BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 1 2013, pages 39–46
doi:10.1093/bioinformatics/bts654

Sequence analysis

Advance Access publication November 9, 2012

DLocalMotif: a discriminative approach for discovering local motifs in protein sequences

Ahmed M. Mehdi[1], Muhammad Shoaib B. Sehgal[2], Bostjan Kobe[1,3,4], Timothy L. Bailey[1] and Mikael Bodén[1,3,*]

[1]Institute for Molecular Bioscience, The University of Queensland, Australia, [2]Microsoft corporation, USA, [3]School of Chemistry and Molecular Biosciences, The University of Queensland, Australia and [4]Infectious Diseases Research Centre, The University of Queensland, Australia

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Local motifs are patterns of DNA or protein sequences that occur within a sequence interval relative to a biologically defined anchor or landmark. Current protein motif discovery methods do not adequately consider such constraints to identify biologically significant motifs that are only weakly over-represented but spatially confined. Using negatives, i.e. sequences known to *not* contain a local motif, can further increase the specificity of their discovery.

Results: This article introduces the method DLocalMotif that makes use of positional information and negative data for local motif discovery in protein sequences. DLocalMotif combines three scoring functions, measuring degrees of motif over-representation, entropy and spatial confinement, specifically designed to discriminatively exploit the availability of negative data. The method is shown to outperform current methods that use only a subset of these motif characteristics. We apply the method to several biological datasets. The analysis of peroxisomal targeting signals uncovers several novel motifs that occur immediately upstream of the dominant peroxisomal targeting signal-1 signal. The analysis of proline-tyrosine nuclear localization signals uncovers multiple novel motifs that overlap with C2H2 zinc finger domains. We also evaluate the method on classical nuclear localization signals and endoplasmic reticulum retention signals and find that DLocalMotif successfully recovers biologically relevant sequence properties.

Availability: http://bioinf.scmb.uq.edu.au/dlocalmotif/

Contact: m.boden@uq.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 7, 2012; revised on October 30, 2012; accepted on November 1, 2012

1 INTRODUCTION

Local motifs are patterns in DNA or protein sequences that occur in a short sequence interval relative to a sequence anchor or landmark. For example, the peroxisomal targeting signal-1 (PTS1; defined by the consensus [SAC][KRH][LA]) occurs at the C-terminus of many proteins that localize to the peroxisome. However, up to 12 residues found upstream of PTS1 are important for localization, but as of yet, no motif is known (Hawkins *et al.*, 2007; Neuberger *et al.*, 2003). Another example is the proline-tyrosine nuclear localization signal (PY-NLS) that is recognized by a specific nuclear import factor Kapβ2 (Lee *et al.*, 2006). The PY-NLS contains a highly conserved PY at the C-terminus of the motif, but residues upstream of this motif are required for the interaction with Kapβ2. Additionally, proteins are retained in the endoplasmic reticulum (ER) owing to the presence of the motif [KH]DEL local to the C-terminus (Elrod-Erickson and Kaiser, 1996). We aim to discover multiple local motifs that co-occur with these anchors.

Existing motif discovery methods typically aim to discover over-represented motifs in DNA and protein sequences (Austin *et al.*, 2007; Bailey *et al.*, 2009; Dogruel *et al.*, 2008; Ettwiller *et al.*, 2007; Linhart *et al.*, 2008; Pavesi *et al.*, 2004; Redhead and Bailey, 2007; Roepcke *et al.*, 2006; Thijs *et al.*, 2002), but do not usually account for positional information and negative data. The few methods that do use sequence distance or position as a feature operate on DNA sequences (Keilwagen *et al.*, 2011; Linhart *et al.*, 2008; Narang *et al.*, 2010; Ohler *et al.*, 2002; Roepcke *et al.*, 2006; Yan *et al.*, 2011) and are thus unsuitable for proteins. Discriminative motif finding methods distinguish functional motifs from randomly occurring sequence patterns by using functionally unrelated 'negative' sequences in which sought motifs are absent (or present owing to chance alone) (Redhead and Bailey, 2007). To our knowledge, none of the available protein motif discovery methods make use of both types of information.

Several local motif discovery methods are designed for inferring motifs that define gene regulatory networks (Ohler *et al.*, 2002; Roepcke *et al.*, 2006; Vardhanabhuti *et al.*, 2007; Xie *et al.*, 2007). Recently, Narang *et al.*, (2010) developed 'LocalMotif' for discovering nucleotide motifs that occur in a short sequence interval relative to transcription start sites. They introduced a novel scoring function to determine the spatial confinement of a DNA motif. Using human promoter data, the authors demonstrated that their method outperformed several other tools such as Amadeus (Linhart *et al.*, 2008), Trawler (Ettwiller *et al.*, 2007), Weeder (Pavesi *et al.*, 2004) and MEME (Bailey *et al.*, 2009) on discovering transcription factor binding motifs in ChIP data. The 'spatial confinement score' (Narang *et al.*, 2010) does not adequately deal with sparse data. With few samples, the spatial confinement score is high at singular counts. To make matters worse, at low numbers of samples under observation, the method

*To whom correspondence should be addressed.

for determining statistical significance is inaccurate (Wilks, 1938). LocalMotif is therefore incapable of distinguishing between real and spurious, low-count motifs.

In this article, we develop a new method inspired by LocalMotif (Narang *et al.*, 2010) that works for the 20-symbol amino acid alphabet enabling *protein* motif discovery. It uses negative data enabling *discriminative* motif discovery, and statistically identifies *local* motifs at realistically low counts. Our new method DLocalMotif discovers motifs in a set of protein sequences that are aligned relative to a defined landmark. We use three scoring functions, namely motif spatial confinement (MSC), motif over-representation (MOR) and motif relative entropy (MRE). To deal with spurious matches, we use pseudo counts for probabilities in the scoring functions. (*Maximum a posteriori* estimates tend to reasonable values when there is little data.) The scoring functions collectively establish whether a motif is enriched in a constrained sequence interval in the positive dataset relative to the negative dataset. To uncover only significant, spatially confined motifs, *P*-values are determined by a (corrected) binomial test of motif location within matched sequences. We believe that DLocalMotif is the only tool for discovering local motifs in protein sequences. We expect that several methods, including Trawler and Amadeus, could be reworked to operate on proteins, but the efficiency and accuracy are yet unknown.

