ABSTRACT

Motivation: A recent development in sequence-based remote homologue detection is the introduction of profile–profile comparison methods. These are more powerful than previous technologies and can detect potentially homologous relationships missed by structural classifications such as CATH and SCOP. As structural classifications traditionally act as the gold standard of homology this poses a challenge in benchmarking them.

Results: We present a novel approach which allows an accurate benchmark of these methods against the CATH structural classification. We then apply this approach to assess the accuracy of a range of publicly available methods for remote homologue detection including several profile–profile methods (COMPASS, HHSearch, PRC) from two perspectives. First, in distinguishing homologous domains from non-homologues and second, in annotating proteomes with structural domain families. PRC is shown to be the best method for distinguishing homologues. We show that SAM is the best practical method for annotating genomes, whilst using COMPASS for the most remote homologues would increase coverage. Finally, we introduce a simple approach to increase the sensitivity of remote homologue detection by up to 10%. This is achieved by combining multiple methods with a jury vote.

Contact: reid@biochem.ucl.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

The identification of remote homologues is a central problem in bioinformatics. New tools to accomplish this task appear frequently, but rarely are they rigorously benchmarked against a range of other software, under varying conditions. Benchmarking is crucial, both to determine the best tool for a particular job and to determine a discriminating E-value threshold.

Brenner et al. (1998) showed that sequence–sequence (often termed pairwise) methods such as BLAST (Altschul et al., 1997) can detect most relationships of >30% sequence identity. Such methods compare individual sequences and use substitution matrices to model residue substitutions which are likely between homologues in general. Park et al. (1998) showed that profile–sequence methods could find three times as many homologues as sequence–sequence methods at sequence identities below 30%. These profile–sequence methods, including HMMer (Eddy, 1996), SAM (Karplus et al., 1998) and PSI-BLAST (Altschul et al., 1997) have become widely used for detecting remote homologues. Profiles are built using a multiple alignment of one of the sequences and its relatives to model sequence changes expected within that family. Additionally, HMMer and SAM employ Hidden Markov Models (HMMs) which incorporate position-specific gap penalties. More recently, profile–profile methods have been introduced which use a profile to search a database of profiles. There is therefore specific evolutionary information in both the query and the target and thus more remote relationships can be detected. Such methods include COMPASS (Sadreyev and Grishin, 2003), prof_sim (Yona and Levitt, 2002), LAMA (Pietrokovski, 1996), PRC (Madera, 2006) and HHSearch (Soding, 2005). Of these methods, COMPASS, HHSearch and PRC are used in this article. COMPASS aligns profiles against profiles, whereas HHSearch and PRC align HMMs. All three use a log-sum-of-odds score, which is a generalization of the log-odds score used by sequence–profile methods. Optimum alignments are determined by COMPASS using the Smith–Waterman algorithm, whereas HHSearch uses the viterbi algorithm and PRC can use either viterbi or forward. They all use some kind of distribution fitting to calculate E-values although HHSearch requires a calibration step. There has not previously been a comprehensive benchmark across sequence–sequence, profile–sequence and profile–profile methods. It is important to determine their relative performance in specific remote homology detection tasks in order to make an informed choice of which methods should be used for different tasks.

The fundamental requirement of a benchmark for remote homologue detection is a gold standard dataset of known evolutionary relationships between protein domains. Previously, 3D structure comparison has been shown to detect more distant evolutionary relationships than sequence comparison (Chothia and Lesk, 1986) and can therefore hint at relationships missed by sequence analysis. Two major domain structure classifications [SCOP—(Murzin et al., 1995), and CATH—(Greene et al., 2007)] use structural similarity to suggest homology which is then further validated by seeking additional evidence, e.g. from sequence profile matches.
or experimental evidence of functional similarity. Structural classifications have therefore been exploited extensively in benchmarking sequence-based remote homology detection methods. SCOP, FSSP (Holm and Sander, 1994) and CATH have all been used in this context by, e.g. Park et al. (1998), Sadreyev and Grishin (2003) and Sillitoe et al. (2005), respectively. Bateman and Finn (2007) used Pfam clans (Finn et al., 2006), which are based on a mixture of sequence and structural evidence.

