BioSystems 142-143 (2016) 32–42

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

A new approach to the automatic identification of organism evolution using neural networks

Andrzej Kasperski[a,*], Renata Kasperska[b]

[a] Faculty of Biological Sciences, Department of Biotechnology, University of Zielona Gora, ul. Szafrana 1, 65-516 Zielona Gora, Poland
[b] Institute of Occupational Safety Engineering and Work Science, University of Zielona Gora, ul. Szafrana 4, 65-516 Zielona Gora, Poland

ARTICLE INFO

Article history:
Received 30 September 2015
Received in revised form 20 January 2016
Accepted 8 March 2016
Available online 11 March 2016

Keywords:
Computational biology
Evolution
Neural network
Phylogenetics
Programming

ABSTRACT

Automatic identification of organism evolution still remains a challenging task, which is especially exiting, when the evolution of human is considered. The main aim of this work is to present a new idea to allow organism evolution analysis using neural networks. Here we show that it is possible to identify evolution of any organisms in a fully automatic way using the designed EvolutionXXI program, which contains implemented neural network. The neural network has been taught using cytochrome b sequences of selected organisms. Then, analyses have been carried out for the various exemplary organisms in order to demonstrate capabilities of the EvolutionXXI program. It is shown that the presented idea allows supporting existing hypotheses, concerning evolutionary relationships between selected organisms, among others, Sirenia and elephants, hippopotami and whales, scorpions and spiders, dolphins and whales. Moreover, primate (including human), tree shrew and yeast evolution has been reconstructed.

1. Introduction

Fully automated identification of organism evolution can be considered as a dream for researchers and sometimes, taking into account the complexity of this task, this aim can be treated as the stuff of science fiction (MacLeod, 2007). In the analysis of organism evolution and their genetic variability, the methods based on Neighbor Joining (NJ), Maximum Parsimony (MP), Maximum Likelihood (ML), Bayesian Inference (BI), supported by, for example, the dot matrix method, are usually used (Finstermeier et al., 2013; Kasperski and Kasperska, 2012, 2014). During these analyses, the number of generated phylogenetics trees which should be considered depends substantially on the number of analyzed organisms. Theoretically, establishing a real conclusion requires analysis of each of the possible trees. This task can be impossible to perform for a larger number of taxa, for example, the number of possible rooted trees for 50 taxa is bigger than the number of atoms in the universe. For this reason, the reconstruction of the real organism evolution is often impossible when trying to determine the best phylogenetics trees. It makes it necessary to seek new methods, which will allow for more reliable determination of organism evolution and their genetic variability. One of the computational tools, that can be used in solving complex real-world problems, are artificial neural networks (ANNs) (Basheer and Hajmeer, 2000). Neural computation can be used in various fields, due to nonlinearity, high parallelism, robustness, fault and failure tolerance, learning, ability to handle imprecise and fuzzy information, and their capability to generalize (Jain et al., 1996). ANN, as a programming method based on a mathematical approximation of the functioning of human brain cells, can be seen as a set of interconnected nodes implementing a mapping function from an input space to one of several output categories. By possibility of a learning and outcome prediction, ANNs can replace traditional statistical techniques in modeling and classification of selected problems (Ahmed, 2005; Hannachi et al., 2003). In biology, ANNs are considered as holding great promise in helping with advanced understanding of biological phenomena and biosystems. For example, the ability of neural networks to learn complex functions from large amounts of data without the need for predetermined models makes them a good tool for a protein structure prediction. ANNs can also support the acquiring of accurate knowledge of quantitative structure-activity relationship (Jalali-Heravi et al., 2011). Moreover, neural networks can be applied to: pattern recognition of DNA, RNA, gene identification, sequence classification, analysis of electron microscopy images of biological macromolecules, prediction of microbial growth, identification of microorganisms and molecules, interpreting pyrolysis

* Corresponding author.
E-mail address: A.Kasperski@wnb.uz.zgora.pl (A. Kasperski).

http://dx.doi.org/10.1016/j.biosystems.2016.03.005
0303-2647/

mass spectrometry, high-performance liquid chromatography, and gas chromatography data (Basheer and Hajmeer, 2000; Pascual-Montano et al., 2003; Wu, 1997). Neural networks can be used to support recognizing and classifying of human and animal diseases, for example, ANNs are able to predict a biopsy result with 87% accuracy and tumor recurrence with 90% overall accuracy, they may also be useful in decision making regarding prostate cancer patients and classifying animal fibers (She et al., 2002; Snow et al., 1994). ANNs can be used to categorize a selected set of bird species based on vocalization, to identify primate vocalizations and analyze their communication (Abewardana and Sonnadara, 2012; Pozzi, 2010).

