Assessing computational tools for the discovery of transcription factor binding sites

Martin Tompa[1,2], Nan Li[1], Timothy L Bailey[3], George M Church[4], Bart De Moor[5], Eleazar Eskin[6], Alexander V Favorov[7,8], Martin C Frith[9], Yutao Fu[9], W James Kent[10], Vsevolod J Makeev[7,8], Andrei A Mironov[7,11], William Stafford Noble[1,2], Giulio Pavesi[12], Graziano Pesole[13], Mireille Régnier[14], Nicolas Simonis[15], Saurabh Sinha[16], Gert Thijs[5], Jacques van Helden[15], Mathias Vandenbogaert[14], Zhiping Weng[9], Christopher Workman[17], Chun Ye[18] & Zhou Zhu[4]

The prediction of regulatory elements is a problem where computational methods offer great hope. Over the past few years, numerous tools have become available for this task. The purpose of the current assessment is twofold: to provide some guidance to users regarding the accuracy of currently available tools in various settings, and to provide a benchmark of data sets for assessing future tools.

A major challenge in molecular biology is to understand the mechanisms that regulate the expression of genes. An important step in this challenge is the ability to identify regulatory elements, notably the binding sites in DNA for transcription factors. Transcription factors are proteins that bind to DNA, typically upstream from and close to the transcription start site of a gene, and regulate the expression of that gene by activating or inhibiting the transcription machinery. The prediction of such regulatory elements is a problem where computational methods offer great hope, and indeed computational biologists have invested considerable effort into solving this problem.

Because little is known about most transcription factors and their target binding sites, even in well studied organisms, we focus here on those computational tools designed for the discovery of novel regulatory elements, where nothing is assumed a priori of the transcription factor or its preferred binding sites. Usually, a user provides a collection of regulatory regions of genes that are believed to be coregulated, and the computational tool identifies short DNA sequence 'motifs' that are statistically overrepresented in these regulatory regions. Accurate identification of these motifs is difficult because they are short signals (typically about 10 bp long) in the midst of a great amount of statistical noise (a typical input being one regulatory region of length 1,000 bp upstream of each gene). To make matters worse, there is sequence variability among the binding sites of a given transcription factor, and the nature of the variability itself is not well understood.

Over the past few years, numerous tools have become available for this task of motif prediction, differing from each other chiefly in their definition of what constitutes a motif, what constitutes statistical overrepresentation of a motif and the method used to find statistically overrepresented motifs. However, the biologist has been offered little guidance in the choice among these tools. The purpose of the current assessment is twofold: to provide some guidance regarding the accuracy of currently available computational tools in various settings, and to provide a benchmark of data sets for assessing future tools.

There have been only a few small-scale assessments of some of these motif discovery tools, for instance, those of Pevzner & Sze[1] and Sinha & Tompa[2]. The current assessment is modeled on earlier large-scale

[1]Department of Computer Science and Engineering, Box 352350, University of Washington, Seattle, Washington 98195-2350, USA. [2]Department of Genome Sciences, Box 357730, University of Washington, Seattle, Washington 98195-7730, USA. [3]Institute for Molecular Biosciences, University of Queensland, Brisbane, Australia. [4]Department of Genetics and Lipper Center for Computational Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [5]ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium. [6]Department of Computer Science and Engineering, University of California, San Diego, California 92093, USA. [7]State Scientific Centre 'GosNIIGenetica,' 1st Dorozhny pr. 1, Moscow, 117545, Russia. [8]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Vavilova 32, Moscow 119991, Russia. [9]Bioinformatics Program, Boston University, Boston, Massachusetts 02215, USA. [10]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064, USA. [11]Department of Bioengineering and Bioinformatics, Moscow State University, Lab. Bldg B, Vorobiovy Gory 1-33, Moscow 119992, Russia. [12]Department of Computer Science and Communication (D.I.Co), University of Milan, Milan, Italy. [13]Department of Biomolecular Science and Biotechnology, University of Milan, Milan, Italy. [14]INRIA Rocquencourt, Domaine de Voluceau B.P. 105, 78153 Le Chesnay, France. [15]SCMB-Université Libre de Bruxelles, Campus Plaine, CP 263, Boulevard du Triomphe, 1050 Bruxelles, Belgium. [16]Center for Studies in Physics and Biology, The Rockefeller University, New York, New York 10021, USA. [17]Department of Bioengineering, University of California, San Diego, California 92093, USA. [18]Bioinformatics Program, University of California, San Diego, California 92093, USA. Correspondence should be addressed to M.T. (tompa@cs.washington.edu).

Published online 6 January 2005; doi:10.1038/nbt1053

assessments for a different computational problem, namely the prediction of genes themselves, reported by Burset & Guigó[3], by Burge & Karlin[4], and by Reese *et al*[5,6]. In this study, we assess 13 motif-discovery tools, all available on the internet, that do not use auxiliary information, such as comparative sequence analysis, mRNA expression levels or chromatin immunoprecipitation results.

In brief, we created data sets containing known binding sites to test these tools. Without revealing the known binding sites, each author with specific expertise on a particular tool then ran that tool on these data sets. Experts were chosen to test each tool so that none would be put at the disadvantage of being run with an uninformed setting of its parameters. The expert predictions were then compared with known binding sites, using various statistics to assess the correctness of the predictions.

The study reported here is a first attempt and therefore by no means perfect. We introduce an experimental design and statistical analyses that solve some of the problems in comparing tools, but we also believe that future assessments will benefit from our mistakes, enabling improved comparisons. We conclude by offering suggestions for such future improvements.

