Published online 15 October 2008

Nucleic Acids Research, 2008, Vol. 36, No. 21 ***e142***
doi:10.1093/nar/gkn738

Evolutionary computation for discovery of composite transcription factor binding sites

Gary B. Fogel[1], V. William Porto[1], Gabor Varga[2], Ernst R. Dow[2], Andrew M. Craven[2], David M. Powers[2], Harry B. Harlow[2], Eric W. Su[2], Jude E. Onyia[2] and Chen Su[2,*]

[1]Natural Selection, Inc., 9330 Scranton Rd., Suite 150, San Diego, CA 92121 and [2]Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, IN 46285, USA

Received July 7, 2008; Revised September 5, 2008; Accepted October 2, 2008

ABSTRACT

Previous research demonstrated the use of evolutionary computation for the discovery of transcription factor binding sites (TFBS) in promoter regions upstream of coexpressed genes. However, it remained unclear whether or not composite TFBS elements, commonly found in higher organisms where two or more TFBSs form functional complexes, could also be identified by using this approach. Here, we present an important refinement of our previous algorithm and test the identification of composite elements using NFAT/AP-1 as an example. We demonstrate that by using appropriate existing parameters such as window size, novel-scoring methods such as central bonusing and methods of self-adaptation to automatically adjust the variation operators during the evolutionary search, TFBSs of different sizes and complexity can be identified as top solutions. Some of these solutions have known experimental relationships with NFAT/AP-1. We also indicate that even after properly tuning the model parameters, the choice of the appropriate window size has a significant effect on algorithm performance. We believe that this improved algorithm will greatly augment TFBS discovery.

INTRODUCTION

Transcription factors (TFs) are key regulatory proteins that are commonly associated with coexpressed genes. Following microarray analysis, examination of the upstream regions of coexpressed genes may lead to the identification of common features, such as TF binding sites (TFBS) that facilitate coregulation (1–3). Computational approaches have been offered to assist in the discovery of these elements (4–6) including evolutionary algorithms (7–14). The low signal-to-noise ratio of the generally short TFBS *n*-mer sequences relative to the much longer upstream region makes identification and prediction of TFBS difficult. Improved modeling of the background sequence distribution for upstream regions that do not contain TFBS can assist in TFBS discovery by increasing the signal-to-noise ratio (15). For example, it is possible to make use of exhaustive calculation to evaluate the frequency of occurrence of all possible *n*-mers to update a nucleotide probability matrix. The distribution of *n*-mers sampled from sets of genes known to be independently regulated can be used as a basis for comparison to the entire distribution of *n*-mers sampled from a genome and any *n*-mers that are more overly abundant than the expectation are labeled as putative TFBSs (6,16–20). This approach is guaranteed to identify *n*-mers with the highest *Z*-scores (a length and composition correct measure of similarity) if there are no substitutions in the matching segments and if $n < 7$ to afford computation in reasonable time (5). It is also possible to specify the *n*-mer as a probability matrix and utilize an iterative approach, such as an expectation maximization (21) or Gibbs sampling procedure (5,22–27). These approaches can increase the *n*-mer length to $n > 8$ nt but are also susceptible to local entrapment during optimization. Improved models of the background sequence distribution are also possible using methods, such as higher order Markov models (15) or discriminative seeding motif discovery (28), but these approaches are organism (or even gene cluster) specific. Approaches for TFBS discovery that are not organism specific, do not require exhaustive calculation, and report back to the user-putative TFBS locations are required by the community.

Previously, we developed an approach for TFBS discovery using evolutionary computation (EC) that met these conditions (29). Testing of the algorithm was with respect to single TFBS considered in isolation (Oct-1 and NF-κB). However it is known that composite binding sites exist in upstream regions, such as the binding site for NFAT/AP-1, a heterodimer form of a TF complex. In general, portions of TF complexes take the form of homodimers or heterodimers, with TFs binding to nonadjacent DNA sites

*To whom correspondence should be addressed. Tel: +1 317 277 9657; Fax: +1 317 276 6009; Email: chen_su@lilly.com

due to DNA tertiary structure. It is therefore a technical challenge to predict the binding sites of TF complexes *in silico*, because the distance between individual TFBSs is not well understood and because different TFs can form complexes with different partners in different biological contexts. Despite these challenges, it is important to understand TF complexes and TFBSs, especially given that for most eukaryotes, TFs work in complexes and because the content of the TF complex is a reflection of the biological conditions of the cell.

ALGORITHM

The code for EC from the previous effort (29), for the discovery of similar TFBS motif 'windows', one for every upstream region, was adopted for TF complex discovery. The sequence information contained in the grouping of these windows was used to calculate a nucleotide likelihood matrix. Fitness was measured with respect to a basis matrix where both a metric for similarity and a metric for complexity were used. Additional details on the implementation can be found in Ref. (29). Revisions and additions to the previous approach are detailed below and afford the possibility for composite TFBS element discovery.

Complexity normalization

The previous approach to TFBS discovery made use of compositional complexity as a portion of the objective function. This remained unnormalized with respect to window size, and therefore maximum complexity was highly dependent on the choice of window size. The compositional complexity calculation was revised such that the maximum possible complexity score is calculated for the user-defined window size prior to EC. During evolution, each complexity score was rescaled between [0, 1] where the maximum possible complexity score was 1. This removed any potential bias in complexity relative to window size.

Ambiguous nucleotide assignment

Ambiguous nucleotide positions resulting from sequencing errors may exist in upstream sequences examined. We determined that these positions can have a significant and undesired effect on the way in which complexity is calculated when using the equation given in Ref. (30). Previously, any ambiguous positions were simply ignored during complexity and fitness calculation. Given this, Ns in the sequence windows could affect fitness by artificially increasing the complexity score for a solution beyond a theoretical maximum complexity sequence composed of nucleotides A, T, C, G. This was corrected by applying a score of 0.25 for any N under the assumption of possibly equal representation of A, T, C, G at that position (Figure 1). The adjustment can be observed in Figure 1C where the last position of sequence 3 is given as an N. Comparison of the scores for complexity and similarity under this revised condition relative to Figure 1A and B show that the revised scoring scheme separates a matched position (a C in position 8 of sequence 3, Figure 1A) from a mismatch (a T in position 8 of sequence 3, Figure 1D)

```
(A) Fitness 1.000000 (Complexity Score 1.000000, Similarity
Score 1.000000):
Sequence 1:   AGTCAGTC
Sequence 2:   AGTCAGTC
Sequence 3:   AGTCAGTC

(B) Fitness 0.766667 (Complexity Score 1.000000, Similarity
Score 0.666667):
Sequence 1:   AGTCAGTC
Sequence 2:   AGTCAGTC
Sequence 3:   NNNNNNNN

(C) Fitness 0.968795 (Complexity Score 0.993204, Similarity
Score 0.958333):
Sequence 1:   AGTCAGTC
Sequence 2:   AGTCAGTC
Sequence 3:   AGTCAGTN

(D)Fitness 0.936490 (Complexity Score 0.982743, Similarity
Score 0.916667):
Sequence 1:   AGTCAGTC
Sequence 2:   AGTCAGTC
Sequence 3:   AGTCAGTT
```

Figure 1. (**A–D**) Complexity and similarity scoring improvement. (A) A set of sequences of identical sequence and length result correctly in a score of 1 for complexity and 1 for similarity. (B) The addition of a region of all N's in sequence 3 causes the complexity to remain the same at 1, while lowering the similarity to 0.67. (C) A score of 0.25 was used for any position that was an N, under the assumption that position truly represented an equal probability of A, T, C or G. This adjustment can be observed in (C) above where only one position (the last position of sequence 3) is an N. Comparison of the scores for (A) and (C) above demonstrates that this method now adjusts both complexity and similarity accordingly. The N position in solution (C) above might contain A, T, C or G. In the case where that position truly is a C, the resulting solution will score as given in (A) above. In the case where that position is A, T or C [shown as a T in (D) above], the solution receives a different and lower complexity and similarity score than where it is an N.

from an N at that same position (Figure 1C) where there is still a possibility for either a C or a T. Any IUPAC symbols other than A, T, C, G or N remain ignored during calculation of fitness given that their occurrence in upstream regions is considered infrequent.

