Representation of DNA sequences with virtual potentials and their processing by (SEQREP) Kohonen self-organizing maps

João Aires-de-Sousa¹,* and Luisa Aires-de-Sousa²

¹Departamento de Química, CQFB, campus Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre, 2829-516 Monte de Caparica, and ²Clínica Universitária de Pediatria, Hospital de Santa Maria, Av. Prof. Egas Moniz, 1699 Lisboa Codex, Portugal

Received on May 29, 2002; revised on July 22, 2002; accepted on July 25, 2002

ABSTRACT
Motivation: We propose representing individual positions in DNA sequences by virtual potentials generated by other bases of the same sequence. This is a compact representation of the neighbourhood of a base. The distribution of the virtual potentials over the whole sequence can be used as a representation of the entire sequence (SEQREP code). It is a flexible code, with a length independent of the sequence size, does not require previous alignment, and is convenient for processing by neural networks or statistical techniques.

Results: To evaluate its biological significance, the SEQREP code was used for training Kohonen self-organizing maps (SOMs) in two applications: (a) detection of Alu sequences, and (b) classification of sequences encoding for HIV-1 envelope glycoprotein (env) into subtypes A-G. It was demonstrated that SOMs clustered sequences belonging to different classes into distinct regions. For independent test sets, very high rates of correct predictions were obtained (97% in the first application, 91% in the second). Possible areas of application of SEQREP codes include functional genomics, phylogenetic analysis, detection of repetitions, database retrieval, and automatic alignment.

Availability: Software for representing sequences by SEQREP code, and for training Kohonen SOMs is made freely available from http://www.dq.fct.unl.pt/qoa/jas/seqrep.

Contact: jas@fct.unl.pt

Supplementary Information: Supplementary material is available at http://www.dq.fct.unl.pt/qoa/jas/seqrep/bioinf2002.

INTRODUCTION

Simple representations of entire DNA sequences by a fixed number of variables, independent of the sequence size and alignment, are highly desired for fast processing of DNA data using neural networks, statistical procedures or pattern recognition techniques. The simplest representation scheme is base composition, which can be used in trivial applications. More popular and sophisticated approaches consist of counting specific words or n-grams with or without gaps in a given sequence. These representations can give rise to a large number of variables depending on the size of the alphabet, the size of the words, and the gap length. Data reduction is usually necessary as a second step (Wu and Shivakumar, 1994; Wu, 1998).

Chaos game representation (CGR) (Jeffrey, 1990) is an iterative mapping technique that assigns to every position of a sequence a set of coordinates in a continuous space. In the case of DNA sequences, the coordinates of a base at a specific position depend on the bases that precede it, and are thus descriptors of the preceding window. Distribution of the coordinates has been proposed as a representation of the entire sequence and a framework for comparing different sequences (Almeida et al., 2001; Almeida and Vinga, 2002). Working in a 2D space, the plot of CGR can be interpreted as a graph, while 3D CGRs are formally comparable to 3D molecular structures. This was the rationale for applying well-established techniques from mathematical chemistry to generate descriptors of sequences, in the very same way they are applied to calculate molecular descriptors that characterize molecular graphs (representing the connectivity of small molecules), or molecular descriptors of 3D structures (Randic and Vracko, 2000; Randic et al., 2000; Randic and Basak, 2001).

In this paper we introduce the concept of virtual potential generated at an individual sequence position by other bases in the same sequence. The distribution (a histogram) of the virtual potentials over the whole sequence, the SEQREP code (SEQuence REpresentation using virtual Potentials), is used as a representation of the entire sequence. It represents a sequence by a code...
with a fixed number of variables (fixed-length code) independently of the sequence size, and it does not require previous alignment.

To evaluate the biological significance of SEQREP, the code was used for training Kohonen self-organizing maps (SOMs) in two applications: (a) detection of Alu sequences, and (b) classification of sequences encoding for HIV-1 envelope glycoprotein (env) into subtypes A–G.

