Published online July 28, 2006

Nucleic Acids Research, 2006, Vol. 34, No. 13 ***3625–3633***
doi:10.1093/nar/gkl484

Genomic scale sub-family assignment of protein domains

Julian Gough*

Unite de Bioinformatique Structurale, Institut Pasteur, 25-28 Rue du Docteur Roux, 75724 Paris Cedex 15, Paris, France

Received March 29, 2006; Revised June 12, 2006; Accepted June 26, 2006

ABSTRACT

Many classification schemes for proteins and domains are either hierarchical or semi-hierarchical yet most databases, especially those offering genome-wide analysis, only provide assignments to sequences at one level of their hierarchy. Given an established hierarchy, the problem of assigning new sequences to lower levels of that existing hierarchy is less hard (but no less important) than the initial top level assignment which requires the detection of the most distant relationships. A solution to this problem is described here in the form of a new procedure which can be thought of as a hybrid between pairwise and profile methods. The hybrid method is a general procedure that can be applied to any pre-defined hierarchy, at any level, including in principle multiple sub-levels. It has been tested on the SCOP classification via the SUPERFAMILY database and performs significantly better than either pairwise or profile methods alone. Perhaps the greatest advantage of the hybrid method over other possible approaches to the problem is that within the framework of an existing profile library, the assignments are fully automatic and come at almost no additional computational cost. Hence it has already been applied at the SCOP family level to all genomes in the SUPERFAMILY database, providing a wealth of new data to the biological and bioinformatics communities.

INTRODUCTION

Hundreds of complete genomes have been sequenced generating a massive quantity of protein amino acid sequences. Information about these can be computationally inferred via homology to other sequences which have had experiments performed on them or which we know something about from some other means. People working on individual proteins, families of related proteins or entire genomes frequently use databases that can annotate their sequences with some information. This is often performed by using similarity searches to discover that the sequence in question is a member of a certain group, represented by a profile, where this group might share common features such as function or three-dimensional (3D) structure.

Most databases and methods for protein sequence annotation are essentially based on a single level. For example most of the member databases of InterPro (1): SUPERFAMILY (2,3) (until now) only at the superfamily level based on evolutionary ancestry, PFAM (4) at the family level based more on sequence similarity and others based on their corresponding systems. In contrast there are classification systems that are hierarchical such as SCOP (5) and CATH (6), having several levels of classification. There is a general problem of placing protein sequences (e.g. from genomes) in the hierarchy of these classification systems. The more specific problem addressed here is, given that a sequence has already been placed into a higher level of the hierarchy, how to subsequently place it into a lower level in one of these classification systems. The desired characteristics chosen for this investigation are that the method used to solve the problem: is fully automatic, is computationally tractable on all completely sequenced genomes and across broad general, inclusive classifications, performs significantly better than existing methods, requires little or no work to update with respect to new sequences and database releases, and provides *E*-value scores suitably reliable for selecting low error rates. Also considered desirable are the ability to suggest a possible closest homologue and the ability to detect cases where an incomplete classification means that a sequence, which although a member of the known higher level group, belongs to a new previously uncharacterized sub-group at the lower level. The aim is not to create a classification system based on statistical principles, but to create a statistical method for annotating sequences into a biologically based classification system which already exists.

The method and results described here are general in nature and could equally well be used to solve the sub-classification problem in answer to other biological questions; however, the SUPERFAMILY database (2,3) using the SCOP (5) hierarchical classification was chosen for testing, implementation on a genome-wide scale and application to existing

*Tel: +33 1 45 88 87 37; Fax: +33 1 45 68 87 19; Email: gough@pasteur.fr

questions. The SCOP domain database for proteins of known structure uses 3D structure, sequence and functional information to classify proteins into a multi-level hierarchy. The SUPERFAMILY database maps the SCOP classification at the superfamily level onto all completely sequenced genomes. This is the level at which sequences with evidence for a common evolutionary ancestor are grouped together. The superfamilies are already sub-classified by SCOP, as part of its hierarchy into the family level, grouping protein domains which are often more similar in sequence and usually share the same or a related function. In release 1.69 of SCOP there are 1539 superfamilies containing 2845 unevenly distributed families including 10 894 proteins with sequences that are <95% identical to each other.

The SUPERFAMILY database was designed to tackle evolutionary problems involving the most distant homologies, but has been heavily used for genome annotation and by biologists working on individual proteins. A family level database such as PFAM may detect less distant relationships, but it provides the functional biologist with more specific information. By adding the family sub-classification to SUPERFAMILY it would substantially enrich the resource from the point of view of the hundreds of scientists more interested in the specific view rather than the broad view. Any method satisfying the criteria laid out as desirable objectives above would also be useful to computational biologists looking to map other existing sub-classification schemes to genome sequences, e.g. G-protein-coupled receptors (7), globins (8,9), zinc fingers (10), phosphoregulators (11) and so on.