Surprisingly, poor performance has been reported for profile–profile methods using benchmarks based solely on the SCOP structural classification (Soding, 2005). On close inspection, it was shown that this was due to potentially homologous domain sequences which had been classified as unrelated. This occurred largely due to previous lack of evidence for homology between these domains. Two domains are homologous if they have descended from a single ancestral domain. Over time homologues tend to diverge in both sequence and structure. In general, it is easier to find similarity in their structures than in their sequences, particularly when they have diverged greatly. However, structural similarity alone is not sufficient to guarantee homology as there may be physical constraints on folding and limited topologies available. Therefore, it is necessary to build up several lines of evidence to describe domains as homologous. CATH or SCOP initially group structurally similar domains into fold groups. Within fold groups, homologous relationships are subsequently recognized using one or more elements of independent evidence. For example, statistically significant sequence similarity or experimental verification of functional similarity (e.g. from the literature, bound ligands or identification of common catalytic residues).

Recent analyses of CATH have shown that homologues can diverge considerably in their structures (Reeves et al., 2006) and in some families a 5-fold or more variation in size is observed between extremely distant relatives. It is now apparent that the most highly sensitive profile–profile methods are detecting significant sequence patterns suggestive of homology between domains which are highly structurally divergent. In these cases, any structural similarity would probably be missed on manual inspection (e.g. in SCOP) or fall below the stringent automatic thresholds currently used for classifying homologues in CATH. These thresholds on structural similarity were determined empirically based on earlier analyses of homologous families, but the more distant relationships recently revealed by profile–profile methods (Sadreyev et al., 2007) are prompting a re-examination of these thresholds.

Consequently, the status of these putative homologues is uncertain and they should therefore not be considered as non-homologous when benchmarking methods for remote homology detection. β-Propellers, for example, are thought to have evolved by duplication of the β-sheet blades and inheritance of blades between structures (Jawad and Paoli, 2002). Their evolution is therefore somewhat unusual as structural embellishments occur in the core of the domain and make superposition difficult. Gough et al. (2001) identified such examples in SCOP by examining false positives produced by SAM and provide a Perl function coding exceptions for a SCOP-based benchmark on the web (http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/ruleset_1.65.html). Soding showed that such examples could be accounted for automatically by using a combination of sequence and structural evidence which is more powerful than taking either measure in isolation (Soding, 2005). He used a measure for local structural similarity to avoid similar problems in SCOP when benchmarking HHSearch. False positive hits were excluded by using a score cut-off from the MaxSub structural comparison algorithm (Siew et al., 2000).

In addition to the choice of structural classification used for benchmarking, test sets of varying difficulty affect the separation of methods. At high sequence identities sequence–sequence and profile–sequence methods detect a more similar number of true relationships than at lower sequence identities. Park et al. (1998) and Sillitoe et al. (2005) used <40 and <35% non-redundant (nr) test sets respectively for benchmarking profile–sequence methods. Soding (2005) used a <20% nr set of query sequences to benchmark the same type of method. Casbon and Saqi (2006) used a SCOP dataset where homologous pairs had <10% sequence identity in benchmarking HHSearch.

Furthermore, benchmarking methods differ in their definition of a false positive. Yona and Levitt consider matches between members of the same SCOP/CATH superfamily to be true and anything else false. Park et al. (1998), Sillitoe et al. (2005) and Soding (2005) consider matches between members of the same fold but differing superfamily to be ambiguous and therefore discounted. Sadreyev and Grishin (2003) used a definition based on FSSP Z-scores. Muller et al. (1999) consider fold matches to be false in benchmarking PSI-BLAST for genome annotation.

Aside from Muller et al. these benchmarks all focus on a general remote homology detection task, i.e. distinguishing homologues from non-homologues. However, remote homologue detection is frequently used to annotate genomes. To our knowledge, there has been no work comparing the abilities of different methods to annotate genomes with structural domains (Muller et al. give no comparator for PSI-BLAST).

