Poster H-48

Chaos Game Representation and Vector Quantization (CGR-VQ): a new computational tool for the identification of transcription factor binding sites

Authors:
Dominik Beck (*Biomathematics Group, ITQB-UNL / INESC-ID*)
Jonas S Almeida (*Dept Biostatistics and Applied Mathematics, Univ. Texas MDAnderson Cancer Center / Biomathematics Gr*)
Ana Teresa Freitas (*IST/ INESC-ID*)
Arlindo L. Oliveira (*IST/ INESC-ID*)
Susana Vinga (*INESC-ID /FCM-UNL*)

Short Abstract: A new computational methodology for the Identification of Transcription Factor Binding Sites in DNA promoter regions is presented. This algorithm combines Chaos Game Representation and cluster analysis using Vector Quantization. This technique was tested on real and artificial datasets, showing good agreement with biological knowledge and other motif finding algorithms.

Long Abstract:

The behavior and properties of cells are determined in great part by the proteins that are expressed. An important process in the synthesis of proteins is the regulation and control of transcription, in which specific genes are transcribed into the initial ribonucleic acid (RNA) transcript, further processed and eventually translated to an expressed protein.

Transcription is usually activated in the promoter regions of genes. These functional regions include a specific short sequence that is called Transcription Factor Binding Site (TFBS). In order to trigger the transcription of a gene, specific proteins called Transcription Factors (TF) need to bind to this region. If the binding between TF and TFBS is strong, transcription is more effective and produces more expressed proteins. It has been observed that TFBS are conserved within functional and phylogenetic similar organisms or sequences. This is due to the fact that similar organisms with comparable function need to synthesize the same or similar proteins. On this basis we can identify TFBS by seeking repetitive patterns in genomic datasets of functional regions such as promoter sequences. However some properties of these motifs sometimes hamper their accurate identification. A problem is that the motifs might not be identical for every input sequence, but can have variations or mutations on one or more nucleotides. They can also appear as structured motifs that consist of two or more boxes separated by a distance of arbitrary nucleotides. In this case the first box motif often initializes and controls transcription, whereas the second box additionally suppresses or represses transcription. Another problem is that the TFBS, or the region that is affected by a TF, is a very short segment, of about 6 to 20 base pairs, compared to the overall lengths of functional regions in genomes. The identification of such TFBS remains a main challenge in functional genomics and computational biology.

The theory of chaos game representation (CGR) has been used to map nucleotide sequences into a two dimensional space, given by a coordinate system in $[0,1]^2$. This is done by successively processing the input symbols with a special two dimensional iterative

function system (IFS) called chaos game representation. Using this system a DNA sequence can be mapped onto a 2D coordinate system where each nucleotide is represented by one point in that space (that is only dependent on the coordinates of its predecessor – the Markov property). The system assigns, for each of the four DNA bases {A,T,G,C}, one of the four vertexes (0,0), (1,0), (0,1), (1,1) in the map.
The initial starting point is usually chosen at the map centre x_0 = (0.5,0.5). However it has been shown that the starting point can be randomly selected, without altering any of the CGR properties. The first nucleotide is mapped from x_0 half way into the direction of its symbols vertex, where it generates the point x_1. The following symbols x_i, i=1,…,SeqLength are mapped in the same way but starting always from their predecessor x_{i-1}. This leads to a fractal like representation of the nucleotide sequence that has the property of mapping similar substrings/suffixes near each other in the space, independently of their prefixes. It has been shown that the distance between similar substrings is reduced by the factor 0.5 with each shared or common symbol.
The proposed algorithm makes use of this feature. It starts by mapping all the input sequences into a CGR map. Since most of the input sequences are expected to share one particular substring the map has a high density of points in the area where this pattern is mapped, i.e., a specific sub-quadrant.
This region can then be identified and extracted using an unsupervised learning technique. We applied the Linde-Bouzu-Grey (LBG) algorithm, a method which is similar to k-means clustering and is frequently used in Vector Quantization for the design of codebooks in areas such as image, speech and signal processing.
The LBG algorithm progresses in 5 steps:

1. Choose number of clusters N;
2. Randomly choose N input vectors (CGR coordinates) and set them as cluster centers;
3. Calculate the memberships of all input vectors using a metric or dissimilarity measure;
4. Calculate for each cluster its centroid and call the centroids new cluster centers;
5. If (new cluster centers == old cluster centers) stop; else goto 3.

After the extraction of the high density area we used the cluster centre coordinates and a reverse function that, starting with a coordinate in CGR space, calculates the nucleotide sequence that could have produced it. The suffix of this sequence if further extracted, up to a desired length, chosen by the user. Given the random initial selection step, the procedure is repeated several times, thus employing different starting initial cluster centers.
Using this combination of CGR representation and machine learning we have obtained good results in both artificial and genomic datasets. We tested 7 artificial datasets divided in two groups. In the first one the motif was present in all the sequences and in the other set the planted motif appeared only in a smaller subgroup, i.e. with a lower occurrence rate (low quorum). In these artificial datasets we could fully recover the 7 implanted motifs and missed 3 with an error of one base.
We also tested a real genomic dataset consisting of the promoter regions of Pseudomonas putida KT2440 and completely identified the 7 bases corresponding to the first motif. We additionally recovered the second motif with a length of 4 bases with an error of one base.
Other known motif finding algorithms were also tested, namely SMILE, Bioprospector and MEME, in order to compare these methodologies in terms of accuracy, complexity and main advantages and drawbacks.
One improvement of the algorithm currently being developed it the use of an automatic

procedure to calculate the number of cluster centers based on information criteria. Other metrics and different clustering algorithms might also conduct to better overall results.