Journal of Advanced Bioinformatics Applications and Research ISSN 0976-2604
Vol 3, Issue 1, 2012, pp 250-261 http://www.bipublication.com

A HIDDEN MARKOV MODEL FOR DOMAINS SPLIT BY INSERT DOMAINS

Sekhar Talluri
Department of Biotechnology, GIT, GITAM University
Corresponding author: Email: stalluri.home@gmail.com, Tel: +9989914868 ;

ABSTRACT:

A Hidden Markov Model (HMM) for discovering and characterizing certain types of split domains is described here. The Hidden Markov model for split domains (HMMsd) will be useful for annotation of novel classes of proteins containing domain insertions and also to search for homologs of proteins containing split domains. The Hidden Markov Model for split domains (HMMsd), described here, models domains that are split due to the presence of insertion of another domain. HMMsd is fully probabilistic and it is derived from the Plan-7 model architecture used in the Pfam HMM profiles for individual domains. The curated parameters characterizing the profile HMMs of constituent domains, obtained from PFAM, are used to calculate the scores for the model described here. HMMsd identifies proteins known to possess domain insertions, such as Syntrophins and Phospholipase Cɣ, with high specificity. The only input required for using HMMsd is the pair of profile HMMs (which can be obtained from Pfam) and the sequence/s of the protein of interest.

Keywords: *Hidden Markov Model; Annotation; Split domains*

[I] INTRODUCTION

The human genome encodes many proteins whose function is still unknown. Many other genomes are known with lower levels of functional annotation. Discovery and demonstration of sequence and structural homology has played an important role in the functional annotation of many protein sequences [1]. Many proteins contain domains which are composed of sequence fragments that are separated by long intervening stretches of sequence that are not part of the structurally compact domain. In some cases, the intervening sequences are long enough to contain a complete domain of a different type. The probabilistic model described here would enable a systematic search for homologs of proteins known to contain split domains and for discovery of novel split domains that have been missed by other methods of annotation.

The domains observed in most proteins are formed from a single contiguous polypeptide chain. However, in certain cases a domain is formed from two or more fragments of a single polypeptide chain; such domains are called discontinuous or split domains [2,3]. Insertions are one particular form of discontinuous domains; the polypeptide sequence of one domain (insert domain) intervenes between the fragments of the single polypeptide chain that define a discontinuous domain or split domain (the parent domain). Domains insertions are the most common form of discontinuous domains and 9% of domain combinations in non-redundant PDB are insertions [3]. For example, the PH domain of Phospholipase C□ contains an insertion consisting of two SH2 domains and one SH3 domain which replace the first helix of the PH domain [4]. Human Syntrophin contains a split PH [5] domain with a PDZ [6] domain insertion [7,8]. The SH3 domain of CaVb is split with an insertion [9] of the HOOK domain before the 5th and last beta strand of the SH3 domain. It has been suggested that such split domains may mediate intermoleculer interactions by interaction of complimentary split domains of different proteins [10,11].

Systematic methods for finding homologous split domains would be useful in studies involving such proteins.

PSI-BLAST [12] is a rapid and sensitive method for finding homologs of proteins. If this method is used for finding homologs of proteins that contain split domains (such as syntrophin), successive iterations will find a large number of proteins that have very good matches to one of the fragments of the protein and the resulting profile would ultimately be either averaged or biased towards one of the two constituent domains, leading to loss of information regarding the synteny of the fragments. The HMM for split domains, described here, can provide a probabilistic description of the two domains and encodes information regarding the nesting of one domain within the other; therefore, this model can be the basis for a systematic method to search for homologs of proteins containing split domains. The probabilistic basis for this model and the use of curated parameters from Pfam HMM profiles imply a high selectivity, sensitivity and reliability for sequence analysis.

Split domains may be discovered by using heuristic methods with a low threshold to discover the fragments, followed by a combinatorial probabilistic evaluation of the results of the search; or by using structure based multiple sequence alignments [13]; or by a rule based approach to treat overlaps [14]. It has also been suggested that split domains may be discovered by fusing (virtually) the 'candidate split domain' to the consensus sequence for the complimentary fragment of the split domain and using these synthetic chimeric sequences to search the Conserved Domain Database (CDD) [12]. However, these methods are difficult to automate and require substantial expertise. The method described here requires no user intervention. The Hidden Markov Model [15,16] (HMM) for split domains (HMMsd), described here, is highly specific for detection of domains containing other inserted domains. The importance of contextual information and the ability of HMMs to exploit this information has been demonstrated in the annotation of *Plasmodium falciparum* [17] using data regarding frequencies of co-occurrence. Contextual information is exploited in a different manner by HMMsd.

[II] MATERIALS AND METHODS

The protein sequences for the Human and Mycobacterium tuberculosis genomes were downloaded from UniProt/SwissProt databases [18,19]. The sequences for other proteins were downloaded either from EBI, NCBI [20] or the RCSB [21] protein databank. The accession numbers for the proteins sequences used for this analysis are: Q13424 (Syntrophin-A), Q13884 (Syntrophin-B1), Q13425 (Syntrophin-B2), Q9NSN8, NP_061840.1 (Syntrophin-G1) , Q9NY99, NP_061841.1 (Syntrophin-G2), P19174 (phospholipase-Cg1), Q8K326 (Agrin). For proteomic analysis the Uniprot-KB/Swissprot human proteome dataset was used and the IPI datasets were used for the mouse (*Mus*) and zebra fish (*Danio rerio*). HMMSEARCH and HMMBUILD programs from the HMMER 2.3 package [22,23] were used for sequence analysis. The HMM profiles for PH (PF00169.21), SH2 (PF00017.16), SH3 (PF00018.20), PDZ (PF00595.16), Laminin G (PF00054), Hook (PF05622.4) and EGF (PF00008.19) domains were downloaded from PFAM [24,25]. Pfam has recently started the use of HMMER 3.0 profiles as default, but the HMMER2 HMM profiles used by HMMsd are available at ftp.sanger.ac.uk/pub/databases/Pfam/releases/Pfam23.0.

