DLocalMotif: A discriminative approach for discovering local motifs in protein sequences
Ahmed M. Mehdi, Muhammad Shoaib B. Sehgal, Bostjan Kobe, Timothy L. Bailey, and Mikael Bodén

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

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

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 "negative sequences" known to not contain a local motif can further increase the specificity of their discovery.

Results: This paper 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 data sets. The analysis of peroxisomal targeting signals uncovers several novel motifs that occur immediately upstream of the dominant PTS1 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/

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 twelve residues found upstream of PTS1 are important for localization but as yet no motif is known (Neuberger et al., 2003; Hawkins et al., 2007). Another example is the proline-tyrosine nuclear localization signal (PY-NLS) that is recognized by a specific nuclear import factor Kap2/2 (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 Kap2/2. Additionally, proteins are retained in the endoplasmic reticulum (ER) due to the presence of the motif [KH]DEL local to the C-terminus (Austin et al., 2007). 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 (Redhead and Bailey, 2007; Bailey et al., 2009; Linhart et al., 2008; Pavesi et al., 2004; Austin et al., 2007; Ettwiller et al., 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 (Narang et al., 2010; Linhart et al., 2008; Roepcke et al., 2006; Ohler et al., 2002; Yan et al., 2011; Keilwagen 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 due 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 (Roepcke et al., 2006; Vardhanabhuti et al., 2007; Xie et al., 2007; Ohler et al., 2002). 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" (SCS) (Narang et al., 2010) does not adequately deal with sparse data. For example, at one occurrence of motif instance, the SCS remains significantly high and therefore incapable of distinguishing between real and spurious, low-count motifs. To make matters worse, at a low number of samples under observation, the method for determining statistical significance is inaccurate (Wilks, 1938).

In this paper, we develop a method inspired by LocalMotif (Narang et al., 2010) that works for the 20-symbol amino acid alphabet enabling protein motif discovery, that utilises negative
data enabling discriminative motif discovery, and that 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 utilise 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 if a motif is enriched in a constrained sequence interval in the positive data set relative to the negative data set. To uncover only significant, spatially confined motifs, p-values are determined by a (corrected) binomial test of motif location within matched sequences.

We use synthetic data sets 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 favorable position of each discovery. We apply our method to several biological data sets 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, \( R \) is the number of mismatches, between the consensus and a “matched” string (both of length \( k \)) and \( R \) is a range \([r_1, r_2]\). Each sequence is aligned to a universal “anchor” position. Specifically, \( s[i, i + k - 1] \) 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 \( \text{match}(s, K, d, i) \) to be true if and only if \( H(s[i, i + k - 1], K) \leq d \), where \( H(\cdot, \cdot) \) is the Hamming distance between two strings. We similarly define \( \text{count}(s, K, d, [i_1, i_2]) = \sum_{i \in [i_1, i_2]} \text{match}(s, K, d, i) \) to be the number of instances that match \( K \) in the interval \([i_1, i_2]\) of \( s \in X \). We define \( \text{match}(s, K, d, [i_1, i_2]) \) to be true if \( \text{count}(s, K, d, [i_1, i_2]) \geq 1 \).

We define \( i^* = \text{argmin } H(s[i, i + k - 1], K) \). When \( \text{match}(s, K, d, [i_1, i_2]) \) is true, we use \( \text{match}(s, K, d, [i_1, i_2])_j \) to access the \( j \)-th 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 Motif spatial confinement (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 \( S^+ = \{ s \in S \mid \text{match}(s, K, d, [b_1, b_2]) \} \) 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 \( P_{[i_1, i_2]}(s) = \text{count}(s, K, d, [i_1, i_2]) / \text{count}(s, K, d, [b_1, b_2]) \). We define the set \( S' = \{ s \in S \mid P_{[r_1, r_2]}(s) \} \) 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 \( \sum_{s \in S'} P_{[r_1, r_2]}(s) \). The number of sequences in this sample as \( N_S' = |S'| \). We define \( c_1 = (N_S' + z \pi_{\text{MRE}}) / (N_S' + z) \), \( z = 1 / \pi_{\text{MRE}} \) is a pseudo count, and \( \pi_{\text{MRE}} = (r_2 - r_1) / (L - k + 1) \) 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, and as above. Their counts are referred to as \( N_{U^*} \) and \( N_{U'} \), respectively. Analogous to \( c_1 \), let \( c_2 = (N_{U^*} + z \pi_{\text{MRE}}) / (N_{U^*} + z) \).

