likely to be transcriptionally functional elements in favor of those that show characteristics more in line with those observed in true transcription factor binding sites.

Figure 4 shows that the core region extension to GAMI finds motifs that have a different structure than those found by the unmodified GAMI. In particular, at CSW settings of 0.5 and 0.9, the motifs were less likely to be homogeneous sequences of a single nucleotide, for example AAAAAAAAAAAA or other simple sequences. Although such sequences can play a functional role [24], they are less likely to be the TFBSs we are searching for. After all, TFBSs function as recognition sequences, and homogeneous sequences carry little information. Nonetheless, a CSW setting of 0.1 recovered as many known TFBSs as a CSW of 0.9, so it may be that the complexity is not as important as the core region in filtering for biological function.

It is interesting that the most extreme CSW settings of 0.1 and 0.9 recovered the most known TFBSs in our dataset. In fact, between the two settings, all but one of the known sites were recovered. This may reflect the varying structure of TFBSs and suggest that different settings will aid in the recovery of different types of TFBSs. For example, a low CSW may help us identify the true TFBSs in areas of higher conservation, while a high CSW may help us identify TFBSs in areas of noise.

Although our results suggest that looking for motifs with highly conserved core regions is useful in identifying biologically functional TFBSs, this study cannot provide conclusive evidence. There are no noncoding regions for which every TFBS has been annotated. Therefore, we cannot be sure that GAMI identified more or fewer functional sites using any of the settings without validating all our solutions in the wet lab. All we can say for certain is that using CSW settings of 0.1 and 0.9 on this data identified three known TFBSs that were missed using the unmodified GAMI.

It seems likely that the core regions extension to GAMI presented here will be the most beneficial for recovering TFBSs that exhibit strongly conserved cores. Matrices in public databases suggest that many, but not all, TFBSs exhibit these cores. However, some TFBSs seem to contain more than one highly conserved cluster that could be called a core region [10], which our current tool does not address. Therefore, we expect our results will vary by dataset.

It is also worth noting that GAMI sometimes finds more than one motif that refers to the same location, or more than one motif equidistant from the same location (e.g. both 1 base away). In all cases we have listed the results for the highest ranked motif.

Our results suggest that incorporating the notion of highly conserved core regions into the search may be helpful in identifying biologically functional elements. We also found that, at least with this data, Core Score Weights of 0.1 and 0.9 lead to better recovery than other settings. Additionally, we found that in longer motifs (≥ 9 nucleotides), CSW settings 0.5 or 0.9 identified motifs of greater complexity than the original GAMI.

VII. Future Work

The work presented here represents a preliminary investigation into the merits of incorporating the idea of highly conserved core regions into the search for candidate functional elements in noncoding DNA. Much work remains to understand this idea futher. For example, it is important to understand how the TFBS that were not found in Table VI scored, relative to others in the final GAMI population to understand why they are not found by the search. Additionally, of course, it is important to assess this approach on additional datasets. Furthermore, the method of selecting the location of the core is unsatisfactory and needs refinement, for example, by adding an integer to the GA string to represent the location of the core, so that that facet of the solution will be subject to search. Finally, it might be advantageous to adapt this idea to handle cases in which there is more than one highly conserved core.

Acknowledgments

This project was supported by grants from the National Center for Research Resources (5 P20 RR024475-02) and the National Institute of General Medical Sciences (8 P20 GM103534-02) from the National Institutes of Health, a National Science Foundation (NSF) CAREER award (#953495), and NSF Cooperative Agreement No. HRD-0833567.

References

1. Congdon CB, Aman JC, Nava GM, Gaskins HR, Mattingly CJ. An evaluation of information content as a metric for the inference of putative conserved noncoding regions in DNA sequences using a genetic algorithms approach. IEEE/ACM Trans Comput Biol Bioinform. 5:1–14.2008; [PubMed: 18245871]
2. Congdon CB, Fizer C, Smith NW, Gaskins HR, Aman JC, Nava GM, Mattingly CJ. Preliminary results for gami: A genetic algorithms approach to motif inference. CIBCB'05. 2005:97–104.
3. Cooper GM, Brudno M, Green ED, Batzoglou S, Sidow A. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. Genome Res. 13:813–820.May.2003 [PubMed: 12727901]
4. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 31:374–378.Jan.2003 [PubMed: 12520026]
5. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC, Maskeri B, Hansen NF, Schwartz MS, Weber RJ, Kent WJ, Karolchik D, Bruen TC, Bevan R, Cutler DJ, Schwartz S, Elnitski L, Idol JR, Prasad AB, Lee-Lin SQ, Maduro VV, Summers TJ, Portnoy ME, Dietrich NL, Akhter N, Ayele K,