Bonus scoring for similarity

Within TFBSs it is generally accepted that 'core' sequences are conserved relative to flanking nucleotide positions (31). To incorporate this notion into our scoring method for similarity, we added a 'central weighting' bonus to the similarity score. This bonus increases the overall fitness of putative motifs that were similar in the core region rather than at the ends. The 'weight' of sequence similarity was adjusted by position over a window following a Gaussian distribution. Each column's similarity score was scaled according to this arbitrarily chosen Gaussian distribution (Figure 2). We also incorporated an 'adjacent conservation' bonus for adjacent columns in the nucleotide likelihood matrix with 100% conservation. A bonus of 1 was given for every two adjacent columns that met this condition and this bonus was applied only to the calculation of similarity (Figure 3).

Self-adaptation

Self-adaptation of the number of variations performed at each generation as well as the probability associated with each variation operator was added to the EC.

```
                    A    A    A    T    G    C    G    G
                    G    G    A    T    G    C    C    C
                    C    C    A    T    G    C    T    T
                    T    T    A    T    G    C    A    A

Column Score:      −.5  −.5    1    1    1    1  −.5  −.5
Motif Score = Sum(Column_score)/(n of Columns) = 2/8 = .25
Weights:            .5  .68  .873   1    1  .873  .68   .5
Revised score:    −.25 −.34  .873   1    1  .873 −.34 −.25
Revised Motif Score = 2.565/8 = .3206
```

Figure 2. Central weighting scheme for similarity scoring. For the example above, four hypothetical sequences of 8 nt each are shown. The score for their similarity is calculated and sums to 0.25. To this similarity score, a weight range is applied from [0.5, 1.0] where positions of similarity in the inner portion of the motif will essentially receive a bonus for their location relative to the outer positions. The original score is multiplied by the weight and the revised score is used (a sum of 0.3206) for similarity. Note that the weight distribution shown above is based on a Gaussian distribution and can be set to any range or distribution that the user desires.

```
            A  C  G  A   T   T   T   T
            A  C  G  A   G   G   G   G
            A  C  G  A   C   C   C   C
            A  C  G  A   A   A   A   A

score:      1  1  1  1 −.5 −.5 −.5 −.5    = 2/8 =  .25

bonus:         +1 +1 +1                   = +3

new score:                                = (2+3)/8 = .625
```

Figure 3. Adjacent conservation bonus. For the sequences here, a bonus of 3 is given for the three pairs (four columns) of adjacent A's that are 100% conserved. This dramatically increases the similarity score for this solution.

Self-adaptation is a method to automatically evolve and tune the parameters associated with the evolutionary algorithm concurrent with the main process of evolution. In this regard, self-adaptation can be thought of as a hierarchical or 'meta-level' evolutionary optimization process (32). In cases where it is difficult to assign proper weights for parameters by hand, self-adaptation can save time and effort during parameter adjustment.

Automated clustering

A main deficiency of the previous approach for TFBS discovery was the requirement for hand-curation of the top solutions from the evolutionary algorithm into clusters of similar solutions. Clustering is time intensive but results in groups of similar putative motif sequences while simultaneously reducing redundancy of those solutions. We extended the previous approach with an automated clustering method (Figure 4). In order for two solutions to be grouped together in a cluster, we imposed a requirement that 90% of the sequences in each group must be identical. This threshold was set after repeated testing and analysis on the ability of the method to recapitulate hand-curated clusters for Oct-1 and NF-κB.

Parallelization

For parallelization, five Compaq Pentium III machines operating with Linux RedHat 9 were used as a testing environment. Parallelization was made using single master and multiple clients using a 'hub and spoke'

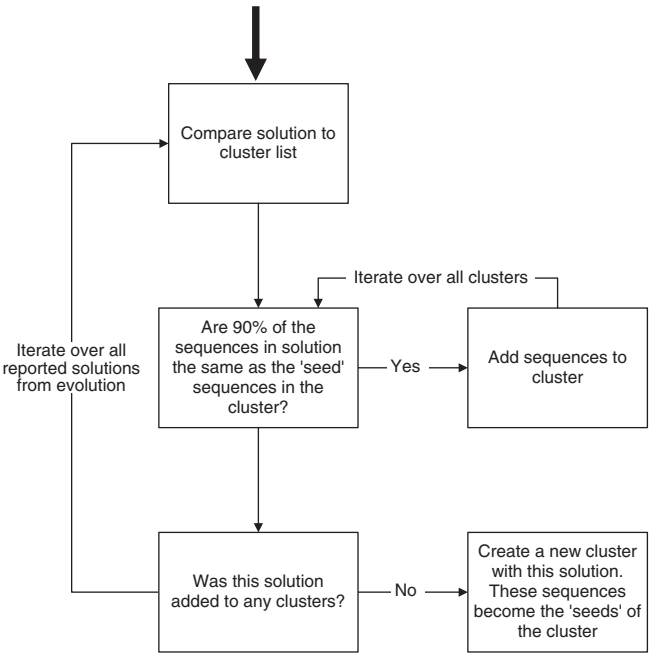

Figure 4. The method of clustering used to assist with final user interpretation of evolved solutions. Starting at the bold arrow, after the final generation of evolutionary optimization, the first two solutions in the output are compared. If ≥90% of the sequences in these solutions are identical, then these two solutions become a 'seed' for a new cluster (cluster #1). If these sequences are <90% identical, then each solution becomes a seed for two separate clusters (clusters #1 and #2). The process of comparison is the iterated with solution #3 in the output file. If solution #3 is ≥90% identical to either cluster #1 or cluster #2, the sequences in solution #3 are merged with the sequences found in the appropriate cluster. Redundant sequences are removed and any sequences remaining are appended to the appropriate cluster. The process is reiterated until all top solutions in the evolutionary search have been partitioned into clusters and the final clustering is presented to the user.

architecture, where the master was the hub and the clients were the spokes. No client processes were run on the master. A user-specified number of client processes was generated each with independent evolutionary optimization. After the first 50 generations, the best results from each client were sent back to the master for sorting and storage. After subsequent 50 generation intervals, best client results were sent to the master process where the results were added to the best result list, the list was sorted and pruned to retain a constant number of best results throughout the evolutionary process. Upon completion of all client evolutionary optimizations, the master process performed an automated clustering of results and output the results to file for user interpretation.

Program implementation

For the experiments presented here, unless otherwise noted, a population of 15 parents and 30 offspring per parent was used with elitist selection. Evolution was allowed to proceed for 1000 generations using all of the above revisions and additions to the code, including central bonus weights, adjacent conservation and self-adaptation. Automated clustering was used to identify clusters of

similar solutions. The window size was then varied starting from 25 and decreasing to 7 to determine if it was possible to find known solutions.

RESULTS

Evaluation of code enhancement using Oct-1 and NF-κB

Complexity scoring. We tested the revised complexity scoring on both the Oct-1 and NF-κB examples from the previous experimentation and determined that for the case of Oct-1, the known Oct-1 TFBS solution dropped from being the top scoring solution to solution #6. In the case of NF-κB, however, the known NF-κB TFBS solution increased in resulting score from solution #29 to solution #6 (Figure 5). Thus, the normalized complexity scoring relative to window sizes and ambiguous nucleotides resulted in the algorithm to be equally successful over both Oct-1 and NF-κB cases with known TFBS motifs recovered in the top 10 solutions.

```
Oct-1 (best)
Solution 1:
Fitness 0.855031 (Complexity Score 0.933438, Similarity Score 0.821429):
                                ****** *
Sequence C00039: (-832,-825)  TTCTGCAC
Sequence C00043: (-533,-526)  TTCTGCAA
Sequence C00048: (-574,-567)  TTCTGCGA
Sequence C00049: (-631,-624)  TTCTGAAA
Sequence C00050: (-41,-34)    GTCTGCAA
Sequence C00158: (-119,-112)  TTCTGCCA
Sequence C00169: (-895,-888)  TTCTGCAA

Oct-1 (truth)
Solution 6:
Fitness 0.849688 (Complexity Score 0.832293, Similarity Score 0.857143):
                                *** ****
Sequence C00039: (-215,-208)  ATGCAAAT
Sequence C00043: (-320,-313)  ATGTAAAT
Sequence C00048: (-904,-897)  ATGTAAAT
Sequence C00049: (-55,-48)    ATGCAAAT
Sequence C00050: (-517,-510)  ATGCAAAG
Sequence C00158: (-806,-799)  ATGCACAT
Sequence C00169: (-96,-89)    ATGCAAAT


NF-kB (best)
Solution 1:
Fitness 0.846272 (Complexity Score 0.870301, Similarity Score 0.833333):
                                ** *****
Sequence C00097: (-808,-801)  TTTCCCAG
Sequence C00099: (-818,-811)  TTTCCAAG
Sequence C00100: (-708,-701)  TTCCCCAG
Sequence C00101: (-582,-575)  TTTCCCAG
Sequence C00152: (-108,-101)  TTTCCCAT
Sequence C00153: (-504,-497)  TTTCCCCG
Sequence C00155: (-883,-876)  TTTCCCAG
Sequence C00165: (-870,-863)  TGTCCCAG
Sequence C00156: (-646,-639)  TTGCCCAG

NF-kB (truth)
Solution 6:
Fitness 0.841463 (Complexity Score 0.959734, Similarity Score 0.777778):
                                ** * ***
Sequence C00097: (-285,-278)  GATATTCC
Sequence C00099: (-138,-131)  GAAATTCC
Sequence C00100: (-247,-240)  GTAACTCC
Sequence C00101: (-675,-668)  GAGATGCC
Sequence C00152: (-790,-783)  GAGATTCC
Sequence C00153: (-789,-782)  GAAAGTCC
Sequence C00155: (-169,-162)  GAAATTCC
Sequence C00165: (-207,-200)  GAAATTCC
Sequence C00156: (-172,-165)  GAGATTCC
```

