NestedMICA: Sensitive inference of over-represented motifs in nucleic acid sequence

Thomas A. Down and Tim J.P. Hubbard
Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA
{*td2,th*}*@sanger.ac.uk*

24^{th} January 2005

Abstract

NestedMICA is a new, scalable, pattern-discovery system for finding transcription factor binding sites and similar motifs in biological sequences. Like several previous methods, NestedMICA tackles this problem by optimising a probabilistic mixture model to fit a set of sequences. However the use of a newly developed inference strategy called Nested Sampling means NestedMICA is able to find optimal solutions without the need for a problematic initialisation or seeding step. We investigate the performance of NestedMICA in a range scenarios – on synthetic data and a well-characterised set of muscle regulatory regions – and compare it to the popular MEME program. We show that the new method is significantly more sensitive than MEME: in one case, it successfully extracted a target motif from background sequence four times longer than could be handled by the existing program. It also performs robustly on synthetic sequences containing multiple significant motifs. When tested on a real set of regulatory sequences, NestedMICA produced motifs which were good predictors for all five abundant classes of annotated binding sites.

INTRODUCTION

Motif-finding is a long standing problem in sequence bioinformatics, with a history going back over twenty years (1). A typical statement of the problem would be "given a set of sequences, which motifs are significantly over-represented with respect to a given background model." The term "motif" could refer to a single, perfectly specified, word, but usually describes a family of words, with at least some positions where several alternate symbols are acceptable. For example, both TATATAAA and TATAAAAA are good TATA-box sequences (2). A classical application for motif-finding software is the discovery of novel transcription factor binding sites in transcriptional regulatory regions, but there are other interesting functional elements in biological sequences – both nucleic acid and protein – which can be found by motif-discovery methods. While the program described here has been developed and tested on DNA sequences, the techniques are all applicable to other types of sequence and therefore we prefer the general term "symbol" to describe an element of a sequence.

Motif-finding strategies can be broadly divided into two classes: those which rely on exhaustively enumerating a set of motifs – for example all nucleotide n-mers, then reporting the most frequent or overrepresented, and those which find the most significant motifs by fitting a probabilistic model to the sequence data. Exhaustive enumeration can be very fast when implemented with optimised data structures like suffix trees (3), and is a good strategy for finding totally constrained motifs (*i.e.* every instance is identical). However, for typical transcription factor binding sites which often have several weakly constrained positions, exhaustive enumeration becomes problematic and the results usually have to be post-processed with some kind of clustering system as in (4). We will not consider exhaustive enumeration here further.

Probabilistic motif-finders treat the supplied sequences as a mixture of interesting motifs and non-interesting – at least from this point of view – background sequence. We therefore refer to them as Sequence Mixture Models (SMMs). In principle, any probabilistic model can be used to represent the interesting motifs, but the usual choice is the Position-Weight matrix (PWM) (2). This is a model which treats each position in the motif independently, and records a probability distribution over symbols which can be observed at that position. PWMs are a good way of modelling

motifs which have a mixture of highly constrained and weakly constrained positions, but they lack any capacity to record possible correlations between positions in the motif – a factor which could be significant in real interactions between proteins and nucleic acid (5). PWMs are often visualised as a pictogram where each position is represented by a stack of letters whose height is proportional to the information content of that position (6). This 'logo' representation of PWMs is used throughout the results section of this paper.

The probabilistic motif-finding problem has classically been reduced to a simple case: considering just one motif at a time, model each sequence with a random background model which may, or may not, contain a single instance of the motif under consideration. This is the zero-or-one occurrences per sequence (ZOOPS) model. It can be easily represented as a hidden Markov model (7), as shown in figure 1, and standard techniques such as expectation maximisation (8) or Gibbs sampling (9) can be used to find high-likelihood sets of model parameters, corresponding to good motif models.

There are several significant concerns about this strategy, which we have tried to address in this work. Firstly, real regulatory regions, and most other contexts where interesting motifs can be found, usually contain more than one distinct functional motif. Many regulatory regions also contain several instances of the same motif – at least in some contexts seeing five or more binding sites for a single transcription factor in less than a kilobase of sequence is not unusual (10). Programs which use a ZOOPS-like model work around these issues by finding the strongest motif in a set, then scanning for all its instances, masking them out, and re-running the process on the remaining sequence. This strategy is greedy and it is by no means clear that its behaviour will be optimal, especially when working on a system where there is a set of closely related, yet still distinct, motif types. In a genomic environment where novel transcription factors are created by gene duplication then diverge to perform a new function, such situations seem quite probable, but we do not know of any investigation into the behaviour of motif-finders when faced with related but distinct motifs.

Another major concern with existing techniques for optimising or exploring sequence mixture models is that that they tend to be strongly local in nature: the optimisation concentrates on regions of the probability landscape close to their starting point. This is clear for expectation-maximisation methods, which always move

in a direction which increases the likelihood of the model. This can lead to a local maximum which can never be left since every direction leads to a lower likelihood. Strategies based on Monte-Carlo sampling methods do not, in theory, suffer from this limitation, but in practice crossing the low-likelihood valley between two high-likelihood peaks tends to be an unlikely event, often to the point where it becomes vanishingly rare.

Here we present a novel method, NestedMICA, which avoids both these issues, firstly by using a sequence model based on the independent component analysis framework to learn models for multiple motifs simultaneously, and secondly by using an alternative inference strategy which is likely to find a globally optimal model in a single run. NestedMICA has also been implemented in a fashion which allows arbitrary background models to be plugged in, allowing the investigation of more sophisticated backgrounds. We discuss a general-purpose background model in this paper, but it is also possible to develop highly specialised backgrounds, for example to search for motifs embedded in protein-coding sequence (B. Leong, unpublished).