Benjamin B, Cariaga K, Brinkley CP, Brooks SY, Granite S, Guan X, Gupta J, Haghighi P, Ho SL, Huang MC, Karlins E, Laric PL, Legaspi R, Lim MJ, Maduro QL, Masiello CA, Mastrian SD, McCloskey JC, Pearson R, Stantripop S, Tiongson EE, Tran JT, Tsurgeon C, Vogt JL, Walker MA, Wetherby KD, Wiggins LS, Young AC, Zhang LH, Osoegawa K, Zhu B, Zhao B, Shu CL, De Jong PJ, Lawrence CE, Smit AF, Chakravarti A, Haussler D, Green P, Miller W, Green ED. Comparative analyses of multi-species sequences from targeted genomic regions. Nature. 424:788–793.Aug.2003 [PubMed: 12917688]

6. Dubchak I, Brudno M, Loots GG, Pachter L, Mayor C, Rubin EM, Frazer KA. Active conservation of noncoding sequences revealed by three-way species comparisons. Genome Res. 10:1304–1306.Sep.2000 [PubMed: 10984448]
7. Elnitski L, Jin VX, Farnham PJ, Jones SJ. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. Genome Res. 16:1455–1464.Dec.2006 [PubMed: 17053094]
8. Zhang Z, Gerstein M. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. J Biol. 2:11.2003; [PubMed: 12814519]
9. Laurila K, Lahdesmaki H. Systematic analysis of disease-related regulatory mutation classes reveals distinct effects on transcription factor binding. In Silico Biol (Gedrukt). 9(4):209–224.2009; [PubMed: 20109151]
10. Fogel GB, Weekes DG, Varga G, Dow ER, Craven AM, Harlow HB, Su EW, Onyia JE, Su C. A statistical analysis of the TRANSFAC database. BioSystems. 81(2):137–154.Aug; 2005 [PubMed: 15941617]
11. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer M, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 34:D108–110.Jan.2006 [PubMed: 16381825]
12. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. 38(Database issue):D105–110.Jan; 2010 [PubMed: 19906716]
13. Jauch R, Ng C, Saikatendu K, Stevens R, Kolatkar P. Crystal structure and DNA binding of the homeodomain of the stem cell transcription factor Nanog. Journal of molecular biology. 376(3): 758–70.Feb; 2008 [PubMed: 18177668]
14. Ciolkowski I, Wanke D, Birkenbihl RP, Somssich IE. Studies on DNA-binding selectivity of WRKY transcription factors lend structural clues into WRKY-domain function. Plant molecular biology. 68(1-2):81–92.Sep; 2008 [PubMed: 18523729]
15. Wang X, Tomso DJ, Chorley BN, Cho HY, Cheung VG, Kleeberger SR, Bell DA. Identification of polymorphic antioxidant response elements in the human genome. Hum Mol Genet. 16(10):1188–1200.May; 2007 [PubMed: 17409198]
16. Moreland JL, Gramada A, Buzko OV, Zhang Q, Bourne PE. The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications. BMC Bioinformatics. 6:21.2005; [PubMed: 15694009]
17. Tahirov TH, Inoue-Bungo T, Morii H, Fujikawa a, Sasaki M, Kimura K, Shiina M, Sato K, Kumasaka T, Yamamoto M, Ishii S, Ogata K. Structural analyses of DNA recognition by the AML1/Runx-1 Runt domain and its allosteric control by CBFbeta. Cell. 104(5):755–67.Mar; 2001 [PubMed: 11257229]
18. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol. 112(3):535–542.May; 1977 [PubMed: 875032]
19. Mirny LA, Gelfand MS. Structural analysis of conserved base pairs in protein-DNA complexes. Nucleic Acids Res. 30(7):1704–1711.Apr; 2002 [PubMed: 11917033]
20. Morozov AV, Siggia ED. Connecting protein structure with predictions of regulatory sites. Proceedings of the National Academy of Sciences of the United States of America. 104(17):7068–73.Apr; 2007 [PubMed: 17438293]

21. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T. MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics. 21:2933–2942.Jul.2005 [PubMed: 15860560]

22. Portales-Casamar E, Arenillas D, Lim J, Swanson MI, Jiang S, McCallum A, Kirov S, Wasserman WW. The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. Nucleic Acids Res. 37(Database issue):54–60.Jan; 2009

23. Stewart AJ, Plotkin JB. Why Transcription Factor Binding Sites are Ten Nucleotides Long. Genetics. Aug.2012

24. Kashi Y, King DG. Simple sequence repeats as advantageous mutators in evolution. Trends Genet. 22(5):253–259.May; 2006 [PubMed: 16567018]

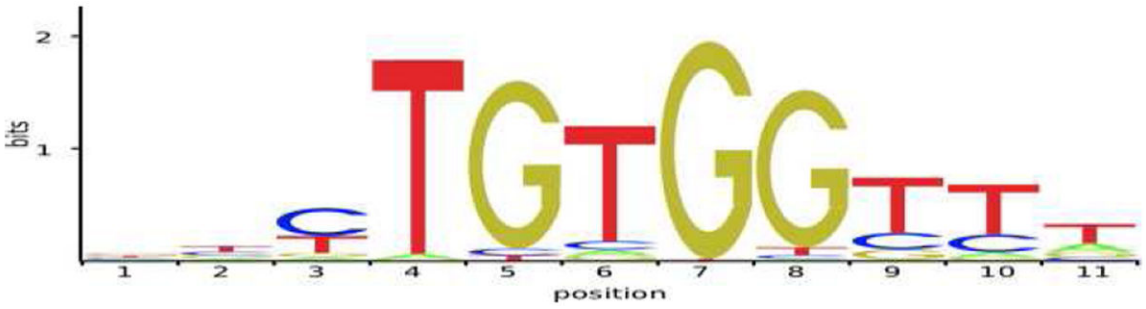

Fig. 1.
Sequence logo of the RUNX1 transcription factor from the JASPAR database.

