DNA PATTERNS LOCALIZATION USING SPECTROGRAMS

Alin VOINA, Petre G. POP
Comm. Dept., Technical University of Cluj-Napoca
26-28 Baritiu str. 400027, Cluj-Napoca, ROMANIA, e-mail: alin.voina@com.utcluj.ro

Abstract. This paper presents the possibilities for processing and analysis of genomic sequences which are converted into a numerical representation. There were used two types of symbol mapping: indicator sequences and quartic mapping. The resulted sequences were analyzed using a spectral analysis, completed with the obtaining of BW spectrogram, technique which offers the possibility for a unique and innovative visualization of periodical properties of genomic sequence. The advantages of this kind of representation are that offers an overview of the whole genomic sequence without using any *a priori* data. Once we have the spectrogram, we can obtain approximate information about positions and length of the repeated sequences.

Keywords: *DNA repeats,spectral analysis, Fourier transform, spectrogram, genomic sequence*

I. INTRODUCTION

The data regarding structural and functional features of the genomes, for various organisms is being accumulated and analyzed in laboratories all over the world, from the small university or clinical hospital laboratories, to the large laboratories of pharmaceutical companies and specialized institutions, both state owned and private. This data is stored, managed, and analyzed on a large variety of computing systems, from small personal computers using several disk files to supercomputers operating on large commercial databases. The volume of genomic data is expanding at a huge and still growing rate, while its fundamental properties and relationships are not yet fully understood and are subject to continuous revision.

The standard symbolical representation of genomic information—by sequences of nucleotide symbols in DNA and RNA molecules or by symbolic sequences of amino acids in the corresponding polypeptide chains (for coding sections)—has definite advantages in what concerns storage, search, and retrieval of genomic information, but limits the methodology of handling and processing genomic information to pattern matching and statistical analysis. This methodological limitation determines excessive computing costs in the case of studies involving feature extraction at the scale of whole chromosomes, multiresolution analysis, comparative genomic analysis, or quantitative variability analysis [1-3].

Conversion of the DNA sequences into digital signals, offers the possibility to apply signal processing methods to analyze genomic information and reveal characteristics of the DNA sequence, which in fact would be difficult to detect using standard statistical methods and symbol patterns of genomic sequences.

In recent years, the analysis of genomic signals has shown potential in the discovery of large scale features of DNA sequences, 10^6-10^8 base pairs, including both coded and non-coding regions of the chromosome or an entire genome.

II. NUMERICAL REPRESENTATION OF DNA SEQUENCES

Although these biological data collection effort is very high, far more significant is the challenge to identify gene expression mechanisms, determining the protein encoded by genes and understanding how they interact with each other.

The main genetic material constituent cells, is represented by DNA molecules that have a basic structure simple and well known [4]. Repetitive units are nucleotides, each consisting of three strong covalent bonds. In the DNA sequence are two types of nitrogen bases, namely: Thymine (T) and Cytosine (C)-which are pyrimidines, Adenine (A) and Guanine (G) - which are purines. As we can see, the alphabet in the case of DNA sequence has a size of 4 and consists of the following letters: A, G, C and T. The conversion of genomic sequences, from the form of symbolic representation, which are in various public databases [5, 6], into a digital genomic signal may be achieved using various processing and analysis techniques used in digital signals. There are many possibilities to map symbols of genomic data in digital genomic signals. One possibility is to map nucleotides in three vectors, symmetric placed in the three-dimensional space (3D), oriented towards the tips of a regular tetrahedron. By choosing coordinates (± 1) for the cube

Manuscript received February 10, 2010; revised April 02, 2010

peaks, the vectors representing the four nucleotides will have the following expressions [5]:

$$\begin{cases} \vec{a} = \vec{i} + \vec{j} + \vec{k} \\ \vec{c} = -\vec{i} + \vec{j} - \vec{k} \\ \vec{g} = -\vec{i} - \vec{j} + \vec{k} \\ \vec{t} = \vec{i} - \vec{j} - \vec{k} \end{cases} \quad (1)$$

This representation is entirely appropriate for well-defined sequences, in which each entry is uniquely specified. Such sequences are found in databases that integrate a large volume of genomic data.