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

\[
\text{MSC}(M, X) = D(c_1 || c_2) = c_1 \log \frac{c_1}{c_2} + (1 - c_1) \log \frac{1 - c_1}{1 - c_2} \tag{1}
\]

Note that, in the absence of a negative data set, \( c_2 \) equals \( \pi_{\text{MRE}} \).

2.1.3 Motif over-representation (MOR)

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

\( \pi_{\text{MOR}} = P(\text{one site}) = 1 - P(\text{zero site}) = 1 - (1 - \pi_{\text{MRE}} )^{L - k + 1} \)

\( c_2 = (N_{U^*} + z \pi_{\text{MRE}}) / (N_{U^*} + z) \) is the proportion of sequences in \( U \) that match the string \( K \) subject to \( d \) mismatches at any position. \( z = 1 / \pi_{\text{MRE}} \) is a pseudo count and \( \pi_{\text{MRE}} \) is the prior probability of finding a match in the sequence, calculated as follows.

Let 0.05 \( b \) be the (uniform) prior probability of finding a match at one position in a sequence. Then \( \pi_{\text{MRE}} = P(\text{zero site}) = 1 - P(\text{zero site}) = 1 - (1 - 0.05)^{L - k + 1} \) and \( c_2 \) is defined as the Kullback-Leibler (KL) divergence (\( D \)) between \( c_1 \) and \( c_2 \). (Equation 2).

\[
\text{MOR}(M, X) = D(c_1 || c_2) = c_1 \log \frac{c_1}{c_2} + (1 - c_1) \log \frac{1 - c_1}{1 - c_2} \tag{2}
\]

Note that, in the absence of negative data set, \( c_2 \) equals \( \pi_{\text{MRE}} \).

2.1.4 Motif relative entropy (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 \( P_M(a, j) \) (Equation 3).

\[
P_M(a, j) = \frac{n(a, j) + z \pi_{\text{MRE}}}{N_S' + z} \tag{3}
\]

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

The MRE is calculated as in Equation 4.

\[
\text{MRE}(M, X) = \frac{1}{k} \sum_{j=1}^{k} \sum_{a \in A} P_M(a, j) \log \frac{P_M(a, j)}{q_a} \tag{4}
\]
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 $F(M,X)$ incorporates three different characteristics of a motif by a simple geometric combination of the aforementioned scores (see Equation 5).

$$F(M, X) = MSC(M, X)MOR(M, X)MRE(M, X)$$  \hspace{1cm} (5)

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 $S^*_r$ (see Equation 6 which refers to Equation 3).

$$W_M(a,j) = \log \frac{P_M(a,j)}{q_a}$$  \hspace{1cm} (6)

2.1.7 Statistical significance of motif Narang and colleagues computed $p$-values for each score individually. The authors used Wilks' theorem, and presented the likelihood ratio (LR) test statistics and finally estimated $p$-values as area under the tail of the $\chi^2$ distribution. However, Wilks' theorem makes inaccurate assumptions for computing LR if the numerator of the $\chi^2$ term is large. 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 $p_{NS} = \sum_{n=0}^{\lfloor N_S \rfloor} \binom{N_S}{n} \left(1 - r_2 \right)^{N_S - n}$. We report the $p$-value corrected for multiple tests $p = 1 - (1 - p_{NS})^{2T}$ 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

For a given value of $d$, and a range of $k \in \mathbb{N}$, DLocalMotif uses greedy search to find non-redundant $k$-mers occurring in positive sequences in sequence intervals with start positions $R = [r_1, r_2]$ where $r_1 \in \{1, \ldots, L - k + 1\}$, $r_2 \leq r_1 + \delta$, where $\delta$ is user specified.