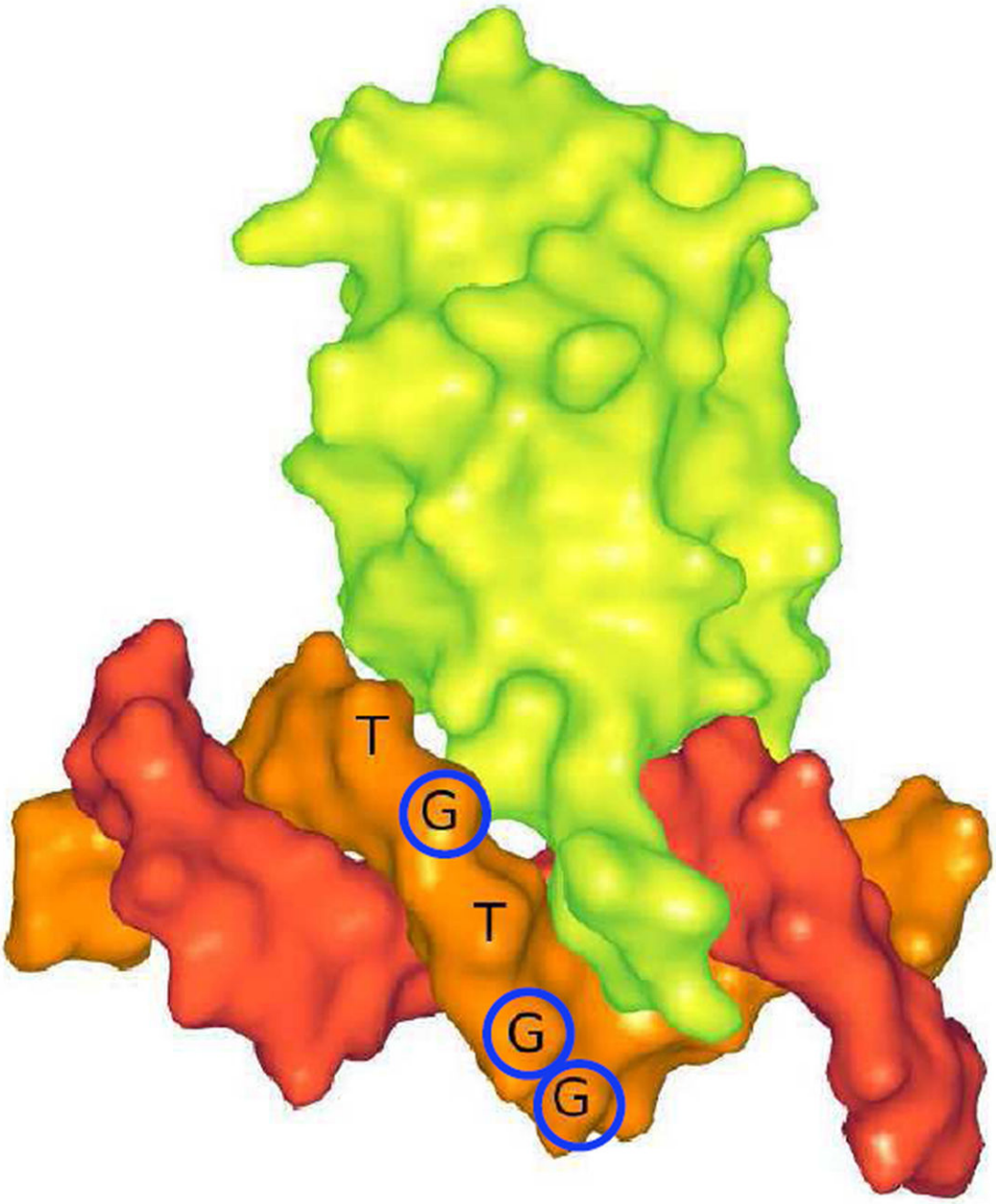

Fig. 2.

A computer generated image of the RUNX1 transcription factor (green) binding to DNA (red and orange). The core region of highest conservation in its JASPAR matrix is labelled on the image. Protein-DNA contact points are circled in blue. The image was generated using the Protein Workshop [16] from data obtained by X-ray crystallography [17] and retrieved from the Protein Data Bank [18] (PDB ID: 1HJC).

