Protein Domain Boundary Prediction from Residue Sequence Alone using Bayesian Neural Networks

DAVID SANCHEZ
Department of Computer Science
California State University
Fullerton, CA 92834
dsanchez@fullerton.edu

SPIROS H. COURELLIS
Department of Computer Science
California State University
Fullerton, CA 92834
shc@ecs.fullerton.edu

Abstract: - Protein domain boundary prediction has been essential in identifying protein domains. Artificial Neural Networks have been used in the prediction of protein domains successfully and have been incorporated in software tools that are widely used. We propose a Feed-forward Artificial Neural Network trained using Bayesian methods. The proposed ANN uses protein residue sequences as inputs and a binary output indicating whether a domain boundary is predicted or not. Training and testing was performed with proteins selected from the SCOP and CATH databases. Preliminary results suggest that the proposed neural networks perform comparably or better than existing neural networks that detect domain boundaries which have been embedded in widely used software tools for the prediction of domain boundaries.

Key-Words: - Protein Domain Boundaries Feed Forward Bayesian Neural Networks MLP

1 Introduction

The identification of protein domains plays a central role in several areas of Biology and Biochemistry. Many different approaches have been developed to identify the region on the aminoacid sequence of a protein where functional domains are located. Although Hidden Markov Models is a frequently used approach to determine domain boundaries, Artificial Neural Network techniques have also been employed successfully, resulting in the successful deployment of software tools such as PPRODO [7] and DOMPro [8]. Protein domain boundary prediction is important in determining the function of a protein as well as determining the three dimensional structure of a protein. A domain is a subunit of a protein that folds independently of the entire protein. It is estimated that there are 1,000 to 10,000 protein domains that exist in nature [1].

Bayesian Neural Networks have been used for a variety of biological applications including protein classification [6]. Although their use, so far, has addressed protein domain classification, they do not address protein domain boundary prediction. In this paper, we present a Feed-Forward Artificial Neural Network trained using Bayesian Methods to predict domain boundaries. We provide a brief background about protein domains and their relationship to protein residues, we define the Artificial Neural Network architecture, we specify its inputs and its output, we introduce variations in the network topology, and we define the training and the testing datasets. We also present preliminary results associated with the convergence of the network, its sensitivity, and its specificity. We conclude with discussion comparing our approach to mainstream comparable approaches reported in the literature.

2 Methodology

A Feed Forward Artificial Neural Network (FFANN) with one hidden layer and one output unit will be considered. The inputs of the FFANN will be binary digits (bits) that will encode each aminoacid in the protein chain. For protein domain boundaries to be detected, successive subsets of its aminoacid sequence will be presented to the FFANN sequentially (the aminoacid sequence will “slide” in front of the FFANN input units). When a boundary is detected, the output will assume the value of “1” otherwise it will be “0”.

2.1 Protein Domain Boundaries

First let us present a brief overview of what protein domain boundaries are. A protein is a chain of amino acids (Fig. 1). Aminoacids in turn are molecules made of an amino group (NH2), a carboxyl group (COOH), and a hydrogen atom attached to an alpha carbon (Fig. 2). Each amino acid has a side chain which is called a residue that distinguishes it from the other amino acids (Fig. 2). There are 20 aminoacids altogether. The
components of the aminoacid that link together are known as the backbone of a protein while the residues that differentiate other amino acids from each other are known as the side chain (rf. Bourne, P. E. et al. for further reading [7]). A domain is a fundamental entity of a protein which folds independently of the entire protein. A protein can have more than one domain which may or may not be contiguous.

2.2 The Neural Network

The input to the Bayesian FFANN consists of binary representations of the amino acids using orthogonal encoding as described by Ma, Q. et al [6]. It takes 20 bits to represent a residue (a residue is equivalent to an amino acid). For example aminoacid “A” can be encoded by the following 20-bit string “00000000000000000001”, aminoacid “B” can be encoded by the 20-bit string “000000000000000000010”, and so on. Orthogonality can be readily verified by the fact that the dot product of any two 20-bit binary aminoacid representations is zero. Using this method we can be sure that the amino acids are independent of each other since they are orthogonal.

In this study, we introduce two neural network architectures. The first neural network (Fig. 3) has 420 binary inputs. That is, 21 consecutive input aminoacids at a time, each encoded by 20 bits yielding a total of 420 binary inputs. It also has 10 hidden units and a single binary output unit. The output is set to “1” when the boundary aminoacid is at the 11th input location between bit 200 and 219, having 10 consecutive neighboring residues on either side of the boundary residue. The second neural network (Fig. 4) has 220 binary inputs. That is, 11 input aminoacids at a time, each encoded by 20 bits yielding a total of 220 binary inputs. The 11 aminoacids are not consecutive. They are obtained by employing a “pick-one, skip-one” interleaving scheme. It also has 5 hidden units and a single output unit.
binary output unit. The output is set to “1” when the boundary aminoacid is at the 6th input location.

![Image](image.png)

**Fig. 4:** Second Neural Network: 220-inputs, 5 hidden units. (interleaving every other aminoacid at the input sequence)

The activation functions are logistical sigmoids between 0 and 1. We trained our neural networks using “Software for Flexible Bayesian Modeling” created by Professor Radford M. Neal of the University of Toronto in Canada.

The output will fire (output value =1) when the boundary aminoacid is in location 11 for the non-interleaved input or 6 for the interleaved input (Fig. 5). Since the number of domain boundaries is extremely small compared to the total number of residues, in addition to the strict mapping of a boundary to one input aminoacid (Fig. 5), we will expand the mapping of the boundary so that the output will fire when a window of aminoacids around the aminoacid of the strict boundary appears at the input. In our study, we employed a window of ±15 residues (i.e., 10 residues – hence window of 10 mapping in our notation) around the strict target (Fig. 6) to increase the domain boundary signal. As described by Ye, L., et al [5], in the training and test data, we will assign all residues within ±15 residues of the true boundary into the boundary class and the rest into the non-boundary class to ease any areas of conflict that may rise from the fact that the CATH and SCOP databases agree on a boundary within ±5 residues of each other.

