Modeling Transcription Factor Binding Sites with Gibbs Sampling and Minimum Description Length Encoding

Jonathan Schug and **G. Christian Overton**
510 Blockley Hall
418 Guardian Drive
Philadelphia PA 19104
{jschug,coverton}@cbil.humgen.upenn.edu

Abstract

Transcription factors, proteins required for the regulation of gene expression, recognize and bind short stretches of DNA on the order of 4 to 10 bases in length. In general, each factor recognizes a family of "similar" sequences rather than a single unique sequence. Ultimately, the transcriptional state of a gene is determined by the cooperative interaction of several bound factors. We have developed a method using Gibbs Sampling and the Minimum Description Length principle for automatically and reliably creating weight matrix models of binding sites from a database (TRANSFAC) of known binding site sequences. Determining the relationship between sequence and binding affinity for a particular factor is an important first step in predicting whether a given uncharacterized sequence is part of a promoter site or other control region. Here we describe the foundation for the methods we will use to develop weight matrix models for transcription factor binding sites.

Introduction

The binding of transcription factors to promoter and enhancer regions plays a major part in controlling gene expression. Transcription factors typically recognize a family of short DNA sequences (typically 4 - 10 base pairs) The binding affinity is determined largely by the DNA sequence. Determining what sequences a factor binds to directly and what effects the context of the sequence may have is an important first step in predicting whether a given uncharacterized sequence is a part of a promoter site or other control region. There are several databases and/or search tools of gene expression-related models such as **TRANSFAC** (Wingender *et al.* 1996) and **Promoter Scan** (Prestridge 1995). We offer a WWW-based service, **TESS**, at **http://agave.humgen.upenn.edu/utess/-tess** which uses **TRANSFAC** to locate factor binding sites.

Clearly the importance of this information is recognized, though its use is still in its infancy. We intend to build weight matrix models for all factors in **TRANSFAC**. These models will serve as a basis for two subsequent activities, the analysis of the **TRANSFAC** data and the building of higher level models of dimers, cooperative/competitive groups, and promoter sites for various classes of genes. We expect that the models of promoter sites will help in both predicting expression patterns of newly sequenced genomes and in understanding the mechanisms and conditions of gene expression.

Models used to describe biopolymer sequences fall into five categories: 1) consensus sequence, 2) weight matrices, 3) hidden Markov models (HMMs), 4) neural nets, and 5) grammars. Consensus sequences though common and useful as mnemonics though do not contain enough information to accurately model the binding process. Weight matrices offer the intuitive simplicity of consensus sequences but with more accurate modeling of base distributions at each position. Weight matrices have the potential problem of assuming that the positions are independent. HMMs offer a means of including inter-position dependencies and of defining richer models in other senses. The first three classes form a strict hierarchy of expressive power. Neural nets offer similar flexibility in describing relations between distant sites. Grammars, augmented with probability information for rule use, provide a very rich, yet intuitive and hierarchical means of describing a model.

We have applied HMMs to modeling transcription factors binding sites (Raman & Overton 1994) and have found that individual binding sites have not yet demanded expressive power beyond that of weight matrices. We have chosen a combination of Gibbs sampling and minimum description length methods to build our binding site models. Fickett has performed similar work in (Fickett 1996). These methods work well and should be extendible to cover the next few levels of models. Indeed. (Grate *et al.* 1994) uses Gibbs sampling to determine parameters for stochastic context-free grammars.

Methods

Weight Matrix Models

Our model for a factor site is a block of contiguous positions, each with its own base distribution, i.e., a weight matrix. Given a multiple alignment (without gaps) of a sequence set, the array $C_A(c,b)$ is the number of occurrences of each base b in column c of the alignment A and $C_A(c) = \sum_b C_A(c,b)$ is the number of sequences that contribute to the data at a given position. Since not all sequence are of the same length and the binding site is not in the same position for all sequences, $C_A(c)$ varies with c. A range $[c_m, c_{m+W-1}]$ of positions as defined to be the conserved binding or binding-related region. Other positions outside this range are considered to be generated from a single background distribution. The weight matrix model $P_M(c,b)$ is then defined as shown in equation (1)

