Using GO terms to evaluate protein clustering

Hugo Bastos*, Daniel Faria, Catia Pesquita and André O. Falcão

University of Lisbon, Department of Informatics, Campo Grande, 1749-016 Lisbon, PORTUGAL

ABSTRACT

Motivation: Protein sequence data is growing at an exponential rate. However a considerable portion of this data is redundant, with many new sequences being very similar to others in the databases. While clustering has been used to reduce this redundancy, the influence of sequence similarity in the functional quality of the clusters is still unclear.

Results: In this work, we introduce a greedy graph-based clustering algorithm, which is tested using the Swiss-Prot database. We study the topology of the protein space as function of the threshold BLAST *e*-values, and the functional characterization of the clusters using the Gene Ontology. Initial results show that seemingly the cluster centers alone can capture a large portion of the information content of the database, therefore largely reducing its redundancy. Also it was found an expected increase of cluster functional coherence and characterization with the stringency of the threshold, as well as the amount of information captured by the cluster centers.

1 INTRODUCTION

The Universal Protein Resource (UniProt) provides a central hub for the collection of protein sequences, with accurate, consistent and rich annotation (The UniProt Consortium, 2007). As such, it can be considered a subset of the universal protein space. The protein space is a metric space containing all protein sequences that uses sequence similarity as a distance function. While it is true that the number of sequences in UniProt is growing exponentially over time, it is also true that a significant proportion of these sequences is redundant, being very similar to others already present in the database.

One way to cope with this large amount of partially redundant data, is through clustering methods, which ideally may reduce the dimension of the protein space without loss of information. In fact, UniProt itself provides clustered versions of its full database with several levels of sequence similarity (UniRef) to facilitate information retrieval and accelerate the querying process. Other purposes for clustering biological sequences include functional annotation, comparative genomics and structural genomics (Petryszak *et al.*, 2005, Nikolski and Sherman, 2007, Yan and Moult, 2005).

Protein sequence clustering methods can be classified according to two aspects: the clustering algorithm type used, which is usually either hierarchical or graph-based; and the criterion used to group sequences, which is usually either domain-based or family-based.

On structural genomics studies, Charette *et al.* (2006) used protein clustering for protein ligand-docking and molecular dynamics, whereas Shen *et al.* (2005) used it for protein class prediction, and Yan and Moult (2005) searched for representative family structural templates.

In the field of functional genomics, Pellegrini *et al.* (1999) used clustering to functionally assign proteins, whereas Nikolski and Sherman (2007) used a consensus algorithm tailored for comparative genomics projects. Hierarchical algorithms include ProtoMap (Yona *et al.*, 1999), ProtoNet (Sasson *et al.*, 2003), and CLUGEN (Ma *et al.*, 2005) which aim at constructing a comprehensive view of the protein space by means of family-based classification. On the other hand, CluSTr's (Petryszak *et al.*, 2005) main objective is automated protein annotation. As for graph-based algorithms, Abascal and Valencia (2003) used this type of clustering to identify families for comparative genomics and protein functional inference, while the Cluster-C algorithm (Mohseni-Zadeh *et al.*, 2004) is used for protein family construction within whole proteomes. SEQOPTICS (Chen *et al.*, 2006) used a different clustering algorithm which performs density-based ordering.

One way to evaluate the biological quality of the clustering process, is by analyzing the amount of information (functional or other) conserved in the cluster centers vs. the reduction in dimension achieved. In this context, a unified and structured vocabulary to describe proteins' functional aspects would provide a unique background for evaluation, facilitating the clusters' functional characterization.

Being already extensively used to annotate several biological databases (including UniProt), the Gene Ontology (GO) (Ashburner *et al.*, 2000) is one of the best choices for this end. Indeed, GO has already been used as a background for functional comparison of proteins (Lord *et al.*, 2003) and to understand the relation between protein sequence and function (Duan *et al.*, 2006).

In this paper we present a graph-based protein sequence clustering algorithm which was tested using the Swiss-Prot database, with a discrete range of BLAST *e*-value cut-offs. We study the reduction in protein space and its topology as

[*] To whom correspondence should be addressed.

function of sequence similarity, and evaluate the biological quality of the clusters using three GO-based parameters which measure different aspects: functional coherence, functional characterization and representativeness of the cluster center.

