11-1-2005

Truncated Profile Hidden Markov Models

Jennifer A. Smith
Boise State University

This document was originally published by IEEE in *Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. Copyright restrictions may apply. DOI: 10.1109/CIBCB.2005.1594926

Truncated Profile Hidden Markov Models

Scott F. Smith, Senior Member IEEE
Department of Electrical and Computer Engineering
Boise State University
Boise, Idaho 83725-2075 USA
sfsmith@boisestate.edu

***Abstract*-The profile hidden Markov model (HMM) is a powerful method for remote homolog database search. However, evaluating the score of each database sequence against a profile HMM is computationally demanding. The computation time required for score evaluation is proportional to the number of states in the profile HMM. This paper examines whether the number of states can be truncated without reducing the ability of the HMM to find proteins containing members of a protein domain family. A genetic algorithm (GA) is presented which finds a good truncation of the HMM states. The results of using truncation on searches of the yeast, E. coli, and pig genomes for several different protein domain families is shown.**

I. Introduction

A profile hidden Markov model (HMM) [1] can be used for very powerful searches of databases for remote homologs. The structure of the profile HMM allows for different distributions of expected residues at each conserved multiple alignment position as well as variable probabilities of insertions and deletions between each of these positions. Given a multiple alignment of a large enough group of protein domains deemed to be homologous, the parameters of the profile HMM can be estimated and other proteins containing this domain searched for using the combined information of all members of the group and not just a single member. Pair-wise alignment algorithms such as Smith-Waterman [2], FASTA [3], and BLAST [4] can not capture the full joint information content of the group even when the multiple-alignment consensus sequence is used as the query. However, the profile HMM can be very slow for database search since the dynamic-programming-based scoring method is very similar to Smith-Waterman, but with a large number of parameters which are likely to be assigned to memory variables and not processor registers.

The computation time needed for profile HMM database search is nearly proportional to the number of HMM states, which in turn is proportional to the number of multiple alignment columns deemed conserved when designing the model. Depending on the protein domain family being modeled, the number of conserved columns tends to range from about ten up into the hundreds. The traditional way to choose if an alignment column is to be conserved (associated with a match state) is to include the column if it is expected to contribute *any* improvement to the signal to noise ratio of the score. There is usually no consideration of the tradeoff between extra computation time and potential gain in actually finding more remote homologs. It is the purpose of this paper to investigate whether it is common that a significant number of these columns could be eliminated with negligible effect on search efficacy and (if so) to present a method for finding which columns to eliminate.

A database of protein domain families and associated profile hidden Markov models is available as Pfam [5]. The HMMER [6] program suite was used to search the UniProt [7] database to find the family members in the Pfam database. A truncated (some conserved columns removed) HMM will be considered acceptable for the purposed of this paper if the truncated model returns exactly the same set of Pfam family members at the top of its score-ranked list as the untruncated model. The rank-order of the family members found is not considered important as long as they are all there and no false positives score higher than the lowest scoring true family member (as defined by the untruncated model and Pfam).

Trying all combinations of excluded columns in an HMM is in impractical undertaking. The number of combinations of column exclusion for a 250 match-state HMM is enormous and evaluating if the resulting model is acceptable by using the model for a database search can take on the order of minutes for each evaluation even for subsets of UniProt. As an alternative, a genetic algorithm (GA) is proposed for finding a good truncation. A good truncation is one that is acceptable as defined above and has a number of states reasonably close to the minimum possible number of states among all acceptable truncations. Since the purpose here is to show than a significant amount truncation is possible, finding the absolute minimum number of states is not needed.

The genetic algorithm details for finding good truncations are given in section II. Results of using the GA for the *S. cerevisiae* (baker's yeast) genome subset of UniProt for a number of Pfam protein domain families are shown in section III. The truncations found using yeast are cross-validated on *E. coli* and *Sus scrofa* (pig) data in section IV. Section V presents some concluding remarks.

II. GA For Finding Truncations

A. Representation of Truncated HMM

The profile HMM is composed of one begin state (B), one end state (E), one insert state at the start of the model (I_0), and any number of stages (indexed with i) each containing a single match state (M_i), a single insert state (I_i), and a single delete

0-7803-9387-2/05/$20.00 ©2005 IEEE

state (D_i). The structure of a three stage profile hidden Markov model is shown in Fig. 1, where the possible transitions are shown with arrows. All insert states (I) also have self-transitions which are not shown in the figure. The B, E, and I_0 states will remain in both the initial and truncated models. The truncated model will exclude states in sets of three (M_i, I_i, and D_i with the same i). If the initial model has N stages, the inclusion or exclusion of a stage (set of three states) can be represented as a binary sequence of length N where a 1 represents inclusion and a 0 represents exclusion of the stage. Using this representation, the initial model is a sequence of N 1s.

