Bioinformatics, 32(3), 2016, 338–344
doi: 10.1093/bioinformatics/btv538
Advance Access Publication Date: 10 October 2015
Original Paper

OXFORD

Sequence analysis

Rapid and enhanced remote homology detection by cascading hidden Markov model searches in sequence space

Swati Kaushik[†], Anu G. Nair[‡], Eshita Mutt[§], Hari Prasanna Subramanian and Ramanathan Sowdhamini*

National Centre for Biological Sciences, Tata Institute of Fundamental Research, GKVK Campus, Bangalore 560065, India

Associate Editor: John Hancock

*To whom correspondence should be addressed.
[†]Present address: University of California, San Francisco, Helen Diller Family Comprehensive Cancer Center, San Francisco, CA 94158, USA
[‡]Present address: School of Computer Science and Communication, Royal Institute of Technology, SE 100 44 Stockholm, Sweden
[§]Present address: Paul Scherrer Institute (PSI), 5232 Villigen, Switzerland

Received on February 19, 2015; revised on August 12, 2015; accepted on September 6, 2015

Abstract

Motivation: In the post-genomic era, automatic annotation of protein sequences using computational homology-based methods is highly desirable. However, often protein sequences diverge to an extent where detection of homology and automatic annotation transfer is not straightforward. Sophisticated approaches to detect such distant relationships are needed. We propose a new approach to identify deep evolutionary relationships of proteins to overcome shortcomings of the available methods.
Results: We have developed a method to identify remote homologues more effectively from any protein sequence database by using several cascading events with Hidden Markov Models (C-HMM). We have implemented clustering of hits and profile generation of hit clusters to effectively reduce the computational timings of the cascaded sequence searches. Our C-HMM approach could cover 94, 83 and 40% coverage at family, superfamily and fold levels, respectively, when applied on diverse protein folds. We have compared C-HMM with various remote homology detection methods and discuss the trade-offs between coverage and false positives.
Availability and implementation: A standalone package implemented in Java along with a detailed documentation can be downloaded from https://github.com/RSLabNCBS/C-HMM
Supplementary information: Supplementary data are available at *Bioinformatics* online.
Contact: mini@ncbs.res.in

1 Introduction

Rapid increase in genome sequencing initiatives has led to the accumulation of a huge amount of sequence information that needs detailed annotation. Structural and functional characterization of new proteins is a critical task and it is often based on the sequence similarity with proteins of known structure and function. But in some cases, proteins diverge to an extent where relationships between them cannot be easily established using direct sequence search based approaches. Such evolutionary distant proteins are called as remote homologues, and detection of such remote relationships is still a challenging task. Accurate detection of remote homologues can facilitate the assignment of putative functions to uncharacterized proteins improving genomic function annotations. In many cases, functional annotation can be achieved using protein structures, as structural divergence is much slower than sequence divergence. However, protein structural information is limited and the growth of structure databases is also relatively very slow (Levitt, 2007). With the continuous expansion of sequence databases, there is a growing need for a protocol to identify such distant homologues by exploiting mere sequence information.

Many methods based on position-specific scoring matrices (PSSM) and profile Hidden Markov Models (HMM) have been developed to detect distant relationships between protein sequences (Altschul *et al.*, 1997; Eddy, 1998, 2011). For example, position-specific iterated BLAST derives a PSSM from multiple sequences identified by BLAST searches, which are further used to search sequence database to identify newer hits in an iterative manner (Altschul *et al.*, 1997). The drawback of PSI-BLAST is that it can converge before finding all the true positive hits. On the other hand, probabilistic model based profile HMMs (Eddy, 1998, 2011) are more sensitive and efficient in remote homolog detection (Birney, 2001; Karplus *et al.*, 1998). HMMER utility provides one such package named Jackhmmer, which iterates profile-HMMs till convergence, to detect distant relationships of proteins (Eddy, 2011; Finn *et al.*, 2011). Profile HMMs are similar to sequence profiles, but in addition to the amino acid frequencies in the columns of a multiple sequence alignment, they also contain position-specific probabilities for insertions and deletions along the alignment. Profile HMMs are more sensitive than BLAST PSSMs, since position-specific gap penalties penalize chance hits more, as compared with true positives. HMM–HMM comparison, as implemented in HHsearch and HHblits, is also an effective method to identify remote homologues (Remmert *et al.*, 2012; Söding, 2005).

In principle, these methods extract information from intermediate sequences that share sequence properties of multiple remotely related proteins to detect remote homologues (Park *et al.*, 1997; Salamov *et al.*, 1999). When the relation between two remote homologues is not evident, these intermediate sequences serve as stepping-stones for hopping through sequence space. For example, hits from the first ‘generation’ serve as intermediate sequences to detect homologues in the second generation of PSI-BLAST. We have shown earlier that the intermediate sequence approach can be utilized more effectively by cascading through PSI-BLAST over various generations to cover more distant relationships at family, superfamily and fold levels (Kaushik *et al.*, 2013; Sandhya *et al.*, 2005). Yet, a major drawback of using PSI-BLAST in cascade searches is the computational time, which increases considerably with the size and complexity of the databases, lending it almost impossible to apply on large sequence databases.

