BIOINFORMATICS

Vol. 30 ECCB 2014, pages i624–i630
doi:10.1093/bioinformatics/btu478

Entropy-driven partitioning of the hierarchical protein space

Nadav Rappoport[1], Amos Stern[1], Nathan Linial[1] and Michal Linial[2,*]

[1]School of Computer Science and Engineering and [2]Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, 91904, Israel

ABSTRACT

Motivation: Modern protein sequencing techniques have led to the determination of >50 million protein sequences. *ProtoNet* is a clustering system that provides a continuous hierarchical agglomerative clustering tree for all proteins. While ProtoNet performs unsupervised classification of all included proteins, finding an optimal level of granularity for the purpose of focusing on protein functional groups remain elusive. Here, we ask whether knowledge-based annotations on protein families can support the automatic unsupervised methods for identifying high-quality protein families. We present a method that yields within the ProtoNet hierarchy an optimal partition of clusters, relative to manual annotation schemes. The method's principle is to minimize the entropy-derived distance between annotation-based partitions and all available hierarchical partitions. We describe the *best front* (BF) partition of 2 478 328 proteins from UniRef50. Of 4 929 553 ProtoNet tree clusters, BF based on Pfam annotations contain 26 891 clusters. The high quality of the partition is validated by the close correspondence with the set of clusters that best describe thousands of keywords of Pfam. The BF is shown to be superior to naïve cut in the ProtoNet tree that yields a similar number of clusters. Finally, we used parameters intrinsic to the clustering process to enrich a priori the BF's clusters. We present the entropy-based method's benefit in overcoming the unavoidable limitations of nested clusters in ProtoNet. We suggest that this automatic information-based cluster selection can be useful for other large-scale annotation schemes, as well as for systematically testing and comparing putative families derived from alternative clustering methods.

Availability and implementation: A catalog of BF clusters for thousands of Pfam keywords is provided at http://protonet.cs.huji.ac.il/bestFront/

Contact: michall@cc.huji.ac.il

1 INTRODUCTION

The explosive growth in the number of sequenced proteins is mostly a result of the breakthroughs in sequencing technologies and the corresponding sequencing of hundreds of organisms. Despite these advances, the structure and function of most of these proteins remains unknown. The most successful method for functional annotation of proteins is by sequence alignment, homology detection and inference techniques. Generalization of such approaches calls for charting the protein space by clustering or classification (Radivojac *et al.*, 2013). If successful, each such group of proteins would represent a 'family'. Classification into families is a critical component in structural and functional genomics. No accepted consensus exists for how many of these protein families might comprise the entire protein-space (Coordinators, 2014). There are ∼30 000 main orthologous groups (Fischer *et al.*, 2011) in addition to rare and peculiar single proteins. With the increased number of complete proteomes, a phyletic partition shows that thousands of families are associated with each multicellular organism. Various expert-based databases provide a good description of protein families (Mi *et al.*, 2013; Punta *et al.*, 2012). For example, InterPro (IPR) is composed of 25 000 models for families and domains and from a structural perspective, there are ∼2600 superfamilies according to CATH classification (Sillitoe *et al.*, 2013).

Classifying the entire protein space into families serves not only as a method for large-scale protein annotations but also to support functional and structural genomic initiatives (Loewenstein *et al.*, 2009). Some prominent examples for protein classification are SYSTERS (Liu and Rost, 2003), CluSTr (Petryszak *et al.*, 2005) and ProtoNet (Rappoport *et al.*, 2013). The shared theme of all these resources is the hierarchical nature of the protein families. Furthermore, while all use BLAST-based statistical distance metrics for the clustering, the implementation, sensitivity, the notion of the distance metrics and consequently the depth of the hierarchical representations are different for each of the underlying algorithm and resource (Liu and Rost, 2003).

An important differentiating factor between various classification systems is the level of granularity. More often than not, definitions regarding the granularity are made arbitrarily, not based on natural or systematic considerations. The importance of granularity stems from the inherent diversity of the protein space. While some families are well defined by their sequence (e.g. Rubisco, Ribosomal proteins), the boundaries of many other protein families are blurred. To complicate things even further, some families may be best defined by a composition of several subfamilies, while others are part of large and multifunctional superfamilies (e.g. AAA superfamily, G protein coupled receptors, protein kinases). Gene Ontology database (GO) (Gene Ontology Consortium *et al.*, 2013) and to some extent IPR (Radivojac *et al.*, 2013) incorporate the notion of parent/child relationships for gene/protein family in cases where such relationships are accepted. For example, the 'Neurotransmitter-gated ion-channel (IPR006201)' is a parent of '5-hydroxytryptamine 3 receptor (IPR008132)', which is itself divided into two related subfamilies of A and B subunits (IPR008133 and IPR008134, respectively).

To tackle the issue of varying granularity, we suggest a method that searches among a hierarchical tree for the 'front' of clusters that have a minimal entropy-based 'distance' from the optimal annotation-based partition. The method can be used for seeking the best matching partition, relative to any existing classification system. In our work, we assess the quality of clusters of a construction of the ProtoNet system, which is a fully automatic tree of protein sequences (Rappoport *et al.*, 2013).

