CMfinder - a covariance model based algorithm [1]

Fayette Shaw

December 7, 2005

We haven't talked about how to come up with model where we are not certain of presence, location and size. For example, given 10-20 sequences of ∼500 bases, many but not all of which may contain a motif of 50-100 bases, how do we find, align and predict the secondary structure? There are a huge number of possibilities of location and secondary structures. In Rfam, most of this is done manually, but a more automated approach is desirable for many purposes.

One approach is comparison:

1. Align sequences first then predict common secondary structures
   - example: CLUSTALW doesn't align SECIS well
   - Alignment without structure information is unreliable
2. Predict structures of single stranded RNA folding, then align
   - 60% accurate for single seqence struture predction
   - motif may appear to interact w/ flanking sequence
   - no widely accepted model for aligning such motifs
3. Do both together - good but expensive
   - Sankoff 1985 - align sequence & structure on evolutionary tree. for 2 sequences this is n^6 algorithm, 3 sequence n^9 algorithm
   - heuristics exist to do this, but still generally too slow except for small data sets and simple structures like single hairpins

Design goal: Search for RNA motifs in unaligned sequences.

- Perform local alignment - given unaligned sequences there exists a repeated/matching local alignment
- Exploit but do not require sequence conservation
- Robust to inclusion of unrelated sequences - may have sequences missing
- Reasonably fast and scalable.
- Produce a probabilistic model of the motif that can be directly used for homolog search.

The last point is important: finding more examples allows for model refinements and better secondary structure predictions. We build a covariance model to use for the search as the results of search go into the covariance model. Secondary sequence folding prediction goes into the alignment. Unaligned sequences go into an initial alignment.

Mutual information is good evidence of secondary structure, but often don't have enough data at an appropriate evolutionary distance. The folding prediction is reasonably reproducible

Can we blend these two approaches? Mutual Information + Folding = Probabilistic model?

An example of a blind test: Tbox. It worked out well, but note (in slides) that there are false positives and negatives. Still, this automated approach compares reasonably well to the hand-created Rfam models. We ended up with roughly 90% sensitivity / specificity on several tests.

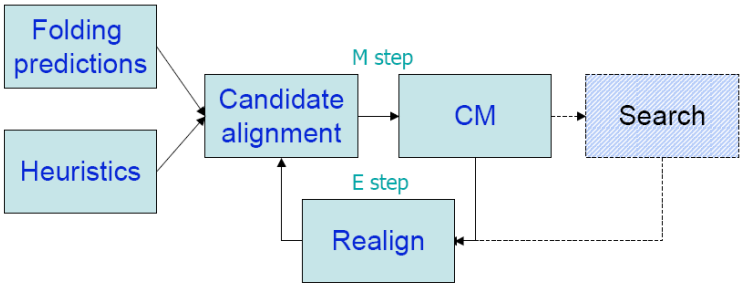

Figure 1: CMfinder procedure

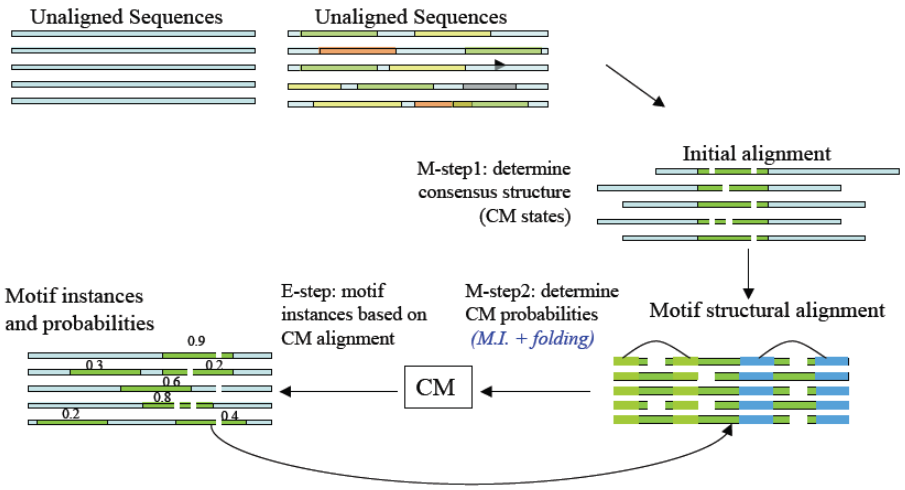

Figure 2: Start with an alignment, produce a CM, realign against the model (EM-like)

References

[1] Zizhen Yao, Zasha Weinberg, and Walter L. Ruzzo. Cmfinder: A covariance model based RNA motif finding algorithm. *Bioinformatics*, 2005. to appear.