We use synthetic datasets to characterize the accuracy of our method and to compare it with alternative methods. The results indicate that DLocalMotif has superior accuracy on protein sequences largely because of its ability to use positional information and negative data. In addition, DLocalMotif finds the most favourable position of each discovery. We apply our method to several biological datasets and uncover novel motifs that co-occur with a variety of protein localization signals.

2 MATERIAL AND METHODS

2.1 Motif description language

A local motif is a tuple $M=(K, d, R)$ where K is a 'consensus' string of k symbols from the 20-amino acid alphabet A, d is an integer representing the maximum accepted Hamming distance, i.e. the number of mismatches, between the consensus and a 'matched' string (both of length k) and R is a range $[r_1,r_2]$ specifying all accepted starting positions of the motif in a sequence (making it 'local').

Our objective function, $F(M, X)$ is a function of a local motif M and a set of sequences $X \in \{S,U\}$, each of length L. X is divided into positive and negative sequence sets, S and U, respectively, with sizes $N_X = N_S + N_U$. The objective function decomposes into three sub-functions described in the subsections later in the text.

Each sequence is aligned to a universal 'anchor' position. Specifically, we designate $b_1 = 1$ to indicate the first position of the sequence that can contain a match to a motif and $b_2 = L - k + 1$ the final position for a match. $s[i,i+k-1]$ is a k-symbol string of any sequence $s \in X$, starting at i, where $i \in [b_1,b_2]$, i.e. any valid subinterval.

We define $match(s, K, d, i)$ to be true if and only if $H(s[i, i+k-1], K) \leq d$, where $H(.,.)$ is the Hamming distance between two strings. We similarly define $count(s, K, d, [i_1, i_2])$ $|\{i \in [i_1, i_2] | match(s, K, d, i)\}|$ to be the number of instances that match K in the interval $[i_1,i_2]$ of $s \in X$. We define $match(s, K, d, [i_1, i_2])$ to be true if $count(s, K, d, [i_1, i_2]) \geq 1$.

We define $i^* = \text{argmin}_i H(s[i, i+k-1], K)$. When $match(s, K, d, [i_1, i_2])$ is true, we use $match(s, K, d, [i_1, i_2])_j$ to access the jth symbol in the 'matched' string $s[i^*+j-1, i^*+j+k]$ where $j \in \{1, \ldots, k\}$.

2.1.1 Problem formulation The *discriminative* local motif finding problem is an extension of the local motif finding problem (Narang *et al.*, 2010). Suppose that instances of an unknown string K, subject to a user-specified maximum of d mismatches, are enriched within a confined interval R in positive sequences relative to negative sequences. Our goal is to establish the parameters of $M=(K, d, R)$ leveraging the differences between positive and negative instances. Below we describe the functions to objectively score parameter values.

2.1.2 MSC MSC measures a motif's enrichment 'inside' an interval R, relative to any other position in a sequence.

For a given string K (subject to d mismatches), we define the set $S^* = \{s \in S | match(s, K, d, [b_1, b_2])\}$ to be all positive sequences with at least one match. We denote the number of sequences in this set as $N^*_S = |S^*|$.

For the same string, consider for each sequence $s \in S^*$, a Bernoulli trial where we find the string either inside or outside an interval R. The probability of picking an occurrence of K subject to d inside an interval $[i_1,i_2]$ is $P_{[i_1,i_2]}(s) = count(s, K, d, [i_1, i_2])/count(s, K, d, [b_1, b_2])$. We define the set $S' = \{s \in S | P_{[r_1,r_2]}(s)\}$ as the sample of positive sequences whose match is inside $[r_1,r_2]$. Note that counting a sequence as having a 'local' motif is a random event. The success of this event is based on the proportion of matches inside (as opposed to outside) the interval. The expected sequence count is thus the sum of these probabilities $\sum_{s \in S^*} P_{[r_1,r_2]}(s)$. We denote the number of sequences in this sample as $N'_S = |S'|$.

We define $c_1 = (N'_S + z\pi_{\text{MSC}})/(N^*_S + z)$. $z = 1/\pi_{\text{MSC}}$ is a pseudo count, and $\pi_{\text{MSC}} = (r_2 - r_1)/(L - k + 1)$ is the (uniform) prior probability of observing a string within the interval $[r_1,r_2]$.

To qualify c_1 using known negatives, we similarly define the sets U^* and U', on basis of the set U, for sequences with matches anywhere and with local matches, respectively (analogous to S^* and S'). Their counts are referred to as N^*_U and N'_U, respectively. Analogous to c_1, let $c_2 = (N'_U + z\pi_{\text{MSC}})/(N^*_U + z)$.

MSC is defined as the Kullback–Leibler (KL) divergence (D) between c_1 and c_2 (see Equation 1).

$$MSC(M, X) = D(c_1||c_2) = c_1 \log\frac{c_1}{c_2} + (1 - c_1)\log\frac{(1 - c_1)}{(1 - c_2)} \quad (1)$$

Note that, in the absence of a negative dataset, c_2 equals π_{MSC}.

2.1.3 MOR MOR is a statistical measure of the abundance of motif instances in positive sequences relative to a background.

$e_1 = (N^*_S + z\pi_{\text{MOR}})/(N_S + z)$ is the proportion of sequences in S that match K subject to d mismatches at any position. $z = 1/\pi_{\text{MOR}}$ is a pseudo count, and π_{MOR} is the prior probability of finding a match in the sequence, calculated as follows.

Let 0.05^k be the (uniform) prior probability of finding a match at one position in a sequence. Then $\pi_{\text{MOR}} = P(\geq\text{one site}) = 1 - P(\text{zero site}) = 1 - (1 - 0.05^k)^{L-k+1}$. $e_2 = (N^*_U + z\pi_{\text{MOR}})/(N_U + z)$ is the proportion of sequences in U that match the string K subject to d mismatches, at any position.

The MOR of M is measured as the Kullback–Leibler divergence between e_1 and e_2 (Equation 2).

$$MOR(M, X) = D(e_1||e_2) = e_1 \log\frac{e_1}{e_2} + (1 - e_1)\log\frac{1 - e_1}{1 - e_2} \quad (2)$$

Note that, in the absence of negative dataset, e_2 is π_{MOR}.

