European Symposium on Computer Aided Process Engineering – 15
L. Puigjaner and A. Espuña (Editors)

Prediction of Folding Type of Proteins Using Mixed-Integer Linear Programming

Metin Türkay*, Fadime Üney and Özlem Yılmaz
Center for Computational Biology and Bioinformatics and College of Engineering
Koç University, Rumelifeneri Yolu, Sarıyer, İstanbul, 34450, Turkey

Abstract

Proteins are classified into four main structural classes by considering their amino acid compositions. Traditional approaches that use hyperplanes to partition data sets into two groups perform poorly due to the existence of four classes. Therefore, a novel method that uses mixed-integer programming is developed to overcome difficulties and inconsistencies of these traditional approaches. Mixed-integer programming (MIP) allows the use of hyper-boxes in order to define the boundaries of the sets that include all or some of the points in that class. For this reason, the efficiency and accuracy of data classification with MIP approach can be improved dramatically compared to the traditional methods. The efficiency of the proposed approach is illustrated on a training set of 120 proteins (30 from each type). The prediction results and their validation are also examined.

Keywords: Data Classification, Protein Structure, Mixed-Integer Linear Programming

1. Introduction

Proteins are the molecules of life that play a key role in realizing the functions of any biological organism. Discovery of the functions of proteins will enable us to understand the principles of life and working mechanisms of any organism. In the case of humans, this discovery will lead to the design of new drugs that will regulate the functions of proteins in order to improve the quality of life. Functions of proteins are highly correlated to their three dimensional structure. There exist some experimental methods to determine the protein structure including X-ray diffraction and nuclear magnetic resonance (NMR). These experimental methods require quite long times and large amounts of resources. In order to overcome these shortcomings of experimental methods, researchers have developed a host of methods to predict the protein structures.
Due to the importance of protein structure in understanding the biological and chemical activities in any biological system, protein structure determination and prediction has been a focal research subject in computational biology and bioinformatics. The objective of this research is to develop novel methods based on mathematical programming to predict the folding types of proteins by using the primary structure data. A mixed-integer programming model is developed to classify a given primary

* Author to whom correspondence should be addressed: mturkay@ku.edu.tr

protein structure into super-secondary structures according to its amino acid composition.

2. Proteins

Proteins are large molecules indispensable for existence and proper functioning of biological organisms. Proteins are used in the structure of cells, which are main constituents of larger formations like tissues and organs. Bones, muscles, skin and hair of organisms are made basically up of proteins.

2.1 Protein Structure

Starting with the sequence of residues in the chain(s) making up protein, there are 4 basic structural phases: primary structure, secondary structure, tertiary structure and quaternary structure. The primary structure is basically the sequence of amino acids that make up the protein.

The secondary structure (folding type) of a segment of polypeptide chain is the local spatial arrangement of its main-chain atoms without regarding to the conformation of its side chains or its relationship with other segments. There are mainly three types of secondary structural shapes: α-helices, β-sheets and other structures connecting these helices and sheets such as loops, turns or coils.

Alpha-helices are spiral strings formed by hydrogen bonds between CO and NH groups in residues i^{th} amino acid in an α-helix bonds to $i+4^{th}$ amino acid (Brandon and Tooze, 1998). The repeat length of an α-helix is 3.6 residues and the rise is 5.6 Å per turn. Residues i+3 and i+4 are the nearest ones to residue i (Jarey and Hanley, 2004). Alpha-helices are very stable and common structures, accounting for 32-38% of all residues in a typical protein (Kabsch and Sander, 1983).

Beta-sheets are plain strands formed by stretched polypeptide backbone. When β-sheets come together, hydrogen bonds are formed between C=O and NH groups of residues of adjacent chains. The average length of a β-sheet is 6 residues where the minimum number of residues is 2. Typically 20-28% of residues of a protein are placed in β-sheets (Kabsch and Sander, 1983).

Connection structures do not have regular shapes; they connect α-helices and β-sheets to each other. Minimum number of residues required to form a connecting structure is 1. Loops approximately contain 21% of residues with general length of 6 to 16 residues in an average protein where turns hold nearly 33%.