MATERIALS AND METHODS

Motif ICA

We treat finding motifs in a set of sequences as a form of independent component analysis (ICA) problem (11, 12). In linear ICA, a matrix of observations, X is approximated as a linear mixture, A, of some sources, s:

$$x = As + \nu \tag{1}$$

(where ν is a noise matrix representing any errors in the linear approximation). A classical example is the cocktail party problem where a set of M microphones record different mixtures of the voices of N speakers. Given samples from these microphones at t time points, ICA methods attempt to factorise the $M \times t$ observation matrix into a $N \times t$ source matrix and a $M \times N$ mixing matrix.

While this straightforward view of mixing as matrix multiplication is clearly not directly applicable to strings such as biological sequence data, if we can find a satisfactory alternative definition for the mixing operator, we can handle a wide variety of problems within an ICA-like mixture modelling framework.

In motif ICA (MICA), the sources are short sequence motifs (currently, but not necessarily, modelled as position-weight matrices (2)), while the observations are larger sequences. There are several possible interpretations of the ICA mixing matrix. In the implementation described here, we use a binary mixing matrix (all coefficients are either 0 or 1), and a given sequence is expected to contain a given motif if the relevant mixing coefficient is 1. The 'noise' part of the ICA model represents all the sequence that is not modelled by one of the motifs.

ICA problems can be handled in a Bayesian probabilistic framework by writing a likelihood function which defines the probability of a set of observations given particular source and mixing matrices (12). For linear ICA a typical likelihood would be a Gaussian distribution centred on As, with the Gaussian modelling the noise part of the ICA model. For each sequence, we collect the set of motifs with non-zero mixing coefficients, and generate a hidden Markov model as shown in figure 2. This is somewhat similar to the ZOOPS HMM except that there is (potentially) more than one motif, and it is possible to pass through a given motif more than once while generating a single observed sequence.

Given a likelihood function – in this case the probability of a sequence being generated by the HMM – we can place priors over the parameters of the model (the source and mixing matrix) then perform Bayesian inference (13) to find likely values for the parameters given a set of data. A number of inference strategies exist, and the choice is important: not all strategies are guaranteed to explore the whole of parameter space. We chose to use a new and powerful inference strategy, nested sampling, which is described below.

Nested sampling

Nested sampling is a novel approach to performing probabilistic inference in a Bayesian framework, proposed recently by John Skilling (unpublished, manuscript available at `http://www.inference.phy.cam.ac.uk/bayesys/`). Along with existing methods such as Metropolis-Hastings and Gibbs Sampling (both described in (13)), it can be classified as a Monte Carlo method, since the process is driven forward by a series of randomly chosen events. However, nested sampling is quite distinct from the family of classic Monte Carlo methods. While Metropolis-Hastings and all its derivatives rely on making an unbiased random exploration of the probability landscape, nested sampling proceeds in a more orderly fashion.

Nested sampling is always applied to an ensemble of states – typically a few hundred – each of which represents a possible solution to the problem at hand. The ensemble is initialised by sampling uniformly from the prior distribution, then sorting the states according to their likelihoods. Each state in the ensemble is considered to be a representative of the set of states with similar likelihoods. If the likelihood of each state is drawn as a contour on the likelihood distribution, we see a nested set of contour lines, converging towards the peaks of the likelihood distribution. We therefore call the ordered set of states a nested ensemble. For each cycle of nested sampling, the least likely state in the ensemble is discarded, and a new state is chosen by sampling uniformly from the prior subject to the constraint that the likelihood of the new state must be greater than or equal to the likelihood of the discarded state. The exact strategy used to draw constrained samples from the prior should not be important for the final results, but the usual strategy – recommended by Skilling and employed in our implementation – is to randomly pick an existing state from the ensemble, duplicate it, then use conventional Monte-Carlo techniques to move the new state to a new point in the prior. Since priors are generally much smoother

than likelihood functions (indeed we use a uniform prior over weight-matrix space), drawing good quality samples from the prior in this way does not pose any great technical difficulties.

Nested sampling has some similarity to simulated annealing techniques (13) in that during the course of the sampling process, we move from a situation where the distribution of states is defined by the prior to a situation where the distribution of states is influenced mainly by the likelihood function. But unlike annealing, there is no temperature parameter to control, and no risk of states becoming trapped because of phase-change events.

In this context, the most exciting property of nested sampling is that, given a reasonably large ensemble, the final sample drawn from a converged nested sampler can be expected to reflect the global optimum of the likelihood landscape. Moreover, in cases where more than one globally significant optimum exists, these should be represented in the sample set in direct proportion to the amount of posterior mass they represent.

Mosaic background sequence models

The background model is an important component of the SMM framework – after all, it will usually be responsible for modelling the majority of the input sequence. The simplest strategy – and still a common one – is to treat all non-motif sites as independent and identically distributed (i.i.d.). In HMM terms, this makes the background model a zeroth-order Markov chain. However experience shows that genomic DNA sequence, even in apparently non-functional areas, is not a good fit to the i.i.d. model. The best known deviation is the dramatic under-representation of CpG dinucleotides in most parts of vertebrate genomes, but other significant effects have been reported (14). In any case, practical experience shows that motif finders equipped with naive background models tend to report low-complexity elements rather than interesting binding sites.

