RAMMER: Accelerating Protein Function Prediction

Brett Poulin*, Duane Szafron, Russell Greiner,
Roman Eisner and Paul Lu
***poulin@cs.ualberta.ca**

PENCE BIOINFORMATICS I

UNIVERSITY OF ALBERTA

www.cs.ualberta.ca/~bioinfo

Why accelerate prediction?

A wide range of hidden Markov models (HMMs), such as those found in the Pfam database [1], have been used to represent and detect many protein families by amino acid sequence. However, profile HMM evaluation against sequences is relatively slow. To accelerate the process, we followed the example of BLAST by using rapid approximations. HMMs can be accelerated by rapid approximation of full HMM evaluation with computationally efficient higher-order Markov chains (MCs). Using MCs in combination with HMMs can speed up the identification of HMM-sequence matches without loss of precision and with little or no loss of recall.

How can we do it?

We use Markov chains as a pre-processing filter for HMM software (HMMER)[2], removing the most unlikely matches (low Markov chain scores) from processing. This technique for accelerating HMM evaluation makes HMM sequence analysis more practical for large-scale and widespread use. This method can be used for both ways of using HMMs ('family' or 'sequence' scans).

What are Markov chains?

Markov chains are simple (and fast!) probabilistic models that score protein sequences by how well they fit the model.

Given a protein sequence *x* where the $x_1 \ldots x_i \ldots x_n$ are residues, the score $S(x)$ is

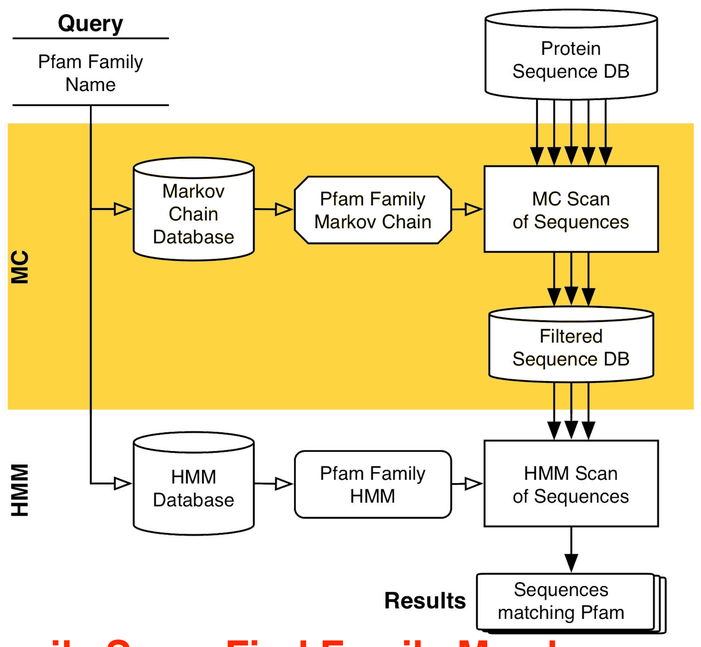
$$S(x) = \log\left(\frac{P(x \mid \text{family model})}{P(x \mid \text{null model})}\right)$$

where the probability is

$$P(x) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1 x_2)\ldots P(x_n \mid x_1 \ldots x_{n-1}) = P(x_1)\prod_{i=2\ldots n} P(x_i \mid x_1 \ldots x_{i-1})$$

