BIOINFORMATICS

Vol. 24 ECCB 2008, pages i193–i199
doi:10.1093/bioinformatics/btn301

Connect the dots: exposing hidden protein family connections from the entire sequence tree

Yaniv Loewenstein[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, Jerusalem 91904, Israel

ABSTRACT

Motivation: Mapping of remote evolutionary links is a classic computational problem of much interest. Relating protein families allows for functional and structural inference on uncharacterized families. Since sequences have diverged beyond reliable alignment, these are too remote to identify by conventional methods.

Approach: We present a method to systematically identify remote evolutionary relations between protein families, leveraging a novel evolutionary-driven tree of all protein sequences and families. A global approach which considers the entire volume of similarities while clustering sequences, leads to a robust tree that allows tracing of very faint evolutionary links. The method systematically scans the tree for clusters which partition exceptionally well into extant protein families, thus suggesting an evolutionary breakpoint in a putative ancient superfamily. Our method does not require family profiles (or HMMs), or multiple alignment.

Results: Considering the entire Pfam database, we are able to suggest 710 links between protein families, 125 of which are confirmed by existence of Pfam clans. The quality of our predictions is also validated by structural assignments. We further provide an intrinsic characterization of the validity of our results and provide examples for new biological findings, from our systematic scan. For example, we are able to relate several bacterial pore-forming toxin families, and then link them with a novel family of eukaryotic toxins expressed in plants, fish venom and notably also uncharacterized proteins from human pathogens.

Availability: A detailed list of putative homologous superfamilies, including 210 families of unknown function, has been made available online: http://www.protonet.cs.huji.ac.il/dots

Contact: lonshy@cs.huji.ac.il

1 INTRODUCTION

Homologous protein sequences (of common evolutionary ancestry) assume similar 3D structure, and carry out related molecular functions—this is the most fundamental premise of protein sequence analysis. This understanding has facilitated the grouping of proteins descending from clear common ancestry into homologous sequence groups known as protein families. Functional and structural relatedness allow alternative objectives for grouping of proteins into so called protein families. These however are not necessarily homologous, as in the case of the fold level in the SCOP structural classification (Murzin *et al.*, 1995), or the ENZYME (Bairoch, 2000) hierarchy based on similar enzymatic functions—these may have evolved independently by seldom events of convergent evolution. Convergent sequence evolution is extremely rare (Doolittle, 1994), and we thus concentrate on homologous families in this manuscript.

Common ancestry (i.e. homology) can be statistically inferred from sequence alignment. Ergo, the importance of pairwise and multiple sequence alignment (MSA) methods for detection and characterization of protein families, were appreciated already 30 years ago. BLAST (Altschul *et al.*, 1997) is the most popular method for detecting homologues for a query sequence. Based on a fast pairwise alignment search, it reports a statistical score (E-score) for the query and candidate sequences. BLAST and other pairwise alignment methods alike, perform poorly for twilight-zone homologous sequence pairs <30–35% sequence identity (Rost, 1999).

It is only in the last decade with the increase in genomics and proteomics data, that systematic methods had been developed to assign sequences to protein families in a genomic scale (Pearson and Sierk, 2005). In Pfam (Finn *et al.*, 2008), and resources alike, a statistical profile (HMM) is built from a semi-manual multiple alignment of seed homologous sequences. The model is then used to scan protein sequences for additional family members. Pfam families are domain based, while other resources like PIRSF (Wu *et al.*, 2006) focus on whole protein homology. SUPERFAMILY (Wilson *et al.*, 2007) scans all proteins for structural domains, using HMMs built from structural alignments of SCOP superfamily level representatives (homologous structures).

Homologous protein sequences diverge faster than structure. As a result, structural superfamilies are often crumbled into distinct protein families based on sequence similarities. These are embodied in distinct sequence signatures (profiles). The extant families are said to be homologous (i.e. evolutionary linked) if they have clear common ancestry, which is manifested by significant structural similarity, and most often also functional relatedness (Finn *et al.*, 2006). The average sequence identity within a Pfam family is often in the twilight-zone. Homologous sequences from different families are even more remote, and are usually neither alignable nor detectable by pairwise alignment methods alone.

It is of great interest to detect relatedness of protein families without requiring their costly experimental 3D-structure elucidation. Hence, computational methods targeting detection of these faint evolutionary links need to rely only on sequence. Recent advances in the field are dominated by methods that include profile–profile alignment (PPA) and profile–HMMs comparisons (Soding, 2005). Profile methods outperform single sequence-based search. However, they are significantly more computationally intensive (slow), and gravely affected by the quality of the underlying MSA (Madera and Gough, 2002; Sadreyev and Grishin, 2004).

*To whom correspondence should be addressed.

In Pfam, superfamilies, a grouping of homologous families, are manually gathered into Pfam clans (Finn *et al.*, 2006) based on PPA methods, the literature and scarce structural data. Further to the evolutionary insight *per se*, detection of homologous families allows to safely transfer costly experimental knowledge from a well-studied family to a large number of proteins in an uncharacterized family.