METHODS

The tools compared in this assessment are AlignACE[7], ANN-Spec[8], Consensus[9], GLAM[10], Improbizer[11], MEME[12], MITRA[13], MotifSampler[14], oligo/dyad-analysis[15,16], QuickScore[17], SeSiMCMC[18], Weeder[19] and YMF[20]. Short descriptions of them are provided in **Table 1**.

Creating good data sets posed some immediate challenges. At one extreme, we could use real genomic promoter sequences containing real annotated transcription factor binding sites. The drawback of this approach is that no one knows the complete 'correct' answer: there could be unannotated binding sites, and programs that correctly predict these would necessarily be penalized. At the other extreme, we could assure that we know the complete correct answer by using artificially constructed sequences. For instance, we could generate random sequences using a Markov chain, and implant at random positions instances of a randomly chosen position-specific scoring matrix. The drawback of this approach is that no one knows the 'correct' stochastic process that nature uses, and so we would be introducing biases that favor certain tools over others.

Table 1 Details about the operation principles, basic technical data and URLs of 13 analyzed tools

Program	Operating principle	Technical data	URL	Reference
AlignACE	Gibbs sampling algorithm that returns a series of motifs as weight matrices that are over-represented in the input set	Judges alignments sampled during the course of the algorithm using a maximum *a priori* log likelihood score, which gauges the degree of overrepresentation. Provides an adjunct measure (group specificity score) that takes into account the sequence of the entire genome and highlights those motifs found preferentially in association with the genes under consideration.	http://atlas.med.harvard.edu/	7
ANN-Spec	Models the DNA-binding specificity of a transcription factor using a weight matrix	Objective function based on log likelihood that transcription factor binds at least once in each sequence of the positive training data compared with the number of times it is estimated to bind in the background training data. Parameter fitting is accomplished with a gradient descent method, which includes Gibbs sampling of the positive training examples.	http://www.cbs.dtu.dk/~workman/ann-spec/	8
Consensus	Models motifs using weight matrices, searching for the matrix with maximum information content	Uses a greedy method, first finding the pair of sequences that share the motif with greatest information content, then finding the third sequence that can be added to the motif resulting in greatest information content, and so on.	http://bifrost.wustl.edu/consensus/	9
GLAM	Gibbs sampling-based algorithm that automatically optimizes the alignment width and evaluates the statistical significance of its output	Since the basic algorithm cannot find multiple motif instances per sequence, long sequences were fragmented into shorter ones, and the alignment was transformed into a weight matrix and used to scan the sequences to obtain the final site predictions.	http://zlab.bu.edu/glam/	10
The Improbizer	Uses expectation maximization to determine weight matrices of DNA motifs that occur improbably often in the input sequences	As a background (null) model it uses up to a second-order Markov model of background sequence. Optionally, Improbizer constructs a Gaussian model of motif placement, so that motifs that occur in similar positions in the input sequences are more likely to be found.	http://www.soe.ucsc.edu/~kent/improbizer	11
MEME	Optimizes the E-value of a statistic related to the information content of the motif	Rather than sum of information content of each motif column, statistic used is the product of the *P* values of column information contents. The motif search consists of performing expectation maximization from starting points derived from each subsequence occurring in the input sequences. MEME differs from MEME3 mainly in using a correction factor to improve the accuracy of the objective function.	http://meme.sdsc.edu/	12
MITRA	Uses an efficient data structure to traverse the space of IUPAC patterns.	For each pattern, MITRA computes the hypergeometric score of the occurrences in the target sequences relative to the background sequences and reports the highest scoring patterns.	http://www.calit2.net/compbio/mitra/	13
MotifSampler	Matrix-based, motif-finding algorithm that extends Gibbs sampling by modeling the background with a higher order Markov model	The probabilistic framework is further exploited to estimate the expected number of motif instances in the sequence.	http://www.esat.kuleuven.ac.be/~dna/BioI/Software.html	14

Table 1 continued on following page

For the binding sites, we decided to use the TRANSFAC database[21] (http://www.gene-regulation.com/pub/databases.html#transfac) to choose real transcription factors, their known binding sites, and the positions and orientations of those binding sites. (Because TRANSFAC contains only eukaryotic transcription factors, we restricted ourselves to eukaryotic data sets, though it would be beneficial to do a similar assessment on prokaryotic data sets.) Each such transcription factor gives rise to one data set of sequences. Each such data set consists of one of three different types of background sequence, with the transcription factor's known binding sites planted at their known positions and orientations. The three types are (i) the binding sites' real promoter sequences (called 'real' in the sequel) (ii) randomly chosen promoter sequences from the same genome (called 'generic') and (iii) sequences generated by a Markov chain of order 3 (called 'markov'). Using some of each type, we intended to avoid systematic effects of the drawbacks described above. No attempt was made to eliminate sequences that might contain additional transcription factor binding sites, since our ability to identify such sites accurately is limited.

The process for selecting transcription factors and binding sites from TRANSFAC was as follows. We selected only transcription factors for which TRANSFAC also lists a binding site consensus sequence. For each factor, we removed duplicate instances of the same binding site, removed binding sites missing sequence or position information, removed binding sites whose position was annotated with respect to anything other than transcription or translation start site, removed binding sites whose position was less than –3,000 bp or greater than 0, and removed sequences with two reported binding sites contradicting each other in sequence and position. Any factor with fewer than five remaining binding sites in a single species was then discarded. (Some data sets lost additional binding sites in subsequent consistency tests against genomic sequence data.) Only fly, human, mouse, rat and yeast had at least four remaining data sets; we discarded the rat sequences, as the rat genome was not yet completely sequenced and these data sets might be too close to mouse.