In this study, we have proposed a new protocol called Cascade-HMM (C-HMM) that can be used to identify remote homologues from protein sequence databases. The primary motivation of developing C-HMM is to permit effective remote homology detection against any available sequence database, irrespective of its size. C-HMM involves cascading of HMMs through various generations to detect true distant homologues even at superfamily and fold levels. We also show that, clustering of protein sequences, followed by profile generation helps in effective detection of remote homologues. In addition, we have compared different remote homology detection methods to provide insights about proper usage of different approaches prior to initiating sequence searches. Our analysis highlighted that BLAST-based methods perform more adequately at the family and superfamily levels, while HMM-based methods are overall more efficient for sequence searches even at the fold level.

2 Methods

2.1 C-HMM protocol

C-HMM protocol is divided into three modules, as described below:

2.1.1 Module 1: C-HMM

The first module of C-HMM involves rigorous sequence searches for many generations (Fig. 1). We implemented cascaded sequence searches by performing five generations of Jackhmmer (Finn *et al.*, 2011). Each generation involves multiple rounds of Jackhmmer searches against a user-defined protein sequence database. In the first generation, the query is scanned against a database with Jackhmmer and the resulting hits are filtered using expectation (E) and inclusion (h) thresholds in order to recognize ‘true’ hits. Since short alignments with good E-values can lead to wrong connections, a length filter is also considered in addition to the E-value criteria. The length filter only permits hits that are >75% of the query length. All the collected homologues that pass the above criteria are defined as ‘true hits’.

All the ‘true hits’ of the first generation are considered as seeds (query sequences) to initiate a second generation of C-HMM. The second generation involves multiple rounds of Jackhmmer, which are initiated with the seed sequences as one query at a time. All the collected hits, obtained through Jackhmmer in the second generation, from each seed sequence are again scanned for the minimum E-value and length filter criteria. Unique hits that pass these criteria and those, which are not identified in the first generation are counted as ‘true hits’ from the second generation. Similarly, third generation of C-HMM is initiated with only filtered unique hits from the previous generations. For each generation, only aligned regions of the query and hits are considered. These generations can be continued until no new hits are obtained.

2.1.2 Module 2: Cluster-HMM

Sequence searches against ever-growing sequence databases, such as the non-redundant (NR) protein sequence database, require a protocol that can process a large number of sequences effectively.

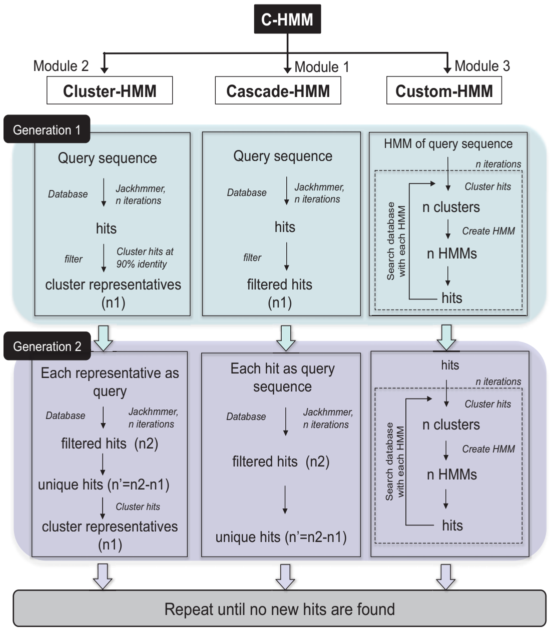

Fig. 1. Flow chart of the C-HMM algorithm. First and second generations are marked separately

C-HMM protocol also offers an option of clustering of the collected homologues (hits), within Cluster-HMM module, while searching against such huge databases. In this module, all the collected true hits obtained after each generation, are clustered using Cd-hit (Li and Godzik, 2006). Only representative sequences of the clusters are employed to initiate subsequent generations. This clustering criterion (Module 2) has been introduced to remove very similar sequences so that sequence searches can be performed in less computational time but with little trade-off on sequence coverage.

2.1.3 Module 3: Custom-HMM

To further enhance the performance of C-HMM, we implemented another module called custom-HMM. Here, similar to Module-2, the filtered hits obtained from the first generation are clustered at a particular sequence identity cut-off using Cd-hit. However, instead of using only representative sequences, entire set of hits within the cluster is considered to create an HMM profile, which is then considered as seed to initiate the next generation. This approach further reduces the search duration without significant loss in sequence information as compared with the other two modules.

