RESEARCH

PDAMIML:A Multi-Instance Multi-Label Learning Framework for Protein Domain Annotation

Jane E Doe[1*†] and John RS Smith[1,2]

*Correspondence:
jane.e.doe@cambridge.co.uk
[1]Department of Zoology, Cambridge, Waterloo Road, London, UK
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

First part title: Text for this section.

Second part title: Text for this section.

Keywords: domain annotation; multi-instance multi-label; SVM; auto covariance transformation

Introduction

One of the most challenging and intriguing problems in the postgenomic era is the characterization of the biochemical functions of proteins. Accurate computational assignment of protein function is becoming a useful resource for both the community at large and the curators that eventually assign function to proteins. It is known that domains appear either singly or in combination with other domains as building blocks in a protein citeApic2001 citeWang2006. Domains play important roles in the process of protein-protein interactions which determine the function of the protein citeBork1999. Identifying functions of domains provides key clues in the annotation of protein functions. Since the number of domains is relatively finite while the number of unannotated proteins is considerably large, it will be much easier to infer the function of proteins if the functions of their component domains are determined. Therefore, there is a great need to develop accurate computational methods for domain function annotation.

Traditional work of annotation of functions is based on the physicochemical properties of special structure and primarily conducted manually, which is a time-consuming, low-efficient and experience-dependent process. Moreover, with the rapid growth of proteins and domains, manual annotation becomes increasing infeasible. To deal with this problem, researchers introduced computational methods into the field of function annotation, in order to simplify and improve the work.

Despite a variety of the strategies and methods have used in identifying protein functions, only a few researchers have paid attentions to the field of domain annotation. For which only a relatively small number of domains have been annotated until now, several innovative methods have been proposed to predict domain function. Schug textitet al. citeSchug2002 described a heuristic algorithm for associating Gene Ontology (GO) citeAshburner2000 defined molecular functions to proteins as listed in the ProDom citeCorpet2000 and CDD citeMarchler2003 databases. The

algorithm generates rules for function-domain associations based on the intersection of functions assigned to gene products by the GO consortium that contain ProDom and/or CDD domains at varying levels of sequence similarity. Lu textitet al. citeLu2004 utilized protein-domain mapping (P2D) features to construct a logistic regression model to investigate the association rules between target domains and GO terms.

Recently, domain-domain interaction information has provided a new way to predict domain function, since it's rational to assume that interacting domains have a high probability to share similar functions. Deng *et al.* citeDeng2002 applied a Maximum Likelihood Estimation method to infer interaction domains that are consistent with the observed protein-protein interaction. Their method showed robustness in analyzing incomplete data sets and dealing with various experimental errors. Moreover, some researchers utilize domain coexisting features to predict domain function. For example, Wang *et al.* citeWang2007 gave a general framework to predict protein interactions by considering the information of both multi-domains and multi organisms, which can also be applied to identify cooperative domains, further to annotate functions of domains.

Based on the previous works, Zhao and Wang citeZhao2008 designed two methods, the threshold-based classification method and the support vector machine method, for protein domain function prediction by integrating heterogeneous information sources, including protein-domain mapping features(P2D) citeLu2004, domain-domain interactions(DDI) citeDeng2002, and domain coexisting features(CDD) citeWang2007, improving not only prediction accuracy but also annotation reliability. During these two methods, the threshold-based classification method outperforms the support vector machine method according to Zhao's citeZhao2008 experiments.

In those traditional learning formalizations, each domain is represented by an instance (or feature vector) and associated with single GO term. Although the above single instance and single label formalization is prevailing and successful, it is not an appropriate model for the actual situation of domain function annotation. In fact, each domain usually exists in multiple proteins, which can be described by a feature vector, and the domain can belong to multiple categories since it is associated with some different functions.

In this paper, we propose a novel Multi-Instance Multi-Label Learning (MIML) citeZhou2012 based framework, PDAMIML, to predict functions of protein domains accurately. PDAMIML combines MIML model, support vector machine citeVapnik1998 and auto-cross covariance citeWold1993 to overcome the multi-label classification problem and effectively utilize the features of Position-Specific Scoring Matrix (PSSM) citeStephen1997. Furthermore, we design an ensemble method, PDAMIML-ensemble, that integrates PDAMIML and other two eminent threshold-based approaches (CDD and P2D citeZhao2008) with majority voting strategy. Our experimental results show that PDAMIML-ensemble significantly outperforms the state-of-the-art domain annotation approaches.

Materials and Methods

Datasets

Relationships between proteins and domains are obtained from InterPro Database citeSarah2011. The function annotations of domains are generated from GOA Database citeEvelyn2003. InterPro contains three main entities: proteins, signatures and entries. The signatures from InterPro comes from 11 member databases, including Pfam citeBateman2002, ProDom citeCorpet1999, etc. The protein-domain-function dataset contains 13137 proteins, 5748 domains, and 2535 GO terms. Around 76% domains are annotated with more than one GO term, and every domain has 2.5 annotations on average. We choose the top 100 domain with the most GO terms as the data set, and then filter sparse labels which are inconvenient for study, select the most frequent 10 GO terms as target labels out of the total 188 labels attached to the top 100 domains according to their frequency. The diagram of frequency of GO terms is showed in Fig.refgoterms and the definition of the 10 GO terms we select is demonstrated in TABLE refdefinition. Each protein can be viewed as an instance and each domain is represented as a bag of instances.

Figure 1 GO terms and their frequency. A short description of the figure content should go here.

Table 1 Selected GO terms and their definition. This is where the description of the table should go.