The problem as specifically laid out above has not previously been solved, but there have been other related approaches, a few of which will be mentioned. The PFAM database has introduced 'clans', which are a manually curated collection of links between families that are thought to be related, forming connected groups; the annotation to a higher level is trivial since it is inherent in the classification. At UCSC computationally intensive support vector machines have been applied to the individual class G-protein-coupled receptor proteins (12), sub-classifying them based on ligand specificity. Phylogenetic methods have been used in attempts to define automatic classifications (13–15) and to a lesser extent for fitting new sequences into existing classifications. Subgroup-specific hidden Markov models (HMMs) have been used at the sub-family level (16) but could be applied at the sub-superfamily level. Many projects have used simple pairwise methods such as BLAST (17) for trivially associating a nearest neighbour, or in a similarly trivial way researchers using SUPERFAMILY have been known to use the family membership of the seed sequence of a model, despite the fact that the models target the superfamily level.

This paper describes a method which offers a solution to the problem. Some details of the development have been included, which are relevant to would-be developers and users of the method, as well as details of how and where biologists can make use of the fruits of application, and an example of an application to evolutionary studies.

MATERIALS AND METHODS

As described in Introduction, the SCOP database is a hierarchical classification in which superfamilies are sub-divided into families. Given a query sequence, the SUPERFAMILY database is responsible for determining which superfamily in SCOP the domains in that query sequence belong to. Given the superfamily, it is desirable to then extend the information to include more specifically which family the domain belongs to.

The hybrid method

The hypothesis which we wish to test for a domain in a query sequence, assuming we know which superfamily it belongs to, is whether it is a member of sub-family 'A'. The null hypothesis is that it belongs to a different sub-family 'B', 'C', etc. To test this hypothesis we take the best pairwise score between the query domain and any sequence in family 'A', and subtract from it the best pairwise score between the query domain and any member of the superfamily which is a member of a family other than 'A'. In this way if a query sequence has comparable strong scores to more than one family, the discrimination between families is weak; similarly to a lesser extent, weak scores to a family will have improved discrimination if all the other families have excessively weak scores.

To carry out this test the method requires pairwise scores from HMMs, which usually give a many-against-one (profile) score. This can be thought of as a hybrid profile/pairwise method. An HMM will give both a score and an alignment between any sequence and the model, so an HMM guided pairwise alignment can be achieved by aligning both the query sequence and a superfamily sequence to the model. The two sequences are aligned to the model and not to each other, but an alignment between them can be inferred from the relative positions of the residues with respect to the model, i.e. if a residue from each sequence aligns to the same position in the model, they are aligned to each other. A raw score can then be calculated from the inferred pairwise alignment using a substitution matrix and affine gap penalties. Optionally, the contributions to the score of each position may be weighted relative to the importance of that position in the model. A description of how the weights are calculated is shown on each model page in SUPERFAMILY, which can be reached by clicking on one of the models listed, e.g. http://supfam.org/SUPERFAMILY/cgi-bin/models_list.cgi?sf=49503.

The *E*-values are calibrated on the benchmark test using an ad hoc two-parameter sigmoidal distribution (18). The fit is shown in Figure 1. Cases that are impossible to classify correctly in the cross-validation (because the family only has a single member) are included for the sake of calibration. The scoring function is as follows:

$$E\text{-value} = \frac{K}{1 + e^{(\ln(n_2 e^{-\lambda S_2}) - \ln(n_1 e^{-\lambda S_1}))}}$$

where K, λ and τ are coefficients, and n is the length of the target sequence, S is the raw score and the subscripts 1 and 2 refer to the top scoring hit and the next highest scoring hit outside the top scoring family, respectively. Sometimes no values for n_2 are available; in this case n_2 is set to 180 (average domain length) and S_2 is set to 0 (no alignment).

The design of this method was not made independently from the intended implementation on an all-genome scale.

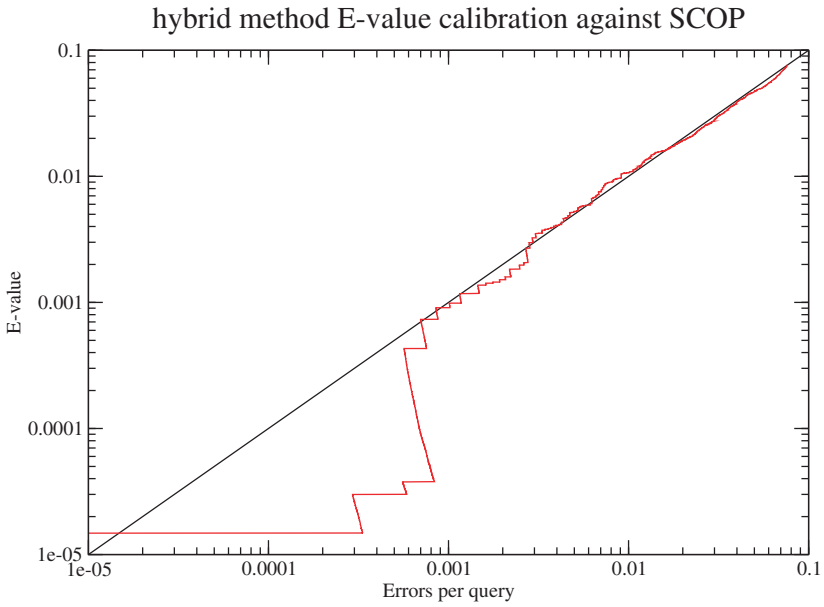

Figure 1. The *E*-value calibration curve of the hybrid method using the sigmoidal distribution on the benchmark test. Anomalies in the SCOP classification at the family level are at least partially responsible for the digression in the lower part of the graph, i.e. a handful of top-scoring false positives.