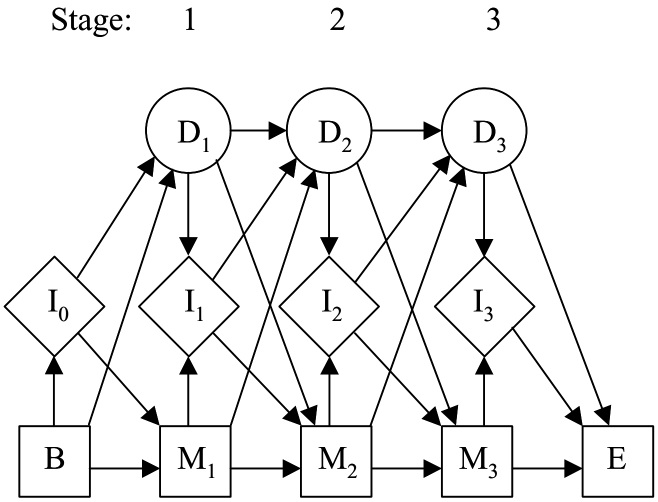

Fig. 1. Structure of 3-stage profile hidden Markov model. Self loops for insert states (I) not shown.

B. Allowed Mutations

The final pattern of excluded states is expected to have relatively contiguous regions of 0s and 1s. This is expected since conserved columns of the multiple alignment tend to come in contiguous regions. These regions often correspond to helices and sheets (often in the core portion of the protein) or to active areas of the protein [8]. It is expected that retaining or removing these regions as a whole is more likely to result in fewer states in conjunction with an acceptable model than a sequence of ones and zeros with no autocorrelation. However, such non-correlated structure should not be totally ruled out since this expectation is only conjecture.

To encourage clumps of 0s and 1s in the hope of faster convergence to a good solution, mutations will take the form of changing a range of values to either all zeros or all ones. A starting point is uniformly chosen along the length of the sequence. The length of the range to be changed is chosen as a value uniformly distributed in the range 1 to N. If the starting point and length of the range imply changes beyond the end of the sequence, changes are made exactly to the end of the sequence. Since this method tends to make changes more often to locations near the N-terminal end of the sequence, the choice of whether to count positions starting on the left or right end of the sequence is made with one-half probability for each. Finally, whether to change the range to all 0s or all 1s is made with one-half probability for each.

C. Other Details of the GA

Each generation contains 100 individuals. The initial population contains one member with all ones (exactly the untruncated model). The remaining 99 initial individuals are broken into three groups of equal size (33 each). The first group gets one mutation relative to the untruncated model. The second group is mutated twice and the third three times.

At the end of each generation the fittest individual is retained without modification. Nine copies of the fittest individual are mutated once. Four copies each of the twenty fittest individuals below the most fit individual (fitness rank 2 through 21) are each mutated once. The remaining ten new individuals are single-point crossovers of any two randomly selected individuals from the top half of the fitness range of the previous generation. The GA is run for a total of ten generations.

The fitness of a individual is evaluated by searching a database with the truncated HMM and by counting the number of 1s in the representation. Before the GA is run, the untruncated HMM is run against the database and a list of Pfam protein domain family members found by the HMM as an uninterrupted series at the top of the ranked score list is recorded. A truncated model is acceptable if the list for the truncated model is the same as for the untruncated model (without regard to order). Unacceptable models get a fitness contribution of -(N+1) and acceptable models a fitness of 0 based on this criterion. The total fitness is the contribution due to acceptability minus the number of 1s in the representation. With this fitness calculation, the untruncated model gets a fitness of -N, all unacceptable models have fitness less than -N, and all acceptable models with some truncation have a fitness of greater than -N. When ranking individuals with the same fitness, the rank order is chosen randomly.

III. Results For Yeast

A. Database and Initial Profile HMM Models

The UniProt database of all known or putative proteins in the *S. cerevisiae* (baker's yeast) genome [9] is used as a search target. This database was selected since it is well studied and therefore expected to be relatively accurate. The database has 5095 entries making it the third largest single organism database in UniProt (after human and mouse). Use of the full UniProt database of all available organisms was too large for this study. Since this paper is only meant to explore the possibilities of truncated HMM models, the shorter database will suffice. Release 47.2 of the database is used.

