The predictive power of the CluSTr database

Robert Petryszak, Ernst Kretschmann, Daniela Wieser and Rolf Apweiler*

EMBL Outstation Hinxton, The European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

Received on April 5, 2005; revised and accepted June 14, 2005

Advance Access publication June 16, 2005

ABSTRACT

Summary: The CluSTr database employs a fully automatic single-linkage hierarchical clustering method based on a similarity matrix. In order to compute the matrix, first all-against-all pair-wise comparisons between protein sequences are computed using the Smith–Waterman algorithm. The statistical significance of the similarity scores is then assessed using a Monte Carlo analysis, yielding Z-values, which are used to populate the matrix. This paper describes automated annotation experiments that quantify the predictive power and hence the biological relevance of the CluSTr data. The experiments utilized the UniProt data-mining framework to derive annotation predictions using combinations of InterPro and CluSTr. We show that this combination of data sources greatly increases the precision of predictions made by the data-mining framework, compared with the use of InterPro data alone. We conclude that the CluSTr approach to clustering proteins makes a valuable contribution to traditional protein classifications.

Availability: http://www.ebi.ac.uk/clustr/
Contact: rolf.apweiler@ebi.ac.uk

1 INTRODUCTION

1.1 Sequence clustering

Protein sequences are classified into families, i.e. groups of proteins that share a significant sequence (Yona et al., 2000) or domain (Mulder et al., 2003) similarity. As a result of many large-scale sequencing projects, the number of publicly available protein sequences is increasing exponentially. This ever-increasing body of data cannot be annotated manually, given the high cost of human curation, and automatic methods are required to increase the coverage on a reasonable level of precision.

Examples of methods in which manual curation is used to refine families found by automatic classification procedures are Clusters of Orthologous Groups (COGs) (Tatusov et al., 2003) and PIR Super-Family (Wu et al., 2003). Fully automatic classification methods, on the other hand, are used in methods such as CluSTr (Kriventseva et al., 2001), ProtoMap (Yona et al., 2000) and TribeMCL (Enright et al., 2002).

The process of automatic sequence classification is known as sequence clustering and involves an algorithmic procedure for building clusters of proteins with a similarity over a particular threshold (e.g. single-linkage—Fig. 1). The similarity score can be based on sequence alignment (CluSTr) or pattern matching (InterPro).

The sequence clustering procedure can return either a set of clusters (TribeMCL) or a cluster hierarchy (CluSTr, ProtoMap). In the latter case, each level of the hierarchy represents a different degree of similarity between clustered proteins (see Fig. 2 for an example of a cluster hierarchy). Hierarchical clustering at all levels of similarity allows one to divine decisions about what degrees of similarity would result in biologically meaningful clusters (A biologically meaningful cluster is defined here as one corresponding to a family of homologous proteins, i.e. those that share a common evolutionary history.) from the clustering process itself. After an initial hierarchy of clusters is generated, other (divergent or even contradictory) methods may be used to detect the clusters and similarity levels that are biologically revealing. Once families of homologous proteins are identified, annotation associated with well characterized members of the family can be assigned to unknown family members. A brief comparison of selected clustering methods is presented in Table 1.

CluSTr classifies protein sequences from both the UniProt Knowledgebase, which includes all known protein isoforms (Apweiler et al., 2004) and the IPI (Kersey et al., 2004). At the time of writing, it covered around 720 000 protein sequences and 200 complete proteomes (including 11 eukaryotes). In order to keep CluSTr up-to-date with changes (New sequences being added to and existing ones being deleted from the proteome member sets.) to member sets of the represented proteomes, an efficient update pipeline was implemented and deployed.

CluSTr employs a fully automatic hierarchical clustering method based on a similarity matrix to yield hierarchies of protein families. In order to compute the matrix, first all-against-all pair-wise comparisons between protein sequences are computed using the Smith–Waterman algorithm, resulting in a set of similarities and their corresponding Smith–Waterman scores.