When applied to the genome analysis in the SUPERFAMILY database, it requires no additional BLAST or HMM scores and alignments to be calculated that are not already performed as part of the superfamily level assignment process. Furthermore, it does not require the creation of any additional objects which would need to be updated with future releases of the database, e.g. family-specific models, phylogenetic trees and trained neural networks. Once implemented the human and computational cost of applying this to all genomes is almost nothing. Furthermore, it is independent of and complimentary to other potential solutions to the problem which have been used or suggested before, such as generating family-specific models; these could simply be used in place of the superfamily models in the current method further improving performance.

A sequence submitted to SUPERFAMILY for analysis is scored against all superfamily models. There is one model for each seed sequence with 95% sequence identity at the SCOP superfamily domain level. Alignments to the models for each domain in the query sequence are necessarily produced as a byproduct of scoring, and these alignments are used as the query side of the hybrid method. For each model, alignments to every family member are pre-calculated and stored so that they may simply be looked up to provide the other side necessary for the hybrid score. Potentially, pairwise alignments between the query sequence and every family member are available for every model without additional calculation. The potential problem of confusing multiple domain hits in a single query sequence is neatly sidestepped via use of information from the superfamily level assignment procedure (2); the individual domain from the whole sequence which each hit is associated with comes out without the need for explicitly chopping the sequence into domains. This is a crucial feature of the approach since detecting domain boundaries is a particularly difficult problem.

The benchmarks

The SCOP database classifies protein domains of known structure using 3D structure information as well as functional information from literature and sequence information. As a 'gold standard' it can and has been used for benchmarking many sequence comparison methods (19–21). However in this work there is a unique difference, inasmuch as a correct superfamily classification is assumed, and the correct family within the superfamily must be determined.

The sequences of each domain in SCOP obtained from the ASTRAL database (22), sharing no more than 95% sequence identity are used for cross-validation. Each sequence in turn is removed from the database, along with the HMM which was built using that sequence as a seed. Then the sequence is compared to all the sequences/models remaining in the superfamily, using BLAST, HMM or the hybrid method to obtain the supposed family level classification and score. The classification obtained by each method is compared to the actual family classification in SCOP to determine whether it is true or false. Cases where a domain is the only member of its family are excluded in the benchmark, since it is impossible to correctly determine the family; these cases are included for *E*-value score calibration since they should score poorly.

The benchmark is excellent for comparing the relative performance of different methods, but it is very difficult to estimate the performance in absolute terms. The dataset is filtered to the 95% level, but this is relatively arbitrary and designed to remove trivially easy cases. Many of the most difficult examples of sub-classification which are tested in this independent (from superfamily level) test would not arise in reality since they are too distantly related to be detected first at the superfamily level, which is a prerequisite.

Benchmarking the closest structural homologue is less easy to do clearly, since there is no reliable automatic measure of structural distance. Parsed and cleaned PDB-style files

for all the domains were obtained from the ASTRAL database (22) and all family members compared with the CE structural alignment program (23). The top-scoring structural domain by HMM, BLAST and the hybrid method were compared to the *Z*-score produced by CE as shown in Figure 4. The top 10 scores from the hybrid method which disagreed with CE were examined, and at least half were due to errors made by CE; in the other cases, both the top hits from CE and the hybrid method seem plausible. Examining 10 at random with lower scores, there were two cases where CE was wrong and one case where the hybrid method was wrong, the rest being plausible for both. The cause of these errors is not due to CE assigning inappropriate *Z*-scores to pairs of structures, but appears to be due to CE failing to detect alignments in some cases. Sometimes the alignment is detected by CE when one structure is compared to another, but not vice versa, i.e. it is asymmetric.

RESULTS

The method was developed and tested on SCOP and SUPERFAMILY versions 1.67 and subsequently implemented in version 1.69. The SCOP classification was used for cross-validation on all domains filtered to 95% sequence identity, giving a true or false result for each. This is similar in concept to previous benchmarks (19–21) and described fully in Materials and Methods.

The family level assignments for all (over 300 but ever increasing) completely sequenced genomes are publicly available via the SUPERFAMILY web server at http://supfam.org/SUPERFAMILY/cgi-bin/gen_list.cgi and users may submit their own sequences for analysis at http://supfam.org/SUPERFAMILY/downloads.html. The data are available for bulk download via FTP; instructions may be found at http://supfam.org/SUPERFAMILY/downloads.html. Also available via FTP is the complete SUPERFAMILY package, which includes family level classification. Those interested in inspecting or using the actual hybrid method sub-family classification code may access it there.

Hybrid method

Each model in the SUPERFAMILY database has a corresponding seed sequence, but in model-building the aim is to produce a model which covers as much of the whole superfamily as possible. There may be several models for each superfamily. Several researchers in their need for family level classification have performed this by taking the family that the seed sequence of the top-scoring SUPERFAMILY model belongs to. This is not expected to work very well, since the model is designed to span all families in the superfamily. The performance of this approach is shown in Figure 2 labelled 'HMM score'. The curves in this figure were obtained by plotting the true versus false positives (see the benchmark in Materials and Methods) in order of increasing score, i.e. decreasing strength of assignment.