$$P_M(c,b) = \begin{cases} \frac{C_A(c+c_m,b)+p(b)}{C_A(c+c_m)+1} & \text{for } 1 \leq c \leq W \\ \sum_{c \notin [c_m, c_{m+W-1}]} \frac{C_A(c,b)+p(b)}{C_A(c)+1} & \text{for } c = 0 \end{cases} \tag{1}$$

where $p(b)$ is the prior probability for base b. This formula results from using a Dirichlet prior on the background base distribution. Matrix $C_M(c,b)$ and array $C_M(c)$ are defined similarly. Thus the parameters of the model are the width of the motif and the base distributions at each column and the background. The parameters of the alignment are the sense and start of each sequence relative to the motif block.

Alignment Using Gibbs Sampling

We use a Gibbs sampling technique to perform the multiple alignment. The algorithm consists of two phases, sampling and clipping, which are repeated until completion. *Sampling* is the process of realigning a single sequence by selecting a new alignment for it based on the likelihood of each possible alignment being correct given the current alignment of the other sequences. A score is computed for each possible alignment of the target sequence which is a function of the sequence, the current alignment, and temperature parameter. One of these alignments is selected nondeterministically according to the probability distribution achieved by normalizing the likelihood scores. We repeat this process for all sequences and consider this to be a *step*. Our algorithm is derived from extensive work by Lawrence and others (Lawrence *et al.* 1993; Neuwald, Liu, & Lawrence 1995; Lawrence *et al.* 1994; Lawrence & Reilly 1996; Liu, Neuwald, & Lawrence 1995; Liu 1994), but tailored to our application. *Clipping* is the process of determining which block of contiguous columns of the alignment shows significant conservation and using this block as the new motif. These steps are alternated until an annealing schedule similar to one in (Geman & Geman 1984) for the temperature parameter has been completed.

The alignment of the sequences takes place in a box which has a width $W_{max} = 2 * L_{max} - 1$ where L_{max} is the length of the longest sequence in the data set. When a sequence is placed in an alignment, it is required to overlap the central position, or to have an end point within two bases of the center. This forces all the sequences to overlap. Were they not forced to overlap or were there no penalty for non overlap, the sequences would be free to spread out so that there is no overlap and therefore no mismatch. We extend this requirement during sampling to penalize alignments that do not overlap the conserved region to a large enough degree. The alignment can be initialized in one of two ways, random and short-to-long sequential alignment. A *random initialization* consists of pseudo-random choices for the start and sense of each sequence, with the restriction that the string must overlap the center of the motif. These are made using a uniform distribution over all legal alignments. In a *short-to-long initialization* we perform a sequential alignment on the sequences ordered by length from shortest-to-longest. The weight matrix is then computed from the initial alignment.

The likelihood score of the data given the alignment is computed using equation (2).

$$P(x|A) = \prod_{c=1}^{W} \prod_{b \in B} P_M(c,b)^{C_M(c,b)} \tag{2}$$

Positions of the sequence which are in background columns are not considered. A multiplicative penalty may be applied in addition if the sequence covers less than half or a third of the motif and it is long enough to cover it all. This weights against alignments and motifs which are one base long.

The likelihood score for an alignment is modified by a temperature parameter: $P_T\{x|A\} = P\{x|A\}^{1/T}$, where we usually have $0 < T \leq 1$. At lower temperatures the selection process closely approximates the deterministic algorithm that always chooses the best alignment. At $T = 1$ we select according to the likelihood score. As $T \rightarrow \infty$ the selection distribution becomes uniform. As mentioned above this is controlled by an annealing schedule as indicated in (Geman & Geman 1984) of the form $T = \frac{C}{\log(1+t)}$ where t is the number of steps performed.

We set a limit on the number of consecutive sampling iterations that will be performed without improvement in the total alignment score, typically 10 to 20. Once this limit is reached, clipping takes place. In preliminary tests, this method is efficient and effective. Clipping is the process where by the best contiguous block of the alignment is chosen to be the motif. It is called clipping, but in fact it is possible for the motif to get wider at a clipping step if the well conserved region extends beyond the current motif boundaries. We have tried several techniques but settled on MDL. The appropriate technique we feel should be either parameterless or have parameters which can be set in a principled manner and have a reasonable physical interpretation.

