Bayesian Models for Gene Prediction

Sudeshna Adak
IBM India Research Lab, Bioinformatics Group
Block 1, Indian Institute of Technology
New Delhi, India
asudeshn@in.ibm.com

1. Introduction

The problem of gene identification from DNA sequences by computational methods has received wide attention recently due to ongoing debates regarding the number of genes present in the human genome. The problems of current in-silico gene prediction methods is however not restricted to human DNA, but applies equally to genomes of all organisms.

Current methods of gene-prediction are either *ab initio* methods that use deterministic/probabilistic models for DNA to determine gene locations or homology-based methods that rely on similarity to known genes to identify genes in new DNA sequences. *Ab Initio* methods which use probability models for DNA to capture the complexity of gene structures have been found to perform well. Examples of well-performing *ab initio* methods include Genscan (Burge and Karlin 1997) which has over 90% sensitivity and 90% specificity for higher eukaryotes and Glimmer (Salzberg et. al. 1998) which has over 97% sensitivity and around 80% specificity for microbial genomes. However, the performance of *ab intio methods* is seen to decline in extremely long sequences of DNA and homology based methods like GeneID+ (Guigó et. al. 1993) outperform such *ab initio* methods. As whole genome sequencing projects are producing extremely long contigs (contiguous peices of DNA), which are subsequently used for gene prediction, the need for reliable methods for gene prediction in whole genomes is required. In this manuscript, a new Bayesian method for gene prediction in whole genomes is proposed which provides a framework for combining the probability models of *ab initio* techniques and homology information.

There are 59 complete and annotated genomes that have been published and there are 338 ongoing whole genome sequencing projects that are ongoing (205 in prokaryotes are 133 in Eukaryotes). This information is based on the latest update (April 8, 2001) of GOLD^{TM}, the Genome On Line Database at http://wit.integratedgenomics.com/GOLD/, which is a world wide web resource for comprehensive access to information regarding complete and ongoing genome projects around the world. As more and more complete genomes are becoming available due to improvement in large-scale sequencing technologies, the need for reliable high-throughput gene prediction methods is also increasing. Moreover a method that make use of homology information from the growing number of completed genomes is clearly of practical significance.

2. Method

Definition 2..1 *A DNA sequence S is a sequence of characters $s_1, s_2, \cdots, s_n$, where each $s_i \in \{A, C, T, G\}$. A whole genome is the entire DNA sequence present in the cell of an organism.*

Definition 2..2 *For a DNA sequence, an annotation is a vector $\Phi[S] = (m, T_1, T_2, \ldots, T_m, \phi_0, \phi_1, \phi_2, \ldots, \phi_m,)$, where ϕ_i $(i = 0, 1, 2, \ldots, m)$ is a descriptive label of the function of the segment $[T_i + 1, T_{i+1}]$ of the DNA sequence S, and $T_0 = 1 \leq T_1 \leq \ldots \leq T_m \leq T_{m+1} = n$ partitions the DNA sequence into functionally homogenous segments. A complete annotated genome is the pair $(S, \Phi[S])$, where S represents the DNA sequence of the whole genome and $\Phi[S]$ is the corresponding annotation.*

Example 1: In the problem of identifying which nucleotides code for proteins, the possible annotation labels are: $C_1 =$ coding region; $C_2 =$ non-coding region. Note that in this case, it is sufficient to specify ϕ_0 as either coding and non-coding and $\phi_1, \ldots, \phi_m$ are determinsitic once ϕ_0 is known.
Example 2: A slightly more complex problem is of identifying coding regions as well as the direction of transcription (i.e. if the particular segment of the sequence is coding in the forward or reverse direction). In that case, the possible annotation labels are: $C_1 =$ coding region (forward strand); $C_2 =$ coding region (reverse strand); $C_3 =$ non-coding region.

Larger lists of annotation labels are required to capture the complexity of gene structure in higher eukaryotes. For example, a list of 27 labels was used in Genscan to describe the gene structure of human DNA.

Problem: Given a query DNA sequence S, and the complete annotated genome Γ of a closely related species, we wish to predict the annotation $\Phi[S]$ of S, by maximizing $P(\Phi|S, \Gamma)$.

Then, using Bayes' formula gives us

$$P(\Phi|S, \Gamma) = \frac{P(S|\Phi, \Gamma).P(\Phi|\Gamma)}{\sum_{\Phi'} P(S|\Phi', \Gamma).P(\Phi'|\Gamma)} \tag{1}$$

Conditional on an annotation Φ, the probability distribution of S can be assumed to be independent of the annotated genome Γ. This assumption is clearly a restatement of the basic premise of *ab initio* methods like GenScan and Glimmer that the probability model for a DNA segment depends only on its functional role. This allows us to substitute $P(S|\Phi)$ for $P(S|\Phi, \Gamma)$ in Equation 1.