2.1.4 MRE MRE is a measure of the information-theoretic content of a motif, relative to a background distribution. To capture the functional importance of residues in a motif, we measure MRE using

background frequencies taken from the negative data. We first generate a probability matrix P_M (Equation 3).

$$P_M(a,j) = \frac{n(a,j) + z\pi_{\text{MRE}}}{N'_S + z} \quad (3)$$

$P_M(a,j)$ is the probability of observing a in the jth position of the motif, i.e. $n(a,j) = |\{s \in S' | match(s,K,d,[r_1,r_2])_j = a\}|$. $z = 1/\pi_{MRE}$ is a pseudo count, and $\pi_{MRE} = 0.05$ is the (uniform) prior probability of observing an amino acid $a \in A$.

The MRE is calculated as in Equation 4.

$$MRE(M,X) = \sum_{j=1}^{k} \sum_{a \in A} P_M(a,j) \log \frac{P_M(a,j)}{q_a} \quad (4)$$

where q_a represents the probability of observing an amino acid a in the negative data irrespective of position.

2.1.5 The objective function F(M,X) The objective function $F(M,X)$ incorporates three different characteristics of a motif by a simple geometric combination of the aforementioned scores (see Equation 5).

$$F_{PROD}(M,X) = MSC(M,X)MOR(M,X)MRE(M,X) \quad (5)$$

We also explored an objective function, which sums all three scores as shown in Equation 6.

$$F_{SUM}(M,X) = MSC(M,X) + MOR(M,X) + MRE(M,X) \quad (6)$$

DLocalMotif tries to find M and its interval of occurrence R that maximizes our objective function $F(M, X)$. Note that in the absence of negative data, we resort to a uniform background.

2.1.6 Positional weight matrix For searching in novel and unaligned sequences, we present the discovered motif in the form of a positional weight matrix (PWM). We construct the PWM, W_M as the 'log-odds' of the position-specific probability and a zero-order background probability of the amino acid a at position j as established from matching M against S' (see Equation 7, which refers to Equation 3).

$$W_M(a,j) = \log \frac{P_M(a,j)}{q_a} \quad (7)$$

2.1.7 Statistical significance of motif Narang and colleagues computed P-values for each score individually. The authors used Wilks' theorem, presented the likelihood ratio test statistics and estimated P-values as area under the tail of the χ^2 distribution. However, Wilks' theorem makes inaccurate assumptions for computing likelihood ratio if the numbers of samples are low (which they tend to be for protein sequences). We use a distinct approach to alleviate such concerns and to focus specifically on spatially confined motifs.

For each sequence in $X = \{S,U\}$, we perform a Bernoulli trial as described in Section 2.1.2. We determine the probability of picking a local string K (subject to d mismatches) in S^* and U^*, by c_1 and c_2, respectively (where 'local' means inside $[r_1,r_2]$). We note that $1-c_1$ and $1-c_2$ is the probability of picking a non-local string in S^* and U^*, respectively.

We calculate the cumulative binomial probability as $p_{\text{BIN}} = \sum_{\kappa=N'_S}^{N^*_S} \binom{N^*_S}{\kappa} c_2^{\kappa}(1-c_2)^{N^*_S-\kappa}$. We report the P-value corrected for multiple tests $p = 1-(1-p_{\text{BIN}})^T$ where $T = (L-k+1)(r_2-r_1)N_S$, the total number of motifs evaluated (Chatfield, 1989).

2.2 Search algorithm and implementation

We do not use an exhaustive enumeration strategy because of its computational demands. For a given value of d, and a range of $k \in \mathbb{N}_{\geq 1}$, DLocalMotif uses a greedy enumeration: DLocalMotif finds all non-redundant k-mers occurring in positive sequences in sequence intervals with start positions $R = [r_1,r_2]$ where $r_1 \in \{1, \ldots, L-k+1\}, r_1 < r_2 \leq r_1 + \delta$, where δ is user specified.

Technically, δ represents tolerance to local motif shifts. In the extreme, if no shift is accepted ($\delta = 0$), a sequence profile of the alignment would suffice to identify the motif. At the other extreme, the motif can shift arbitrarily over the sequence ($\delta = L-k+1$), meaning that no guidance is provided by an alignment.

From candidate motifs in different sequence intervals, the method constructs a consensus string and subsequently a PWM, both of which are used to evaluate the objective function. As the candidate motifs are being scored in different position intervals, a list of top m scores is maintained at each position, where m can be set depending on available memory, here we chose $m = 50$ in our default setting.

If two motifs overlap by 25% ($o.p \geq 0.25$ as defined by Equation 8) and share some common instances, DLocalMotif discards the lower scoring motif. Finally, the best motifs (according to the objective function) and their optimal PWMs are reported. Importantly, motifs without statistical support (with corrected $P > 0.001$) are simply discarded.

In Supplementary Section 4.1, we vary the number sequences and sequence length to illustrate how processing time is influenced.

We implemented the DLocalMotif algorithm using the Java programming language. The program is freely available in the form of jar files at: http://bioinf.scmb.uq.edu.au/dlocalmotif/. The user can adjust different parameters to discover motifs and locations, including (i) length of motif (default $k = 4$–11); (ii) maximum number of local motifs to be discovered (default 10); (iii) number of allowable mismatches (default $d = k-3$); and (iv) motif shift (default $\delta = 4$). The algorithm presents discovered motifs and identifies their location, the three individual scores (MOR, MRE and MSC) and the combined score.

2.3 Datasets

2.3.1 Synthetic datasets Consider motif discovery problems falling between two extremes: On the one extreme, sequences are highly enriched with a particular motif, but motifs are not spatially confined. Such problems can be addressed by available motif discovery methods. The other extreme has sequences with only weakly enriched motifs, but when they occur, they are spatially confined in relation to a landmark. We do not expect traditional motif discovery methods to handle such problems well. DLocalMotif is specifically designed to address the latter type of problem.

Inspired by the study of DEME (Redhead and Bailey, 2007), we constructed two datasets that present discovery problems that lay between these extremes. Each dataset contains 50 uniformly generated amino acid sequences each of length L (varied from 30 to 200). We inserted instances of local motifs each with d mutations in t% of sequences (t% varies between 10–100%), in an interval $R = [r_1,r_2]$ relative to each C-terminus. We further generated data according to motifs with $\delta \leq 4$ varied uniformly when applicable. It has been shown previously that MEME and DEME perform equally well with up to three point mutations in planted motifs (Redhead and Bailey, 2007). In the synthetic datasets, we uniformly chose d to have up to a maximum of three point mutations. Other variables (r_1 and t) were also selected uniformly. For each length, we generated 50 datasets. Additional details of synthetic dataset construction are provided in (Redhead and Bailey, 2007).