2.2. Assessment of C-HMM and dataset creation

For the assessment of the C-HMM protocol, we used PALI+(Phylogeny and Alignment of homologues protein structures) database (Balaji *et al.*, 2001), which is a SCOP-derived (Murzin *et al.*, 1995) protein sequence database of known 3D structures and their sequence homologues. In PALI+, each member in a family has been structurally aligned with every other member in the same family (Gowri *et al.*, 2003). Each sequence of PALI+ is annotated with the respective SCOP fold, thereby permitting the identification of true positive sequences easily. Hence, the total numbers of true connections are known *a priori*. To assess the performance of C-HMM at different levels of sequence similarity, we analyzed all the collected hits at protein family, superfamily and fold levels. We assessed the performance of C-HMM on 16 superfamilies, belonging to different classes of SCOP database, at different E-values (10^{-3}, 10^{-4} and 10^{-6}) for five generations (Supplementary Table S1). These superfamilies belong to 30 most populated metafolds, representing the folds for approximately half of the non-redundant subset of Protein Data Bank (PDB) (Berman *et al.*, 2000; Day *et al.*, 2003). We also performed sequence searches against the Swiss-Prot database (Bairoch and Apweiler, 2000). In addition, we examined the effect of different E-values on the performance of C-HMM during different generations.

2.3 Available parameters and selection

C-HMM provides certain user-defined parameters (E-value, h-value, length filter and clustering threshold) to perform variations in sequence searches. Maximum number of hits to be considered for the next generation runs can be decided before initiating sequence searches. This parameter is useful for highly dispersed protein families.

2.4. Comparison of C-HMM with other methods

We compared the performance of C-HMM with other available remote homology detection methods. We chose four methods: Jackhmmer, PSI-BLAST, HHblits and Cascade PSI-BLAST (Sandhya *et al.*, 2005), to compare C-HMM against PALI database. For this comparison we used a dataset of 18 large superfamilies covering different classes of SCOP database (Supplementary Table S2). Next, we also compared the performance of C-HMM with protein 3D structure comparison method Dali (Holm *et al.*, 2006; Holm and Rosenström, 2010). These searches were performed using larger SCOP folds (Triose phosphate isomerase (TIM), Ferredoxin and Immunoglobulin folds) with many superfamilies against PDB. Additionally, we also assessed the performance of C-HMM, against SwissProt, with a protein domain alignment based method PSI-Search (Li *et al.*, 2012), which can avoid homologues over extensions (HOE) in the sequence alignments (Gonzalez and Pearson, 2010). We employed SUPERFAMILY database (Gough *et al.*, 2001; Gough and Chothia, 2002) as a benchmarking dataset to classify the hits as true positives or false positives, while performing sequence searches against PDB (Dali) and Swiss-Prot (PSI-search) databases.

2.5 Performance measures

We assessed the performance of C-HMM using coverage and precision scores. Coverage was defined as the true associations (true positives) identified using a sequence search method by the total number of true associations present in PALI+ database at family, superfamily and fold levels.

$$\text{Coverage} = \frac{\text{TP} \quad (\text{family/superfamily/fold})}{\text{No. of sequences in database (family/superfamily/fold)}}$$

Any hit that did not belong to the same fold as query sequence was considered as a false positive. Precision scores were also calculated to identify the fraction of hits that are relevant to the sequence search.

$$\text{Precision} = \frac{\text{TP} * 100}{\text{TP} + \text{FP}}$$

where, TP and FP are the number of true positives and false positives identified during sequence searches.

3 Results

3.1 Performance of C-HMM at family, superfamily and fold levels

3.1.1 Detection of homologues within family

Proteins at family level are structurally related and share high sequence similarity. Therefore, relationships at the family level can be identified more easily by pairwise or profile-based methods. Using C-HMM, when we performed sequence searches at E-values of 10^{-3} and 10^{-4}, most of the trusted homologues could be obtained with high coverage scores as shown in Figure 2A. For most of the families, C-HMM could cover 90% of the homologues within family, while for some of them (including Chromo-domain like, Galactose-binding domain and P-loop containing nucleoside triphosphate hydrolases), C-HMM could obtain ~80% of the members. On an average, C-HMM could cover ~94, ~91 and ~82% of homologues within family level at different E-values of 10^{-3}, 10^{-4} and 10^{-6}, respectively. Overall coverage scores of the families were improved during subsequent generations as shown in Supplementary Figure S1.

We noticed that it is very important to propagate sequence searches to many generations to identify remote relationships even at the family level (Fig. 2C). For example, in chromo-domain family, only 10% of homologues were obtained in the first generation, while coverage was increased to 85% when cascaded searches were performed till third generation. For α/β plait and Ferredoxin-like families, continuation of sequence searches till fifth generation also enabled in identifying distant family members that were not captured in the first generation. For five families, however, a single generation was sufficient to identify all the homologues of the respective families.

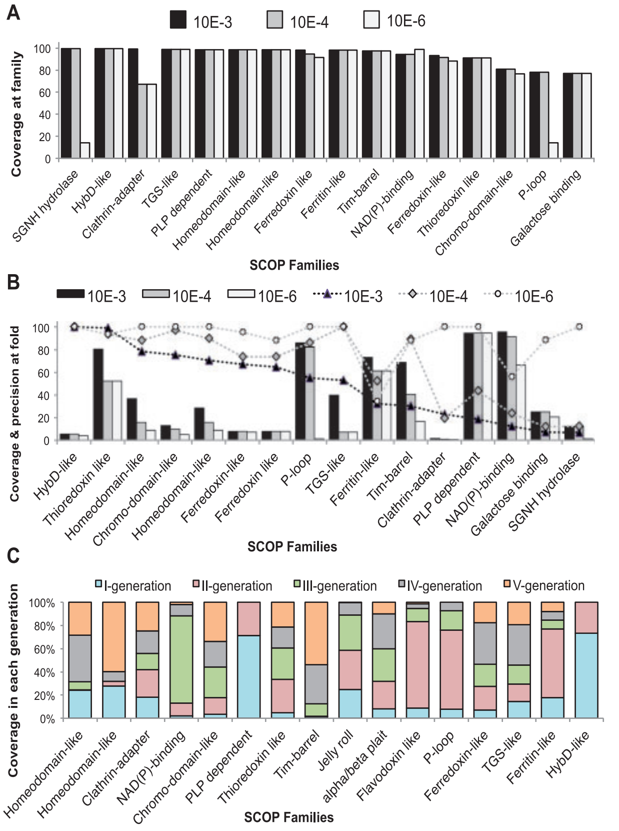

Fig. 2. Coverage of C-HMM at SCOP family and fold levels. **(A)** Coverage of C-HMM at three different E-values at family level. **(B)** Coverage and precision scores at fold level. **(C)** Importance of different generations of C-HMM at fold level. Generation I to V are shown from bottom to top (see also Supplementary Table S2)

3.1.2 Detection of homologues within superfamily

Protein sequences related by structure and function are commonly grouped together to form a superfamily. Distantly related domains share poor sequence identity and could still be related at the superfamily level by means of similar structure and biological function. Hence, the application of cascaded sequence searches becomes more meaningful at that level. Moreover, different protein domain families are differentially represented within a superfamily in terms of sequence dispersion. Our analysis showed that the C-HMM coverage at the superfamily level varied from 90% for a few superfamilies to only 25% (as in Galactose-binding domain superfamily). We also observed variations in coverage score at different E-values, as observed at the family level. For example, in Homeodomain-like superfamily, high coverage (∼58%) was captured at a relaxed E-value of 10^{-3}, while at very stringent E-value of 10^{-6} only 30% coverage was obtained. However, for some of the superfamilies, coverage scores remained the same, even when sequence searches were performed at different E-values (Supplementary Figure S2). Propagation of sequence searches till fifth generation could capture many more homologues at the superfamily level too, except in TGS-like superfamily where 100% coverage could be obtained even within the first generation (Supplementary Figure S3).

3.1.3 Detection of homologues within fold

Detection of fold-level relationships is the most challenging since functions and sequences are far more diverged and share poor similarities even at the structural level. Therefore, obtaining higher coverage is more difficult without increasing false positives due to very distant relationships (Sandhya *et al.*, 2005), and most of the sequence search methods fail to find fold-level relationships in the sequence databases. Interestingly, we observed that for three of the superfamilies, C-HMM could achieve >80% of the coverage at fold-level. This includes superfolds like Rossman fold, PLP-dependent transferase and P-loop containing nucleoside triphosphate hydrolases (Fig. 2B) showing the remarkable power of C-HMM approach. On an average, 42% coverage was captured using C-HMM at the E-value of 10^{-3}, while coverage was reduced to 23% at the E-value of 10^{-6}. Overall, contribution of each generation in accumulating true homologues was improved at fold level as well (Fig. 2C). In 10 out of 16 superfamilies, fifth generation contributed to ∼30% increase in coverage as compared with previous generations.

We also calculated precision scores to observe the performance of C-HMM at different E-values. As expected, we obtained higher precision score of ∼90% at stringent E-value of 10^{-6}, while at the relaxed E-value of 10^{-3}, precision at the fold level was reduced to 50%.

3.2 Comparison of performance of Jackhmmer and C-HMM

C-HMM involves cascaded sequence searches using Jackhmmer. To test if cascaded searches improve the performance of a simple Jackhmmer, we compared C-HMM with simple Jackhmmer with respect to true homologues captured at different E-values. When examined at the family level, C-HMM performed very well in some of the families, while Jackhmmer outperformed C-HMM in other families. For example, for queries within the Rossman fold, C-HMM could identify 70% more homologues at the family level, as compared with simple Jackhmmer searches (Supplementary Figure S4). Most likely, due to our stringent length-filter criteria, some true homologues are also filtered out in C-HMM searches, leading to less coverage for some families. However, detection of homologues at the superfamily level was 19% higher for C-HMM searches as compared with Jackhmmer, which could cover 64.5% of sequence space (Supplementary Figure S5). Similarly, comparison of the two methods at fold level depicted 20% more true homologues with cascade searches (Supplementary Figure S6). These collected homologues are highly distributed all over the sequence space (Supplementary Figure S7).

When we applied different E-values to perform sequence searches, we noticed a clear trade-off between precision and coverage. For example, at E-value of 10^{-3}, coverage score of C-HMM was 20% higher, but precision was dropped by 40% than Jackhmmer. Similarly, at 10^{-4}, coverage was 13% higher, but precision was decreased by 30%, while at 10^{-6} there was 7% increase in coverage, but precision was reduced to 9%.

Comparison of the families identified by Jackhmmer and C-HMM was also carried out by mapping number of families obtained by each method in selected superfolds. For example, in clustering diagram of P-loop family members, Jackhmmer could cover 18 out of 24 families of this superfamily, while 6 other families were not identified with single generation sequence searches of Jackhmmer. However, cascaded searches (E-value = 10^{-3}) could identify four additional families. Still, two of the families named Atu3015-like (c.37.1.13) and Bacterial cell division inhibitor (c.37.1.22) could not be identified using both search methods, suggesting the need for further improvement in such sequence search strategies (Supplementary Figure S8A). The same scheme was also employed to detect how many superfamilies are obtained by these methods (Supplementary Figure S8B).

3.3 C-HMM searches against Swiss-Prot database

Performance of C-HMM was also tested against Swiss-Prot database, which is one of the highly annotated protein sequence databases. During C-HMM searches, some of the classical examples of distant homologues were captured, where Jackhmmer failed. For example, cupredoxin superfamily (SCOP id: b.6.1) consists of seven families, one of which is Plastocyanin/azurin-like family. Azurin and plastocyanin proteins of this family share only 16% sequence identity, but similar protein fold reflecting a common evolutionary origin. When simple Jackhmmer searches were carried out at E-value of 10^{-4} (till convergence), azurin (PDB id: 2CCW) could not identify plastocyanin (PDB id: 1PLC) as a hit, whereas with C-HMM searches this relationship could be established after three generations. This was enabled using two intermediate sequences (bacterial copper containing proteins, auracyanin and rusticyanin, involved in electron transfer reactions), identified during first and second generations, respectively (Fig. 3). This sequence search could also identify 186 other plastocyanin proteins from diverse set of organisms. Another classical example, which we could capture, was of TIM and aldolase folds, which are a pair of analogous folds (Supplementary Figure S9).

3.4 Comparison of available remote homologue detection methods

3.4.1 Comparison with BLAST and HMM based methods

We next compared the performance of C-HMM with other popular methods of remote homology detection. To this end, we applied PSI-BLAST, Jackhmmer, HHblits and Cascade PSI-BLAST on 18 diverse families of SCOP database (Fig. 4). These methods were selected, since they can be applied on any available sequence database. Comparative analysis of coverage at the fold level revealed better performance of C-HMM, as compared with the four other methods. C-HMM could cover more true relationships as compared with direct sequence search methods such as PSI-BLAST and Jackhmmer. Cascade PSI-BLAST outperformed C-HMM in three of the families (Fibronectin-type-III, Concanavalin like lectins and PDZ-domain-like), where only 1.8, 3 and 12% more homologues were captured than C-HMM, respectively. HHblits could capture 12% more homologues as compared with C-HMM in Concanavalin-like lectin family. On an average, however, PSI-BLAST, Cascade PSI-BLAST, Jackhmmer, HHblits and C-HMM could cover 12.5, 37.4, 36.2, 31.6 and 50.5% true distant homologues at fold level, respectively (Fig. 4).

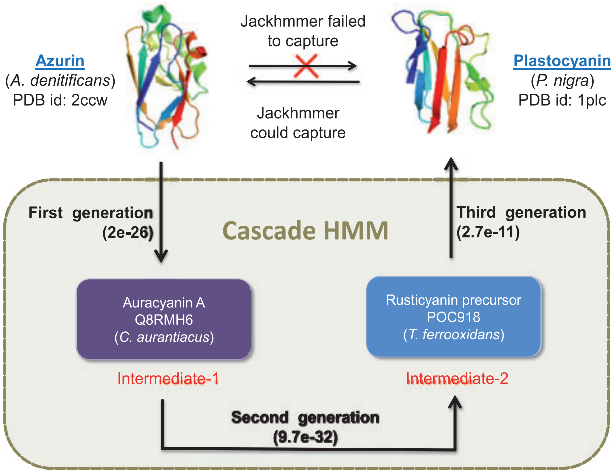

Fig. 3. C-HMM identified the relationship between Azurin and Plastocyanin in three generations using Auracyanin A and Rusticyanin precursors as intermediates

Similarly, C-HMM could cover more homologues as compared with other methods at the superfamily level. HHblits and Cascade PSI-BLAST performed better than C-HMM for four families of β-rich folds, i.e. Fibronectin-type-III, PDZ-domain-like, Concanavalin-like lectins, RNA-binding domains and E-set domains, (Supplementary Figure S10). However, on an average, C-HMM could identify 14% more homologues than Cascade PSI-BLAST, 17% more homologues than Jackhmmer, 17% more homologues than HHblits and 48% more true relationships than PSI-BLAST. For one of the families (periplasmic-binding protein of the doubly wound fold type), 60% increase in coverage was observed with C-HMM searches.

Performance of the selected programs varied at the family level (Supplementary Figure S11). For some of the families, BLAST-based methods performed better than HMM-based approaches. On an average, PSI-BLAST, Cascade PSI-BLAST, Jackhmmer, HHblits and C-HMM could cover 61, 75 67, 70 and 74%, respectively. For Immunoglobulin-like family, all the members within the same family were not captured by Jackhmmer and C-HMM searches, but homologues were covered from other families of the same superfamily.

Precision scores were also calculated at the fold level for all the methods. It was noticed that high precision values (99.89%) were obtained for direct sequence search based methods (e.g. PSI-BLAST), while when sequence searches were cascaded, precision values were dropped (C-HMM = 43%). Highest precision scores were incurred from PSI-BLAST, followed by Jackhmmer (97.72%) and HHblits (94.96%) searches, representing the expected trade-off between coverage and precision score. This showed that BLAST-based approaches performed better for finding direct relationships i.e. finding members of same family, while HMM-based approaches could cover more distant members, including superfamily and fold level relationships more effectively along with family members.

3.4.2 Comparison with protein structure-based methods

We also compared the performance of C-HMM with protein structure comparison-based method, Dali, to identify if C-HMM can relate two superfamilies comparably to Dali. We performed this analysis on SCOP folds with many superfamilies (TIM, Immunoglobulin and Ferredoxin folds). We found that Dali can efficiently make cross-superfamily connections using 3D structures of proteins. Yet, C-HMM is comparable even with mere sequence information. For example, in TIM fold, Dali could identify 16 superfamilies, while C-HMM could connect 12 superfamilies (Fig. 5). Both Jackhmmer and Cascade PSI-BLAST could not identify any cross superfamily connections. In Ferredoxin fold, Dali and C-HMM could cover 25 and 22 superfamilies, respectively (Supplementary Figure S12), while in Immunoglobulin fold, Dali searches could not identify any cross superfamily connections but both Jackhmmer and C-HMM could capture sequence neighbors from Fibronection type III superfamily (Supplementary Figure S13).

3.4.3 Comparison with protein domain alignment based methods

In iterative sequence searches, misleading alignments such as extension of alignments in unrelated proteins domains can lead to contamination of PSSM resulting in higher number of false connections. PSI-Search is a method which masks errors caused by HOE. Therefore, we also compared the performance of C-HMM with

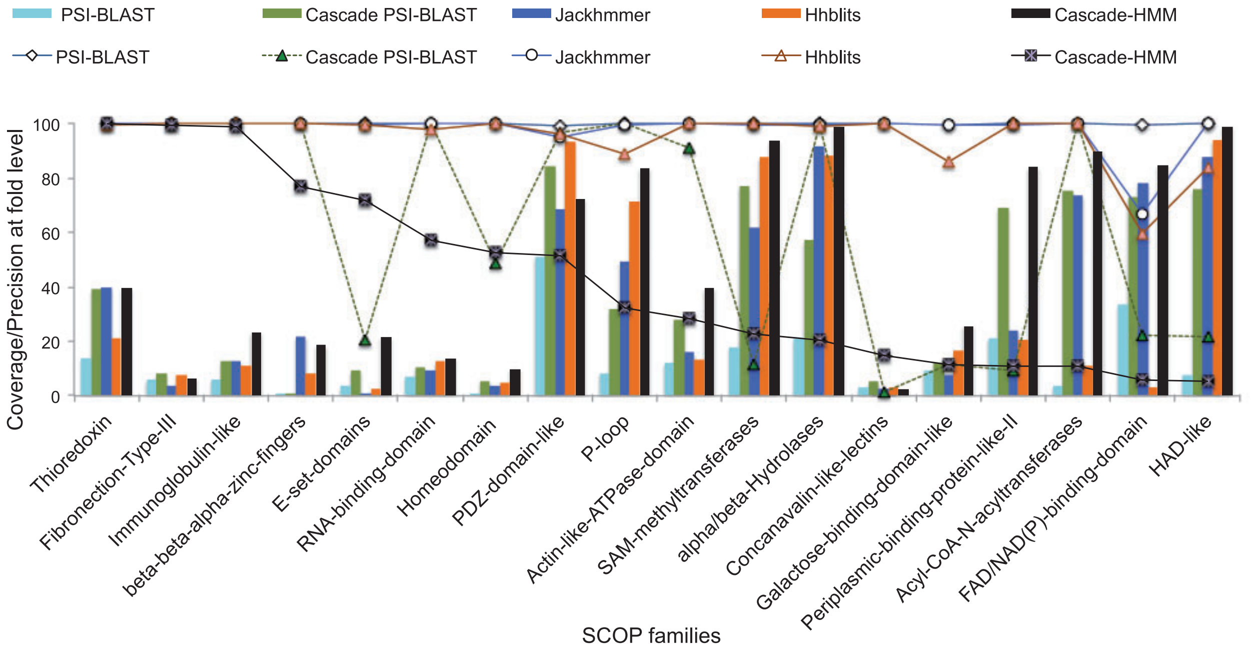

Fig. 4. Comparison of performance of PSI-BLAST, Cascade PSI-BLAST, Jackhmmer, HHblits and C-HMM at fold level. Bars and dashed lines represent coverage and precision scores, respectively