To date, there has been no coherent benchmark encompassing the wide range of methods now available for sequence-based remote homology detection. Although Bateman and Finn (2007) benchmarked a range of profile–profile methods using Pfam clans, there is no comparison with profile–sequence methods and Pfam clans provide a relatively sparse test set compared to CATH or SCOP. The various benchmarks described in the literature differ significantly and are not easily comparable. Here, we explore the relative performance of seven methods from among the sequence–sequence, profile–sequence and profile–profile classes in more depth than previous work. This article aims to measure the effect of datasets and benchmarking strategies when assessing the performance of homologue recognition methods. Two types of benchmark are performed, reflecting different applications of remote homology detection. The alipos benchmark models the general task of separating homologues from non-homologues whilst the tophit benchmark captures the task of annotating proteomes with structural domains. In order to account for
the erroneous high error rate of profile–profile methods, we introduce a heuristic method based on a leading structural comparison algorithm (SSAP) to exclude false positives with good $E$-values and high structural similarity. The method is compared to a set of exceptions defined by expert analysis. Park et al. (1998) have shown that different homologue detection methods identify similar sets of homologues and relatively dissimilar sets of non-homologues. This makes intuitive sense in two ways. First, the methods have been trained to detect the same types of homologues, while false hits are the result of imperfections of the particular approach or implementation. Second there are more false positives to choose from. Here we show that by combining different methods with a simple jury approach, coverage can be increased by up to 10%.

2 METHODS

2.1 Datasets

The broadest dataset (nr35) consists of representative sequences from CATH v3.0.0. Initially, all CATH sequences were clustered into families at 35% sequence identity with 80% overlap cutoff by directed multi-linkage clustering. Cluster representatives were chosen by picking the member with the highest resolution structure and with a length closest to the average. This set contains many pairwise relationships which are trivial examples of remote homologues, in the sense that they can be detected by sequence–sequence methods (i.e. BLAST). To better examine the ability of the most sensitive methods the nr10 dataset was created by clustering at 10% sequence identity within the nr35 set. A 60% overlap cutoff was used to take account of the greater diversity at this evolutionary distance. Cluster representatives for this set were selected by greatest length. Both datasets were filtered to exclude sequences with no superfamily partner. The nr35 dataset comprised 6570 sequences in 937 superfamilies with 210 075 homologous pairs by the allpos rule and 6570 by the tophit rule. The nr10 dataset comprised 1434 sequences in 288 superfamilies with 16464 homologous pairs by the allpos rule and 1434 by the tophit rule.

2.2 Profile/model building

Each sequence in the nr35 dataset was used as a seed for the SAM3.4 target2k program (Karplus et al., 1998) to build an HMM representing the superfamily of that seed. This procedure initially performs a BLAST on the GenBank non-redundant database of all known protein sequences (at max $E$-value 400) to produce a reduced-size database within which to detect homologues. An iterative HMM procedure then builds an alignment, using more and more relaxed $E$-values, up to an $E$-value of 0.005. These alignments were used for benchmarking PSI-BLAST. The alignments were filtered to remove positions aligned to gaps in the seed sequences before being used to build COMPASS profiles using mk_compass_db (part of the COMPASS software), as recommended by Ruslan Sadreyev (personal communication). The SAM3.4 program w0.5 was used to build HMMs from the original alignments and these were used for benchmarking SAM. The SAM models were converted to HMMer format using Madera’s convert.pl script (http://www.mrc-lmb.cam.ac.uk/ genomes/julian/convert/convert.html) and calibrated with 1000 random sequences. These HMMer models were used in benchmarking both HMMer and HHSearch. PRC was benchmarked using the SAM models converted to PRC format using the convert_to_prc program (provided with PRC).

2.3 Benchmarking procedure

Both datasets (nr35, nr10) were compared, all-against-all, using each method. In the case of BLAST, this meant sequences were scanned against sequences, for SAM/HMMer HMMs against sequences, for COMPASS profiles against profiles and for PRC/HHSearch HMMs against HMMs. For PSI-BLAST, profiles were scanned against sequences, allowing up to 20 iterations. HHSearch was used without structural information (ssm = 0). Local or local-local scoring was used for all methods. Although domains are being compared in the benchmark, we are assaying performance in annotating genomes and detecting very remote relationships within families, for which local scoring is most appropriate. As suggested by Madera and Gough (2002) other parameters are defaults, such that the relatively inexperienced user can achieve the same performance. Equally, the methods presented here are those most easily available for download from the web.