Representation of the nucleotides may be reduced to a bi-dimensional form by designing the nucleotide tetrahedron on the right design. This plan can be mapped to the complex plane, thereby achieving a complex representation of nucleotides [5]:

$$\begin{cases} a = 1 + j \\ c = -1 - j \\ g = -1 + j \\ t = 1 - j \end{cases} \quad (2)$$

In this representation, complementarity of base pairs A-T and C-G is expressed by symmetry to the real axis (complex conjugate representations are: t= a *, g = c *) and pairs purine / pyrimidine have the same imaginary parts.

One of the major mapping techniques that are used frequently is that in which the sequence is mapped using indicators. These sequences contain the numbers 0 and 1 to indicate the absence or presence of the symbol in the original sequence.

Thus, if we consider a sequence S of length N containing the symbols A, G, C and T, the four indicator sequences are:

$$\begin{cases} S_A(n) = 1 \text{ if } S(n) = A \text{ or } 0, \text{otherwise} \\ S_C(n) = 1 \text{ if } S(n) = A \text{ or } 0, \text{otherwise} \\ S_C(n) = 1 \text{ if } S(n) = C \text{ or } 0, \text{otherwise} \\ S_T(n) = 1 \text{ if } S(n) = T \text{ or } 0, \text{otherwise} \end{cases} \quad (3)$$

A new mapping technique, Cumulative Categorical Periodogram (CCP) [7] can be defined for any definite sequence. CCP (i), $1 \le i \le N$, for a sequence of length N, is the number of occurrences of cycles of period i. In the context of categorical time series, a cycle is considered to be achieved when a state is encountered previous cycle and is defined as one plus the number of events occurred. Mathematically, it is expressed as:

$$CCP(i) = \sum_{j=i+1}^{N} \begin{cases} 1, \text{if } S(j) = S(j-i) \text{ and } 1 \le k < j-i, S(k) \ne S(i) \\ 0, \text{otherwise} \end{cases} \quad (4)$$

Reducing the size of the nucleotide sequence, codon and amino acids can be achieved by using a real mapping, one-dimensional. Digits (0, 1, 2, and 3) can be associated to the four nucleotides. Thus, a full DNA sequence can be seen as a huge number written on the basis of four. There are 4! = 24 possibilities for attachment of digits 0-3 to nucleotides A, C, G, T. The best allocation for nucleotides is given in Table 1 and results from the condition to obtain the most monotonic mapping of the codons 0–63 to the amino acids plus the terminator 0–20, leading to best autocorrelated intergene genomic signals [5,10].

Pyrimidines	**Purines**
Thymine=T=0	Adenine=A=2
Cytosine=C=1	Guanine=G=3

Table 1. Nucleotide representation with the help of digits in base 4

Another type of quartic mapping it is with the help of EIIP potentials (Electron Ion Interaction Potentials), where nucleotides were mapped into values described in Table 2.

Nucleotides	**EIIP value**
A	0.1260
C	0.1340
G	0.0806
T	0.1335

Table 2. EIIP values assigned to nucleotides

III. DNA SPECTROGRAM ANALYSYS

To obtain global information about the existence of periodicity from the analysis of genomic sequence is quite difficult in a frame analysis. Thus, we need a technique to detect the beginning and the end of a periodicity in a region of sequence. Once we detect a local periodicity and we estimate its fundamental period, we can determine the subsequent analysis window corresponding to the intervals identified. To achieve this, we make an analysis of genomic sequence through the help of *spectrograms*.

The spectrogram is an "image" which shows the spectral density variation of a signal. The most common format in which spectrograms are represented, is a geometric graph with two dimensions: horizontal axis represents position and the vertical axis represents frequency; a third dimension with the role of indicating the magnitude of a particular frequency at a given moment, is the intensity or color of each point in the image.

The spectrogram can be obtained by applying *Discrete Fourier Transform (DFT)* on the signal represented in time domain.