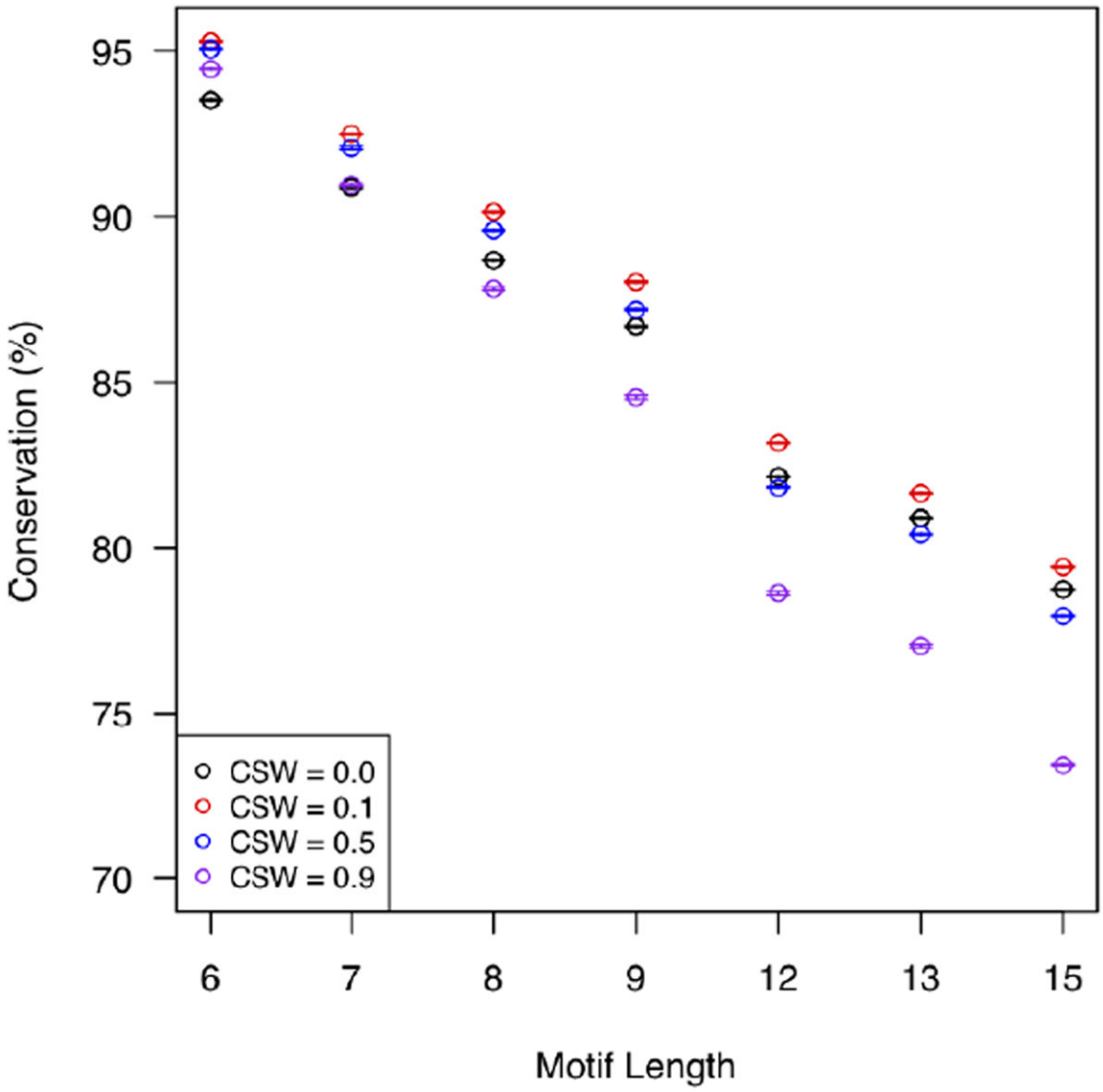

Fig. 3.

Mean conservation of motifs across all runs. The conservation is given as a percentage of matches through the input sequences. A Core Score Weight (CSW) of 0.0 refers to the unmodified GAMI.