Kohonen self-organizing maps (SOMs), or Kohonen neural networks (Kohonen, 1988), were employed because they learn by unsupervised training. This means that the network self-organizes during the training, and distributes the sequences in a map without knowing to which classes they belong. Only at the end of the training are the sequences labelled with their known classes. The neurons are then assigned to classes according to the sequences that excited them, and by inspection of the resulting map it is possible to verify if clustering of the different classes occurred. Because the self-organization of the networks is based on the similarity between SEQREP codes, this technique is particularly useful to investigate whether there are any correlations between SEQREP codes and the biological properties of the sequences they represent.

In the first example, SOMs were trained with sequences from two classes, Alu sequences and non-Alu sequences, to verify if separation between the two types was obtained, and if the trained SOMs could be used for automatic detection of Alu sequences. Alu sequences are the largest family of short interspersed repetitive elements (SINEs) in human genome accounting for more than 5% of it. They are about 300 bp long and are distributed randomly throughout the genome (Rowold and Herrera, 2000). Many genes contain one or more Alu sequences within one or more of their larger introns, but they are rarely present in protein-coding regions of mature RNA (Sukarova et al., 2001). Alu repeats affect the genome in several ways, causing insertion mutations, recombination between elements, gene conversion and alterations in gene expression (Batzer and Deininger, 2002; Li and Schmid, 2001). Computational tools for detection of these sequences are thus of importance for automatic annotation of DNA sequences.

In the second example, a SOM was fed with sequences coding for HIV-1 envelope glycoprotein of isolates belonging to different subtypes. Within the major group of HIV-1 (group M), different genetic subtypes (clades) have been identified by phylogenetic analyses of the nucleotide sequences of the envelope (env) and core (gag) genes. The different subtypes are unequally distributed by geographical regions, which provides useful information for epidemiological analysis (Osmanov et al., 2002). The envelope protein on the viral surface is of potential value in the development of a vaccine as the outer gp120 segment carries a site that recognizes and binds to the CD4 molecule on macrophages, T lymphocytes, and other HIV target cells. In this context, genetic variability of the env gene is a central issue even if the genetic subtypes are not necessarily equivalent to antigenic or immunological subtypes (Alaeus, 2000; Nyambi et al., 2000; Lee et al., 2001; McMichael et al., 2002). In this application, SOMs were applied for classification of env gene sequences into subtypes.

MATERIALS AND METHODS

SEQuence REpresentation using virtual Potentials (SEQREP)

The SEQREP encoding method is based on the idea of a virtual potential generated at a given position of a DNA sequence by other bases of the same sequence. The virtual potential generated at position Y by the base at position X, VP\(_{X \rightarrow Y}\), is defined by Equation (1)

\[
VP_{X \rightarrow Y} = 1/d_{XY}
\]

where \(d_{XY}\) is the distance between positions X and Y. In this investigation only virtual potentials generated by bases up to 10 positions away were considered (\(d_{XY} \leq 10\)), and the potentials at a given position Y are considered to be generated only by bases that appear before that position (\(Y > X\)).

At every position in the sequence, four virtual potentials are defined: potentials generated by A, C, G, and T bases. The potential generated by A bases at a given position is the sum of all the potentials at that position generated by A bases (Figure 1). The same is done for potentials generated by C, G, and T bases (Figure 2). In the whole sequence, the virtual potentials can be divided into 16 types—the potentials at positions occupied by bases A, C, G, and T generated by bases A, C, G, and T. We represent by A→G, the potentials generated at positions with G bases by A

Fig. 1. Calculation of the virtual potential at a position (indicated with the arrow) generated by A bases.

The SEQREP code consists of the histograms of the 16 types of virtual potentials. In this investigation, the histogram of a virtual potential contains 30 classes of width 0.1, covering the range of virtual potentials between

![Virtual potential generation](image)
Fig. 2. Virtual potentials (VP) generated by A, C, G, and T bases at each of the last six positions of the sequence used as an example in Figure 1.

