Introduction to Motifs and Motif Finding

Jeremy Buhler

July 29, 2014

1 What do we Mean by a Sequence Motif?

We can talk about motifs from a biological or a computational standpoint.

- Biologically speaking, a *motif* in DNA or RNA (or protein) sequence is a short functional sequence element.
- Examples found in genomic DNA include
  - transcription factor binding sites
  - small noncoding RNAs
  - small repetitive elements (e.g. inverted repeats)
  - common mRNA elements, such as Shine-Dalgarno or Kozak sequences, splice enhancers and suppressors.
- We will be strongly tempted to talk about *conserved* DNA sequence motifs. But we must be careful to recognize that many motifs, especially short transcription factor binding sites, are known to arise convergently. Hence, the various instances of a motif, while similar in appearance, may not actually be homologs. Don't say "conserved" unless you mean it!

To understand motif-finding algorithms, it will help to consider a more abstract conception of what a "motif" is.

- A motif is a short sequence element that is repeated, perhaps with variation, multiple times in a collection of sequences.
- Typical motif lengths are five to a few tens of bases.

- What does "repeated with variation" mean? We need to define this idea more precisely in order to build motif-finding algorithms.

A motif is described by a computational *model* that specifies how it may vary across its instances. We will consider several important questions for working with motif models:

1. Given a putative motif instance, how well does it fit the model?
2. Given a collection of putative instances, can we derive a model that fits them well?
3. When is a putative motif *interesting* enough to warrant further study?

Let's focus for a minute on what "interesting" means.

- Given a putative motif, we would like to reject the null hypothesis that it is just a bunch of unrelated sequences.
- More precisely, we typically assume a *background* model for sequences that are not part of the motif, and we want to reject the hypothesis that the putative motif's sequences are unrelated samples from this background.
- Intuitively, a putative motif whose instances display less variation should be more interesting, since the instances are more similar than would be expected for unrelated sequences drawn from the background.
- Often, the background is assumed to consist of i.i.d. (independent and identically distributed) random DNA bases, with each base chosen according to some fixed multinomial distribution.
- We will sketch quantitative measures of interest for motifs, but how to turn these measures into p-values is beyond the scope of this talk. Different programs use different approaches.

2 A Simple Model of Motif Sequence Variation

To precisely describe motifs, we introduce two formal models: the *consensus* and the *weight matrix model*. We consider the consensus model first.

- Let C be a "Platonic ideal" sequence of the motif. For example, for a transcription factor binding site, C might be the sequence that most strongly binds its target protein domain.

- Every instance of the motif is assumed to be a variant of C. Instances might deviate from C by some number of *differences* (substitutions, insertions, and/or deletions).
- For example, if C is "agtagc", the actual instances of the motif might be, e.g., "agcagc", "attac", and "agtggc".
- We call C the motif's *consensus* sequence.

Let's consider the questions outlined above for the consensus model.

- A putative motif instance can be aligned to C using the Smith-Waterman algorithm to determine how many differences separate it from C.
- To infer a model C from a collection of instances, we (multiple-) align them and then let each position of the aligned sequences "vote" on what base should be given in the consensus.
- We can describe degenerate consensus positions using IUPAC ambiguity symbols. For example, "R" represents a purine (a or g).
- Inferring C is trivial if motif instances can differ from it only by substitution. If they can have insertions and deletions, finding the "best" C is computationally hard but can be approximated.

OK, what about a measure of motif interest?

- One useful measure of interest for a motif in the consensus model is the *radius*: if we align all instances of the motif against its consensus C, the motif's radius is the smallest r such that every instance differs from C by at most r differences.
- (In the above example, the radius is 2.)
- We also need to know the proposed motif's length to measure interest, since 2 differences in 5 is a lot less interesting than 2 differences in 20.
- Finally, we need to know how many instances exist. A motif with 100 close instances in a given amount of sequence is a lot more interesting than one with just 3.
- To turn this multi-part measure of interest into a p-value, a program must compute the chance that a set of putative motif instances would all arise independently by chance in a given amount of background sequence.

3 Another Model of Motif Sequence Variation

The consensus model, while appealing to people who like combinatorics, is limited in its ability to describe variation in a motif.

- Not every position of a motif may be equally variable. For example, eukaryotic splice donor sites have an almost invariant "gt" at the intron boundary, but the surrounding sequence, while not invariant, exhibits more similarity across splice sites than would be expected by chance.
- Not all differences from the consensus may be equally likely. IUPAC ambiguity symbols cannot capture an arbitrary set of base frequencies in a particular position.