An alternative to profile-based methods is the ProtoNet database. It offers a hierarchical classification of proteins based on a tree that captures evolutionary relatedness among protein families (Kaplan *et al.*, 2004). It is based on agglomerative average-linkage clustering of all protein sequences, based on BLAST E-scores from an exhaustive all-against-all comparison. Clusters in the tree show high correspondence with homologous sequence (i.e. Pfam and InterPro), functional (i.e. EC classification) and structural (i.e. SCOP) families (Kifer *et al.*, 2005). It serves as a resource for discovery of overlooked and new functional connections (Schueler-Furman *et al.*, 2006). The tree construction is fully automatic, and is based only on the protein sequences. It provides protein groupings in continuous evolutionary granularities—from closely related subfamilies (high percent identity) to hardly alignable distant superfamilies, in contrast to the limited granularity provided by standard resources (e.g. Pfam).

We have recently reported a new clustering algorithm (MC-UPGMA), which can cluster millions of sequences with a mathematical exactness guarantee (Loewenstein *et al.*, 2008). Using it, we constructed a hierarchical tree (ProtoNet5.1) for 1.8 million non-redundant (UniRef90, maximum 90% sequence identity) proteins, that represent 2.5 million UniProtKB (Wu *et al.*, 2006) proteins. Clusters in the tree correspond to protein families as defined by external resources including Pfam and InterPro (Mulder and Apweiler, 2007). A total of 61% of the tree sequences (UniRef90) are assigned to at least one family by Pfam (here, 8168 families).

Herein, we present a systematic approach to suggest undetected relations between homologous protein families based on this tree. We take a radically different approach for this task, which does not require a family profile, nor the hard task of multiple alignment of remote homologues. Instead, we rely on the varying tree granularity, and on its ability to grasp homologous superfamilies from BLAST similarities. We calibrate the tree for the granularity of each inspected family, and then test for other families in the same putative superfamily which is suggested by the tree.

We control for the possibility of false transitivity in the instance of multi-domain proteins, by taking into account co-occurrence patterns of the inspected domain families. The capacity of our simple protocol to identify hundreds of overlooked protein family connections is reported and exemplified by some new biological findings.

2 APPROACH

Our method is based on a binary tree representing sequence evolution by protein sets of varying evolutionary granularities. The tree construction is fully automatic and requires only the protein sequences and no external prior knowledge (Kaplan *et al.*, 2004).

Our current effort is based on the robust tree stemming from accurately clustering the mass of all pairwise similarities in an all-against-all permissive BLAST comparison of all protein sequences (Loewenstein *et al.*, 2008).

We identify junctions (tree clusters) that represent an evolutionary breakpoint in an ancestral protein family, into two sub-clusters which correspond well with two different protein families, A and B (Fig. 1). Such a cluster, an *AB*-pair, represents a sequence superfamily—a super-set of two existing sets (protein families) from which both have descended. Proteins in *A* and *B* are remote homologues, most often in sequence alignment twilight zone (Rost, 1999), and are identified by distinct sequence signatures (here, Pfam HMMs). However, *A* and *B* have homologous 3D structures, and are often functionally related. We take advantage of this fact to propose new functional and structural assignments.

3 METHODS

3.1 Evolutionary-driven sequence clustering

ProtoNet provides an evolutionary-driven tree constructed using UPGMA—an agglomerative hierarchical clustering average-linkage method (Sokal and Michener, 1958). Here, the clustering includes the entire set of 1.8 million non-redundant UniRef90 protein, for which an all-against-all permissive BLAST comparison yields 1.5 billion unique sequence similarities as reported in (Loewenstein *et al.*, 2008). We have used the MC-UPGMA algorithm which provides the mathematically correct UPGMA tree for very large data, i.e. the tree that best captures the evolutionary process as reflected by BLAST sequence similarities.

The clustering is able to trace very faint relations between homologous families, which are otherwise not discernible from noise (e.g. on a single sequence basis). The huge mass of similarities is embodied into a comprehensive tree by the clustering, and allows for the identification of hidden family connections as reported here. We are now able to leverage the entire sample size of sequenced proteins in a family, which is translated in turn to highly sensitive predictions.

The clustering process is unsupervised—it does not rely on any external knowledge such as protein family assignments, but rather on sequence similarities alone. Thus, sequences which are not assigned to any known protein family may still provide valuable similarity data to guide the clustering process (Figs 1 and 3).