Negative random problem: Unique local motif instances were implanted in positive data sequences as aforementioned. The negative dataset was generated from a uniform distribution, thus negatives do not contain any useful information. The local motif can thus be identified with or without discrimination.

Decoy motif problem: Positive and negative sequences have one or more local motifs in common. The positive sequences contain one unique local motif. The negatives thus contribute by identifying motifs

that are not unique to the positives, leaving only one motif to be discovered discriminatively.

2.3.2 Biological datasets To evaluate the ability of DLocalMotif to discover local motifs, we studied five biological datasets, assembled using standard data curation practices. Details about each of the datasets can be found in the Supplementary Material.

PTS1: The PTS1 dataset contains known peroxisomal protein sequences with actual peroxisomal targeting signals at their C-termini (positives), and non-peroxisomal proteins with PTS1-like C-termini (negatives). As discussed further in Section 3, several studies have suggested that there are additional, complementary 'signals' upstream to the PTS1, and we expected DLocalMotif to be able to find them.

The initial dataset contained 124 positive sequences and 182 negative sequences identified by Hawkins *et al.* (2007). We updated the dataset with more recent peroxisomal and non-peroxisomal protein sequences in Uniprot, using the same approach as that of Hawkins and colleagues (see Supplementary Material). We extracted 15 residues upstream the PTS1 (or PTS1-like) C-terminus and applied 30% redundancy reduction. The final dataset contained 209 positive and 240 negatives.

ER retention signal: The classical ER retention signal is known to occur at the C-termini of proteins and influences their retention in the ER, possibly in concert with additional signals. We used the C-terminus as an anchor to align sequences and used DLocalMotif to discover retention motifs.

We first filtered 172 proteins (from Uniprot) with evidence of ER retention signals. We then extracted 20 residues upstream the C-terminus to capture additional signals (Qiu *et al.*, 2009). We finally applied 30% redundancy reduction on the filtered sequences. The final dataset contained 130 positive sequences. No negative data were used.

Type-1 copper proteins: The type-1 copper (blue) proteins are involved in electron transport in various systems such as photosynthesis (Giri *et al.*, 2004). These proteins contain a variable-spaced motif with conserved C-terminal glutamine or methionine residues. All positive and negative data for type-1 copper protein were taken from PROSITE database (Sigrist *et al.*, 2010). We aligned all protein sequences relative to the C-terminal residues. The final dataset contains 86 positive sequences and 69 negative sequences.

PY-NLS: The PY-NLS is recognized by the nuclear import factor Kapβ2. Literature reports a poorly defined motif with a highly conserved proline-tyrosine pair PY at the C-terminus of the motif (Lee *et al.*, 2006). We aligned all sequences relative to PY and used DLocalMotif to discover local motifs that co-occur with this anchor.

We first constructed a non-redundant mouse nuclear [NUCPROT; Fink *et al.* (2008)] and non-nuclear (from Uniprot) protein set (both with a maximum sequence redundancy of 30% [Huang *et al.*, 2010)]. We then identified potential PY-NLSs by matching each sequence with defined regular expressions (REs) (Lee *et al.*, 2006). The final dataset contained 297 positive (nuclear proteins that match the REs) and 240 negative sequences (proteins with a known location, which is not nuclear; sequence match the REs).

Bipartite classical nuclear localization signal: The bipartite classical nuclear localization signal (bipartite cNLS) consists of two clusters of basic amino acids, separated by a linker of variable length and composition (Dingwall and Laskey, 1991; Kosugi *et al.*, 2009). We aligned all nuclear localization signals relative to the C-termini and used DLocalMotif to discover complementary local motifs. We expected to at least recover the N and C termini clusters of basic residues.

We first constructed a non-redundant mouse nuclear (NUCPROT). A sequence redundancy of 30% was also applied (Huang *et al.*, 2010). We then identified bipartite cNLSs by matching each sequence with defined REs (Kosugi *et al.*, 2009). The final dataset contained 237 positive sequences (nuclear proteins that match REs). No negative data were used.

2.4 Statistical enrichment analysis of PTS1 motifs

For each discovered motif, we identified a group of proteins that 'have-motif' and a group that 'do-not-have-motif'. We counted the number of proteins in each group, distinguishing between proteins that are assigned a specific property [have a specified Gene Ontology (GO) term or taxonomy term] from those that do not.

The null hypothesis for each motif, and each assigned property, is that the 'have-motif' proteins do not differ in terms of assigned property from those of the 'do-not-have-motif' proteins. Fisher's exact test establishes a *P*-value, the total probability of observing data as extreme or more extreme, given that the null hypothesis is true. From this analysis, we identified terms that have $P \leq 0.05$. The GO terms were retrieved from http://www.geneontology.org (January, 2012). The Taxonomy IDs were retrieved from http://www.uniprot.org (January, 2012).

2.5 Performance metric for synthetic datasets

The synthetic problems discussed in Section 2.5.1 intend to illustrate how well DLocalMotif discovers planted local motifs in protein-like sequences. The top motif in each dataset is used to evaluate the prediction accuracy. Let $I_a = [r_1, r_2 + k]$ represent the *actual* range used for a planted motif M and $I_p = [r'_1, r'_2 + |K'|]$ be the *predicted* range for the top motif with consensus string K'. Similar to Narang and colleagues, we calculated the accuracy of DLocalMotif by measuring the overlap percentage ($o.p$) between the actual (I_a) and predicted (I_p) intervals (see Equation 8).

$$o.p = \frac{I_a \cap I_p}{\max(|I_a|, |I_p|)} \tag{8}$$

We compare the performance of DLocalMotif with the available methods MEME (Bailey *et al.*, 2009), DEME (Redhead and Bailey, 2007) and NestedMICA (Dogruel *et al.*, 2008). All of these methods find over-represented motifs in unaligned sequences, with no regard to spatial arrangements of motifs. In addition, DEME and NestedMICA also consider background data to find motifs that are discriminative.