The Viterbi algorithm for the Plan-7 HMM (included in the HMMER package version 2.3.2 [23] was modified as per the requirements of the HMMsd model. The time required for analyzing 630 proteins belonging to the globin family for split globin domains using the current implementation of the program was 3-5 minutes using a PC with Intel Celeron 600MHz processor and 256MB of memory. In addition, the entire human proteome could be scanned for a specified pair of domains in less than 30 minutes by using the filter option of HMMsd, using a PC equipped with an AMD Sempron processor. There is substantial scope for further speed up of the current implementation without changing the Model architecture. The program will be freely available to everyone after publication of the manuscript.

2.2 Algorithm and implementation

Hidden Markov Model for split domains (HMMsd): The model architecture for split domains was derived from the Plan-7 architecture used for Pfam profile HMMs. A graphical representation of the proposed Hidden Markov Model for split domains is shown in Figure 1. All transitions start from the S state. The only possible transition from S state is to the N state (representing residues at the N-terminus). Two types of transitions are possible from the N state. A transition can occur from the N state to the B state, without emission of any symbols; or a transition can occur to the N state itself with emission of a symbol. Transitions can occur from the B state, to any of the M states (M) of this model or to the D1 state. The number of M states depends upon the profile being modeled and is usually close to the number of columns with conserved residues in the multiple sequence alignment. The kth M state is referred to as M_k. Associated with each M_k state is a D_k state (standing for deletion). An I_k state is associated with each M_k state except the last one. Transitions can occur from an M_k state to an M_{k+1} state, or to an I_k state or to a D_{k+1} state or to the E state. Transitions can occur from a D_k state to state D_{k+1} or to M_{k+1}. Similarly, transitions can occur from state I_k to I_k, or to M_{k+1}, or to B* state. B* represents the begin state for the inserted domain. Transitions occur from the E state to the C state or to the J state. Transitions can occur from the J state either to the B state or the J state. Transitions can occur from the C state either to the C state or to the T (terminal) state. The architecutre of the HMM module representing the inserted domain is similar to that of the parent domain except that the N, C and J states are not present and that transitions from I state are limited to transitions to the same I state or to the subsequent M state. The transitions in gray (Figure 1) are effective if the alignment is local with respect to the Hidden Markov model; the corresponding transition probabilities are set to zero for alignments that are global with respect to the HMM. In the present study only alignments global with respect to the HMM were investigated.

All states except the S, B, B*, D, D*, E, E* and T states emit symbols. The emission probabilities of all states and all state transition probabilities (except the transition probabilities of the I->B* transition) were obtained from the PFAM profile HMMs. Details regarding the derivation of these probabilities and the construction of PFAM profile HMMs are described elsewhere [22,23,25]. The I->B* transition probability can be set to 1/L, where L is the length of the sequence being aligned to the HMM. We expect that a transition will occur from one of the states of the parent domain to the B* state of the insert domain; however, a priori, we do not have any knowledge regarding the location or length of the insert in the sequence of length L, therefore, the I->B* transition probability can be chosen to be 1/L, so that the sum of these probabilities of insertion at all residues adds up to one. The I->B* transition probability may be chosen to have a different value, e.g., 1/(L-N*) or 1/(N-1); The former corresponds to the assumption that an insert of expected length N* occurs somewhere in the sequence of the split domain of length L-N* and the later value corresponds to the assumption that the transition occurs from any one of the (N-

1) insert states of the split domain. However, proteins such as Phospholipase C□ may contain more than one copy of the insert domain. Therefore, a moderate value (1/ (L-N*)) for the I->B* transition probability was chosen for all the experiments described in this manuscript. This is the default value in this implementation and can be changed by using a command line argument if necessary. This is the only adjustable parameter required by this model and the choice of this parameter does not affect the final score (unless there is a change in the optimal path), if the Viterbi scores are corrected (See section describing the scoring method). The magnitude of the I->B* transition probability is quite low (usually less than 0.01), therefore, it was not necessary to renormalize the state transition probabilities obtained from Pfam profile HMMs (that are calculated for the Plan-7 model architecture).