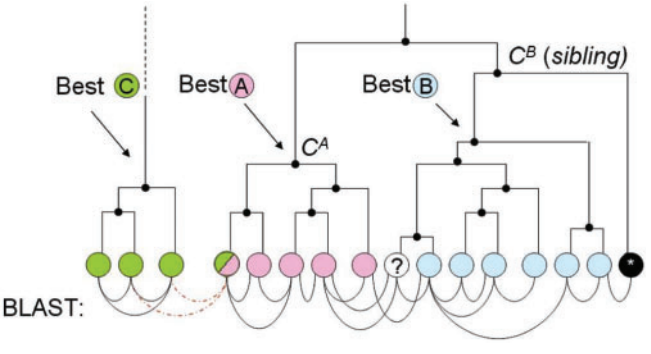

Fig. 1. Superfamily tree search illustration. Pink and blue represent proteins in homologous families *A* and *B*, while green and black denote other families *C* and *D*. Reported BLAST similarities are depicted by curved edges (bottom). *A* and *C* coincide on a multi-domain protein (pink and green protein) which may induce false-transitivity—falsely clustering *A* with non-homologous *C* due to local BLAST similarities of multi-domain protein (red edges). Correct merging of *A* and *B* is aided by transitive similarities of an unassigned protein (white).

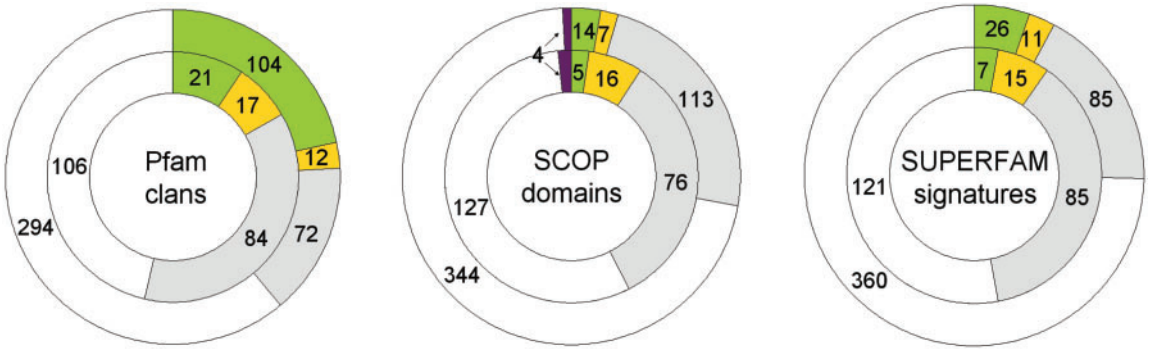

Fig. 2. Agreement of external family groupings with *AB*-pairs. Inner and outer circle represent 228 coinciding *AB*-pairs (*A* and *B* might occur on the same protein) and 482 non-coinciding *AB*-pairs, respectively. Green and orange represent agreement (TP) and disagreement (FP) respectively. Grey and white represent pairs for which agreement could not be tested due to either (*A* or *B*) or both missing external classifications, respectively. SCOP domains contain additional black band for ambiguous Pfam to SCOP mappings in a classifiable pair (see text). Non-coinciding *AB*-pairs are consistently of higher agreement with external classifications (see text).

3.2 Tree correspondence with protein families

3.2.1 Correspondence score In this report, identification of protein families in the tree is based on a given external reference assignment of keywords (here, Pfam) to proteins. A keyword corresponds to an external protein family. A protein might contain several keywords, for instance in the case of domain-based families and a multi-hetero domain protein. The correspondence of a cluster c to a keyword k is given by the Jaccard score

$$J(c,k) = \frac{|c \cap k|}{|c \cup k|} = \frac{TP}{TP+FP+FN} \quad (1)$$

In cluster c, a protein having keyword k is a true positive (TP), and a keyword having protein without k is a false positive (FP). A protein having the keyword k, which is not in c, is a false negative (FN). This score (J) ranges from 0 (no intersection) to 1 (perfect correspondence), and is a lower bound on both specificity and sensitivity.

3.2.2 Best cluster The 'best cluster' for each keyword, is defined to be the cluster with the highest correspondence score. The correspondence of the tree to this keyword ($J(k)$) is measured by the best cluster, as described in Kaplan *et al.* (2004). This criterion allows for (i) finding the cluster granularity that best matches each external reference family; (ii) scoring the correspondence quality. We have shown the biological relevance and the high quality of functional inference based on this criterion in the past.

Keyword A (having a best cluster C^A) with $J(A) < J^A_{cut}$ (implies that specificity or sensitivity are $< J^A_{cut}$) is not considered of high correspondence with the tree, and is thus not further evaluated.

3.3 Homologous family search criterion

3.3.1 'Good' sibling For each keyword A, having a high quality cluster C^A (i.e. passing the J^A_{cut} filtration), we have inspected the sibling in the binary clustering tree (Fig. 1)—the nearest cluster with whom it was merged. In cases where the sibling cluster C^B corresponds well ($J \geqslant J^B_{cut}$) with another protein family keyword B, keywords (protein families) A and B are hypothesized to be evolutionary related (i.e. homologous).

Given the correct tree (assuming one exists), and protein family assignments which fully agree with it, this procedure will trace all speciation events which have diverged ancestral superfamilies into extant families, all having clusters with perfect correspondence ($J=1$). In practice however, it is clear that huge root clusters are often meaningless artifacts of the clustering, rather than homologous groups. Domain combinations introduce further complications.