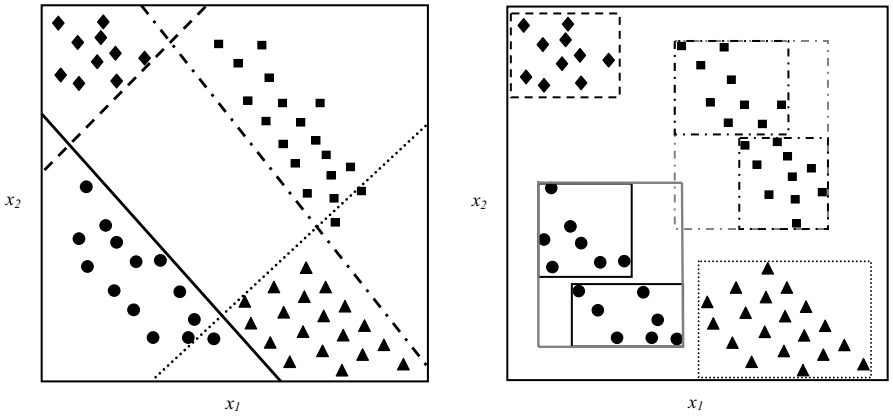
2.2 Classification of Proteins According to Their Secondary Structures

Proteins are classified according to their secondary structure content, considering α-helices and β-sheets. Levitt and Chothia (1976) were the first researchers who propose such a classification with four basic types (Mount, 2001). “All-alpha” proteins consist almost entirely (at least 90%) of α-helices. “All-beta” are the ones composed mostly of β-sheets (at least 90%) in their secondary structures. There are two intermediate classes which have mixed α-helices and β-sheets. “Alpha/beta” proteins have approximately alternating, mainly parallel segments of α–helices and β-sheets. The last class, “alpha+beta” has mixture of all alpha and all beta regions, mostly in an antiparallel fashion.

3. Folding Type Prediction

The overall folding type of a protein depends on its amino acid composition (Nakashima *et. al.,* 1986). There have been several methods proposed to exploit this theory for predicting folding type of a protein (Chou, 1995; Bahar *et. al.,* 1997; Cai *et. al*., 2001). These methods conduct a statistical analysis and separate multi-dimensional amino acid composition data into several folding types.

Classification of multi-dimensional data plays an important role in the decision determining main characteristics of a set. Traditional approaches to data classification use separating hyperplanes as shown in Figure 1a. Although these methods can be efficient in classifying data into two sets, they are inaccurate and inefficient when the data needs to be classified into more than two sets. Mixed-integer programming allows the use of boxes for defining boundaries of the sets that include all or some of the points in that set as shown in Figure 1b. Therefore, the efficiency and accuracy of data classification can be improved dramatically compared to traditional methods. This section consists of details of the mixed-integer programming model and the predicted folding type results of the proteins.



a) using hyperplanes *b) using hyper-boxes*

Figure 1: Schematic representation of classification of data

It may be necessary to use more than one hyper-box in order to represent a single class as shown in Fig. 1.b. When the classes that are indicated by square and circle data points are both represented by single hyper-box respectively, the boundaries of these hyper-boxes overlap. If a region of the attribute space is assigned to more than one class, it is impossible to classify a new data point into a single class. In order to eliminate this possibility, more than one hyper-box must be used to include all of the data points that belong to a class into the same class.