Figure 5. Solutions for Oct-1 and NF-κB after normalization of the complexity scoring to include ambiguous positions and to account for window sizes. Sequence information is provided with the COMPEL identifier (C00###) and start and end position indices for the motifs relative to the transcription start site.

Bonus scoring for similarity. To test the central bonusing method, two sequences were generated with identical scores under the original scoring method: one where the outer region was conserved and the inner region was not, and another where the inner region was conserved and the outer region was not. For initial tests, the weights for the complexity and similarity scores were 0 and 1, respectively. Sequence complexities were made to be identical so that only the effects of similarity could be observed (Figure 6).

To test the adjacency scoring method, two sets of sequences were generated that scored identically under the original scoring technique. One set of sequences contained no adjacency in the conserved columns whereas the other set of sequences had four adjacent columns that were conserved. The weights used for the complexity and similarity scores were 0 and 1, respectively (Figure 7). The new scoring method indeed generated higher fitness values to putative motifs with conserved 'core' regions.

We evaluated the worth of these modifications on the Oct-1 and NF-κB examples used in Ref. (29). After preliminary testing, the central weights were scaled from 0.8 to 1.0. Using these settings the known Oct-1 solution was reported as solution #6 (Figure 8) and the top two NF-κB solutions were nearly identical to the known NF-κB solution (Figure 9). This suggests that similarity scoring for 'core' regions can drive the algorithm to find more useful solutions. Figures 10 and 11 demonstrate that self-adaptation reduces the number of generations required for the convergence of the population to the known best solutions for the cases of Oct-1 and NF-κB. The output for the automated clustering of the Oct-1 solutions is shown in Figure 12. The known Oct-1 solutions were found in the fourth cluster.

```
Scores Without Central Bonusing:

Outer Conserved Regions:
Fitness 0.250000 (Complexity Score 0.633080, Similarity Score 0.250000):

                 **    **
Sequence 1:      AAATGCGG
Sequence 2:      AATGCAGG
Sequence 3:      AAGCATGG
Sequence 4:      AACATGGG

Inner Conserved Regions:
Fitness 0.250000 (Complexity Score 0.633080, Similarity Score 0.250000):

                   ****
Sequence 1:      AAATGCGG
Sequence 2:      GGATGCCC
Sequence 3:      CCATGCTT
Sequence 4:      TTATGCAA


Scores With Central Bonusing:

(Gaussian distribution scaled from .5 to 1.0)
In this case, the sequences with inner conserved regions score better than
sequences with outer conserved regions.

Outer Conserved Regions:
Fitness 0.061028 (Complexity Score 0.633080, Similarity Score 0.061028):

              **    **
Sequence 1:   AAATGCGG
Sequence 2:   AATGCAGG
Sequence 3:   AAGCATGG
Sequence 4:   AACATGGG

Inner Conserved Regions:
Fitness 0.320613 (Complexity Score 0.633080, Similarity Score 0.320613):

                ****
Sequence 1:   AAATGCGG
Sequence 2:   GGATGCCC
Sequence 3:   CCATGCTT
Sequence 4:   TTATGCAA
```

Figure 6. Fitness of motifs with inner and outer conservations with and without 'central weighting' bonus.

```
Without Adjacency Bonus:
Both set of sequences score the same.

Solution 1:
Fitness 0.250000 (Complexity Score 0.287310, Similarity Score 0.250000):

              * * * *
Sequence 1:   ATATATAT
Sequence 2:   AGAGAGAG
Sequence 3:   ACACACAC
Sequence 4:   AAAAAAAA


Solution 2:
Fitness 0.250000 (Complexity Score 0.287310, Similarity Score 0.250000):

             ****
Sequence 1:   AAAATTTT
Sequence 2:   AAAAGGGG
Sequence 3:   AAAACCCC
Sequence 4:   AAAAAAAA

With Adjacency Bonus Score:
Sequences with adjacent conserved columns scores better.

Solution 1:
Fitness 0.250000 (Complexity Score 0.287310, Similarity Score 0.250000):

             * * * *
Sequence 1:   ATATATAT
Sequence 2:   AGAGAGAG
Sequence 3:   ACACACAC
Sequence 4:   AAAAAAAA

Solution 2:
Fitness 0.6250000 (Complexity Score 0.287310, Similarity Score 0.625000):

             ****
Sequence 1:   AAAATTTT
Sequence 2:   AAAAGGGG
Sequence 3:   AAAACCCC
Sequence 4:   AAAAAAAA
```

Figure 7. Test of the operation of the adjacency bonus method. Solutions 1 and 2 have equivalent scores when using the former method, but with the addition of the adjacency bonus, Solution 2 scores higher in terms of similarity and fitness.

Composite element discovery: NFAT/AP-1

The nuclear factors of activated T-cells (NFAT) family of TFs are known to be involved with the upregulation of T-cell genes following antigenic stimulation (33). Cooperation of NFAT with activating protein 1 (AP-1) provides a model system for combinatorial transcriptional regulation, as NFATp/c factors are known to play a role in cytokine regulation during immune response. Most of the experimental data for NFAT and AP-1 demonstrate that they bind at closely juxtaposed sites.

The NFAT family [NFAT_Q6 in TRANSFAC® Professional version 11.4 (34)] is based on 26 NF-ATp/c binding sites with a resulting nucleotide distribution matrix 12 bp in length (Table 1). NFAT TFBSs are composite elements of an NFATp/c and an AP-1 binding site. The recognition of these NFAT TFBS elements has been investigated previously in the literature (35) providing a very useful means for the evaluation of algorithms designed for composite TFBS discovery.

The nucleotide distribution matrices for NFAT and AP-1 [Figure 1, Ref. (35)] differ only with respect to AP-1. For NFAT, the core was identified as 'GGAAA' with a highly conserved W just upstream of the core. For AP-1, the core was identified as 'TGASTCA' (Table 2). Experimentally determined NFATp/c binding sites are available from a variety of sources in the literature. In addition, 13 known NFAT TFBSs specific for activated T-cells, many of which are NFATp/AP-1 composite elements have been identified [Table 5, Ref. (35)]. We used this information to determine if these motifs could be identified using our approach for TFBS discovery.