3 RESULTS

3.1 Evaluating *P*-values on randomly generated data

Motif discovery methods may uncover highly significant motifs even when tested on random datasets (Harbison *et al.*, 2004). To allow the user to distinguish between spurious and biological relevant significant motifs, before assessing the accuracy of discoveries, we set out to illustrate that *P*-values assigned to discovered motifs are statistically meaningful. We thus evaluated *P*-values on random datasets. We generated random protein sequences of different lengths, applied DLocalMotif and extracted the motif with the minimum *P*-value. In particular, we varied the length of sequences from 50 to 100 with each experiment repeated 200 times. The results are shown as Q–Q plots, i.e a plot between calculated and ranked *P*-values (Supplementary Fig. S1). Owing to observed *P*-value versus rank spread, we show that our method neither over- nor underestimates statistical significance.

3.2 Evaluating DLocalMotif on synthetic datasets

We investigated the performance of DLocalMotif discovering motifs in synthetic datasets containing randomly placed local motifs. In particular, we studied the effects of varying the length L of sequences. For comparison, we considered MEME (Bailey *et al.*, 2009), DEME (Redhead and Bailey, 2007) and

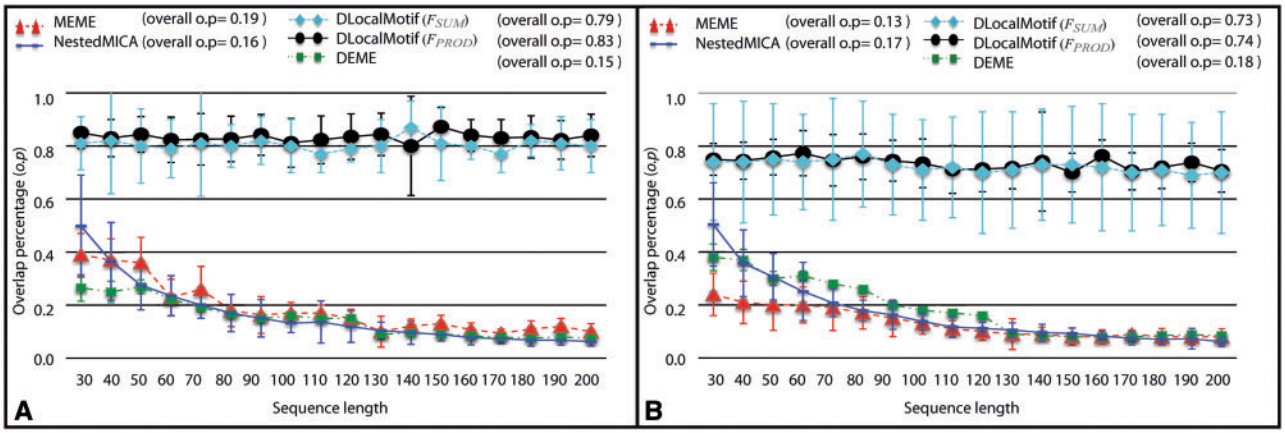

Fig. 1. Comparison of DLocalMotif, DEME (Redhead and Bailey, 2007), MEME (Bailey *et al.*, 2009) and NestedMICA (Dogruel *et al.*, 2008) on (**A**) 'negative random' and (**B**) 'decoy' motif problems. Each plot shows the accuracy of discovered motifs as overlap percentage (*o.p*). Each data point represents the average (±standard deviation) *o.p* on 50 independent runs. In all cases, DLocalMotif, DEME and NestedMICA are using positive and negative data, whereas MEME is using positive data only

NestedMICA (Dogruel *et al.*, 2008). The performance of MEME was based on positive data only but DEME, NestedMICA and DLocalMotif were tested on positive and negative data. In Figure 1, each data point represents the average *o.p* from 50 datasets. The overall accuracy was calculated by averaging the *o.p* for all data sequences. For DLocalMotif, the accuracy using both objective functions are reported (Equation 5† and Equation 6‡).

For the negative random datasets, DLocalMotif achieved higher accuracy than standard algorithms with average *o.p* of 0.83† and 0.79‡ as compared with 0.15 (DEME), 0.19 (MEME) and 0.16 (NestedMICA) (see Fig. 1). MEME either outperformed DEME or performed equally well, as DEME in the test that has no additional information in the negative data. NestedMICA outperformed MEME and DEME for short sequences (30 residues).

For the decoy motif datasets, DLocalMotif outperformed MEME, DEME and NestedMICA in terms of average overlap percentage with an average *o.p* of 0.74† and 0.73‡, compared with 0.18 (DEME), 0.13 (MEME) and 0.17 (NestedMICA). DEME and NestedMICA consider negative data and are thus able to identify decoys. The results further illustrate the ability of DLocalMotif to discover local motifs that discriminate between positive and negative sequences, by identifying the motifs that are only available in the positive data. Note that DEME, MEME and NestedMICA perform relatively well with shorter length sequences. However, the overall accuracy of standard algorithms decreases significantly with the increase in the sequence length. DLocalMotif's ability to discover local motifs does not change with sequence length.

It needs to be emphasized that most other protein motif discovery tools, including MEME, DEME and NestedMICA, are not designed for local motif discovery. We also believe that the accuracy of NestedMICA may be improved by optimizing its setting to each scenario, as is illustrated in previous studies (Dogruel *et al.*, 2008). As much as possible, we used default parameter settings, to illustrate the typical behaviour of the method. We do not claim that standard motif discovery algorithms are inaccurate, rather point out that they are not designed for discovering 'local motifs' in protein sequences. When such motifs are discovered, they are discovered to an extent that is well below that of DLocalMotif. We have shown that DLocalMotif effectively recovers spatially confined motifs. F_{PROD} is slightly superior to F_{SUM} as objective function, but both perform well above the baseline provided by MEME, DEME and NestedMICA. In the following, we will use the former objective exclusively.

To check the accuracy of DLocalMotif with increasing number of background sequences, we planted a decoy motif of length 8. The total number of residues in each sequence (L) was set to 100. The number of sequences in the positive dataset was fixed to 100, whereas the number of sequences in the negative dataset was varied from 0 to 150 and each experiment was repeated 15 times. The results are shown in Supplementary Figure S5. The results indicate that DLocalMotif is highly accurate when the number of background sequences is $\geq$70% of the number of foreground sequences. This illustrates the tool's ability to use negative data.