This resulted in 52 data sets. Six of the data sets are from fly, 26 from human, 12 from mouse and 8 from yeast. As negative controls, we added 4 additional data sets of type markov containing no planted binding

Table 1 Continued

Program	Operating principle	Technical data	URL	Reference
Oligo/dyad-analysis	Detects overrepresented oligonucleotides with oligo-analysis[15] and spaced motifs with dyad-analysis[16]	These algorithms detect statistically significant motifs by counting the number of occurrences of each word or dyad and comparing these with expectation. Most crucial parameter is choice of appropriate probabilistic model for the estimation of occurrence significance. In this study, a negative binomial distribution on word distributions was obtained from 1,000 random promoter selections of the same size as the test sets	http://rsat.scmbb.ulb.ac.be/rsat/	15,16
QuickScore	Based on an exhaustive searching algorithm that estimates probabilities of rare or frequent words in genomic texts	Incorporates an extended consensus method allowing well-defined mismatches and uses mathematical expressions for efficiently computing z-scores and *P* values, depending on the statistical models used in their range of applicability. Special attention is paid to the drawbacks of numerical instability. The background model is Markovian, with order up to 3.	http://algo.inria.fr/dolley/QuickScore/	17
SeSiMCMC	Modification of Gibbs sampler algorithm that models the motif as a weight matrix, optionally with the symmetry of a palindrome or of a direct repeat, and optionally with spacers	Includes two alternating stages. The first one optimizes the weight matrix for a given motif and spacer length. The algorithm changes the positions of the motif occurrences in the sequences and infers the motif model from the current occurrences. These changes are used to optimize the likelihood of sequences as being segmented into the (Bernoulli) background and the motif occurrences. The optimization is organized via a Gibbs-like Markov chain, which samples positions in sequences one by one, until the Markov chain converges. The second stage looks for best motif and spacer lengths for obtained motif positions. It optimizes the common information content of motif and of distributions of motif occurrence positions.	http://favorov.hole.ru/gibbslfm/	18
Weeder	Consensus-based method that enumerates exhaustively all the oligos up to a maximum length and collects their occurrences (with substitutions) from input sequences	Each motif evaluated according to number of sequences in which it appears and how well conserved it is in each sequence, with respect to expected values derived from the oligo frequency analysis of all the available upstream sequences of the same organism. Different combinations of 'canonical' motif parameters derived from the analysis of known instances of yeast transcription factor binding sites (length ranging from 6 to 12, number of substitutions from 1 to 4) are automatically tried by the algorithm in different runs. It also analyzes and compares the top-scoring motifs of each run with a simple clustering method to detect which ones could be more likely to correspond to transcription factor binding sites. Best instances of each motif are selected from sequences using a weight matrix built with sites found by consensus-based algorithm.	http://159.149.109.16/Tool/ind.php	19
YMF	Uses an exhaustive search algorithm to find motifs with the greatest *z*-scores	A *P* value for the z-score is used to assess significance of motif. Motifs themselves are short sequences over the IUPAC alphabet, with spacers ('N's) constrained to occur in the middle of the sequence.	http://bio.cs.washington.edu/software.html#ymf	20

Box 1 Study design

In a real application, a biologist would select one of these tools and perhaps pursue a number of the top motifs reported. In this study, however, we allowed only one 'best' motif per data set. The explanation for this decision is tied to the fact that no single statistic is perfect for measuring the correctness of predictions. If we allowed multiple predictions, on what basis would we compare the performance of two tools on a given data set? Even with only one predicted motif per tool, there are still many appropriate statistics with which they may be compared.

We restricted our assessment to tools that do not make use of auxiliary data, such as comparative sequence analysis (also known as phylogenetic footprinting). Although we believe such comparative analysis to be important and effective, the fact is that many tools (including all of the participating tools in this assessment) are not designed to exploit it, and users of these tools need to know how accurate they are. If these tools are not sufficiently accurate, then a detailed analysis of the data sets on which they fail will point out problems in our current approaches and hopefully the path to improving them. An assessment of tools designed for phylogenetic footprinting would be equally important, but must necessarily be the subject of a separate study.

As outlined in **Box 2**, there are numerous statistics available to measure a tool's correctness on a data set, various ways of summarizing those statistics over a collection of data sets and numerous interesting collections of data sets over which to summarize.

sites, and added 2 of them to the fly collection and 2 of them to the yeast collection. For each species, about one-third of its data sets are of each vof the types real, generic and markov. To 31 of the 38 data sets of type generic or markov, we added 1 to 4 additional sequences with no planted binding sites, so that each input sequence contains 0 or more planted binding sites. The number of sequences per data set varies from 1 to 35 with mean 7, and the individual sequence length per data set varies from 500 bp to 3,000 bp. The total size of each data set varies from 1 to 70 kb with mean 8 kb. The number of planted binding sites per data set varies from 0 to 76 with mean 9.

The data sets are available as a benchmark at the assessment web site http://bio.cs.washington.edu/assessment/ (see also **Supplementary Data** online). In fact, each of the 52 data sets (excluding the 4 negative controls) is available there in each of the data set types real, generic, and markov, although the assessment described here used only one of those three types for each data set.

Each of the 56 data sets was supplied for testing as a FASTA file with an indication of their species of origin, but with no indication of the type (real, generic or markov) or any other information. For each data set, the prediction tools were required either to select the single best motif and report the positions and sequences of that motif's occurrences or to report that the data set contains no significant motif. It was permissible to vary parameter settings from data set to data set, mask repeats in the input sequences, postprocess the output to eliminate low-complexity motifs and generally perform any pre- and postprocessing deemed appropriate. Neither consultation of TRANSFAC nor the employment of methods that would not be available in a real application of novel motif discovery were permitted. Data sets were supplied in November 2003 and prediction results returned in February 2004 (see **Box 1**).