![Image](image.png)

**Fig. 5:** Strict mapping of the target value. The output fires when the boundary aminoacid appear in location 11 (or 6).

**Fig. 6:** Broadening the boundary definition to a neighborhood of ±5 (i.e., 10) aminoacids (Window 10 mapping).

### 2.3 Training and Testing Datasets

The training and testing datasets were formed by acquiring a total of 491 protein chains from the SCOP and CATH databases. As in the papers by Sim, J., et al [4] and Ye, L., et al [5], the proteins forming the training and testing datasets were determined by only selecting proteins based on the following criteria: (1) the protein domains have to consist of 40 or more residues but less than 500 residues; (2) the protein...
chain must consist of only two domains; (3) each domain within a chain must be contiguous. The protein chain domain boundary must exactly match its counterpart in SCOP or CATH. The training set was comprised of 241 protein chains, which contained over 800 domain boundaries. The testing set included 250 protein chains, which contained also over 800 domain boundaries.

3 Preliminary Results

Preliminary results of the performance of the proposed neural networks are summarized on Table 1.

<table>
<thead>
<tr>
<th>Neural Network</th>
<th>Correctly Predicted Boundaries</th>
<th>False Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-interleaved input Neural Network (Window 10 Mapping)</td>
<td>88.0%</td>
<td>10.9%</td>
</tr>
<tr>
<td>Non-interleaved input Neural Network (Strict Mapping)</td>
<td>72.7%</td>
<td>10.2%</td>
</tr>
<tr>
<td>Interleaved input Neural Network (Window 10 Mapping)</td>
<td>64.2%</td>
<td>14.4%</td>
</tr>
<tr>
<td>Interleaved input Neural Network (Strict Mapping)</td>
<td>66.5%</td>
<td>9.0%</td>
</tr>
</tbody>
</table>

Table 1: The neural network that performed the best at predicting a domain boundary was the non-interleaved neural network with the window 10 mapping.

In addition to correct predictions, Table 1 also contains information indicating when each neural network incorrectly classified a residue as a domain boundary. Correctly predicted boundaries are evaluated via the true positive fraction \( \frac{N_{cb}}{N_{cb} + N_{fb}} \) where \( N_{cb} \) is the number of correctly predicted boundaries and \( N_{fb} \) is the number of domain boundaries that exist in the test set. False positives are evaluated via the false positives fraction \( \frac{N_{fb}}{N_{fb} + N_{pb}} \) where \( N_{fb} \) is the number of falsely predicted boundaries and \( N_{pb} \) is the number of predicted boundaries. The highest the value of the true positive fraction the better the performance of the FFANN is. Also, the lowest the value of the false positive fraction the better the performance of the FFANN is. Ideally, the best FFANN is the one with the highest true positive fraction and the lowest false positive fraction. The highest true positive fraction of about 88% was exhibited by the FFANN with non-interleaved input and window of size \( \pm 5 \). The lowest true positive fraction of about 64.2% was exhibited by the FFANN with interleaved input and window of size \( \pm 5 \). The highest false positive fraction of about 14.4% was exhibited by the FFANN with interleaved input and window of size \( \pm 10 \). The lowest false positive fraction of about 9% was exhibited by the FFANN with interleaved input without input windowing, but its true positive fraction of 66.5% was low compared to the rest of the networks. The best overall performance was exhibited by the network with non-interleaved inputs employing windowing. It has the second worst false positive fraction of 10.9%, which is only 1.9 percentage points worse than the best false positive fraction of 9%. However, its true positive fraction of 88% is 21.5 percentage points better than that of the FFANN exhibiting the smallest false positive fraction of 9%. Its true positive fraction is also 15.3 percentage points higher than that of the FFANN with the second best true positive fraction of 72.7%. The worst performance was exhibited by the network that employs interleaving at the input and windowing. The networks not employing interleaving at their input performed better than the networks that employed interleaving.

4 Discussion

We proposed a Feed-forward Artificial Neural Network for predicting protein domain boundaries, which we trained employing Bayesian methods. We also introduced variations of the network in terms of input and output representation. We used the true positive fraction and the false positive fraction to evaluate the performance of the proposed FFANNs. The summary of the performance measures shown on table 1 reveal that one FFANN stands out in terms of its true positive fraction having a reasonably acceptable (compared to the rest of the FFANNs) false positive fraction.

All our architectures performed in the range of 60% of correct predictions or better, which is comparable to the performance exhibited by mainstream software tools for the prediction of
protein domain boundaries such as PPRODO and DOMPro ([4], [5], [6]). Our neural network with non-interleaving at the input, employing a window of ±5 residues around the true boundary residues assigned to the boundary class reached an 88% success rate well above the other networks and the existing mainstream tools. However, interleaving did not yield favorable results. It appears that interleaving may be appealing because it reduces the dimensionality of the input space, but it does not perform as well. It is probably an indicator of the significance that adjacent aminoacids play in determining a domain boundary.

We would like to extend our work to perform more extensive testing, to employ other input attributes in addition to residues, as they may improve prediction, to consider variations of the proteins we trained with reflected in the evolutionary chain, and to introduce proteins with more than two domains.

Acknowledgments
We would like to thank Professor Radford M. Neal for making the “Software for Flexible Bayesian Modeling” available for download as well as the documentation for using the software. This work was partially supported by the CalState Fullerton MARC U*STAR Program (NIH; 2 T34 GM008612-12).

References: