Poster H-33

Probabilistic reasoning on occurrences of related binding site model matches

Authors:
Rainer Pudimat (*University of Freiburg*)
Rolf Backofen (*University of Freiburg*)

Short Abstract: We present a novel approach for modeling clusters of protein binding sites on nucleotide sequence (DNA, RNA). Single binding motifs are represented as PSSM-like Naive Bayes classifiers. These models are connected by probabilistic evaluating the fulfillness of logical expressions which describe the relations between models.

Long Abstract:

Background:

The prediction of binding sites for proteins on nucleotide sequences is a key task of computational biology. Since single hits of simple probabilistic sequence models tend to be false positives, recent research is concentrating on additional biological knowledge for separating real binding sites from background [4]. Especially the preference of binding sites to occur in groups (clusters) seems to have exquisite filtering properties [2,3]. Clusters of protein binding sites can be observed in promoter sequences where transcription factors bind to DNA and cooperate with each other in order to regulate gene transcription. Further clusters of binding sites are common within splice enhancers and silencers which are bound by various members of spliceosome complexes.

Among current approaches to make use of these site grouping effects there are mainly two groups: (1) parallel connections of various sequence models and special gap states in an HMM [2] and (2) Counting putative hits in a sliding sequence windows [3]. The latter approach needs to much adjustments of free parameters (i.e. windows width) which are not biologically explainable. The first approach introduces an artificial construct of binding site modules together with a combined scoring scheme. Each motif is equally likely and there is no possibility to consider ordering and orientation constraints.

Approach:

We introduce an alternative approach for modeling arbitrary relations between cooperating binding sites and other textual presented knowledge on input sequences. Such relations are formulated as logical expressions (i.e. "factor A binds iff B and C binds in close neighborhood." or "factor A binds when the gene is liver-related"). Probabilistic representations of these expressions are modeled in a specially constructed Bayesian network. This network is able to evaluate the soft, "probabilistic" satisfaction of the relations and provides an a priori distribution over all considered sequence models (e.g. PSPMs) at each sequence position. The a priori probability P(M) at one position is a superposition of the probabilistic satisfactions of all relations concerning model M. It expresses the expectation of observing a match for that model without having seen the nucleotides at the current position. On the other side each model M delivers sequence probabilities P(seq | M) for each position of the input sequence meaning how the subsequence starting there is fitting the PSSM

columns. Both parts (the a priori expectation for models M and their positional sequence score) are combined using Bayes' rule. This results in an a posteriori modeling P(M|seq) of motif occurrences.

Results:
Given an optimal match P(seq|M) to a sequence model (e.g. a high PSSM score), the a priori model will penalize it if the relational constraints for that model are not fulfilled (e.g. a co-acting factor is missing a suitable binding site). Otherwise the match is favored by the a priori model.

The effect is similar to that of [2,3] in favoring clusters of binding sites. In contrast to these approaches, we are able to model biological relations between sequence models more adequate and more flexible. Whereas [2,3] prefer clustering of sites by default, our approach will penalize a single site within a cluster when there is no relation to the other cluster members.

Literature:
[1] T.L. Bailey & W.S. Noble: Bioinformatics, 2003, 19 Suppl 2, II16-II25.

[2] M.C. Frith, M.C. Li & Z. Weng: Nucleic Acids Research, 2003, 31(13), 3666-8.

[3] R. Durbin, S. Eddy, A. Krogh & G. Mitchison: Biological sequence analysis, Cambridge University Press, 1998.

[4] R. Pudimat, E.G. Schukat-Talamazzini, R. Backofen: Bioinformatics, 2005, 21(14):3082-8