RESULTS

Figure 1a shows the results of all seven statistics (see **Box 2**)—nucleotide-level sensitivity (*nSn*), nucleotide-level positive predictive value (*nPPV*), nucleotide-level performance coefficient (*nPC*), nucleotide-level correlation coefficient (*nCC*), site-level sensitivity (*sSn*), site-level positive predictive value (*sPPV*) and site-level average site performance (*sASP*)—summarized over all 56 data sets (regardless of species, data set type). The values *nPPV* and *sPPV* should be noted with some caution. As described in **Box 2**, these statistics are undefined on each data set for which a program predicts no motif. As a result, these *PPV* values will be exaggerated for those programs that make no predictions on hard data sets. **Table 2** summarizes the number of data sets on which each tool made no prediction.

Figure 1b breaks down the data sets according to species (regardless of data set type), using the correlation coefficient *nCC* as a proxy for correctness. **Figure 1c** breaks down the data sets according to type real, generic or markov (regardless of species). This figure suggests greater difficulty with the real type data sets, likely for the reasons described in the Methods section above. Because of this, **Figure 1d** recapitulates the seven statistics of **Figure 1a** over just the data sets of types generic and markov. **Figure 1e** shows *nSp* for the four negative control data sets containing no planted motif.

Finally, **Table 3** shows the improvement possible in the correlation coefficient *nCC* when a pair of tools' predictions are used rather than a single tool, summarized over all 56 data sets (regardless of species, data set type). The purpose of this table is not to simulate what a biologist might do with two tools, but rather to demonstrate that tools may complement each other: on some data sets the first tool will have better predictions and on others the second will have better predictions. The primary tool T is listed in the row header and the secondary tool T' in the column header. For each individual data set D, we choose the nucleotide level scores *nTP*, *nFN*, *nFP* and *nTN* of whichever of T or T' has the greater *nCC* score on D, if T predicts some motif on D, otherwise we choose the nucleotide level scores of T on D. (Note the asymmetry when primary tool T predicts no motif on D.) We then add these chosen nucleotide level scores over all 56 data sets, as described for the 'Combined' method of summarizing in **Box 2**, and compute the correlation coefficient *nCC* for the combined scores. This is the value shown in **Table 3** in row T and column T'. If $T = T'$, then the value is equal to the individual *nCC* score from **Figure 1a**.

DISCUSSION

We have described an assessment of 13 different computational tools for *de novo* prediction of regulatory elements. Data in **Figure 1a–d** reveal that the absolute measures of correctness of these programs are low: site sensitivity *sSn* is at most 0.22 and correlation coefficient *nCC* is at most 0.20 in **Figure 1d**, for example. This should not be taken as an indictment of computational methods for prediction of regulatory elements, for a very great number of reasons:

1. Most importantly, the underlying biology of regulatory mechanisms is very incompletely understood. We lack an absolute standard against which to measure the correctness of tools (unlike the crystal structures