To evaluate the performance of this 'Bioinspired Evolutionary Algorithm for Genes with Linked Expression' BEAGLE algorithm, 1000 nt of sequence information upstream of the transcription start site from the

```
Experiment 2:

Oct-1:

Solution 1:
Fitness 0.724142 (Complexity Score 0.659195, Similarity Score 0.751976):

                                  ****** *
Sequence C00039:    (-833,-825)   TTCTGCAC
Sequence C00043:    (-534,-526)   TTCTGCAA
Sequence C00048:    (-575,-567)   TTCTGCGA
Sequence C00049:    (-632,-624)   TTCTGAAA
Sequence C00050:    (-42,-34)     GTCTGCAA
Sequence C00158:    (-120,-112)   TTCTGCCA
Sequence C00169:    (-896,-888)   TTCTGCAA


Solution 2:
Fitness 0.722575 (Complexity Score 0.653972, Similarity Score 0.751976):

                                  *******
Sequence C00039:    (-385,-377)   TCCACAGA
Sequence C00043:    (-814,-806)   TCTACAGA
Sequence C00048:    (-87,-79)     TCCACAGT
Sequence C00049:    (-469,-461)   TGCACATA
Sequence C00050:    (-898,-890)   TCCACAGA
Sequence C00158:    (-963,-955)   TCCACAGG
Sequence C00169:    (-606,-598)   TCCACAGA

Solution 6:
Fitness 0.716322 (Complexity Score 0.587766, Similarity Score 0.771418):

                                  *** ****
Sequence C00039:    (-216,-208)   ATGCAAAT
Sequence C00043:    (-321,-313)   ATGTAAAT
Sequence C00048:    (-905,-897)   ATGTAAAT
Sequence C00049:    (-56,-48)     ATGCAAAT
Sequence C00050:    (-518,-510)   ATGCAAAG
Sequence C00158:    (-807,-799)   ATGCACAT
Sequence C00169:    (-97,-89)     ATGCAAAT
```

Figure 8. Results with reduced bias for inner conservation on the test example of Oct-1. In this case, the known Oct-1 sequences were recovered in Solution 6.

```
Solution 1:
Fitness 0.670436 (Complexity Score 0.650509, Similarity Score 0.681165):

                                   ***** *
Sequence C00097:    (-501,-493)   TGAAATTC
Sequence C00099:    (-140,-132)   GGAAATTC
Sequence C00100:    (-300,-292)   GTAAATGC
Sequence C00101:    (-499,-491)   GGAAATGC
Sequence C00152:    (-519,-511)   GGAAACTC
Sequence C00153:    (-765,-757)   GGAAATGC
Sequence C00155:    (-171,-163)   GGAAATTC
Sequence C00165:    (-209,-201)   AGAAATTC
Sequence C00156:    (-377,-369)   GGAAATCC

Solution 2:
Fitness 0.670224 (Complexity Score 0.673703, Similarity Score 0.668350):

                                  ** ** *
Sequence C00097:    (-500,-492)   GAAATTCT
Sequence C00099:    (-139,-131)   GAAATTCC
Sequence C00100:    (-397,-389)   GGAATACC
Sequence C00101:    (-498,-490)   GAAATGCT
Sequence C00152:    (-791,-783)   GAGATTCC
Sequence C00153:    (-764,-756)   GAAATGCC
Sequence C00155:    (-170,-162)   GAAATTCC
Sequence C00165:    (-208,-200)   GAAATTCC
Sequence C00156:    (-173,-165)   GAGATTCC
```

Figure 9. The NF-κB solutions with a revised complexity score in addition to central weight bonus scaling. The top two solutions contain the known NF-κB motifs.

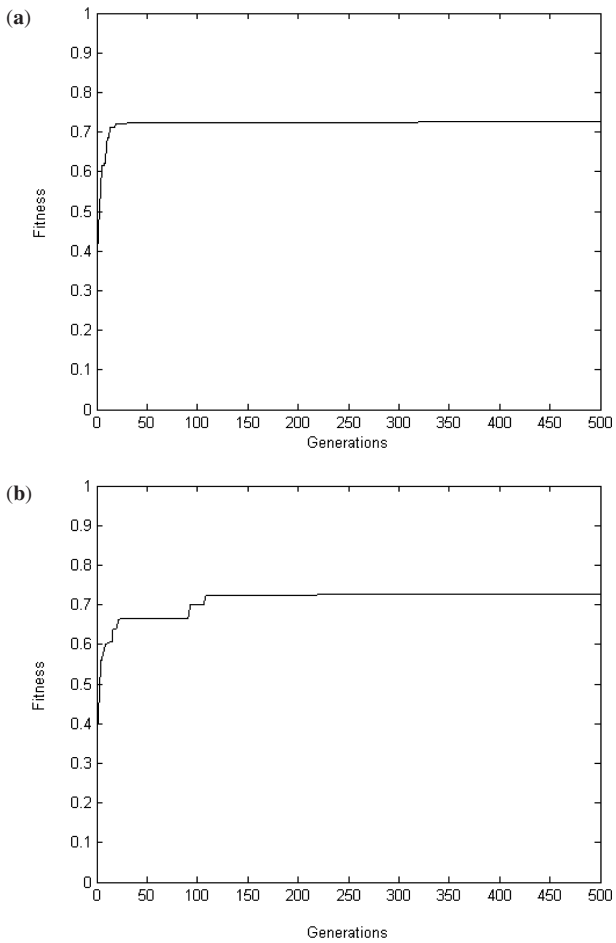

Figure 10. Mean best fitness of evolutionary algorithm (**a**) without self-adaptation and (**b**) with self-adaptation for the Oct-1 example. Note that self-adaptation arrives at the best solution in approximately 125 generations, with a slight further increase in performance at generation 225 (b) whereas the evolutionary algorithm without self-adaptation rapidly becomes stuck in a local optimum until approximately generation 325.

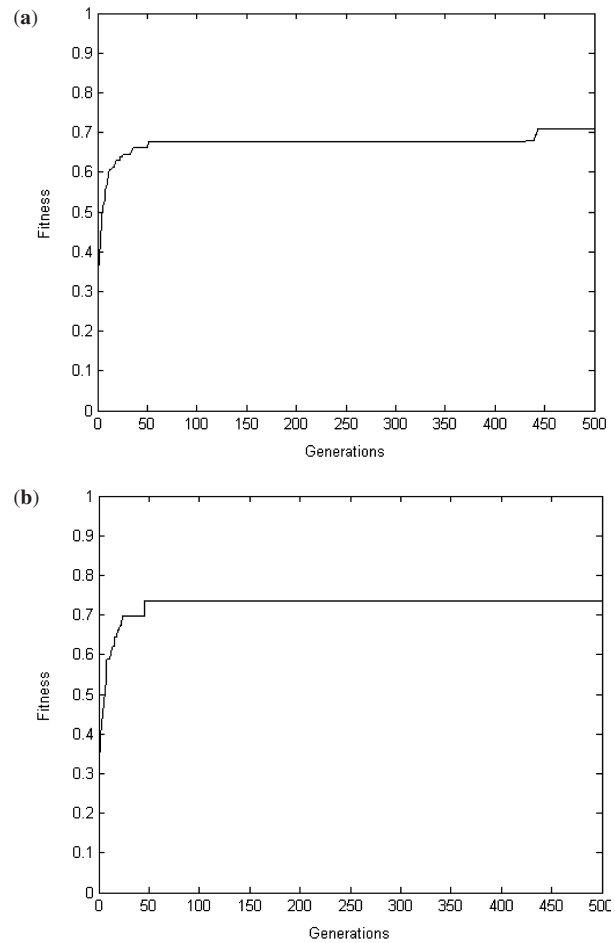

Figure 11. Mean best fitness of evolutionary algorithm (**a**) without self-adaptation and (**b**) with self-adaptation for the NF-κB example. Note that self-adaptation arrives at the best solution in approximately 50 generations and that this solution is better than the best solution discovered without self-adaptation over 500 generations.

following genes: human GM-CSF (NM_000758.2; reverse complement), human IL-2 (NM_000586.2), human IL-4 (NM_000589.2; reverse complement), human TNF member 2 (NM_000594.2; reverse complement), mouse GM-CSF (X03020), mouse IL-2 (X14473), mouse IL-4 (X05064) and mouse IL-5 (D14461) (35). The weights for similarity and complexity were varied over the ranges [0.2, 0.4] and [0.8, 0.6], respectively. Weights of 0.25 for similarity and 0.75 for complexity provided the best solutions after repeated testing.