The relationship between each pair of domains was defined by the CATH domain structure classification. If a method of remote homology detection matches two members of the same superfamily, it is considered to have recognized a true relationship. Matches between members of different superfamilies that are within the same topology are treated as ambiguous and are not counted as true or false. Matches between superfamilies that are not within the same topology are counted as false hits. The benchmark is an all-against-all search within each dataset and it is therefore necessary to exclude trivial matches between a query and itself, or the profile/model for which it was the seed. Exceptions to the benchmarking procedure (identified by expert analysis or the SAS8 heuristic) are effectively counted as topology hits and therefore ignored.

Two different rules for counting hits were implemented. With the allpos rule every non-trivial relationship is counted. This creates a benchmark that reflects how well a method can separate homologues from non-homologues. With the tophit rule only the first non-trivial hit for each query is counted. This simulates the case of genome annotation, where only the best scoring hit is considered.

2.4 Coverage versus error plots

For each method, all the hits from the all-against-all scan are sorted by $E$-value and for successive 10-fold $E$-value cutoffs, the coverage and error rate are plotted. This is somewhat like the traditional ROC curve, which plots the proportions of true and false positives. Here, however, error rate (or Errors Per Query, EPQ) is the number of false positives divided by the number of false positives plus the number of true positives for a certain $E$-value. This gives a more intuitive reading of the results than plotting the proportion of total errors in the dataset, or the raw number of false positives, i.e. for 0.05 EPQ there are 5% false positives and 95% true positives. Coverage, on the y axis, shows the proportion of true positives found for a particular $E$-value. For the allpos rule, this is calculated using the total number of pairwise homologues in the dataset. For the tophit rule, the number of queries is used.

2.5 Combining different methods to increase specificity

A simple approach was employed to combine the results of multiple methods. All-against-all hits up to an $E$-value of 10 for two or more methods are compared. Hits are discarded unless all methods have scored that hit. For those remaining hits, the combined $E$-value (CEV) for each is calculated thusly,

$$\text{CEV} = 10 \sum_{i=1}^{n} E_i,$$

where $E_1...n$ are the $E$-values from each method for a specific pairwise hit and $n$ is the number of methods.
All possible permutations of methods (excluding BLAST) were subjected to this analysis, using different datasets and the allpos rule. Only permutations of profile–sequence methods (appropriate for genome annotation) were used with the tophit rule.

3 RESULTS AND DISCUSSION

3.1 A heuristic exception rule exploiting structural similarity to recognize putative homologues

Profile–profile, sequence-based methods for remote homologue detection have previously been found to detect relationships between domains which are classified in different superfamilies in SCOP but are potentially homologous on close inspection (Soding, 2005). In these cases, homology may be confirmed by identifying structural similarity or evidence of functional similarity in combination with the significant sequence similarity already detected. We wanted to create a general, heuristic, rigorously derived rule which would allow these ambiguous relationships to be identified and excluded from a benchmark of sequence-based methods of remote homology detection. We explored the validity of a heuristic method based on structural comparison scores, similar to Soding’s approach of using a MaxSub cutoff. To do this, we used a well established structural comparison algorithm and compared the performance of this approach with that of using a manually validated set of exceptions.

For all domain pair matches (identified by PRC on the nr35 dataset up to an E-value of 0.1) involving different CATH topologies, manually curated exceptions to the CATH classification were determined by examining the structural superpositions (to determine whether sufficient topological features suggest the domains evolved from a common ancestor), sequence alignments, functional annotations, catalytic residues and the literature. Those pairs with several lines of evidence to suggest homology were considered exceptions. The aim was to make a fairer benchmark by excluding these putative homologues.

For the heuristic rule, we have used the SSAP structural comparison algorithm (Orengo and Taylor, 1996) and the SAS score (Subbiah et al., 1993) to score structural alignments. The SAS score has proven to be a better discriminator at the fold level than native structural comparison scores (Kolodny et al., 2005).

