Journal of Theoretical Biology 377 (2015) 75–84

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Discovering short linear protein motif based on selective training of profile hidden Markov models

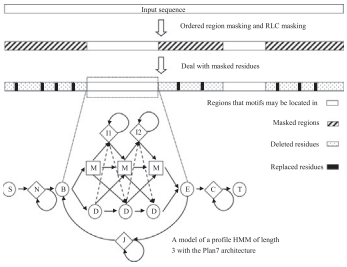
Tao Song, Hong Gu *

Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, Liaoning 116024, China

HIGHLIGHTS

- A short linear motif discovery algorithm based on profile HMMs is proposed.
- We apply a new method to deal with the masked residues obtained by ordered region and RLC masking.
- We adopt the selective training of HMMs to make full use of evolutionary information.
- Profile HMM-based method complements the existing algorithms and provides another way to analyze motifs.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 23 June 2014
Received in revised form
6 March 2015
Accepted 7 March 2015
Available online 17 March 2015

Keywords:
Intrinsic disorder prediction
Relative local conservation
Masked residues processing
Evolutionary weighting
Statistical significance

ABSTRACT

Short linear motifs (SLiMs) in proteins are relatively conservative sequence patterns within disordered regions of proteins, typically 3–10 amino acids in length. They play an important role in mediating protein–protein interactions. Discovering SLiMs by computational methods has attracted more and more attention, most of which were based on regular expressions and profiles. In this paper, a *de novo* motif discovery method was proposed based on profile hidden Markov models (HMMs), which can not only provide the emission probabilities of amino acids in the defined positions of SLiMs, but also model the undefined positions. We adopted the ordered region masking and the relative local conservation (RLC) masking to improve the signal to noise ratio of the query sequences while applying evolutionary weighting to make the important sequences in evolutionary process get more attention by the selective training of profile HMMs. The experimental results show that our method and the profile-based method returned different subsets within a SLiMs dataset, and the performance of the two approaches are equivalent on a more realistic discovery dataset. Profile HMM-based motif discovery methods complement the existing methods and provide another way for SLiMs analysis.

1. Introduction

Globular domains of proteins contribute to various molecular functions such as catalysis and ligand binding. Many bioinformatics tools have been developed for studies of globular domains including Pfam (Finn et al., 2004, 2011), SMART (Letunic et al., 2012), CDD (Marchler-Bauer et al., 2011), SCOP (Andreeva et al., 2004) and CATH (Sillitoe et al., 2013). In the past twenty years, SLiMs have attracted more and more attention as they can constitute another class of module that contribute to molecular functions of the proteins. SLiMs are predominantly found in natively disordered regions of protein sequences (Russell and Gibson, 2008). One property of SLiMs is their short length, typically 3–10 amino acids long, which results in low-binding affinities when SLiMs interact with other modules of proteins. However, this low affinity is ideal for transient interactions

* Corresponding author.
E-mail address: guhong@dlut.edu.cn (H. Gu).

http://dx.doi.org/10.1016/j.jtbi.2015.03.010
0022-5193/

in signal transduction or for quickly responding to a stimulus (Davey et al., 2010). Another property of SLiMs is linear which means that residues in one motif are adjacent in the primary sequence of the protein. This is unlike the globular domains which are required to make distant segments of the protein sequence being in close proximity in the tertiary structure of the protein (Diella et al., 2008). In addition to these, the conservation of these motifs varies: some are highly conserved while others allow substitutions, and the level of their conservation is lower than globular domains. However, they are more conserved than surrounding non-functional residues because of purifying selection. Due to the short length, high flexibility and low-binding affinities of SLiMs, *de novo* SLiMs discovery is more difficult than globular domains identification which is based on high-quality multiple alignments.

Computational prediction of SLiMs is to detect over-represented pattern in a set of sequences with a common attribute (*e.g.*, biological function, sub-cellular location or a common interaction partner). *De novo* SLiMs discovery can be divided into different categories according to motif representation: REs (Regular Expressions), PWMs (Position Weight Matrices), PSSMs (Position-Specific Scoring Matrices), profiles, HMMs and profile HMMs (Bailey, 2007; Eddy, 1998). There are two commonly used assumptions of motif distribution for motif discovery: (i) the one occurrence per sequence (OOPS) assumes that each sequence in the dataset contains exactly one occurrence of each motif, (ii) the zero or one occurrence per sequence (ZOOPS) assumes that each sequence may contain at most one occurrence of each motif (Bailey and Elkan, 1995a; Bailey et al., 2009; Lawrence and Reilly, 1990). The ZOOPS assumption is useful and practical for datasets in which some motifs may be missing from some of the sequences. In that case, the motifs found will be more accurate than using the OOPS assumption.