Fig. 3. Histogram of the virtual potentials at positions occupied by T bases, generated by A bases, for a given sequence of bases.

0 and 3 (the maximum possible virtual potential is $\approx 2.929$). This means that if all the 16 types of VP are used, the SEQREP code contains $16 \times 30 = 480$ values. If only a selection of potentials is used, the code will be shorter (a multiple of 30). For example, considering only VP at positions occupied by T bases generated by A bases ($A \rightarrow T$), the length of the SEQREP code will be 30 (Figure 3).

The SEQREP code for a sequence can be seen as a vector, each component being a value from the histogram. Such vectors were normalized (by dividing each component by the magnitude of the vector) and used as input to Kohonen self-organizing maps.

Kohonen self-organizing maps

Kohonen self-organizing maps (SOM) can be used for automatic classification of objects, represented by a vector. Here we used SOMs to classify DNA sequences represented by their SEQREP codes. SOMs learn by unsupervised training, revealing similarities between objects (SEQREP codes), this being the reason why it was chosen to evaluate the biological significance of the new code. An additional advantage is the speed of training when compared to other classification algorithms.

A Kohonen SOM consists of a grid of so-called neurons, each containing as many elements (weights) as there are input variables (Figure 4). In the investigations here described, the input variables are the values of the SEQREP codes. Before the training starts, the weights take random values.

During the training, each individual DNA sequence is mapped into the neuron that contains the most similar weights compared to its SEQREP code. This is the central neuron, or winning neuron. It is said that the winning neuron was excited by the sequence, and its weights are then adjusted to make them even more similar to the SEQREP code of the presented sequence. Not only the winning neuron has its weights adjusted, but also the neurons in its neighbourhood. The extent of adjustment depends, however, on the topological distance to the winning neuron—the closer a neuron is to the central neuron the larger is the adjustment of its weights. The objects of the training set are iteratively fed to the network, the weights corrected, and the training is stopped when a pre-defined criterion is met (e.g. a certain number of cycles or a measure of stability).

After training, all the objects of the training set are mapped, and the neurons are assigned to the classes of the sequences that excited them. If sequences belonging to different classes excited the same neuron, this is assigned to the class that excited it more times. A trained Kohonen neural network will reveal similarities in the objects of a data set in the sense that similar objects (similar SEQREP codes) are mapped into the same or closely adjacent neurons. When a new DNA sequence is presented to the trained network, it can be classified according to the class of the winning neuron, or the most frequent class of the adjacent neurons.

Selection of potentials

The ability of each of the 16 virtual potentials to discriminate between sequences of the different classes was tested separately. A Kohonen neural network was trained with the sequences of the training set, encoded by the SEQREP method using only one of the 16 virtual potentials. After the training, the degree of separation between the different classes was measured by a clustering factor, which was calculated by counting, for every neuron, the number of adjacent neurons assigned to its class, and subtracting the number of adjacent neurons assigned to a different class. This summation (over all neurons) was then divided by the total number of neurons to obtain the clustering factor. In order to obtain more reliable clustering factors, they were averaged over ten runs, and then averaged with the minimum clustering factor obtained in the ten runs. This procedure was repeated for all the 16 virtual potentials, and the 16 potentials were ranked according to their clustering ability.
Respecting this ranking, an increasing number of virtual potentials (from 1 to 16) were then selected for training SOMs, and the selections were evaluated by the clustering factor: one network was trained using the potential with the best discriminating ability, then a second network was trained using the two best potentials, then a third using the three best potentials and so on, until the last network was trained with all the 16 potentials. The clustering factor typically peaked at some point and then tended to stabilize slightly below the maximum. The set of virtual potentials giving rise to the peak were further used as the selected potentials.