Development of a full set of truncated models using the full UniProt database would require the use of a grid computing environment or large cluster of computers. The model truncation would need to be done only once whereas the speed improvement of the truncated models would be observed repeatedly with every new query sequence tested against the model set. It is possible to select only those HMM models that are very large for truncation since the truncation processing time is likely to give the greatest benefit for these models. It is also possible to truncate the model set while the model set is in use, slowly improving search time with each release with computing resources applied to truncation as available.

Initial (untruncated) profile models are obtained from Pfam for those protein domain families in the "top twenty" classification that had any members associated with yeast. The "top twenty" domains are those which are most numerous in the database. Table I shows the models used, the number of occurrences of yeast for that model in the database, the number of stages in the model, the average length of sequences used to form the model, and the percentage residue identity in multiple alignment columns. The models are from release 17.0 of Pfam. The number of HMM stages in typically larger than the average length of family member sequences since insertions which appear in far fewer than half of the member sequences are often assigned stages (with relatively small penalties for visiting the D states of that stage). Comparing the percent identity column with the number of stages shows that many of the easily identified families (high percent identity) also have many stages. These large high-identity models may well be orders of magnitude more powerful than they really need to be, whereas short low-identity models may need all the power they contain.

TABLE I
PFAM PROTEIN DOMAIN FAMILY MODELS

Pfam Name	Yeast Members	HMM Stages	Average Length	Percent Identity
RVT1	4	241	167	74
zf-C2H2	42	23	23	37
RVP	1	110	93	86
LRR1	10	24	23	26
CytoChBN	6	209	154	69
WD40	92	37	38	19
COX1	8	488	229	48
Ank	20	32	30	27
ABCtran	35	205	185	26
CytoChBC	1	111	89	74
Pkinase	122	287	228	24
TPR1	25	33	33	18
PPR	2	34	32	20
zf-CCHC	8	17	17	51

B. *Retained HMM Model States*

The results of the GA-based truncation are shown in Table II. The number of truncated stages is shown along with the ratio of truncated stages to original HMM stages. The included stages column shows which of the original HMM stages were retained in the truncated model. Even though the GA frequently generates groups of four or five retained stages during execution, the final solution never has more than three groups and in most cases has one or two groups. This reinforces the idea that using a GA that prefers large clumps of retained states, but which does not exclude smaller clumps, is likely to be more efficient than a GA that does not prefer large clumps.

Those models which have many stages in the original tend to be the ones that can be significantly truncated. For instance, the RVT1 family with 241 initial stages was reduced to 53 truncated stages, whereas the TPR1, zf-C2H2, and zf-CCHC families with 33, 23, and 17 initial stages respectively allowed for very little proportional reduction in size. This is probably due to models with less than about 15-20 stages having insufficient discriminatory power against random sequences of amino acids.

TABLE II
TRUNCATED MODELS

Pfam Name	Truncated Stages	Fraction of Orig. Size	Included Stages
RVT1	53	0.220	51-103
zf-C2H2	19	0.826	5-23
RVP	81	0.736	17-73, 83-106
LRR1	22	0.917	1-22
CytoChBN	108	0.517	32-91, 99-104, 122-163
WD40	36	0.973	2-37
COX1	103	0.211	27-106, 135-141, 170-185
Ank	30	0.938	1-21, 24-32
ABCtran	163	0.795	6-47, 58-178
CytoChBC	30	0.270	12-34, 87-93
Pkinase	246	0.857	1-235, 243-253
TPR1	32	0.970	1-32
PPR	23	0.676	1-2, 5-25
zf-CCHC	15	0.882	3-17

C. In Depth Discussion of WD40 Domain

The least amount of size reduction for any of the fourteen protein domain families examined is for the WD40 domain. This domain has highly conserved residues at stages 2-3 (consensus G and H), stage 7 (V), stage 14 (P), stage 23 (L), stages 25-27 (S, G, and S), stage 29 (D), and 36-37 (W and D). Of these, the three most highly conserved stages are 3, 29, and 36. Since the conservation pattern is spread out over the entire 37 stage model, it is not surprising that model truncation is difficult. Only stage 1 was easily cleaved off the end of the model. In general, internal groups of stages are harder to truncate than ends. This is due the information contained in the state transition probabilities. Single conserved positions that need to be a specific number of residues apart require the intervening states to maintain this separation information even if the emission probabilities of these states is nearly uninformative. This same reasoning helps explain the poor ability to reduce states in the models of other binding protein domains such as zf-C2H2 and zf-CCHC (two types of zinc finger domains).