2 DATA AND IMPLEMENTATION

In order to evaluate the cluster algorithm all the Swiss-Prot portion of the UniProt Knowledgebase (Release 9.6) was used. After filtering out all segment sequences, an all-against-all BLAST homology search was performed with the remaining 2.51×10^5 sequences. The maximum *e*-value accepted in the BLAST step was 10^{-4}, which resulted in about 5.48×10^7 pairwise BLAST comparisons. A simple greedy graph partitioning clustering algorithm was then applied to the BLAST results, as is described next.

For a given *e*-value threshold an edge (weighted by *e*-value) between two nodes (sequences) is said to exist if the *e*-value between the two proteins is below that threshold, and a list of sequences is constructed, containing for each the total number of edges (cardinality). The algorithm then proceeds as follows:

(1) Each node is sorted descendingly according to the number of edges it has (cardinality).

(2) The node with the highest cardinality is selected from the list. If it does not belong to a cluster, a new one is created with this node as its center. All proteins linked to that protein by an edge are then assigned to this new cluster.

(3) The clustering proceeds iteratively down the cardinality list until no more nodes can be assigned.

The clustering procedure was run with a discrete set of *e*-value thresholds ranging from 1×10^{-4} (most permissive threshold) to 1×10^{-100} (most stringent threshold). For biological validation of the clustering, the GOA-UniProt release 45.0 and the GO release 2007-02 were used. Only the *molecular function* aspect of GO was considered, since the goal was to characterize the clusters in functional terms, and the relation between the other GO aspects and functional similarity is unclear.

3 RESULTS AND DISCUSSION

Within the chosen range of *e*-value thresholds, the size of the protein space achieved with clustering varied non-linearly with the threshold, growing from 9% of the original size (22799 protein clusters) with the most permissive threshold to 25% (62254 protein clusters) with the most stringent (Figure 1). As would be expected, this variation was accompanied by a reduction in the average cluster size from 18 to 3 proteins, as well as a growing fraction of singletons from 42 to 62%, from the most permissive to most stringent threshold.

As the goal of this work was to evaluate the clustering process at the functional level, using *molecular function* GO terms, only non-singleton clusters with at least one annotated protein were considered. While the fraction of proteins annotated with *molecular function* GO terms in the dataset is 84%, the fraction of annotated clusters grew with threshold stringency (from 72 to 87%) although the absolute number of non-annotated clusters was approximately constant (~4000 clusters).

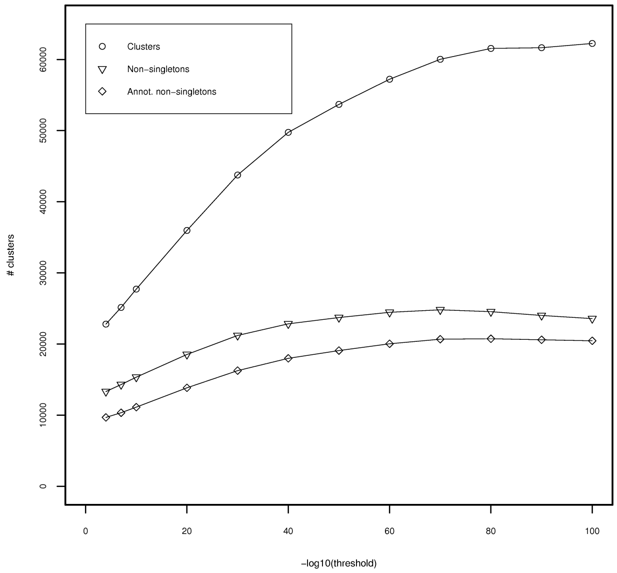

Fig. 1. Number of clusters, non-singleton clusters and annotated non-singleton clusters according to clustering *e*-value threshold.

In order to evaluate the clustering process, three GO term-based parameters were developed to measure different aspects of cluster quality: *GOoccurrence*, *GOscore* and *GO-center*. For a given cluster *C*, *GOoccurrence* is given by the average frequency of annotation within the cluster of each of the cluster's GO terms:

$$GOoccurrence(C) = \mathrm{AVG}_{term \in C}[\mathrm{freq}_C\,(term)]$$

and measures how coherent is the cluster functionally. The maximum for this parameter is achieved when all terms are annotated to all of the cluster's proteins (i.e. when all proteins are functionally identical). However, as annotations are considered at all levels of the GO graph (i.e. direct annotations and their ancestors) this parameter is slightly biased by the more general terms (which usually have greater frequency of annotation).

