Prediction of Nucleotide Sequences by using Genomic Signals

PAUL CRISTEA, RODICA TUDUCE
Biomedical Engineering Center
University “Politehnica” of Bucharest
Spl. Independentei 313, sect. 6, 060042 Bucharest
ROMANIA
{pcristea, trodica}@dsp.pub.ro

VALERI MLADENOV, GEORGI TSENOV,
sIMONA PETRAKIEVA
Dept. of Theoretical Electrical Engineering
Technical University of Sofia
8 Kliment Ohridski St., 1000 Sofia
BULGARIA
{valerim, gogotzenov}@tu-sofia.bg

Abstract: - The conversion of symbolic sequences into complex genomic signals allows using signal processing methods for the handling and analysis of nucleotide sequences. This methodology reveals surprising regularities, both locally and at a global scale, allowing to predict nucleotides in a sequence when knowing the preceding ones. Such experiments have a biologic significance, as they explore the possibility and the efficiency of error correction in processes like replication, transcription and translation.

Key-Words: - Genomic signals, Sequence prediction, Time series prediction, Neural networks

1 Introduction

The conversion of symbolic sequences into complex genomic signals [3] allows using signal processing methods for the handling and analysis of nucleotide sequences. This methodology reveals surprising regularities, both locally and at a global scale, which would be difficult to identify by using only statistical analysis and pattern matching, as currently done for symbolic sequences. The approach is useful for studying large scale features of chromosomes [5,10,11], detect mutations in small and medium genomes such as those of pathogens [4,6,7,8,9] and model some ribosome functional properties [5, 14].

The regularities in the distribution of nucleotides [1] and pairs of nucleotides, reflected in the low values and predictable variation of the nucleotide imbalance (cumulated phase) and nucleotide pair imbalance (unwrapped phase), show that a genome has a multi-level ordered structure, despite the gene low compressibility. Mutations, such as those in the genomic signals of related pathogen strains, tend to mutually compensate, so that overall regularities of chromosomes are conserved. A direct consequence of this statistical regularity is that SNPs appear rarely isolated and more often in correlated groups, sometimes placed at large distances along the nucleotide sequences [3].

On the other hand, the regularity of genomic signals allows using techniques similar to those used in time series prediction [15] to estimate the nucleotides in a sequence when knowing the preceding ones [5,14]. Such experiments have a biologic significance, as they explore the possibility and the efficiency of error correction in processes like replication, transcription and translation.

In previous work, the efficiency of the prediction has been improved by using a two step procedure which comprises a principal analysis step (PCA) and a feed-forward ANN [13]. The PCA retains only the high variance components of the input signal, which are used by the ANN to perform the actual prediction.

The present paper uses either directly feed forward neural networks, or preceding step for feature extraction. This improves the prediction accuracy and increases the efficiency of the system. The results of experiments with this prediction model show a remarkable low error rate. This behavior is the direct effect of the regularities in the structure of the genomic sequences [1,3,11].

2 Nucleotide Representation

The mapping we are using [3] is a one-to-one (bijective) unbiased representation of nucleotide equivalence classes, which attaches complex numbers to adenine, cytosine, guanine and thymine nucleotides:

\[ a = 1 + j, \quad c = -1 - j, \quad g = -1 + j, \quad t = 1 - j. \]  \hspace{1cm} (1)

While conserving the information in the initial symbolic sequence, this mapping introduces no artifacts related to specific assumptions on the types of interaction that characterize the nucleotides.

Correspondingly, the distribution of nucleotides along a sequence is described by the nucleotide imbalance:
\[ N = 3(n_G-n_C)+(n_T-n_T), \]

where \( n_A, n_C, n_G \) and \( n_T \) are the numbers of adenine, cytosine, guanine and thymine nucleotides in the sequence, from the first to the current entry, while the distribution of nucleotide pairs by the nucleotide pair imbalance:

\[ P = n_+ - n_-, \]

where \( n_+ \) is the number of positive pairs (A→G, G→C, C→T, T→A), and \( n_- \) the number of negative pairs (A→T, T→C, C→G, G→A).