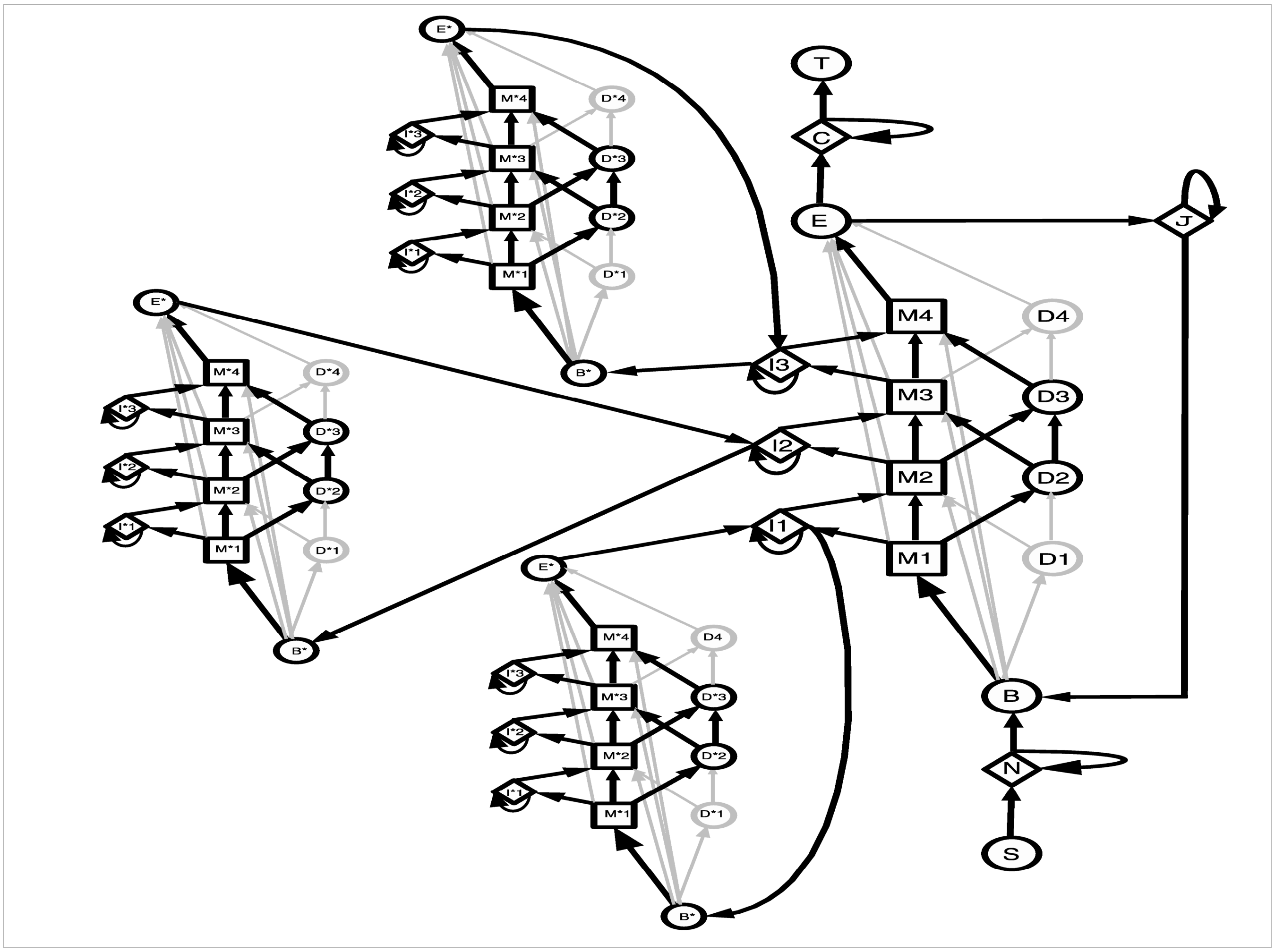

Figure 1. Architecture of the Hidden Markov model for split domains. The states marked with an asterisk are for alignment of the inserted domain (if there is one) in the sequence. The remaining states model the fragments of the parent domain sequence and N- and C-terminal segments. The number of M, I, D, M*, I* and D* states depends on the number of columns containing conserved residues in the multiple sequence alignment of representative parent and inserted domains. The model shown above represents the case where there are four match states (M, M*) for each domain.

Viterbi algorithm for Hidden Markov Model of split domains

In HMMsd, transitions to the states of the inserted HMM can only occur from the insert states of the parent HMM (see Figure 1). Therefore, the Viterbi score for HMMsd can be calculated using a modification of the Viterbi algorithm used for calculation of the Viterbi score for the Plan-7 HMM in the HMMER package [22,23]. In HMMsd, the Viterbi scores, FM[][] are calculated in the same manner as in the Viterbi algorithm used to compute scores for the Plan-7 architecture profile HMMs in HMMER [22,23]. In HMMsd, the scores for the insert states are calculated as follows:

$$
\begin{array}{l}
FI[i][k] = \max \{ \\
\quad FM[i-1][k] + P(M_k\text{->}I_k) + e_{Ik}(x[i]) , \\
\quad FI[i-1][k] + P(I_k\text{->}I_k) + e_{Ik}(x[i]) , \\
\max_{j=2 \ldots i-1} \{ FI[i-j][k] + P(I_k\text{->}B^*) + \\
\qquad FXB^*_{i-j,j} \} \\
\qquad \}
\end{array}
\quad (1)
$$

$FXB^*_{p,q}$ is the optimum score for alignment of the subsequence x[p]...x[q] to the profile HMM of the inserted domain.

Computation of $FXB^*_{p,q}$: Only the parameters of the profile of the inserted domain are required for computation of $FXB^*_{p,q}$. The state transition and emission probabilities of the profile of the inserted domain are marked with an asterisk.