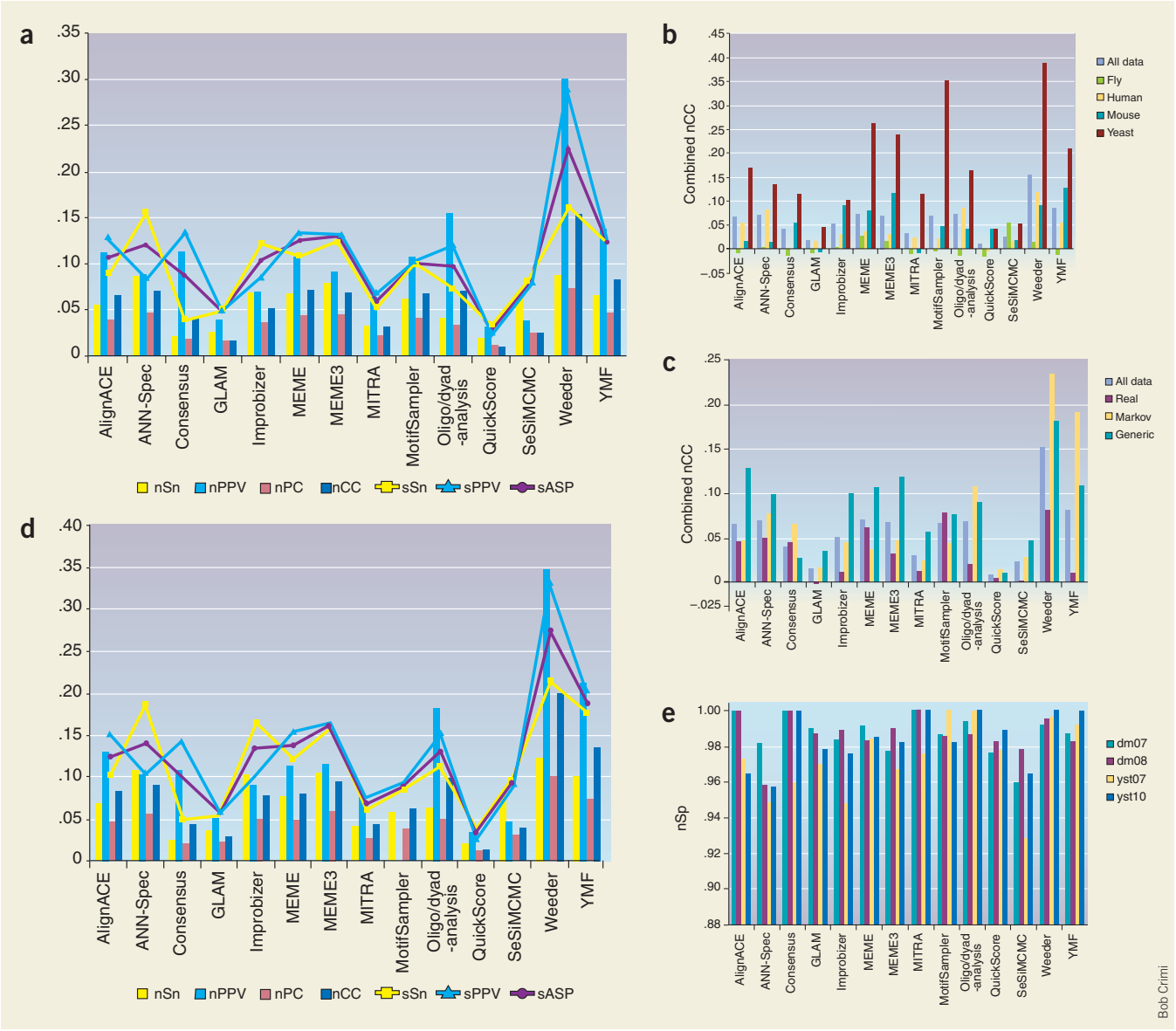

Figure 1 Representative statistics comparing the accuracy of the 13 tools assessed in this analysis. (**a**) Combined measures of correctness over all 56 data sets. See **Tables 1** and **2** for details on the individual tools, Methods section for an explanation of data set types and **Box 2** for definitions of all statistics. (**b**) Correlation coefficient (*nCC*) by species. (**c**) Correlation coefficient (*nCC*) by data set type. (**d**) Combined measures of correctness over generic and Markov data sets. (**e**) Specificity (*nSp*) on four negative control data sets.

used in the Critical Assessment of Techniques for Protein Structure Prediction[22]). For these reasons, our benchmark of data sets is likely a poor approximation for the biological truth.

2. Each participant was required to predict a single motif per data set (or none), this choice necessarily being subjective and sometimes arbitrary. In practice, one might instead pursue the top several motifs predicted by any given tool. This has a dramatic effect on sensitivity.
3. The assessment allowed no comparative sequence analysis among species, a powerful method for the prediction of regulatory elements.
4. The assessment allowed no exploitation, except possibly in the data sets of type real, of the fact that the binding sites of multiple transcription factors often occur in close proximity to each other.
5. The assessment depends on TRANSFAC[21] as its standard for the true binding sites; any such database is fallible and biased.
6. Many of the binding sites cataloged in TRANSFAC[21] are unusually long: 35 of the binding sites used in this assessment were each 31–71 bp in length. This may reflect lack of precision in the experimental method used, with the true binding site actually a shorter subsequence of the cataloged site. Such long cataloged sites have a detrimental effect on measured sensitivity, both at the nucleotide and site levels.
7. The assessment allows only one known motif for each data set, despite the fact that the 18 data sets of type real are likely to have binding sites for multiple transcription factors.

In addition, in comparing the performance of the tools, one must keep in mind the fact that each predicted set of motif instances was subject to human choices of parameters and pre- and postprocessing, and that the amount of time and effort invested by the participants

Box 2 Statistics used to assess tool performance quality

For each tool *T* and each data set *D*, we now have the set of known binding sites and the set of predicted binding sites. The correctness of *T* on *D* can be assessed both at the nucleotide level and at the site level. Specifically, at the nucleotide level define true positives (nTP), false negatives (nFN) and others as follows:

- *nTP* is the number of nucleotide positions in both known sites and predicted sites,
- *nFN* is the number of nucleotide positions in known sites but not in predicted sites,
- *nFP* is the number of nucleotide positions not in known sites but in predicted sites, and
- *nTN* is the number of nucleotide positions in neither known sites nor predicted sites.

We will say that a predicted site overlaps a known site if they overlap by at least one-quarter the length of the known site. (Although this cutoff is somewhat arbitrary, the motivation is that, if an experimentalist were to remove the predicted site, enough of the known site would be deleted so that one might be able to see a difference in expression.) At the site level, then, let:

- *sTP* be the number of known sites overlapped by predicted sites,
- *sFN* be the number of known sites not overlapped by predicted sites, and
- *sFP* be the number of predicted sites not overlapped by known sites.

At either the nucleotide ($x = n$) or site ($x = s$) level, one can then define:

- *Sensitivity*: $xSn = xTP/(xTP + xFN)$, and
- *Positive Predictive Value*: $xPPV = xTP/(xTP + xFP)$.

The sensitivity gives the fraction of known sites (or site nucleotides) that are predicted, and the positive predictive value gives the fraction of predicted sites (or site nucleotides) that are known.

At the nucleotide level one can also define:

Specificity: $nSP = nTN/(nTN + nFP)$.

Finally, it is enlightening to consider various single statistics that in some sense average (some of) these quantities. Following Pevzner & Sze[1], define the (nucleotide level) performance coefficient as:

- $nPC = nTP/(nTP + nFN + nFP)$.

Following Burset & Guigó[3], define the (nucleotide level) correlation coefficient as:

$$nCC = \frac{nTP \cdot nTN - nFN \cdot nFP}{\sqrt{(nTP + nFN)(nTN + nFP)(nTP + nFP)(nTN + nFN)}}$$

and the (site level) average site performance as:

- $sASP = (sSn + sPPV)/2$.

