Cumulative Local Cross-Correlation – an Algorithm for the Decomposition of Sequence Patterns

Simon Kogan[1]

Keywords: cross-correlation, pattern, repeat, sequence

1 Introduction.

Nucleotide sequences (DNA), a store of biological inheritance information, contain multiple codes (messages) responsible for an organism's functioning and structure [8]. A nucleic sequence can be seen as a signal and investigated by formal methods of signal processing theory: Fourier [6] and wavelet [1] transforms, cross-correlation analysis [2, 7], etc. Cross-correlation is especially suited for the indication of common (frequently encountered) sequence parts (i.e., repeats). A dispersed repeat (frequent combination of oligonucleotides following each other at some specific distance with an arbitrary sequence in between) will be captured as well.

Examining cross-correlation between AA and TT dinucleotides in the human genome reveals a peak in lag '12' that suggests an existence of sequence repeat(s) containing AA dinucleotide followed by TT dinucleotide at a distance of 12 bases. One can find all positions of such a motif (AA-12-TT) in a given sequence, and then, calculate a nucleotide local distribution in the vicinity of the motif. If there is only one unique repeat containing this motif, the distribution is actually a pattern representing the repeat. If there are several different repeats containing the motif, the distribution is an overlapping of all of the patterns. To reconstruct one of the patterns or all of them, one has to decompose the distribution picture. This is exactly the objective of the algorithm developed in this work.

2 The algorithm.

In order to calculate cross-correlation functions and also use the Cumulative Local Cross-Correlation (i.e., CLCC) algorithm, one needs to represent an investigated nucleic sequence by 4 discrete signals: the first one for 'A' nucleotide positions ('1' – where it is present, '0' – where it is absent); the second one for 'C' nucleotide positions, etc.

The algorithm is as follows:

For a given nucleic sequence, calculate nucleotide local distribution in the vicinity of the initial motif (AA-12-TT for example).

For each occurrence of the motif in the sequence, calculate the cross-correlation (including auto-correlation) between the signals in each pair of positions inside a window enclosing the motif. The result of this operation is a symmetric matrix (local cross-correlation instance).

Calculate cumulative local cross-correlation matrix as a sum of local cross-correlation instances. The number of CLCC matrices is equal to the number of cross-correlation functions (16 in our nucleotide example). Each element of CLCC matrix characterizes correlation between a pair of positions in the local distribution.

Now, in order to know, whether two specific peaks in the local distribution are related to the same pattern or not, one just needs to check a value of CLCC matrix element associated with them (larger value for correlated peaks than for uncorrelated ones). Theoretically, by checking CLCC matrix elements associated with all pairs of peaks in a local distribution and collecting mutually correlated peak groups, all patterns can be separated from each other.

[1] *Genome Diversity Center, Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel. E-mail: skogan@research.haifa.ac.il*

3 Application to human genomic sequences.

The following example illustrates the CLCC application to one contig (7MB) of human chromosome **1**. The development of this algorithm took place during the investigation of a nucleosome pattern based on dinucleotides [5]. There are 16 local distributions and 256 cross-correlation functions (and CLCC matrices as well) in dinucleotide case (against 4 of the former and 16 of the latter in nucleotide case).

AA-12-TT was chosen as an initial motif and CLCC matrices (and local dinucleotide distributions) in its vicinity were obtained as described in the previous section. Analysis of CLCC matrix elements associated with several pairs of peaks in the local distribution permitted an isolation of mutually correlated group of peaks. The new, more detailed motif was constructed from these peaks and dinucleotide local distributions in the vicinity of the new motif were calculated.

Knowing, that the new distribution is a non overlapped one, we converted it to a nucleic sequence (the non-overlapping is a necessary condition for the conversion). The resulting sequence was found to be very similar to the well-known Alu tandem repeat widely present in the human genome [3]. Therefore, we were able to align the sequence with Alu consensus [4] to validate the CLCC technique (see Fig. 1 for the alignment result).

```
AAAAAAAAAAAATTCAGGCgcGGgGCGGgGGggCcCcCgccTAAaCCCccacCcTTGGGgGGggGgGGgG
                  GGCcgGGcGCGGtGGctCaCgCctgTAAtCCCagcaCtTTGGGaGGccGaGGcG

GGgGGgaacCCcGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACcTGGTGAAACCCCGgCTCTACTAAAA
GGcGGatcaCCtGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACaTGGTGAAACCCCGtCTCTACTAAAA

ATACAAAAATTAGCCGGGCGTGGTGGCGCGCGCCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGA
ATACAAAAATTAGCCGGGCGTGGTGGCGCGCGCCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGA

ATCGCTTGAACCCGGGAGGCGGAGGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGAC
ATCGCTTGAACCCGGGAGGCGGAGGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGAC

AGAGCGgGACcCCGTCcCAAAAAAAAAAAAAAAAAAAAAAA
AGAGCGaGACtCCGTCtCAAAAAAAA
```

Fig. 1: Sequence alignment: upper (underlined) rows – the extracted nucleic sequence; lower rows – Alu consensus.

References

[1] Arneodo, A., E. Bacry, P.V. Graves, and J.F. Muzy. 1995. Characterizing long-range correlations in DNA sequences from wavelet analysis. Physical Review Letters. 74(16): pp. 3293-3296.

[2] Herzel, H., E.N. Trifonov, O. Weiss, and I. Grosse. 1998. Interpreting correlations in biosequences. Physica A. 249(1-4): pp. 449-459.

[3] Hwu, H.R., J.W. Roberts, E.H. Davidson, and R.J. Britten. 1986. Insertion and/or deletion of many repeated DNA sequences in human and higher ape evolution. Proc Natl Acad Sci U S A. 83(11): pp. 3875-9.

[4] Jurka, J. and T. Smith. 1988. A fundamental division in the Alu family of repeated sequences. Proc Natl Acad Sci U S A. 85(13): pp. 4775-8.

[5] Kato, M., Y. Onishi, Y. Wada-Kiyama, T. Abe, T. Ikemura, S. Kogan, A. Bolshoy, E.N. Trifonov, and R. Kiyama. 2003. Dinucleosome DNA of human K562 cells: experimental and computational characterizations. J Mol Biol. 332(1): pp. 111-25.

[6] Peng, C.K., S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H.E. Stanley. 1992. Long-range correlations in nucleotide sequences. Nature. 356(6365): pp. 168-70.

[7] Trifonov, E.N. and J.L. Sussman. 1980. The pitch of chromatin DNA is reflected in its nucleotide sequence. Proc Natl Acad Sci U S A. 77(7): pp. 3816-20.

[8] Trifonov, E.N. 1989. The multiple codes of nucleotide sequences. Bull Math Biol. 51(4): pp. 417-32.