To investigate how the accuracy of DLocalMotif varies with extent of motif shift, we performed additional tests on the decoy motif data. We varied $\delta \in \{6,8,10,12\}$ and ran 15 tests for each sequence length (see Supplementary Fig. S6). DLocalMotif is less accurate for shorter sequences (30 and 40 residues) when $\delta = 12$. The accuracy is highest when δ is small and close to the size of the interval used to plant motifs. A greater δ (ultimately the length of the sequence) leads to similar performance as standard non-local motif discovery methods (Supplementary Fig. S6).

3.3 Evaluating DLocalMotif on biological datasets

We collated five biological datasets. For the PTS1 and PY-NLS datasets, we aimed to discover novel local motifs not found by existing methods. Using the ER retention, bipartite cNLS and type-1 copper protein datasets, we aimed to evaluate DLocalMotif's ability to recover known local motifs.

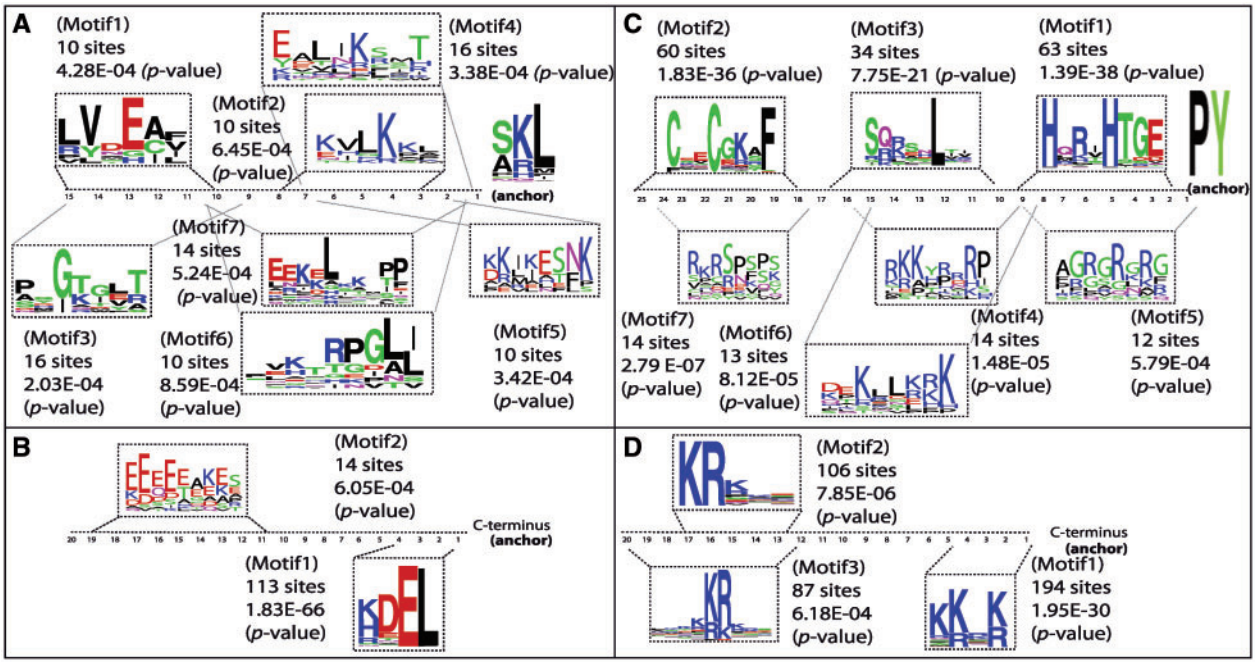

Fig. 2. Discovered local motifs in (**A**) PTS1 dataset (**B**) ER retention signal dataset (**C**) PY-NLS dataset and (**D**) bipartite cNLS dataset. The *P*-values for each score are also shown. The discovered motifs are numbered according to their overall rank based on their combined score. The *x*-axis represents distance relative to anchor. The logos are generated using WebLogo (Crooks *et al*., 2004)

All sequences that contain a match with each discovered motif were used to generate sequence logos (Crooks *et al*., 2004).

3.3.1 Discovering motifs occurring with PTS1 Figure 2A summarizes the logos of local motifs and their position relative to the anchor, as discovered in the PTS1 dataset by DLocalMotif. We found seven motifs that co-occur with PTS1, here named Motif1-7. We validated the position of discovered motifs using literature.

There is a high prevalence of hydrophobic residues at 5–11 upstream of the C-terminus in Motif1, Motif2, Motif6 and Motif7, in agreement with functionally relevant observations (Neuberger *et al*., 2003). It has been shown in the literature that in *Candida boidinii*, basic residues are found upstream of PTS1 (Mullen and Trelease, 2000), as observed in Motif2, Motif4 and Motif5. Neuberger and colleagues also observed basic residues at 1–7 upstream of PTS1. In Motif4 and Motif7, threonine is prevalent at one and two residues, respectively (relative to the anchor), which matches with observation of Neuberger and colleagues.

The MOR of all discovered motifs shown in Figure 2A was low, whereas the MSC was high. By allowing small variation in interval length, DLocalMotif was able to discover motifs that are unavailable to standard motif discovery tools. To investigate whether motifs are biologically meaningful and (if so) perform a defined function, we evaluated the statistical enrichment of the functions of proteins containing the instances of discovered motifs.

Supplementary Tables S2–8 show the statistical enrichment of GO and taxonomical terms of proteins in different groups. We generated each group by filtering proteins that contain discovered motifs. Each motif co-occurred with PTS1 independently of other motifs. Proteins with Motif3, Motif4, Motif6 and Motif7 are enriched with plant (*peroxidase activity, Liliopsida, Arabidopsis thali ana*) and fly (*Drosophila melanogaster*)-related terms, indicating that they are prevalent in these species. In contrast, Motif1 contains non-plants terms (*D-amino-acid oxidase activity, Cetartiodactyla*). Motif6 is also prevalent in flowering plants (*Poaceae*). We note that Motif2 and Motif5 occur in proteins involved in assimilation of acetyl co-enzyme A (acetyl-CoA), an essential process in many bacteria that proceed via the ethylmalonyl-CoA pathway (Erb *et al*., 2010).

When run on the same dataset, MEME, DEME and NestedMICA discover either one or at best two motifs similar to Motif1 and Motif3 (see Supplementary Fig. S8).

3.3.2 Recovering known motifs in ER and copper proteins Using Uniprot annotations to validate, DLocalMotif successfully recovered 113 of 130 classical ER retention signals. Using PROSITE annotations, 99 of 113 known instances of the ER retention signals were recovered (Fig. 2B).

