Jinkuk Choi 6/4/2004

BIOCHEM218 Final Project

A Critical Review of Computational Approaches on Finding Transcriptional Factor Binding Sites

Introduction

Many algorithms to identify binding sites for a regulatory protein in non-coding regions of DNA have been developed ever since the computational approaches became available to biology. Recently, adoption of computational methods to this area has grown to be more demanding because of the birth of high-throughput transcriptional regulation assay, called microarray technique. With the power of gene clustering data from microarrays, now we can investigate transcriptional regulatory networks in the context of the whole genome, and finding of transcriptional factor binding site is naturally thought as the next step that will further our knowledge a far.

The main challenge of this problem always has been the development of effective algorithm that can treat all the intrinsic complexities associated with the nature of binding sites. The difficulty mainly arises from the nature of the binding site. The base sequence of most binding sites is often called a regulatory motif, and as we can easily infer from this name, it is a tiny, highly variable sequence. What makes it much worse to identify is that they usually have gaps and repeats in them. Their insufficient informational content still matters even if we can facilitate gene clustering data.

Knowing these difficulties, many experts have developed various techniques that implement

advanced statistics. In this paper, we will survey through the characteristics of major algorithms that are developed and currently used. We also will track their improvements, and point out their strengths and weaknesses mainly by looking their assumptions. Finally, we will discuss the future direction of the development of new algorithm.

Expectation Maximization Algorithm: an old but faithful probabilistic motif model

Expectation Maximization (EM) algorithm was proposed by Lawrence and Reilly in 1990[1]. It assumes that the binding sites for a regulatory protein have a constant length *k*. We can think this assumption as a critical limitation of this one of the earliest, but still being used, old, model. However, when we think of protein and DNA interaction as the usual lock and key argument, it is the rule because all binding domains should fit into the same physical portion of the regulatory protein. This model treats a motif as a matrix of size $4\times k$ of probability values, each entry giving the probability of the letters A, C, G, T occurring in that column position. This is called a probabilistic model of a motif, and the format of this motif model that is actually used in calculation is called Position Weight Matrix (PWM). Motifs are then contrasted to "background" sequence, where letters are chosen independently from a common distribution. In a typical data set, each observed upstream sequence is assumed to have a single instance of the motif, but its exact location is unknown. EM algorithm use this missing data and it iteratively maximizes the expected *log likelihood* over the conditional distribution of the missing data, given the observed data and current estimates of parameters θ and λ (explained below).

In 1993, EM algorithm has been extended for fitting finite mixture models by Bailey and Elkan, and the new algorithm was named MM[2]. MM is the version of EM that are implemented to MEME (http://meme.sdsc.edu/meme/website/intro.html).

MM is basically same as EM except for the improvement that it relaxes the assumption that each sequence in the dataset contains one occurrence of the motif. In a rough sketchy, MM algorithm follows these three steps to find likeliest motif and background models and the best classification of words. (1) It initializes parameters θ and λ. θ is a set of motif and background, and λ is a priory probability of occurrence of a motif. They try different values of λ. (2) They repeat Expectation step and Maximization step until change in θ = (M,B), λ falls below ε (user-set threshold). (3) Report results for several good λ s. It means that it will report whenever you find a good motif and background, you report the motif.

EM is a gradient descent method. Therefore, EM has a weakness that it cannot escape local optimum. However, it also means that it always converges in a predictable, relatively small number of iterations. By contrast, original Gibbs sampling algorithms combine gradient search steps with random jumps in the search space, so they can spend an unpredictable number of iterations before converging.

There are several apparent limitations of EM, one is that all motifs found are length of *k*, according to initial assumption. The second is that we cannot know the number of different motifs present in a dataset. Also, EM does not allow gaps and it does not estimate the statistical significance of their alignments. Therefore, EM is an old model with a certain limitation. However, many researchers are still using it because of its ample ability to identify new motifs.