Second, a Monte-Carlo simulation is performed to assess the statistical significance of the Smith–Waterman scores, yielding Z-values (Comet et al., 1999). These Z-values are used as a measure of sequence similarity in the similarity matrix.

Given the set of similarities and their associated Z-values, a single-linkage clustering procedure (for example, see Fig. 1) takes place, which yields a hierarchy of clusters. When the hierarchy is traversed from children to parents, cluster sizes get progressively larger; the corresponding Z-values, and hence the similarity levels, get smaller. Clustering is performed in single-species groups and selected multi-species groups, e.g. ‘human and mouse’ and ‘all against all’ (the latter refers to clustering based on similarities involving all sequences that currently exist in UniProt Knowledgebase or IPI, i.e. it excludes any sequence that has been deleted from these databases). Each group is clustered into a separate hierarchy. In the case of species-specific groups, the corresponding hierarchy is a
importing a subset of the 'all against all' hierarchy from CluSTr into knowledgebases. Therefore, cluster members were cross-referenced with other available sources of reliable data. This was achieved by it being possible to mine for the common annotation and achieve sensible results. Only if the clusters are classifications of biological relevance is one of the sections to draw conclusions about the biological value of the CluSTr data content.

2 MATERIAL AND METHODS

This section describes the methodology that underpins the predictive power of CluSTr in greater detail. The procedure of quantifying the data mining potential of CluSTr using the UniProt automated annotation pipeline is presented and a set of cross-validation experiments is defined. These are used in later sections to draw conclusions about the biological value of the CluSTr data content.

2.1 The clustering algorithm

As mentioned in the introduction, the similarity matrix used in the clustering method uses Z-values as a similarity measure. The benefits of using Z-values in clustering are 3-fold.

First, Z-values estimate the statistical significance of Smith–Waterman scores. A Z-value of 8.0 has been shown (Comet et al., 1999; Bastien et al., 2004) to be a conservative estimate of the cutoff, above which the Z-value for a given similarity is not likely to be obtained by chance. Such cutoff (rounded to 10 for pragmatic reasons) is used in CluSTr to restrict the set of similarities taken into account when clustering.

Second, the Monte-Carlo simulation ensures that the derived Z-values depend only on the compared sequences and not on the size and composition of the sequence database (unlike similarity measures such as BLAST comparisons). This allows an incremental update of the CluSTr database by keeping all scores of unchanged sequences and only calculating ‘new-against-new’ and ‘new-against-unchanged’ similarities, thus avoiding time-consuming recalculations.

Third, Z-values have also been shown to be much less dependent on the lengths of the sequences than Smith–Waterman scores (Comet et al., 1999). Given the set of similarities and their associated Z-values, a single-linkage clustering procedure takes place (Fig. 1).

This procedure starts off by sorting protein sequence similarities by Z-value in descending order, and then creating singleton clusters for all proteins that take part in those similarities. The singletons are assigned the maximum level of similarity (identity) and are simply artefacts of the clustering procedure. Smaller clusters, with higher corresponding levels of similarity (Z-values), are then merged into bigger clusters, with lower Z-values. Under the principle of single-linkage, a new similarity $s_{ab}$ between sequences $a$ and $b$ at Z-value $Z_{ab}$ will result in the merging of clusters $C_a$ and $C_b$ (created in previous steps, at Z-value higher than $Z_{ab}$) into cluster $C_{ab}$ if sequence $a$ is a member of $C_a$ and sequence $b$ is a member of $C_b$.

$C_{ab}$ is then on referred to as a parent of $C_a$ and $C_b$, and $C_a$ and $C_b$ as children of $C_{ab}$. The process of creating parent clusters continues until the set of sequence similarities is exhausted.

The resulting hierarchy of clusters is a binary forest (Fig. 2) in which each parent has only two children, but where a number of parent-less clusters (a.k.a. ultimate predecessors) may exist.
Note that since all similarity levels are considered during the single-linkage clustering, deep hierarchies result, which make their traversal unwieldy via the CluSTr web interface. In addition, the sheer volume of data that results from a UniProtKB accession to cluster identifier mapping, needed for incorporating CluSTr data into the UniProt warehouse, motivated us to find a way of pruning the CluSTr hierarchy while keeping the loss of valuable information to a minimum.

