

New Secondary Structure Prediction software package using automatically trained Bayesian Networks

Jorge Moraleda
(moraleda@stanford.edu, moraleda@data-digest.com)

Abstract

This paper describes a new software package for prediction of secondary structure of proteins. Two main contributions are presented: A novel approach to protein secondary structure prediction based on the usage of Bayesian Networks whose structure and parameters are learned automatically from existing data, and a new data representation tool in which information beyond the most likely classification of each residue is provided.

Introduction to protein structure

A protein is a string of amino acids (or residues). The catalytic, binding, and structural roles played by proteins are all dependent on the correct folding of the protein into a unique 3-D structure. Figure 1 shows what a protein looks like:

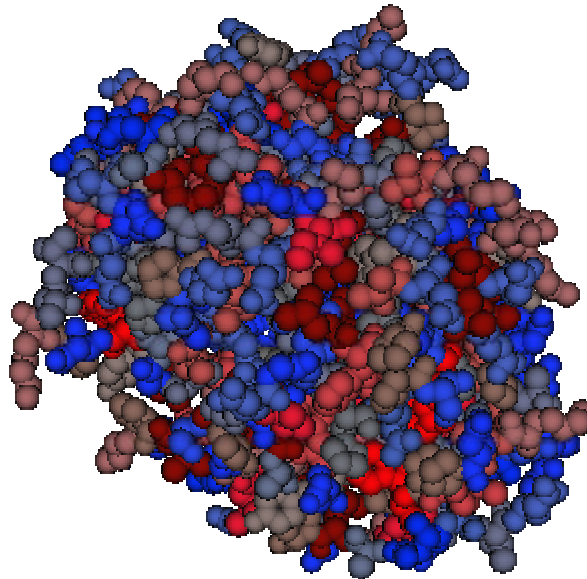


Figure 1. Spatial fill of murine adenosine deaminase (PDB Id: 1ADD)

As can be observed from the figure, protein structures, even of the smallest proteins, can appear at first glance of overwhelming complexity. An efficient route to the goal of understanding structure is to organize thinking into a hierarchy. Unfortunately not everyone agrees precisely in what this hierarchy should be.

One commonly accepted hierarchy is:

- **primary structure:** The linear amino acid sequence of the polypeptide chain including post-translational modifications and disulfide bonds.
- **secondary structure:** Local structure of linear segments of the polypeptide backbone atoms without regard to the conformation of the side chains.
- **super secondary structure (motif):** Associations of secondary structural elements through sidechain interactions.
- **domains:** Associations of lower order structure
- **tertiary structure:** The three-dimensional arrangement of all atoms in a single polypeptide chain.
- **quaternary structure:** The arrangement of separate polypeptide chains (subunits) into the functional protein.

Viewing protein structures at the various hierarchical levels mentioned above is an essential part of understanding the overall and the detailed aspects of protein structure and function. Figures 2 and 3 show schematic representations of how protein structure can be represented and viewed.

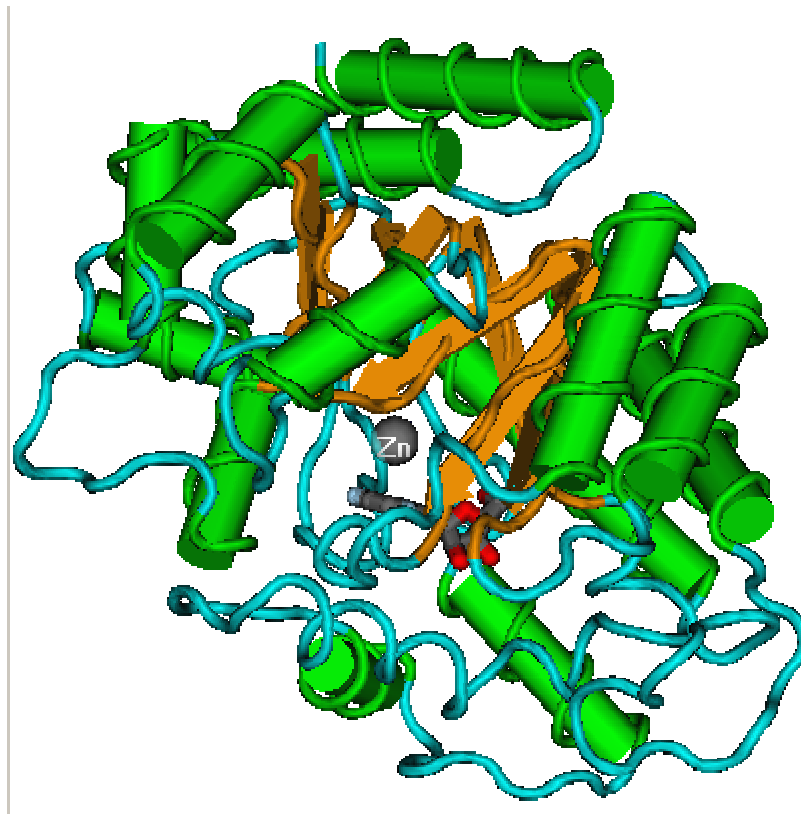


Figure 2. 3D structure of murine adenosine deaminase (PDB Id: 1ADD)