The following indices are used to model the training part of the data classification problem by using hyper-boxes:

i samples (*i=Sample1, Sample2, ..., SampleI*)

k class types (*k=Class1, Class2, ..., ClassK*)

l hyper-boxes that encloses a number of data points belonging to a class (*l=1,..,L*)

m attributes ($m=1,..,M$)
n bounds ($n=lo, up$)

The data points are represented by using the parameter a_{im} that denotes the value of attribute m for the sample i. The class k that the data point i belongs to are given by the set D_{ik}. The following variables are defined for the data classification problem:

X_{lmn} the continuous variable that models bounds n for box l on attribute m
$XD_{l,k,m,n}$ the continuous variable that models bounds n for box l of class k on attribute m
yb_l binary variable to indicate whether the box l is used or not
ypb_{il} binary variable to indicate whether the data point i is in box l or not
ybc_{lk} binary variable to indicate whether box l represent class k or not
ypc_{ik} binary variable to indicate whether the data point i is assigned to class k or not
$ypbn_{ilmn}$ binary variable to indicate whether the data point i is within the bound n with respect to attribute m of box l or not
$ypbm_{ilm}$ binary variable to indicate whether the data point i is within the bounds of attribute m of box l or not
$yp1_{ik}$ type 1 binary variable to indicate the misclassification of data points
$yp2_{ik}$ type 2 Boolean variable to indicate the misclassification of data points

The following MIP formulation models the multi-class data classification problem by the use of hyper-boxes:

$$\min \quad z = \sum_i \sum_k \left(yp1_{ik} + yp2_{ik} \right) + c \sum_l yb_l \tag{1}$$

subject to

$$XD_{lkmn} \le a_{im} ypb_{il} \quad \forall iklmn \,|\, n = lo \tag{2}$$

$$XD_{lkmn} \ge a_{im} ypb_{il} \quad \forall iklmn \,|\, n = up \tag{3}$$

$$XD_{lkmn} \le M \; ybc_{lk} \quad \forall klmn \tag{4}$$

$$\sum_k XD_{lkmn} = X_{lmn} \quad \forall lmn \tag{5}$$

$$ypbn_{ilmn} \ge (1/M)(X_{lmn} - a_{im}) \quad \forall ilmn \,|\, n = up \tag{6}$$

$$ypbn_{ilmn} \ge (1/M)(a_{im} - X_{lmn}) \quad \forall ilmn \,|\, n = lo \tag{7}$$

$$\sum_l ypb_{il} = 1 \quad \forall i \tag{8}$$

$$\sum_k ypc_{ik} = 1 \quad \forall i \tag{9}$$

$$\sum_l ypb_{il} = \sum_k ypc_{ik} \quad \forall i \tag{10}$$

$$\sum_k ybc_{lk} \le yb_l \quad \forall l \tag{11}$$

$$ybc_{lk} - \sum_i ypb_{il} \le 0 \quad \forall lk \tag{12}$$

$$ybc_{lk} - \sum_i ypc_{ik} \le 0 \quad \forall lk \tag{13}$$

$$\sum_n ypbn_{ilmn} - ypbm_{ilm} \le \mathrm{N} - 1 \quad \forall ilm \tag{14}$$

$$\sum_{m} ypbm_{ilm} - ypc_{ik} \le \mathrm{M} - 1 \quad \forall ilk \tag{15}$$

$$ypc_{ik} - yp1_{ik} \le 0 \quad \forall ik \notin D_{ik} \tag{16}$$

$$ypc_{ik} + yp2_{ik} \ge 1 \quad \forall ik \in D_{ik} \tag{17}$$

$$X_{lmn}, XD_{lkmn} \ge 0,\ yb_{l}, ybc_{lk}, ypb_{il}, ypc_{ik}, ypbn_{ilmn}, ypbm_{ilm}, yp1_{ik}, yp2_{ik} \in \{0,1\} \tag{18}$$

4. Results

A better training database is important for improving the accuracy of prediction. Therefore, the selection of proteins for the training database is carried out according to following points: (i) a typical or distinguishable feature for each of the folding types concern, (ii) a good quality of structure, (iii) as many non-homologous structures as possible. Training database is composed of 120 proteins, 30 members from each class as shown in Table 1.