Initialization:

$$
\begin{array}{lll}
FB_p^*[0] = 0 & & \\
\multicolumn{3}{l}{\text{(probability p=1, hence log p = 0)}} \\
FM_p^*[0][k] & = -INF & \text{for } k=0..N^* \\
FI_p^*[0][k] & = -INF & \text{for } k=0..N^* \\
FD_p^*[0][k] & = -INF & \text{for } k=0..N^* \\
FM_p^*[i][0] & = -INF & \text{for } i=0..L-p \\
FI_p^*[i][0] & = -INF & \text{for } i=0..L-p \\
FD_p^*[i][0] & = -INF & \text{for } i=0..L-p
\end{array}
\quad (2)
$$

Recursion:

$$
\begin{array}{l}
FM_p^*[i][k] = \max \{ \\
\quad FM_p^*[i-1][k-1] + \\
\qquad P^*(M_{k-1}\text{->}M_k) + e^*_{Mk}(x[i+p]), \\
\quad FI_p^*[i-1][k-1] + \\
\qquad P^*(I_{k-1}\text{->}M_k) + e^*_{Mk}(x[i+p]), \\
\quad FD_p^*[i-1][k-1] + \\
\qquad P^*(D_{k-1}\text{->}M_k) + e^*_{Mk}(x[i+p]), \\
\quad FB_p^*[i-1] + \\
\qquad P^*(B\text{->}M_k) + e^*_{Mk^*}(x[i+p]) \}
\end{array}
\quad (3)
$$

$$
\begin{array}{l}
FD_p^*[i][k] = \max \{ \\
\quad FM_p^*[i][k-1] + P^*(M_{k-1}\text{->}D_k), \\
\quad FD_p^*[i][k-1] + P^*(D_{k-1}\text{->}D_k) \}
\end{array}
\quad (4)
$$

$$
\begin{array}{l}
FI_p^*[i][k] = \max \{ \\
\quad FM_p^*[i-1][k] + P^*(M_k\text{->}I_k)+e^*_{Ik}(x[i+p]), \\
\quad FI_p^*[i-1][k] + P^*(I_k\text{->}I_k) + e^*_{Ik}(x[i+p]) \\
\qquad \}
\end{array}
\quad (5)
$$

Termination:

$$
FXB^*_{p,i} = FM_p^*[i][N^*] \quad \text{for } i=1..L-p \quad (6)
$$

Notation:

HMMp = profile HMM of domain that may be split

HMMi = profile HMM of domain that may be present inserted into the split domain

N = number of match states in HMMp

N* : number of match states in HMMi

L = length of protein sequence

P(A->B) = log of probability of transition from state A to state B.

M_k, I_k, D_k = Match, Insert and Deletion states in HMMp

M^*_k, I^*_k, D^*_k = Match, Insertion and Deletion states in HMMi

S*, N*, B* = Start, N terminal and Begin states in HMMi

C*, J*, T* = C terminal, loop back and Terminal states in HMMi

x[i] = residue at position i in sequence

$e_A(C)$ = log of ratio of probability of emission of residue C from state A in HMMp compared to probability of emission from a random model.

The $FXB^*_{p,q}$ values are computed and stored. Therefore, the time requirement of the overall algorithm is only $O((N+N^*)L^2)$. However, there is an increase in the memory requirement of $O(L^2)$. The total memory requirement for this algorithm is $O((N+N^*)L + L^2)$.

HMMsd scores

The score obtained by application of the Viterbi algorithm for finding the optimal alignment of the sequence to the HMM will be referred to as S0.

$S0 = \log_2 (P(seq \mid HMMsd) / P(seq \mid R))$ (7)

Where, P(seq | HMMsd) is the probability of the sequence based on the Hidden Markov model of the split domain architecture (illustrated in Figure 1); P(seq | R) is the probability of the sequence based on the (null) Random model. Since the S0 scores are bit scores, P-values and E-values can be calculated by assuming that the scores follow an extreme value distribution [22,23]. However, the actual form of the distribution is not known and therefore P-values and E-values are not used as a basis for identification of split domains. Positive values of S0 indicate that HMMsd is a better model than the random model for describing the given sequence. However, it is not necessary that the optimal path should pass through the match states of the inserted domains because these transitions are optional. Therefore, high S0 scores may also be obtained when there are no insertion domains, if the sequence matches the unstarred match states of the HMMsd. The score S1, defined below, can be used to distinguish between these two possibilies.

$S1 = \log_2(P(seq \mid HMMsd) / P(seq \mid HMMp))$ (8)

P(seq | HMMp) is calculated using the same model architecture as in Figure 1, with all I->B* probabilities set to zero (as in the Pfam Plan-7 HMM model architecture [22,23]. P(seq | HMMp) is the probability of the sequence from the Hidden Markov model where all insertions are modeled as loops for the given sequence. The transition and emission probabilities for evaluating this probability are obtained from the Pfam HMM profile of the parent domain. S1 is the log odds score that the specified (parent) domain contains an inserted domain compared to the probability of a Hidden Markov model where all insertions are modeled as loops (for the given sequence). Similarly,

$S2 = \log_2(P(seq \mid HMMsd) / P(seq \mid HMMi))$ (9)

P(seq | HMMi)/ P(seq | R) is calculated by using the Viterbi algorithm for the Plan-7 model [22,23]. This probability is evaluated by using the parameters of the Pfam profile HMM of the domain specified as the inserted domain. If both S1 and S2 are positive, it indicates that the split domain HMM model is a better model of the sequence than the Plan-7 HMM model for either domain. Finally, the score S3 is used to ascertain the independence of the probabilities of the two types of domains.