The rational for the proposed selection criteria are manifold. First, since the partition into C^A and C^B is supported by external expert knowledge (J above threshold) it is considered solid. Furthermore, the suggested relatedness of A and B stems mostly from true family members since C^A and C^B are specific, and on the majority of family members since they are sensitive as well, and is thus supported by entire uncontaminated families. Notwithstanding, permitting some FPs in the process (cutoffs <1) allows to sustain false family membership, e.g. proteins with missing family assignments, or minor clustering errors which are expected to be negligible due to the robustness of averaging over entire families at clustering time. Here, we use $J^A_{cut} = J^B_{cut} = 0.6$. In effect, this requirement implies that most good siblings are also best clusters.

3.4 Error-proneness due to multiple domains

The pitfalls of false transitivity while grouping multi-hetero-domain protein domains have been long known (Portugaly *et al.*, 2006). Seldom coincidence of families A and B on the same protein sequences may cause them to cluster together, for a reason other than homology (Fig. 1). Keyword coincidence is thus a marker for relations which are more prone to have stemmed not from homology. We thus mark these putative *AB*-pairs for more careful analysis.

3.5 Pfam protein families

The Pfam families were selected for this study, as it is one of the most prominent high quality and high coverage sources of homologous protein families. Unlike e.g. InterPro families, Pfam signatures are reconciled to never overlap by definition, and are thus mutually exclusive. Hence, Pfam does not contain trivial links between families and presents an extensive and consistent test case for our method. However, our method is applicable to any protein family resource. Previous studies for homologous family recognition have been based on Pfam as well (Pandit *et al.*, 2002).

Pfam assignments for UniProtKB sequences are fully automatic, and are thus prone to have some (i) false family assignments and (ii) missing family assignments. The evolutionary granularity of different Pfam families is dependent for instance on manual selection of HMM seed sequences, and overlap reconciliation considerations. Only about 40–45% of Pfam families are currently represented by a solved structure, and about 2300 domains belong to the 'Domain of unknown function' (DUFs) or 'Uncharacterized protein family' category according to (Grabowski *et al.*, 2007).

The tree corresponds with 1 791 417 non-redundant proteins (<90% sequence identity) having some BLAST alignment (Loewenstein *et al.*, 2008). A total of 8168 Pfam families, were assigned to 61% of these sequences, based on the UniProtKB data files (rel. 9.0). The average size of Pfam families on the non-redundant tree sequences is 178 ± 567 and 6882 families contain at least 10 members.

As our clustering process is not aware of Pfam assignments, it may incorporate family members which have not been detected by Pfam (Figs 1 and 3).

3.6 Automatic validation

To quantify the quality of our results, we compare our predictions to external resources. Pfam have recently introduced the concept of Pfam clans—a partial grouping of putative homologous families. This grouping is still very much in flux, and currently contains 283 clans (rel. 22.0) used for evaluation herein. In SCOP however, there are currently 1777 structural superfamilies (rel. 1.73) for the much smaller set of solved structures. Therefore, the current coverage of superfamilies by existing clans seems to be far from being complete. To bridge this gap, and to test our predictions by structural references, we have carefully tailored two high quality custom reference sets that provide structurally driven groupings of Pfam families.

First, we validate our predictions based on SCOP classifications. These serve as a standard of truth based on manual classification of solved 3D structures. Second, we test the agreement of our predictions with that of the SUPERFAMILY predictions. Hereby, we elaborate the exact design of these benchmarks.

3.6.1 Domain agreement The level of agreement of two domains on a single sequence is measured by a standard agreement score (Portugaly *et al.*, 2006)

$$agreement(k_1, k_2) = \frac{|k_1 \cap k_2|}{|k_1 \cup k_2|} \quad (2)$$

On the inspected sequence, this is the sequence length ratio of the two domains overlap to the total coverage of both domains. This score, ranges from 0 for no overlap, to 1 for full agreement on domain boundaries. It will be low if the overlap between the two signatures does not cover the majority of both.

3.6.2 Assessment by SCOP A mapping of 3624 Pfam families to the sequences (and domain boundaries) of all PDB structures and of PDB to SCOP domains were downloaded from the Pfam (rel. 22.0) and SCOP (rel. 1.73) websites, respectively. SCOP domain assignments are derived from 3D structural proximities and thus occasionally incident on multiple PDB chains (different polypeptides) or on non-consecutive sub-sequences. Only the former cases were not considered for our analysis. The following SCOP classes i, j and k (low resolution structures, peptides and designed respectively) were not used in our analysis.

A Pfam is mapped to a SCOP code, if their agreement score (including non-continuous domains) exceeded 0.75, or if the SCOP domain covered the entire respective PDB chain. This resulted with a total of 1489 Pfam to SCOP one-to-many mappings. The number of mappings was not sensitive to the agreement threshold parameter (1543 and 1345 for 0.5 and 0.9 agreement). However, 34 (26, 35) mappings contained more than one SCOP fold (class, superfamily). These Pfam mappings (2.3%) are ambiguous and inconsistent with the SCOP structural classification. We note that this might be an underestimate of Pfam inconsistency with structural classification due to the low structural coverage of protein sequences which are classified by Pfam, even though the issue of structural coverage by Pfam has been recently addressed (Finn *et al.*, 2008).

