H - 48

Chaos Game Representation and Vector Quantization (CGR-VQ)

- a new computational tool for the identification of transcription factor binding sites

Dominik Beck[1,2], Jonas S. Almeida[3*,2], Ana Teresa Freitas[1,4], Arlindo L. Oliveira[1,4], Susana Vinga[1,5,2]

1 ALGOS Group, INESC-ID (*Portugal*)
2 Biomathematics Group, ITQB/UNL Instituto de Tecnologia Química e Biológica (*Portugal*)
3 Dept Biostatistics and Applied Mathematics, Univ. Texas MDAnderson Cancer Center (*USA*)
4 IST Instituto Superior Técnico / Universidade Técnica de Lisboa (*Portugal*)
5 FCM/UNL Faculdade Ciências Médicas / Universidade Nova de Lisboa (*Portugal*)

* Presenting author

dbeck@itqb.unl.pt , *jalmeida@mdanderson.org , atf@algos.inesc-id.pt , aml@inesc-id.pt , svinga@algos.inesc-id.pt

1 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 alignment-free scale-independent technique was tested on real and artificial datasets, showing good agreement with biological evidence and reference motif finding algorithms.

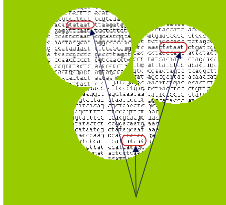
2 Introduction

Transcription is mainly influenced by transcription factors that bind in specific promoter regions of genes, called transcription factor binding sites (TFBS). It is broadly considered that these binding sites are conserved in functional and phylogenetic similar datasets. On this basis we can identify TFBS by seeking repetitive patterns in the dataset under study. However these patterns are not 100% identical for each sequence but can vary as regulatory factor. The accurate and complete identification of such TFBS remains a main challenge in functional genomics and computational biology.

3 Methods and algorithms

DATA

Sequences from phylogenetic similar organisms sharing the same TFBS



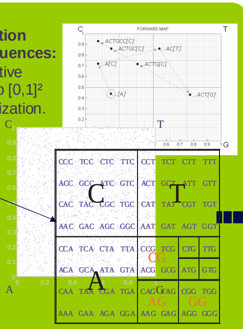
- •Set of Prokaryotic sequences sharing the Pribnow or **TATA-box**.
- •This TFBS identifies the exact nucleotide of transcription initiation.

CGR

VECTOR MAPS

Chaos game representation (CGR) of nucleotide sequences: DNA mapping via an iterative function system (IFS) onto [0,1]²
→ Markov Chain generalization.

Suffix property – Strings ending in a specific suffix are in the sub-quadrant labeled with the suffix


FORWARD MAP
C
T
G
A
ACTGCC[C]
ACTGC[C]
...AC[T]
...A[C]
ACTG[C]
...[A]
...ACT[G]
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.2
0.4
0.6
0.7
0.8
0.9
1
CCC TCC CTC TTC CCT TCT CTT TTT
ACC GCC ATC GTC ACT GCT ATT GTT
CAC TAC CGC TGC CAT TAT CGT TGT
AAC GAC AGC GGC AAT GAT AGT GGT
CCA TCA CTA TTA CCG TCG CTG TTG
ACA GCA ATA GTA ACG GCG ATG GTG
CAA TAA CGA TGA CAG TAG CGG TGG
AAA GAA AGA GGA AAG GAG AGG GGG
CG
AG
GG

IFS function :

$$\begin{cases} x_0 = (0.5, 0.5) \\ x_i = x_{i-1} + \frac{1}{2}(y_i - x_{i-1}), i = 1, \dots, N \end{cases} \quad \text{where} \quad y_i = \begin{cases} (0,0) \text{ if } s_i = 'A' \\ (0,1) \text{ if } s_i = 'C' \\ (1,0) \text{ if } s_i = 'G' \\ (1,1) \text{ if } s_i = 'T' \end{cases}$$

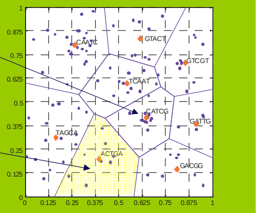
VQ

CLUSTER ANALYSIS

Vector quantization: automatic detection of important "words" with biological meaning, for example in promoter regions: TTGACA--TATAAT

Cluster centroids or codevectors: represent the motifs and can be recovered univocally

Voronoi regions: Over-expressed motifs are grouped in high density regions

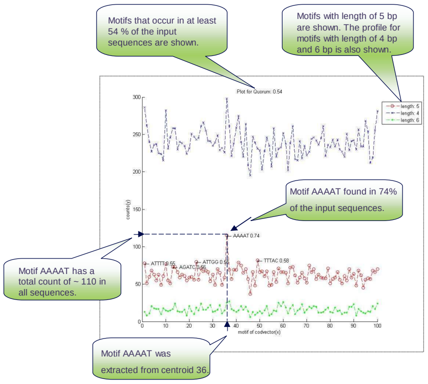


Linde-Buzo-Gray (LBG) algorithm:

```
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 center == old cluster centers) stop; else goto 3;
```

4 Results and Discussion

Result graph for the Dataset bioC



Artificial Datasets

Dataset	Expected motif	Found motif
M3	ATC	ATC
M4	ATCG	ATCG
M5	ATCGA	ATCGA
M7	ATC-X-AGC	ATC, AGC

σ54 regulon of Pseudomonas putida

Sigma54	TGGCACG TTGC	TGGCACG, TGGC

Artificial Datasets

Dataset	Expected motif	Found motif
bioA	TTTTA	TTTTA
bioB	AAAAT CCCCT	AAAAT CCCGT
bioC	AAAAT TTTTA CCCCT	AAAAT TTTAC GCCCC

Similar results with programs MEME, SMILE and Bioprospector

5 Conclusions and future work

- **Find optimal number of centroids**
  - **→ Use Information Theory to automatically estimate these parameters**
- **Estimate motif length to be extracted**
  - **→ Use biological knowledge and graph densities**

- **Combination of CGR and VQ is a good and flexible method for the extraction of short conserved motifs in biological sequences.**
- **Results show good agreement with biological knowledge and other motif finding algorithms**

Acknowledgements:

D.Beck thankfully acknowledges the financial support by grant D/2004/PL-42057-S from the European Union Leonardo da Vinci Program. This work was partially funded by project MaGiC (IE02ID01004) from INESC-ID (A.T.Freitas, PI).