The genomic signal approach reveals large scale features of DNA sequences that are maintained over distances of \( 10^6 \) – \( 10^8 \) base pairs, including both coding and non-coding regions [3,5,10,11]. The methodology is also adequate for the study of pathogen variability and the identification of multiple drug resistance, important for fast diagnoses and prompt socio-medical decisions in contamination with pathogens such as Human immunodeficiency virus (HIV) [9], Avian influenza virus (H5N1) [8] and Mycobacterium tuberculosis (MT) [2,4,6].

3 Protease Gene of HIV-1 Clade F

The genomic signal representation has been used for the analysis and nucleotide prediction study of the protease (PR) gene of the Human immunodeficiency virus type 1 (HIV-1) clade F isolated from Romanian patients. The PR gene is an essential functional gene of the HIV-1 virus, which encodes the two identical peptide chains in the structure of the protease enzyme. The PR gene consists of 297 base pairs, located along the 1799...2095 bp segment of the NC001802 sequence [16].

The PR enzyme is a small dimension dimer that plays a vital role in the HIV life cycle by chopping up the long strands of polypeptide synthesized by the HIV RNA. After being generated by using the human ribosome in the attacked cell, the polypeptide molecules are cut by PR in smaller pieces corresponding to the viral proteins, at the proper sites and proper timing. The two PR enzyme chains, each 99 amino acids long, form a tunnel that holds the polypeptide, which is cut by the enzyme active site located in the center of the tunnel. Because of its importance, protease has been selected as a preferred target for anti-HIV drugs. The protease inhibitor drugs bind to PR to block its action.

4 Genomic Signals of PR Gene vs. Global Representation

The PR gene of HIV-1 subtype F virions isolated from Romanian patients have been sequenced at the National Institute of Infectious Diseases “Prof.Dr.Matei Bals”, Bucharest.

The symbolic sequences of the PR gene for 30 patients have been converted into complex genomic signals using (1). The nucleotide imbalance \( N \) (cumulated phase \( \theta_1 = \pi N / 4 \)) and the nucleotide pair imbalance \( P \) (unwrapped phase \( \theta_2 = \pi P / 2 \)) have been computed for the genomic signals using (2) and (3), respectively.

The signals have been classified in three groups taking into account the clinical behavior of the patients with respect to the response to current antiretroviral treatment:

- Sensitive (S) - the patient responds to the antiretro-viral treatment (wild type pathogen),
- Resistant (R) - patient resistant to one anti-HIV drug,
- Multiresistant (M) - patient resistant to several anti-HIV drugs.

The data set that we have used in this work contains equal groups of each type (10 patients who are sensitive, 10 patients who are resistant and 10 patients who are multiresistant).

Fig.1 shows the nucleotide imbalance \( N \) (cumulated phase) and the nucleotide pair imbalance \( P \) (unwrapped phase), for the sensitive, resistant and multiresistant patients.

The data have also been analyzed directly from the point of view of the nucleotide classes content. Figures 2-4 show the relative single and double nucleotide class content of the 30 individual sets of data, using the mapping given in Table 1.

<table>
<thead>
<tr>
<th>Class</th>
<th>IUPAC symbol</th>
<th>Complex representation</th>
<th>Scalar label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenine</td>
<td>A</td>
<td>[1, 1]</td>
<td>1</td>
</tr>
<tr>
<td>Guanine</td>
<td>G</td>
<td>[1, -1]</td>
<td>2</td>
</tr>
<tr>
<td>Cytosine</td>
<td>C</td>
<td>[-1, 1]</td>
<td>3</td>
</tr>
<tr>
<td>Thymine</td>
<td>T</td>
<td>[-1, -1]</td>
<td>4</td>
</tr>
<tr>
<td>Week bond</td>
<td>W</td>
<td>[1, 0]</td>
<td>5</td>
</tr>
<tr>
<td>Purines</td>
<td>R</td>
<td>[0, 1]</td>
<td>6</td>
</tr>
<tr>
<td>Strong bond</td>
<td>S</td>
<td>[-1, 0]</td>
<td>7</td>
</tr>
<tr>
<td>Pyrimidines</td>
<td>Y</td>
<td>[0, 1]</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 1 The data set mapping used for data presentation