- Likelihood Specification: In the following algorithm, the "likelihood" specification used is:

$$\begin{aligned} P(S|\Phi) &= P(S|m, T_1, T_2, \ldots, T_m, \phi_0, \phi_1, \phi_2, \ldots, \phi_m) \\ &= \prod_{i=1}^{m} P^{\phi_i}(S_{[T_i+1, T_{i+1}]}) \end{aligned} \tag{2}$$

Thus, the probability distribution is identical for segments of DNA that are in the same functional category. In the case of example 1 above, this would mean that all coding regions have the same probability distribution. Markov models for DNA have been used traditionally, and we can furthermore assume that $P^{\phi}(\cdot)$ to be markov of order p. This leads to "phased" markov models as are used in GenMark (Borodovsky et. al. 1993). The parameters of the markov model are ascertained using maximum likelihood rather than estimating them through a Bayesian mechansim

- Prior Specification: Specifying a prior using the information from a closely related species is based on the same premise as homology based methods. The prior is specified, using Γ, as follows:

$$\begin{aligned} & P(m, T_1, T_2, \ldots, T_m, \phi_0, \phi_1, \phi_2, \ldots, \phi_m | \Gamma) \qquad (3) \\ = \; & P(m) \left\{ P(\phi_0) \prod_{i=1}^{m} P(\phi_j | \phi_{j-1}, \Gamma) \right\} P(T_1, T_2, \ldots, T_m | m, \phi_0, \phi_1, \phi_2, \ldots, \phi_m), \text{ where} \end{aligned}$$

  - The prior probability distribution for $\phi_0, \phi_1, \phi_2, \ldots, \phi_m$ is a markov chain where the transition probabilities $P(\phi_j | \phi_{j-1}, \Gamma)$ are estimated from the observed transitions in the genome Γ.
  - The prior for $T_1, T_2, \ldots, T_m$ is specified in terms of a prior distribution for $W_0, W_1, \ldots, W_m$, where $T_j = T_{j-1} + W_{j-1}$. The prior, $P(W_0, W_1, W_2, \ldots, W_m | \phi_0, \phi_1, \phi_2, \ldots, \phi_m, m, \Gamma)$ is multinomial$(n, \pi_0, \pi_1, \pi_2, \ldots, \pi_m)$. The hyper-prior parameter π_j is chosen depending on state ϕ_j and the annotated genome Γ.
  - The prior for ϕ_0 is taken to be uniform over the set of all annotation labels and the prior for m will be assumed uniform on {1,2,...,M}, where M is a large number and a possibly upper bound on the number of transitions. This number M will be chosen depending on the length of the sequence S.

Whole Genome Gene Prediction Algorithm: As M can be very large for whole genomes, we use a windowing approach.

1. A sliding window (of length L and shift parameter K) is used to first divide the whole genome into segments.
2. For the DNA sequence of each segment (call it S), we estimate the posterior mode as follows: For each value of $m = 1, 2, \ldots, M$, we determine the posterior mode of $P(T_1, T_2, \ldots, T_m, \phi_0, \phi_1, \phi_2, \ldots, \phi_m | S.\Gamma)$ and then choose the value of m to maximize the posterior modes.
3. Consecutive windows are then combined for a prediction for the whole genome. In case of conflict in overlapping regions, the prediction used is the one with the higher posterior probability.

3. Results and Discussion

We use the complete annotated genome of Escheria coli K-12 as deposited in Genbank (sequence id U00096) to predict genes in the whole genome of Haemophilus infuenzae Rd (sequence id L42023). They are reasonably closely related species as they are both bacteria and of the proteobacteria, gamma subdivision kind. They occur next to each other in the taxonomy classification of the complete, annotated genomes available today.

While H. infuenzae is a complete, annotated genome, it was used in order to verify the results of the algorithm. We used the above algorithm with $L = 1000$ and $K = 500$ and for these choices of L, K, we used $M = 10$. We used the annotation label of coding versus non-coding and psterior distributions were estimated via Gibbs Sampling. Preliminary results showed prediction accuracy of 93%.

Extensions of the algorithm to incorporate (1) homology information and (2) multiple genomes is discussed in a separate manuscript (Adak, 2001). Furthermore, speed up of the algorithm for convergence needs to be explored.

REFERENCES

Adak, S. (2001). Bayesian Models for Gene Prediction. *IBM Technical Report*, In preparation.

Borodovsky, M. and McIninch, J. (1993). GenMark: Parallel gene recognition for both DNA strands. *Computers and Chemistry*, **17**, 123-133.

Burge, C. And Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*,**268**, 78-94.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, **85**, 398-409.

Guigó, R. Knudsen, S. Drake, N. and Smith, T. (1992). Prediction of gene structure. *Journal of Molecular Biology*,**225**, 141-157.

Salzberg, S.L., Delcher, A.L., Kasif, S., and White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, **26(2)**, 544-548.

RESUME

Dr. Sudeshna Adak leads the Bioinformatics group at IBM India Research Lab. She has a Ph.D. from Stanford and worked in Dana-Farber Cancer Institute and Harvard School of Public Health before joining IBM India Research Lab. Her interests include microarray gene expression informatics, computational algorithms for prediction of gene and protein structure, and computational methods for protein-protein interactions.