Model Selection Using the Minimum Description Length Principle

The minimum description length principle holds that the best model is the one which minimizes the total description costs of the data as well as the model. Rissanen, in (Rissanen 1983), derives equation (3) to measure the total description length of a set of data and its model.

$$L(x,\theta) = -\ln P(x|\theta) + \ln^*(C(k)[(\theta', M(\theta)\theta)]^{k/2}) \quad (3)$$

Here k is the number of model parameters $((W+1)*3$ in our case), θ is a column vector of the parameters, $M(\theta)$ is the second partials matrix of $-\ln P(x|\theta)$ with respect to θ, $C(k)$ is the volume of the unit k-sphere, and $\ln^*(x) = \ln(x) + \ln(\ln(x)) + \ldots$ where only the positive terms are used.

The first term in equation (3) corresponds to the description length of the data, the second term is the description length of the model. Minimizing the first term alone corresponds to choosing the model that best explains the data. In our case, the best model is one that has width W_{max}.

We use the base probabilities for each column and the background distribution as the model parameters to be transmitted. We leave out the motif start indices since we make no prior judgment as to their value. The motif width W may be included in the model using either Rissanen's universal prior for integers or by a more biologically motivated prior. In either case this is an additive term in equation (3).

A significant contribution of (Rissanen 1983) is determining the precision required for specifying real-valued parameters. Without limiting the precision of a real-valued parameter an infinite amount of information would be required to specify it. Selecting the precision of the parameters is important since it directly influences the model cost. If the model cost is too high then given a number of observed bases, it will be cheaper to transmit the bases using a single distribution for all columns. If the model cost is too low, then there is no pressure to choose a small model and W will expand to W_{max}. (Rissanen 1983) makes a choice for the precision based on first principles which we adopt. The precision for each parameter is different depending on the effect the parameter has on the value of $-\ln P(x|\theta)$.

The log-likelihood of the data given the model is expressed as shown in equation (4).

$$-\ln P_M(x|\theta) = \sum_{c=0}^{W}\sum_{b\in B} C_M(c,b)P_M(c,b) \qquad (4)$$

(Rissanen 1983) requires that the parameters' range be $(-\infty,\infty)$ and that there be no redundant parameters. The four parameters of a DNA base distribution do not meet either of these criteria; their ranges are $[0,1]$ and there are only 3 independent parameters. Hence we first re-parameterize the model as

$$P(c,b) = \frac{e^{q_{cb}}}{Z_c} \qquad (5)$$

where $Z_c = \sum_{b\in B} e^{q_{cb}}$ to meet the range requirement. To remove the redundant parameter we set $q_{cb_1} = 0.0$ and do not consider q_{cb_1} a parameter of the model in the MDL context. It does not matter which base is chosen to be removed from the parameterization; the results are the same for all choices. One consequence of equation (1) is that $\forall c,b\ 0 < P_M(c,b) < 1$, hence the q_{cb} are not required to take on values of $-\infty$ or ∞.

Using the re-parameterization as shown in equation (6), the partials of $-\ln P(x|\theta)$ are shown in equation 7.

$$-\ln P(x|\theta) = -\sum_{i=0}^{W}\left[\sum_{b\in B} C_M(i,b)q_{ib} - C_M(i)\ln Z_i\right] \qquad (6)$$

$$\frac{\partial^2 - \ln P(x|\theta)}{\partial q_{ic}\partial q_{jd}} = \begin{cases} 0 & \text{if } i \neq j \\ -\frac{e^{q_{ic}}e^{q_{id}}}{Z_i^2}C_M(i) & \text{if } c \neq d \\ \left[\frac{e^{q_{ic}}}{Z_i} - \left(\frac{e^{q_{ic}}}{Z_i}\right)^2\right]C_M(i) & \text{else} \end{cases} \qquad (7)$$

To select the most conserved region, we let the width and start position vary over all legal combinations and choose the combination with the shortest total description length.