Assessment is only possible whenever each Pfam family (*A* and *B*) in a predicted *AB*-pair are mapped to exactly a single SCOP fold. The prediction is correct if both are mapped to the same fold, false otherwise.

3.6.3 Assessment by SUPERFAMILY The SUPERFAMILY method applies HMMs built from SCOP superfamily-level representatives (homologous structures) to scan all protein sequences for putative structural domains. This allows for high-quality structural prediction for protein sequences having no structural representative. Generally, SUPERFAMILY (*SSF*) domains are of coarser evolutionary granularity than Pfam. They can thus be leveraged to structurally relate several Pfam families from sequence.

The mapping of Pfam domains to SSF domains is based on InterPro scan (rel. 12.6, from InterPro's ftp) for Pfam and SSF assignments on the full UniProtKB data. Since the volume of SSF predicted domains (sequence space) is orders of magnitude larger than SCOP, a more quantitative mapping policy was appropriate. Whenever a SSF and a Pfam domain coincided on the same protein, the domains' boundary agreement was recorded [Equation (2)]. We required that (i) the average agreement between the matched signatures is at least 0.5 (implying that at least half of each is covered by the other), (ii) the SSF signature appears at least at 50% of the Pfam occurrences (prevents spurious SSF-Pfam co-occurrences). Here, we have used a lower agreement threshold, to accommodate for (i) fuzzier domain boundaries due to two local HMM searches (as opposed to structural determination), and (ii) improved stability since, unlike SCOP mappings, SSF is typically assigned to a large sample of sequences.

Out of 1563 mappings (average agreement $\geq$0.5), only 97 covered <50% of the Pfam occurrences. The high quality of this mapping is further demonstrated by the fact that no Pfam signature is mapped to more than one SSF signature. Whenever both *A* and *B* are mapped to some SSF signature (here, 59 pairs), we say that a predicted homologous pair by our method is correct if it is the same SSF signature, false otherwise.

4 RESULTS AND DISCUSSION

4.1 Pfam tree correspondence and *AB*-pairs

Pfam families were very well captured by our new tree. For 8095 (out of 8158) non-trivial families (at least two members with <90% sequence identity) the tree achieved an average Pfam correspondence score (*J*) 0.89 $\pm$ 0.17, specificity 0.96 $\pm$ 0.09 and sensitivity 0.92 $\pm$ 0.16. Single domain or fixed domain architectures, comprise the majority of the data, and are captured better by the tree, compared to a handful of domain families which appear in promiscuous domain architectures. The latter are more prone to clustering mistakes (Fig. 1).

From this set, our method predicts a total of 710 unique *AB*-pairs (i.e. *AB* = *BA*) linking putative homologous families. For this subset of Pfam families, the tree achieves average correspondence of 0.93 $\pm$ 0.09 (specificity 0.98 $\pm$ 0.05, sensitivity 0.95 $\pm$ 0.08) for the best clusters (*A*) and 0.88 $\pm$ 0.12 (specificity 0.93 $\pm$ 0.10, sensitivity 0.94 $\pm$ 0.09) for the siblings (*B*).

BLAST similarities between sequences in the same Pfam family are sparse, i.e. only some of the pairs are reported (Loewenstein *et al.*, 2008). At the Pfam clan level, connections are extremely sparse, yet they are reliably picked up by the clustering. For example, 80% of the best clusters for 283 Pfam clans had <10% BLAST linkage (proportion of reported pairs) even at a very permissive threshold (*E* = 100). For the reported 710 *AB*-pairs there was 14% linkage on average. For comparison, the average BLAST linkage is 64% in Pfam families' best clusters (12% have <10% linkage).

4.2 Validation—clans, SCOP and SUPERFAM

Pfam clans have provided the most extensive validation for our predictions—154 of 710 predictions could be automatically evaluated by existence of a clan for both *A* and *B*. The number of *AB*-pairs that could be automatically validated by SCOP and SUPERFAMILY but not existence of a clan is rather limited due to (i) low coverage and (ii) significant overlap with existing clan groupings. Out of 710 pairs, 125 were confirmed by existence

Table 1. *AB*-pair predictions containing obsolete ('dead') Pfam families

A	B	New family	Comments
PF00252	***PF00826***	PF00252	
PF01598	***PF03897***	PF04116	
PF01892	PF04608	PF04068	DUF correctly classified
PF02502	***PF06562***	PF02502	DUF correctly classified
PF03240	***PF06981***	PF08666	
PF03692	***PF05779***	PF03692	DUF correctly classified
PF04000	***PF07493***	PF04000	
PF04032	***PF08296***	PF04032	
PF04132	PF04157	PF04157	
PF04956	***PF06921***	PF04956	
PF04965	PF07025	PF07025	DUF correctly classified

Obsolete families (bold italics) were merged into existing or new families by the Pfam curators. All 11 cases were correctly classified by our method as being homologous, including 4 DUF containing pairs.

