Poster H-30

Prediction of Transcription Factor Binding Sties by a Novel Computational Algorithm

Authors:
Daekwan Seo (*Virginia Commonwealth University*)
Moritoshi Yasunaga (*University of Tsukuba*)
Insook Kim (*University of Araknsas Medical Sciences*)
Jung H. Kim (*University of Arkansas*)

Short Abstract: The identification of transcription factor binding sites (TFBS) still remains a fundamental challenge in modern biology. The proposed algorithm predicts TFBS based on a hybrid Dynamic Programming combined with statistical method. We predicted experimentally verified TFBS and unknown but highly putative TFBS in each group of Dictyostelium gene sequences.

Long Abstract:
Since the rapid advances in genome sequencing projects, the identification of transcription factor binding sites (TFBS) still remains a fundamental challenge in a modern biology and genomics study. Gene expression is a complex process to produce proteins that is an inevitable process to sustain its life. Gene expression is initially regulated by trans-acting protein complexes that bind to TFBS in a proximal region of gene sequence. The prediction of these TFBS is a hard and time consuming task in a classical biology method. For this reason, we propose a novel computational approach to predict TFBS. The TFBS problem is defined as “a given set of k unaligned co-regulated sequences, a distance measure d and a threshold value t for d, a typical problem in pattern discovery is to find all patterns that occur in at least q sequences out of k within distance t from the sequences”. Since gene sequences contain lots of mono- and dimeric repeats such as TTTTTTTT and ATATATAT, it is easy to predict fake TFBS in co-regulated sequences (85% A+T rich in Dictyostelium gene sequences of intergenic region). The proposed algorithm is based on a hybrid Dynamic Programming matching (DP matching) algorithm with a statistical method. Differently with previous proposed algorithms, the proposed algorithm includes several different sets of co-regulated sequences (m) and tried to predict unique TFBS in each set of co-regulated sequences not other sets of co-regulated sequences. The proposed algorithm consists of two different steps. (1) Evaluation of candidate TFBS and (2) Selection of putative TFBS based on a proposed statistical method. At the 1st step of the proposed algorithm, the method generates every possible candidate TFBS for a given length of TFBS and figures out their evaluation score (E-score) based on a hybrid DP matching algorithm. The evaluation score of a candidate TFBS contains a distance measure between proximal regions of set of co-regulated genes and a candidate TFBS that considers mutation process and positional importance. That is, the algorithm calculates m different E-scores for each candidate TFBS (m is the number of sets of co-regulated sequences). At the 2nd step, the algorithm selects putative TFBS among every possible candidate TFBS by a statistical method. Since the goal of the proposed algorithm is to predict TFBS that work only in a target set of co-regulated gene sequences, the algorithm uses the other m-1 sets of co-regulated sequences as reference sets to predict TFBS in a target set of co-regulated sequences. We applied the

proposed algorithm to Dictyostelium gene sequences. Dictyostelium discoideum (Dd) is a soil living amoeba that lives on bacteria. It is haploid organism with 12,500 estimated genes. Many of the known genes show a high degree of sequence similarity to genes in vertebrate species. Dd is a facile system for basic biomedical research in cell and developmental biology, having unique advantages for studying fundamental cellular processes either absent or less accessible in other organisms with powerful molecular genetic tools. Dd has clearly different life stages. There are 4 different life stages in Dd gene sequences. Dd lives on forest soil and eats other bacteria and yeasts and shows a single cell organism life. When they do not have enough food (starvation condition), these bacteria aggregate each other and develop as a true multicellualr organism. After aggregation, a multicellualr mass behaves as a single organism, generating a motile slug form to seek more optimal environmental conditions and then producing a fruiting body cellular stalk and spores. Finally it germinates and goes back to a single cell bacterium. We classified the life stages of Dd into the vegetation stage, aggregation stage, slug stage, and culmination stage and collected genes according to their life stage from the website of Dicty cDNA project of the University of Tsukuba at Japan (http://dictycdb.biol.tsukuba.ac.jp/). We prepared the proximal region of each gene sequence (5' untranscribed region of gene sequences) from Dictyostelium BLAST search in San Diego State University (http://dicty.sdsd.edu). The proposed algorithm was tested as to whether it could predict developmentally related Dd TFBS that have already been experimentally proven from the literature. The literature search resulted in five known TFBS as follows: Psp-3B variant Harwood, ras S-promoter, Gene 7E psp-AT type, ecmB Repressor and MMF Elements. We found these five known TFBS in our simulation results. Among putative TFBS, we predicted five highly putative TFBS in each life cycle of Dd. "tcatgattgtc," "tatctcatcg," "acttttgag," "acagatttgt," and "gccaatttga" were predicted in the vegetation stage; "ctagaactaa," "taacaatacc," "tgggtgtcaa," "attctaccac", and "ttcatttgag" were predicted in the aggregation stage; "tattgcataa," "ttggaaactg," "caataatgtg," "tggtgggcga," and "atggtgggcg" were predicted in the slug stage; "tttgcccaac," "aaatggaagg," "cccttaaacg," "ttaaatggta," and "tcaaacgata" were predicted in the culmination stage. We chose these TFBS within top twenty-five putative TFBS in each life stage of the putative TFBS length 10 that were also predicted as putative TFBS in the length 8 and 9bp. With these five putative TFBS, one can confirm the prediction power of the proposed approach. The proposed algorithm can be applied to other species to predict their TFBS in co-regulated gene sequences such as vertebrate and invertebrate species.