$S3 = \log_2(P(seq \mid HMMsd) / (P(seq \mid HMMp) * P(seq \mid HMMi))$ (10)

This calculation involves comparison of scores from profile HMMs of different lengths, therefore, a correction for edge length effects is applied. However, this is optional and the correction can be ignored by using the –nolencor option for HMMsd.

If the optimal alignment requires a transition into the states of the insert domain, the final Viterbi score calculated by using the algorithm described above includes a term for the I->B* transition with a default probability of 1/L. However, if the location of the insert domain is known (a posteriori) this probability should be 1. Therefore, a $\log_2(L)$ correction can be applied, optionally, to the S2 and S3 scores to obtain the log-odds of a sequence for a model that describes a domain split by one insert domain. Use of the correction implies that the final score would be independent of the choice of the transition probability of the I->B* transition, provided that the traceback path is not altered. This correction can be controlled by the –S3cor option of HMMsd. This correction was not used in the the results described in this manuscript. Based on these definitions, S0, S1, S2 and S3 are expected to be positive for proteins containing domain insertions, for the correct choice of split (parent) and insert domain profiles.

In addition to the global scores described above HMMsd can also calculate fragment specific scores. These are of particular use in proteins that contain a large number of domains. HMMsd (used with –slow option) identifies the fragment containing the insert using traceback and recalculates the scores for the fragment containing the split domain. This computation increases the computational time but increases the specificity of the method. Fragment scores are based on global alignments, although the analysis is restricted to a certain fragment which has been identified to match the insert domain. This helps to focus attention on the region likely to contain the split domain.

[III] RESULTS

The Hidden Markov model for split domains (HMMsd) was tested initially by using syntrophins. Syntrophins were selected for testing HMMsd, as it has been demonstrated that they contain insertions of PDZ domains in PH domains. Table 1 shows the results of the sequence analysis of five human syntrophins. The parameters of the Pfam PH domain were used for the parent domain and the parameters of the Pfam PDZ domain were used for the inserted domain. It is evident that both S0 and S1 scores are greater than 50 for all five syntrophins, indicating that the split domain Hidden Markov model, described here, is highly probable for these proteins. In addition, the S2 and S3 scores are positive for Syntrophin-A, Syntrophin-B1 and Syntrophin-B2, indicating unambiguous identification of the PH domain containing an inserted PDZ domain. However, S2 scores are negative for the Syntrophin-Gs (although S0, S1 and S3 are positive). This is not particularly surprising, since sequence based domain assignment programs have difficulty in identifying the first PH domain of Syntrophin-Gs which is split and contains the inserted PDZ domain [26]. For example, the HMMSEARCH program, when used in the local alignment or fragment search mode with default parameters, is unable to locate the fragments of the split PH domain for Syntrophin-G2, although it finds the second fragment of the split PH domain for all other Syntrophins.

Table 1. Sequence analysis of Human Syntrophins using HMMsd with PH domain as parent and PDZ domain as insert domain.

Sequence name	S0	S1	S2	S3	I
Syntrophin-A1	97.0	92.2	21.5	24.2	1
Syntrophin-B1	94.9	102.5	15.3	31.5	1
Syntrophin-B2	85.4	100.0	0.3	23.4	1
Syntrophin-G1	62.9	75.1	-10.0	8.7	1
Syntrophin-G2	58.1	83.5	-17.7	15.2	1

I greater than 0 indicates that the optimal path passes through the states of the insert domain.

The ability of HMMsd to distinguish incorrect assignments of domains was tested as follows. SH2 domain was incorrectly specified as the parent domain and the PDZ domain was specified as the inserted domain; although Syntrophins are known to contain a domain insert of a PDZ domain in a PH domain. The results of this study are shown in Table 2. From these results it is evident that S2 scores are strongly negative for the wrong choice of the parent domain for all Syntrophins. In addition, S3 scores are also negative for all Syntrophins (except Syntrophin-G1). These results demonstrate the ability of HMMsd to identify the correct type of parent.

Table 2. Sequence analysis of Human Syntrophins using HMMsd with incorrect specification of parent domain as SH2.

Sequence name	S0	S1	**S2**	S3	I
Syntrophin-A1	**-1.0**	58.4	**-77.5**	**-9.0**	1
Syntrophin-B1	0.7	54.4	**-78.8**	**-16.0**	1
Syntrophin-B2	14.7	71.0	**-70.5**	**-5.0**	1
Syntrophin-G1	9.7	68.6	**-65.2**	2.8	1
Syntrophin-G2	**-3.1**	58.4	**-79.9**	**-9.3**	1

Furthermore, the ability of HMMsd to distinguish incorrect assignment of insert domains was tested as follows. PH domain was specified as the parent domain and the SH2 domain was incorrectly specified as the inserted domain. The results of this experiment are shown in Table 3. From these results it is evident that S1 scores are zero for the wrong choice of the parent domain for all Syntrophins.

Table 3. Sequence analysis of Human Syntrophins using HMMsd with incorrect specification of insert domain as SH2.

Sequence name	S0	**S1**	S2	S3	I
Syntrophin-A1	37.9	**0.0**	97.3	68.4	0
Syntrophin-B1	61.9	**0.0**	115.6	62.9	0
Syntrophin-B2	31.8	**0.0**	88.2	65.5	0
Syntrophin-G1	21.4	**0.0**	80.3	68.0	0
Syntrophin-G2	19.5	**0.0**	81.0	70.6	0

Finally, the parent domain was incorrectly specified as a protein Hook domain and the insert domain was incorrectly specified as SH2 domain, for the Syntrophins, to investigate the effect of incorrect specification of both parent and insert domain. The results are shown in Table 4. It is evident that both S0 and S2 are strongly negative and that S1 is zero for the wrong assignment of domains.