X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR) techniques are capable of producing structures at atomic resolution. However, these methods are quite laborious and expensive and each is beset with its own unique set of limitations (and advantages).

Predicting the three-dimensional shape of proteins from their amino acid sequence is widely believed to be one of the hardest unsolved problems in molecular biology. It is also of considerable interest to pharmaceutical companies who generate more candidate proteins than they can image via NMR.

Prediction of secondary structure provides valuable guidance to physical chemists as they select compounds for crystallography. Protein fold recognition using sequence profile searches frequently allows prediction of the structure and biochemical mechanisms of proteins with an important biological function but unknown biochemical activity.

There are three common secondary structures in proteins, namely **alpha helices**, **beta sheets**, and **turns**. A number of other secondary structures types have been proposed, however they represent a small fraction of residues and may not be a general structural principle of proteins. Nonetheless we decided to attempt a classification system that would use a richer alphabet, namely the DSSP classification system:

- The DSSP code
- H = alpha helix
- B = residue in isolated beta-bridge
- E = extended strand, participates in beta sheet
- G = 3-helix (3/10 helix)
- I = 5 helix (pi helix)
- T = hydrogen bonded turn
- S = bend
- _ = other

As we will see in actuality only the three common secondary structures, and amino acids not belonging to any secondary structure (other) can be predicted with any accuracy from primary structure, but they amount to over 80% of all the residues of the average protein.

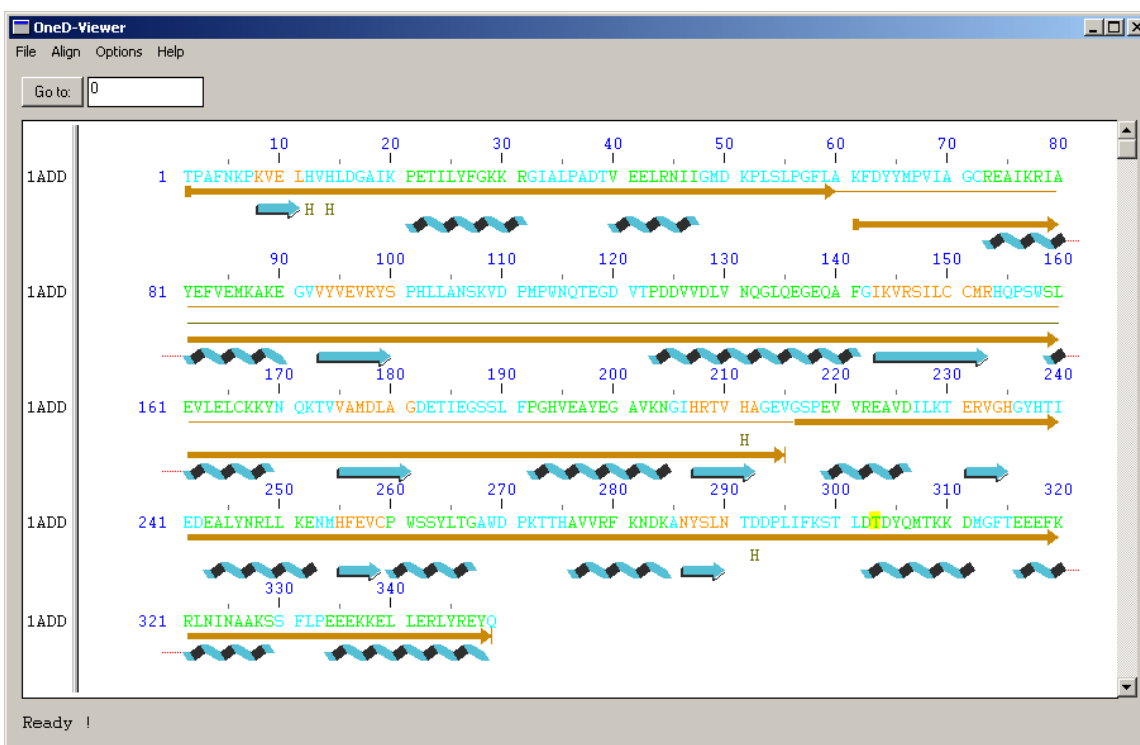


Figure 3. Secondary structure of murine adenosine deaminase (PDB Id: 1ADD)