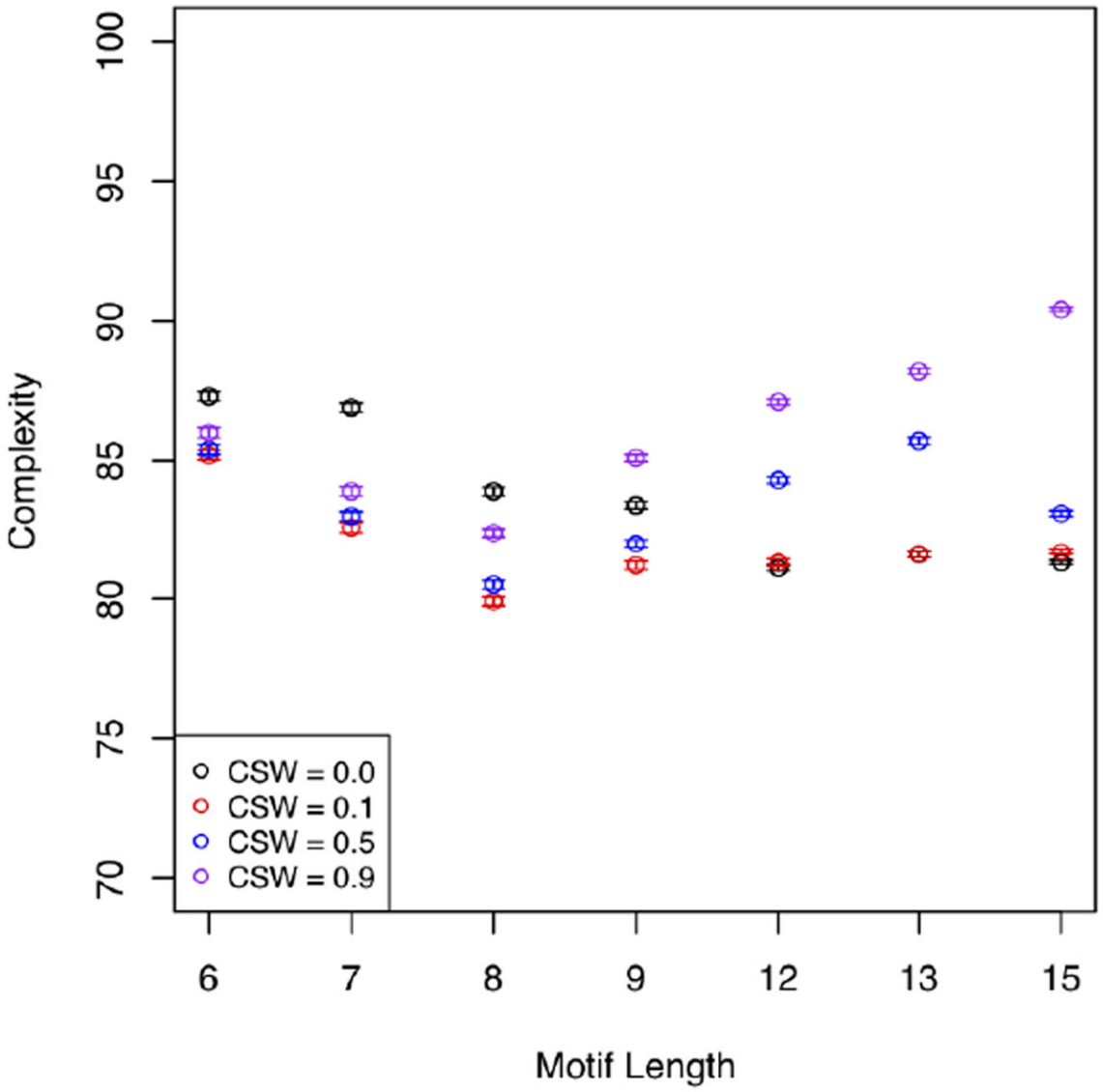

Fig. 4.
Mean complexity of motifs across all runs. Complexity is calculated as in Equation 1.

TABLE I

Position Frequency Matrix for the RUNX1 Transcription Factor from fhe JASPAR Database

Base	1	2	3	4	5	6	7	8	9	10	11
A	287	234	123	57	0	87	0	17	10	131	500
C	496	485	1072	0	75	127	0	42	400	463	158
G	696	467	149	7	1872	70	1987	1848	251	81	289
T	521	814	656	1936	53	1716	13	93	1339	1325	1053

IEEE Symp Comput Intell Bioinforma Comput Biol Proc. Author manuscript; available in PMC 2019 April 19.

TABLE II

Motif AATTAGTG matching short sequences from the PAZAR database [22] that contain binding sites for HOX5A. The motif location is shown in blue in the context of the sequence, and then as a set of local-alignment windows, with the mismatched bases shown in red. In this example, the motif matches 32/40 bases, resulting in a fitness of 80%.

PAZAR Regseq ID	Sequence	Motif	Matches
RS0001715	ttgtcacggcggatAATTTATGag	AATTTATG	6/8
RS0001716	acagcgttactAATTACAGcccccc	AATTACAG	6/8
RS0001717	ttgtcaggggagatAATTTATGgg	AATTTATG	6/8
RS0001718	acagcgttactAATTAGAGcccccc	AATTAGAG	7/8
RS0001719	ccaactccccCATTAGTGcacgag	CATTAGTG	7/8

IEEE Symp Comput Intell Bioinforma Comput Biol Proc. Author manuscript; available in PMC 2019 April 19.

TABLE III

Close up of the Motif column from Table II. However, in this table, the motif is scored by column rather than by sequence. In either case, the motif matches 32/40 bases, resulting in a fitness of 80%. The window of the highest conservation (the core region) is highlighted in yellow. If *core weight* is set to .5, then the fitness is (19 × .5 × 100 / 20) + (32 × .5 × 100 / 40) = 87.5%