Creating a spectrogram using DFT is usually a digital process. Digitally sampled data, in time, are divided into smaller windows which usually overlap, and then DFT is applied to calculate the amplitude spectrum of each window. Thus, each window corresponds to a vertical line image which represents a measure of the amplitude vs. frequency at a specific time. Spectra obtained are then attached one by one to form an "image". The spectrogram is given by the square amplitude of the DFT:

$$Spectrogram\ (t, \omega) = |\ DFT\ (t, w)\ |^2 \qquad (5)$$

To obtain the spectrogram of a genomic signal, the next steps were followed:

- genomic sequence was converted into digital signal by one of the representations described above;
- the sequence number obtained was divided into *sliding windows* (of size 2's power) on which we applied DFT;
- we've set the number of rows and columns of the spectrogram according to the size and the overlap degree corresponding to the *sliding window*;
- we've performed an amplitude spectrum calculation for each analysis window and we've map the spectral components to a level of gray;
- the resulted gray levels associated with the corresponding row and column, form the spectrogram;

The mapping consists of multiplying each indicator sequences (shown above) with one factor which represents one of the numerical values assigned to the four symbols A, C, G, T that are found in genomic sequence. The expression of the sequence becomes:

$$X_{seq}(n) = C_A X_A(n) + C_C X_C(n) + C_G X_G(n) + C_T X_T(n) \qquad (6)$$

The values for C_A, C_C, C_G, C_T coefficients are those described in Table 3, n =0..N-1 , where N is the size of analysis window.

Coefficient	Coefficient
C_A =1	C_G =2
C_C =3	C_T =4

Table 3. C_A, C_C, C_G, C_T values

To reveal periodicity in the case of quartic mapping, we used the genomic sequence AC017075.8 [6] as a case study.

In Figure 1, is a graphical representation of the genomic sequence AC017075.8 [6] for a quartic mapping.

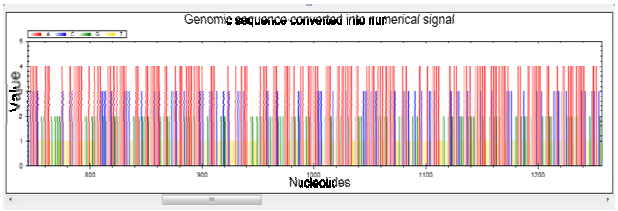

Figure 1. Genomic sequence AC017075.8 for a quartic mapping

The presence of periodicity in the spectrum is indicated by the presence of peaks which can be identified on the same position while scrolling the genomic sequence with an analysis window of size N.

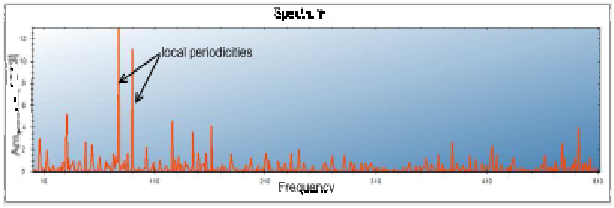

Figure 2. Local periodicities for AC017075.8 in case of quartic mapping

In Figure 2, we can identify these local periodicities for an analysis window of size 512 from genomic sequence AC017075.8 [6].

Since not all peaks are significant, for a better identification of periodicities, in Figure 3, we've filtered the spectrum by setting a threshold $T = T_{th} * X_{avg}$, where T_{th} is a threshold value between (0…4) and X_{avg} is the average value of the spectrum.

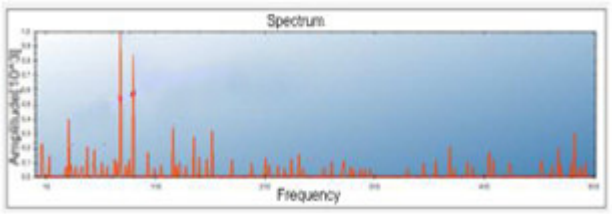

Figure 3.The resulted spectrum of an analysis window of 512 nucleotides, on genomic sequence AC017075.8, for quartic mapping and T_{th}=0.75

Next we illustrate the results obtained on several genomic sequences, making observations about the performance obtained from the variation of analysis

parameters. Thus, in Figure 4, is presented the spectrogram of genomic sequence AC017075.8 which was downloaded from GenBank database [8]. The genomic data used for generating the spectrogram, was mapped using indicator sequences. The following parameters were used:

- spectrogram component-frameworks were generated using a Fast Fourier Transform in 1024 points;
- the percentage of overlap of the *sliding window*, was set at 5%;
- the filtering threshold was set to T = 1.8;