Introduction to Bayesian Networks

A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest. When used in conjunction with statistical techniques, the graphical model has several advantages for data analysis. One, because the model encodes dependencies among all variables, it readily handles situations where some data entries are missing. Two, a Bayesian network can be used to learn causal relationships, and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. Three, because the model has both a causal and probabilistic semantics, it is an ideal representation for combining prior knowledge (which often comes in causal form) and data. Four, Bayesian statistical methods in conjunction with Bayesian networks offer an efficient and principled approach for avoiding the over fitting of data.

In real learning problems, however, we are typically interested in looking for relationships among a large number of variables. The Bayesian network is a representation suited to this task. It is a graphical model that efficiently encodes the joint probability distribution (physical or Bayesian) for a large set of variables. In this section, we define a Bayesian network and show how one can be constructed from prior knowledge. A Bayesian network for a set of variables $X = \{X_1 ; : : ; X_n\}$ consists of (1) a network structure S that encodes a set of conditional independence assertions about variables in X , and (2) a set P of local probability distributions associated with each variable. Together, these components define the joint probability distribution for X . The network structure S is a directed acyclic graph. The nodes in S are in one_to_one correspondence with the variables X . The lack of possible arcs in S encode conditional independencies.

Bayesian networks for secondary structure prediction

Our goal is to predict secondary structure from primary structure. I.e. given the sequence of residues of a protein we would like to assign a DSSP classification to each residue. Because secondary structure is largely depending on local relationships we chose to construct Bayesian networks with conceptually two types of nodes. Some nodes (R) will represent a series of consecutive residues, and some other nodes (D) will represent the classification of some of these amino acids.

The JPred distribution material (<http://jura.ebi.ac.uk:8888/dist/>) was used for this experiment. It consists of 513 non-redundant sequences. 396 sequences are derived from the 3Dee database of protein domains plus 117 proteins from the Rost and Sander set of 126 non-redundant proteins. All sequences in this set have been compared pair wise, and are non redundant to a 5SD cut-off.

The set of 513 proteins was split into a training set and a validation set. From the training set a structure was automatically learned using Data Digest (www.data-digest.com) proprietary technology for automatic learning of Bayesian Networks from data. Figure 4 shows the network under 10 nodes total with the highest score found automatically by the software.

When observing its structure it is interesting to remark that it is that of a Naïve Bayes Classifier, implying that the probability of a given residue in a chain is independent of its neighbors given the classification. Nonetheless it is important to remark that Naïve classifiers emerge as the best prediction structure also in cases when not enough data exists to support more complex models.

It is curious to observe that the system has found that when taking 7 consecutive amino acids, we predict better the classification of the 3rd in the sequence than that of the 4th, which would be the middle one.

Using the test set, the confusion matrix shown in table 1 was created (“?” appear in the confusion matrix as it appeared in some of the test data, indicating a lack of knowledge of the classification of the corresponding amino acid).

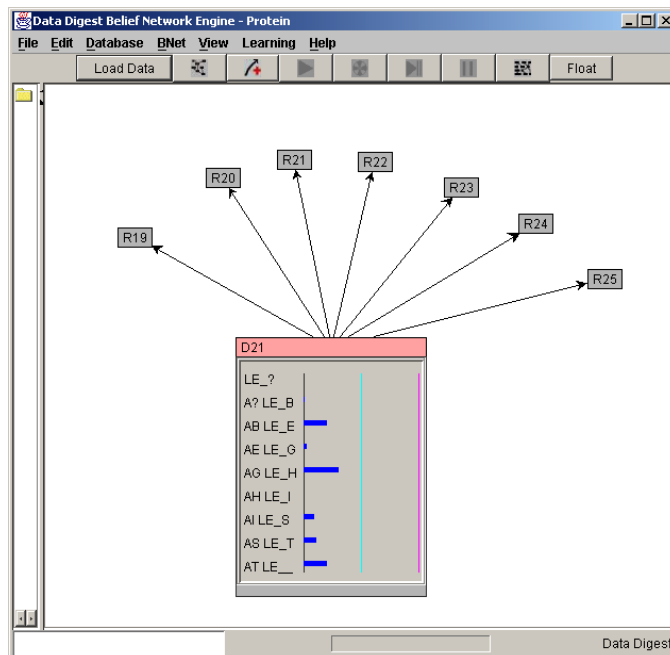


Figure 4. 8 node network (7R,1D)