There are a few examples in literature where more than one ER retention signal is present upstream of the C-terminus (Qiu *et al*., 2009). Interestingly, DLocalMotif finds one motif upstream classical ER retention motif (see Fig. 2B; Motif2). The novel motif contains cluster of acidic residues and occurs at a distance 19 residues upstream the ER retention motif. Indeed, the literature suggests that proteins that are efficiently retained in the ER are often distinguished by the presence of acidic amino acid residues at that location (Munro and Pelham, 1987; Yun and Eipper, 1995). Rose-John *et al* claimed that an acidic residue signal (EEDDD) that occurs 14 residues

upstream the C-terminal of IL-6-PDI contributes to the efficiency of ER retention (Rose-John *et al*., 1993).

We also tried MEME, DEME and NestedMICA on the same data: The classical ER retention signal was recovered in each case, but additional motifs were not found (see Supplementary Fig. S8).

When tested on type-1 copper proteins, again taken from PROSITE, DLocalMotif recovered 72 of 86 copper ligand sites. We show a recovered known motif in Supplementary Figure S7.

3.3.3 Discovering motifs occurring with PY-NLSs PY-NLSs are recognized by the transport factor Kapβ2. We used DLocalMotif to investigate the existence of local motifs that supplement the PY anchor that appear in almost all Kapβ2 cargo. DLocalMotif discovered several novel motifs that figured strongly upstream of the anchor (see Fig. 2C).

Motif1, Motif2 and Motif3 occur at a distance 9, 25 and 17 residues upstream the PY anchor. We found that these three motifs correspond to zinc finger (Zf) motifs, and a manual analysis using Pfam (Finn *et al*., 2010) suggested that they belong to the C2H2 class of Zfs. Literature evidence also suggested that Zf domains can efficiently act as NLSs and are recognized by karyopherins (Lee *et al*., 2000; Saijou *et al*., 2007; Yamasaki *et al*., 2005). We also searched the literature to find evidence of Kapβ2 interacting with Zf domains (specifically the C2H2 class). We found that ADR1, which contains C2H2 Zf domains, interacts with Kap104, the Kapβ2 ortholog in yeast (Stark *et al*., 2006). It is not known whether C2H2 domains are necessary or sufficient for Kap104 binding.

Motif4 and Motif6 contain clusters of basic amino acids and are found to be prevalent in the proteins that contain basic PY-NLSs. In contrast, Motif5 and Motif7 contain many hydrophobic amino acids, and our manual analysis revealed that they are prevalent in proteins that contain hydrophobic PY-NLSs. (Two more motifs were found but are not shown owing to space limitations).

One or at best two fragments of the Zf domain were detected by MEME, DEME and NestedMICA (see Supplementary Fig. S8).

3.3.4 Discovering motifs within bipartite cNLSs Here, we consider the C-terminus of the bipartite cNLS motif as anchor to improve our understanding of the variable-length linker-region and the basic N- and C-termini clusters (Dingwall and Laskey, 1991; Kosugi *et al*., 2009). Studies have indicated that the linker region contributes to nuclear localization activity (Engelmann *et al*., 1996), but so far, specific motifs have not been identified. We thus used DLocalMotif to discover motifs relevant to nuclear import.

DLocalMotif discovered three motifs (Motif1 at C-terminus Motif2 and Motif3 at N-terminus; see Fig. 2D). Motif2 and Motif3 co-occur with Motif1 containing clusters of basic residues (see Fig. 2D). Motif1 is a purely basic residue motif. Motif2 and Motif3 are 5 and 9 residue long motifs, respectively, and also contain basic residues.

We also ran MEME (Bailey *et al*., 2009), DEME (Redhead and Bailey, 2007) and NestedMICA (Dogruel *et al*., 2008) on the same dataset. These methods were able to uncover only one motif consisting of basic residues. The spatial confinement of the C-terminal motif is high, making both motifs easy targets for DLocalMotif.

4 CONCLUSION

In this article, we address the motif discovery problem when motifs are only weakly enriched overall, but biological expertise suggests that they are confined to an approximate, but defined position. For example, structural constraints of protein conformation make fragments distant in sequence come together in space. DLocalMotif discovers such 'local motifs' in a set of protein sequences that are aligned to a predefined anchor, and their appearance is linked to their position within the alignment. Unlike similar current methods, DLocalMotif is specifically designed for proteins, and to solve problems where negative data are available.

To evaluate the performance of the proposed method, we investigated a series of protein translocation problems where targeting signals are assisted by additional, often spatially related, but otherwise more subtle properties. To enable DLocalMotif to adequately deal with sparse data, we re-designed the scoring functions of Narang *et al*. by introducing pseudo counts. We formulated three discriminative scoring features, MSC, MOR and MRE. These features establish whether a motif is positioned in a sequence interval in positive data and is generally absent in negative data. The new formulation gives a quantitative evaluation of a motif's relevance, considering its over-representation, relative entropy and spatial confinement. Importantly, our search strategy removed all motifs with non-significant spatial confinement *P*-values determined using a robust binomial test of motif location.

Although DLocalMotif has many parameters that can be tuned, we have shown that default parameters settings are effective for discovering biologically significant motifs. To examine the performance of DLocalMotif, we planted *random negative* and *decoy* motifs in artificial datasets. The results underscored that DLocalMotif is able to accurately discover the location of a planted motif's occurrence, independently of sequence length. The results also demonstrated that DLocalMotif will outperform standard motif discovery algorithms, here represented by MEME (Bailey *et al*., 2009), DEME (Redhead and Bailey, 2007) and NestedMICA (Dogruel *et al*., 2008) when motifs are spatially confined. It is important to note, however, that standard motif discovery algorithms are not expected to discover local motifs any better than non-local motifs, and their performance thus degrades with the increase in sequence length.

On biological data with limited over-representation of motifs, DLocalMotif discovered multiple local motifs. We present seven novel PTS1 local motifs, some of which appear to be species-distinct. DLocalMotif discovered three entirely novel PY-NLS local motifs that overlap with C2H2 Zf domains, associated with nuclear trafficking. We believe these motifs may further our understanding of PY-NLS-mediated translocation.