Classification of sequences at the SCOP superfamily level is the ultimate remote homology detection problem, but given the superfamily, determining the family level classification within the SCOP hierarchy is much easier since the sequences are more closely related. For this reason, a pairwise method such as BLAST may be sufficiently powerful (whereas superior profile methods such as HMMs are needed at the superfamily level). The performance of family level assignment via the family membership of the sequence in SCOP with the best BLAST score is also shown in Figure 2.

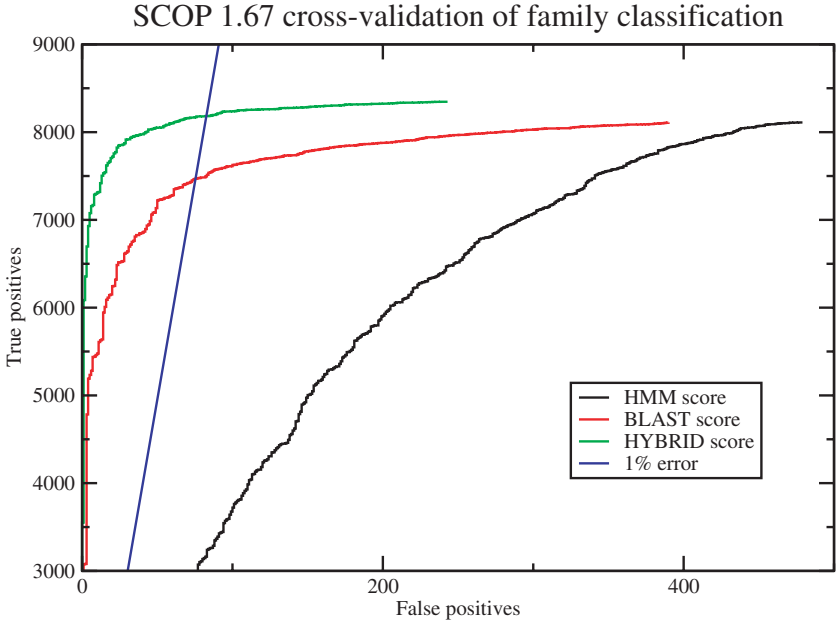

Figure 2. Out of a possible 8775 family classifications in SCOP 1.67 between domains with <95% sequence identity, these are the numbers of true and false positives plotted with increasing value for three different scores: the family of the seed sequence of the top-scoring HMM, the family of the top scoring BLAST hit and the family chosen by the hybrid method. Approximately, the last 400 classifications for each method are not shown due to a requirement of *E*-value < 100.

Since the sub-family classification problem is relatively easy, if BLAST was used it would provide useful additional information to associate with the superfamily assignments; this in itself is a worthwhile end. However this would merely be an informatics gain, rather than actually achieving something beyond that which is possible with existing bioinformatics resources and some effort. Although it is not entirely needed for a useful family level classification, a general hybrid method was developed which surpasses both the HMM scores and the BLAST scores, and is more widely applicable to other sub-classification problems. The results complete the trio shown in Figure 2. BLAST is expected to perform very well on closely related sequences, and HMMs are expected to have the capacity to detect the more distant relationships that are beyond the reach of most pairwise methods (19,21). It was hoped that a hybrid HMM-based method that can be made pairwise would perform at best as well as each individual method at the two extremes, but these hopes were exceeded as the hybrid method outperformed both individual methods across the whole range. A full description of the algorithm can be found in Materials and Methods.

Rarely in biological sequence analysis can anything be inferred from a negative result. The power of the hybrid method is such that there is a potential to predict the presence of new families from a negative result. The benchmark test was repeated going from the worst scores to the best, and now including single-member families. The single member families represent true positives in the cross-validation since they are unique, and all other sequences from the first benchmark represent false positives regardless of whether the correct family is assigned or not. As shown in Figure 3 one in three sequences with an *E*-value >0.05 would belong to a new family for which there is no structural representative.

Development

The development process of the method represents a large part of this work. The process itself may not be interesting but it is worth noting a few of the results and lessons learnt from it.

(i) The hybrid method, like BLAST, uses a substitution matrix. Various substitution matrices were tested but the default BLOSUM62 matrix was found to be the best.

(ii) Similarly to pairwise methods affine gap penalties are used. Parameterization led to the selection of a 'gap open' penalty of 3 and a 'gap extend' penalty of 0.8. The gap penalties, in particular 'gap open', are lower than the BLAST default values. This is due to the fact that the alignment is handed down from the superfamily-based HMM and, being aware of other sequences in the superfamily, is less locally refined on the pair; the choice of gap penalty does not affect the alignment.

(iii) Key (conserved) sites in an HMM contribute much more to the HMM score than other (more variable) sites. The relative importance of each position in the HMM can be calculated and used as a weight. The value from the substitution matrix for each position in the pair alignment was multiplied by the weight for that position in the model to include this information. The more weighting is used, the more the method performs similar to the original HMM score (as you would expect), and the optimal performance was ultimately achieved without weighting.

(iv) The *E*-value scores from the method were calibrated on the cross-validation data, and using a sigmoidal distribution a reasonable fit was produced, except at low error rates, which can potentially be explained by a small number of misclassifications in SCOP. The calibration on SCOP 1.67 was checked against version 1.69 and the difference was not great enough to merit changing it.