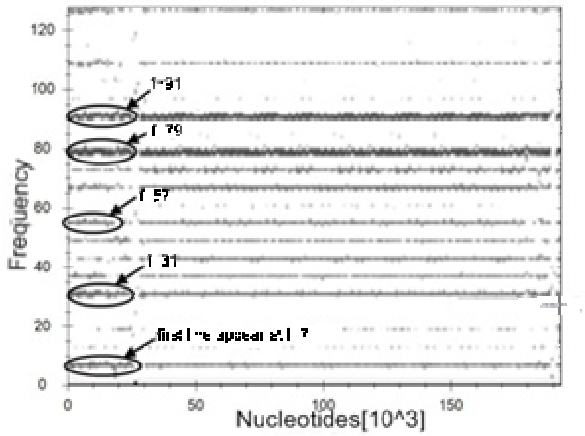

Figure 4. Spectrogram for AC017075.8 [6],overlap 5%, T=1.8, indicator sequences mapping

From Figure 4, can be observed that genomic sequence AC017075.8, have repeat regions containing long repeated sequences, shown by the large number of lines present in the spectrogram. First line of spectrogram is indicating the presence of repetitive sequences found at frequency f = 7, indicating a repetitive sequence of about 1024 div 7 > 100 nucleotides. This information on the number of repeat regions is confirmed by existing data in genomic databases (GenBank [6]) presented in Table 4.

Repetition region	Number of base pairs
6..18745	18740
18747..18857	111
18858..25768	6911
25771..26983	1213
28247..188848	160602
188849..189040	192
189044..190699	1656
190700..193275	2576

Table 4. Repetition region for genomic sequence AC017075.8 as described by GenBak [6]

Repeat regions in Table 4, are present also in spectrogram from Figure 4, but because these regions are very close, it's difficult to separate the regions of repetition from the spectrogram.

In Figure 5, the spectrogram of the same DNA sequence (*AC017075.8*) is realized by using quartic mapping and represented with the help of a *sliding window* with an overlap factor set to 15% and a filtering threshold set to T=2.5 .

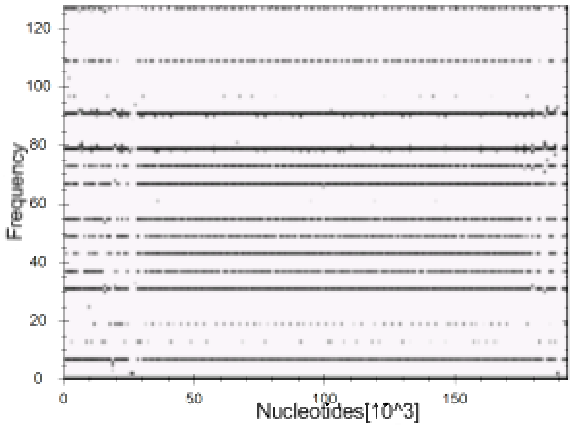

Figure 5. Spectrogram for AC017075 overlap 15%, T=2.5, quartic mapping

If we set the filtering threshold at T = 3.5, we obtain spectrogram from Figure 6, in which we can identify only lines corresponding to frequencies f = 7, f = 31, f = 79 and f = 91. With this filter, we've practically reduced spectrum peaks and have remained only the most significant, highlighting the global periodicity occurring in the analyzed genomic sequence.

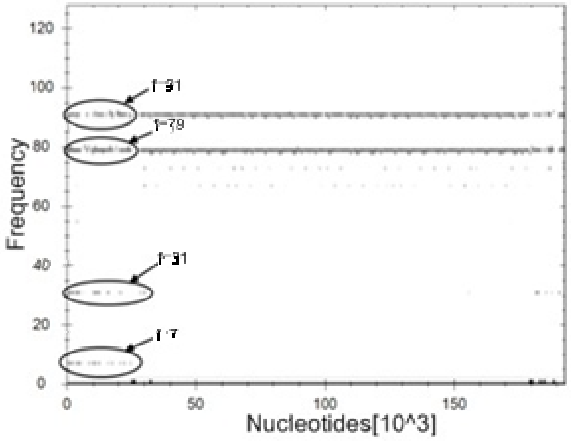

Figure 6. Spectrogram for AC017075.8, overlap 15%, T=3.5, quartic mapping

Thus, it is noted especially the presence of repetitive regions between position 6 and 26,983, according to Table 3.

In Figure 7, is represented the spectrogram for *AC017075.8* for the case of EIIP mapping, with a threshold set to T=1 and the overlap of sliding window set to 15%.