The first obvious improvement is to replace the zeroth order Markov chain with a first order chain (*i.e.* the background probability of observing a particular symbol at position n depends on the symbol at position $n-1$). This model is good at capturing anomalies like the CpG underrepresentation. Success with first order background models has led some researchers to investigate higher order models. One investigation of Markov chain backgrounds can be found in (15): this concludes

that pentanucleotide frequency tables (*i.e.* fourth-order Markov chains) are optimal. However, there are two concerns about this result: firstly, it leaves an open question about what these high-order correlations in background sequence mean (and why fourth-order models appear to outperform fifth-order). Also, training a background model generally requires sequence proportional to the number of free parameters in the model. Fifth order models, with 768 parameters, therefore require large amounts of sequence. Moreover, it is desirable to train the background model on sequence which does not contain target motifs, since a fifth order model could easily capture some information about this motifs, thereby reducing the sensitivity of the motif-finding process. But it is hard to find large amounts of representative background training sequence which doesn't contain interesting motifs.

A different way to generalise the naive background model is to allow several different classes of sequence, each with its own particular base distribution (which could be zeroth-order or higher-order). We call these *mosaic models*, since their underlying assumption is that genome evolution includes some set of constraints which act non-uniformly, even on background sequence.

To test the effect of mosaic models, we took a set of 192 non-redundant human promoter sequences from release 69 of the Eukaryotic Promoter Database (16). These were split into 142 training sequences and 50 test sequences. For each model architecture, the parameters were optimised on the training sequences using the Baum-Welch algorithm (7), as implemented in the BioJava HMM library, then the likelihood of the test sequences given those learned parameters was calculated. Test likelihoods for a variety of class numbers and Markov chain orders are shown in figure 3. Considering just the one class 'mosaics' – equivalent to classical Markov chain background models – we repeat the previously reported observation that higher order Markov chains are better models of genomic DNA. However, we also see large increases in likelihood when moving to larger numbers of mosaic classes. Interestingly, the lines for zeroth-order and first-order models run almost parallel: this suggests that the benefits of mosaic models are almost orthogonal to the benefits of first-order models. However, this is not true when moving beyond first-order models.

Based on these results, we recommend the use of a four class, first order, mosaic background model for most motif-finding applications on mammalian genomic sequence. In practice, the four classes appear to include a C+G rich class (corresponding to classically reported CpG islands), a purine-rich class, a pyrimidine-rich

class, and a final relatively neutral class. This four-class background model is used for all subsequent NestedMICA tests in this paper, and is available to download from the NestedMICA web site.

Synthetic data spiked with known motifs

Non-repetitive intergenic regions of various lengths were extracted randomly from the human genome (release NCBI34) using gene and repeat annotation from the Ensembl human database release 20.34 (17). To generate test sequences for motif-finding programs, we selected experimentally derived transcription factor weight matrices from the JASPAR database (18), and generated target motifs by sampling from the weight matrices, assuming each position of the motif is independent. Target motifs were inserted into the intergenic regions at random positions. In cases where more than one motif was inserted into a single sequence, non-overlapping positions were chosen.

Muscle regulatory regions with annotated binding sites

Sequences for muscle regulatory regions, as described in (19) were downloaded from `http://www.cbil.upenn.edu/MTIR/DATATOC.html`. We took binding site annotation from the HTML pages linked from `http://www.cbil.upenn.edu/MTIR/HomePage.html` and manually mapped it back to the FASTA-formatted sequence file.

Weight matrix ROC curves and scores

Log-likelihood weight matrix scores were calculated for each possible position in the set of test sequences, then the complete list of hits was sorted by score. Hits were classified as correct if they overlapped the annotated binding sites for a target transcription factor, incorrect otherwise. We calculated accuracy (proportion of correct hits) and coverage (proportion of annotated binding sites covered by at least one hit) for successively larger head-lists (initially just the highest scoring hit, then the best two, and so on until the complete list is used). Plotting accuracy against coverage gives a form of receiver operating characteristic (ROC) curve.

For comparison purposes, the area under ROC curves was calculated by direct summation. At the same time, we calculated the expected ROC score if high-scoring hits were distributed randomly along the sequence.

NestedMICA implementation

NestedMICA was implemented in Java, with a small amount of C code for loops in the dynamic programming code responsible for calculating sequence likelihoods. The primary motivation for using C was the availability of optimising compilers which could rewrite the key loops to use vector processing capabilities of certain modern CPUs (e.g. Pentium 4s). The BioJava library (http://www.biojava.org/) was used for loading sequence data and manipulating motifs and PWMs. The main program was developed on Linux and Mac OS X machines, but should be easy to port to any platform with a good Java implementation. Source code, documentation, and test datasets can be downloaded from http://www.sanger.ac.uk/Users/td2/nmica/

Analysing a 70kb sequence set takes around 3-4 hours on one Pentium IV processor at 2.8GHz. Processing time is dominated by the dynamic programming routines which evaluate the likelihood of the sequence set. Execution time therefore scales linearly with the number of sequences, meaning that analysis of large datasets is feasible. In addition, the likelihood of each sequence can be calculated independently, which offers a natural and efficient way of dividing the workload up between multiple processors.

RESULTS

Testing on simple synthetic data

Evaluating the relative performance of motif-finding software on real data is difficult, because there are very few large collections of sequences where we can be confident that every functional binding site has been accurately annotated. Therefore we generated synthetic evaluation sequences containing a known number of known sequence motifs. To make the synthetic data as realistic as possible, our synthetic data was based on sequence fragments taken from intergenic regions of the human genome, into which we inserted experimentally derived human transcription factor binding sites from the JASPAR collection (18).

Our basic test strategy was to take a set of 100 intergenic sequences of a particular length, then spike the known motif into 50 of these. We chose to focus on sets of 100 sequences because this is the typical order of magnitude for clusters of co-regulated genes selected from contemporary experiments such as microarrays – for example (4). We only placed the target motif into half of these sequences since this makes the motif-finding problem considerably more challenging – it becomes necessary to determine which sequences contain motifs, rather than merely discover their locations – and because it is rare to obtain a large set of sequences which are known with 100% certainty to contain the same functional element.

