Bioinformation
www.bioinformation.net
Volume 5
Issue 2
open access
Hypothesis

Early diagnosis of systemic lupus erythmatosus using ANN models of dsDNA binding antibody sequence data

Mohamad Hasan Bahari[1], Mahmoud Mahmoudi[2], Asad Azemi[3], Mir Mojtaba Mirsalehi[4], Morteza Khademi[5]

[1] Katholieke Universiteit Leuven/ESAT, Leuven, Belgium. Email: m.h.bahari@ieee.org; [2] Immunology Research Center, Mashhad University of Medical Science, Mashhad, Iran. Email: mahmoudim@mums.ac.ir; [3] Penn State University /Engineering Department, Delaware, USA. Email: Azemi@psu.edu; [4] Ferdowsi University of Mashhad/Electrical Engineering Department, Mashhad, Iran. Email: mirsalehi@um.ac.ir; [5] Ferdowsi University of Mashhad/Electrical Engineering Department, Mashhad, Iran. Email: khademi@um.ac.ir.

Received March 28, 2010; revised April 30, 2010; accepted June 08, 2010; published July 06, 2010

Abstract:
In this paper a new method based on artificial neural networks (ANN), is introduced for identifying pathogenic antibodies in Systemic Lupus Erythmatosus (SLE). dsDNA binding antibodies have been implicated in the pathogenesis of this autoimmune disease. In order to identify these dsDNA binding antibodies, the protein sequences of 42 dsDNA binding and 608 non-dsDNA binding antibodies were extracted from Kabat database and encoded using a physicochemical property of their amino acids namely Hydrophilicity. Encoded antibodies were used as the training patterns of a general regression neural network (GRNN). Simulation results show that the accuracy of proposed method in recognizing dsDNA binding antibodies is 83.2%. We have also investigated the roles of the light and heavy chains of anti-dsDNA antibodies in binding to DNA. Simulation results concur with the published experimental findings that in binding to DNA, the heavy chain of anti-dsDNA is more important than their light chain.

Keywords: Anti-dsDNA; Antibody; General Regression Neural Network (GRNN); Systemic Lupus Erythematosus

Background:
Systemic Lupus Erythematosus (SLE or 'lupus') is a major autoimmune rheumatic disease where autoantibodies frequently target against intracellular antigens of the cell nucleus (double and single stranded DNA). According to the result of an experiment on more than five million U.S. Armed Forces personnel, in 115 of the 130 patients with SLE (88 percent), at least one SLE autoantibody tested was present 3.3 years before the diagnosis **[1].** This fact suggests that SLE can be predicted several years before the diagnosis.

In patients with SLE, a wide variety of antibodies against nuclear antigens can be found, including antibodies to nucleic acids, histones and non-histone nuclear proteins. However, anti-dsDNA antibodies are considered the most important autoantiboies in SLE. These antibodies were identified, for the first time, nearly forty years ago in the serum of patients with SLE **[2]**. Based on similar findings, many researchers have concluded that there is a close relationship between disease activity and levels of anti-dsDNA antibodies. In other words, dsDNA binding antibodies have been implicated in the pathogenesis of this autoimmune disease **[3].**

Anti-dsDNA antibodies seem highly likely to cause tissue damage in patients with SLE. They are present in the serum of patients several years before its diagnosis. Therefore, their identification and analysis may be useful in the identification of patients who would benefit from early diagnosis, as well as patients who do not require further evaluation. In addition, a powerful method of their analysis may lead to design of new drugs that interfere with antibody–DNA interactions, which might have therapeutic applications.

Recently, bioinformatics approaches, including statistical techniques and intelligent systems, have tried to obtain a better grasp of the DNA-binding structures and utilize them in an early prediction of autoimmune diseases such as SLE. Although these kinds of works have been able to explain many biological phenomena, lead to the development of mathematical models for prediction and lead to reduction in the cost of experimental research, there is still no accurate mathematical method for identifying dsDNA binding antibodies **[4].** We can trace the roots of this difficulty in the long amino acid sequence of antibodies, lack of suitable mathematical tools, and more importantly the vast diversity of antibodies.