To capture these position- and sequence-dependent effects, we introduce the *weight matrix model* (WMM). For simplicity, we will assume that motifs modeled by a WMM have fixed length – their instances cannot exhibit insertions and deletions relative to the model.

- A WMM for a motif of length ℓ is a $4 \times \ell$ matrix W of probabilities. The four rows of W are labeled with the four DNA bases, while the columns are labeled $1 \ldots \ell$.
- $W(c, i)$ is the probability that position i of a motif instance is the base c.
- Instances of the motif are assumed to be drawn uniformly at random from W. More precisely, the ith base of an instance is drawn independently from the multinomial distribution $W(*, i)$.
- Here's an example W of a WMM with $\ell = 6$:

	1	2	3	4	5	6
a	0.4	0.1	0.25	0.3	0.1	0.1
c	0.1	0.7	0.25	0.2	0.1	0.4
g	0.4	0.1	0.25	0.2	0.1	0.1
t	0.1	0.1	0.25	0.3	0.7	0.4

Once again, let's consider our fundamental questions about motif models.

- For any sequence s of length ℓ, we can compute $\Pr(s \mid W)$, the probability that s arises as an instance of W, by multiplying together the probability of seeing each base in s.

- For example,

$$\begin{aligned}\Pr(\text{acagtc} \mid W) &= 0.4 \times 0.7 \times 0.25 \times 0.2 \times 0.7 \times 0.4 \\ &\approx 0.004.\end{aligned}$$

- Given a collection of putative instances of common length n, we can infer a motif model W by setting the probabilities directly from the counts of each base at each position.

- That is,

$$W(c, i) = \frac{\#\text{ of instances with base } c \text{ at position } i}{\text{total } \#\text{ of instances}}.$$

- The resulting W is guaranteed to give the highest total probability for the observed instances among all possible weight matrices; that is, it is a *maximum-likelihood estimate* of W given the instances.

- (In practice, we don't want to let any entry of W be zero, since we infer these probabilities from only a finite number of examples.)

How do we measure how interesting a putative motif is in this model?

- Let W be a motif inferred from n putative instances $s_1 \ldots s_n$.

- For any s_j, we can compare $\Pr(s_j \mid W)$ to the chance $\Pr(s_j \mid B)$ that s_j arose from some background base distribution B.

- The quantity $\log \frac{\Pr(s_j|W)}{\Pr(s_j|B)}$ is the *log-likelihood ratio* (LLR) for comparing hypotheses W and B given s_j. If this quantity is greater than 0, s_j is more likely to have come from the motif W than from B. If it is less than 0, the opposite is true.

- Our measure of interest for W is the *total LLR* score

$$\sum_{j=1}^{n} \log \frac{\Pr(s_j \mid W)}{\Pr(s_j \mid B)}.$$

- Note that the total LLR of a collection of putative instances is the *sum* of the LLR for each; equivalently, it is the log of the *product* of the probability ratios for the instances.

- Motifs with higher total LLR scores look less like the background and more like their most plausible common WMM!
- Greater similarity among instances makes the probability distributions in W more sharply peaked and hence raises LLR scores].
- To assign a p-value to a putative motif, a program must compute the probability of seeing a motif with total LLR of at least some σ in a given amount of background sequence.

There are many ways to elaborate the basic WMM to describe more complex motifs. An example is the class of symmetric binding sites with a spacer, which are common in bacteria due to the prevalence of homodimeric transcription factors. One can also permit motif instances to exhibit insertions and deletions vs. the WMM by expanding it into a more general *hidden Markov model*.

4 Gathering Evidence for a Motif

We now transition from describing motifs to finding them.

- The models described above provide ways to quantify how interesting a motif is if we *know* its instances and the background in which it occurs.
- In the *motif-finding* problem, we are given some sequences that are believed to contain a motif, but we do *not* know which parts of the sequences are motif instances and which are background.
- There are two main variants of this problem: either the sequences are assumed to be unrelated to each other, or they are assumed to be homologous (and hence alignable). The algorithms used for these two cases are quite different.

We'll start with the case of unrelated sequences.

- Consider, for example, searching for a transcription factor binding site that occurs in the promoters of several different genes. These promoter regions are not usefully homologous, so we treat them as unrelated background sequences.