DLocalMotif successfully recovered ER retention motifs and the bipartite NLS, despite the absence of negative data. Specifically in ER retention data, we found a motif consisting of acidic residues that occurs immediately upstream the classical ER retention signal. Literature indicates that the same motif may

contribute to the efficiency of ER retention. With many motif discovery tools unable to deal with large motifs with variable linker regions, DLocalMotif offers a compromise by detecting multiple smaller but spatially interlinked motifs.

Conflict of Interest: none declared.

REFERENCES

Austin,R.S. *et al.* (2007) C-terminal motif prediction in eukaryotic proteomes using comparative genomics and statistical over-representation across protein families. *BMC Genomics*, **8**, 191.

Bailey,T.L. *et al.* (2009) MEME suite: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.

Chatfield,C. (1989) *Statistics for Technology: a Course in Applied Statistics*. 3rd edn. Chapman and Hall, London/New York, 1983.

Crooks,G.E. *et al.* (2004) Weblogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

Dingwall,C. and Laskey,R.A. (1991) Nuclear targeting sequences–a consensus? *Trends Biochem. Sci.*, **16**, 478–481.

Dogruel,M. *et al.* (2008) NestedMICA as an ab initio protein motif discovery tool. *BMC Bioinformatics*, **9**, 19.

Elrod-Erickson,M.J. and Kaiser,C.A. (1996) Genes that control the fidelity of endoplasmic reticulum to golgi transport identified as suppressors of vesicle budding mutations. *Mol. Biol. Cell.*, **7**, 1043–1058.

Engelmann,J. *et al.* (1996) Early stage monitoring of miltefosine induced apoptosis in KB cells by multinuclear NMR spectroscopy. *Anticancer Res.*, **16**, 1429–1439.

Erb,T.J. *et al.* (2010) The apparent malate synthase activity of rhodobacter sphaeroides is due to two paralogous enzymes, (3s)-malyl-coenzyme a (coa)/beta-methylmalyl-coa lyase and (3s)- malyl-coa thioesterase. *J. Bacteriol.*, **192**, 1249–1258.

Ettwiller,L. *et al.* (2007) Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nat. Methods*, **4**, 563–565.

Fink,J.L. *et al.* (2008) Towards defining the nuclear proteome. *Genome Biol.*, **9**, R15.1–R15.8.

Finn,R.D. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.

Giri,A.V. *et al.* (2004) Functionally specified protein signatures distinctive for each of the different blue copper proteins. *BMC Bioinformatics*, **5**, 127.

Harbison,C.T. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Hawkins,J. *et al.* (2007) Identifying novel peroxisomal proteins. *Proteins*, **69**, 606–616.

Huang,Y. *et al.* (2010) CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.

Keilwagen,J. *et al.* (2011) De-novo discovery of differentially abundant transcription factor binding sites including their positional preference. *PLoS Comput. Biol.*, **7**, e1001070.

Kosugi,S. *et al.* (2009) Six classes of nuclear localization signals specific to different binding grooves of importin α. *J. Biol. Chem.*, **284**, 478–485.

Lee,B.J. *et al.* (2006) Rules for nuclear localization sequence recognition by karyopherin beta 2. *Cell*, **126**, 543–558.

Lee,J.Y. *et al.* (2000) Characterization of a zinc finger protein ZAN75: nuclear localization signal, transcriptional activator activity, and expression during neuronal differentiation of P19 cells. *DNA Cell Biol.*, **19**, 227–234.

Linhart,C. *et al.* (2008) Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. *Genome Res.*, **18**, 1180–1189.

Mullen,R.T. and Trelease,R.N. (2000) The sorting signals for peroxisomal membrane-bound ascorbate peroxidase are within its C-terminal tail. *J. Biol. Chem.*, **275**, 16337–16344.

Munro,S. and Pelham,H.R. (1987) A c-terminal signal prevents secretion of luminal er proteins. *Cell*, **48**, 899–907.

Narang,V. *et al.* (2010) Localized motif discovery in gene regulatory sequences. *Bioinformatics*, **26**, 1152–1159.

Neuberger,G. *et al.* (2003) Motif refinement of the peroxisomal targeting signal 1 and evaluation of taxon-specific differences. *J. Mol. Biol.*, **328**, 567–579.

Ohler,U. *et al.* (2002) Computational analysis of core promoters in the Drosophila genome. *Genome Biol.*, **3**, 1–8.

Pavesi,G. *et al.* (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.*, **32**, W199–W203.

Qiu,S. *et al.* (2009) An endoplasmic reticulum retention signal located in the extracellular amino-terminal domain of the NR2A subunit of N-Methyl-D-aspartate receptors. *J. Biol. Chem.*, **284**, 20285–20298.

Redhead,E. and Bailey,T.L. (2007) Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. *BMC Bioinformatics*, **8**, 385.

Roepcke,S. *et al.* (2006) Identification of highly specific localized sequence motifs in human ribosomal protein gene promoters. *Gene*, **365**, 48–56.

Rose-John,S. *et al.* (1993) Intracellular retention of interleukin-6 abrogates signaling. *J. Biol. Chem.*, **268**, 22084–22091.

Saijou,E. *et al.* (2007) Nucleocytoplasmic shuttling of the zinc finger protein EZI is mediated by importin-7-dependent nuclear import and CRM1-independent export mechanisms. *J. Biol. Chem*, **282**, 32327–32337.

Sigrist,C.J.A. *et al.* (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.*, **38**, D161–D166.

Stark,C. *et al.* (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.

Thijs,G. *et al.* (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol.*, **9**, 447–464.

Vardhanabhuti,S. *et al.* (2007) Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. *Nucleic Acids Res.*, **35**, 3203–3213.

Wilks,S.S. (1938) A the large-sample distribution of the likelihood ratio for testing composite hypotheses. *Proc. Natl Acad. Sci. USA*, **1**, 60–62.

Xie,X. *et al.* (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc. Natl Acad. Sci. USA*, **104**, 7145–7150.

Yamasaki,H. *et al.* (2005) Zinc finger domain of Snail functions as a nuclear localization signal for importin β-mediated nuclear import pathway. *Genes Cells*, **10**, 455–464.

Yan,R. *et al.* (2011) A tree-based approach for motif discovery and sequence classification. *Bioinformatics*, **27**, 2054–2061.

Yun,H.Y. and Eipper,B.A. (1995) Addition of an endoplasmic reticulum retention/retrieval signal does not block maturation of enzymatically active peptidylglycine alpha-amidating monooxygenase. *J. Biol. Chem.*, **270**, 15412–15416.