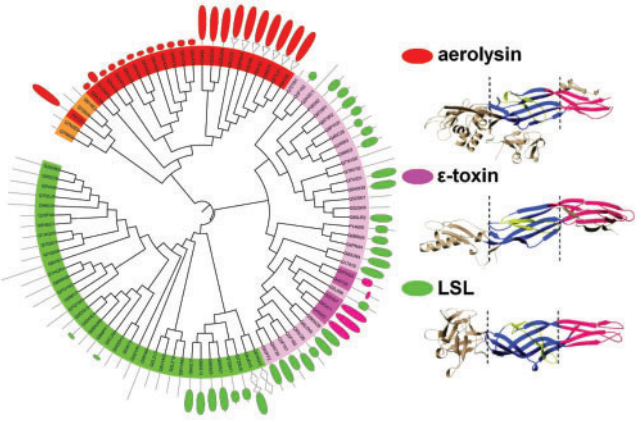

Fig. 3. Subtree of *AB*-cluster for aerolysin-ETX-MTX2**.** subclusters are color coded by Pfam correspondences. Cluster C^A (red) is best for (*A*) the Aerolysin toxin family (PF01117) and is merged with C^B (pink and green), corresponding with (*B*) ETX_MTX2 family of Clostridium epsilon toxin and Bacillus mosquitocidal toxins (PF03318). Pfam annotated sequences are marked by darkened leafs (red and pink, for *A* and *B*, respectively). Multiple family-less proteins are included in the clusters due to sequence similarities. Both families assume homologous 3D structures (right)—SCOP superfamily f.8.1. The green cluster contains no significant Pfam correspondence, and is a newly discovered putative eukaryotic family (*C*) in this superfamily. A solved structure (fungal LSL toxin) (right) and 9 SUPERFAMILY predictions ($E \leqslant 1$) for this SCOP superfamily in the green cluster support our findings. Other cluster members, share significant sequence similarity with predicted pore-forming proteins. Outer ring depicts domain architectures for Pfam (red and pink ovals denote *A* and *B* domains) and SSF predictions for f.8.1 ($E \leqslant 1$ green ovals).

of a clan, but only 15 extra non-clan pairs were added by SUPERFAMILY and SCOP altogether.

On this set, the clans seem to extract most, but not all, of the information which could be derived from structure—12 of 33 SUPERFAMILY (3 of 19 SCOP) validated homologies were not incorporated into a clan. The mapping of Pfam to PDB still seems to not fully agree with SCOP. Considering the high overlap of correct predictions by all three methods, and the fact that the larger clan assessment has yielded notably more favorable error estimates than SCOP and SUPERFAMILY, we deduce that these automatically constructed benchmarks are of possible lower quality than the manual clans. This finding is also supported by ambiguities in SCOP to Pfam mappings. Our analysis indicates that Pfam clans are currently the most extensive resource for homologous sequence families, yet they are still far from being complete. The majority of our predictions are indeed novel, since they could not be automatically validated by any of the inspected existing resources.

We note that the mapping of Pfam to SUPERFAMILY offers a powerful way to propose structurally driven evolutionary links between protein families from sequence.

4.3 Validation—Pfam re-annotations

The Pfam set of families is constantly updated. For instance, Pfam families are occasionally merged by Pfam curators, as they are identified in retrospect as two sub-families of a single common family. As a result, some family signatures are pronounced as obsolete ('dead'). We have inspected all *AB*-pairs containing a Pfam that 'died' in the time frame from our analysis to present time. Table 1 shows that our fully automatic predictions were all judged as correct by manual re-definitions occurring at Pfam.

4.4 Biological example—pore-forming toxins

We demonstrate an application of our method to reveal new biological findings, which are not captured by Pfam clans. The clustering suggested that the following families are related.

A—Aerolysin (Pfam PF01117) is a family of toxins from Gram-negative bacteria which oligomerize to form pores in membranes leading to host cell lysis. It is involved in diarrhoeal diseases and deep wound infections.

B—The Pfam ETX_MTX2 (PF03318) family encompasses epsilon toxins originating from the Gram-positive bacteria *Clostridium perfringens*, and related insecticidal toxins from the Gram-positive bacteria *Bacillus thuringiensis*, which have been expressed in a variety of genetically modified crops (e.g. Bt-corn) for pest control.

This *AB*-pair is picked up at 3.6% BLAST linkage. Our finding is supported by solved structures from both families, which are classified into the same SCOP superfamily (Fig. 1). This example also illustrates how unannotated proteins (i.e. that do not belong in any Pfam family) are still instrumental in the clustering process—these are putative undetected family members proposed by the clustering.