The diagrams in Figures 2, 3 and 4 show that the nucleotide class content of the data for patients belonging to the same group of resistance to drugs is almost the same, as expected. Additionally, Fig. 5 shows that the average values of the nucleotide class content for different sets of data is also almost the same.
This proves the need for an adequate description and analysis of genetic information. The global statistical analysis of the nucleotide class content of the gene data does not detect the specific changes resulting in drug resistance. As shown in previous work [6,7,8,9], the development of the phenotypic resistance to drugs, at clinical scale, can be consistently linked to mutations in certain loci of the gene involved, at molecular scale. The specific locations of the mutations determine the drug resistance, while the nucleotide class content of the genes for various groups of drug resistant patients, and their average values do not.
The negative result of this investigation, performed at the level of global statistical features, confirms the previous work that clearly linked pathogen resistance to treatment, on one side, to the detailed structure of the nucleotide sequences of characteristic genes, on the other side. The specificity of the genomic signals prompted us to further develop a new approach to the nucleotide prediction experiments, complementary to the work reported in [5].
5 Prediction of Nucleotide Sequences

As shown in [5], the prediction of nucleotide sequences in a way similar to time series prediction is also an investigation of the theoretical possibility of ribosomes to use the redundancy in the genomic sequences to correct errors when synthesizing polypeptide chains based on the information brought by mRNA. The same question can be formulated about DNA replication and transcription to mRNA.

The description of the ribosomal machine holds that a human ribosome maintains the contact with the previous 35 nucleotides and with the next 3 nucleotides, when processing the current nucleotide. In order to avoid contradictions and conceptual difficulties, the model we adopted for checking the possibility to predict nucleotide sequences uses only the information from the previous nucleotides. The previous 35 samples of the nucleotide sequence, converted into complex numbers, i.e., 70 real and imaginary components, are applied at the input of the system, which is trained to produce an estimate of the current nucleotide.

The input was either applied directly to the NN input, or after preprocessing step to extract essential features, with a significant increase in system performance.

5.1 ANN Sequence Prediction

After trying a large variety of candidate systems, the MATLAB toolbox has been used to create a feed-forward neural network with 70 input neurons, one hidden-layer with a variable number of neurons and two output neurons. The first 35 inputs consist of the real parts of the previous 35 samples of the complex genomic signal, while the last 35 inputs are the corresponding imaginary parts. The two outputs are the real and imaginary parts of the predicted sample. The number of hidden neurons has been varied to find the optimal network architecture. The neurons were provided with the MATLAB toolbox activation function.

The total set of 7830 instances composing the supervised training set, i.e., pairs of input vectors and target outputs, has been randomly divided in three equally sized subsets: one for training, one for validation (early stopping to avoid overtraining), and one for testing (to evaluate network performance). The neural network has been trained by using the MATLAB toolbox training algorithm for its computational and performance advantages.

The performance measured by the sum of the squared errors is given in Fig. 6. The total data set has been divided in three subsets: one third for training, one third for validation and one third for testing. The first subset is the training set, which is used for computing the gradient and updating the network weights and biases. The second subset is the validation set. The test set is not used during the training, but it is used to compare different models.

A surprisingly high prediction accuracy of 96% has been obtained for a network with 105 neurons in the hidden layer. The marked decrease for large number of hidden neurons is caused perhaps due to overfitting.