Gibbs Sampling Algorithm: the Bayesian version of probabilistic model

In 1993, Lawrence et al. published a Bayesian version of the motif model which applies Gibbs sampling to estimate parameters and motif locations [3]. Gibbs sampling can stochastically find out motifs from a set of sequences which are believed to contain the motif.

Gibbs sampling approach is essentially a special numerical approximation method, Markov Chain Monte Carlo (MCMC), which enables one to draw samples of high-dimensional random variables in an iterative fashion. It first removes one sequence from the set by randomly fashion, and it creates an initial temporary motif by randomly aligning the remaining sequences. It also is a probabilistic model (like EM), so it computes the probabilities of base composition for each position and makes a matrix. The windows of sequences are then moved back and forth and probability matrix is calculated at each time. The optimal alignment is found when the ratio of the motif probability to the background probability reaches maximum. Then the removed sequence is put back into the set. The start position of the window of returning sequence is estimated by scoring each segment of the sequence against the optimal matrix. Each segment is then assigned with its weight using the motif. To bias the selection, the start position is picked at random using their weights. It is repeated, and the iteration will end when the residue frequencies in the columns of matrix no longer changes.

Gibbs sampling is a very effective method for detecting weak and complex signals in a set of sequences, so people developed several web-based motif finding tools utilizing Gibbs Sampling algorithm. One of which is AlignACE (**Align**s Nucleic **A**cid **C**onserved **E**lements), developed by the Church group[4]. The Brutlag group also developed Gibbs Sampling based regulatory motif finding software, BioProspector[5].

BioProspector uses an improved version of Gibbs sampling algorithm. First improvement is that it relaxes the assumption that every sequence contains the motif. Second improvement is that it can now handle multiple copies of the same motif within the sequence. Their improvement is that BioProspector can search for a two-block motif, a subset of gapped motif, where the motif is separated into two parts with a small gap in-between them. BioProspector requires to have the correct background model because it generates a 3rd order hidden markov

model based on the given background model. This new aspect of BioProspector is biologically relevant since we now know that different genomes have different base compositions, and also different patterns of repeats. Therefore, the background data should be obtained from the same genomic sequences. Background data should also be from different promoter regions than query sequences. BioProspector can get stuck in a local maximum like EM, so multiple run is required to achieve the true optimal alignment. BioProspector is relatively slow, so we cannot us it for the whole genomes-wide application.

Gibbs Sampling algorithm is one of the pattern sampling method like EM, which is always probabilistic. The characteristic of pattern sampling method is that it is usually fast, but approximatively. Therefore, it does not always give the best possible solution. However, the general goal of pattern discovery is to find a local alignment of width *x* of sites that maximizes information content in reasonable time. Therefore, regulatory motif finding is usually done by Gibbs sampling or EM (expectation maximization) methods. Also, there are several more strong points in probabilistic method that transcriptional factor binding sites are not words, computational efficiency is the best, and can be intentionally influenced by biological data or existing knowledge.

Word based Algorithm : Moby Dick

Bussemaker et al.[6] first proposed a word based approach to motif recognition. In their model, DNA sequence data is viewed as a concatenation of different words, each word randomly selected from a dictionary with specified probabilities. Their algorithm estimates the probabilities of all of the words in a fixed dictionary and sequentially builds a dictionary from data. They named it as Moby Dick after the name of English novel they used to test their

algorithm. One drawback of this model is its assumption that each word has a unique spelling, because it practically doesn't allow misspelling of words, particularly with a 4 letter alphabet. This approach is totally different from previously described approaches in a sense that it is more likely to be an exhaustive way of motif finding rather than probabilistic. Therefore, Moby Dick can effectively find important motifs but it is so slow that its application is severely limited.