*To whom correspondence should be addressed.

We choose for this purpose Pfam, an extensive documentation of domains and families that represents one of the most reliable sources for protein families (Finn *et al.*, 2014). Pfam is a semiautomatic family database with a strict quality control. It contains ~14 000 models for protein families and covers almost 80% of known protein sequences with ~50% coverage of amino acids. As such, Pfam is an extremely valuable resource for addressing questions concerning the quality of structural and functional protein families.

The classification provided by ProtoNet is based on a bottom-up unsupervised agglomerative hierarchical clustering (Kaplan *et al.*, 2005; Sasson *et al.*, 2003). Specifically, it provides a full range of cluster granularity; from single proteins to huge clusters that carry minimal biological coherence (the root clusters). It is possible to match the majority of Pfam families to specific clusters within the ProtoNet clustering. Moreover, we search for 'natural' clustering partitions, from which the optimal granularity can be inferred. The intuition for such a natural partition is for it to be 'entropically close' to a partition based on Pfam annotations of the proteins.

In this article, we describe the theoretical basis for a novel entropy-based procedure for *best front* (BF) searching. To allow for a priori prediction of the BF, we introduce additional intrinsic clustering parameters that partially separate the entire set of clusters present and those that are included in the BF. We show that a combination of two such parameters is enough to strongly enrich the BF from all clusters. Our method to identify optimal granularity allows for automatic and systematic definition of the set of proteins that correspond to an orthologous family. Such automatic definition could supplement current techniques in genome-wide annotation projects, which are mostly based on expert annotation (Barker *et al.*, 2001). Furthermore, it will serve to define families using the matching of a large number of classification systems such as MetaFam (Silverstein *et al.*, 2001), Superfamily (Wilson *et al.*, 2009) and more.

We show that the automatic information-based cluster selection of the BF is extremely useful for a systematic comparing of clustering methods. Furthermore, the supervised approach can be applied to biochemical functions (such as enzyme classification), as well as for structural superfamilies. We provide a complete catalog of the BF clusters for >10 000 IPR keywords.

2 METHODS

2.1 Databases and sources

2.1.1 ProtoNet tree The tools and methods described in this article were applied to the ProtoNet protein classification system. ProtoNet implements agglomerative hierarchical clustering using several merging strategies. For the sake of simplicity, we choose to discuss only one of the merging strategies offered by ProtoNet, called the 'Arithmetic' merging strategy (Sasson *et al.*, 2003). ProtoNet (version 6.1) provides a classification hierarchy that covers ~9 000 000 proteins from the UniProtKB database (release 15.4). We have not discussed the expanded version of ProtoNet that covers ~20 million proteins (Rappoport *et al.*, 2013). We clustered 2 478 328 representative proteins as defined by UniRef50. In the clustering tree, there are 4 929 553 clusters and 27 103 roots (mostly singletons). ProtoNet is available at http://www.protonet.cs.huji.ac.il.

2.1.2 ProtoNet clustering measurements The agglomerative hierarchical clustering scheme defines a set of terms that are intrinsically associated with the process. In such a scheme, each cluster is created from smaller clusters, which are captured as its descendants in the clustering tree.

ProtoLevel (PL) ranges from 0 to 100 and is used as a standard quantitative measure of the relative height of a cluster in the merging tree. Indirectly, the PL of a cluster reflects the global average of the sequence similarity BLAST E-score between proteins in the cluster. The PL of the leaves of the tree is defined as 0, whereas the PL of a root equals 100. The larger the PL, the later the merging that created the cluster took place. Therefore, the PL scale is considered as an 'internal timer' of merges during the clustering process.

The *lifetime* (LT) of a cluster is the difference between PL at its creation and its termination. The LT of a cluster reflects its remoteness from the clusters in its 'vicinity'. Explanations for additional terms that describe the clustering process such as depth, connectivity and compactness are available on the ProtoNet Web site (see above).

2.1.3 Protein family annotations A total of 10 337 annotations from Pfam and a total of 11 327 keywords from IPR Family were used as external protein family annotation sources. Of the 2 478 328 UniRef50 proteins, 50% have at least one Pfam annotation.

2.2 Keyword correspondence scores

To measure the correspondence between a given cluster and a specific annotation, we define the notion of a *correspondence score* (CS). The CS for a certain cluster C and a given keyword K measures the correlation between the cluster and the keyword, using the well-known intersect-union ratio:

$CS(C, K) = |c \cap k|/|c \cup k| = TP/(TP + FP + FN)$ Where c is the set of annotated proteins in cluster C, and k is the set of proteins annotated with the keyword K.

TP, FP and FN stand for true positives, false positives and false negatives, respectively.

- TP is the number of proteins in cluster C that have keyword annotation K.
- FP is the number of annotated proteins in cluster C that do not have keyword annotation K.
- FN is the number of proteins not in cluster C that have keyword annotation K.