Using a window size of 25, the top cluster of 26 resulting clusters contained 27 similar putative TFBS motifs, four of which could be identified as known NFAT binding motifs [motifs N2, N3, N13 and N38 from Table 1, Ref. (35)]. Despite the length of the window size relative to the smaller known TFBS, known motifs with conserved similarity as small as length 5 not only appeared in the top solution, but also were grouped automatically in the top cluster. The other 23 solutions in the top cluster had broad similarity in terms of repeated A's central to, or on the 5′-end of, the window. Window sizes as small as 13 nt could be used to recover the known NFAT binding sites. However, when the window size was made to be $\leq$12 positions, the known NFAT binding motifs appeared further down on the list of best clusters in the output from the evolutionary algorithm. With window sizes as small as 7 positions, NFAT binding sites were no longer observed but other, apparently well-conserved sequences, were discovered. These well-conserved short sequences were considered novel putative TFBSs.

For window sizes 25, 13, 11 and 7, a complete examination of clusters was made such that any output sequences containing known NFAT TFBS motif of GGAAA were identified (Tables 3–6). Any pairs of putative TFBSs were determined as potential TFBS motifs for other transcription binding factors. Four of the known NFAT sites can be discovered with this method, two of which also contain known composite elements (Table 3), and the NFAT motif can be centralized to a smaller window of 13 for seven of the eight input sequences (Table 4, panel a). In some cases, the similarity over all 13 bp in these motifs is striking (e.g. mouse IL-2 relative to human GM-CSF). Four sets of similar sequences were discovered with a

```
Cluster: 1
Number of Sequences: 14
Solution Numbers: 1 19 20 21 22 31
67 74
C00039 (−832,−825)  TTCTGCAC
C00043 (−533,−526)  TTCTGCAA
C00048 (−574,−567)  TTCTGCGA
C00049 (−631,−624)  TTCTGAAA
C00050 (−41,−34)    GTCTGCAA
C00158 (−119,−112)  TTCTGCCA
C00169 (−895,−888)  TTCTGCAA
C00049 (−199,−192)  GTCTGCAC
C00158 (−500,−493)  GTGTGCAA
C00049 (−57,−50)    GTATGCAA
C00049 (−417,−410)  GTATGCAA
C00048 (−547,−540)  GTCTGCGA
C00049 (−432,−425)  TTGTGTAA
C00049 (−84,−77)    TTATGTAA


Cluster: 3
Number of Sequences: 17
Solution Numbers: 4 5 7 10 12 29
30 42 48 50 55
C00039 (−384,−377)  TCCACAGA
C00043 (−813,−806)  TCTACAGA
C00048 (−86,−79)    TCCACAGT
C00049 (−468,−461)  TGCACATA
C00050 (−897,−890)  TCCACAGA
C00158 (−962,−955)  TCCACAGG
C00169 (−605,−598)  TCCACAGA
C00158 (−159,−152)  TCCAGAGA
C00049 (−663,−656)  TCCACAAT
C00049 (−974,−967)  TCCATATA
C00049 (−15,−8)     TCCACACC
C00049 (−196,−189)  TGCACATG
C00049 (−802,−795)  TACACTGG
C00049 (−198,−191)  TCTGCACA
C00049 (−173,−166)  TTCAGACA
C00049 (−134,−127)  TGGACTGA
C00049 (−459,−452)  TGCTGAGA
```

```
Cluster: 2
Number of Sequences: 26
Solution Numbers: 2 3 5 8 9 11 32 33 34
41 43 45 46 47 49 51 52 53 54 95
C00039 (−384,−377)  TCCACAGA
C00043 (−813,−806)  TCTACAGA
C00048 (−86,−79)    TCCACAGT
C00049 (−175,−168)  TCTTCAGA
C00050 (−897,−890)  TCCACAGA
C00158 (−159,−152)  TCCAGAGA
C00169 (−605,−598)  TCCACAGA
C00158 (−962,−955)  TCCACAGG
C00049 (−468,−461)  TGCACATA
C00049 (−663,−656)  TCCACAAT
C00049 (−974,−967)  TCCATATA
C00049 (−15,−8)     TCCACACC
C00158 (−687,−680)  TCCTGAGA
C00049 (−802,−795)  TACACTGG
C00049 (−196,−189)  TGCACATG
C00049 (−198,−191)  TCTGCACA
C00048 (−590,−583)  TCCCGAGA
C00049 (−323,−316)  TCCAGTAA
C00158 (−704,−697)  TCCTTAGA
C00049 (−173,−166)  TTCAGACA
C00049 (−459,−452)  TGCTGAGA
C00049 (−134,−127)  TGGACTGA
C00158 (−150,−143)  TCATCAGA
C00049 (−530,−523)  TTGACAGG
C00049 (−811,−804)  TGCACATT
C00158 (−294,−287)  CCCAAAGA
```

Figure 12. Automated clustering of the Oct-1 solutions showing the known solutions in bold in cluster 4. Each cluster is identified by the number of sequences in the cluster, the solution numbers from which those sequences were derived, COMPEL identifier for the upstream sequences, position upstream of the transcription start site, and putative TFBS motif identified. In this case, the known Oct-1 solutions appear together in cluster 4, in addition to other 8mer windows that share some unspecific commonality, typically with a 'AAA' repeat in the 3′-half of the sequence.

window size of 13 that were non-NFAT motifs (Table 4, panel b). In one case, this represented a complete match of 13 nt between human IL-2 and mouse IL-2 of TTG TCCACCACAA. In another case, 12 of 13 positions were conserved between mouse IL-4 and human IL-4 (CATTGGAAAWTTT).

With a window size of 11 the difficulty associated with discovering the known NFAT solution for all eight sequences increased. The motif GGAAA was discovered only in human TNF and GM-CSF sequences, and mouse IL-4 (Table 5, panel a). Other very similar pairs of motifs of length 11 bp also exist (Table 5, panel b). For window length 7, no NFAT sequences were observed in the top 100 solutions, however other very well-conserved sequences of 7 bp were discovered. Many of these small motifs were conserved at 100% between human and mouse. The motif AATKGCT appears to be highly conserved and repeated in these sequences. The motif CTGAGDV also appears to be highly conserved but not as repeated and is found in all eight sequences known to harbor the NFAT/AP-1 complex.

MatInspector v5.0 (36) was used to determine if any of the conserved motifs were previously experimentally determined (Table 7). For the putative TFBS elements, we then searched the literature to determine any TFs that were known to have some relation to NFAT. Nine of these putative TFs had no known relation to NFAT. These included nuclear matrix protein 4 (NMP4)/Cas-interacting zinc-finger protein (CIZ), PAX2, c-Ets-1, E2F, Basonuclin, Atp1a1 regulatory element binding factor 6 (AREB6), Fork head-related activator-2 (FOXF2), Binding site for S8 type homeodomains

```
Cluster: 4
Number of Sequences: 37
Solution Numbers: 6 13 14 15 16 17
18 23 24 25 26 27 28 37 38 39 40 56
57 60 63 64 65 66 70 71 72 73 84 96
97
C00039 (-215,-208)   ATGCAAAT
C00043 (-320,-313)   ATGTAAAT
C00048 (-904,-897)   ATGTAAAT
C00049 (-55,-48)     ATGCAAAT
C00050 (-517,-510)   ATGCAAAG
C00158 (-806,-799)   ATGCACAT
C00169 (-96,-89)     ATGCAAAT
C00043 (-331,-324)   ATCCAAGT
C00043 (-667,-660)   ATGCTAAC
C00043 (-598,-591)   ATGCACTT
C00048 (-428,-421)   GTGCGAAT
C00048 (-829,-822)   GGGCAAAT
C00048 (-397,-390)   ATTCAAGT
C00043 (-722,-715)   ATACAGAT
C00050 (-113,-106)   ATGCAGCT
C00043 (-314,-307)   ATGCTTAT
C00050 (-788,-781)   GTGCACAT
C00043 (-308,-301)   ATGTAAAC
C00043 (-531,-524)   CTGCAAAA
C00048 (-276,-269)   TTCCAAAT
C00048 (-76,-69)     CTCCAAAT
C00043 (-970,-963)   ATGGGAAT
C00049 (-812,-805)   ATGCACAT
C00158 (-737,-730)   GTGCAACT
C00043 (-991,-984)   ACTCAAAT
C00169 (-413,-406)   ATGCTAAT
C00050 (-695,-688)   GAGCAAAT
C00048 (-712,-705)   TTGTAAAT
C00158 (-882,-875)   ATGCATAC
C00158 (-906,-899)   ATGCATGT
C00158 (-848,-841)   ACGCACAT
C00158 (-812,-805)   ACGCACAT
C00158 (-697,-690)   ATGCAACC
C00050 (-596,-589)   GTGTAAAT
C00050 (-885,-878)   ATGAGAAT
C00158 (-743,-736)   ATGAAAGT
C00158 (-762,-755)   GTGAAAAT
```

Figure 12. Continued.

and Sox-5. It remains unclear if any of these represented motifs share coregulation with NFAT, AP-1 or any other TF in these upstream regions.