An intuitive solution was chosen in order to obtain a slimmed down subset of CluSTr, referred to as CluSTr Slim (Fig. 3).

The principle behind pruning CluSTr relies on an observation that many of the large clusters in the CluSTr hierarchy are parents of one very small cluster and of another cluster that is almost as large as its parent. Arguably, such a split of the parent cluster is not revealing and contributes little information to that carried by the parent itself. Cases where the parent cluster splits into two children with comparable sizes are more interesting. The precise cutoff operation applied to prune CluSTr hierarchy was to eliminate clusters whose members formed 90% or more of the member set of their respective parents. Additionally, singletons and ultimate predecessors were pruned off. In total, 11% of non-singleton clusters were excluded from CluSTr Slim.

A confirmation of the value of pruning CluSTr using the above method was shown in a recent mapping from CluSTr to GO, performed via the manually curated InterPro to GO mapping (see http://www.ebi.ac.uk/GOA/).

The clusters to be mapped were selected from the full CluSTr set (which by definition subsumes CluSTr Slim), based on their degree of overlap with InterPro families and domains. Specifically, a cluster was selected for mapping if at least 70% of its member proteins also covered at least 70% of proteins in an InterPro domain or family (Fig. 4) Next, the GO terms, which had been assigned to that InterPro entry in the manually curated InterPro to GO mapping, were also assigned to the corresponding cluster.

All clusters mapped to GO using the above principle turned out to be in CluSTr Slim. This result confirmed that CluSTr Slim was both a pragmatic choice (thus restricting the amount of information to be processed during an automated annotation run) and a biologically pertinent one when it comes to comparing its predictive ability with that of InterPro.

2.2 Measuring the predictive power of CluSTr

It is no trivial task to present the data that underpin the quality of a data source. Especially for sets generated in a fully automated way such as CluSTr, the biological value is not obvious immediately. To make it explicit, the contents of the clusters in CluSTr were mapped to their biological relevance using...
The predictive power of the CluSTr database

Three tests were performed to analyse the predictive power of the CluSTr content and to compare it with other classification approaches available from InterPro. The main evaluation concerns the predictive power of an integrated approach, i.e. the usage of data from InterPro and CluSTr rather than the usage of data from InterPro alone. To make the results comparable, all tests were performed on the set of UniProtKB/Swiss-Prot proteins that are covered by CluSTr, i.e. all proteins from fully available literature curation process. Those are for instance keywords, descriptions and comments found in the KW, DE and CC lines of the UniProtKB/Swiss-Prot data. Core data types that were included in the data mining experiments are, the factual data types present for each protein, such as the organism from which it was extracted, the sequence of the protein, sequence signatures provided by InterProScan (Zdobnov and Apweiler, 2001) and the belonging of a protein to a cluster in CluSTr. Annotation data are the descriptive data types usually added by a literature curation process. These are for instance keywords, descriptions and comments found in the KW, DE and CC lines of the UniProtKB/Swiss-Prot entries, respectively.

2.2.1 Measuring the added value Both Spearmint and Xanthippe use the C4.5 algorithm to produce decision trees, which are then further processed to obtain annotation rules. Ideally, the algorithm is trained on 50–500 examples to have a solid statistical footing on the one hand and sufficient runtime performance on the other. Since many of the InterPro families and domains generally consist of protein sets of these sizes, they were chosen as fundamental training sets. It was shown in cross-validations that this approach leads to high quality predictions, which is only achievable due to the biological relevance of the InterPro families and domains. This is only an indirect method to access this relevance but it reveals a good estimation of the data quality contained in the individual protein classes.