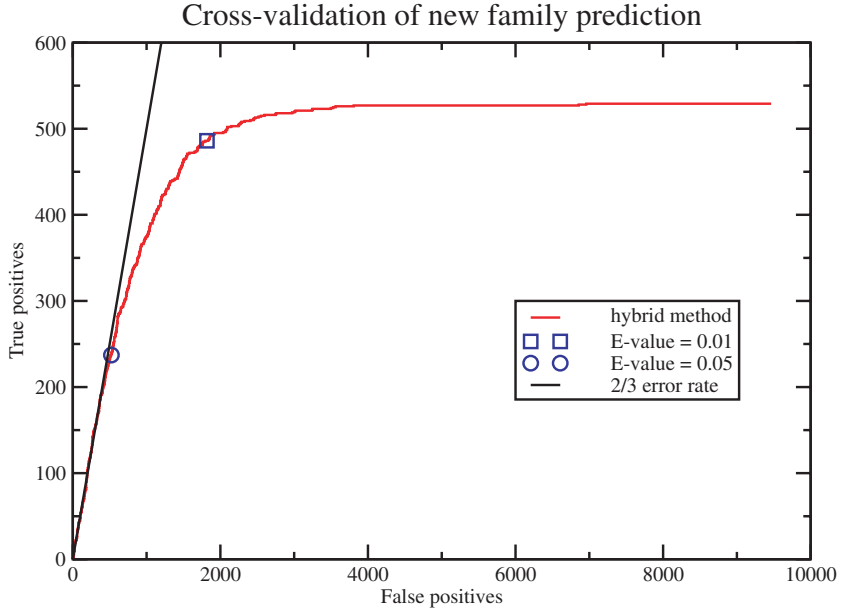

Figure 3. The ability of the hybrid method to predict when a query sequence is the member of a new family within the superfamily as tested on SCOP 1.67. The red curve starts at the origin with the highest score and then decreases in value from left to right.

(v) As alluded to above, inspection of the top-scoring false positives has suggested inconsistently classified domains. Some of these will be re-classified in the next version of SCOP, e.g. sigma factors 1rp3 (N-terminus of chain A) and 1l0o (chain C). For others there is already a note in SCOP, e.g. restriction endonucleases BstyI have local sequence similarity to BglII.

(vi) Four different selection methods for the simple score and null score were tested. The most comprehensive compares the query sequence against each of the superfamily member sequences in turn, and repeats this once for the alignment produced by each HMM; the top two scores to different families from all comparisons are used. The time taken scales with the square of the superfamily size. The selection process which was chosen, however, only does the comparisons for alignments from the two top-scoring HMMs from different families. The time taken scales with double the superfamily size and has almost no loss in performance, but has a significant advantage over using a single alignment (which would only halve the time taken).

Application to SUPERFAMILY

The hybrid method has been integrated into the SUPERFAMILY database as a solution to classifying protein domains at the SCOP family level given an existing superfamily-level assignment; this generates a wealth of data with numerous applications. Query sequences submitted on-line now have the family information as well as the superfamily information that was provided before. Family level assignment has also been added to the scripts and files which are available for external users to install locally. Furthermore, family classification has been calculated for over 300 completely sequenced genomes, and some other sequence sets such as Uniprot (24). These genome assignments are available for browsing via the SUPERFAMILY server web interface as well as on the FTP site for bulk downloads; researchers have already accessed these data before publication and are working with it for genome annotation (25).

Family level classification within SCOP gives far more function-specific information than the broader superfamily level, so it is used in projects such as this one on transcription factor predictions (26), since some superfamilies contain both families which are and are not transcription factors. Future studies of specific gene families and superfamilies (e.g. (10,27) and many more) will benefit not only from having their members identified in the genomes, but also already broken down into their constituent sub-families. Specifically, the grouping of immunoglobulin genome sequences into their sets (V-set, I-set, etc.) is being used by Chothia *et al*. (private communication).

Application of Gene Ontology (GO) (28,29) to SCOP and SUPERFAMILY has been ongoing at the European Bioinformatics Institute (EBI) for some time, but has been hampered by the fact that the superfamily level does not map well to functional ontologies. It is expected that the more functionally specific family level classification will allow a serious improvement in the GO and ultimately lead to the generation of new terms.

Structural genomics projects (30,31) focus on amongst other things, solving the structures of new folds. For sequences of unknown structure which have an assignment to a known superfamily, the hybrid family level classification is able to suggest those which could be members of a new sub-family of the superfamily, for which there is currently no 3D representative. These sequences with the poorest scores at the family level to existing families make excellent targets for structural genomics projects that are interested in solving new members of known folds. In addition to family level classification, the pairwise nature of the HMM-guided hybrid method suggests a closest structural homologue for each domain. This information is useful for selecting templates for 3D protein structure prediction in homology modelling (32). Figure 4 shows the improvement achieved by the hybrid using the CE structural comparison (23) as a benchmark. Unfortunately all automatic structural alignment algorithms are imperfect, so this benchmark can only give some qualitative comparison.