In this paper, we extended the computer-assisted autoantibody analysis methods by proposing a new approach with the capability of identifying dsDNA binding antibodies. The proposed identification system is constructed using a General Regression Neural Network (GRNN).

In the proposed method, we encode the amino acid sequence of antibodies using Hydrophilicity values of amino acids. These signals are fed to a GRNN. In order to better interpret the role of different parts of the anti-dsDNA antibodies, we have investigated the roles of light and heavy chains and their bindings to DNA. In general, the proposed method provides an efficient and accurate method to identify the dsDNA.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 5(2): 58-61 (2010)

© 2010 Biomedical Informatics

Bioinformation **Volume 5**
www.bioinformation.net Issue 2 **Hypothesis**
open access

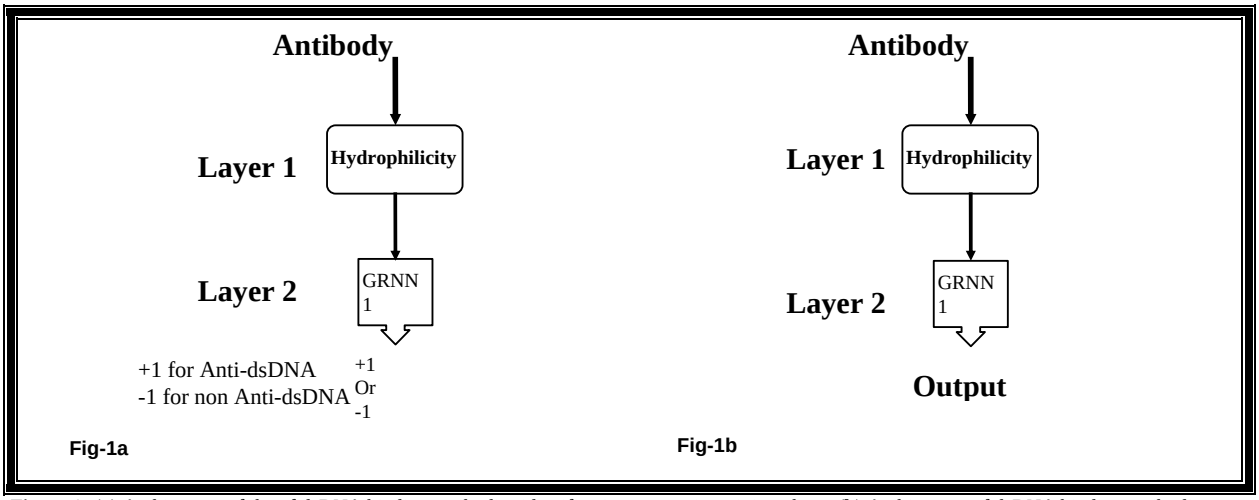

Figure 1: (a) Architecture of the of dsDNA binding antibodies identification system in training phase. **(b)** Architecture of dsDNA binding antibodies identification system in testing phase

Methodology:

Dataset

In order to use ANN to identify dsDNA binding antibodies the training set needs to include a set of anti-dsDNA, as the positive set, and a set of other non dsDNA binding antibodies, as the negative set. Our training data set was constructed using the Kabat database. Kabat antibody sequence database contains primary structure and sequence information on antibodies and other proteins of immunological interest **[5].** There are 42 sequences of anti-dsDNA (both heavy and light chains) available in the Kabat database, which we used as positive patterns, and we took 608 of the other non dsDNA binding antibody sequences (both heavy and light chains) as negative patterns.