An indication of the quality delivered by CluSTr is the amount and quality of the generated predictions, if it is used alongside the original data types (Table 2). To measure this, the Spearmint technology was used in three variations. First, all the data types that are usually exploited (Set 1) were used to generate a set of annotation rules. Second, these data types were combined with the CluSTr dataset (Set 2) to produce another set of annotation rules. Both sets were then cross-validated against UniProtKB/Swiss-Prot. Third, Pfam signature hits were chosen to be removed from Set 2 (obtaining Set 3). This allows a rough comparison of the value added by the CluSTr data to that added by Pfam, a commonly used and well-established dataset.

2.2.2 Measuring the rate of avoided errors With Xanthippe, a technology is available that detects errors in protein annotation. The method is used as a post-processing step on the predicted data and reduces the number of errors produced by automated annotation considerably.

Xanthippe uses the fact that most proteins are classified into more than one group. If an annotation is predicted based on one group, it checks whether there are conflicts with any of the other groups the protein belongs to. If there are, the predicted annotation is removed. To make this method work, the individual protein groups need to be clustered in the strict sense that all the members of a cluster share the same biological property or set of properties and all the non-members do not.

Xanthippe was generated in two modes, the original one using the InterPro data alone, and the extended one using CluSTr on top of it. Both these contradictory rule sets were then applied on the output from the original Spearmint method (Set 1) and on the extended Xanthippe method (Set 2).

It is shown that CluSTr also classifies in a strict sense, i.e. such that proteins outside a given cluster do not share the biological qualities of the ones within it, if the Xanthippe set that uses CluSTr is able to detect more annotation errors than the original one.

3 RESULTS

To evaluate the impact of the examined datasets to the performance of the data mining procedure, the values for false discovery rate (FDR) and recall were computed from a cross-validation against Swiss-Prot, where

\[ \text{Recall} = \frac{TP}{TP + FN} \]
\[ \text{FDR} = \frac{FP}{TP + FP} \]

TP is number of true positive predictions, i.e. overlaps between predicted and original UniProtKB/Swiss-Prot annotation. FP is number of false positive predictions, i.e. predictions not belonging to TP. FN is number of false negative predictions, i.e. original UniProtKB/Swiss-Prot annotation that could not be predicted. Recall and FDR values depend on each other, i.e. the higher the recall rate the higher the FDR and vice versa. By introducing a new core dataset to mine on, both values are likely to change simultaneously, which makes comparisons of performances difficult. To obtain a thorough analysis, the construction of an ROC curve would be required. This is an expensive procedure and would measure the performance of

![CluSTr cluster and InterPro Family or Domain](image)

Fig. 4. An example of a cluster chosen as suitable for mapping to GO according to the 70% rule. Note that the intersection covers at least 70% of members of both the cluster and the InterPro family or domain.

### Table 2. Core data types that were included in the data mining experiments

<table>
<thead>
<tr>
<th>CluSTr</th>
<th>Pfam</th>
<th>Original Data Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1</td>
<td>v</td>
<td>v</td>
</tr>
<tr>
<td>Set 2</td>
<td>v</td>
<td>v</td>
</tr>
<tr>
<td>Set 3</td>
<td>v</td>
<td>v</td>
</tr>
</tbody>
</table>

The original data types, i.e. Pfam and the types listed in the last column of the table, were used to produce Set 1. CluSTr was included in the mining process to produce Set 2 and Pfam was removed to obtain Set 3.
the predictors themselves. Hence, a simplified analysis was chosen, which will illustrate the quality of the CluSTr dataset. To be able to compare the results with each other, parameters were chosen to keep recall values constant and measure the behaviour of the FDR. Figure 5 shows the virtually constant recall rates for all the three training sets employed. Since statistical values differ between the individual data types, small variations are unavoidable. The set that excludes Pfam (Set 3) has an overall recall of 59%, while the others lie at ∼62%.

The main focus of this work is the examination of the CluSTr dataset. Figure 6 illustrates the drop of the FDR, once CluSTr data is taken into account. In total a drop of 1.9% was observed, which translates to an avoidance of 27% of all false predictions. The protein name annotation in the DE line of the UniProtKB/Swiss-Prot entries benefits most of the CluSTr integration with a drop of the FDR from 9.8% to 6.4% (34% of annotation errors are avoided).