However, HMGI(Y) high-mobility-group protein I (Y) is known to be involved in the suppression of IL-4 transcription whereas NFAT-1 is involved in the enhancement of IL-4 promoter activity (37). Myocyte enhancer factor 2 (MEF2) binding sites have recently been found to be co-located with NFAT in the regulation of prion protein gene (PRNP) and its normal product PrP(C) (38). POU factor Brn2 (BRN-2) has been identified as being upstream of KCNN3 in a region of the human genome implicated in schizophrenia by linkage (39). NFAT and AP-1 TFBSs have also been discovered upstream of KCNN3 (39). MEF2 has the greatest number of references in the literature in common with NFAT. NFAT appears to be a key regulator of MEF2 in skeletal muscle, which in turn regulates GLUT4 whole-body insulin action (40).

Table 1. Nucleotide distribution matrix for the matrix V$NFAT_Q6 in TRANSFAC derived from 26 experimentally verified binding sites

Pos.	1	2	3	4	5	6	7	8	9	10	11	12
A	6	13	5	12	2	0	26	25	25	15	9	5
C	11	4	5	1	0	0	0	0	1	5	6	6
G	2	5	8	2	23	26	0	1	0	2	2	6
T	7	4	8	11	1	0	0	0	0	4	9	9
IUPAC	N	N	N	W	G	G	A	A	A	A	N	N
C_i	22.2	23.0	15.5	35.2	73.2	100.0	100.0	89.9	89.9	30.4	21.1	15.4

This is identical to the matrix presented in Ref. (13) for NFAT.

In addition, it appears that both MEF2 and NFAT play key roles in the T-cell receptor-mediated signal transduction pathways leading to IL-2 transcription (41). While calcineurin and NFAT were known to have roles in mediation of the calcium signaling required for IL-2 transcription regulation, MEF2 has only recently been implicated and may serve as a novel target for development of immunosuppressants (41). These and other sources from the literature indicate that our method of TFBS discovery can assist in not only the identification of composite elements but also in the identification of neighboring conserved elements that are experimentally verified as TFBSs.

CONCLUSIONS

A previous approach for TFBS discovery making use of evolutionary computation was extended with a variety of additional features. These refinements include incorporation of complexity normalization, ambiguous nucleotide assignment, bonuses in the scoring function for similarity within different regions of the TF window considered to be the 'core', self-adaptation of the evolutionary parameters associated with the optimization method, a procedure for automated clustering of evolved solutions and parallelization of the entire evolutionary search. In this article, we have evaluated the ability of this revised approach to discover composite TFBS elements using NFAT/AP-1 as an example. The discovery of composite TFBS elements was not possible with the prior approach.

Our results demonstrate that by using appropriate existing parameters such as window size and novel scoring methods such as central bonusing, TFBSs of different sizes and complexity can be identified as top solutions. Identification of NFAT/AP-1 elements was possible when using window sizes of 25-nt positions. However, the probability of identifying composite elements decreased with decreasing window length. While it was still possible to detect the NFAT element with window size 13, by window size 11 this became nearly impossible and by window size 7 other small motifs were discovered that had more similarity than any of the true NFAT/AP-1 composite sites. These results suggest that the choice of window size can directly influence the success of any approach for TFBS discovery including those that are capable of finding composite elements. Such a result is important not only with respect to the current approach but for any TFBS search methods that make use of

Table 2. Nucleotide distribution matrix for the matrix AP-1 from Ref. (13) derived from 47 experimentally verified binding sites with position relative to the transcription start site

Pos.	−6	−5	−4	−3	−2	−1	0	+1	+2	+3	+4	+5	+6
A	6	14	19	4	4	36	3	0	2	47	2	10	15
C	15	12	3	2	2	4	13	0	44	0	8	24	12
G	19	13	16	5	33	2	29	0	0	0	24	5	11
T	7	8	9	36	8	5	2	47	1	0	13	8	9
IUPAC	N	N	R	T	G	A	S	T	C	A	K	C	N
Ii	0.079	0.014	0.116	0.432	0.335	0.432	0.305	1.0	0.799	1.0	0.181	0.126	0.012

Table 3. NFAT motifs discovered when using window size 25

Human IL-2 (−51,−27)	AA**AGGAGGAAAAACT**GTTTCATACA [Table 1, N2][Table 5, N5]
Mouse IL-4 (−953,−929)	AAC**CAAGGGAAAATGA**GTTTACATT [Table 1, N13]
Human IL-4 (−563,−539)	AACCGAGGGAAAATGAGTTTACATT
Human IL-4 (−989,−965)	AACTGAGGAAACTTCCTACAAACCC
Human TNF (−923,−899)	ACC**AGCGGAAAACTT**CCTTGGTGGA [Table 1, N38]
Mouse IL-2 (−471,−447)	ACCATCTTGAAACAGGAAACCAATA
Mouse IL-2 (−300,−276)	CA**AAGAGGAAAATTT**GTTTCATACA [Table 1, N3] [Table 5 N6]
Human TNF (−941,−917)	CGGGGAAAGAATCATTCAACCAGCG
Human TNF (−947,−923)	GGAGGGCGGGGAAAGAATCATTCAA
Human TNF (−818,−794)	GGATTTGGAAAGTTGGGGACACACA

Motifs that are similar to known NFAT TFBS are given in bold and their relationship is given in Table 1 from Ref. (35). In addition, two sequences had exact similarity to NFAT composite elements given in Table 5, Ref. (35). These were NFATp; JunB/Fra-1 (N5) and NFATp; AP-1 (N6), respectively.

Table 4. NFAT motifs discovered with window size 13 (Panel a). Non-NFAT motifs discovered when using a window size of 13 on the NFAT data set that have high pairwise similarity (Panel b)

Panel a	
Mouse IL-2 (−298,−286)	AAGAGGAAAATTT
Human IL-2 (−49,−37)	AGGAGGAAAAACT
Mouse IL-5 (−119,−107)	CACTGGAAACCCT
Human GM-CSF (−966,−954)	CAGAGGAAATGAT
Human TNF (−921,−909)	CAGCGGAAAACTT
Mouse IL-4 (−932,−920)	CATTGGAAAATTT
Human IL-4 (−542,−530)	CATTGGAAATTTT
Human IL-4 (−561,−549)	CCGAGGGAAAATG
Human IL-4 (−987,−975)	CTGAGGAAACTTC
Panel b	
Mouse IL-5 (−716,−704)	TTTTTAAGCACAG
Mouse IL-5 (−203,−191)	TTTTTAAGCAGGG
Human IL-2 (−143,−131)	TTGTCCACCACAA
Mouse IL-2 (−385,−373)	TTGTCCACCACAA
Human IL-4 (−895,−883)	TTCTCTAGCAGCT
Human IL-2 (−302,−290)	TTCTCTAGCTGAC
Mouse IL-4 (−932,−920)	CATTGGAAAATTT
Human IL-4 (−542,−530)	CATTGGAAATTTT

window lengths and suggests that researchers should evaluate a range of window sizes for TFBS discovery. A problem with short TFBS motifs is that other, longer motifs of equal conservation may exist within the upstream regions, providing solutions that score higher than the known shorter 'truth'. These other conserved regions may still be interesting, but only with the inclusion of

Table 5. NFAT motifs discovered with window size 11 (Panel a). Non-NFAT motifs discovered when using a window size of 11 on the NFAT data set that have high pairwise similarity (Panel b)

Panel a	
Human TNF (−941,−931)	CGGGGAAAGAA
Human TNF (−919,−909)	GCGGAAAACTT
Mouse IL-4 (−934,−924)	TACATTGGAAA
Human GM-CSF (−964,−954)	GAGGAAATGAT
Human TNF (−438,−428)	CAGGAAAGGCT
Panel b	
Mouse IL-4 (−360,−350)	CCTTAGACAGA
Mouse IL-2 (−711,−701)	CCTTAGATACA
Human GM-CSF (−157,−147)	TCTCAGGTACA
Human IL-2 (−763,−753)	TCTCTGAGACA
Human IL-2 (−547,−537)	GAGGTAAAGAC
Mouse IL-4 (−23,−13)	GAGGTACTCAT
Mouse IL-5 (−465,−455)	GAGGTATACAT