We investigated a number of human motifs from JASPAR, representing binding sites from a range of major transcription factor families. We analysed each set of sequences using the NestedMICA program as described here, and also with MEME version 3.0.4 (8). Both methods were run with default options. For NestedMICA, background model generation is a separate step. We used a general four-class human background model, learned from the EPD sequences discussed previously.

Both programs tested here tended to fail rapidly. By this, we mean that, below a certain threshold sequence length (which depends on the method) the recovered motif was always very similar to the target, while above the threshold length a dramatically different motif is found. Examples of this are shown in figure 4. This rapid failure makes it possible to quantify the performance of a method for finding a particular motif by identifying the longest set of sequences from which it can be successfully recovered.

Results for a selection of JASPAR motifs are shown in tables 1-3. For reference, the subset of JASPAR used in the tests published here is shown in figure 5.

NestedMICA proved to be significantly more sensitive in most cases. The extent of the difference varies depending on the motif in question. In the case of HLF, NestedMICA successfully retrieves the expected motifs from sequences four times as long as the longest handled by MEME. At the other extreme, the sensitivity of both methods was similar when searching for the HFH-1 motif. Considering these two motifs, we note that HFH-1 has a highly constrained core, with a central GTTT sequence which is conserved in all instances. On the other hand, HLF has no such obvious core, and indeed the JASPAR profile contains no single position which is totally constrained. We suggest that motifs with highly constrained cores may be favoured by MEME's seeding heuristics

Synthetic data with decoy motifs

Real regulatory regions do not contain single instances of single motifs. Therefore, we also tested the response of MEME and NestedMICA on sequences containing multiple motifs. We picked two of the JASPAR motifs discussed above (CREB and Tal1beta) and spiked 50 instances of each into independently chosen subsets of the intergenic background sequences (*i.e.* about a quarter of the sequences were spiked with both motifs, and a quarter contained no motifs). MEME and NestedMICA were run with the same parameters as before, except that they were told to find two motifs (`-nmotifs 2` for MEME, `-numMotifs 2` for NestedMICA).

We assessed the ability of NestedMICA and MEME to find the CREB binding sites in the presence of the decoy Tal1 sites. Results are shown in table 4.

The presence of a decoy motif makes little difference to the discovery of CREB by NestedMICA. But while MEME can successfully find this motif in 400 base sequences with no decoy, it fails in the presence of the Tal1beta decoy. We suggest that the presence of multiple overrepresented motifs makes it harder to pick a good starting point for expectation maximisation algorithms.

Analysis of muscle regulatory regions

Finding real biological sequence with comprehensive, high-quality annotation of transcription factor binding sites is difficult, but some such data does exist. One well-

known collection is a set of confirmed regulatory for muscle-specific genes, curated by Wasserman and Fickett (19). This is still a relatively small dataset: 43 sequences, mostly of around 300 bases in length, with significant redundancy (orthologous regions from related species). Binding sites for a number of transcription factors are well-annotated within these regions, allowing formal testing of motif-finding software.

We ran NestedMICA on the complete set of 43 sequences with default options, requesting 20 motifs of up to 12 bases in length. A four-class mosaic background model learned from a large set of human upstream regions was used for this test. We also ran MEME on the same sequences, again requesting 20 motifs of 12 bases with default options.

Weight matrices can be used to scan sequence and provide a score at each position, which we hope is indicative of the affinity of transcription factor binding at that position (20). To predict a set of sites, it is necessary to specify a score threshold. The choice of threshold controls the trade-off between accuracy and coverage. This makes evaluating the quality of weight matrices (and other predictive models) from different sources difficult since it is not obvious whether a model which gives high coverage at low accuracy has more or less predictive power than another model which gives much better accuracy at the expense of coverage. The solution is to consider the receiver operating characteristic (ROC) curves for each model. These are graphs of accuracy against coverage for a variety of score thresholds. Having obtained the data for a ROC curve, we can either inspect them visually or calculate the total area under the curve (sometimes called the ROC score), which gives a threshold-independent measure of a model's predictive power. A model which can predict all the sites in the data set (100% coverage) with no false positives (100% accuracy) will receieve the maximum possible ROC score of 1.0. On the other hand, a model with no predictive power will be given a ROC score equal to the fraction of positions in the dataset which are considered to be correct. Since in this case the targets are a relatively small number of annotated binding sites in a large set of sequences, the expected random ROC scores are rather low (less than 0.05 even in the case of the most abundant binding site, MyoD). A model which predicted all the sites with a uniform 50% accuracy would get a score of 0.5 but – perhaps more realistically – a method which found half the sites with a very high accuracy but only found the remainder with a very relaxed threshold and consequently much lower accuracy would also score around 0.5. It should be understood that it is not

necessarily realistic to hope for a ROC score of exactly 1.0: in particular, there may be some real binding sites which have been missed by annotation, which will lead to apparent false positives and prevent 100% accuracy being reached. Nevertheless, higher ROC scores are a good indication of better predictive power.

We calculated ROC scores as described in the methods section for all the motifs learned by both MEME and NestedMICA against each factor for which more than 5 binding sites were annotated. For each factor, we picked the highest-scoring motif from each method. ROC scores are listed in table 5, and examples of complete ROC curves for SRE sites are shown in figure 6.

In all but one case, MEF2, the NestedMICA weight matrix received a higher ROC score than the equivalent MEME weight matrix. In the case of MyoD, none of the MEME motifs had any significant predictive power – a surprising result since MyoD was the most common motif in the dataset with 40 annotated instances.