Actual	Predicted									SUMROW		True positives
	?	B	E	g	h	l	S	t	_			
?	0	0	0	0	5	0	0	2	3	10	0.12%	0.00%
B	0	0	43	0	35	0	0	15	50	143	1.71%	0.00%
E	0	0	939	1	607	0	14	71	277	1909	22.79%	49.19%
G	0	0	43	1	159	0	3	42	101	349	4.17%	0.29%
H	0	0	384	0	1680	0	11	71	184	2330	27.82%	72.10%
I	0	0	1	0	1	0	0	0	3	5	0.06%	0.00%
S	0	0	117	1	231	0	19	109	338	815	9.73%	2.33%
T	0	0	115	1	307	0	20	256	258	957	11.43%	26.75%
_	0	0	298	4	510	0	27	144	818	1801	21.50%	45.42%
SUMCOL	0	0	1940	8	3535	0	94	710	2032	3713	99.33%	
	0.00%	0.00%	23.16%	0.10%	42.21%	0.00%	1.12%	8.48%	24.26%	99.33%		
False positives	0.00%	0.00%	51.60%	87.50%	52.48%	0.00%	79.79%	63.94%	59.74%			
Number Correct	3713											
Total	8375											
Fraction Correct	44.3%											

Table 1. Confusion matrix for 8 node network (7R,1D)

There are many interesting results that can be inferred analyzing this data:

First, the ability to correctly detect alpha helices is higher than 70%, as high as the state of the art results published in the literature. Note that albeit with a lower success rate, beta sheets are also detected with statistical significance. The author is not aware of prior attempts to do this.

We also observe that predictions never are equal to B, I, and very rarely to G. There exist two likely explanations, either these features most likely do not depend on local properties of the peptidic chain, and can thus not be detected, or their frequency of appearance is so low so that not enough training examples were present. Probably both causes are present to a certain degree.

Figure 5 shows the best network found with less than 15 nodes. Given that the amount of data available for training was very limited, it did not make statistical sense to go beyond this number.