Technically, $\delta$ 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. If two motifs occur in the same sequence interval and in the same sequences, 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.

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 4 to 11); (ii) maximum number of local motifs to be discovered (default 10); (iii) number of allowable mismatches (default $d = k - 3$); (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 Data sets

2.3.1 Synthetic data sets 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 data sets that present discovery problems that lay between these extremes. Each data set 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 $\%$ of sequences, in an interval $R = [r_1, r_2]$ relative to each C-terminus. We further generate data according to motifs with $\delta \leq 4$ and $0 \leq d \leq 3$, varied uniformly when applicable. Other variables ($\tau_1$ and $\tau_2$) were also selected uniformly. For each length, we generated 50 data sets. Additional details of synthetic data set construction are provided in (Redhead and Bailey, 2007).

2.3.2 Biological data sets To evaluate the ability of DLocalMotif to discover local motifs, we studied four biological data sets, assembled using standard data curation practices: A peroxisomal targeting signal data set, an ER retention signal data set, a proline-tyrosine nuclear localization signal data set and a bipartite nuclear localization signal data set. Details about each of the data sets can be found in the Supplementary Material.

Peroxisomal targeting signal (PTS1): The PTS1 data set 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 Results, 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 data set contained 124 positive sequences and 182 negative sequences identified by Hawkins and colleagues. We updated the data set with more recent peroxisomal and non-peroxisomal protein sequences in Uniprot, using the same approach as that of Hawkins et al. (see Supplementary Material). We extracted 15 residues upstream the PTS1 (or PTS1-like) C-terminus and applied 80% redundancy reduction. The final data set contained 209 positive and 240 negatives.

Endoplasmic reticulum (ER) retention signal: The classical ER retention signal is known to occur at the C-termini of proteins and influences their retention in the endoplasmic reticulum, 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 (Qu et al., 2009). We finally applied 80% redundancy reduction on the filtered sequences. The final data set contained 132 positive sequences. No negative data was used.

Proline-tyrosine nuclear localization signal (PY-NLS): The proline-tyrosine nuclear localization signal is recognized by the nuclear import factor Kap32. 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 data set 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 (bipartite cNLS):** The bi-partite cNLS consists of two clusters of basic amino acids, separated by a linker of variable length and composition (Kosugi et al., 2009; Marfori et al., 2010). 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 terminus 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 bi-partite cNLSs by matching each sequence with defined REs (Kosugi et al., 2009). The final data set contained 237 positive sequences (nuclear proteins that match REs). No negative data were used.

### 2.4 Statistical enrichment analysis on PTS1 data set

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 the “have-motif” proteins do not differ in terms of assigned property from those that do not.

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

### 2.5 Performance metric on synthetic data set

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 data set is used to evaluate the prediction accuracy. We collated four biological data sets. For the PTS1 and PY-NLS data sets, we aimed to evaluate DLocalMotif's ability to recover known local motifs. We found seven local motifs and their position relative to PTS1. We found seven motifs that co-occur with the PTS1 anchor, here named Motif1-Motif7. We validated the position of discovered motifs by searching the literature.

There is a high prevalence of hydrophobic residues at 5 to 11 upstream of the C-terminus in Motif1, Motif2, Motif5, 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 et al., 2003). It has been shown in the literature that in Candida boidinii, basic residues are found upstream of PTS1 (Mullen et al., 2003).

### 3 RESULTS

#### 3.1 Evaluating p-values on randomly generated data

Motif discovery methods may uncover highly significant motifs even when tested on random data sets (Harbison et al., 2004). To distinguish between spurious and biologically relevant significant motifs, the p-value of a motif discovery method should be evaluated extensively on random data sets. We generated random protein sequences of different lengths and 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. In Supplementary Fig. S1, 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 data sets 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) and DEME (Redhead and Bailey, 2007). The performance of MEME was based on positive data only but both DEME and DLocalMotif were tested on positive and negative data. In Fig. 1, each data point represents the average o.p from 50 data sets. The overall accuracy was calculated by averaging the o.p for all data sequences. 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 data sets, we uniformly chose $d$ to have up to a maximum of three point mutations.