Figure 1 shows how accurately the manually curated exceptions are captured by using a heuristic based on SSAP structural alignment and SAS scores. A SAS score of 8 gives 86% coverage of the curated exceptions with 12% additional exceptions not identified in the manually curated set. Lower thresholds result in much poorer coverage of the manually curated exceptions while higher thresholds cause a rapid increase in errors with relatively little gain in coverage. This high error rate is in part due to the difficulty in accurately scoring structural similarity in general. The errors are explored in detail subsequently.

Figure 2 shows that the performance of PRC using the heuristic exception rules SAS 8, 9 or 10 is very similar to that achieved using the manually curated exceptions rule. SAS 8 is, however, the most appropriate rule since it is less error prone than SAS9 with no significant loss of coverage. Although SAS9 achieves closer performance to the curated exceptions in a benchmark, fitting less closely to PRC will reduce the bias incurred by benchmarking the exceptions solely on this method. Details of both manually curated and SAS8 exceptions are presented in Supplementary Table 1.

By far, the most common matches between putative homologues for both manually curated and SAS8 exceptions are between the FAD/NAD(P) binding domain topology (3.50.50) and the Rossman fold (3.40.50). These folds are from the ββα and αβα sandwich architectures, respectively. Previous analyses have suggested that these may be very remote homologues (Harrison et al., 2002). The SAS8 heuristic captures all of the β-propeller exceptions (2.115, 2.120, 2.130 and 2.140 architectures) and all box/horseshoe exceptions (3.70 and 3.80 architectures). Thus, 87.6% of the heuristic exceptions are accounted for. Several smaller classes of exception were noted during manual curation, however, these were either re-classified in CATH or appeared at E-values ≥0.1.
Fig. 3. Performance of all methods using the alipos and SAS8 rules on the nr35 (a) and nr10 (b) datasets. Runs were performed to an E-value of 10.

Several exceptions identified by the SAS8 rule were not recognized by manual curation, comprising 12.4% of the SAS8 exceptions. The most common three of these classes, comprising nearly half, are discussed here. The Aminoglycoside 3’-phosphotransferase (3.90.1200) versus Phosphotransferase (1.10.510) class is the first of these (2.3% of the SAS8 exceptions). Both topologies consist of a single superfamily, each implicated in protein kinase activity. Only two members of the 3.90.1200.10 superfamily were in the nr35 dataset and one of these (1nd4A01) is now being reclassified in CATH. The other (2b0kC02) does show some local structural similarity to members of the 1.10.510.10 superfamily following superposition. For the best match (with 1wbsA02), the SAS score is 6.37 and the PRC E-value is $1.7E-5$. The overall topologies are clearly different however.

The Class II MHC (3.10.320) and Class I MHC (3.30.500) folds (2.1% of the SAS8 exceptions) each contain only one superfamily. Most of the domains within each fold hit the other fold below an E-value of 0.01. The full crystal structures (e.g. 1ktd versus 1keg) show these folds are highly similar, with very good superposition. Half the domain for the 3.10.320 examples is however provided by a different chain. There is homology between the CATH domains, but the functional domain has been split between two chains in one case. This is therefore a very reasonable exception.

Matches between the TIM Barrel (3.20.20) and Aspartate Aminotransferase (3.40.640) folds (1.3% of the SAS8 exceptions) only occur at E-values $>$0.0005. The best match by PRC (1hx0A01 versus 1p3wA02) only just makes the SAS score cutoff with a score of 7.965 and has an RMSD of 15.93. On superposition, members of these superfamilies have no apparent structural similarity other than $\alpha\beta$ motifs. Both superfamilies are large and we believe that these matches have genuinely occurred by chance.

3.2 Detecting remote homologues
3.2.1 Distinguishing homologues from non-homologues (alipos rule) All seven methods (BLAST, PSI-BLAST, HMMer, SAM, HHSearch, COMPASS and PRC) were benchmarked with both nr35 and nr10 datasets using the alipos scoring rule (Fig. 3) which captures the ability of the methods to distinguish all homologues from all non-homologues. The SAS8 exceptions rule is used here and throughout the rest of the article to exclude false positives (i.e. putative homologues) with SAS score $<8$ for all methods.