- We hypothesize that there exists a common motif with instances in some or all of these promoter sequences, though we don't know where the instances are.
- Our goal is to find the *most plausible* motif – a set of putative instances matching a motif that is as unlike the background (high total LLR, or small radius for length) as possible.
- Such a motif (hopefully) explains its putative instances significantly better than does the null hypothesis.

Finding the best possible motif in a set of sequences is a computationally hard problem!

- In the consensus model, we typically fix a motif length ℓ, a minimum instance count k, and a radius r, then seek sets of at least k instances that are all within radius r of some common consensus.
- Algorithms that do this search are often *enumerative* – they enumerate all possible consensus sequences C and check whether there is a good motif (as defined above) in the data matching each C.
- Many sneaky tricks can be used to speed up this enumerative search, but its cost is fundamentally exponential in the motif length ℓ and/or the input sequence length.
- For the WMM, we fix a motif length ℓ and seek a set of putative instances that maximize total LLR of their best WMM vs. the background (which is assumed to consist of all sequence not part of a motif instance).
- We may make any of several assumptions as to how motif instances are distributed in each input sequence – one instance per sequence, at most one instance per sequence, multiple instances per sequence, etc.
- While enumerative search can be used here as well, a more common approach is *local search*: guess an initial WMM W_0 matching a set S_0 of instances present in the input, then progressively make small changes to S_0 and W_0 so as to improve the motif's total LLR.
- There are two well-known, principled strategies for local search: *expectation maximization* and *Gibbs sampling*. They are guaranteed to find the best motif in the "neighborhood" of the original guess.

- These approaches can yield good (though not necessarily globally optimal) answers much faster than enumeration. They are the basis for well-known programs like MEME and GibbsDNA.

5 Using Homologous Sequences in Motif Finding

Conservation can sometimes provide an additional signal to make motif finding easier.

- Because a motif is a functional sequence element, it is subject to evolutionary pressures, in particular conservation against change.
- As with any functional feature, we expect that a motif instance will be more strongly conserved against mutation than the surrounding sequence (assuming the latter is not itself functional!).
- Hence, if we have several homologous sequences, each of which contains a motif instance at the same location, aligning them should cause the motif to stand out as better conserved than its surroundings.
- Using this approach to locate motifs is called *phylogenetic footprinting* – the motif leaves a "footprint" of higher-than-normal conservation in the alignment.
- An example of this approach is the EvoPrinter tool.

The footprinting approach is subject to several challenges.

- First, the input sequences must be globally alignable with high confidence; otherwise, errors due to incorrect placement of indels could result in instances of the motif not being aligned at all!
- Second, the motif should appear at the same place in each of the sequences being aligned. Recall that motifs can arise convergently; they can also "disappear" from one place in a promoter or enhancer region and arise independently elsewhere in the same region.
- Hence, it's important to align homologous sequences that diverged on a time scale shorter than that of motif instance migration.
- Third, it must be possible to distinguish the enhanced conservation of the motif from the background. But the null model for *homologous* background sequences is different from that for unrelated sequences.

- Hence, we are looking for "somewhat more" (motif) vs. "somewhat less" (background) conserved regions, which means the signal-to-noise ratio is reduced.

Assuming we can overcome these issues, how do we evaluate a putative motif found through footprinting?

- Measures of interest are still useful, but they use a weak null hypothesis (unrelated background sequences) that is not right for footprinting.
- Alternatively, estimate a "null" level of conservation from the background and show that the motif is better conserved than that.
- The sophistication of this approach can range from low (estimate the typical information content of each alignment column in the background) to high (use a known phylogenetic tree on the sequences to estimate relative rates of evolution at motif vs. background positions).

At least one tool (Magma) combines traditional motif finding with phylogenetic footprinting by searching for motifs that appear in the promoter regions of multiple genes, such that each instance is conserved in the orthologous sequences of those regions from multiple species.

6 Practical Computational Motif Finding

- Some tools offer p-values or E-values for putative motifs; others do not. For each tool, know what null model is being tested by the p-value calculation. Is it strong enough?
- Motif finding works best when the search is restricted to a small region, on the order of a few thousand bases or less, and can survey a large number of motif-containing sequences (5-20 would be nice).
- As the region gets larger and the number of instances smaller, the desired signal may be drowned out by noise. Algorithms will not find motifs, and those that are found will receive poor p-values.
- You can use footprinting to identify regions of high conservation that are larger than a single motif (say, tens to hundreds of bases), then search just these regions using a traditional motif finder for unrelated sequences. This approach has much higher power than searching long sequences without first filtering for conservation.