![Fig.6](image)

**Fig.6.** Average correct predicted nucleotides for various numbers of hidden neurons of the ANN.

5.2 Feature Extraction

In this approach the prediction problem is converted into a classification problem. The extraction of relevant features allows us to make a code prediction on basis of the extracted information. Theoretically, it could be possible to have certain statistical common properties of various data sets that would facilitate code prediction. However, because of the large variability of the code, it is necessary to perform some form of feature extraction to overcome this variability or stick to the classic forecasting solutions as shown previously in 5.1.

Obtaining a feature set space can simplify the classification task. Relevant features lead to accurate classification.

The prediction task has been simplified by reducing the set of input sample values. The double nucleotide classes have been replaced by single nucleotide ones in the input and, correspondingly, in the output (prediction) data. Specifically, the reduction replacements W → A, Y → T, S → C and R → G has been operated. Taking into account the low frequency of double nucleotide classes in the current sequencing data, the effect of this reduction is marginal.

To reduce the dimension of the input vectors, the complex representation has also been replaced for the prediction task with the real number coding A → -3, T → -1, C → 3, G → 1.

Using features such as number of the zero crossings, slope sign changes, distance between vectors, mean, sum of squared elements, etc, we form a feature matrix in the way described in the following.

Data from 5 patients in each patient group are used for training and data from the remaining 5 patients are used for testing. Again, 35 data points represent the input vector, while the 36th sample is target value. For every nucleotide A,T,C,G we form two arrays:
- the network input matrix (if we have 80 occurrences of A, we shall have a 80x35 matrix divided into two 40x35 matrices – one for training and one for testing).
- the target vector (for these 80 occurrences of A, we shall have a vector of 80 ‘-3’ elements, and for the current coding we divide it in two 40 element vectors for training and testing).

Seven different kinds of features such as mean value of a data window, norm, number of zero crossings in data window, slope sign changes, median, distance between the vectors and sum of the squares in a data window, are fed into the inputs of the chosen neural network.

A feed-forward neural network with seven neurons in the first layer, 3 neurons in the hidden layer, and one output neuron has been used. The activation functions are MATLAB® \texttt{tansig} for the hidden neurons, and MATLAB® \texttt{pureline} for the output neuron.

Training was conducted using the data from half of the patients, and then we tested the network performance with data from the other half.

![Fig.7. Correct prediction for the feature extraction approach.](image)

As shown in Fig. 7, using the whole feature set we obtained a correct classification rate above 90% successful prediction for all the patient groups. Best classification results have been obtained for the Multiple Resistant group, while worst results was gained for the Drug Resistant group. Again, surprisingly high prediction accuracy has been obtained, as we did using ANN sequence prediction approach, and comparing both of them the feature extraction approach appeared to be better by 1%.

5 Conclusions

The correlation of genomic signals allows using techniques similar to those used in time series prediction [5, 16] to estimate the nucleotides in a sequence when knowing their preceding ones [5,14]. Such experiments have a biologic significance, as they explore the possibility and the efficiency of error correction in processes like replication, transcription and translation.

The efficiency of the prediction is improved by using a two step procedure, comprising feature extraction and nucleotide prediction. The paper presents results of experiments with this prediction model that show a remarkable good efficiency. This behavior is the direct effect of the regularities and correlations in the structure of the genomic sequences.

Acknowledgment

The work was supported in part by grants from the Ministry of Education and Research of Romania, the CEEX Program contract 50/2005 and CNCSIS 963-5/2007, the Ministry of Education and Science of Bulgaria, in the framework of the bilateral Scientific, Technological and Innovative Co-operation with Romania - project BPC-08, University “Politehnica” of Bucharest and by the Franqui Research Project ADSI 133 / 2007, ETRO, Vrije Universiteit Brussel, Belgium, the Technical University - Sofia, Research Project No:08004ni-8/2008.

References