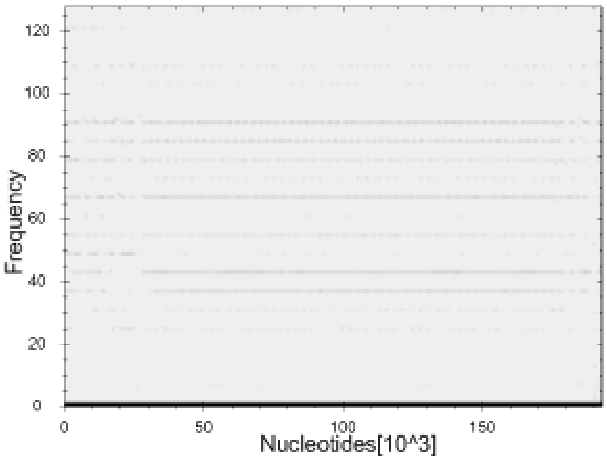

Figure 7. Spectrogram for AC017075.8, overlap 15%, T=1, EIIP mapping

Next figures show spectrograms for sequence M13882.1 which is a human alpha satellite from the centromeric region DNA. This microsatellite presents several repeat regions reported by Genbank, with size and localization presented in Table 5.

Repetition region	Number of base pairs
1-166	166
167-337	171
338-508	171
509-675	167
676-850	175
851-1016	166
1017-1187	171
1188-1358	171
1359-1529	171
1530-1700	171
1701-1870	170
1871-2041	171
2042-2208	167
2209-2375	167
2376-2541	166
2542-2712	171

Table 5. Repetition region for genomic sequence M13882.1 as described by GenBak [6]

The spectrograms were realized for different mapping methods, using FFT in 512 points, with 15% overlap of the sliding window and a threshold set to T=1.5 .

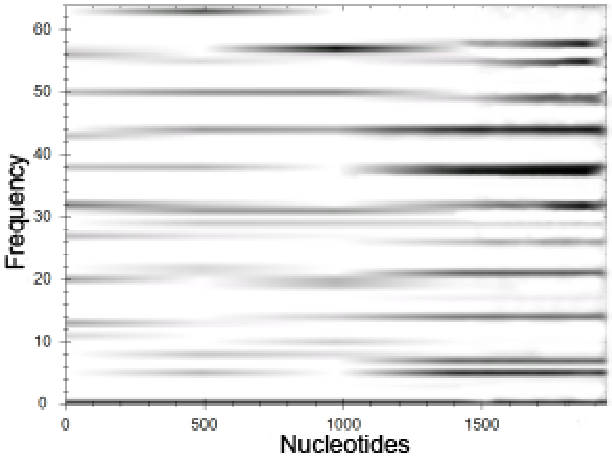

Figure 8. Spectrogram for M13882.1, overlap 15%, T=1.5, indicator sequence mapping

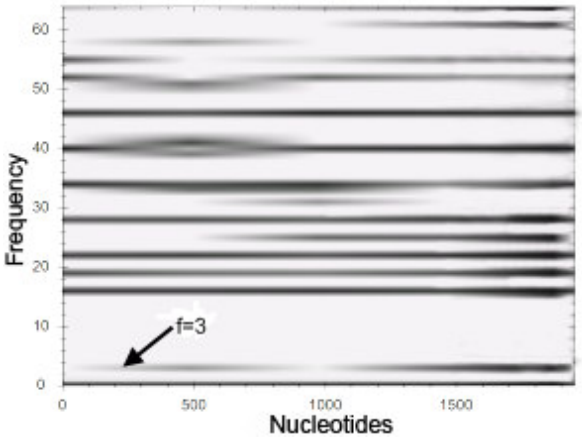

Figure 9. Spectrogram for M13882.1, overlap 15%, T=1.5, quartic mapping

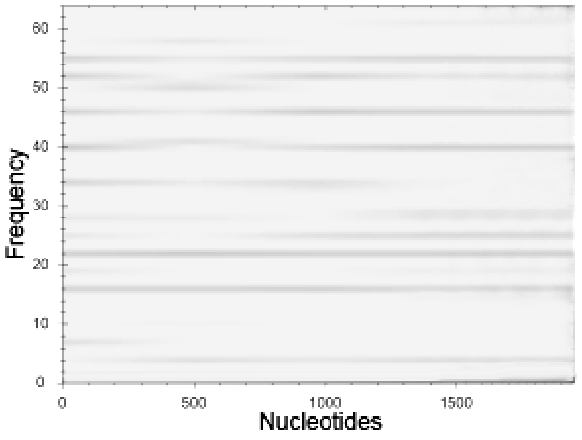

Figure 9. Spectrogram for M13882.1, overlap 15%, T=1.5, EIIP mapping

From above spectrograms, can be observed that genomic sequence M13882.1, have repeat regions containing several repeated sequences. Continuous lines

found in the spectrogram suggest that the analyzed sequence presents repeat regions very close to each other. The best results were obtained for quartic mapping, spectrogram in Figure 9. First line of spectrogram is indicating the presence of repetitive sequences found at frequency f = 3, indicating a repetitive sequence of about 512 div 3 ≅ 171 nucleotides (confirmed also by Table 5).