Choosing the Final Model

Since the Gibbs sampler is a stochastic technique and we are performing a finite number of iterations, we repeat the annealing schedule some number of times with different initial configurations. Typically, for a large data set, we would find about 15 different solutions for 50 runs. Our heuristic is to choose the model that appears most often, considering a model and its reverse complement to be the same.

Discussion

Equation (2) applies to sequences which do not contain any ambiguous base characters. When ambiguous bases are allowed $P(x|\theta)$ is computed somewhat differently and the expressions for the partials of $-\ln P(x|\theta)$ are slightly more complex but reduce to those given in equation (7) when the input data sequences have no ambiguous bases.

We will probably switch to a Bernoulli sampler (Liu, Neuwald, & Lawrence 1995) for alignments which will allow multiple instances of a motif in a single sequence since this is common in TRANSFAC. This entails a change in the alignment parameter structure, but not in the model, since we will still use a block model. A block model is likely to be appropriate for most factors, however it can not model dimer binding where the conserved motif consists of two block motifs separated by a variable distance. In this case we can identify the parts of the binding site, possibly using prior information, and then model the spacing between the two sites in a second-level model.

In work by Lawrence, (Lawrence *et al.* 1993), there was a broad peak in the average information content per column around the natural motif width. The width which achieved the maximum value was then selected to be the criterion for choosing the motif width. For transcription factors, there does not appear to be such a peak; the shorter the motif, the higher the information content. Typically there is a very short region, of 3 of 4 bases which gives a very high score. As the motif gets wider, the average information content per column drops steadily but not precipitously.

Acknowledgements

This work was supported by NIH grants R01-RR04026-07A2 and 1R01-HG01539-01. The authors wish to thank Chip Lawrence for many enlightening conversations and suggestions and the ISMB-97 referees for their helpful comments.[1]

References

Fickett, J. 1996. Quantitative Discrimination of MEF2 Sites. *Molecular and Cellular Biology* 16(1):437–441.

Geman, S., and Geman, D. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6(6):721–741.

Grate, L.; Herbster, M.; Hughey, R.; Main, I.; Noller, H.; and Haussler, D. 1994. RNA Modelling Using Gibbs Sampling and Stochastic Context Free Grammars. *ISMB94* 235:1501–1531.

Lawrence, C., and Reilly, A. 1996. Likelihood Inference for Permuted Data With Application to Gene Regulation. *JASA* 91(133):76–85.

Lawrence, C.; Altschul, S.; Boguski, M.; Liu, J.; Neuwald, A.; and Wootton, J. 1993. Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Sequence Alignment. *Science* 262:208–214.

Lawrence, C.; Altschul, S.; Wootton, J.; Boguski, M.; Neuwald, A.; and Liu, J. 1994. A Gibbs Sampler for the Detection of Subtle Motifs in Multiple Sequences. *Proceeding of the 27th Hawaii International Conference on System Sciences.*

Liu, J.; Neuwald, A.; and Lawrence, C. 1995. Bayesian Models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies. *JASA* 90(432):1156–1170.

Liu, J. 1994. The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem. *JASA* 89(427):958–966.

Neuwald, A.; Liu, J.; and Lawrence, C. 1995. Gibbs motif sampling: Detection of bacterial outer membrane protein repeats. *Protein Science* 4:1618–1632.

Prestridge, D. 1995. Prediction of Pol II Promotor Sequencesusing Transcription Factor Binding Sites. *JMolBio* 249:923–932.

Raman, R., and Overton, G. C. 1994. Application of Hidden Markov Modelling to the Characterization of Transcription Factor Binding Sites. *Proceedings of the 27th Annual Hawaii International Confererence on System Sciences.*

Rissanen, J. 1983. A Universal Prior for Integers and Estimation by Minimum Description Length. *The Annals of Statistics* 11(2):416 – 431.

Wingender, E.; Dietze, P.; Karas, H.; and Knüppel, R. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. *NAR* 24(1):238–241.