The cluster receiving the maximal score for keyword K is considered to be the cluster that best represents K within the ProtoNet tree. The score for a given cluster on keyword K ranges from 0 (no correspondence) to 1 (a cluster containing exactly all of the proteins with keyword K maximally corresponds to the keyword).

For annotation keywords from several external sources, we define the cluster with the best CS for each keyword as the *best cluster* for this keyword.

The sources used for defining the best clusters as well as their CS are IPR (families and domains), Pfam, SCOP (fold, superfamily, family and domain levels), GO (in three categories—molecular function, cellular process and cellular localization), CATH (architecture, class, homology and topology) and ENZYME (four levels of the EC hierarchy). For a detailed description on the database and structure of the annotation sources see (Rappoport *et al.*, 2012). For simplicity, we describe in details only the results of the information-theoretical method for Pfam and IPR.

An interactive table is available at www.protonet.cs.huji.ac.il/best_cluster/

2.3 Information-theoretic approach for searching optimal protein partition

For a given keyword annotation type, we would like to find, within the hierarchical ProtoNet tree, the set of clusters whose partition of the set of annotated proteins maximally corresponds to the partition of the proteins induced by their annotations. We use an information-theoretic–based approach to find the set of protein-disjoint clusters having the minimal 'distance' from the keyword-induced partition of the proteins. Each protein in the system has ≥ 0 keyword annotations, as defined by a given external source.

Let P be the set of all proteins in the system, KW the set of all keywords in the system and CL the set of all ProtoNet clusters.

For a protein $p \in P$, define $k(p)$ as the set of all $k \in KW$ s.t. k annotates p.

We thus define the following:

$$\Omega = \{(k, p) | p \in P, k \in KW\}$$

Ω = set of couples (keyword, protein) where the protein p has the specific annotation k.

c = the set of the proteins in cluster C.

f = a given front in the ProtoNet tree

= a protein-disjoint set of ProtoNet clusters that together cover the whole space

= a partition of the protein space.

We define the mapping function: $C_f : \Omega \to CL$ as

$C_f(k, p) =$ the unique cluster of p in f

And define $K : \Omega \to KW$ as $K(k, p) = k$.

These are the projections on the first and second coordinate, respectively.

Our underlying probability distribution is uniform:

$$\forall (k, p) \in \Omega : \Pr\,(k, p) = 1/|\Omega|$$

Thus for any $k \in KW, c \in CL$:

$$P(K = k, C_f = c) = \begin{cases} \sum\limits_{(k,p) \in \Omega, p \in c} \Pr(k, p) = \dfrac{\sum\limits_{(k,p) \in \Omega, p \in c} 1}{\Omega}, & c \in f \\ 0, & \text{otherwise} \end{cases}$$

We then define, using the Rokhlin metric (Katok, 1995), the 'distance' between the keyword-induced partition (K) and the partition defined by f (C_f) as

$$\Phi(K, C_f) = H(K|C_f) + H(C_f|K)$$

$H(Y|X)$ is the conditional entropy of Y given X, defined as

$$-\sum_{x,y} p(X = x,\ Y = y) \log\,(p(Y = y | X = x))$$

Minimizing this distance captures the intuition of attempting to find a generalized 'equivalence' between specific protein clusters in f and specific keyword annotations.

We then attempt to find the optimal ProtoNet front f^* such that

$$f^* = \arg\min{}_f \Phi(K, C_f)$$

We used an algorithm to calculate f^* ('the best front' = BF) from the leaves of the tree upward, using the fact that the score of a front f is simply the summated score contributions of its member clusters. The algorithm has linear ($O(n)$) time complexity in the number of clusters in the tree.

Note that any cluster consisting solely of non-annotated proteins are transparent to this method due to the fact that all such cluster's proteins have an effective probability of 0.

Once the BF has been calculated, we can analyze in several ways. We can compare its clusters' CS to individual keyword annotations to see which of the clusters in the front are practically equivalent to a specific annotation. This is done for each keyword using the above-defined *keyword CS*. We also compare its maximally obtained scores to the maximum scores in the whole ProtoNet tree. In addition, we look for intrinsic clustering parameters that separate the BF clusters from non-BF clusters.

3 RESULTS

3.1 ProtoNet clusters assignment as Best Global, Best Front and Best Cut

ProtoNet relies on unsupervised automatic agglomerative clustering method. Figure 1 illustrates the scaffold of the ProtoNet tree. The leaves (i.e. individual proteins) represent the UniRef50 representative proteins. The term 'Best Clusters' is assigned for clusters according to a specific annotation resource (e.g. Pfam) and its keywords. We label clusters as 'best' according to the maximal CS. For each keyword the CS ranges from 0 to 1.0 (see Methods, Section 2.2). We used three partitions and definitions throughout the analysis: (i) Best Global (BG) clusters, covering the best CS cluster among all ProtoNet clusters; (ii) Clusters that have a maximal CS at a predetermined PL [Best Cut (BC) cluster] and (iii) the newly developed entropy-based partitions, called the BF clusters.