Traditionally, motif finding problem has been dominated by the methods based on REs. The TEIRESIAS algorithm (Rigoutsos and Floratos, 1998) finds motifs in two phases: scanning and convolution. During the scanning phase, elementary patterns (smaller pieces of motifs) are identified, and they must appear in at least K distinct protein sequences in the training sets. These elementary patterns are combined into maximal patterns for the convolution phase. The DILIMOT (Neduva et al., 2005; Neduva and Russell, 2006) firstly deals with sequences to remove regions unlikely to contain motif instances, such as globular domains, signal peptides, transmembrane and coiled-coil regions defined by using SMART and Pfam with default parameters. Then it finds motifs in the remaining sequence based on the TEIRESIAS algorithm and uses the binomial distribution to score and rank found motifs. The SLiMDisc (Davey et al., 2006) is another TEIRESIAS-based method. It is similar to DILIMOT in the part of data acquisition, while in the part of data analysis it further filters the found motifs according to a number of optional criteria (evolutionary relatedness, information content and surface probability) and ranks the motifs according to information content. The SLiMFinder (Edwards et al., 2007) algorithm consists of two parts: SLiMBuild constructs dimers and then combines them with amino acid ambiguity and variable-length wildcard spacers in unrelated protein clusters; SLiMChance estimates the probability of returned motifs arising by chance and assigns a significance value to each motif using the binomial distribution.

When extending the regular expression-based approaches to the methods based on profiles, the found motifs can obtain a richer representation. The MEME algorithm (Bailey and Elkan, 1994, 1995b) adopts the expectation maximization (EM) algorithm to maximize the probability of unaligned sequences given the profile. Firstly, it chooses starting points systematically, then runs to convergence from the point with the highest likelihood, and finally masks the appearances of the found motifs for further discovering multiple non-redundant motifs. However, it can only find contiguous motifs (motif with variable-length gaps are considered as two separate motifs) and performs relatively poor for motifs in disordered regions of proteins. The MemeFinder algorithm (Haslam and Shields, 2012) applies evolutionary weighting that accounts for redundancy amongst homologous proteins and relative local conservation (RLC) masking (Davey et al., 2009) in the MEME algorithm to discover motifs with richer and more accurate representations.

In this paper, we put forward a *de novo* motif discovery method, where SLiMs are represented by profile HMMs with the Plan 7 architecture (Eddy, 1998). The assumption of OOPS was adopted in this paper which is controlled by setting the transition probability from the end state to joining segment unaligned sequence state to zero. The proposed method was inspired by MemeFinder. However, different from MEME and MemeFinder which employ subsequences, our method uses the full-length sequences during training models. The evolutionary weightings, which reflect the evolutionary relationships of the query sequences to each other and account for redundancy amongst homologous proteins, were applied in the training of profile HMMs. This was achieved by the selective training of HMMs, a method that has been widely used in speech recognition (Foo and Dong, 2003; Foo et al., 2004; Meyer and Schramm, 2006) and proved that if weightings are added to the training sequences, the convergency of the maximum likelihood estimation (MLE) still holds (Arslan and Hansen, 1999). The ordered residues of sequences were predicted by IUPred (Dosztányi et al., 2005) and filtered out to increase the likelihood of discovering SLiMs in the query protein sequences. In addition, the RLC score was applied to mask out unconserved residues because residues of motif instances are more conserved than their surrounding residues. A new method was proposed to process the masked residues since it is more complicated for training HMMs by the Baum–Welch algorithm (Rabiner, 1989) than pattern enumeration-based and profile-based methods. This approach is equivalent to the way that treats the masked residues as special events of profile HMMs when running the Baum–Welch algorithm. Although it is simple, this trick effectively reduces the algorithmic complexity of training HMMs as well as the impact of long ordered regions. Finally, statistical significance of the found motifs was returned by the Mann–Whitney U test (Tanaka et al., 2014) which needed to calculate the log-odds scores of sequences from both the input set and the null set.

An experimentally validated SLiMs dataset collected from the Eukaryotic Linear Motifs (ELM) resource (Dinkel et al., 2014) and a realistic protein–protein interaction dataset downloaded from the Human Protein Reference Database (HPRD) (Prasad et al., 2009) have been used to compare different methods. The results indicate that the performance of our method is better than the profile-based method at both the residue level and the site level on the SLiMs dataset. However, they recovered different subsets of annotated SLiMs within it which means the two methods can complement each other. Although, our method used OOPS as the assumption of motif distribution at the present stage, the results on the realistic dataset show that its performance is approximately equivalent to the profile-based method with the ZOOPS assumption. In general, we believe that the profile HMM-based SLiMs discovery method proposed in this paper complements the existing algorithms and provides another way to analyze the SLiMs.

2. Methods and materials

2.1. Problem formulation

We represent the motif by HMM and model the defined positions (positions that cannot tolerate an amino acid substitution or can tolerate a limited number of amino acid substitutions that usually share some physicochemical or structural property)