<<生物信息学>>

图书基本信息

书名：<<生物信息学>>

13位ISBN编号：9787030176400

10位ISBN编号：7030176405

出版时间：2006-10

出版时间：科学出版社

作者：[美] D.W.芒特

页数：582

译者：曹志伟

<<生物信息学>>

内容概要

当前生物信息学研究重点是对基因组序列、蛋白质组学和数组技术所产生的大量数据的计算分析。
本书对DNA、RNA和蛋白质数据的计算提供了丰富的演算方法，并指出了在解决生物学问题中这些方法的优缺点及应用策略。
本书的第一版是在Mount博士讲稿的基础上进行整理出版的，在全球范围内用作教材。
第二版对内容进行了全面的修订，由专业教师提供导读，最大程度地适用本科生和研究生教学。
? 本书为高等院校生物信息学专业本科生和研究生提供理想的学习材料。
同时，本书也适宜科研人员、信息专家自学使用。

<<生物信息学>>

书籍目录

CHAPTER 1 历史简介和概论CHAPTER 2 Collecting and Storing Sequences in the LaboratoryCHAPTER 3 Alignment of Pairs of SequencesCHAPTER 4 Introduction to Probability and Statistical Analysis of Sequence AlignmentsCHAPTER 5 Multiple Sequence AlignmentCHAPTER 6 Sequcence Database Searching for Similar SequencesCHAPTER 7 Phylogenetic PredictionCHAPTER 8 Prediction of RNA Secondary StructureCHAPTER 9 Gene Prediction and RegulationCHAPTER 10 Protein Classification and Structure PredictionCHAPTER 11 Genome AnalysisCHAPTER 12 Bioinformatics Programming Using Perl and Perl ModulesCHAPTER 13 Analysis of MicroarraysIndex

章节摘录

The object is to adjust these parameters so that the model represents the observed variation in a group of related protein sequences. A model trained in this manner will provide a statistically probable msa of the sequences.

One problem with HMMs is that the training set has to be quite large (50 or more sequences) to pro-duce a useful model for the sequences. A difficulty in training the HMM residues is that many different parameters must be found (the amino acid distribu-tions, the number and positions of insert and delete states, and the state transition frequenaes add up to thousands of parameters) to obtain a suitable model,and the purpose of the prior and training data is to find a suitable estimate for all these parameters.When trying to make an alignment of short sequence fragments to produce a profile HMM, this problem is worsened because the amount of data for training the modelis even further reduced. Algorithms for calculation of an HMM. As illus-trated in Figure 5.16, the goal is to calculate the best HMM for a group of sequences by optimizing the transition probabilities between states and the amino aad compositions of each match state in the model.The sequences do not have to be aligned to use the method. Once a reasonable model length reflecting the expected length of the sequence alignment is cho-sen, the model is adjusted incrementally to predict the sequences. Several methods for training the modelin this fashion have been described (Baldi et al.1994; Krogh et al. 1994; Eddy et al. 1995; Eddy 1996;Hughey and Krogh 1996; Durbin et al. 1998). For example, the Baum-Welch algorithm, previously used in speech recognition methods, adjusts the parameters of HMMs for optimal matching to sequences, as discussed below. This HMM is devel-oped as follows: 1.The model is initialized with estimates of transi-tion probabilities, the probability of moving from one state to another particular state in the model(e.g., the probability of moving from one match state to the next), and the amino aad composi-tion for each match and insert state. If an initial alignment of the sequences is known, or some other kinds of data suggest which sequence posi-tions are the same, these data may be used in the model. For other cases, the initial distribution of amino acids to be used in each state is described below. The initial transition probabilities that are chosen generally favor transitions from one match state, a part of the model that represents one column in an msa, to the next match state,representing the next column. The alternative of using transitions to insert and delete states, which would delete a position or add another sequence character,is less favored because this builds more uncertainty into the HMM sequence model, 2.All possible paths through the model for generat-ing each sequence in turn are examined. There are many possible such paths for each sequence. This procedure would normally require a huge amount of time computationally. Fortunately, an algo-rithm, the forward-backward algorithm, reduces the number of computations to the number of steps in the model times the total length of the training sequences. This calculation provides a probability of the sequence, given all possible paths through the model, and, from this value, the probability of any particular path may be found. The Baum-Welch algorithm, referred to above,then counts the number of times a particular state-to-state transition is used and a particular amino acid is required by a particular match state to generate the corresponding sequence position. 3.A new version of the HMM is produced that uses the results found in step 2 to generate new transi-tion probabilities and match-insert state compo-sitions. 4.Steps 3 and 4 are repeated up to ten more times to train the model until the parameters do not change significantly. 5.The trained model is used to provide the most likely path for each sequence, as described in Figure 5.16. The algorithm used for this purpose,the Viterbi algorithm, does not have to go. Through all of the possible alignments of a given sequence to the HMM to find the most probable alignment,but instead can find the alignvment by a dynamic programming technique very much like that used for the alignment of two sequences, as discussed in Chapter 3. The collection of paths for the sequences provides an msa of the sequences with the corresponding match, insert, and delete states for each sequence. The columns in the msa are defined by the match states in the HMM such that ammo acids from a particular match state are placed in the same column. For columns that do not correspond to a match state, a gap is added. 6. The HMM may be used to search a sequence database for additional sequences that share the same sequence variation. In this case, the sum of the probabilities of all possible sequence align- ments to the model is obtained. This probability is calculated by the forward component of the forward-backward algorithm described

above in step 2. This analysis gives a type of distance score of the sequence from the model, thus providing an indication of how well a new sequence fits the model and whether the sequence may be related to the sequences used to train the model. In later derivations of HMMs, the score was divided by the length of the sequence because it was found to be length dependent.A z score giving the num-ber of standard deviations of the sequence length-corrected score from the mean length-corrected score is therefore used (Durbin et al.1998). Recall that for the Bayes block aligner, the initial or prior conditions were amino acid substitution matrices, block numbers, and alignments of the sequences. The sequences were then used as new data to examine the model by producing scores for every possible combination of prior conditions. By using Bayes' rule, these data provided posterior probability distributions for all combinations of prior informa-tion. Similarly, the prior conditions of the HMM are the initial values given to the transition values and amino aad compositions. The sequences then pro-vide new data for improving the model. Finally, the model provides a posterior probability distribution for the sequences and the maximum posterior prob- ability for each sequence represented by a particular path through the model. This path provides the alignment of the sequence in the msa;i.e.,the sequence plus matches, inserts, and deletes, as described in Figure 5.16. (Bayes' rule is discussed in Chapter 4, p. 148, along with related terms of condi-tional probability including prior and posterior probability.) ……

<<生物信息学>>

版权说明