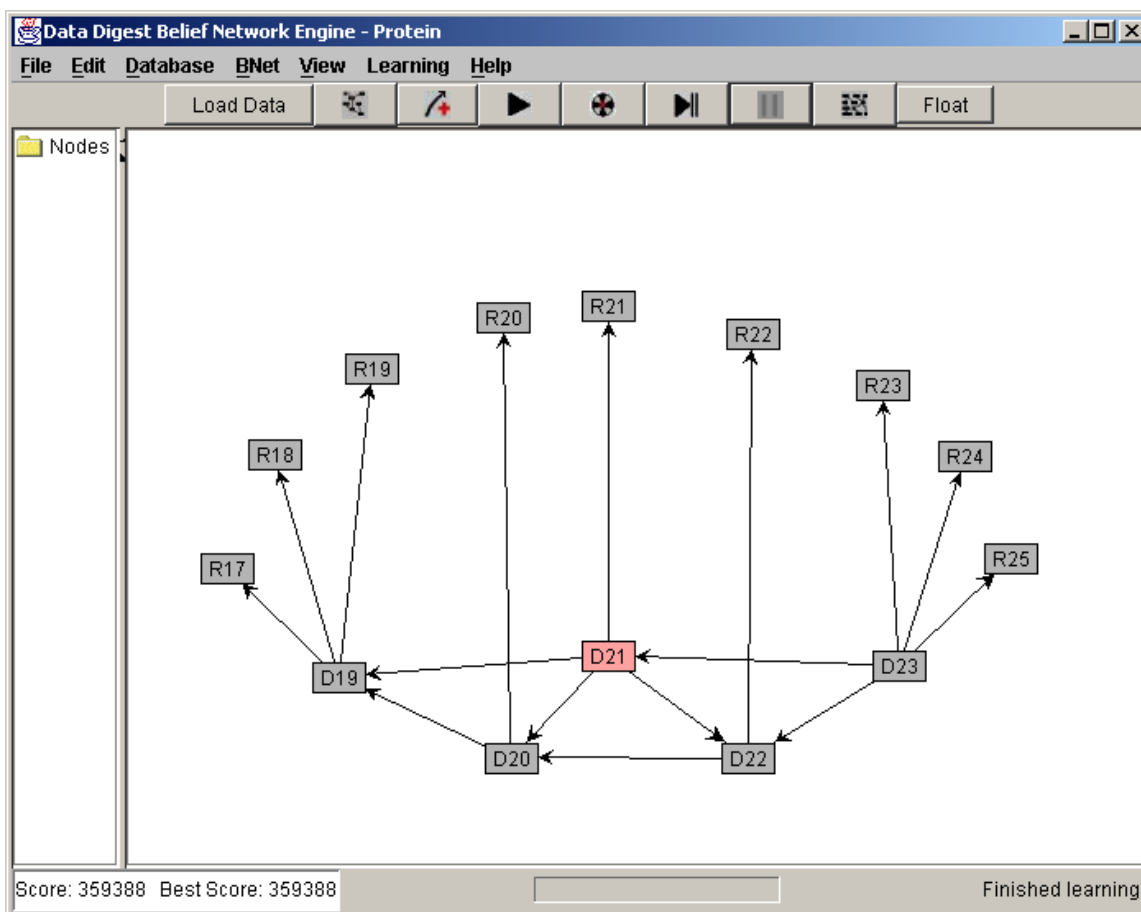


Figure 5. 14 node network (9R, 5D)

Table 2 shows the confusion matrix obtained when testing this network, by setting evidence on all R variables and trying to predict all D variables. The confusion matrix was plotted for D21.

Actual	Predicted									SUMROW		True positives
	?	b	e	g	h	i	s	t	_			
?	0	0	0	0	4	0	0	1	5	10	0.12%	0.00%
B	0	0	35	0	42	0	0	11	55	143	1.71%	0.00%
E	0	0	990	0	601	0	0	59	259	1909	22.79%	51.86%
G	0	0	39	0	157	0	0	23	130	349	4.17%	0.00%
H	0	0	379	0	1700	0	0	38	213	2330	27.82%	72.96%
I	0	0	3	0	0	0	0	0	2	5	0.06%	0.00%
S	0	0	135	1	249	0	0	85	345	815	9.73%	0.00%
T	0	0	107	0	326	0	0	165	359	957	11.43%	17.24%
_	0	0	314	0	569	0	1	108	809	1801	21.50%	44.92%
SUMCOL	0	0	2002	1	3648	0	1	490	2177	3664	99.33%	
	0.00%	0.00%	23.90%	0.01%	43.56%	0.00%	0.01%	5.85%	25.99%	99.33%		
False positives	0.00%	0.00	0.00%	100.00%	0.00%	0.00	100.00%	66.33%	62.84%			
Number Correct	3664											
Total	8375											
Fraction Correct	43.7%											

Table 2. Confusion matrix for 14 node network (9R,5D)

We can observe that the results obtained are very similar to those obtained with the smaller network. The small amount of training data, probably does not justify the additional complexity. Note also that the previous smaller network contains enough consecutive amino acids for two complete alpha helix turns, which is a biologically inspired justification for that particular model.

The software package

To make the system of use for biologists I developed a wrapper around the inference engine that will use one of the above models to make DSSP secondary structure predictions given a residue sequence. If the classification for the sequence is known, it can be entered for evaluation purposes.

The system is intuitive and easy to use, and it will provide not only the most likely classification for each residue, but an ordered list of all the classifications with likelihood above a user settable threshold. In the confusion matrices results discussed above this feature was not used, and only the most likely prediction was taken into account, however a trained biologist will find the additional information very insightful, which will provide additional value and usability. Different structures can be highlighted for ease of visualization

Figure 6 shows a screen shot of the protein secondary structure prediction package.



Figure 6. Secondary structure prediction of murine adenosine deaminase (PDB Id: 1ADD)

Conclusion

The goal of this research was to investigate the usage of Bayesian Networks in Bioinformatics, in particular to predict secondary structure of proteins from primary structure. I.e. given the sequence of residues of a protein we would like to assign a DSSP classification to each residue. Because secondary structure is largely depending on local relationships we chose to construct Bayesian networks with conceptually two types of nodes. Some nodes (R) will represent a series of consecutive residues, and some other nodes (D) will represent the classification of some of these amino acids.

The results obtained are very promising, showing an ability to predict alpha helices higher than 70%, which is state of the art in the literature, as well as the ability to predict other secondary structures, which to the author's knowledge has not been done before.

An advanced interactive software tool has been developed specifically for this project. It presents results beyond the most likely classification of each amino acid, including alternative high probability classifications.

The approach taken in this initial investigation was direct and lends itself multiple enhancements. E.g. no multiple sequence alignments were used despite that it is clearly shown in the literature that they increase prediction success by a few percentage points. Even more sophisticated approaches in which properties of amino acids such as hydrophobicity, size, charge, etc. are taken into account are currently being investigated by the author. In particular, one possible approach being generating artificial data in which actual data is slightly modified and then fed into the training system. The possibility of including explicitly variables in the network to account for properties of residues is also being investigated, as well as the usage of this technology to construct models for DNA micro array data.