Supplementary Table 2 shows E-value cutoffs for each method at differing error rates. It has been noted previously for alipos-style benchmarks (Park et al., 1998) that profile–sequence methods achieve up to 3 times more coverage than BLAST for a fixed error rate when considering homologous pairs with sequence identities <30%. That result is replicated here with PSI-BLAST (25.9%) and SAM (24.4%) achieving 3 times the coverage of BLAST (8.6%) with the nr35 dataset. On all datasets PRC is the best method, performing almost 2.5 times better than the best profile–sequence method at 0.05 EPQ (22.1% coverage versus 8.8% for SAM) on the more difficult dataset (nr10). This almost equals the increase in performance seen for profile–sequence methods over BLAST, although only on very remote homologues (<10% sequence identity). All profile–profile methods are better than all profile–sequence methods on all datasets at an error rate of 0.05 EPQ. For low error rates (e.g. 0.01 EPQ), PSI-BLAST and SAM achieve similar performance to profile-profile methods on the nr35 dataset.

3.2.2 Annotating genomes (tophit rule) All seven methods were benchmarked with both nr35 and nr10 datasets using the tophit rule, which only scores the first true positive for each query (Fig. 4). Supplementary Table 3 shows E-value cutoffs for each method at differing error rates.

On the nr35 dataset (Fig. 4a), profile–profile methods slightly outperform profile–sequence methods. The best profile-profile method is PRC which achieves 4.8% greater coverage than SAM (the best profile-sequence) at 0.01 EPQ (COMPASS has almost equal coverage to PRC). Interestingly, PSI-BLAST performs as well as HMMer (81.4 and 81.2%, respectively at 0.05 EPQ). BLAST’s performance of 70% coverage at 0.05 EPQ shows that this is a relatively easy dataset. In fact, with the tophit rule, it is relatively easy for all methods to get high coverage because only the nearest neighbour need be identified.

The superior sensitivity of profile–profile methods is much clearer on the more difficult dataset (nr10—Fig. 4b).
COMPASS achieves ~30% greater coverage than HMMer at 0.01 EPQ and ~20% greater coverage than SAM at 0.05 EPQ. COMPASS outperforms PRC on these more difficult datasets in contrast to the benchmark using the allpos rule, where PRC is always the superior method.

3.2.3 Synthesis Overall, COMPASS is the best performing method, and as might be expected all profile–profile methods outperform all profile–sequence methods. However, for the nr35 dataset, SAM performs almost as well as the profile–profile methods at 0.02 EPQ. In practice, it is too computationally expensive for most bioinformatics labs to use profile–profile methods to annotate genomes, since a profile would have to be built for each non-redundant protein sequence. However, for a subset that does not score well with profile–sequence methods or where more powerful computing services are available it may be worthwhile using COMPASS or PRC. COMPASS produces >20% increase in coverage for the most remote homologues at 0.05 EPQ over the closest profile–sequence method (SAM).

Interestingly, PSI-BLAST performs better than HMMer using both the tophit rule and the allpos rule. To our knowledge this has not been reported before in the literature. The reason may be the way in which PSI-BLAST was used. Generally, PSI-BLAST is used to build a profile from a single query sequence. In this case, however, an initial profile was built using target2k. Therefore, PSI-BLAST had the advantage of HMM technology in building what may have been a more powerful profile than can be built by PSI-BLAST itself. Additionally, COMPASS is shown to be more powerful than HHSearch in contradiction to other reports (Bateman and Finn, 2007; Soding, 2005). This may also be due to HMM-based profile building.

3.3 Combining methods improves performance by excluding false positives

The effect of combining different methods of remote homology detection was explored to see whether this improved performance by filtering out false positives. Figure 5a shows the best performing combined methods using the allpos rule, on the nr10 dataset. The best performing combination was COMPASS and PRC, giving an increase of 2–3% over the best single over a large range of error rates. For moderate homologues, a
Combined E-value (CEV; see Methods section for definition) cutoff of $3E^{-21}$, relating to 0.05EPQ is suggested.