Figure 1 shows that some clusters may be 'best' for more than one keyword. Evidently, for a specific keyword the BG can be assigned to both parent and child clusters but this is not legitimate for the BF and BC clusters.

3.2 Automatic identification of a BF for Pfam-annotations

Using the information-theoretic entropy-based algorithm described above (see Methods), a BF was created for Pfam annotations. A total of 26 891 clusters (including 2122 singletons)

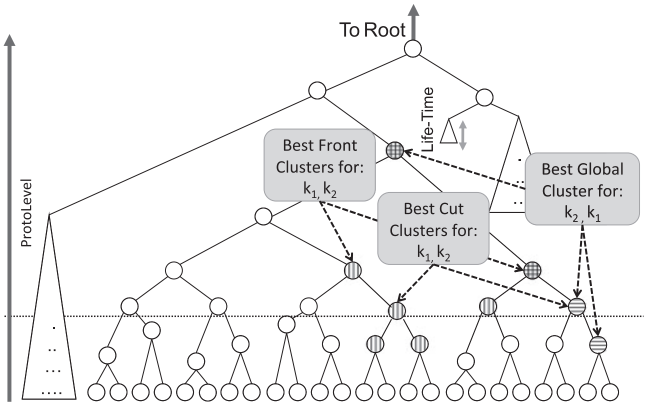
To Root
ProtoLevel
Life-Time
Best Front Clusters for: k_1, k_2
Best Cut Clusters for: k_1, k_2
Best Global Cluster for: k_2, k_1

Fig. 1. ProtoNet clusters assignment. An illustration for 'best' clusters according to BF, BG and BC clusters. The best clusters are defined according to CS (see Methods). Note that the BG Clusters will have the maximal CS values. However, the BF and BC clusters are restricted by the partition of proteins into disjoint clusters. Some clusters maybe 'best' for more than one keyword. The illustration is according to a specific annotation resource (e.g. Pfam). The same scheme applies for other annotation resources [e.g. the Homology level in CATH classification (Cuff *et al.*, 2009)]. The PL (ranges 0–100) and the Life Time (LT, range 0–100) are internal measurements of the ProtoNet tree

were defined as the BF. This number reflects an average compressing factor of 92 from the number of initial proteins included in the analysis. Recall that using the UniRef50 protein marked a 5.7 compression level on average (not shown). Thus, the BF partition of UniRef50 sequences yields an effective compression of 300–500 folds. In addition, the number of clusters in the BF is only 0.5% of the original number of clusters. These 26 981 clusters contained 2 119 556 of the 2 478 328 proteins. Of these 2 119 556 BF proteins (86% of all proteins), 17 237 proteins (0.8%) had no annotations.

The family size and features of homologous protein families have been extensively investigated. In Figure 2, the size distributions of the clusters in the BF and the keyword groups in IPR family. The two size distributions are similar, but the groups of IPR keywords tend to be slightly smaller on average. Almost all of the clusters in the BF contained ≤500 proteins (Fig. 2). The paired *t*-test of sizes' distribution does not reject the null hypothesis (*P*-value = 0.7). The intersection between the clusters of IPR BG clusters and the BF clusters is surprisingly low (only 623 clusters, 12% of the IPR clusters). While the size of the cluster is a good indicator for overall familial correspondence, it does not provide any direct information on the purity and quality of the BF clusters.

3.3 Quality assessment of BF clusters

In assessing the resulting BF on a larger scale, we compared the maximal CS within the BF to the maximal CS for all ProtoNet clusters, for each IPR family keyword annotation.

Such a comparison can indicate how well the hierarchical tree was 'compressed' into the BF, vis-à-vis its performance for each IPR keyword. Such compression, if it is sufficiently informative in terms of its CS on specific keywords, is obviously more useful than a global search for each keyword's maximum scoring cluster (defined as BG clusters). The former yields a disjoint partition of the protein space that does not act as an essential constrain for the BG maximum scoring CS clusters (Fig. 1).

The BG maximal CS distributions for ProtoNet clusters and the BF clusters are shown (Fig. 3). A leftward shift of the score distribution is observed, due to the fact that the maximal scoring cluster is not necessarily included in the BF. However, since the shift is modest, it is also apparent that the success of the BF, as measured by its individual CS values, is high. The high match of the BF clusters with the IPR keywords applied for many of the IPR keywords, indicating that the BF retained significant knowledge regarding the nature of the individual IPR keywords. Weighting the CS by the size of the clusters suggests that the average CS of the BF cluster is higher than the non-weighted CS average value. The average maximal CS was ~0.71 and the statistical significance for achieving such correspondence is *P*-value < 1E-300.

A close inspection of hundreds of examples of protein families revealed that many of the BF clusters merges with the maximal correspondence-scoring (BG) clusters for those families. For example, cluster 4802079 corresponds to 'BolA-like protein' keyword (CS = 0.98). Cluster 4768763 corresponds to the 'Cysteine dioxygenase type I' keyword (CS = 0.99). Such examples were relatively abundant, indicating that the BF contained clusters that, in addition to being part of the optimal global solution for the BF, were optimal for a specific keyword.