GOscore is given by the maximum of term information content (IC) times term annotation frequency:

$$GOscore(C) = \mathrm{MAX}_{term \in C}[\mathrm{freq}_C(term) \times \mathrm{IC}(term)]$$

and measures how well characterized is the cluster functionally, by capturing the most representative functional aspect: one that is simultaneously specific (with high IC) and frequent within the cluster. Note that the product of IC times frequency is actually proportional to the logarithm of the probability of the term occurring in the cluster with that frequency.

GOcenter is given by the fraction of the cluster's terms which are annotated to the cluster center protein, and provides a measure of how much of the cluster's functional aspects are captured by the center (*i.e.* how representative the center is of the cluster).

When analyzing the average variation of these three evaluation parameters with the threshold *e*-value used for clustering, they were found to increase non-linearly with threshold stringency (Figure 2), showing an apparent asymptotic behavior as they reach high stringency values.

The increases in *GOoccurence* (32% increment) and in *GOscore* (14% increment) agree with what would be expected of the clustering process: by increasing the level of sequence similarity required for a protein to be in a cluster, we obtain clusters that are functionally more coherent and better characterized.

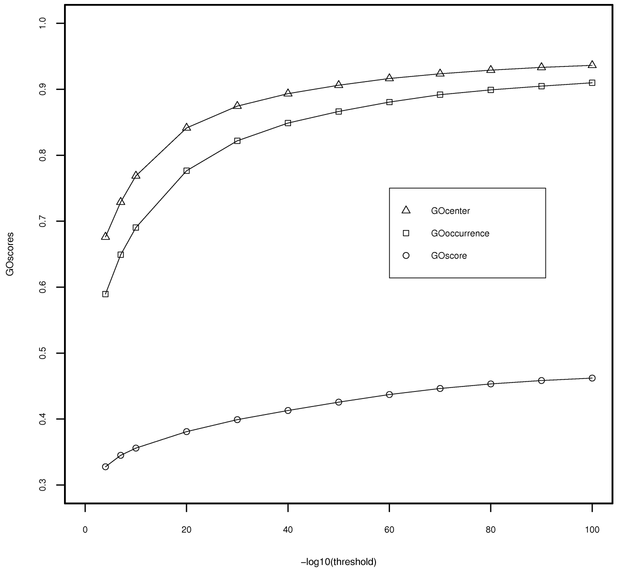

Fig. 2. Average GO term-based evaluation parameters as function of the threshold *e*-value used for clustering.

While the absolute values of *GOscore* may appear low, considering that their theoretical maximum value is 0.48 (average *IC* of the most specific term annotated to each protein in the dataset), their range corresponds to 59-96% of the maximum, which is similar to the ranges of the other parameters.

Thus, both *GOoccurrence* and *GOscore* support the quality of the clustering process, and provide a biological validation for the clustering algorithm used.

As for the *GOcenter* value, its increase (from 0.68 to 0.94) means that the representativity of the cluster centers is increasing with stringency, supporting the notion that the protein space can be successfully reduced through clustering without significant loss of functional information. This also shows that the algorithm is successful, since the cluster centers it selects capture most of the information in the clusters.

One issue that can be raised is that GO annotations inferred by sequence similarity lead to circularity of the results. However, the number of such annotations which are curated (evidence code ISS) amounts only to 1.5% of all annotations, and while it is likely that a portion of the electronic annotations (90.1% of all annotations) are also inferred by sequence similarity, that portion is difficult to estimate. Therefore, while the issue is acknowledged, there is no way to avoid it save ignoring all ISS and electronic annotations, which would imply losing the vast majority of the available information. Furthermore, the main consequence of circularity will likely be an increase in the absolute values of the evaluation parameters used, which is counterbalanced by the fact that proteins with no annotations cause a decrease in those parameters. It is unlikely that the patterns observed (Figure 2) are uniquely an artifact of data circularity, since that would mean that the majority of annotations are not only inferred by sequence similarity but also erroneous.

Interestingly, the approach developed and specifically the parameters *GOscore* and *GOoccurence* can be used to evaluate and improve the quality of annotations (by identifying protein clusters which have very incoherent annotations, and by providing a safer basis for inferring functional annotations).

