BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–7

Detection of new protein domains using co-occurrence: application to *Plasmodium falciparum*

Nicolas Terrapon [1,2], Olivier Gascuel [1], Éric Maréchal [2] and Laurent Bréhélin [1,*]

[1]Méthodes et algorithmes pour la Bioinformatique, LIRMM, Univ. Montpellier 2, CNRS, 161 rue Ada 34392 Montpellier Cedex 5 France
[2]CEA Grenoble iRTSV/LPCV, 17 rue des Martyrs, 38054 Grenoble cedex 9 France

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Hidden Markov Models (HMMs) have proved to be a powerful tool for protein domain identification in newly sequenced organisms. However, numerous domains may be missed in highly divergent proteins. This is the case for *Plasmodium falciparum* proteins, the main causal agent of human malaria.

Results: We propose a method to improve the sensitivity of HMM domain detection by exploiting the tendency of the domains to appear preferentially with a few other favorite domains in a protein. When sequence information alone is not sufficient to warrant the presence of a particular domain, our method enables its detection on the basis of the presence of other Pfam or InterPro domains. Moreover, a shuffling procedure allows us to estimate the false discovery rate associated with the results. Applied to *P. falciparum*, our method identifies 585 new Pfam domains (versus the $3\,683$ already known domains in the Pfam database) with an estimated error rate below 20%. These new domains provide 387 new Gene Ontology annotations to the *P. falciparum* proteome. Analogous and congruent results are obtained when applying the method to related *Plasmodium* species (*P. vivax* and *P. yoelii*).

Availability: Supplementary Material and a database of the new domains and GO predictions achieved on *Plasmodium* proteins are available at http://www.lirmm.fr/∼terrapon/codd/

Contact: brehelin@lirmm.fr

1 INTRODUCTION

Among relevant annotations that can be attached to a protein, domains occupy a key position. Protein domains are sequential and structural motifs that are found independently in different proteins, in different combinations and, as such, seem to be functional sub-units of proteins above the raw amino-acid sequence level (Richardson, 1981). Several approaches have been developed to define and identify domains. Some are based on observed distinct structural classes of proteins (Murzin *et al.*, 1995). Others are inferred by clustering conserved subsequences (Mulder *et al.*, 2007; Finn *et al.*, 2008). One of the most widely used domain schemata is the Pfam database (Finn *et al.*, 2008). In this database, each domain family is defined with a set of distinct representative protein sequences, manually selected and aligned, and used to learn a *Hidden Markov Model* (HMM) (Durbin *et al.*, 1998) of the domain. The current release of the Pfam database (version 23.0 of July 2008) offers a large collection of $10\,340$ HMMs/domains, which account for over 73% of all proteins in the Uniprot database (UniProt Consortium, 2009). Some Pfam HMMs have been annotated by the InterPro consortium (Mulder *et al.*, 2007) in the *Gene Ontology* (GO) (Gene Ontology Consortium, 2000). According to the InterPro annotation policy, a domain is annotated with a given GO term if all proteins where this domain is known also share this GO term. This stringent rule allows, when a new domain is detected in a protein, transfer of its annotations to this protein.

When analyzing a new protein sequence, each Pfam HMM is used to compute a score that measures the similarity between the sequence and the domain. If the score is above a given threshold provided by Pfam (score thresholds differ depending on the HMMs), then the presence of the domain can be asserted in the protein. However, when applied to highly divergent proteins, this strategy may miss numerous domains. For example, with *Plasmodium falciparum*, the main causal agent of human malaria, no Pfam domains are detected in nearly 50% of its proteins, while many domain types seem to be missing from the *P. falciparum* library—see Supp. Table 1 for a comparison of the protein coverage and domain numbers between several eukaryotes. This can be partly explained by the highly atypical genome of *P. falciparum*, which is composed of above 80% A+T, and involves long low-complexity insertions of unknown function believed to form non-globular domains (Pizzi and Frontali, 2001). This strongly biases the amino-acid composition of *P. falciparum* proteome, in which six amino acids account for more than 50% of the protein composition (Bastien *et al.*, 2005). In this article, we propose a new method to increase the sensitivity of Pfam domain detection in divergent proteins like those of *P. falciparum*. Our method involves lowering the thresholds provided by Pfam for detecting domains. This enables more domain detections, but at the expense of numerous false positive predictions. The core of the method is a filter procedure based on domain co-occurrence properties which selects the potential domains that are most likely true.

*to whom correspondence should be addressed