The correlation coefficient *nCC* is the Pearson product-moment coefficient of correlation in the particular case of two binary variables, also called the 'phi coefficient of correlation.' The two binary variables are the characteristic vectors of the known nucleotide positions and of the predicted nucleotide positions, so that this statistic measures the correlation between those two sets of positions. The value of *nCC* ranges from –1 (indicating perfect anticorrelation) to +1 (indicating perfect correlation). Thus, if the predicted motifs exactly coincide with the known binding sites, *nCC* will be +1. If each nucleotide position were predicted to be in the motif randomly and independently, then the expected value of *nCC* would be 0, indicating no correlation.

No single statistic captures correctness perfectly. For those who want to compute other statistics, the seven 'raw scores' *nTP*, *nFN*, *nFP*, *nTN*, *sTP*, *sFN* and *sFP* are tabulated at the assessment web site (http://bio.cs.washington.edu/assessment/) for each data set and each tool.

For the four negative control data sets that have no planted binding sites, $TP + FN = 0$, so *nSn*, *nCC*, *sSn*, and *sASP* are undefined, and *nPPV*, *nPC* and *sPPV* are uninformative. We will simply inspect the specificity *nSp* on these four data sets separately. A far greater analysis problem arises in those cases in which some tool predicts no motif in a data set. In these cases, $TP + FP = 0$, so *nPPV*, *nCC*, *sPPV* and *sASP* are undefined, and *nSn*, *nPC* and *sSn* are uninformative. This hampers any straightforward attempt to compare the tools across this data set.

In any case, we need a way of summarizing the performance of a given tool over a collection of data sets, where that collection might be all the data sets, or all the yeast data sets, or all the generic data sets. For each tool *T*, each statistic *M*, and each collection *C* of data sets of interest, we summarize *T*'s performance on *C* by each of the methods below. (If statistic *M* is undefined for tool *T* on a particular data set in *C*, then omit that data set when summarizing for *T*, except in the Combined method where this omission is unnecessary.)

1. **Average.** The usual arithmetic mean of the *M* scores.
2. **Normalized.** For each data set, normalize the *M* scores by subtracting the mean and dividing by the standard deviation over all the programs on that data set. Average these normalized scores over the data sets in *C*. This method puts easy and hard data sets on the same scale.
3. **Combined.** Add *nTP*, *nFP*, *nFN*, *nTN*, *sTP*, *sFP* and *sFN* over the data sets in *C*, and compute the measure *M* as though *C* were one large data set. For measures such as *Sn* and *PPV*, this is exactly a weighted average, where each term is weighted by its denominator. This method has the advantage that the measure *M* is rarely undefined. However, the problem posed when a tool makes no prediction on a data set is still present, since the method treats that data set as weighted by zero in the weighted average.

There were few qualitative differences among these three methods of summarizing, except that averaging *nPPV*, *sPPV* and *nCC* scores tends to reward programs that make no predictions on many data sets. The results presented here all use the 'Combined' method.

The assessment web site (http://bio.cs.washington.edu/assessment/) provides tools for computing these statistics on predictions made by developers who want to test a new tool on the benchmark data sets.

varied dramatically. These factors of judgment will have had an impact on each algorithms' performance. When comparing tools using the statistics *nPPV*, *nCC*, *sPPV* and *sASP*, one must also keep in mind the fact that these measures are affected by the data sets on which the tools predicted no motif, as explained in **Box 2**.

With the caveats outlined above, several interesting observations on the results can be made. First, an inspection of **Figure 1** reveals that two different versions of the tool MEME were run independently by two experts. It is gratifying to see that, despite the great room for human judgment in the choice of parameters and the final choice of a single (or no) motif per data set, the two collections of MEME results are remarkably consistent across all the measurements.

The tool Weeder outperformed the other tools in most domains and by most measures in this assessment. We believe that some part of

Table 2 Number of data sets for which each tool predicted no motif[a]

Tool	Total (56)	Fly (8)	Mouse (12)	Human (26)	Yeast (10)
AlignACE	32	7	5	17	3
ANN-Spec	3	1	0	1	1
Consensus	37	4	3	26	4
GLAM	3	0	1	2	0
Improbizer	0	0	0	0	0
MEME	6	1	2	2	1
MEME3	14	0	5	8	1
QuickScore	20	2	4	14	0
SeSiMCMC	0	0	0	0	0
MITRA	11	7	3	0	1
MotifSampler	7	2	2	0	3
Oligo/dyad-analysis	23	1	5	13	4
Weeder	17	3	3	10	1
YMF	7	0	2	4	1

[a]The total number of data sets is given parenthetically in the column header.

Weeder's success is due to judicious choices regarding when to predict no motif in a data set: Weeder was run in a 'cautious mode,' where only the strongest motifs were reported. A few small exceptions to Weeder's domination are shown in **Figure 1b**, where SeSiMCMC did somewhat better on the fly data sets, and MEME3 and YMF somewhat better on the mouse data sets.

What is most striking about **Figure 1b** is the fact that so many tools perform much better on the yeast data sets than on other species. This suggests that computational biologists have been more successful at modeling binding sites in yeast than in metazoans. Little significance should be read into the slightly negative *nCC* values in **Figure 1b**: these are so close to zero that they should be interpreted simply as no correlation between the known and predicted binding sites.