4 CONCLUSION

A simple, greedy algorithm to cluster proteins based on sequence similarity was developed. A range of BLAST *e*-values was used as threshold for the clustering process so as to study the topology and functional quality of the cluster space as function of the sequence similarity.

Three parameters, based on GO *molecular function* annotations, were developed to evaluate different aspects of cluster quality: coherence, functional characterization and center representativeness.

Cluster quality in all these aspects was found to increase with threshold stringency, validating the biological significance of the clustering algorithm used and also the apparent ability of present method to reduce the protein space without significant loss of functional information.

Future work will focus on using additional biological validation methods to complement those based on GO, so as to overcome the issue of coverage (only 84% of the sequences in the dataset have GO annotations). The clustering algorithm will also be improved by using the proteins' level of annotation as an additional criterion to select the cluster centers. Furthermore, clusters will be characterized taxonomically and using GO-based functional semantic similarity.

ACKNOWLEDGEMENTS

This work was partially supported by the Peptides project of the LaSIGE research group, and also by grant ref. SFRH/BD/29797/2006 awarded by the Portuguese 'Fundação para a Ciência e a Tecnologia'.

REFERENCES

Abascal, F. and Valencia, A. (2003) Automatic annotation of protein function based on family identification. *Proteins*, **53**(3), 683–692.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet*, **25**(1), 25–29.

Charette, B., Macdonald, R., Wetzel, S., Berkowitz, D. and Waldmann, H. (2006) Protein structure similarity clustering: Dynamic treatment of pdb structures facilitates clustering. *Angew Chem Int Ed Engl*, **45,** 7666-7770.

Chen, Y., Reilly, K., Sprague, A., and Guan, Z. (2006) SEQOPTICS: a protein sequence clustering system. *BMC Bioinformatics*, **7** (Suppl 4), S10.

Consortium, The Uniprot (2007) The universal protein resource. *Nucl Acids Res*, **35** (Suppl 1), D193–197.

Duan, Z., Hughes, B., Reichel, L., Perez, D. and Shi, T. (2006) The relationship between protein sequences and their gene ontology functions. *BMC Bioinformatics*, **7** (Suppl 4), S11.

Lord, P.W., Stevens, R.D., Brass, A. and Goble, C.A. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**(10), 1275–1283.

Ma, Q., Chirn, G.-W., Cai, R., Szustakowski, J.D. and Nirmala, N. (2005) Clustering protein sequences with a novel metric transformed from sequence similarity scores and sequence alignments with neural networks. *BMC Bioinformatics*, **6**, (242-255).

Mohseni-Zadeh, S., Brezellec, P. and Risler, J.-L. (2004) Cluster-C, an algorithm for the large-scale clustering of protein sequences based on the extraction of maximal cliques. *Computational Biology and Chemistry*, **28**(3), 211–218.

Nikolski, M. and Sherman, D. (2007) Family relationships: should consensus reign?--consensus clustering for protein families. *Bioinformatics*, **23**(2), e71–76.

Pellegrini, M., Marcotte, E., Thompson, M., Eisenberg, D. and Yeates, T. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA*, **96**(8), 4285–4288.

Petryszak, R., Kretschmann, E., Wieser, D. and Apweiler, R. (2005) The predictive power of the CluSTr database. *Bioinformatics*, **21**(18), 3604–3609.

Sasson, O., Vaaknin, A., Fleischer, H., Portugaly, E., Yonatan, B., Linial, N. and Linial, M. (2003) Protonet: hierarchical classification of protein space. *Nucleic Acids Research*, **31**(1), 348–352.

Shen, H.-B., Yang, J., Liu, X.-J. and Chou, K.-C. (2005) Using supervised fuzzy clustering to predict protein structural classes. *Biochemical and Biophysical Research Communications*, **334**(2), 577–581.

Wetlaufer, D. (1973) Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci USA*, **70**(3), 697–701.

Yan Y. and Moult, J. (2005) Protein family clustering for structural genomics. Journal of Molecular Biology, **353**(3), 744–759.

Yona, G., Linial, N. and Linial, M. (1999) Protomap: Automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. PROTEINS: Structure, Function, and Genetics, **37**(3), 360–378.