Interestingly, before *B* is merged with *A*, it is merged with a cluster (Fig. 3, green cluster) of a new putative family (*C*) in this superfamily. This cluster contains (i) the natterin and hydralysin eukaryotic toxin families from venomous fish and hydra, (ii) the homologous plant gene *Hfr-2* which has been shown to be up-regulated on larvae feeding, (iii) fungal LSL hemolytic toxins and (iv) uncharacterized proteins from the human pathogens *Schistosoma japonicum* and *Legionella pneumophila* which are causes of schistosomiasis and the legionnaires disease. We have thus shown the evolutionary relatedness of these families, including

possible virulent factors of major health concerns. SUPERFAMILY predictions and a solved structure also support this finding—*A*, *B* and *C* are homologous. SUPERFAMILY predictions also support the inclusion of some Pfam unannotated proteins in the ETX_MTX2 cluster (Fig. 3, pink cluster).

4.5 False-transitivity—coinciding families

We note that the non-coinciding *AB*-pairs produced very high accuracy for the hard task of clan prediction—104 correct predictions versus only 12 wrong predictions (366 novel predictions could not be automatically validated by Pfam clans). Furthermore, through all three automatic validation sets, non-coinciding pairs have had fewer errors (Fig. 2). Hence, our results show that coincidence of keywords is a good proxy to automatically warn against false implications from sequence clustering due to false-transitivity (Fig. 1). We note however that coinciding families, albeit of lower quality in general, still reveal otherwise overlooked meaningful connections.

4.6 Characterization of a high quality subset

We have inspected several features to further separate correct from wrong hypotheses. Requiring that the average correspondence (of *A* and *B*) is at least 0.95, in addition to no-coincidence of *A* and *B*, has yielded an almost perfect assignment of clans—only one false assignment out of 48 instances that could be validated by clan assignments for both *A* and *B*. This set still includes 254 of the original 710 predictions, including 82 DUF containing pairs, and is further inspected throughout this section as a higher quality subset.

We have manually inspected the top and bottom ranked predictions for this set which could not be automatically verified by a clan, SSF signature or SCOP mapping in order to qualitatively characterize our best and worst predictions of clear novelty in this manageable set.

We tested if our predictions are supported by profile comparison methods, known related functions, literature scan, conservation patterns of functional signatures, intrinsic clustering features, structure prediction and more. This effort is summarized in Table 2, for 20 *AB*-pairs. We have assigned a prediction as true (Table 2) only when there was enough independent support for the prediction. Probably true connections, which could not be supported with high confidence, were assigned as 'possible' (P), and the rest were assigned as 'false' (F).

We have identified a handful of new overlooked connections. For some of these cases, a significant MSA could not be deduced from profile and secondary structure comparison, but border-line cases of short well-conserved signatures were identified as candidates for carrying out shared function (e.g. ligand binding).

Notably, the two groups (top and bottom 10 test cases) are of very different character. The top predictions are enriched with viral families and often contain small clusters. The bottom ranked predictions are characterized by large family clusters (50% with at least 100 proteins). Furthermore, the fraction of DUFs is much lower while the BLAST percent linkage is generally higher, indicating different conservation patterns for the two distinctive groups.

Many of the putative partners of viral families lead to interesting evolutionary suggestions on virus–host co-evolution. For instance, mammalian and viral families of interferon-γ receptors (PF04903 and PF07140) are matched. The vaccinia virus interferon-γ receptor is secreted from infected cells. The viral protein efficiently inhibits the interferon-dependent immune response, and leads to increased infectivity. The common evolutionary source of the two families suggests that the viral family has originated from the host proteins. This novel connection is not indicated by any of Pfam's profile comparison tools. Interesting biology is revealed by analyzing *AB*-pairs involving DUFs as well.

Table 2. Top and bottom 10 ranked *AB*-pairs in manual inspection of the high quality subset (see text)—homology predictions are assigned as true, possible (P) or false (F)

A	B	PDB	DUF	Linkage	Number	Profile	Manual	Viral	Correct
PF04541	PF05900								True
PF04903	PF07140								True
PF05307	PF05946								P
PF05780	PF02723								True
PF06358	PF06716								True
PF05733	PF06606								True
PF06193	PF06909								True
PF06285	PF07190								True
PF06147	PF06914								True
PF03158	PF01671								P
PF07357	PF01862								P
PF01190	PF03251								F
PF04953	PF06857								True
PF08470	PF07952								F
PF02495	PF02452								F
PF03025	PF05776								True
PF01785	PF04269								P
PF07406	PF02691								F
PF04237	PF04944								True
PF02221	PF06011								True

Predictions are ranked by average correspondence score (*J*) of *A* and *B*. Each category is marked by three grey levels, indicating low, medium and high levels for each category: linkage (<5%, 5–15%, >15%), number (cluster size <30, 30–100, >100) and profile (HHalign/PRC profile comparison *E*-score >1e-2,1e-2 – 1e-5, <1e-5). Similarly, grey levels indicate the appearance of a category in none (white), one (grey) or both *A* and *B* (dark grey) for viral, PDB structure and DUF. The manual column indicates the support level of further evidence (dark grey indicates more than one independent evidence) such as protein and family descriptions, literature scan, active site comparison, phylogenetic distributions and fold prediction when available.