Table 4. Sequence analysis of Human Syntrophins using HMMsd with incorrect specification of parent domain and insert domains.

Sequence name	S0	**S1**	S2	S3	I
Syntrophin-A1	**-636.6**	**0.0**	-568.4	59.4	0
Syntrophin-B1	**-632.2**	**0.0**	-569.6	53.8	0
Syntrophin-B2	**-658.8**	**0.0**	-593.6	56.4	0
Syntrophin-G1	**-666.1**	**0.0**	-598.4	58.9	0
Syntrophin-G2	**-675.4**	**0.0**	-605.1	61.5	0

I=0 indicates that the optimal path does not pass through the states of the insert domain.

The Hidden Markov model for split domains (HMMsd) was also tested for the ability to locate the domain boundaries. The results, indicating the ability to correctly locate the domain boundaries, are shown in Table 5.

Table 5. HMMsd results for domain boundaries.

Sequence name	HMM profile for parent	HMM profile for Insert domain	Location of insert domain (hmmsearch)	HMMsd Location of insert
Syntrophin-A1	PH-domain	PDZ	87 - 167	87-167
Phospholipase-C□□	PH-domain	SH2	550-639, 668-741	550-639 *
Phospholipase-C□□	PH-domain	SH3	794-849	794-849
Agrin (Mouse)	Laminin_G	EGF	123 -157	123 -157

*HMMsd indicates that two copies of the insert domain were located; however, the location of only one copy is printed out.

Split domains may also be located by using a local alignment method to locate the individual fragments that match the profile of interest; subsequently a score can be calculated for the complete set of ordered fragments [27]. The sensitivity of a such a method, based on the use of a single domain, was compared to that of HMMsd which uses a pair of domains. Sequence homologs of Syntrophin-A, containing 1-200 random mutations per sequence, were generated from the fragment (1-300) of human Syntrophin-A, which is known to contain a split PH domain. The results for searching this data set by using HMMsearch (of the HMMER package) in the fragment mode and those for the HMMsd program are shown in

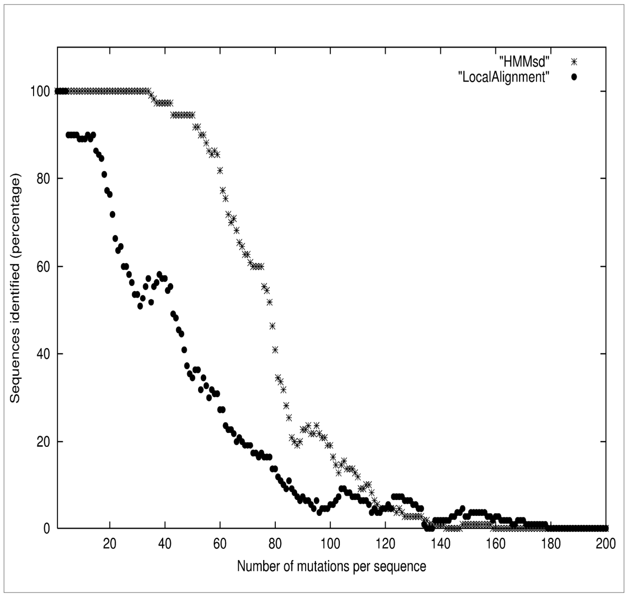

Figure 2.

Figure 2. Comparison of HMMsd and local alignment method for identification of a split parent domain using 3000 homologs of the 1-300 fragment of Syntrophin A. Each sequence was mutated *n* number of times, with *n* varying from 1-300. A sequence was considered to be identified correctly if the calculated E-value was less than 0.05.

An E-value threshold of 0.05 was used as the criterion for identification in both cases. The fraction of sequences that can be correctly identified depends on the mean number of mutations in both cases. It is obvious that both methods are able to identify sequences having a small number of mutations. As the number of mutations per sequence increase, HMMsd is able to identify a greater fraction of the sequences compared to the local alignment method. However, for very large number of mutations, the local alignment method shows a slight increase in sensitivity.

Proteomics

The annotated proteomes of human, mouse and zebra fish were used to identify a set of proteins containing domains split by domain insertions. The ability of HMMsd to identify the members of this selected set from a proteomic scan was evaluated. The analysis of the HMMsd scores for this experiment is shown in Figures 3 and 4.

The E-values obtained from the Viterbi scores (S0) are not adequate for identification of split domains. This is because, high (compared to random) scores of S0 can be obtained either due to the presence of a protein containing a split domain or due to other proteins that contain single or multiple instances of non-split parent or insert domains. In order to distinguish these cases from genuine instances of split domains, identification is based upon the combined use of the scores S0, S1, S2 and S3. Figure 3 shows the ROC plot for the S2 score; the remaining scores S0, S1 and S3 were required to be positive (i.e. a threshold of zero).