In some cases, reference weight matrices were already available, based on manual alignment of the curated site sequences. In figure 7, the reference MEF2 weight matrix is compared to the best-scoring matrices from MEME and NestedMICA. In this case, both programs generate a weight matrix which is instantly recognisable as being similar to the reference motif. The NestedMICA motif is shifted by one base to the right compared to the reference and MEME motifs, and it is possible that this explains the slightly higher predictive power of the MEME motif in this case. Nevertheless, the results are generally extremely similar.

A rather different situation can be seen in figure 8. Once again, visual inspection shows that the NestedMICA result is very similar to the reference motif, so the good ROC score is unsurprising. However, the highest-scoring MEME motif has no obvious similarity. In this case, the most surprising result is that the MEME motif got a high ROC score at all. Looking at the ROC graph for this motif in figure 6, we see that although three instances of SRE are covered with good accuracy, the remaining 8 instances are not detected even with much more relaxed thresholds. We therefore believe that the MEME motif may be discovering some feature which lies close to several of the annotated SRE sites, rather than the sites themselves.

We cannot say for certain why MEME finds MEF2 but misses SRE. One possibility is simple numbers: there are 13 annotated MEF2 sites but only 11 SREs. However, this does not seem like a particularly large difference, especially considering that MEME also fails to find the 40 MyoD sites. An alternative consideration is that

the MEF2 site has a high-information core, including a perfectly constrained TAT sequence, while SRE does not have a clear core. Preferential seeding of motifs with high-information cores by MEME is consistent with the results from our synthetic data tests.

DISCUSSION

We were able to compare different motif-finding methods in a quantitative fashion by searching for known motifs in synthetic datasets. Since these were based on real intergenic sequence and experimentally-derived binding sites, we believe that these results should be representative for real data. One possible criticism is that the synthetic motifs are sampled from weight matrices while assuming that each position in the motif is independent. This assumption is known to be incorrect in at least some cases (5). However, since this assumption is built into the sequence models for both MEME and NestedMICA, we do not expect it to significantly affect the comparisons we provide here.

NestedMICA outperforms existing methods such as MEME when discovering most known regulatory motifs from the JASPAR database. In general, we suggest that MEME (and methods which use similar seeding strategies) will perform best when searching for motifs with a core of very highly constrained bases. The advantages of using a non-seed-based strategy are greatest when considering motifs with few positions that are 100% constrained, such as the HLF motif shown in figure 5. Extending the analysis to the somewhat more realistic case of a dataset containing two different known motifs, we find that NestedMICA responds robustly, and still finds the expected element as well as the decoy. This result makes us optimistic that NestedMICA will also perform well when faced with a real set of regulatory sequences, containing a variety of functional motifs.

Real data sets that are sufficiently well annotated to allow rigorous evaluation are currently rare and limited in size. We have, however, tested NestedMICA on one small but high-quality set of muscle regulatory regions. We learned weight matrices which detected many of the experimentally annotated binding sites with good predictive power, and which also agreed well with weight matrices determined directly from curated sets of binding sites. In four out of five cases, PWMs from NestedMICA outperformed those from MEME.

During the preparation of this manuscript, we learned about a very different scheme for evaluating motif finders, described in (21). This is interesting because it includes evaluation results from 13 different sets of predictions on a single set of test sequences (including two sets of MEME predictions, submitted by different experts using different post-processing strategies). The benchmark is not ideally

suited to NestedMICA, since our program is designed to find PWMs rather than sets of motif instances. We were also concerned about the data preparation, in particular the fact that some datasets consisted of unrealistically small numbers of sequences. Nevertheless, we ran NestedMICA on the human portion of the benchmark set (26 out of 52 datasets), then predicted motif instances by picking the highest-scoring PWM hit from each sequence. If other positions had scores within 0.5 bits of the maximum, these were reported too. Given our previous experiences on synthetic datasets, where both MEME and NestedMICA reported motifs other than the target when running on long sequences, we chose to be conservative and made no predictions for sequences of 1000 bases or longer. Other than this, we did not use any expert input or per-dataset adjustments, although these were permitted in the original assessment. Using the web-based evaluation software described in (21), we saw a correlation coefficient (nCC) of 0.149. This compares favourably with the winner of the assessment (both overall and on the human subset), an exhaustive enumeration method called Weeder (22) which scored 0.115. For comparison, the best of the two MEME entries scored 0.034.

The NestedMICA program has been designed to scale to large sets of data. It can run on symmetric multiprocessor machines and clusters if performance becomes an issue. We hope that the improved sensitivity and ability to learn multiple patterns simultaneously will ultimately allow us to extract near-complete sets of regulatory motifs from large amounts of genomic sequence. Searching for large, general, sets of regulatory motifs presents new challenges in evaluating the results. We are encouraged by the recent publication of a large (1367 binding sites for 87 factors) collection of *Drosophila* binding site annotation (23), and believe that this will be a powerful resource for evaluating motif-discovery on a large scale.

We are considering a number of refinements to the method. One direction is to couple the motif-based sequence model with models of other, associated data, such as gene expression patterns. Our use of the ICA framework can help here: models already exist for ICA of microarray gene expression data (24), and it is possible to couple multiple ICA systems together by using a shared mixing matrix.

Another direction, which may prove powerful when analysing large data sets, is to learn rules about the co-occurrence of groups of separate motifs – sometimes called regulatory modules (10). While applications such as STUBB (25) can search sequences for clusters of motifs, and learn about co-occurrence of known motifs,

it seems reasonable to assume that the sensitivity of motif-finding methods could be improved by including co-occurrence in the underlying model. However, the computational cost of adding such an extension to our model would be significant.