For the negative random data sets, DLocalMotif achieved higher accuracy than MEME and DEME with average o.p of 0.83 as compared to 0.15 (DEME) and 0.19 (MEME). (See Fig. 1.) MEME either outperformed DEME or performed equally well as DEME in this test that has no additional information in the negative data.

For the decoy motif data sets, DLocalMotif outperformed MEME and DEME in terms of average overlap percentage with an average o.p of 0.74, compared to 0.18 (DEME) and 0.13 (MEME). MEME does not consider negative data, whereas DEME does and is 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.

It needs to be emphasized that most other protein motif discovery tools, including MEME and DEME, are not designed for local motif discovery. We have shown that DLocalMotif effectively recovers spatially confined motifs.

### 3.3 Evaluating DLocalMotif on biological datasets

We collated four biological data sets. For the PTS1 and PY-NLS data sets, we aimed to discover novel local motifs not found by existing methods. Using the ER retention and bi-partite cNLS data sets we aimed to evaluate DLocalMotif’s ability to recover known local motifs. All sequences that contain a match with each discovered motif were used to generate motif logos.

#### 3.3.1 Discovering motifs occurring with PTS1

We ran DLocalMotif on the PTS1 data. Fig. 2-A summarizes the logos of discovered local motifs and their position relative to PTS1. We found seven motifs that co-occur with the PTS1 anchor, here named Motif1-7. We validated the position of discovered motifs by searching the literature.

There is a high prevalence of hydrophobic residues at 5 to 11 upstream of the C-terminus in Motif1, Motif2, Motif5, 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 et al., 2003).
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 Fig. 2-A was low while the MSC was quite 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-S8, 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 thaliana) 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).

### 3.3.3 Discovering motifs occurring with PY-NLSs

Prolinetyrosine nuclear localization signals are recognized by the transport factor Kap32. We used DLocalMotif to investigate the existence of local motifs that supplement the PY anchor that appear in almost all Kap32 cargo. DLocalMotif discovered several novel motifs that figured strongly upstream of the anchor (see Fig. 2-C).

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 (Yamasaki et al., 2005; Saitou et al., 2007; Lee et al., 2000). We also searched the literature to find evidence of Kap32 interacting with Zf domains (specifically the C2H2 class). We found that ADR1, which contains C2H2 Zf domains, interacts with Kap104, the Kap32 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.

### 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-terminii clusters (Kosugi et al., 2009; Marfori et al., 2010). 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. 2-E). Motif2 and Motif3 co-occur with Motif1 containing clusters of basic residues (see Fig. 2-E). 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) and DEME (Redhead and Bailey, 2007) on the same data set. Both methods were able to uncover only one motif consisting of basic residues. Without the
anchor, the N terminal motifs are only weakly over-represented. The spatial confinement of the C-terminal motif is high, making both motifs easy targets for DLocalMotif.

4 CONCLUSION
In this paper 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, motif spatial confinement (MSC), motif over-representation (MOR) and motif relative entropy (MRE). These features establish if 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.

While 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 data sets. 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, MEME (Bailey et al., 2009) and DEME (Redhead and Bailey, 2007) 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-different. DLocalMotif discovered three entirely novel PY-NLS local motifs that overlap with C2H2 zinc finger 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 bi-partite 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, we note that DLocalMotif offers a compromise by detecting multiple smaller but spatially interlinked motifs.

REFERENCES

Erb, T. J., Ferriech-Revermann, L., Fuchs, G., and Alber, B. E. (2010). The apparent malate synthase activity of throbacter sphaeroides is due to two paralogous enzymes, (3x)-malyl-coenzyme a (coa)/(beta-methylmalyl-coa hyase and (3x)-maly- 


Fig. 2. Discovered local motifs in A) PTS1 data set B) ER retention signal data set C) PY-NLS data set D) and E) bipartite cNLS data set. 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).