A	A	T	T			T	G
A	A	T	T	A			G
A	A	T	T			T	G
A	A	T	T	A	G		G
	A	T	T	A	G	T	G
4/5	5/5	5/5	5/5	3/5	2/5	3/5	5/5

TABLE IV

G6PC Sequences Used in This Work

Common Name	Database	Accession Number	Version	Indices	Strand
Human	Gene	GI:224589808	GRCh37.p9	41051814-41052813	+
Gorilla	Ensembl	ENSGGOG00000021956	68.31	41133495-41134494	−
Rabbit	Ensembl	ENSOCUG00000016018	68.3	43594064-43595063	+
Guinea Pig	Ensembl	ENSCPOG00000013409	68.3	12488849-12489848	+
Mouse	Gene	GI:372099099	GRCm38 C57BL/6J	101366730-101367729	+
Rat	Gene	GI:389675119	Rnor_5.0	89083064-89084063	+
Squirrel	Ensembl	ENSSTOG00000022070	68.2	17604940-17605939	+
Hamster	Gene	GI:351517471	CriGri_1.0	1022475-1023474	+
Microbat	Ensembl	ENSMLUG00000004107	68.2	1069470-1070469	−
Horse	Ensembl	ENSECAG00000000509	68.2	20213706-20214705	−
Cow	Gene	GI:258513348	Bos_taurus_UMD_3.1	43589846-43590845	+
Dog	Gene	GI:357579622	CanFam3.1	20143643-20144642	−
Anole	Ensembl	ENSACAG00000013598	68.2	62425618-62426617	+
Turkey	Ensembl	ENSMGAG00000004239	68.21	4698910-4699909	+
Fugu	Ensembl	ENSTRUG00000006732	68.4	339213-340212	+
Medaka	Ensembl	ENSORLG00000018711	68.1	1325950-1326949	−
Tetraodon	Ensembl	ENSTNIG00000014413	68.8	1345113-1346112	−
Tilapia	Ensembl	ENSONIG00000018273	68.1	3320436-3321435	+
Zebrafish	Gene	GI:312144718	Zv9	6717275-6718274	−
Lamprey	Ensembl	ENSPMAG00000002488	68.7	9548-10547	−

TABLE V

Known TFBSs in the Human Genome Upstream of G6PC from PAZAR Database

PAZAR Regseq ID	Transcription Factor	Binding Site Sequence	Strand
RS0001890	CEBP	CAACCT	+
RS0001887	HNF3	TGTGTGC	+
RS0001888	HNF3	TGTTTGC	+
RS0001891	HNF3	ACAAACG	+
RS0001893	HNF3	CCAAAGA	+
RS0001889	CRE	TTGCATCA	+
RS0001892	HNF3	GTTTTTGAG	+
RS0001885	HNF4	AAGAAGCATGCC	+
RS0001886	HNF4	GCCAAAGTTAAT	+
RS0144884	HNF4	AGTGCAAGGGTCT	-
RS0144885	HNF4	AGGACAGAGTCTA	-
RS0001884	HNF1	AGTTAATCATTGGCC	+

TABLE VI

Recovery of Known TFBSs Using Various Core Score Weight (CSW) Settings. CSW = 0.0 shows the results for the unmodified GAMI. If the TFBs was recovered within one base position, it is indicated with a score of -1, 0, or +1, showing how close GAMI's solution was to the known TFBs location. Therefore, A 0 indicates a perfect match. Non-recovery within that range is indicated with a dashed line.

PAZAR Regseq ID	Binding Site Sequence	CSW = 0.0	CSW = 0.1	CSW = 0.5	CSW = 0.9
RS0001890	CAACCT	+0	-1	-1	+0
RS0001887	TGTGTGC	+0	+0	+0	+0
RS0001888	TGTTTGC	+0	+0	+0	+0
RS0001891	ACAAACG	+0	-1	+0	+0
RS0001893	CCAAAGA	+0	+0	+0	+0
RS0001889	TTGCATCA	+0	+0	+0	+0
RS0001892	GTTTTTGAG	+0	+0	+0	+0
RS0001885	AAGAAGCATGCC	–	–	–	–
RS0001886	GCCAAAGTTAAT	+0	–	–	-1
RS0144884	AGTGCAAGGGTCT	–	–	–	-1
RS0144885	AGGACAGAGTCTA	–	+1	–	–
RS0001884	AGTTAATCATTGGCC	–	-1	–	–

IEEE Symp Comput Intell Bioinforma Comput Biol Proc. Author manuscript; available in PMC 2019 April 19.