**Data sets**

For the first application, a training set was built containing 327 Alu sequences and 327 non-Alu sequences. The Alu sequences were downloaded from NCBI ftp site (ftp.ncbi.nih.gov) in the directory /blast/db/ (also available from http://www.infobiogen.fr/services/GoldenPath/root/xrefs/LISTS/DB/alu/ALUfasta/). The non-Alu sequences were randomly generated by the following procedure. The probability of generating a given base was set equal to the proportion of that base in the data set of Alu sequences. The length of the random sequences was allowed to vary within the same range of the sequences in the Alu data set, around their average value. The statistics concerning base composition and sequence size are specified in the file HELP.Alu.325 in the above-cited Infobiogen web site. The test set consists of 136 Alu sequences and 141 randomly generated sequences. 125 Alu sequences out of the 136 were downloaded from EMBL ftp site (ftp.ebi.ac.uk/pub/databases/alu) and the other 11 Alu sequences were retrieved from http://alces.med.umn.edu/tables/BR3.5. The 141 random sequences were generated as for the training set.

For an alternative experiment, in which the non-Alu sequences were taken from real genome data, a collection of 463 non-Alu sequences from *Homo sapiens* was randomly selected from fragments of chromosomes 15, 19, and 20. The data was retrieved from ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/. The sequences were selected in order that their lengths vary in the same range and have the same average as the Alu sequences in the training set. They were subsequently screened by RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html) and by CENSOR (Jurka et al., 1996, http://www.girinst.org/) to ensure that no Alu sequence was detected. 327 sequences out of the 463 were used for training and the remaining 136 for testing.

For the study of *env* HIV-1 subtypes, we used a data set with 31 sequences of subtype A, 124 of subtype B, 24 of subtype C, 11 of subtype D, five of subtype E, four of subtype F, and three of subtype G. The sequences were published by Op de Coul et al. (2001) and are freely available from PubMed (http://www.ncbi.nlm.nih.gov). This data set was randomly divided into a training set containing 15 sequences of subtype A, 20 of subtype B, 15 of subtype C, eight of subtype D, three of subtype E, two of subtype F, and two of subtype G, and a test set with the remaining sequences.

**Software**

A Java application (SEQREP1) was developed for (a) computation of SEQREP codes from sequences in FASTA format, (b) training Kohonen SOMs with the SEQREP codes, and (c) application of the knowledge extracted during the training to new sequences. SEQREP1 was built with Java code specifically written for generating SEQREP representations, and with Java code already used in JATOON applets (Aires-de-Sousa, 2002) for general
application of neural networks. SEQREP1 is made freely available on the Internet and can be downloaded from http://www.dq.fct.unl.pt/qoa/jas/seqrep/. As an indication, running SEQREP1 over 15 epochs to train a $15 \times 15$ Kohonen neural network with 654 sequences represented by 6 virtual potentials, using an initial learning span of seven neurons, took 45 s on a PC with an AMD Thunderbird 1 GHz CPU and Microsoft Windows 2000 Professional as the operating system.

RESULTS AND DISCUSSION

Identification of Alu sequences

The ability of each of the 16 virtual potentials to discriminate between Alu and randomly generated sequences was tested separately, with the 654 sequences of the training set, and $16 \times 10 \times 10$ Kohonen neural networks, each network using the histograms of only one of the 16 types of potential. The following ranking of potentials was obtained in decreasing order of clustering ability:

$$\begin{align*}
A \to A, & \quad G \to A, \\
A \to G, & \quad A \to T, \\
G \to G, & \quad C \to T, \\
C \to G, & \quad C \to T, \\
G \to T, & \quad T \to G, \\
C \to A, & \quad G \to C, \\
T \to C, & \quad T \to A, \\
T \to T.
\end{align*}$$

The obtained clustering factors varied between 5.6 (for $A \to A$) and 0.52 (for $T \to T$). The fact that A bases are involved in the four best potentials agrees with the recognition that Alu sequences have typically an A-tail.