Figure 7 illustrates a comparison of the behaviours of the set excluding Pfam (Set 3) and the set excluding CluSTr (Set 1) normalized on the set containing all data types (Set 2). Set 3 yielded a slight decrease of the FDR by 0.34%, of which erroneous EC predictions contribute the most. Over all, the inclusion or exclusion of Pfam as a core data type affects recall and FDR only marginally. If CluSTr is left out from the core dataset, the FDR increases by 36%. The most prominent result is the effect of the CluSTr dataset for protein name predictions (DE), where 54% of all the wrong predictions can be avoided.

Not shown are the influences of the CluSTr data towards the performance of the Xanthippe contradictive systems. The inclusion of this data type led to an increase of the error detection by 3%, while the rate of removed correct predictions remained constant. Again, protein name annotation benefited in particular with an increase of detected errors by 9.7% and a decrease of wrongly removed correct annotations by 33%.

Overall, predictions of the automated annotation pipeline benefit largely by the inclusion of CluSTr as a core data type. The recall rate increased slightly by 1.3%, while the impact on precision is noteworthy. More than 30% of all annotation errors can be avoided, as illustrated in Figure 8.

4 DISCUSSION

A considerable decrease in FDR produced by the Spearmint system was observed, when CluSTr data is included in the mining procedure. This indicates that the clustering algorithm employed...
by CluSTr is able to classify proteins more precisely than the combined approaches used in the InterPro member databases. This aspect is a clear proof that the CluSTr database contains data clusters of biological relevance, a fact that is exploited by our data mining procedures.

A direct comparison between CluSTr and Pfam, a well established and frequently used methodology, was performed. We found that in the employed environment the inclusion or exclusion of the Pfam method altered the predictive power only marginally. This can be attributed to the availability of methods like, for instance, Smart, PROSITE and PRINTS that cover similar aspects of the biological relevance contained in the Pfam data. They all start out from a semi-manual assembly and alignment of proteins sharing functional similarities. For this, the functional behaviour of the proteins in the training sets has to be well-characterized. CluSTr adopts an essentially different approach by not taking advantage of the available functional annotation of the protein data to initiate the clustering routine. This leads to a comprehensive overall coverage without a skew towards well-characterized groups. An interesting fact to mention in this context is that 16% of the CluSTr clusters exclusively contain proteins without any InterPro correlation. The results of our experiments prove CluSTr’s capability to produce valuable classifications. These go beyond what is covered by traditional methods, while the Pfam appears to have a significant overlap with other traditional methods.

The results of the experiments presented in this work are intended to show the value of the CluSTr approach. The collected data allow to draw qualitative conclusions only, the set-up was insufficient to allow quantitative analysis. This analysis is highly desirable to provide answers to the following questions:

- In which protein groups were similarities accessible through CluSTr but not through traditional methods?
- How does CluSTr data perform in well-characterized groups, i.e. does it pick up biological relevance in a comparable way to traditional methods?
- How does the CluSTr perform in recall rates when the FDR is kept constant, i.e. are there other biological functions detected that are inaccessible using traditional methods?

5 CONCLUSIONS

The data shown was derived from the CluSTr ‘all against all’ dataset. This set contains all proteins of fully sequenced proteomes and is hence not comprehensive to all proteins available from UniProt.

We are planning to increase the coverage of CluSTr to all sequences from UniProt.

The use of CluSTr data for mining purposes improves the performance of automated annotation systems considerably and hence we intend to include this dataset into the automated annotation production system.

ACKNOWLEDGEMENTS

The CluSTr project was funded by the European Commission as the TEMBLOR, contract-no. QLRI-CT-2001-00015 under the RTD programme ‘Quality of Life and Management of Living Resources’. The automated annotation work was supported by the National Institutes of Health (NIH) grant 1 U01 HG02712-01.

Conflict of Interest: none declared.

REFERENCES