Figures 5b shows a benchmark on the nr10 dataset using the tophit rule. All the profile-sequence methods are shown individually and in combination. Profile-profile methods are excluded as they are not practical for large-scale genome annotation. A significant increase in coverage of 10% can be achieved at a low error rate of 0.01EPQ using PSI-BLAST combined with SAM on very remote homologues (nr10). At higher error rates (>0.02EPQ) SAM is again the best performer. This is because the combination of methods allows for an increase in specificity, but is not expected to greatly increase sensitivity. For very remote homologues, a CEV (Combined E-Value) cut-off of $5E^{-7}$ is appropriate for PSI-BLAST combined with SAM.

For both tophit and allpos, the best combined methods are the two best performing single methods. This might be because they are also complementary, as they are based on different technologies (PSSMs and HMMs). In the tophit case this is SAM and PSI-BLAST, in the allpos case this is COMPASS and PRC.

4 CONCLUSIONS

4.1 Heuristic exceptions rule

The need to make exceptions to structural classifications in benchmarking remote homology detection methods reflects a shift in our abilities to detect homology from sequence. Until recently, many extremely remote homologues could only be detected from structural similarity which was then validated by manual analysis, e.g. inspection of binding sites and catalytic residues or consideration of experimental data reported in the literature. Structural classifications like CATH and SCOP were accepted as reliable gold standards of homology. It now appears that some extremely remote homologues diverge beyond easy detection by available structure comparison methods whilst a sequence signal can still be detected by the more recently developed and highly sensitive profile-profile based methods. We have presented an effective solution to this problem. By considering both the statistical significance of a sequence match and using a SAS score cutoff for structural similarity, a reliable benchmark can be produced for these very sensitive profile-profile methods. To our knowledge there is no similar approach as rigorously derived. In the future, structural classifications such as CATH and SCOP will clearly benefit from using these methods to detect more remote homologues.

4.2 The importance of benchmarking for application

In benchmarking methods of remote homology detection, it is important to bear in mind the task for which they will be used. Different rules for counting true and false positives reflect whether one is interested in annotating genomes (only the closest homologue is required) or whether a score cutoff is needed to determine whether individual domain pairs are homologous (separation of all homologues from non-homologues). Methods perform more or less well at different tasks and this knowledge is invaluable for ensuring data quality. E-value cutoffs for methods other than BLAST are very different for different applications and some are not at all accurate over the ranges often used. Suitable E-value cutoffs for each method and application are presented in supplementary material.

4.3 Relative performance of methods

Having established the S4S8 exception rule and suitable benchmarks to model both the simple scoring of homologues and genome annotation, PRC was shown to be the best method for distinguishing homologues and non-homologues and thus for clustering together families of sequence domains. In fact, for distant homologues (<10% sequence identity) PRC is 2.5 times better than the best profile-sequence method at clustering homologues into superfamilies. Profile-profile methods and, especially, PRC are greatly increasing our ability to recognize remote homologues. COMPASS was shown to be the best method overall for annotating genomes at low sequence identities (<10%). At sequence identities of <35%, however, PRC is equally effective and there is only ~5% increase in coverage over the best profile sequence method (SAM). It is not practical to use profile-profile methods for annotating whole genomes, but they could perhaps be applied to annotate the most remote homologues where profile-sequence methods fail. PSI-BLAST performed surprisingly well in the benchmarks presented. This may be due to the way in which it was used. Profiles were built using SAM’s T2K program and then used in up to 20 iterations of PSI-BLAST. The use of SAM T2K brings an element of HMM technology into PSI-BLAST.

4.4 Combining methods improves performance

When determining relatives for a protein of interest, an investigator may compare the results of multiple methods to increase the likelihood of a correct assignment. Here, an automated approach has been described using the combined results of multiple methods to improve assignment. When annotating genomes, combining profile-sequence methods gives a large increase in performance of 10% at a 1% error rate. This increase is achieved with a combination of SAM and PSI-BLAST.

ACKNOWLEDGEMENTS

A.J.R. acknowledges funding by the BBSRC. C.Y. acknowledges funding by the EU Biosapiens Network of Excellence.

Conflict of Interest: none declared.

REFERENCES