Figure 3 shows a quality assessment of the BF clusters for all 5150 IPR clusters (a minimal size of 20 proteins for each IPR keyword). We compared the CS value relative to BG clusters that are not restricted to any partition of the tree and in principle can have a maximal CS occurs in a parent–child hierarchy. Interestingly, for the BF clusters, the dominating bins cover the CS > 0.9. In this bin there are 52 and 73% of the BF and BG clusters, respectively.

3.4 A comparison of the BF to a PL-defined partition

As shown before, the BF is only a small subset of the ProtoNet clusters (0.5%).

To compare an alternative protein partition offered by a different partition in ProtoNet, we compared the partition

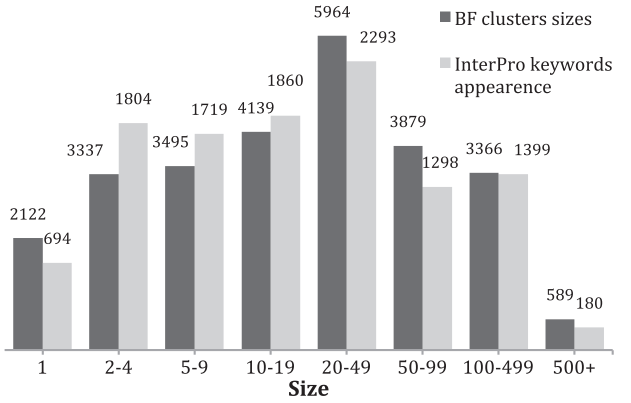

Fig. 2. Distribution of cluster sizes in the BF (dark gray) relative to the distribution of IPR (light gray) family keyword sizes (columns' heights are normalized). Most clusters in the BF contain ~10–100 proteins (the mean and median cluster size is 78.8 and 21, respectively). The mean and median sizes of the number of protein appearances in IPR keyword groups are 63 and 16, respectively

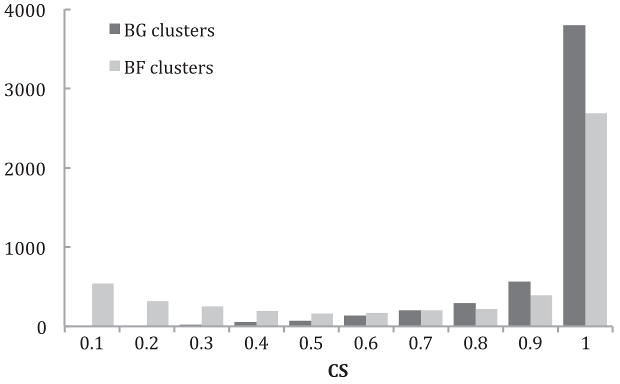

Fig. 3. Distribution of the BG clusters (dark gray) with a maximal CSs from the entire ProtoNet tree for each of the IPR family keywords (having ≥20 appearances). The (unweight) average maximal CS was ~0.91. The distribution of the maximal CS of all BF clusters (light gray) for each of the IPR family keywords (having ≥20 appearances) is shown. The average maximal CS for the BF was ~0.71

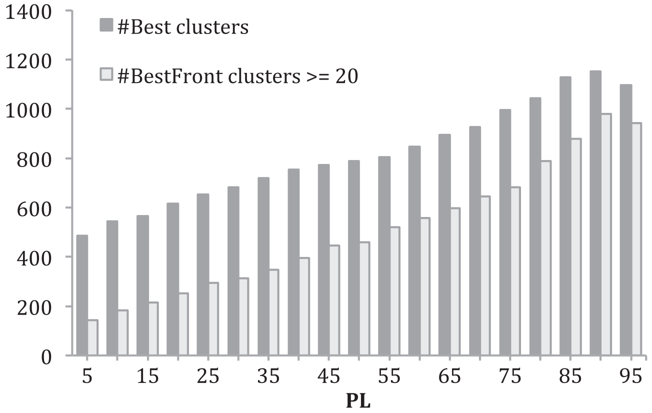

Fig. 4. The distribution of the BG clusters with at least 20 appearances (dark gray) and BF (light gray) clusters that include at least 20 proteins according to the PL (X-axis). The BF clusters show a similar distribution along the entire scale of PL values. PL90 covers the highest number of BG clusters. Evidently, the performance of the BF clusters on any keyword would always be bounded by the performance of all ProtoNet clusters

(i.e. ‘cut’) that have the largest intersection with the best clusters for IPR family. Figure 4 shows that the BC merges with PL 90 (i.e. the clusters were created before PL90 and were terminated after PL90). The set of clusters in PL90 achieves the highest median CS (0.89) for IPR Family keywords and the value is substantially lower for the IPR keywords that cover domains (32% of all keywords). Among all the BF clusters, 3071 (979 clusters with size ≥20) are also included in PL90. Hence, it might be expected, due to its similar number of clusters and ~22% overlap rate, that the performance of the maximally correspondence-scoring clusters in PL90 would be similar to the performance of the maximally correspondence-scoring clusters in the BF. However, for most individual keywords, the maximally scoring BF cluster has a score equal to or greater than the maximally scoring PL90 cluster, with a median of 0.91 versus 0.89 for BF clusters and PL90 clusters accordingly (*t*-test *P*-value of 3E-27), and median sensitivity of 0.99 versus 0.98 (*t*-test *P*-value of 4E-193). We compared the performance (measured by the CS or the sensitivity) of PL90 BC and the BF clusters (Fig. 5, top). The density of points above and around the bounding line indicates that for many keywords the BF corresponds to the keyword outperforms the maximally scoring cluster of BC of PL90.