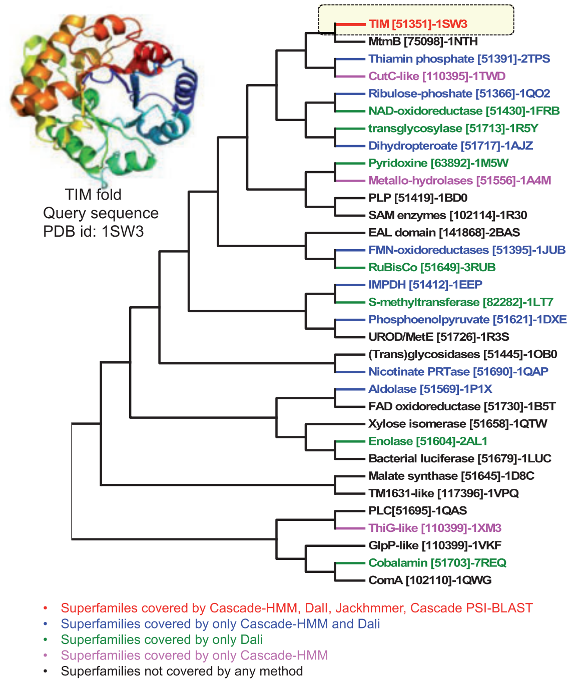
TIM fold
Query sequence
PDB id: 1SW3
TIM [51351]-1SW3
MtmB [75098]-1NTH
Thiamin phosphate [51391]-2TPS
CutC-like [110395]-1TWD
Ribulose-phoshate [51366]-1QO2
NAD-oxidoreductase [51430]-1FRB
transglycosylase [51713]-1R5Y
Dihydropteroate [51717]-1AJZ
Pyridoxine [63892]-1M5W
Metallo-hydrolases [51556]-1A4M
PLP [51419]-1BD0
SAM enzymes [102114]-1R30
EAL domain [141868]-2BAS
FMN-oxidoreductases [51395]-1JUB
RuBisCo [51649]-3RUB
IMPDH [51412]-1EEP
S-methyltransferase [82282]-1LT7
Phosphoenolpyruvate [51621]-1DXE
UROD/MetE [51726]-1R3S
(Trans)glycosidases [51445]-1OB0
Nicotinate PRTase [51690]-1QAP
Aldolase [51569]-1P1X
FAD oxidoreductase [51730]-1B5T
Xylose isomerase [51658]-1QTW
Enolase [51604]-2AL1
Bacterial luciferase [51679]-1LUC
Malate synthase [51645]-1D8C
TM1631-like [117396]-1VPQ
PLC[51695]-1QAS
ThiG-like [110399]-1XM3
GlpP-like [110399]-1VKF
Cobalamin [51703]-7REQ
ComA [102110]-1QWG
• Superfamiles covered by Cascade-HMM, Dall, Jackhmmer, Cascade PSI-BLAST
• Superfamilies covered by only Cascade-HMM and Dali
• Superfamilies covered by only Dali
• Superfamilies covered by only Cascade-HMM
• Superfamilies not covered by any method

Fig. 5. Comparison of cross-superfamily connections (in TIM fold) identified by various methods

PSI-Search on ten superfamilies. We found that PSI-Search is a highly selective method and it did not identify any false connections, while C-HMM could identify higher number of hits but was not as selective as PSI-Search for four superfamilies (Fig. 6). We also found that C-HMM was more advantageous in connecting different superfamilies as compared with PSI-Search. For example, when we performed C-HMM searches we could also identify many cross superfamily connections among Immunoglobulin, Fibronectin type III and PKD domain superfamilies, all of which belongs to Immunoglobulin fold.

3.5 Comparison of performance of C-HMM with Custom-HMM

We then compared the performance of C-HMM and Custom-HMM with respect to time required for each sequence search. When coverage was compared at the fold level, we noticed better or equal performance of Custom-HMM in comparatively lesser time as shown in Supplementary Figure S14. In Ferredoxin and P-loop containing nucleoside triphosphate hydrolases families, we could gather equal coverage scores by both the modules of C-HMM, but the time taken by Custom-HMM was relatively very less. However, for β-grasp and galactose-binding families, we also captured higher coverage scores in comparatively lesser time. This shows that implementing Custom-HMM module can effectively reduce the computational timings for the sequence searches without compromising much on coverage.

4 Conclusions

We have addressed the classical problem of remote homology detection and proposed a generalized protocol that can be used against any protein sequence database to identify distant evolutionary relationships. Our approach also highlights the importance of intermediate sequences to understand difficult protein relationships. C-HMM approach, described in this article, has been demonstrated to be powerful in identifying connections across protein domains that share the same fold but are structurally and functionally very diverse. It is possible to search the large non-redundant sequence databases using this approach to achieve maximum coverage. Implementation of C-HMM searches on generalized databases