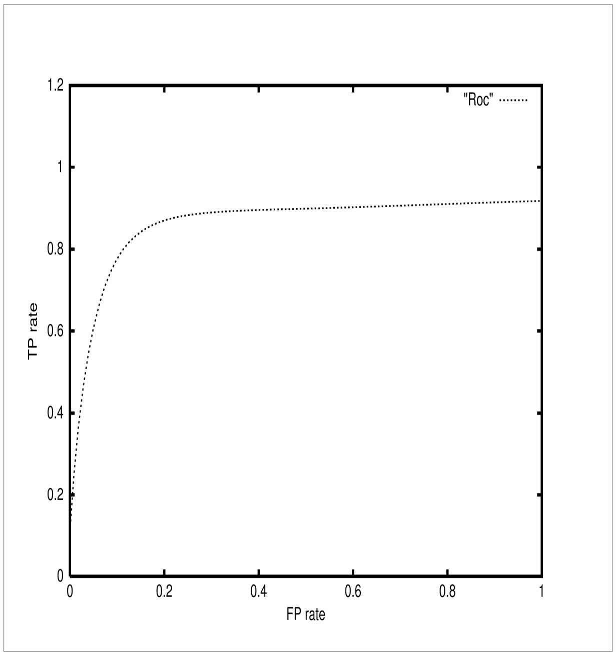

Figure 3. ROC plot for score S2. S2 was used as a threshold to evaluate the ability to identify split domains in the proteomes of human, mouse and zebra fish. The area under the curve (AUC) is 0.89.

Figure 4 demonstrates the high specificity (specificity > 0.998) and the low False discovery rate of this method.

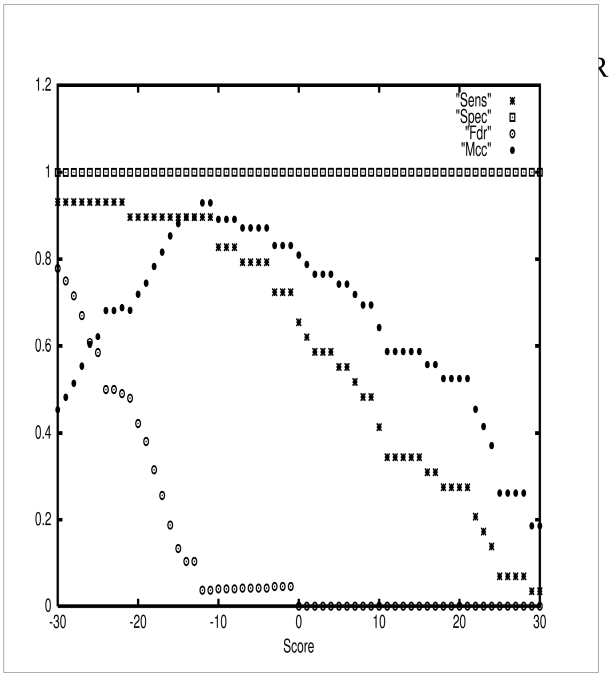

Figure 4. Evaluation of S2 score for predicting split domains in the proteomes of human, mouse and zebra fish. The false discovery rate (Fdr) is zero for all positive values of S2. The specificity (Spec) is greater than 0.998 for the range of scores displayed. Sensitivity (Sens) decreases continuously with increasing values of the threshold for S2. Matthew's correlation coefficient (Mcc) reaches a peak of 0.93 for S2 score threshold of -11, corresponding to a sensitivity of 0.9.

[IV] DISCUSSION

Based on the definitions of the scores and the results of experiments, the recommended method for evaluation of results for an unknown protein is as follows. First check that S0 is positive. If S0 is not positive then the model for split domains is a poorer description of the sequence than a random model. If S0 is positive then check that both S1 and S2 are positive. This ensures that the probability of the Split domain model is higher than the probability of either domain (parent or insert) considered in isolation. Finally, a positive value of S3 ensures that the split domain model with the correct specification of the parent and insert domain is the best model for explaining the (sequence) data. Large positive values for the log odds scores S0, S1, S2 and S3 indicate that there is a high probability that the protein sequence contains a split domain with a domain insertion. HMMsd is based on a probabilistic model; therefore, matches can be identified and evaluated reliably. The examples described here demonstrate that positive values of S0, S1, S2 and S3 scores can be used to identify domains split by the presence of other insert domains in a protein sequence. The sensitivity of this method can be increased further (accompanied by an increase in the false discovery rate) by considering S2 scores that are slightly negative. If this is required, then it would be advisable to combine this information with other methods of sequence analysis.

The transition and emission probabilities for this model are almost completely determined by the parameters of the constituent profile Hidden Markov Models of the two specified domains [22,23]. These emission and transition probabilities for the constituent profile HMMs [22,23] can be obtained from PFAM [25]. Therefore, this method does not require any user parameterization. However, the parameters can be changed if required. This method allows us to leverage the substantial manual curation of the parameters describing the PFAM HMM profiles.

The time requirement of the Viterbi algorithm for the Hidden Markov model for split domains (HMMsd) is O((N+N*)L^2); N is the number of match states in the HMM profile of the parent domain, N* is the number of match states in the HMM profile of the inserted domain and L is the length of the sequence. If an entire proteome is to be searched for all possible split domains, then the computational time required by using this method is quite high. However, the proteomic search can be speeded up considerably, by using a simple filter based on the ability to find the insert domain with a low threshold – the insert domain is not split, therefore it can be detected with a reasonably high level of sensitivity by using a single profile.

In a proteomic scan, sequences that have a reasonable chance of containing the insert domain are identified and these are then analyzed by using the complete model described here – this approach results in substantial speed up with little loss of sensitivity. Using this approach, the human proteome can be subjected to HMMsd analysis in less than half an hour on a PC (for a specified pair of parent and insert domains).