IV CONCLUSIONS

The main idea of this study was to present some possibilities for conversion of the DNA sequence into a numerical representation and to get an overview of the analyzed sequence with the help of spectrograms.

The most relevant results were obtained by substituting the nucleotides symbols (A, C, G and T) with the numerical values from Table 3 or with electron-ion interaction pseudopotentials (EIIP) of the nucleotides, described in Table 2. The computational overhead was reduced with 75%, because instead of using four indicator sequences, we used only one. Also, by analyzing different DNA sequences, we found that some repeated regions (identified in case of spectrograms generated with a quartic mapping) were not reported by spectrograms obtained when indicator sequences were employed.

Many of the methods used in genomic signal processing are based on Fourier analysis (analysis also used in the present study), but it provides information only in frequency. The conventional Fourier analysis can indicate only the overall periodicity for stationary signals [9].

A significant challenge in bioinformatics is to find the best ways to manage the quantity and complexity of information from genome. Analysis of the genomic sequence with the help of the spectrogram is providing a unique view of the DNA sequence. The method can be used for preliminary investigation (screening) of the genomic sequences to determine the occurrence of repeated sequences. The method is robust to mutations and does not require prior information about the original sequence (pattern, position). By viewing the spectrogram, the analyzed region of the DNA sequence provides a unique signature, which proves particularly useful in early recognition of an area of interest.

ACKNOWLEDGMENT

This work was partly supported by CNCSIS – UEFISCSU, project number PNII – IDEI 903/2007.

REFERENCES

[1] A. Ben-Dor, N. Friedman, and Z. Yakhini, “Scoring genes for relevance,” Tech. Rep. AGL-2000-13, Agilent Laboratories, Palo Alto, Calif, USA, 2000.

[2] L.Wernisch, S. L. Kendall, S. Soneji, et al., “Analysis of whole-genome microarray replicates using mixed models,” *Bioinformatics*, vol. 19, no. 1, pp. 53–61, 2003.

[3] Y. Tu, G. Stolovitzky, and U. Klein, “Quantitative noise analysis for gene expression microarray experiments,” *Proc. Natl. Acad. Sci. USA.*, vol. 99, no. 22, pp. 14031–14036, 2002.

[4] R. Durbin, S. Eddy, A. Krogh, and G.Mitchison, *Biological Sequence Analysis.*
ProbabilisticModels of Proteins and Nucleic Acids, Cambridge University Press, Cambridge,UK, 1998.

[5] Edward R. Dougherty, Ilya Shmulevich, Jie Chen, and Z.JaneWang, *Genomic Signal Processing and Statistics*, Hindawi Publishing Corporation, 2005.

[6] The Genome Data Base, http://gdbwww.gdb.org/,
Genome Browser, http://genome.ucsc.edu,
European Informatics Institute, http://www.ebl.ac.uk,
Ensembl, http://www.ensembl.org.

[7] Achuthsankar S. Nair and T. Mahalakshmi*,” Are categorical periodograms and indicator sequences of genomes spectrally equivalent?* ”, In Silico Biology 6, 0019 (2006).

[8] National Center for Biotechnology Information, National Institutes of Health, National Library of Medicine, http://www.ncbi.nlm.nih.gov/genoms .

[9]Jianchang Ning, Charles N.Moore, J.Clare Nelson, *Preliminary Wavelet Analysis of Genomic Sequences*, IEEE Computer Society Bioinformatics Conference (CSB'03), 2003.

[10] P. Tamayo, D. Slonim, J. Mesirov, et al., “Interpreting patterns of gene expression with selforganizing maps: methods and application to hematopoietic differentiation,” *Proc. Natl. Acad.Sci. USA*, vol. 96, no. 6, pp. 2907–2912, 1999.