Direct comparisons of these results in SUPERFAMILY to independent work are not possible since this exact problem has not been addressed before. However, a brief comparison of results to one of the related pieces of work listed in Introduction (15) (Bayesian evolutionary tree estimation or Bete classifier) was performed. This method differs greatly in conception from the work of this paper since it attempts to define the sub-grouping de novo, rather than place sequences into an existing classification. The Bete classifier was applied to SH2-domain containing proteins with the result that they suggest 'a new subfamily assignment for Src2_drome and a suggested evolutionary relationship between Nck_human and Drk_drome, Sem5_caeel, Grb2_human and Grb2_chick'. As a result of the work presented here the SUPERFAMILY database has sub-family classifications for most known sequences across all families of known structure; thus, it is trivial to look up these sequences. The hybrid method assigns these sequences to families which group the above sequences in the same way as the Bete classifier, so it is most likely that the original grouping referred to in the Bete paper was a mistaken annotation in SwissProt. In fact, this mistake has subsequently been rectified and and they are now annotated in agreement with both the hybrid method and the Bete method.

Paralogy versus orthology

In the long run the aim of this work is to help further our understanding of the evolution of proteins. As an example of such an application, divergence within a superfamily was examined. We define a superfamily as grouping protein domains with evidence for a common evolutionary ancestor, so there are no evolutionary relationships between superfamilies. Domains within a superfamily however, may diverge from each other over evolution to create new families within the superfamily. Figure 5 shows the relative distribution of paralogous versus orthologous evolutionary events which have lead to the divergence of families within 374 superfamilies in 259 genomes in the SUPERFAMILY database. Comparing paralogy and orthology (33) between families in this way gives an informative and novel overview, since comparisons between individual protein sequences will be

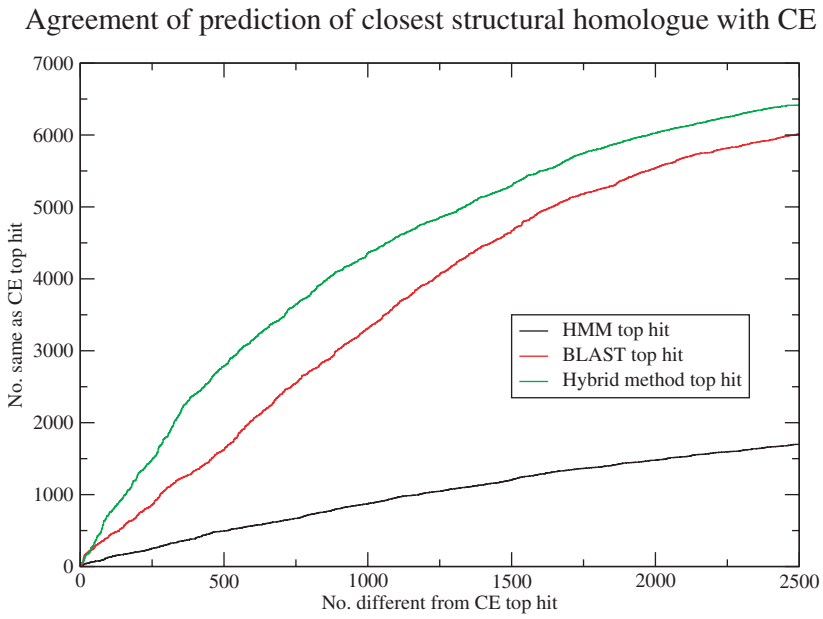

Figure 4. With increasing score, all family members were compared via 3D structure alignment with the CE program. This graph shows the agreement between CE and the other three methods in predicting the closest structural homologue. The top-scoring structures from the three methods were considered to be 'the same' as CE (in agreement) if the structure was assigned a score equal to, or within 0.1 of the top *Z*-score by CE. Sometimes CE assigns equal scores to several structures in the family.

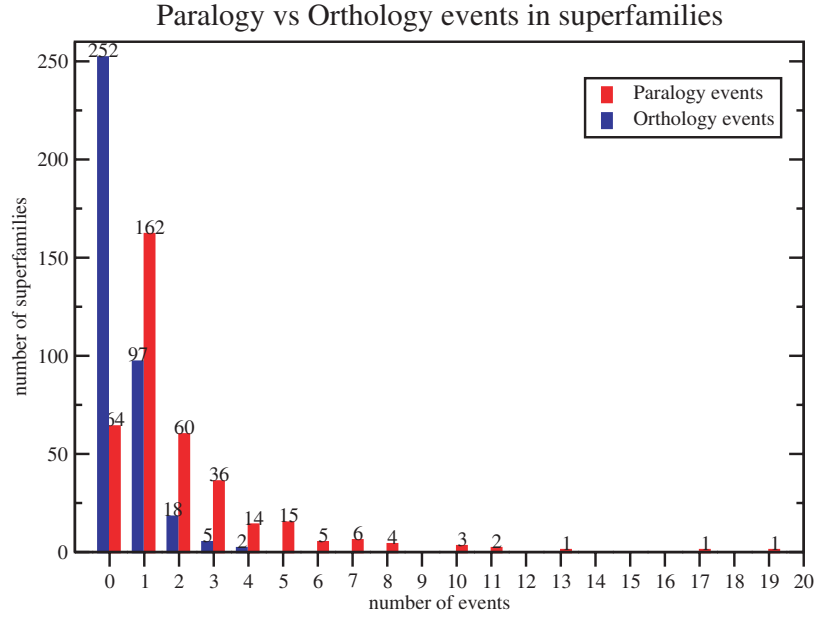

Figure 5. Paralogous versus orthologous evolutionary events that lead to the divergence of a new family within a superfamily. The data for each are shown independently although multiple paralogous and orthologous events can occur in the same superfamily. Superfamilies with no data are not shown, e.g. those with only one known family.

dominated by how you choose to define the proteome, particularly with respect to numbers of splice variants.