After training networks with increasing numbers of potentials (from 1 to 16), it was found that the clustering factor peaked at the set with seven potentials. It should be noted that, in this last series of 16 experiments, the clustering factor varied only slightly, between 5.9 and 6.3. On the basis of this analysis, it was decided to select the first seven potentials ($A \to A, G \to A, A \to G, A \to T, G \to G, C \to T, A \to C$) to train a final network. A $15 \times 15$ Kohonen neural network was trained with the 654 sequences, the initial learning span was set to seven neurons, and the training was stopped after 15 presentations of the entire set to the network (15 epochs). All the objects of the training set were then submitted one by one to the trained network, and the neurons were assigned to classes according to the sequences that excited them. In Figure 5 the resulting map is represented. It should be noted that the surface of the Kohonen SOM has toroidal topology, which means that neurons in the top row are considered to be adjacent to those at the bottom, and the neurons in the first column are considered to be adjacent to those in the last. It is thus clear that the Alu sequences were mapped into a region that is distinct from the one where the randomly generated sequences were mapped. Only three neurons, placed near the border between the two regions, could not be assigned due to conflicts (they were excited by the same number of Alu and non-Alu sequences).

In order to prove that the knowledge extracted by the SOM during the training could be applied to new situations, the sequences of the test set were encoded in the same way and presented to the trained neural network. Each sequence was classified according to the class of the neuron that it excited. If no class had been assigned to the winning neuron, counting the occurrence of each class in the adjacent neurons did the classification, and if a winner class could not be found, the sequence was not classified (undecided). The sequence was also undecided if the excited neuron had been labelled as a conflict. The global results are shown in Table 1.

![Fig. 5. Top view of the Kohonen toroidal surface after training with Alu sequences and randomly generated sequences. The neurons were assigned to classes on the basis of the sequences that excited them.](https://academic.oup.com/bioinformatics/article-abstract/19/1/30/316857/)

<table>
<thead>
<tr>
<th>Class</th>
<th>No. of examples</th>
<th>Correct</th>
<th>Undecided</th>
<th>Wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alu</td>
<td>136</td>
<td>135 (99.3%)</td>
<td>1 (0.7%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Random</td>
<td>141</td>
<td>133 (94.3%)</td>
<td>3 (2.1%)</td>
<td>5 (3.5%)</td>
</tr>
<tr>
<td>Total</td>
<td>277</td>
<td>268 (96.8%)</td>
<td>4 (1.4%)</td>
<td>5 (1.8%)</td>
</tr>
</tbody>
</table>
classes, which suggests that the position of the winning neuron within a class region can be interpreted as a measure of reliability of the prediction. These results compare extremely well with those obtained, for the same data sets, by Ma et al. (Ma et al., 1998) using a feed-forward neural network (using supervised learning) combined with a pattern matching technique (false negative error rate 0%, false positive error rate 3%), or using BLAST (false negative error rate 0%, false positive error rate 9%). In those experiments, the same Alu data sets were used as in ours, while the non-Alu sequences were generated in an analogous way.

The method was further tested using real non-Alu sequences from the human genome instead of randomly generated sequences, and the results are presented below. The sequences were checked by RepeatMasker and CENSOR to guarantee that no Alu sequence was detected. The fact that no Alu sequence was detected does not absolutely guarantee that no Alu sequence is present, as those methods should not be 100% accurate. However, this experiment gives an indication about the robustness of our method when applied to real data.

The following five potentials were selected by the previously described procedure: A→A, G→A, A→G, T→C, G→C, in decreasing order of discriminating ability. As can be seen, the best three potentials are the same as for the experiment with randomly generated sequences. A 15×15 Kohonen neural network was trained as in the first experiment. The map resulting from feeding the training set to the trained network is shown in Figure S1 of Supplementary Information. Excellent clustering was achieved, as well as very good predictions for the test set: 98.5% of Alu sequences were detected, one (0.7%) was undecided, and one (0.7%) was wrongly classified as non-Alu, while 94.9% of non-Alu sequences were correctly classified, 4.4% were wrongly classified, and one was undecided.