Figure 5 (bottom) shows that a complex IPR keywords of ‘kinase’(141, total of 217 000 annotated proteins) resulted in average CS of 0.68 relative to 0.59 that is associated with the BC of PL90. The other ‘cuts’ (PL80, 85, 95) performs lower than that for the BC partition.

Table 1 shows a sample of IPR keywords where the CS of the BF clusters is substantially higher than the best cluster in the PL90 cut. There are >120 clusters (covers 44 000 proteins) for which the improvement of the CS of BF relative to the BC is substantial (additional of >0.3 units of the CS). Note that the BF clusters are substantially small when compared with the BC clusters (Table 1). Capturing the keyword accurately by the BF is

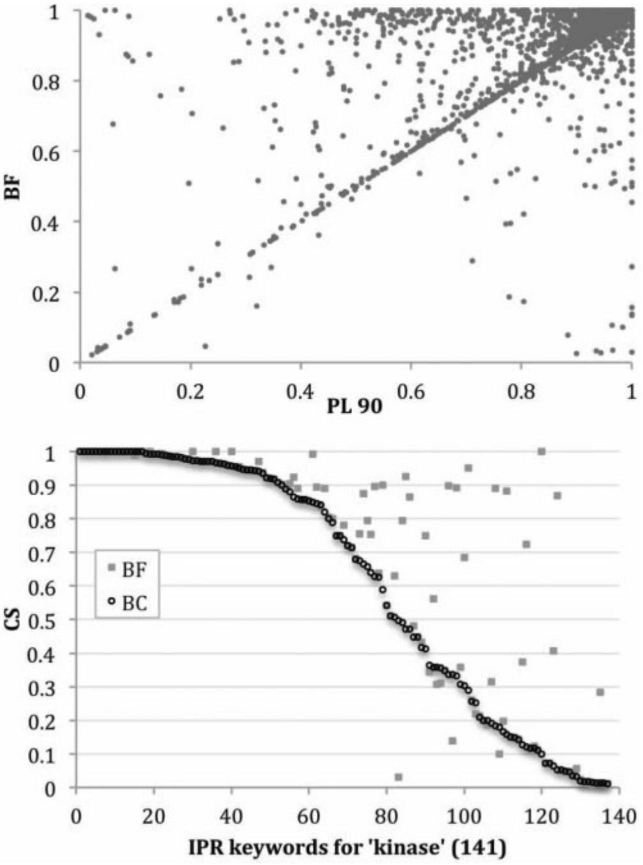

Fig. 5. Comparison of scores of the BC and the BF clusters. We compared the sensitivity scores (ranges 0–1) for ProtoLevel 90 (PL90) for the BC in view of the BF clusters (top). For each of the IPR family keyword (5131 mapped, above 20 appearances), the maximal sensitivity of PL90 clusters is plotted against the sensitivity of the cluster with maximal CS among the BF clusters. Among all the mapped IPR keywords, 141 belong to families of kinases (bottom). All the 141 kinases families are sorted according to the maximal CS of the BC clusters (PL90). Note the improvement in CS values for the BF clusters. Maximal gain in CS corresponds to high-quality clusters (CS > 0.5)

Table 1. Maximal score (CS) for a sample of IPR keywords for BF and BC clusters (sample of 15 IPR keywords)

IPR keyword	BF score	BF size	BC score	BC size
PcrB-like protein	1.00	50	0.26	294
Non-structural protein NS1, parvovirus	1.00	38	0.47	78
Type-IV secretion system protein TraC	1.00	47	0.39	330
Cyanophycin synthetase	1.00	35	0.16	299
Protein of unknown function DUF191	1.00	35	0.34	112
Ferredoxin-like, FixX	1.00	32	0.04	3245
Very-long-chain 3-ketoacyl-CoA synthase	1.00	51	0.13	543
AbgT putaLve transporter	1.00	57	0.30	188
Glycoside hydrolase, family 63	1.00	45	0.25	237
Herpesvirus capsid shell protein VP19C	1.00	60	0.39	10
Short chain fatty acid transporter	1.00	28	0.15	188
Nickel insertion ATPase/GTPase, CooC type	1.00	87	0.08	3136
Glutathione synthetase, prokaryotic	1.00	42	0.20	299
Deoxyribodipyrimidine photolyase-related	1.00	53	0.29	433
Imidazole glycerol phosphate synthase, H	1.00	170	0.34	545

best explained by matching with the subfamily level of the keywords.