Combination of Word-based and Pattern-sampling Algorithm: MDscan

The Brutlag group developed another motif finder that utilizes both word-based algorithm, and patter-sampling algorithm, named MDscan[7]. MDscan facilitates a deterministic algorithm, so it always converges to the same result for a given set of data. MDscan has been developed for the study of clustered genes obtained by microarray analysis. Therefore, MDscan works best when sequences can be separated into two groups, that one group of sequences is believed to contain the motif while the other group of sequences does not contain the motif. MDscan starts with a word-enumeration to look for w-mers in the top sequences. It then enumerates each seed (non redundant w-mer) and searches for all w-mers from the top sequences until at least m base pairs matching the seed. For each seed the top sequences are used to form a PWM. Therefore, MDscan is different from Moby Dick and it is not a exhaustive model but a probabilistic model. The weight matrix is evaluated by a maximum posterior scoring function which uses a measure of (1) how often the matrix appears, (2) how well the matrix is conserved, (3) the probability of finding the motif by chance. The top 10-50 picks are used as a standard to scan the remaining sequences. A new w-mer is added or removed from the weight matrix if it increases the score of the matrix. This algorithm usually stabilizes in around 10 interactions and the top candidate motifs are reported.

The advantage of MDscan is that it is much faster than other methods such as BioProspector and is tractable for searching entire genomes because the search time increases only quadratically with respect to the total number of bases in the top sequences and linearly with respect to the number of bases in the remaining sequences.

Conclusion

Currently there are two major approaches for the identification of transcriptional factor binding sites. One uses pattern-sampling approach ((ex) MEME, AlignACE, BioProspector, MDscan) that only uses probabilistic approach, and another is word-based approach ((ex) Moby Dick) which uses either exhaustive or probabilistic approach. In this review, we examined and evaluated each tools with a computational approach.

Finally, there are some considerations we should bear in mind when we think of the computational approach to tackle this biological problem. One is a futility theorem, which means we still don't have any good methods, other than traditional molecular biology, to find out whether or not our predictions of individual transcriptional factor binding sites have any relationships to an *in vivo* function. Another is that pattern discovery methods are severely restricted by the Signal-to-Noise problem because information content of regulatory motif is severely limited by its intrinsic nature discussed in Introduction chapter. Therefore, all observed patterns must be carefully considered. Also development of motif discovery algorithms is limited by inadequate reference collections (number and quality). It has been clearly seen by the fact that in the development of MDscan, the author had to do a simulation to validate the capability of her own algorithm.

As Frith et al. pointed out in their recent paper, we cannot develop a much better motif

finding algorithm solely by computational approach because the main problem lies not on inadequacy of the alignment optimization procedure, but on the motifs' intrinsic subtlety[8]. It seems that overall balanced development of the entire field of motif discovery should be preceded before we will be able to make another breakthrough in this field.

---

[1] Lawrence, C.E., and Reilly, A.A., An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences, *Proteins* **7**, 41-51 (1990).

[2] Bailey, T.L., and Elkan, C., Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Sec. Intl. Conf. Int. Sys. Mol. Biol.* 28-36 (1994).

[3] Lawrence, C.E., Altschul, S.F., Bogouski, M.S., Liu, J.S., Neuwald, A.F., and Wooten, J.C., Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment, *Science* **262**, 208-214 (1993).

[4] Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M., Computational identification of Cis-regulatory elements associated with groups of fuctionally related genes in *Saccharomyces cerevisiae*, *J. Mol. Biol.* **296**, 1205-1214 (2000)

[5] Liu, X., Brutlag, D.L., and Liu, J.S., Bioprospector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Proc Pac Symp Biocomput* ,127-38 (2001).

[6] Bussemaker, H.J., Li, H., and Siggia, E.D., Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *PNAS* **97**, 10096-10100 (2000).

[7] Liu, X.S., Brutlag, D.L., and Liu, J.S., An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotech.* **20**, 835-839 (2002).

[8] Frith, M.C., Hansen, U., Spouge, J.L., and Weng, Z. Finding functional sequence elements by multiple local alignment. *Nuc. Acids Res.* **32**(1), 189-200 (2004).