4.7 Clustering versus profile methods

The stronghold of our method is also its most prevalent shortcoming—it is reliant on the underlying tree. It is thus not suitable for families which are not well captured by the tree. For instance, domains of promiscuous architectures are especially prone for bad clustering due to false application of transitivity rather than true homology. Occasionally, these may not be captured by a tree at all.

So, why does sequence clustering reveal relations undetectable by seemingly stronger and more complex profile methods? We mark several profound differences between the two regimes.

Profile methods represent families as a single statistical object while the clustering is based on pairwise comparison of individual family members. Hence, the clustering is able to detect similarities which are expressed by only a few cluster members (supported by the low linkage of our clusters). These scarce similarities may be

out-weighted in the process of constructing and comparing profile representations. Furthermore, the clustering is aided by putative family members picked up in the clustering agglomeration process, which are not included in the profiles (Figs 1 and 3). Third, we note that profile methods are largely dependable on the quality of the underlying MSA. Automatic inference in genomic analysis is crucially dependent on the MSA quality (Wong *et al.*, 2008). However, MSA is a significantly harder task than pairwise alignment from both theoretical and practical standpoints.

To summarize, we point out that the clustering process considers the entire sequence similarity space as a whole. The global nature of competing agglomeration forces at clustering time, leads to coherent families and superfamilies of thereof, which may be overlooked on a per-family (profile)-based approach.

5 CONCLUSIONS

We presented a straight-forward and intuitively appealing method to induce evolutionary hypotheses from large-scale sequence clustering data. Remarkably, our method is able to detect very hard cases of remote homologous families from a clustering of simple BLAST searches. Most of the suggested connections were overlooked by state of the art more complex methods. Nevertheless, we are able to confirm many of the suggested relations, and characterize markers for prediction accuracy of others. Our results await curation, and incorporation into resources such as Pfam clans.

Our method is automatic and computationally scalable to any size of data. The method is expected to produce more hypotheses, as sequence and annotation data continue to accumulate.

Further to evolutionary insight *per se*, we have shown how our method can produce practical contributions as well. Exposed evolutionary links could be translated into functional and structural predictions for hard cases of uncharacterized families.

ACKNOWLEDGEMENTS

We thank Menachem Fromer for critically reading the manuscript, and the entire ProtoNet group. We also thank the iToL team, for making their tree visualization tool publicly available.

Funding: Y.L. was granted a fellowship from the SCCB, the Sudarsky Center for Computational Biology. This research was supported by the BioSapiens NoE (EU Fr6).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.

Doolittle,R.F. (1994) Convergent evolution: the need to be explicit. *Trends Biochem. Sci.*, **19**, 15–18.

Finn,R.D. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.

Finn,R.D. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.

Grabowski,M. *et al.* (2007) Structural genomics: keeping up with expanding knowledge of the protein universe. *Curr. Opin. Struct. Biol.*, **17**, 347–353.

Kaplan,N. *et al.* (2004) A functional hierarchical organization of the protein sequence space. *BMC Bioinformatics*, **5**, 196.

Kifer,I. *et al.* (2005) Predicting fold novelty based on ProtoNet hierarchical classification. *Bioinformatics*, **21**, 1020–1027.

Loewenstein,Y. *et al.* (2008) Efficient algorithms for exact hierarchical clustering of huge datasets: tackling the entire protein space. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **24**, i41–i49.

Madera,M. and Gough,J. (2002) A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res.*, **30**, 4321–4328.

Mulder,N. and Apweiler,R. (2007) InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol. Biol.*, **396**, 59–70.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Pandit,S.B. *et al.* (2002) SUPFAM–a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes. *Nucleic Acids Res.*, **30**, 289–293.

Pearson,W.R. and Sierk,M.L. (2005) The limits of protein sequence comparison? *Curr. Opin. Struct. Biol.*, **15**, 254–260.

Portugaly,E. *et al.* (2006) EVEREST: automatic identification and classification of protein domains in all protein sequences. *BMC Bioinformatics*, **7**, 277.

Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.

Sadreyev,R.I. and Grishin,N.V. (2004) Quality of alignment comparison by COMPASS improves with inclusion of diverse confident homologs. *Bioinformatics*, **20**, 818–828.

Schueler-Furman,O. *et al.* (2006) Is GAS1 a co-receptor for the GDNF family of ligands? *Trends Pharmacol. Sci.*, **27**, 72–77.

Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.

Sokal,R.R. and Michener,C.D. (1958) A statistical method for evaluating systematic relationships University of Kansas Science Bulletin 38, 1409–1438.

Wilson,D. *et al.* (2007) The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res.*, **35**, 308–313.

Wong,K.M. *et al.* (2008) Alignment uncertainty and genomic analysis. *Science*, **319**, 473–476.

Wu,C. and Nebert,D.W. (2004) Update on genome completion and annotations: protein information resource. *Hum. Genomics*, **1**, 229–233.

Wu,C.H. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.