Although the shapes of the curves are very similar in **Figure 1a** and **Figure 1d**, the scale is different. Nearly all tools performed better according to most of the seven measures when the data sets of type real were removed. For example, the correlation coefficient *nCC*, averaged over all tools, improved by 39% from **Figure 1a** to **Figure 1d**. This seems to say more about the experimental design than about the tools themselves: it is likely that the data sets of type real contain functional motifs other than the single TRANSFAC binding site on which they were scored, and that tools that discovered other functional motifs were unduly penalized. The tool most affected by this is YMF, whose seven measures each improved between 45% and 67% when the real data sets were removed. Interestingly, there is one tool that did not improve by this removal: MotifSampler's performance was somewhat better on the data sets of type real than on the others. This aspect of MotifSampler can also be seen in **Figure 1c** for the measure *nCC*.

We have not discovered any simple feature, such as type of motif search, that determines the accuracy of these tools. Nor should we expect such a simple conclusion: the tools are based on algorithms and motif models that are varied and complex, and predicting their performance on complex data is beyond our current analytical ability.

Table 3 shows some very interesting complementary behaviors among certain pairs of tools. For example, MotifSampler's predictions complement well the predictions of MEME, oligo/dyad-analysis, ANN-Spec and YMF, improving their individual *nCC* scores by 64–92%. It is also informative to see that MEME's predictions improve the individual *nCC* score of MEME3 by 53%. This gives some idea of the improvement possible by allowing a given tool to predict two motifs rather than just one.

Exploiting comparative sequence analysis, using tools not covered in this assessment, provides a powerful adjunct to these methods. As an example, a recent tool called PhyME that combines intraspecies overrepresentation and interspecies conservation reported success[23] in predicting the binding sites for one of the most difficult human data

Table 3 Correlation coefficient (*nCC*) for all pairs of tools[a]

	Quick score	GLAM	SeSi MCMC	MITRA	Consen	Improb	Align ACE	Motif sampler	MEME3	MEME	Oligo/ dyad	ANN-Spec	YMF	Weeder
QuickScore	0.009	0.020	0.042	0.030	0.025	0.052	0.068	0.072	0.072	0.074	0.038	0.064	0.061	0.084
GLAM	0.031	0.016	0.060	0.037	0.039	0.068	0.066	0.084	0.088	0.086	0.052	0.082	0.090	0.113
SeSiMCMC	0.049	0.059	0.024	0.068	0.042	0.083	0.071	0.091	0.081	0.088	0.058	0.103	0.104	0.092
MITRA	0.042	0.041	0.072	0.031	0.054	0.082	0.084	0.097	0.106	0.105	0.070	0.101	0.103	0.131
Consensus	0.067	0.060	0.075	0.053	0.042	0.077	0.079	0.109	0.084	0.077	0.074	0.082	0.081	0.098
Improbizer	0.065	0.069	0.083	0.077	0.056	0.052	0.089	0.117	0.096	0.098	0.083	0.112	0.091	0.117
AlignACE	0.088	0.084	0.089	0.090	0.085	0.111	0.068	0.097	0.102	0.091	0.088	0.091	0.115	0.119
MotifSampler	0.071	0.092	0.107	0.097	0.077	0.103	0.099	0.068	0.112	0.119	0.103	0.127	0.130	0.134
MEME3	0.089	0.094	0.092	0.102	0.074	0.102	0.093	0.124	0.069	0.106	0.094	0.129	0.126	0.114
MEME	0.091	0.090	0.100	0.102	0.077	0.091	0.095	0.120	0.100	0.073	0.104	0.123	0.121	0.121
Oligo/dyad	0.073	0.088	0.111	0.088	0.082	0.082	0.099	0.136	0.119	0.112	0.071	0.106	0.107	0.130
ANN-Spec	0.085	0.091	0.111	0.094	0.090	0.100	0.085	0.122	0.114	0.110	0.089	0.074	0.118	0.117
YMF	0.094	0.095	0.112	0.101	0.093	0.100	0.114	0.146	0.121	0.129	0.092	0.131	0.084	0.137
Weeder	0.164	0.169	0.162	0.167	0.157	0.171	0.166	0.186	0.168	0.164	0.173	0.167	0.167	0.156

[a]The primary tool is listed in the row header and the secondary tool in the column header. The score shown for the same tool on both axes (that is, along the main diagonal) is the individual *nCC* score from **Figure 1.** Numerical values are categorized by color, ranging from dark blue (poorer predictions) to red (better predictions).

sets from this assessment (data set hm20, corresponding to the human transcription factor Sp1) on which all the assessed tools performed extremely poorly.

Despite considerable effort to date, prediction of regulatory elements remains a wonderful and complex challenge for computational biologists. Biologists would be well advised to use a few complementary tools in combination rather than relying on a single one and to pursue the top few predicted motifs of each rather than the single most significant motif. Even so, more work is clearly needed to optimize tool performance, particularly in the modeling of regulatory elements in the metazoans. The assessment web site (see above) can be used to identify particularly challenging data sets, those on which none of the tools succeeded in predicting the known binding sites with any accuracy. Further testing on these data will hopefully lead to improved prediction methods.

One of the surprises resulting from this assessment is the realization that the design of a good assessment is itself far from straightforward. Constructing representative data sets, when we do not understand the full truth about transcription factor binding sites, is problematic from the outset. Choosing the most appropriate statistics for evaluating the correctness of predictions is also challenging. This is particularly the case in light of the reality that different tools may predict zero, one or more significant motifs on a given data set, and that real promoter sequences may indeed contain binding sites for zero, one or more distinct transcription factors.

For the next such assessment of motif discovery tools, we suggest the following changes in experimental design: (i) eliminate the data sets of type 'real,' (ii) eliminate the negative control data sets that contain no planted binding sites (iii) require each tool to predict exactly (say) three motifs per data set and (iv) for each tool and data set, choose the predicted motif with the greatest *nCC* score to represent that tool. These changes will eliminate the great difficulties caused by undefined statistics and will raise the sensitivity of the tools closer to what it would be in practice, when a user pursues the top few motifs rather than just the top one.