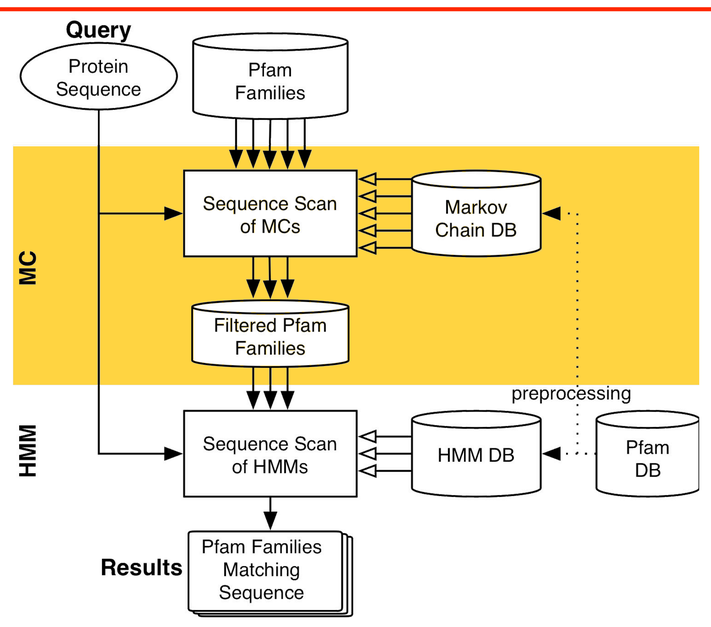
Using a Markov assumption that allows us to simplify to use

$$P_{0\text{-order}}(x_i \mid x_1 \ldots x_{i-1}) \approx P(x_i)$$
$$P_{1\text{-order}}(x_i \mid x_1 \ldots x_{i-1}) \approx P(x_i \mid x_{i-1})$$
$$P_{2\text{-order}}(x_i \mid x_1 \ldots x_{i-1}) \approx P(x_i \mid x_{i-2} x_{i-1})$$
$$\vdots$$



Family Scan: Find Family Members

A 'family scan' corresponds to use of the HMMER program 'hmmsearch'. A RAMMER 'family scan' uses a Markov chain (MC) representing the family to filter sequences from a database. The sequences that the MC evaluates as likely members of the protein family are then passed to an HMM representing the family. The coloured portion of the figure indicates the RAMMER addition to a typical HMM evaluation.



Sequence Scan: Find a Protein's Family

A 'sequence scan' corresponds to use of the HMMER program 'hmmpfam'. A RAMMER 'sequence scan' uses Markov Chains to filter protein families from a database (such as Pfam). The protein families that are most likely to match the query sequence are then evaluated using an HMM for each of the likely families. The coloured portion of the figure indicates the RAMMER addition to a typical HMM system.

How fast is RAMMER?

The speedup is calculated using Pfam families against the entire SwissProt database (42.7). The cumulative speedup (all sequences against all families) was 25.2X (65.4 days for HMMER and 2.6 days for RAMMER).

Cumulative Speedup	25.2 X
Average Speedup	39.7 X
Minimum Speedup	0.70 X
Median Speedup	32.2 X
Maximum Speedup	463.4 X

The following results show a sample of 'family scans' using some sample Pfam families.

Pfam Family (HMM len)	HMMER	RAMMER	Speedup
zf-C2HC5 (56)	220.1 s	18.2 s	12.1 X
Ras (174)	682.5 s	79.1 s	8.6 X
Acetyltransf_2 (263)	973.1 s	32.2 s	30.2 X
Voltage_CLC (455)	1656.1 s	24.5 s	67.6 X
AA_permease (528)	1916.0 s	56.2 s	34.1 X
RNA_pol (562)	2033.7 s	22.2 s	92.5 X

Similar, though less dramatic, speedups are observed for sequence scans. Great speedups are also observed when scanning multiple sequences against multiple families in high-throughput processing.

Why should you use it?

The acceleration method reported here will be especially helpful for high-throughput automated analysis. The RAMMER (hybrid MC+HMM) approach benefits from the fact that no false positives will be reported when compared against HMMER alone (since HMMER is used in both cases). Depending on the application, MC score thresholds may be selected to favor recall or speed. The use of efficient Markov chains shows great promise for making HMM evaluation more practical and popular as a tool to be used along with the ubiquitous sequence alignment.

Overall performance improvements:

- ~40X average speedup for family scan
- ~4X average speedup for sequence scan
- Range of speedups (4-40X) for matching many sequences against many families.

RAMMER: Rapid Approximating Markov Models

ALBERTA INGENUITY CENTRE FOR MACHINE LEARNING

iCORE

NSERC CRSNG

Sun microsystems

References

1) Bateman *et al.* 2004. Pfam. NAR 32:D138-D141.
2) Eddy *et al.* HMMER. http://hmmer.wustl.edu.