Encoding:

As we know, Immunoglobulins are heavy plasma proteins, composed of 20 different amino acids. From a one-dimensional point of view, a protein sequence contains characters of these 20 amino acids. In other words, function of a protein depends on its specific amino acid sequence called the primary sequence structure **[6]-[8].**

An important issue in applying computer-based systems to identify dsDNA binding antibodies is how to encode protein sequences; i.e., how to represent the protein sequences as the input of a system. This is crucial to the success of neural network learning process. Several different methods have been used for encoding proteins. For example, Wu **[9]** has utilized the 2-gram encoding method, which extracts various patterns of two consecutive amino acid residues in a protein sequence and counts the number of occurrences of the extracted residue pairs. Huang **[10]** has employed another encoding technique using six physiochemical properties (attributes) of amino acids, namely, composition, predicted secondary structure, hydrophobicity, normalized Van Der Waals volume, polarity, and polarizability.

In fact, amino acids have many physiochemical properties such as isoelectric pH, surface area, hydrophilicity, polarity and the ability of reaction between the ion and the electron **[11]-[14].** In this work, we have used just one physiochemical property of amino acids, which called Hydrophilicity, for encoding the protein sequence. In other words, the normalized number of the Hydrophilicity of the amino acid is used instead of each amino acid in the protein sequence. Consequently, each antibody was transformed into an array, which is used as the input of the GRNN.

Evaluation Criterion:

As mentioned, new method based on ANNs, is introduced for identifying pathogenic antibodies in SLE. The efficacy of the proposed method is evaluated by calculating the following three parameters: positive accuracy, negative accuracy, and general accuracy. **(See supplementary material)**

Anti-dsDNA Identifier:

As mentioned, for the simulation a GRNN was used. For training and testing of the proposed system, we have formed a database, from the KABAT database, consisting of 42 anti-dsDNAs, as the positive data set, and 608 other antibodies, as the negative data set.

A. Training Phase:

To train the GRNN, a training set is formed by randomly choosing 33 positive and 99 negative samples from our database. **Figure. 1(**a) illustrates the general architecture of the system in training phase. In the first layer, each antibody is encoded into an array, based on normalized quantities of the amino acids, Hydrophilicity. In the second layer, a GRNN is trained by arrays obtained from the previous layer so that +1 is assigned as the output for positive samples and -1 is assigned as the output for negative samples.

B. Testing Phase:

To test the proposed scheme, a testing set is formed by randomly choosing 9 positive and 27 negative samples from the part of our database that was not used in the training set.

Architecture of dsDNA binding antibodies identification system in testing phase is illustrated in Fig. 1.b. In the first layer, like the training phase, each antibody is encoded into a discrete signal, based on normalized quantities of the amino acids Hydrophilicity property. In the second layer, signals from the previous step are fed to the neural network. The output of each network will be a number between -1 and +1. By considering a threshold number between -1 and +1, we can separate the anti-dsDNAs from other antibodies. This threshold was selected to be -0.9.

Discussion:

It is important to note that no mathematical method has been introduced for identifying anti-dsDNA antibodies yet. Therefore, we could not compare the effectiveness of proposed method with any other introduced scheme.

In order to assure the validity of the results, the training, and testing phases of the proposed system were repeated 100 times and the final results were calculated by averaging over all 100 runs.

Positive Accuracy, *Negative Accuracy* and *General Accuracy* obtained using the proposed methods are 78.55, 82.92 and 81.83, respectively. For

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 5(2): 58-61 (2010)
© 2010 Biomedical Informatics

Bioinformation **Volume 5**
www.bioinformation.net Issue 2 **Hypothesis**
open access

investigating the roles of light and heavy chains of anti-dsDNA antibodies in binding to DNA, an experiment was also designed. Heavy chain and light chain of antibodies were extracted from the Kabat database. Next, we repeated the last simulation using heavy and light chains. **Table 1** contains these results **(see supplementary material)**.