An important aspect of NestedMICA is its use of multi-class (mosaic) background models, rather than the single class Markov chains described elsewhere. We find that human genomic sequence can be partitioned into four distinct classes, one of which appears to correspond to the widely reported CpG islands. We are still uncertain about the biological significance of the three remaining classes, but they add an intriguing extra dimension to the genome landscape. We suggest that the previously-reported benefits of using high-order (greater than first-order) Markov chain background models may actually be a reflection of the mosaic structure in the genome rather than a result of real high-order constraints in genomic sequence. If several of the previous bases in a sequence were, for example, purines, then this suggests that the current context might be a purine-rich region and therefore the chance of the next base being a purine is higher than would otherwise be expected. As the Markov chain order increases, the chance of being able to correctly guess the local compositional bias increases.

ACKNOWLEDGEMENTS

We thank Bernard Leong, Remo Sanges, and Jenny Mattison for testing and comments on the NestedMICA software, Casey Bergman for helpful advice on the manuscript, and the Wellcome Trust for support.

REFERENCES

1. Stormo, G. D., Schneider, T. D., Gold, L., and Ehrenfeucht, A. (1982) Use of the 'perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res,* **10**, 2997–3011.

2. Bucher, P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol,* **212**, 563–578.

3. Marsan, L. and Sagot, M. F. (2000) Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J. Comp Biol,* **7**, 345–362.

4. Vilo, J., Brazma, A., Jonassen, I., Robinson, A., and Ukonnen, E. (2000) Mining for putative regulatory elements in the yeast genome using gene expression data. *Proceedings of the Eighth Internation Conference on Intelligent Systems for Molecular Biology,* pp. 384–394.

5. Barash, Y., Elidan, G., Friedman, N., and Kaplan, T. (2003) Modelling dependencies in protein-DNA binding sites. *Proceedings of RECOMB,* pp. 28–37.

6. Schneider, T. D. and Stephens, R. M. (1990) Sequence logos: a new way to display consensus sequence. *Nucleic Acids Res,* **18**, 6097–6100.

7. Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998) Biological Sequence Analysis, Cambridge University Press, .

8. Bailey, T. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology,* pp. 28–36.

9. Thompson, W., Rouchka, E., and Lawrence, C. (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res,* **31**, 358–3585.

10. Arnone, M. I. and Davidson, E. H. (1997) The hardwiring of development: organization and function of genomic regulatory systems. *Development,* **124**, 1851–1864.

11. Comon, P. (1994) Independent Component Analysis: a new concept?. *Signal Processing,* **36**, 287–314.

12. Miskin, J. Ensemble Learning for Independent Component Analysis PhD thesis Astrophysics Group, Cavendish Laboratory, University of Cambridge (2000).

13. MacKay, D. J. C. (2003) Information Theory, Inference, and Learning Algorithms, Cambridge University Press, .

14. Burge, C., Campbell, A. M., and Karlin, S. (1992) Over- and under-representation of short oligonucleotides in DNA Sequences. *Proc Natl Acad Sci,* **89**, 1358–1362.

15. Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P., and Moreau, Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics,* **17**, 1113–1122.

16. Perier, R. C., Praz, V., Junier, T., Bonnard, C., and Bucher, P. (2000) The Eukaryotic Promoter Database (EPD). *Nucleic Acids Res,* **28**, 307–309.

17. Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., L, C., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., M, P., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., I, V., and Clamp, M. (2002) The Ensembl genome database project. *Nucleic Acids Res,* **30**, 30–31.

18. Sandelin, A., Alkema, W., Engstrm, P., Wasserman, W. W., and Lenhard, B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res,* **32**, D91–D94.

19. Wasserman, W. W. and Fickett, J. W. (1998) Identification of Regulatory Regions which Confer Muscle-Specific Gene Expression. *J Mol Biol,* **278**, 167–181.

20. Benos, P. V., Bulyk, M. L., and Stormo, G. D. (2002) Additivity in protein-DNA interactions: how good an approximation is it?. *Nucleic Acids Res,* **30**, 4442–4451.

21. Tompa, M., Li, N., Bailey, T. L., Church, G. M., Moor, B. D., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., vanHelden, J., Vendenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology,* **23**, 137–144.

22. Bavesi, G., Mereghetti, P., Mauri, G., and Pesole, G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res,* **32**, W199–W203.

23. Bergman, C. M., Carlson, J. W., and Celniker, S. E. (2004) Drosophila DNase I footprint database: A systematic genome annotation of transcription factor binding sites in the fruitfly, D. melanogaster. *Bioinformatics,*.

24. Saidi, S. A., Holland, C. M., Kreil, D. P., MacKay, D. J. C., Charnock-Jones, D. S., Print, C. G., and Smith, S. K. (2004) Independent Component Analysis of microarray data in the study of endometrial cancer. *Oncogene,* **23**, 6677–6683.

25. Sinha, S., vanNimwegen, E., and Siggia, E. (2003) A probabilistic method to detect regulatory modules. *Proceedings of the Eleventh Internation Conference on Intelligent Systems for Molecular Biology,* pp. 292–301.

TABLES

Length	100	150	200	300	400	500	600	700
MEME	y	y	n	n	n	n	n	n
N'MICA	y	y	y	y	y	y	y	n

Table 1: Discovery of the HLF motif from sets of 100 synthetic sequences of various lengths. A 'y' indicates that the correct motif was found, 'n' indicates failure.

Length	200	300	400	500	600
MEME	y	y	n	n	n
N'MICA	y	y	y	y	n

Table 2: Discovery of the c-FOS motif from sets of 100 synthetic sequences of various lengths.