The data shown in the graph are conservative, inasmuch as only events which we can be certain about are shown. If a superfamily contains two families which are both present in the genomes containing that superfamily, then we can presume that a paralogous event has taken place. If a superfamily contains two families which occur in separate groups of genomes, i.e. never both in the same genome, then we can presume that an orthologous event has taken place. If, however a superfamily has four families (A, B, C, D), and two families (A, B) occur in one group of genomes with the other two (C, D) occurring in a separate group of genomes, but neither 'A' nor 'B' is ever seen with 'C' or 'D' then we can presume one orthologous event, one paralogous event, plus a third event which could be either orthologous

or paralogous; in this case only one of each event is counted although a third unknown event exists.

The observations shown in Figure 5 support the model of domain duplication followed by subsequent divergence as the dominant evolutionary process creating new function within a protein domain via single-point mutations. This is distinct from previous work (34,35) which addresses gene duplication followed by subsequent recombination in multi-domain proteins via splicing. We also observe that some superfamilies have several or many divergence events, whereas others have none; some structures will inherently be more stable to variations than others. Divergence of domains via paralogy events being more common than via orthology events tells us that an organism is more able to adapt evolutionarily via expansion than by consolidation, although (data not shown) this is more striking in eukaryotes than bacteria.

DISCUSSION

Assignment of sequences to an existing family level classification is exceedingly useful to the biological and bioinformatics communities, and until now there existed no method to achieve this on an all-genome, all-family scale. The problem of classifying sequences into pre-defined subfamilies is relatively easy to solve compared with, for example, initial superfamily level assignment. Therefore even a moderately successful method can be very useful, the critical points are simply that it has been performed, and that it can be easily implemented on an all-genome, all-family scale. Although the most common approach until now has been to use superfamily level HMM scores, these are not good enough; this is actually a testament to the fact that the models are operating (as designed) at the superfamily level. BLAST scores would be a mediocre solution to the problem, but fail on more distant homologues, require an additional scoring step and may need the sequence to be first broken into domains somehow. The hybrid method requires no additional scoring or aligning steps, and the fact that it out-performs (across the whole range of homology distances) classic profile and pairwise methods, is an added bonus; it gives 'money for nothing'.

The hybrid procedure itself was designed to be general, and could equally well be applied at further levels of hierarchy to give sub-sub-family classification. This sort of thing will be essential for the next major release of SCOP which will contain a more fluid and less structured hierarchy (SCOP authors, personal communication). As such the procedure itself does not explicitly attempt to decide by computational and statistical means how families should be divided and grouped but applies a known biologically based classification that more accurately reflects what is seen in nature. However, the hybrid method can be used to test the self-consistency of the classification and has been used to suggest improvements which will be used in subsequent releases of SCOP. Another aspect of the hybrid method which is general, is that it is independent of the models used. PFAM and some other databases use family level models, and it has been suggested in the past that SCOP family level models could be built in addition to the superfamily level models. The hybrid method does not preclude the use of family level models (or other augmentations) which could further improve performance. These would come at the cost of the additional work required to develop and maintain the extra models; the hybrid method could even be used to group sequences for such model building.

A few of the immediately obvious applications (most of which are already underway) include functional annotation of genomes, furthering development of GO for SCOP, studies of individual sequence families, prediction of new families for structural genomics, suggesting the most closely related structure for homology modelling, working with functional sets of domains such as transcription factors, improving the existing hierarchy in SCOP and application to other databases such as Gene3D (36) and PFAM. The hybrid method can be applied to the sub-level of any existing classification for which there is some (preferably profile-based) homology search at a higher level, and should out-perform classical pairwise and profile methods regardless of the clustering properties and qualities of the classification. This work will be an invaluable tool for research into the evolution of proteins, and we can already see in the results here a new and informative overview of the balance of paralogous versus orthologous divergence, and the implications that has for expansive versus consolidatory molecular evolution.

ACKNOWLEDGEMENTS

The author would like to thank Michael Nilges for his support during this work. Funding to pay the Open Access publication charges for this article was provided by the Pasteur Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D.R. *et al.* (2001) InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
2. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all protein of known structure. *J. Mol. Biol.*, **313**, 903–919.
3. Madera,M., Vogel,C., Kummerfeld,S., Chothia,C. and Gough,J. (2004) The superfamily database in 2004: additions and improvements. *Nucleic Acids Res.*, **32**, 235–239.
4. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) The PFAM protein families database. *Nucleic Acids Res.*, **28**, 263–266.
5. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
6. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
7. Horn,F., Weare,J., Beukers,M.W., Horsch,S., Bairoch,A., Chen,W., Edvardsen,O., Campagne,F. and Vriend,G. (1998) Gpcrdb: an information system for *g* protein-coupled receptors. *Nucleic Acids Res.*, **26**, 277–81.
8. Vinogradov,S.N., Hoogewijs,D., Bailly,X., Arrendondo-Peter,R., Guertin,M., Gough,J., Dewilde,S., Moens,L. and Vanfleteren,J.R. (2005) Three globin lineages belonging to two structural classes in genomes from the three kingdoms of life. *Proc. Natl Acad. Sci. USA*, **102**, 11385–11389.
9. Bashford,D., Chothia,C. and Lesk,A.M. (1987) Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.*, **196**, 199–216.