[V] CONCLUSION

The search for split domains can be automated since HMMsd does not require any parameters other than those available in the Pfam HMM profiles. Hence, the human involvement would be quite minimal and the search for specific pairs of domains can be carried out in parallel using a large number of computers, if necessary.

FINANCIAL DISCLOSURE

None required.

ACKNOWLEDGEMENT

The author would like to acknowledge S.R.Eddy for free access to the HMMER package (Eddy, 2003) of programs released under the GNU public licence (GPL) and the accompanying documentation. The current implementation of the HMM model described here was derived from the programs and utilities provided in the HMMER package. The author would also like to acknowledge the use of PFAM, NCBI, EBI and Swiss-Prot sequence databases; protein structure (RCSB) database; Pubmed, Google Scholar and PubMedCentral (PMC) literature databases; Free software: Linux (Debian, Ubuntu, Knoppix, Puppy) and OpenOffice.

REFERENCES

[1] Brayton et al. Complete genome sequencing of Anaplasma marginale reveals that the surface is skewed to two superfamilies of outer membrane proteins. *Proc. Natl. Acad. Sci. USA.* **102**: 844-849, 2005.

[2] Wetlaufer DB, Nucleation, rapid folding and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. USA* **70**:697-701, 1973.

[3] Aroul-Selvam R, Hubbard T, Sasidharan R, Domain insertions in protein structures. *J. Mol. Biol.* **338**:633-641, 2004.

[4] Chang J-S, Interaction of elongation factor-1 alpha and pleckstrin homology domain. *J. Biol. Chem.* **277**:19697-19702, 2002.

[5] Gibson TJ, Hyvonen M, Musacchio A, Saraste M and Birney E, PH domain: the first anniversary. *Trends Biochem. Sci.* **19**:349-353, 1994.

[6] Ponting CP, Phillips C, Davies KE, Blake DJ, PDZ domains: targeting signalling molecules to submembranous sites. *Bioessays.* **19**:469-479, 1997.

[7] Yan, J. et al. Structure of the split PH domain and distinct lipid-binding properties of the PH-PDZ supramodule of A-syntrophin. *EMBO Journal.* **24**:3985-3995, 2005.

[8] Adams ME, Dwyer TM, Dowler LL, White RA and Froehner SC, Mouse alpha1-and beta2-Syntrophin Gene Structure, Chromosome Localization, and Homology with a Discs large domain. *J. Biol. Chem.* **270**:25859-25865, 1995.

[9] Richards MW et al. The HOOK-domain between the SH3 and GK domains of CavB subunits contains key determinants controlling the calcium channel inactivation. *Channels.* **1**:92-101, 2007.

[10] Lemmon MA,Pleckstrin Homology Domains: two halves make a hole? *Cell.* **120**:574-576, 2005.

[11] van Rossum, D.B., Patterson, R.L., Sharma, S., Barrow, R.K., Kornberg, M., Gill, D.L. and Snyder SH, Phospholipase CG1 controls surface expression of TRPC3 through an intermolecular PH domain. *Nature.* **434**:99-104, 2005.

[12] Altschul SF et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nuc. Acids Res.* **25**:3389-3402, 1997.

[13] Jiang H and Blouin C. Insertions and emergence of novel protein structure: a structure based phylogenetic study of insertions. *BMC Bioinformatics.* **8**:444-457, 2007.

[14] Fong JH and Marchler-Bauer A. Protein subfamily assignemnt using the conserved domain database. *BMC Research Notes.* **114**:114-120, 2008.

[15] Krogh A, *Computational methods in molecular biology*. Ed. Salzberg et al. Elsevier. 45-63, 1998.

[16] Durbin R, Eddy S, Krogh A, Mitchison G, *Biological Sequence Analysis. Probabilistic models.* Cambridge University Press, 1998.

[17] Terrapon N, Gascuel O, Marechal E and Breehelin L. Detection of new proteins using co-occurrence: application to Plasmodium falciparum. *Bioinformatics*. **25**:3077-3083, 2009.

[18] The Uniprot Consortium The Universal Protein resource (UniProt). *Nuc. Acids Res.* **36** :D190-195, 2008.

[19] Junker VL, Apweiler R and Bairoch A. Representation of functional information in the SWISS-PROT data bank. *Bioinformatics*. **15**:1066-1067, 1999.

[20] Sayers EW, Database resources of the National Center for Biotechnology Information. *Nuc. Acids Res.* **37**:D5-D15, 2009.

[21] Berman HM, Westbrook J, Feng Z, Gilliland, G., Bhat, T.N., Weissig, The Protein Data Bank. *Nuc. Acids Res.* **28**:235-242, 2000.

[22] Eddy SR, Profile Hidden Markov Models. *Bioinformatics*. **14**:755-763, 1998.

[23] Eddy SR, HMMER User's guide. 2.3.2, 1-90, 2003.

[24] Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe KL, Marshall M and Sonnhammer ELL. The Pfam protein families database. *Nuc. Acids Res.* **28**:235-242, 2000.

[25] Finn RD *et al.* The Pfam Protein Families Database. *Nuc. Acids Res.* **36**:D281-D288, 2008.

[26] Piluso G *et al*, G1- and G2-syntrophins, two novel dystrophin binding proteins. *J. Biol. Chem.* **275**:15851-15860, 2000.

[27] Karlin S and Altschul SF, Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. USA*. **90**:5873-5877, 1993.