GO term ID	Definition
GO:0005737	cytoplasm
GO:0003677	DNA binding
GO:0005524	ATP binding
GO:0005506	iron ion binding
GO:0055114	oxidation-reduction process
GO:0000166	nucleotide binding
GO:0003700	DNA binding transcription factor activity
GO:0020037	heme binding
GO:0006355	regulation of transcription, DNA-dependent
GO:0005634	nucleus

In each experiment, the whole data set is randomly partitioned into two parts, a training set, accounting for 70% of the data set, and a test set, 30%. The training set is used to build classifiers, and the test set is used to evaluate the performance of the corresponding classifier. The whole experiment is repeated for 10 times to get an average and common performance, and every time, all the model parameters are tuned with 10-fold cross validation on the training set to optimize the model performance.

MIML for domain annotation

In this section, we first describe in detail how to formulate protein domain annotation as an MIML problem.

A protein domain is a conserved part of a given protein sequence and structure that can evolve, function, and exist independently of the rest of the protein chain. Many proteins consist of several structural domains; meanwhile one domain may appear in a variety of different proteins. Molecular evolution uses domains as building blocks and these may be recombined in different arrangements to create proteins

with function citeHeringa2005. Besides acting as the structural unit of proteins, domains often form functional units. In a multi-domain protein, each domain may fulfill its own function independently, or in a concerted manner with its neighbors citeHeringa2005. Since proteins interact with each other through domain interactions, the functions of domains determine the functions of their host proteins. From the analysis above, we realize that domains act as the structural and functional units, and proteins and functions are connected by those particular domains, as though some departments with special functions if the whole protein is compared to a complex factory. Each domain is associated with multiple proteins and multiple GO terms derived from its various interaction with other domains.

However, the recent work of domain function annotation still focuses on single label prediction and uses one vector to represent one domain, since there was no explicit relationship between some particular domains and some specific function and no efficient framework to deal with multi-instance and multi-label problem. In addition, the annotation work in particular domain with multi-instance is much more sophisticated and difficult than the work with single-instance.

As illustrated in Fig.refmodel, we propose a novel MIML citeZhou2007 based method, which combines support vector machine and auto-cross covariance transformation, to improve the prediction performance in domain anntation.

Figure 2 Sample figure title. A short description of the figure content should go here.

PDAMIML framework

Auto-cross covariance transformation

Majority voting

Evaluation measures

Results and discussions

The impacts of LG

Performance comparison with the state-of-the-art approaches

Performance comparison on different single labels

Conclusion

Acknowledgements

Sub-sub heading for section

Text for this sub-sub-heading ...

Sub-sub-sub heading for section Text for this sub-sub-sub-heading ... In this section we examine the growth rate of the mean of Z_0, Z_1 and Z_2. In addition, we examine a common modeling assumption and note the importance of considering the tails of the extinction time T_x in studies of escape dynamics. We will first consider the expected resistant population at vT_x for some $v > 0$, (and temporarily assume $\alpha = 0$)

$$E\big[Z_1(vT_x)\big] = E\bigg[\mu T_x \int_0^{v\wedge 1} Z_0(uT_x)\exp\big(\lambda_1 T_x(v-u)\big)\,du\bigg].$$

If we assume that sensitive cells follow a deterministic decay $Z_0(t) = xe^{\lambda_0 t}$ and approximate their extinction time as $T_x \approx -\frac{1}{\lambda_0}\log x$, then we can heuristically estimate the expected value as

$$\begin{aligned} E\big[Z_1(vT_x)\big] &= \frac{\mu}{r}\log x \int_0^{v\wedge 1} x^{1-u} x^{(\lambda_1/r)(v-u)}\, du \\ &= \frac{\mu}{r} x^{1-\lambda_1/\lambda_0 v}\log x \int_0^{v\wedge 1} x^{-u(1+\lambda_1/r)}\, du \\ &= \frac{\mu}{\lambda_1-\lambda_0} x^{1+\lambda_1/rv}\left(1-\exp\left[-(v\wedge 1)\left(1+\frac{\lambda_1}{r}\right)\log x\right]\right). \quad (1) \end{aligned}$$

Thus we observe that this expected value is finite for all $v > 0$ (also see [1, 2, 3, 4, 5]).

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Text for this section . . .

Acknowledgements
Text for this section . . .

Author details
[1]Department of Zoology, Cambridge, Waterloo Road, London, UK. [2]Marine Ecology Department, Institute of Marine Sciences Kiel, Düsternbrooker Weg 20, 24105 Kiel, Germany.

References
1. Koonin, E.V., Altschul, S.F., Bork, P.: Brca1 protein products: functional motifs. Nat Genet **13**, 266–267 (1996)
2. Kharitonov, S.A., Barnes, P.J.: Clinical Aspects of Exhaled Nitric Oxide. in press
3. Zvaifler, N.J., Burger, J.A., Marinova-Mutafchieva, L., Taylor, P., Maini, R.N.: Mesenchymal cells, stromal derived factor-1 and rheumatoid arthritis [abstract]. Arthritis Rheum **42**, 250 (1999)
4. Jones, X.: Zeolites and synthetic mechanisms. In: Smith, Y. (ed.) Proceedings of the First National Conference on Porous Sieves: 27-30 June 1996; Baltimore, pp. 16–27 (1996). Stoneham: Butterworth-Heinemann
5. Margulis, L.: Origin of Eukaryotic Cells. Yale University Press, New Haven (1970)

Figures

Figure 3 Sample figure title. A short description of the figure content should go here.

Figure 4 Sample figure title. Figure legend text.

Tables

Table 2 Sample table title. This is where the description of the table should go.

	B1	B2	B3
A1	0.1	0.2	0.3
A2	...	..	.
A3	..	.	.

Additional Files
Additional file 1 — Sample additional file title
Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title
Additional file descriptions text.