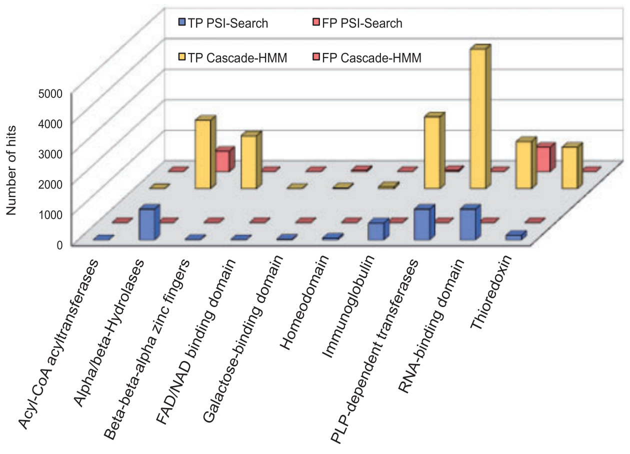

Fig 6. Comparison of performance of C-HMM and PSI-search

showed the universal applicability of this protocol. Hence, it can be used on growing genomic databases to assign probable functions to putative and uncharacterized sequences in the sequenced genomes. Development of such algorithms is also important for protein fold assignment and can further facilitate genome annotation process. However, a clear trade-off between coverage and precision score was observed, suggesting that prior selection of sequence search algorithm could lead to efficient remote homology detection. For instance, PSI-BLAST or HHblits can be used to recognize close homologues, since these methods can find direct relationships more easily and accurately as compared with any other available methods. On the other hand, if we are interested in finding relationships within superfamily/fold level, cascaded sequence searches are more promising.

Although C-HMM approach can effectively scan large sequence space, there is need for further improvement, since a complete coverage was not captured for all families. Inclusion of other protein parameters, such as secondary structures and sequence motifs, can also be implemented in available algorithms to scan the sequence databases. By reducing the number of hits to be used as queries for the next generation, a significant reduction in the search time against large databases (e.g. NCBI nr database) could be achieved by our custom-HMM module. Nevertheless, time required for the Cascade searches are higher than other methods (Supplementary Table S3). Inclusion of structural parameters, along with mapping domain boundaries and masking of HOE, will also help in the reduction of cross-connections among other members of same classes of SCOP database thereby reducing false positives. Another source of limitation is lack of intermediate sequences available in protein sequence databases, which could lead to lesser coverage. Alternate approaches, such as generation of artificial sequences, appear to be highly attractive to fill sequence space, especially when combined with cascaded sequence searches (Mudgal *et al.*, 2014). We believe that the C-HMM protocol makes significant advance in the area of remote homology detection, and will be very useful approach to assign putative functions and folds to protein sequences. We recommend its usage for genome annotation pipelines due to its speed, reliability, efficiency and database independency.

Acknowledgements

The authors thank Prof N. Srinivasan of Indian Institute of Science for useful discussions and NCBS for infrastructural facilities.

Funding

S.K. was supported by Department of Biotechnology scholarship. E.M. was supported by Bioinformatics National Certificate (BINC) fellowship. A.N. was supported by Centre for Excellence Grant by Department of Biotechnology (BT/01/COE/09/01).

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. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Balaji,S. *et al.* (2001) PALI–a database of Phylogeny and ALIgnment of homologous protein structures. *Nucleic Acids Res.*, **29**, 61–65.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Birney,E. (2001) Hidden Markov models in biological sequence analysis. *IBM J. Res. Develop.*, **45**, 449–454.

Day,R. *et al.* (2003) A consensus view of fold space: Combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Sci.*, **12**, 2150–2160.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.

Finn,R.D. *et al.* (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.

Gonzalez,M.W. and Pearson,W.R. (2010) Homologous over-extension: a challenge for iterative similarity searches. *Nucleic Acids Res.*, **38**, 2177–2189.

Gough,J. *et al.* (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.

Gough,J. and Chothia,C. (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.*, **30**, 268–272.

Gowri,V.S. *et al.* (2003) Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database. *Nucleic Acid Res.*, **31**, 486–488.

Holm,L. *et al.* (2006) Using Dali for structural comparison of proteins. *Curr Protoc Bioinformatics*, **Chapter 5**, Unit 5.5.

Holm,L. and Rosenström,P. (2010) Dali server: conservation mapping in 3D. *Nucleic Acids Res.*, **38**, W545–W549.

Karplus,K. *et al.* (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.

Kaushik,S. *et al.* (2013) Improved detection of remote homologues using cascade PSI-BLAST: influence of neighbouring protein families on sequence coverage. *PLoS One*, **8**, e56449.

Levitt,M. (2007) Growth of novel protein structural data. *Proc. Natl. Acad. Sci. USA*, **104**, 3183–3188.

Li,W. *et al.* (2012) PSI-Search: iterative HOE-reduced profile SSEARCH searching. *Bioinformatics*, **28**, 1650–1651.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Mudgal,R. *et al.* (2014) Filling-in void and sparse regions in protein sequence space by protein-like artificial sequences enables remarkable enhancement in remote homology detection capability. *J. Mol. Biol.*, **426**, 962–979.

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.

Park,J. *et al.* (1997) Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.*, **273**, 349–354.

Remmert,M. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Salamov,A.A. *et al.* (1999) Combining sensitive database searches with multiple intermediates to detect distant homologues. *Protein Eng.*, **12**, 95–100.

Sandhya,S. *et al.* (2005) Assessment of a Rigorous Transitive Profile Based Search Method to Detect Remotely Similar Proteins. *J. Biomol. Struct Dyn.*, **23**, 283–298.

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