Length	800	1000	1200	1400	1600
MEME	y	y	y	n	n
N'MICA	y	y	y	n	n

Table 3: Discovery of the HFH-1 motif from sets of 100 synthetic sequences of various lengths.

Length	100	200	300	400	500	600	800
MEME	y	y	y	y	n	n	n
N'MICA	y	y	y	y	y	y	n
MEME decoy	y	y	n	n	n	n	n
N'MICA decoy	y	y	y	y	y	y	n

Table 4: Discovery of the CREB motif in the presence and absence of a decoy Tal1beta motif

Factor	Random score	MEME score	N'MICA score
MyoD	0.045	0.05	0.31
SRE	0.016	0.21	0.64
CArG	0.014	0.05	0.17
MEF2	0.020	0.44	0.36
M-CAT	0.0093	0.42	0.50

Table 5: ROC scores of best MEME and NestedMICA motifs for binding sites annotated in the muscle regulatory region set. The 'random' column gives the expected score for a factor if predictions were made randomly along the sequences.

FIGURES

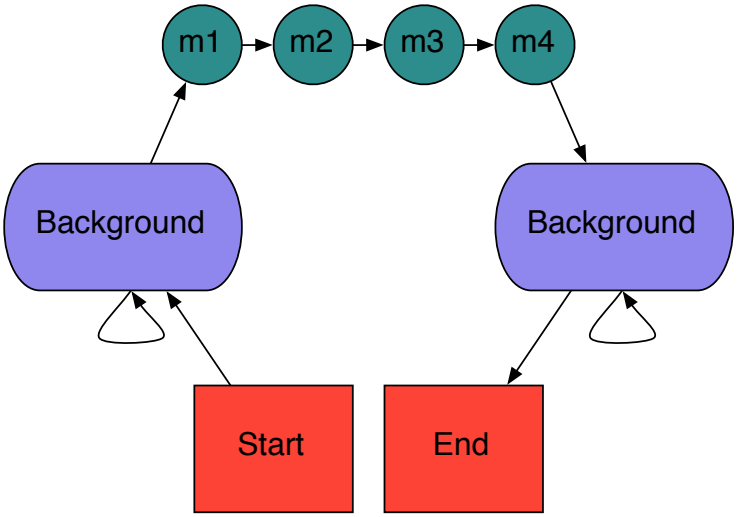

Figure 1: The Zero or One Occurrences per Sequence (ZOOPS) sequence mixture model, represented as a hidden Markov model. The states labelled m1-m4 are responsible for modelling the interesting motif while the other states model the non-interesting remainder of the sequence.

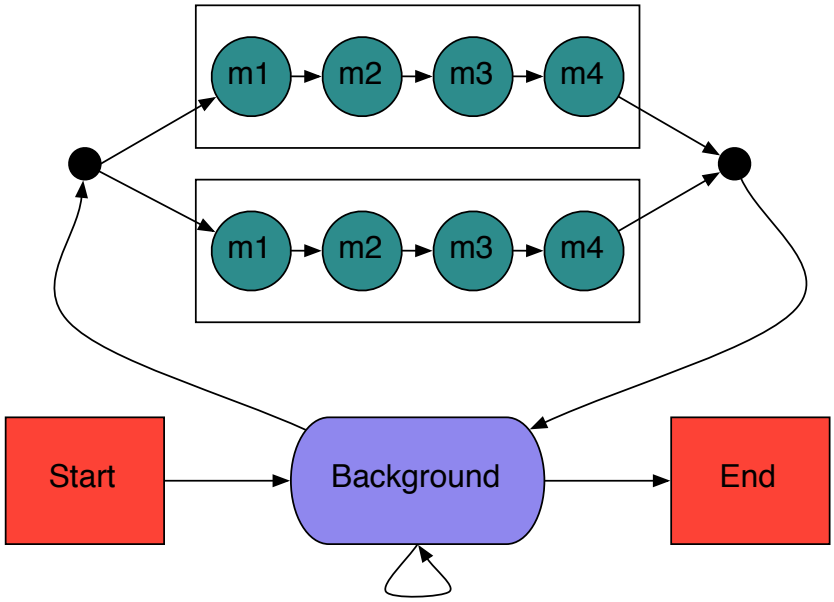

Figure 2: A multiple-uncounted Sequence Mixture Model containing two motifs. The black dots are silent states which are not responsible for modelling any part of the sequence.

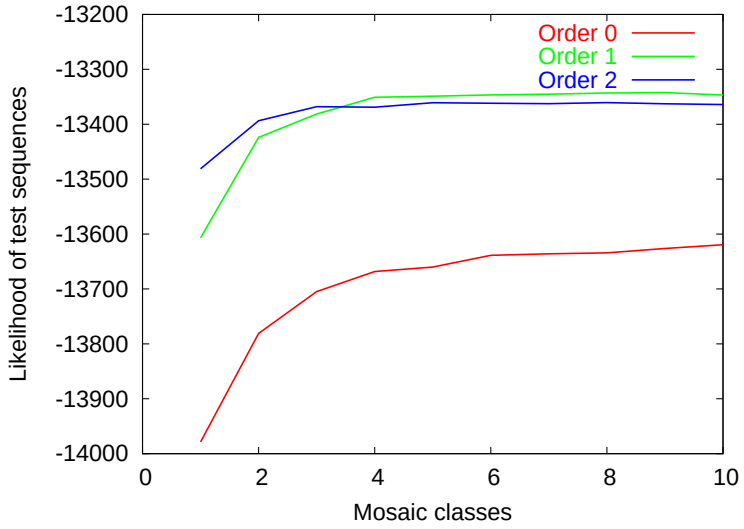

Figure 3: Likelihoods of a set of test sequences given mosaic background models of various orders and class-numbers.