Many different loci on the mitochondrial genome such as 12S rRNA, 16S rRNA, COI, COII, and others have been used for organisms identification (Alessandrini et al., 2008; Ascunce et al., 2003; Balitzki-Korte et al., 2005; Borisenko et al., 2008; Dubey et al., 2009; Hebert et al., 2003; Melton and Holland, 2007; Mitani et al., 2009; Roe and Sperling, 2007). However, cytochrome b has been the main locus used in organisms discrimination (Irwin et al., 1991; Kocher et al., 1989; Tobe et al., 2010). Cytochrome b phylogenies can help in the genus assignment of newly described organisms (Giao et al., 1998). The sequence variability of cytochrome b makes it most useful for the comparison of organisms in the same genus or the same family (Castresana, 2001). Cytochrome b can be used for analysis of phylogenetic relationships within mammals (Irwin et al., 1991; Johns and Avise 1998; Meyer 1994). For example, the use of cytochrome b has led to the proposition of new classification schemes that better reflected the phylogenetic relationships between the true seals, South American echimyid rodents, African mole-rats, delphinid cetaceans, family Bovidae (Arnason et al., 1995; Faulkes et al., 1997; Lara et al., 1996; LeDuc et al., 1999; Matthee and Robinson, 1999). Presently, complete mitochondrial DNA (mtDNA) genomes, including cytochrome b sequences, have been sequenced from more than a dozen remains of ancient humans and archaic hominins (including *Homo heidelbergensis*, the Neanderthals and Denisovans) (Gansauge and Meyer, 2014). As a result, various studies on primate (including human) phylogeny have been presented, but the primate molecular phylogeny is still not completely resolved (Chatterjee et al., 2009; Fabre et al., 2009; Perelman et al., 2011; Springer et al., 2012; Finstermeier et al., 2013).

This work sets out to use artificial neural networks taught by cytochrome b sequences to determine an evolutionary relationship between organisms. This new approach consists of the recognition of evolutionary relationships based on a neural network trained by patterns. This idea is contrary to the other known methods, such as the NJ, MP, ML, BI methods, in which evolutionary relationships are calculated using mathematical formulas. These formulas are implemented within 'never ending' loops (for bigger number of taxa), which have to be terminated using various heuristics (Hall, 2011). All these methods need 'a brutal power of calculation', and it should be added that the ML method is more complex than the NJ and MP methods, and the BI method is even much more complex than the ML method. Additionally, before using the NJ, MP, ML, BI methods, an alignment of used sequences has to be made and correct parameters have to be set (Tamura et al., 2013). All this results in that the NJ, MP, ML, BI methods look 'computationally heavy' in comparison to a recognition of evolutionary relationships using an artificial neural network. After a long process of neural network training, a recognition of evolutionary relationships occurs very quickly, what causes that this method looks 'computationally light' in comparison to the NJ, NP, ML, BI methods. Reconstructions of primate, tree shrew and yeast evolution are presented as examples of using this idea to reach a recognition of evolutionary relationships. The other exemplary results of a recognition of evolutionary relationship of organisms are presented in Section 3 to prove the propriety of the idea.

2. Materials and methods

Cytochrome b amino-acid sequences selected for this study were taken from the protein databases NCBI and Protein BLAST and are available at:

http://www.uz.zgora.pl/~akaspers/BioEvolution/sequences.pdf.

2.1. Methods

All calculations presented in this article were made using the EvolutionXXI program (http://www.uz.zgora.pl/~akaspers/BioEvolution/readme.pdf), which contains implemented neural network. The EvolutionXXI program was written by the authors in Java using the Joone framework. The EvolutionXXI program running on any platforms with installed Java Virtual Machine (JVM) may be freely obtained from the authors.

2.2. Design of the artificial neural network

The artificial neural network has been designed as a full synapse three layer neural network with sigmoid transfer function $y = 1/(1 + \exp(-x))$ (Heaton, 2005). It means that each node of the input layer is connected to each node of the hidden layer through an abstract connection called the synapse with an associative weight, and in the same way each node of the hidden layer is connected to each node of the output layer (Haykin, 2009). It is known that a neural network containing at least three layers can approximate an arbitrary nonlinear function after establishing appropriate internal weights (Basheer and Hajmeer, 2000). This means that most practical problems (consisting in the function approximation) can be solved with a desired accuracy using the tree layer neural network. In the designed neural network the number of neurons in the input layer depends on the number of amino-acids (AA) in the sequences used for teaching and recognizing. Because the number of amino-acids in the cytochrome b sequences is not bigger than 400 AA for almost all organisms, in this work it is assumed that the length of sequences in the input of the neural network is equal to 400 AA. If the number of amino-acids in the sequence is less than 400, then this sequence is aligned by addition of the "character at the end to lengthen it to 400 AA. After the alignment, each sequence is converted to the binary form by each character changing to a five-positional binary number, i.e. "is converted to "00000", "A" (Alanine) is converted to "00001", "B" (Aspartate or Asparagine) is converted to "00010", "C" (Cysteine) is converted to "00011", "D" (Aspartic acid) is converted to "00100" and so on. After converting the new binary form of sequence (which is entered into the neural network input layer) has a length of 2000. For this reason the number of neurons in the neural network input layer is equal to $n = 2000$. The number of neurons in the output layer is equal to the number of organisms used for ANN teaching, i.e. in this work it is equal to $k = 32$. The number of neurons in the hidden layer is calculated by the geometric pyramid rule proposed by Masters (Masters, 1993), so for a three layer network with n input and k output neurons, the hidden layer has got $\mathrm{sqrt}(n \times k)$ neurons. In this work the number of neurons in the hidden layer is calculated as $m = \mathrm{round}(\mathrm{sqrt}(n \times k))$, that gives $m = 253$.

2.3. Teaching of the neural network

The neural network was taught by 32 successive cytochrome b sequences (Table 1), until it converged to the desired mapping. During teaching the ANN built a predictive model that reflected a minimization in a general error (calculated as RMSE, i.e. Root Mean Squared Error) of comparison of the network's prediction with known 32 teaching outputs. A teaching output for the