As illustrated in **Table 1 (see supplementary material)**, the heavy chain of anti-dsDNA is more important than the light chain in binding to DNA. Our simulation results confirm the experimental studies **[16, 17].** Results indicated in Table II show that the proposed method, using heavy and light chains, provides a more accurate identification of anti-dsDNA than the heavy chain or light chain alone. The reason is that by considering a more inclusive model, we are including more information and therefore, we should expect better results. Results of Table I also indicate higher numbers for negative accuracies. This is because we have more negative (non anti-dsDNA) samples. Deleted

We should mention that in this work negative accuracy (non-dsDNA binding antibodies) is more important than positive accuracy. Furthermore, the available data corresponding to the protein sequence for non-dsDNA binding antibodies are much more available than their counterparts. Using unequal number of data points for non-dsDNA and dsDNA during training will introduce some bias, but as it was mentioned since the negative accuracy is more important this action is justified. Simulations using equal number of data points for non-dsDNA and dsDNA produced much lower negative and general accuracy.

Conclusion:

In this paper, we have introduced a new method for identifying pathogenic antibodies in SLE based on GRNN. For identification of dsDNA binding antibodies, the protein sequence of 42 dsDNA binding and 608 non-dsDNA binding antibodies were extracted from the KABAT database. Next, they were encoded using Hydrophilicity values of their amino acids. Coded antibodies were used to train a GRNN. The simulation results indicate that the proposed method is very accurate in recognizing anti-dsDNA antibodies. We have also investigated the roles of light and heavy chains of anti-dsDNA antibodies in binding to DNA. Our simulation results confirmed the experimental findings that the heavy chain is more important than the light chain in regard to binding to DNA.

References:

[1] MR Arbuckle *et al. N Engl Med.* **349**: 1526 (2003)
[2] DA Isenberg *et al. Lupus*, **6**: 290 (1997) [PMID:9296771]
[3] PE Spronk *et al. Lupus*, **4**: 86 (1995) [PMID:7795629]
[4] JS Fetrow *et al. Proteins*, **27**: 249 (1997) [PMID: 9061789]
[5] CR Martin, *Proteins: Struct Funct Genet.* **25**: 130 (1996)
[6] F Sanger, *Adv* Protein *Chem.* **7**: 1 (1952) [PMID: 14933251]
[7] H Paulus, *Chem Soc* Rev. **27**: 375 (1998)
[8] JS Fruton, *Ann NY Acad Sci.* **325**: 1 (1979) [PMID: 378063]
[9] H Wu *et al. Protein Science*, **1**: 667 (1992) [PMID: 1304365]
[10] D Huang *et al. IEEE Trans. Nanobioscience*, **2**: 221 (2003) [PMID: 15376912]
[11] Radzicka & R Wolfenden, *Biochemistry*, **27**: 1664 (1988)
[12] TP Kyte & RF Doolittle, *J Mol Biol*. **157**:105 (1982) [PMID: 7108955]
[13] R Grantham, *Science*, **185**: 862 (1974) [PMID: 4843792]
[14] Cosic, *IEEE Trans. Biomedical Eng.* **41**: 1101 (1994)
[15] F Specht, *IEEE Trans. Neural Networks*, **2**: 568 (1991)
[16] SM Barbas *et al. Proc Natl Acad Sci.* **92**: 2529 (1995) [PMID: 7708679]
[17] P Roben *et al. J Clinical Investigation*, **98**: 2827 (1996) [PMID: 8981931]

Edited by P. Kangueane
Citation: Bahari *et al,* Bioinformation 5(2): 58-61 (2010)
License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 5(2): 58-61 (2010)
© 2010 Biomedical Informatics

Bioinformation **Volume 5**
www.bioinformation.net Issue 2 **Hypothesis**
open access

Supplementry material:

Table 1: Associated accuracies using heavy chain, light chain and the proposed method

	Positive Accuracy	Negative Accuracy	General Accuracy
Heavy chain	80.77	76.48	77.55
Light chain	78	69	71.3
Proposed method (using both heavy and light chains)	78.55	82.92	81.83

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 5(2): 58-61 (2010)

© 2010 Biomedical Informatics