D. In Depth Discussion of COX1 Domain

The COX1 (Cytochrome c oxidase) domain model showed the greatest proportional size reduction of the fourteen models truncated. The alignment of this family shows many contiguous highly conserved regions. The GA has chosen to remove a relatively weak region from the start of the model, but retained the very highly conserved R at stage 33. Several very strongly conserved regions after stage 185 have been truncated. Major truncation of this family was a very simple task and some very informative regions of the HMM have been discarded simply because the original model was far more powerful than necessary. This model could have potentially benefited from a longer run of the GA since it would appear to have many local minima.

IV. Cross Comparison with E. Coli and PIG

The results of the previous section have shown that hidden Markov models of protein domain families can be made significantly shorter without any effect on the ability of the models to discriminate between proteins containing the domain family and proteins that do not. However, it is not clear whether the truncation based on the yeast training set can be generalized to other organisms. This section examines how well the truncations determined using the GA and yeast data work when applied to protein sequences of two other organisms which were not used during the truncation selection process. The two organisms are *Escherichia coli* (a bacterium) [10] and *Sus scrofa* (pig). The data for pig was not available as a stand-alone file from the UniProt database, so it was retrieved via the UniProt Sequence Retrieval System [11].

TABLE III
E. COLI RESULTS USING YEAST TRUNCATED MODELS

Pfam Name	E. Coli Members	Members with False Positive Above for Truncated Model
RVT1	1	0
zf-C2H2	0	-
RVP	0	-
LRR1	1	0
CytoChBN	0	-
WD40	0	-
COX1	1	0
Ank	1	0
ABCtran	78	0
CytoChBC	0	-
Pkinase	0	-
TPR1	4	0
PPR	0	-
zf-CCHC	0	-

A. Cross Comparison of Yeast Truncation on E. coli Data

Table III shows the number of true positives for each family as determined by the untruncated HMM and reference to the Pfam database. It also shows the number of true positives within each family that were ranked lower than at least one false positive in the score list generated by the truncated model. Eight of the fourteen protein domain families do not appear anywhere in the *E. coli* data and could therefore not be evaluated. This is due to the significant differences in biochemical processes between the eukaryote *S. cerevisiae* and the prokaryote *E. coli*. For the other six families the results are perfect. In spite of the limited number of overlapping families between the two organisms, it was deemed important to compare organisms from two different kingdoms. The next subsection will look at organisms with more overlap, but also more similarity.

B. Cross Comparison of Yeast Truncation on Pig Data

All except one of the fourteen protein domains for which truncated models were found using yeast are also found in the pig data. Table IV shows the number of pig protein sequences which the untruncated HMM found to contain at least one copy of the protein domain family. The rightmost column of the table shows the number of true positives which have at least one false positive ranked above it. There is a question as to where the ranked score list should be cut off using the untruncated HMM to separate true family members versus sequences deemed not to contain the domain family. To resolve this, the ranked list generated with the untruncated HMM was compared to the family members as listed by Pfam. All sequences ranking at or above the location of the lowest-ranking sequence on the Pfam list were taken a true positives. This is important due to the fact that sometimes Pfam does not list a protein that the untruncated HMM gives a high score to. This can happen for at least two reasons. First, the UniProt data was obtained in June 2005 and the last Pfam update at that time was generated from March 2005 Uniprot data. Sequences added after March 2005 to Uniprot sometimes score very high on the untruncated HMM and the Uniprot annotation normally indicates that the sequence should indeed be a family member. Second, high-scoring sequences may not have been included in Pfam if the sequence is known by an expert to not contain the domain (in spite of its high score). A good truncated HMM will tend to also assign a high score these non-Pfam listed sequences, so they are taken as true positives for the purpose of this study.

TABLE IV
PIG RESULTS USING YEAST TRUNCATED MODELS

Pfam Name	Pig Members	Members with False Positive Above for Truncated Model
RVT1	4	0
zf-C2H2	20	0
RVP	21	0
LRR1	35	0
CytoChBN	36	1 (Q5YLL5)
WD40	13	0
COX1	7	0
Ank	12	1 (Q9TSY1)
ABCtran	6	0
CytoChBC	33	0
Pkinase	67	1 (Q9N0K8)
TPR1	3	0
PPR	0	-
zf-CCHC	4	0