Table 6. Non-NFAT motifs discovered when using a window size of 7 on the NFAT data set that have high pairwise similarity

Human GM-CSF (−141,−135)	AATGGCT
Human GM-CSF (−692,−686)	AATGGCT
Mouse IL-4 (−153,−147)	AATGGCT
Mouse IL-4 (−340,−334)	AATGGCT
Mouse IL-4 (−311,−305)	AATTACT
Human IL-2 (−17,−11)	AATTGCA
Mouse IL-2 (−37,−31)	AATTGCC
Human IL-4 (−572,−566)	AATTGCT
Mouse IL-5 (−30,−24)	AATTGCT
Human IL-2 (−704,−698)	AATTGTT
Human GM-CSF (−968,−962)	CACAGAG
Mouse IL-5 (−708,−702)	CACAGAT
Human TNF (−80,−74)	CATTGCT
Mouse GM-CSF (−650,−644)	CATTGCT
Mouse GM-CSF (−479,−473)	CTCAGAA
Mouse IL-5 (−45,−39)	CTCAGAG
Human GM-CSF (−23,−17)	CTGACAA
Mouse GM-CSF (−168,−162)	CTGACAA
Mouse IL-2 (−693,−687)	CTGAGAA
Mouse IL-4 (−346,−340)	CTGAGAA
Human IL-2 (−760,−754)	CTGAGAC
Human IL-4 (−857,−851)	CTGAGAG
Human IL-4 (−987,−981)	CTGAGGA
Human TNF (−504,−498)	CTGAGGC
Human TNF (−557,−551)	CTGAGTC
Mouse IL-5 (−108,−102)	CTGAGTT
Mouse GM-CSF (−129,−123)	CTGATAA
Human IL-4 (−29,−23)	GCCTAGT
Mouse GM-CSF (−986,−980)	GCCTGGT
Mouse GM-CSF (−494,−488)	GTGAGAA
Mouse IL-5 (−331,−325)	GTGAGCA
Human TNF (−20,−14)	TGTGGCC
Human GM-CSF (−171,−165)	TGTGGCT

Table 7. Conserved TF binding motifs discovered in proximity to NFAT/AP-1 binding sites

Family/matrix	Description	Strand	Core similarity	Matrix similarity	Sequence
AREB/AREB6.01	AREB6 (Atp1a1 regulatory element binding factor 6)	–	1	0.941	tgtACCTgaga
TBPF/ATATA.01	Avian C-type LTR TATA box	+	1	0.808	ttttTAAGcacag
TBPF/ATATA.01	Avian C-type LTR TATA box	+	1	0.808	ttttTAAGcaggg
TBPF/ATATA.01	Avian C-type LTR TATA box	–	1	0.882	tgtatcTAAGg
BNCF/BNC.01	Basonuclin, cooperates with USF1 in rDNA PolI transcription)	+	1	0.882	tTGTCcacca
HOMF/S8.01	Binding site for S8 type homeodomains	–	1	0.976	agTAATt
BRNF/BRN2.01	Brn-2, POU-III protein class	+	1	0.932	cattggAAATttt
ETSF/ETS1.01	c-Ets-1 binding site	+	1	0.932	cagAGGAaatgat
E2FF/E2F.02	E2F, involved in cell-cycle regulation, interacts with Rb p107 protein	+	0.857	0.884	cagcggAAAActt
E2FF/E2F.01	E2F, involved in cell-cycle regulation, interacts with Rb p107 protein	+	1	0.776	ccgaggGAAAatg
E2FF/E2F.02	E2F, involved in cell-cycle regulation, interacts with Rb p107 protein	+	0.857	0.884	gcggAAAActt
FKHD/FREAC2.01	Fork head-related activator-2 (FOXF2)	+	1	0.853	gaggTAAAgac
GATA/GATA1.03	GATA-binding factor 1	+	1	0.968	ctGATAa
HEAT/HSF1.01	Heat shock factor 1	–	0.910	0.939	NGAAgtttcct
HEAT/HSF1.01	Heat shock factor 1	+	0.916	0.936	GGAAacttc
SORY/HMGIY.01	HMGI(Y) high-mobility-group protein I (Y), architectural TF organizing the framework of a nuclear protein–DNA transcriptional complex	–	1	0.941	aAATTttcctc
SORY/HMGIY.01	HMGI(Y) high-mobility-group protein I (Y), architectural TF organizing the framework of a nuclear protein–DNA transcriptional complex	+	1	0.924	gaaAATTt
SORY/HMGIY.01	HMGI(Y) high-mobility-group protein I (Y), architectural TF organizing the framework of a nuclear protein–DNA transcriptional complex	–	1	0.946	aAATTttccaa
SORY/HMGIY.01	HMGI(Y) high-mobility-group protein I (Y), architectural TF organizing the framework of a nuclear protein–DNA transcriptional complex	+	1	0.932	ggaAATTtt
MEF2/MMEF2.01	Myocyte enhancer factor	–	1	0.902	ctgtgctTAAAaa
MEF2/MMEF2.01	Myocyte enhancer factor	–	1	0.914	ccctgctTAAAaa
MYT1/MYT1.02	MyT1 zinc-finger TF involved in primary neurogenesis	–	1	0.918	AAGTtttccg
MYT1/MYT1.02	MyT1 zinc-finger TF involved in primary neurogenesis	–	1	0.883	gAAGTttcctc
CIZF/NMP4.01	NMP4 (nuclear matrix protein 4)/CIZ (Cas-interacting zinc-finger protein)	+	1	0.992	ggAAAAact
NFAT/NFAT.01	Nuclear factor of activated T-cells	+	1	1	agagGAAAatt
NFAT/NFAT.01	Nuclear factor of activated T-cells	+	1	0.982	ggagGAAAaac
NFAT/NFAT.01	Nuclear factor of activated T-cells	+	1	0.997	attgGAAAatt
NFAT/NFAT.01	Nuclear factor of activated T-cells	+	1	0.971	attgGAAAttt
SORY/SOX5.01	Sox-5	–	1	0.983	aaCAATt
PAX2/PAX2.01	Zebrafish PAX2 paired domain protein	+	1	0.781	cactggAAACcc

Positions in capital letters in the sequence column refer to the 'core' positions as identified in MatInspector. Core similarity (Core sim.) and Matrix similarity (Matrix sim.) are also statistical features generated within MatInspector.

larger sequence windows is there enough 'noise' in the surrounding nucleotides to make the smaller NFAT element score similar to the larger putative TFBS motifs.

Given the knowledge that the appropriate window size is very likely unknown for each upstream region, we propose a dual approach for TFBS discovery. The first approach is to use a top-down approach with large window sizes to find putative composite elements in upstream regions and minimize the window size over time. The second approach is to use a bottom-up approach with small window sizes to find putative single TFBS motifs and then increase the window size over time, continually reseeding with the best results from the previous window size and iterating the evolutionary process. Any commonality between these two methods lends additional credence to the discovered motifs.

The COMPEL database (42) contains examples of experimentally verified composite TFBS motifs. In the future, we plan on reviewing this database to determine if there are any changes to the evolutionary algorithm strategy that can assist when specifically looking for composite elements. For instance, statistics on known motif–motif distances could be added to the fitness function. Distances between putative motifs that fall within the distribution of known samples could be given a bonus, further driving the evolutionary optimization toward reasonable solutions.

The approach to automated clustering drastically reduced the time required to comb through the output of the evolutionary algorithm and identify key similar motifs. However, it was also clear that the resulting clusters were in many cases still redundant. We made use of an

initial threshold of 90% similarity over all sequences in each solution to all sequences in a cluster before that solution could be added to a cluster. The cutoff of 90% may be too stringent in some cases. Thus, we envision a subsequent version that iterates the clustering, first with a highly stringent threshold of 90%, followed by repeated rounds of merging clusters with ever lower thresholds to allow for similar cluster contents to be merged successfully, minimizing the number of clusters visible to the user and aiding in results interpretation.