10. Ravasi,T., Huber,T., Zavolan,M., Forrest,A., Gaasterland,T., Grimmond,S., Hume,D. and RIKEN GER Group,G.M. (2003) Systematic characterization of the zinc-finger-containing proteins in the mouse transcriptome. *Genome Res*., **13**, 1430–1442.
11. Forrest,A.R., Ravasi,T., Taylor,D., Huber,T., Hume,D.A., Grimmond,S. and RIKEN GER Group,G.M. (2003) Phosphoregulators: protein kinases and protein phosphatases of mouse. *Genome Res*., **13**, 1443–1454.
12. Karchin,R., Karplus,K. and Haussler,D. (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, **18**, 147–159.
13. Whelan,S., Debakker,P.I.W. and Goldman,N. (2003) Pandit: a database of protein and associated nucleotide domains with inferred trees. *Bioinformatics*, **19**, 1556–1563.
14. Thomas,P.D., Campbell,M.J., Kejariwal,A., Mi,H., Karlak,B., Daverman,R., Diemer,K., Muruganujan,A. and Narechania,A. (2003) Panther: a library of protein families and subfamilies indexed by function. *Genome Res*., **13**, 2129–2141.
15. Sjolander,K. (1998) Phylogenetic inference in protein superfamilies: analysis of SH2 domains. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., 165–174.
16. Brown,D., Krishnamurthy,N., Dale,J.M., Christopher,W. and Sjolander,K. (2005) Subfamily HMMs in functional genomics. *Pac. Symp. Biocomput*. 2005, 322–333.
17. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
18. Karplus,K., Karchin,R., Shackelford,G. and Hughey,R. (2005) Calibrating *E*-values for hidden Markov models with reverse-sequence null models. *Bioinformatics*, **21**, 4107–4115.
19. Park,J., Karplus,K., Barrett,C., Hughey,R., Haussler,D., Hubbard,T. and Chothia,C. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol*., **284**, 1201–1210.
20. Brenner,S.E., Chothia,C. and Hubbard,T.J. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.
21. Madera,M. and Gough,J. (2002) A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res*., **30**, 4321–4328.
22. Brenner,S.E., Koehl,P. and Levitt,M. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res*., **28**, 254–256.
23. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*., **11**, 739–747.
24. Leinonen,R., Diez,F.G., Binns,D., Fleischmann,W., Lopez,R. and Apweiler,R. (2004) Uniprot archive. *Bioinformatics*, **20**, 3236–3237.
25. Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R., Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T. *et al*. (2006) *Escherichia coli* K-12: a cooperatively developed annotation snapshot—2005. *Nucleic Acids Res*., **34**, 1–9.
26. Kummerfeld,S.K. and Teichmann,S.A. (2006) DBD: a transcription factor prediction database. *Nucleic Acids Res*., **34**, D74–D81.
27. Podar,M., Eads,J.R. and Richardson,T.H. (2005) Evolution of a microbial nitrilase gene family: a comparative and environmental genomics study. *BMC Evol. Biol*., **5**, 1–13.
28. Camon,E., Barrell,D., Brooksbank,C., Magrane,M. and Apweiler,R. (2003) The Gene Ontology Annotation (GOA) project: application of GO in Swiss-Prot, Trembl and Interpro. *Comp. Funct. Genom*., **4**, 71–74.
29. Camon,E., Barrell,D., Lee,V., Dimmer,E. and Apweiler,R. (2003) Gene Ontology annotation database—an integrated resource of GO annotations to Uniprot knowledgebase. *In Silico Biol*., **4**, 2.
30. Bertone,P., Kluger,Y., Lan,N., Zheng,D., Christendat,D., Yee,A., Edwards,A., Arrowsmith,C., Montelione,G. and Gerstein,M. (2001) SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res*., **29**, 2884–2998.
31. Brenner,S., Barken,D. and Levitt,M. (1999) The presage database for structural genomics. *Nucleic Acids Res*., **27**, 251–253.
32. Moult,J., Pedersen,J.T., Judson,R. and Fidelis,K. (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins*, **23**, ii–v.
33. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet*., **39**, 309–338.
34. Vogel,C., Berzuini,C., Bashton,M., Gough,J. and Teichmann,S. (2004) Supra-domains—evolutionary units larger than single domains. *J. Mol. Biol*., **336**, 809–823.
35. Gough,J. (2005) Convergent evolution of domain architectures (is rare). *Bioinformatics*, **21**, 1464–1471.
36. Buchan,D.W., Shepherd,A.J., Lee,D., Pearl,F.M., Rison,S.C., Thornton,J.M. and Orengo,C.A. (2002) Gene3d: structural assignment for whole genes and genomes using the cath domain structure database. *Genome Res*., **12**, 503–514.