There are three cases in Table IV where a true positive protein sequence had a least one false positive ranked above it: the sequence Q5YLL5 with a domain in the CytochromBN family, sequence Q9TSY1 with a domain in the Ank family, and sequence Q9N0K8 with a domain in the Pkinase family. In all three cases the protein is a protein fragment and comes from the TrEMBL supplement to the Swiss-Prot database. The Q5YLL5 sequence is predicted to contain only the final 49 residues of the Cytochrome b N-terminal domain using the untruncated HMM model. Q5YLL5 had an original rank of 36 out of 36 using the full model and has at least 6 false positives ahead of it using the truncated model (with E-value greater than the display cutoff of 10.0). The Q9TSY1 sequence contains a partial copy and a full copy of the Ankyrin repeat domain according to the full HMM. Most sequences found by the full model have at least two full copies of the repeat and some as many as six full copies of the repeat. The original rank of Q9TSY1 was 11 out of 12 and the sequence has two false positives above it in the truncated-model ranking. The Q9N0K8 sequence is not listed as containing a domain in the protein kinase family according to the Pfam 17.0 database. It is not clear why this is, since the sequence ranks 62 out of 67 using the original HMM and the protein is listed as being a protein kinase in UniProt and the sequence was added before March 2005. Q9N0K8 has two false positives ranked above it using the truncated model.

While not perfect, the models truncated using yeast data did extremely well on the pig data. The tradeoff of possibly missing an occasional marginal protein in the database search might be worthwhile if database search response time is important.

V. Conclusions

It has been found that significant truncation of Pfam hidden Markov models can be done with extremely little adverse effect on the ability of the models to detect sequences containing domains of the model family. The truncation is done by simply deleting triples of HMM states from the model (one M, one I, and one D state in each triple) using a genetic algorithm. The GA can generate the truncation using a small subset of known family members as the training set (such as the subset of protein sequences from yeast) and the resulting truncation works well on other organisms. While the GA to select the truncation is rather computationally intensive, it only needs to be done once on each model. Subsequent use of the truncated models to search databases is accelerated in proportion to the size reduction of the hidden Markov models. For the fourteen families tested in this study, the size reduction appears to be on the order of thirty percent.

The GA works well in selecting truncated models, however it is not yet known whether other solution finding methods might outperform the GA in terms of solution quality or computation time. It is also not yet clear if there is a better way to initialize the search. Since the solutions tend to be correlated with conserved regions of the multiple alignment, perhaps starting with highly conserved regions as an initial guess might speed up solution finding.

The method presented is greatly simplified by the fact that HMM states are simply removed from the model without a new estimation the model with the new structure. In other words, the transition probabilities from the last retained stage of a retained block to the first retained stage of the next block are not optimal. Further investigation is needed to determine if a new model parameter estimation after each mutation would significantly change the chosen retained stages.

Acknowledgment

The project described was supported by NIH Grant Number P20 RR016454 from the INBRE Program of the National Center for Research Resources.

References

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

[2] T. Smith and M. Waterman, “Identification of Common Molecular Subsequences,” *Journal of Molecular Biology*, vol. 147, pp. 195-197, 1981.

[3] W. Pearson and D. Lipman, “Improved Tools for Biological Sequence Comparison,” *Proceedings of the National Academy of Sciences*, vol. 85, pp. 2444-2448, 1988.

[4] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, “Basic Local Alignment Search Tool,” *Journal of Molecular Biology*, vol. 215, pp. 403-410, 1990.

[5] A. Bateman, L. Coin, R. Durbin, R. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. Sonnhammer, D. Studholme, C. Yeats, and S. Eddy, “The Pfam Protein Families Database,” *Nucleic Acids Research*, vol. 32, pp. D138-D141, 2004.

[6] HMMER, http://hmmer.wustl.edu.

[7] A. Bairoch, R. Apweiler, C. Wu, W. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. Martin, D. Natale, C. O'Donovan, N. Redaschi, and L. Yeh, “The Universal Protein Resource (UniProt),” *Nucleic Acids Research*, vol. 33, pp. D154-D159, 2005.

[8] C. Branden and J. Tooze, *Introduction to Protein Structure*, 2nd Ed., Garland Publishing, 1999.

[9] UniProt Yeast Database, ftp://us.expasy.org/databases/complete_proteomes/fasta/eukaryota/yeast.fas, Release 47.2, 07 June 2005.

[10] UniProt E. coli Database, ftp://us.expasy.org/databases/complete_proteomes/fasta/bacteria/ecoli.fas, Release 47.2, 07 June 2005.

[11] UniProt Sequence Retrieval System, http://us.expasy.org/srs5/, accessed 07 June 2005 to obtain S. scrofa Release 47.2 sequences.