Table 1: The PDB codes of the 4x30=120 representative proteins in the training database (the5th letter indicates the chain id)

Number	α type	β type	α+β type	α/β type
1	1AVHA	1ACXA	1CTFA	1ABA
2	1BABB	1CD8A	1DNKA	1BKSB
3	1C5AA	1CDTA	1EMEA	1CISA
4	1CPCA	1CIDA	1FXIA	1DBPA
5	1CPCL	1DFNA	1FXIB	1DHRA
6	1ECOA	1HILA	1FXIC	1EAFA
7	1FCSA	1HIVA	1FXID	1ETUA
8	1FHAA	1MAMH	1HSBA	1GPBA
9	1FIAB	1PAZA	1LTSA	1KKJA
10	1HBGA	1REIA	1PPNA	1OFVA
11	1HDDC	1TENA	1RNDA	1OVBA
12	1HIGA	1TFGA	2AAKA	1PFKA
13	1LE4A	1TLKA	2ACHA	1Q21A
14	1LTSC	2ALPA	2ACTA	1S01A
15	1MBCA	2AVIA	2PHYA	1SBPA
16	1MBSA	2AYHA	2PRFA	1SBTA
17	1RPRA	2BPAA	2RNTA	1TIMA
18	1POCA	2BPAB	2VAAA	1TREA
19	1TROA	2LALA	3IL8A	1ULAA
20	256BB	2ILAA	3MONA	2CTCA
21	2CCYA	2OMFA	3RUBS	2FOXA
22	2LH1A	2SNVA	3SICI	2HADA
23	2LHBA	2VAAB	3SSIA	2LIVA
24	2LIGA	3CD4A	4BLMA	2PGDA
25	2ZTAA	3HHRC	4LZTA	2TMDA
26	2MHBA	4GCRA	4TMSA	3GBPA
27	2MHBB	7APIB	5H0HA	4CPAA
28	3HDDA	8FABA	5TLIA	5P21A
29	4MBAA	8FABB	8CATA	8ABPA
30	4MBNA	8I1BA	9RSAA	8ATCA

The proposed method categorized the 120 proteins into their corresponding classes with accuracy value of 100%. The best prediction rates for the α, β, $\alpha+\beta$, and α/β classes are 67, 91, 81, and 67%, respectively (Bahar et al., 1997). The main factor for obtaining high accuracy value with proposed method is the use of boxes (i.e., hyper-boxes) that allows using discrete regions for classification rather than continuous regions separated by hyperplanes.

5. Conclusions

In this paper a new data classification method based on mixed-integer programming is developed. The method uses hyper-boxes rather than hyper-planes to define the boundaries of the classes. Although, hyper-planes are very effective in separating data that belongs to two distinct classes, they perform very poorly in the case of multi classes. The method described in this paper is very efficient for separating the data into more than two classes. The performance of the method is illustrated on protein structure classification problem that contains four distinct classes in 20-dimensional space. The best classification accuracy reported in the literature for this folding type problem was 81% on the average. The prediction accuracy with the proposed method is 100% illustrating the effectiveness of the method.

References

Bahar, I., Atilgan, A.R., Jernigan, R.L., and Erman, B. (1997), Understanding the Recognition of Protein Structural Classes by Amino Acid Composition, *Proteins: Structure, Function, and Genetics* **29,**172-185.

Brandon, C., and Tooze, J. (1998), Introduction to Protein Structure, Garland Publishing, Inc., New York.

Cai, Y.D., Liu, X.J., Xu, X.B., and Zhou, G.P. (2001), Support Vector Machines for predicting protein structural class, *BMC Bioinformatics* **2,** 3.

Chou, K.C. (1995), Does the folding type of a protein depend on its amino acid composition?, *FEBS Letters* **363,** 127-131.

Jarey, J., and Hanley, V., Proteins - Biophysical society on-line text book

Kabsch, W., and Sander, C. (1983), A dictionary of protein secondary structure, *Biopolymers* **22,** 2577-2637.

Levitt, M., and Chothia, C. (1976), Structural patterns in globular proteins, *Nature* **261,** 552-558.

Mount, D. W. (2001), Bioinformatics: Sequence & Genome Analysis, Cold Spring Harbor Laboratory Press, Woodbury, New York.

Nakashima, H., Nishikawa, K., and Ooi, T. (1986), *J. Biochem.* **99,** 152-162.