Poster H-85

Motif Discovery with Arbitrary Insertions and Deletions

Authors:
Martin C Frith (*IMB / RIKEN*)
Timothy L Bailey (*IMB*)

Short Abstract: We present GLAM2, the first Gibbs sampling-based motif discovery algorithm that allows arbitrary indels. Our approach is novel in the way in which it models the probability of indels and optimizes the number of aligned columns. We show that GLAM2 is substantially more sensitive than other methods.

Long Abstract:

Discovering functional regions in biosequences remains a grand scientific challenge. We present GLAM2, the first motif discovery algorithm that allows arbitrary indels and uses Gibbs sampling [1]. Our approach is novel in the way in which it models the probability of indels and optimizes the number of aligned columns.

GLAM2's motif model has position-specific residue and indel probabilities. Since these probabilities are unknown, we eliminate them by integrating the posterior probability of the alignment over all possible values of these parameters. We use a Dirichlet mixture prior for residue probabilities [2], and a simple Dirichlet prior for transition probabilities.

GLAM2 uses Gibbs sampling in conjunction with stochastic traceback and column sampling to find a local alignment of the sequences that maximizes the likelihood ratio of the motif model versus a zero-order null model. Starting from an arbitrary alignment, we repeatedly remove a sequence from the alignment and re-align it using a stochastic traceback [3], which tends to increase the likelihood ratio, but may temporarily decrease it so as to avoid local optima. To avoid local optima in the placement of match columns,
we perform column sampling. This removes one column (so its residues become considered as unaligned insertions) and stochastically re-inserts it at some position. To optimize the motif width, columns may be added (by
duplication) and deleted (by merging). The probabilities of duplication and merging are carefully tuned to maintain detailed balance.

Tested on 58 sets of protein sequences containing Prosite patterns [4], GLAM2 is substantially more sensitive than other methods (SAM-T2K, HMMER 1.8.5, and PRATT 2.1 [2,3,5]). The test set is biased in favor of PRATT, which finds Prosite pattern-like regular expression motifs. Nevertheless, GLAM2 has many fewer cases where the positive predictive value (PPV) (correctly aligned residue pairs / aligned residue pairs from the algorithm) is very low compared to the other methods tested. In some alignments with high sensitivity and low PPV, GLAM2 appears to find legitimate extensions of the Prosite motif as a result of its more expressive motif model. Thus, GLAM2 offers great promise for deciphering biological sequences by discovering motifs without restrictions on indels.

[1] Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262: 208-214.
[2] Krogh A, Brown M, Mian, IS, Sjolander K, Haussler D (1994) Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 235: 1501-31.
[3] Eddy SR (1995) Multiple alignment using hidden Markov models. Proc Int Conf Intell Syst Mol Biol 3: 114-20.
[4] Hulo N, Sigrist CJA, Le Saux V, Langendijk-Genevaux PS, Bordoli L, et al. (2004) Recent improvements to the PROSITE database. Nucleic Acids Res 32: 134-137.
[5] Jonassen I, Collins JF, Higgins DG (1995) Finding flexible patterns in unaligned protein sequences. Prot Sci 4: 1587-1595.