3.5 Intrinsic features of the hierarchical clustering

Once we assessed the quality of the BF based on its CS performance for IPR keywords, we tested whether the clusters of the BF could be characterized by intrinsic parameters of the clustering process.

The underlying motivation is that ProtoNet is created by a fully automatic procedure with no prior knowledge of the protein features or annotations, besides their pairwise alignment scores. But, as a high percentage of the clusters in ProtoNet contain at least one of the 10 337 Pfam family annotations (50% of proteins have at least one such annotation), it would be desirable to characterize a set of clusters, in terms of their 'expected correlation' to an annotation source such as Pfam, by learning the intrinsic features of the clustering procedure. Such learned features could be used to predict, a priori the most likely front that best describes Pfam (or any other selected knowledge-based annotations). Such a compressed representation of the hierarchical clustering would simplify the protocol for restricted search of informative clusters in the hierarchical protein space.

An intrinsic value of the clustering process is a cluster's Life Time (LT, see Methods). We tested the feature of the LT in view of the set of BF clusters as well as the entire ProtoNet tree. Figure 6 quantifies the observation that the BF clusters can be partially inferred from their LT. For example, at a LT of 0.5: only 29% of all clusters have LT > 0.5; however, 59% of BF clusters have LT > 0.5. Clearly one such clustering parameter cannot replace the knowledge-based information. We therefore search an additional parameter that corresponds with the BF clusters. A useful parameter in defining the BF is the cluster size. The intrinsic ProtoNet parameters include the combination of the threshold of LT > 0.5, with a minimal number of proteins in a cluster be >10.

Applying the two thresholds to all clusters yielded 269 233 clusters, while application to the BF clusters yielded only 6609 clusters. The original fraction of the 26 981 BF clusters among all clusters is a mere 0.5%, while the fraction following the minimal thresholds of cluster size >10 and LT >0.5 is 3.5%. Thus, using these internal parameters we significantly enriched a set that includes many of the BF clusters. In fact, each cluster is associated with a larger set of parameters, such as its size, depth, LT, compactness and connectivity. The idea that a combination of such features could characterize a valid robust cluster is appealing and was extensively used to prune many of the less informative clusters (Rappoport *et al.*, 2013).

4 DISCUSSION

The novel method described here offers a systematic framework to address the quality of protein classification systems. While we use ProtoNet as our test bed, it is important to note that the tools and methods presented here are by no means limited in their scope to ProtoNet. These can be tailored to other hierarchical classification systems.

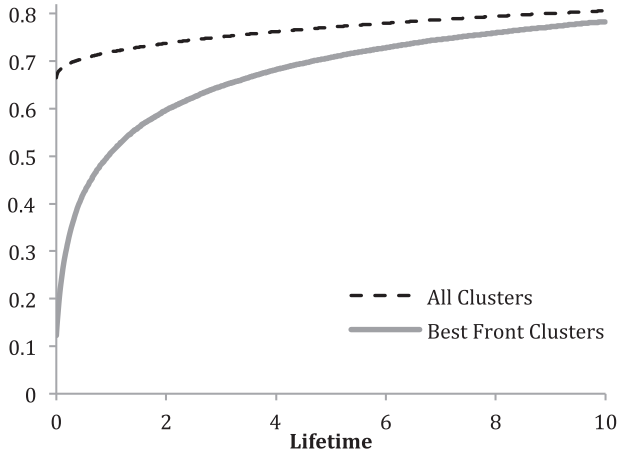

Fig. 6. A cumulative frequency cluster LT histogram. LTs are shown for clusters in the Pfam-based BF (solid line), and for all clusters within ProtoNet (dashed line). Note that while >80% of all clusters are below a LT of 0.3, only <1% of clusters of the BF are characterized by such a LT

The procedure is valid and expandable to other types of annotations. It can possibly serve to suggest the BF for structural and functional annotations (such as SCOP, CATH). However, it must be noted that the procedure is currently only suited to annotation types with approximately one level of granularity. Application of the method to a notably hierarchical annotation type, such as GO protein annotations, has determined (unpublished results) that the BF is the root containing all the proteins. This is due to the fact that many GO annotations are not separable; i.e. there is a large, inherent, minimal level of 'entropy' within the GO annotations [see discussion in (Radivojac *et al.*, 2013)].

Our work shows that application of the annotation-based optimal partitioning procedure to the ProtoNet tree yielded a highly compressed number of protein clusters, which are often highly correlated with individual IPR keywords. In addition, we showed that an inherent property of clustering, the LT, could be used to remove clusters that will probably not be part of the BF partition. The combination of these two results suggests the possibility of an automatic classification method that would be capable of reconstructing the vast majority of Pfam/IPR knowledge (in terms of family boundaries). An example for such automatic support could be submission of novel unannotated sequences to the ProtoNet BF partition, and assigning (with a certain probability) the sequence to one of the BF clusters (each essentially equivalent to a certain protein family).