An additional bonus could be incorporated that makes use of the database of known motifs from TRANSFAC® and composite elements from COMPEL to give bonuses for any window that contains sequences with strong similarity to known TFBSs. For instance, if the goal of a research project is to discover completely novel elements, then any previously experimentally determined TFBS motifs (either represented as specific sequences or statistical matrices) can be given a strong penalty. The resulting solutions will, in theory, contain a greater percentage of new motifs if those motifs exist upstream. If, however, the goal of the research project is to discover previously known elements, then large bonuses can be given to experimentally determined TFBS, driving the algorithm away from discovery of novel TFBS motifs, but potentially resulting in the discovery of novel composite elements or new TFBS motifs that neighbor known TFBS motifs. For any of these approaches to make use of the information in TRANSFAC® or COMPEL, there is a clear benefit to using the nucleotide distribution matrices that have been derived from as many phylogenetically diverse sequences as possible.

FUNDING

Funding for open access charge: Eli Lilly and Co.

Conflict of interest statement. None declared.

REFERENCES

1. Brazma,A. and Vilo,J. (2000) Gene expression data analysis. *FEBS Lett*., **480**, 17–24.
2. Bucher,P. (1999) Regulatory elements and expression profiles. *Curr. Opin. Struct. Biol*., **9**, 400–407.
3. Zhang,M.Q. (1999) Large-scale gene expression data analysis: a new challenge to computational biologist. *Genome Res*., **9**, 681–688.
4. Ohler,U. and Niemann,H. (2001) Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet*., **17**, 56–60.
5. Liu,X., Brutlag,D.L. and Liu,J.S. (2001) Bioprospector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput*., **6**, 127–138.
6. Jensen,L.J. and Knudsen,S. (2000) Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics*, **16**, 326–333.
7. Congdon,C.B., Aman,J., Nava,G., Gaskins,H.R. and Mattingly,C. (2008) 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. Bioinf*., **5**, 1–14.
8. Lones,M.A. and Tyrell,A.M. (2007) Regulatory motif discovery using a population clustering evolutionary algorithm. *IEEE/ACM Trans. Comput. Biol. Bioinf*., **4**, 403–414.
9. Mauri,G., Mosca,R. and Pavesi,G. (2004) A GA approach to the definition of regulatory signals in genomic sequences. *Proc. GECCO 2004*. LNCS 3102, Springer, Berlin, 380–391.
10. Carlos,R.-R. (2006) Finding DNA motifs using genetic algorithms. In: *Proceedings of the Fifth Mexican International Conference on Artificial Intelligence*. IEEE, Piscataway, New Jersey.
11. Paul,T.K. and Iba,H. (2006) Identification of weak motifs in multiple biological sequences using genetic algorithm. In: *Proceedings of GECCO 2006*. New York, ACM, pp. 271–278.
12. Wei,Z. and Jensen,S.T. (2006) GAME: detecting cis-regulatory elements using a genetic algorithm. *Bioinformatics*, **22**, 1577–1584.
13. Chan,T.-M., Leung,K.-S. and Lee,K.-H. (2008) TFBS identification based on genetic algorithm with combined representations and adaptive post-processing. *Bioinformatics*, **24**, 341–349.
14. Kato,M. and Tsunoda,T. (2007) MotifCombinator: a web-based tool to search for combinations of cis-regulatory motifs. *BMC Bioinformatics*, **8**, 100.
15. Thijs,G., Lescot,M., Marchal,K., Rombauts,S., De Moor,B., Rouzé,P. and Moreau,Y. (2001) A higher-order background model improves detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, **17**, 1113–1122.
16. Atteson,K. (1998) Calculating the exact probability of language-like patterns in biomolecular sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **6**, 17–24.
17. Tompa,M. (1999) An exact method for finding short motifs in sequences with application to the ribosome binding site problem. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **7**, 262–271.
18. Brazma,A., Jonassen,I., Vilo,J. and Ukkonen,E. (1998) Predicting gene regulatory elements in silico on a genomic scale. *Genome Res*., **8**, 1202–1215.
19. Vanet,A., Marsan,L., Labigne,A. and Sagot,M.F. (2000) Inferring regulatory elements from a whole genome. An analysis of *Helicobacter pylori* sigma(80) family of promoter signals. *J. Mol. Biol*., **297**, 335–353.
20. Kiełbasa,S.M., Korbel,J.O., Beule,D., Schuchhardt,J. and Herzel,H. (2001) Combining frequency and positional information to predict transcription factor binding sites. *Bioinformatics*., **17**, 1019–1026.
21. Lawrence,C.E. and Reilly,A.A. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, **7**, 41–51.
22. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
23. Bailey,T. and Elkan,C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach. Learn*., **21**, 51–80.
24. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol*., **296**, 1205–1214.
25. Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci*., **4**, 1618–1632.
26. Roth,F.P., Hughes,J.D., Estep,P.W. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequence clustered by whole genome mRNA quantitation. *Nat. Biotechol*., **16**, 939–945.
27. Tharakaraman,K., Mariño-Ramírez,L., Sheetlin,S., Landsman,D. and Spouge,J.L. (2005) Alignments anchored on genomic landmarks can aid in the identification of regulatory elements. *Bioinformatics*, **21(Suppl. 1)**, i440–i448.
28. Fauteux,F., Blanchette,M. and Strömvik,M.V. (2008) Seeder: discriminative seeding DNA motif discovery. *Bioinformatics*, **24**, 2303–2307.
29. Fogel,G.B., Weekes,D.G., Varga,G., Dow,E.R., Harlow,H.B., Onyia,J.E. and Su.,C. (2004) Discovery of sequence motifs related to coexpression of genes using evolutionary computation. *Nucleic Acids Res*., **32**, 3826–3835.
30. Wootten,J.C. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol*., **266**, 554–571.

31. Fogel,G.B., Weekes,D.G., Varga,G., Dow,E.R., Craven,A.M., Harlow,H.B., Su,E.W., Onyia,J.E. and Su,C. (2005) A statistical analysis of the TRANSFAC database. *BioSystems*, **81**, 137–154.
32. Fogel,D.B. (2006) *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*, 3rd edn. IEEE Press, Piscataway, NJ.
33. Shaw,J.P., Utz,P.J., Durand,D.B., Toole,J.J., Emmel,E.A. and Crabtree,G.R. (1988) Identification of a putative regulator of early T cell activation genes. *Science*, **241**, 202–205.
34. Matys,V., Fricke,E., Geffers,R., Gösling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC®: transcriptional regulation, from pattern to profiles. *Nucleic Acids Res.*, **31**, 374–378.
35. Kel,A., Kel-Margoulis,O., Babenko,V. and Wingender,E. (1999) Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. *J. Mol. Biol.*, **288**, 353–376.
36. Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector – new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.
37. Chuvpilo,S., Schomberg,C., Gerwig,R., Heinfling,A., Reeves,R., Grummt,F. and Serfling,E. (1993) Multiple closely-linked NFAT/octamer and HMG I(Y) binding sites are part of the interleukin-4 promoter. *Nucleic Acids Res.*, **21**, 5694–5704.
38. Premzl,M., Delbridge,M., Gready,J.E., Wilson,P., Johnson,M., Davis,J., Kuczek,E. and Marshall Graves,J.A. (2005) The prion protein gene: identifying regulatory signals using marsupial sequence. *Gene*, **349**, 121–134.
39. Sun,G., Tomita,H., Shakkottai,V.G. and Gargus,J.J. (2001) Genomic organization and promoter analysis of human KCNN3 gene. *J. Hum. Genet.*, **46**, 463–470.
40. McGee,S.L. and Hargreaves,M. (2004) Exercise and myocyte enhancer factor 2 regulation in human skeletal muscle. *Diabetes*, **53**, 1208–1214.
41. Pan,F., Ye,Z., Cheng,L. and Liu,J.O. (2004) Myocyte enhancer factor 2 mediates calcium-dependent transcription of the interleukin-2 gene in T lymphocytes: a calcium signaling module that is distinct from but collaborates with the nuclear factor of activated T cells (NFAT). *J. Biol. Chem.*, **279**, 14477–14480.
42. Kel-Margoulis,O.V., Romashchenko,A.G., Kolchanov,N.A., Wingender,E. and Kel,A.E. (2000) COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation. *Nucleic Acids Res.*, **28**, 311–315.