The fact that our results were obtained by unsupervised training shows that a similarity in SEQREP codes corresponds, in this application, with a similarity of biologically relevant features of DNA sequences.

**Phylogenetic classification of DNA sequences encoding for the HIV-1 envelope (env) gene**

The sequences of the training set were represented by their SEQREP codes calculated using the virtual potentials A→T, A→C, C→A, T→C, C→G, C→T, T→A, G→A, G→C, G→T, A→A (these potentials were selected by the procedure described in Materials and Methods, as for the first application). They were submitted to a 14×14 Kohonen SOM with toroidal topology. An initial learning span of seven neurons was chosen, and the training was stopped after 50 epochs. At the end of the training, all the objects of the training set were submitted to the network and the neurons were assigned to classes according to the sequences that excited them. The result is shown in Figure S2 of Supplementary Information. The subtypes were globally mapped into distinct regions of the surface, which means that the SEQREP codes allow to distinguish between the different HIV-1 subtypes, and suggests that the trained network can be used for classifying unknown sequences. Only the region corresponding to subtype E is not clearly separated from the region of subtype G, and a conflict exists between a sequence of class F and another of class B. These problems are probably due to insufficient number of examples for classes E, F and G (3, 2, and 2 respectively). It must be noted again that the training of the network was unsupervised, i.e. the network had no access to the subtype corresponding to each sequence before the end of the training. The only stage of the process that was supervised was the selection of potentials.

The sequences of the test set were encoded in the same way and were submitted to the trained network. These sequences had not been used neither for the selection of potentials nor for training the network. 88% of the cases were correctly classified, 9% were undecided, and only 3% were wrongly classified. The detailed results are presented in Table S1 of Supplementary Information. Inspection of the neurons excited by the test set revealed that the undecided and wrongly classified examples were mapped into neurons at the borders of different regions. Furthermore they were mapped close to the correct regions, with the exception of the isolate of subtype G and one isolate of subtype B. The excellent results obtained for the test set reinforce the conclusion that a correlation exists between SEQREP codes of env genes and HIV-1 subtypes.

This application also illustrates how a Kohonen neural network can be used to extract phylogenetic information both automatically and by visual inspection.

**CONCLUSIONS**

The concept of virtual potential generated at an individual sequence position by other bases of the same sequence provides a method to represent individual residues by real numbers, and to represent entire sequences by a fixed-length code (SEQREP code). It is demonstrated that SEQREP codes can be correlated with relevant biological properties of DNA sequences in two applications—identification of Alu sequences and classification of HIV-1 env genes according to subtypes. In both examples, Kohonen self-organizing maps learned by unsupervised training to classify sequences represented by their SEQREP codes. The knowledge acquired during the training allowed making accurate predictions for the test sets. Kohonen SOMs revealed that generally a similarity between SEQREP codes of DNA sequences corresponds
to a similarity of biologically relevant features of the sequences, in the studied examples.

A virtual potential can be interpreted as a measure of the abundance and proximity of a given type of base in the preceding neighbourhood of an individual base. It is thus a compact representation of a neighbourhood, which relies upon a different approach from the methods commonly used today. SEQREP codes are an alternative method for the representation of biological sequences, with possible application in many areas of bioinformatics such as functional genomics, phylogenetic analysis, detection of repetitions, or retrieval of sequences from databases. Virtual potentials can also be used as guides for automatic alignment of sequences.

After this exploratory investigation, SEQREP codes must be processed by algorithms employing supervised learning for classification tasks, in order that the predictions can be adjusted to the experimental observations. Furthermore, SEQREP codes must now be extensively tested for different purposes to precisely determine their spectrum of application.

ACKNOWLEDGEMENTS

J.A.S. acknowledges Fundação para a Ciência e a Tecnologia (Lisboa, Portugal) for a post-doctoral fellowship.

REFERENCES