ACKNOWLEDGMENTS

We thank Mathieu Blanchette, Ari Frank, Phil Green, Susan Hewitt, S.N. Maheshwari, Larry Ruzzo, Terry Speed, Gary Stormo and the organizers and participants of the 2002 Bellairs Workshop on Computational Biology for their important contributions to this project. Martin Tompa and Nan Li were supported by National Science Foundation (NSF) grant DBI-0218798 and by National Institutes of Health (NIH) grant R01 HG02602. Alexander Favorov, Andrei Mironov and Vsevolod Makeev were supported by Howard Hughes Medical Institute grant 55000309, Ludwig Cancer Research Institute grant CRDF RBO-1268-MO-02, Russian Fund of Basic Research grant 04-07-90270 and support from the Russian Academy of Sciences Presidium Program in Molecular and Cellular Biology, project no. 10. Yutao Fu, Martin C. Frith and Zhiping Weng were supported by NSF grant DBI-0116574 and NIH NHGRI grant 1R01HG03110. Giulio Pavesi and Graziano Pesole were supported by the Italian Ministry of University and Scientific Research's Fondo Italiano per la Ricerca di Base project 'Bioinformatica per la Genomica e la Proteomica' and by Telethon. Nicolas Simonis and Jacques van Helden were supported by the European Communities grant QLRI-199-01333, by the Action de Recherches Concertées de la Communauté Française de Belgique and by the Government of the Brussels Region. Saurabh Sinha was supported by a Keck Foundation Fellowship.

Gert Thijs and Bart De Moor were supported by Geconcerteerde Onderzoeks-Acties Mefisto-666 and Ambiorics, InterUniversity Attraction Pole V-22, and several funded projects of the Institut voor de aanmoediging van Innovatie door Wetenshap en Technologie in Vlaanderen, Fonds voor Wetenshappelijk Onderzoek, and European Union. Zhou Zhu is a Howard Hughes Medical Institute predoctoral fellow. Zhou Zhu and George Church were supported by the Department of Energy and the Lipper Foundation.

COMPETING FINANCIAL INTERESTS

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/

1. Pevzner, P. & Sze, S.-H. Combinatorial approaches to finding subtle signals in DNA sequences. in *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology* (ed. Altman, R. *et al.*). 269–278 (AAAI Press, Menlo Park, CA, 2000).
2. Sinha, S. & Tompa, M. Performance comparison of algorithms for finding transcription factor binding sites. in *3rd IEEE Symposium on Bioinformatics and Bioengineering* (ed. Bourbakis, N.G.). 214–220 (IEEE Computer Society, New York, 2003).
3. Burset, M. & Guigó, R. Evaluation of gene structure prediction programs. *Genomics* **34**, 353–367 (1996).
4. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
5. Reese, M.G. *et al.* Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* **10**, 483–501 (2000).
6. Ashburner, M. A biologist's view of the *Drosophila* genome annotation assessment project. *Genome Res.* **10**, 391–393 (2000).
7. Hughes, J.D., Estep, P.W., Tavazoie, S. & Church, G.M. Computational identification of *cis*-regulatory elements associated with functionally coherent groups of genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **296**, 1205–1214 (2000).
8. Workman, C.T. & Stormo, G.D. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. in *Pacific Symposium on Biocomputing* (ed. Altman, R., Dunker, A.K., Hunter, L. & Klein, T.E.). 467–478 (Stanford University, Stanford, CA, 2000).
9. Hertz, G.Z. & Stormo, G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563–577 (1999).
10. Frith, M.C., Hansen, U., Spouge, J.L. & Weng, Z. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res.* **32**, 189–200 (2004).
11. Ao, W., Gaudet, J., Kent, W.J., Muttumu, S. & Mango, S.E. Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science* **305**, 1743–1746 (2004).
12. Bailey, T.L. & Elkan, C. The value of prior knowledge in discovering motifs with MEME. in *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*. 21–29 (AAAI Press, Menlo Park, CA, 1995).
13. Eskin, E. & Pevzner, P. Finding composite regulatory patterns in DNA sequences. *Bioinformatics* (Supplement 1) **18**, S354–S363 (2002).
14. Thijs, G. *et al.* A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* **17**, 1113–1122 (2001).
15. van Helden, J., Andre, B. & Collado-Vides, J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* **281**, 827–842 (1998).
16. van Helden, J., Rios, A.F. & Collado-Vides, J. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.* **28**, 1808–1818 (2000).
17. Régnier, M. & Denise, A. Rare events and conditional events on random strings. *Discrete Math. Theor. Comput. Sci.* **6**, 191–214 (2004).
18. Favorov, A.V., Gelfand, M.S., Gerasimova, A.V., Mironov, A.A. & Makeev, V.J. Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length and its validation on the ArcA binding sites. in *Proceedings of BGRS 2004* (BGRS, Novosibirsk, 2004).
19. Pavesi, G., Mereghetti, P., Mauri, G. & Pesole, G. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.* **32**, W199–W203 (2004).
20. Sinha, S. & Tompa, M. YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.* **31**, 3586–3588 (2003).
21. Wingender, E., Dietze, P., Karas, H. & Knüppel, R. TRANSFAC: a Database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* **24**, 238–241 (1996).
22. Moult, J., Fidelis, K., Zemla, A. & Hubbard, T. Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins* **53**, 334–339 (2003).
23. Sinha, S., Blanchette, M. & Tompa, M. PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformat.* **5**, 170 (2004).