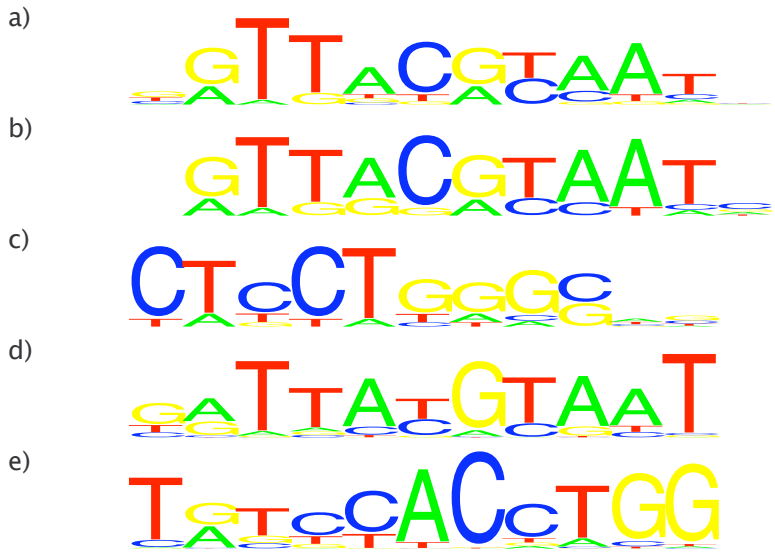

Figure 4: a) The original HLF motif from JASPAR. b) results for searching for HLF in a set of 150 base sequences using MEME. c) MEME with 200 base sequences. d) NestedMICA with 600 base sequences. e) NestedMICA with 700 base sequences.

Figure 5: A selection of mammalian JASPAR weight matrices that used for synthetic data tests.

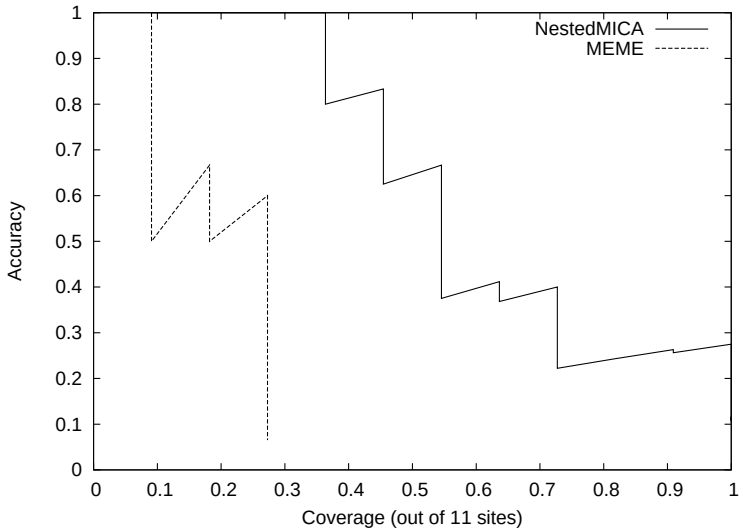

Figure 6: ROC curves for the best matches to the SRE sites in the NestedMICA and MEME results.

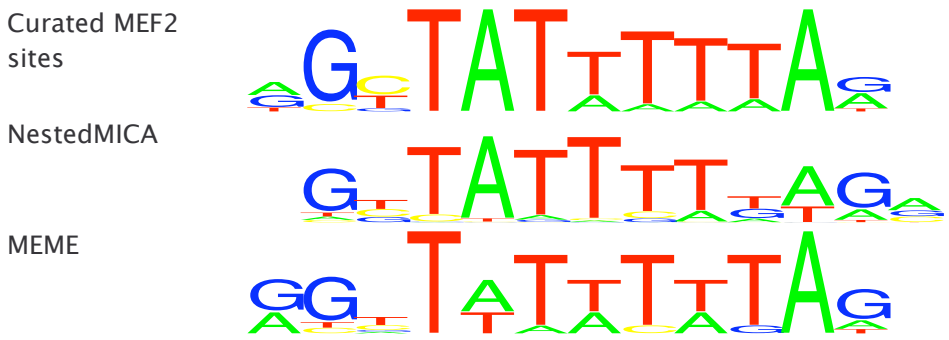

Figure 7: The MEF2 motif derived from curated sites, and the corresponding high-scoring motifs from NestedMICA and MEME.

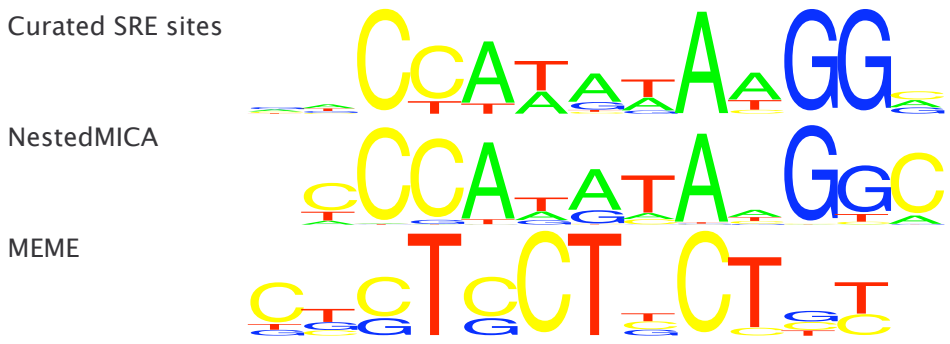

Figure 8: The SRE motif derived from curated sites, and the corresponding high-scoring motifs from NestedMICA and MEME.