An issue not yet addressed is the application of the method to a general set of protein annotations. The distance measure defined here has a theoretical minimum of 0; however, this value will not be the true lower bound for most annotation systems, as even having 1 protein with 2 different annotations (perhaps due to its being composed of multiple non-overlapping domains) will not allow the $H(K|Cf)$ term to go to 0, for any partition of the proteins. Allowing the splitting of a protein into its domains could address this problem. Additionally, calculation of the minimal theoretical value of the distance could be used to put the BF score in the context of its feasible values, implicitly part of the annotation system used. This theoretical lower bound essentially defines the 'entropy' inherent in an annotation system.

The ProtoNet tree that is the basis for the information-theoretic approach is based on UniRef50. This implies that each cluster maybe weighted by hundreds of sequences. For example, in mammals there are 1036 protein sequences named 'hemoglobin' and 3625 named 'hydrolase'. These sets in UniRef50 are compressed to 62 and 840 representatives, respectively. Therefore, the current analysis for ProtoNet that is composed of 2.5 million representatives is challenging with respect to the task in this study. However, extending the analysis for the expanded version of UniProtKB meets the limitation of computation feasibility (not shown). The current UniProtKB release 3_2014 already contains 54 million sequences and ∼10 million in UniRef50. The statistical- and information-driven method presented allowing a rational navigating in such large resource. In future work we will further validate BF hierarchies using statistical learning methods. Such tests are meant to reflect how well the BF would perform in an unsupervised setting. Preliminary results (not shown) indicate that the BF method is both robust and valid.

The main drawback of the approach described above is its need for externally defined protein annotations. Therefore, we would like to be able to learn the intrinsic parameters of a given system that predict, with minimal error, the clusters in a characteristic BF. Fortunately, the amount (and quality) of protein annotation is rapidly increasing, with Pfam already including ∼14 200 annotations. The information-theoretic approach developed here can only benefit from such trends, without sacrificing computational efficiency or accuracy.

ACKNOWLEDGEMENTS

The formulation of the entropy-based method was initiated by Noam Kaplan and Menachem Fromer. The concept of an entropy-based partition was carefully tested by Menachem Fromer, and the performance of the best-front was also determined by him for an early version of ProtoNet that was based on sequences from SwissProt (114K proteins). The current research uses the concept and the working flow for the most recent version of ProtoNet. We thank the ProtoNet team and Solange Karsenty for design and operation of the Web site. We thank Samuel Sayag for developing updates to the platform. We thank Dan Ofer for critical reading and suggestions. This research is partially supported by ERC grant (to N.L). The initial phase of this research was supported by Prospects, the EU FRVII consortium.

Conflict of Interest: none declared.

REFERENCES

Barker,W.C. *et al.* (2001) Protein information resource: a community resource for expert annotation of protein data. *Nucleic Acids Res*., **29**, 29–32.

Coordinators,N.R. (2014) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **42**, D7–D17.

Cuff,A.L. *et al.* (2009) The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res*., **37**, D310–D314.

Finn,R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res*., **42**, D222–D230.

Fischer,S. *et al.* (2011) Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr. Protoc. Bioinform*., **Chapter 6**:Unit 6.12.1–19.

Gene Ontology Consortium *et al.* (2013) Gene Ontology annotations and resources. *Nucleic Acids Res*., **41**, D530–D535.

Kaplan,N. *et al.* (2005) ProtoNet 4.0: a hierarchical classification of one million protein sequences. *Nucleic Acids Res*., **33**, D216–D218.

Katok,A. and Hasselblatt,B. (1995) *Introduction to the modern theory of dynamical systems*. Cambridge, Cambridge.

Liu,J. and Rost,B. (2003) Domains, motifs and clusters in the protein universe. *Curr. Opin. Chem. Biol*., **7**, 5–11.

Loewenstein,Y. *et al.* (2009) Protein function annotation by homology-based inference. *Genome Biol*., **10**, 207.

Mi,H. *et al.* (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res*., **41**, D377–D386.

Petryszak,R. *et al.* (2005) The predictive power of the CluSTr database. *Bioinformatics*, **21**, 3604–3609.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res*., **40**, D290–D301.

Radivojac,P. *et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods*, **10**, 221–227.

Rappoport,N. *et al.* (2012) ProtoNet 6.0: organizing 10 million protein sequences in a compact hierarchical family tree. *Nucleic Acids Res*., **40**, D313–D320.

Rappoport,N. *et al.* (2013) ProtoNet: charting the expanding universe of protein sequences. *Nat. Biotechnol*., **31**, 290–292.

Sasson,O. *et al.* (2003) ProtoNet: hierarchical classification of the protein space. *Nucleic Acids Res*., **31**, 348–352.

Sillitoe,I. *et al.* (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res*., **41**, D490–D498.

Silverstein,K.A. *et al.* (2001) The MetaFam Server: a comprehensive protein family resource. *Nucleic Acids Res*., **29**, 49–51.

Wilson,D. *et al.* (2009) SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res*., **37**, D380–D386.