CLUSTERING BINARY CODES TO EXPRESS THE BIOCHEMICAL PROPERTIES OF AMINO ACIDS

Huaiguo Fu, Engelbert Mephu Nguifo
CRIL-CNRS FRE2499, Université d'Artois, Rue de l'université SP 16, 62307 Lens cedex. France
{fu,mephu}@cril.univ-artois.fr

Abstract We study four kinds of binary codes of amino acids (AA). Two codes of them are based respectively on biochemical properties, and the two others are generated with artificial intelligence (AI) methods, and are based on protein structures and alignment, and on Dayhoff matrix. In order to give a global significance of each binary code, we use a hierarchical clustering method to generate different clusters of each binary codes of amino acids. Each cluster is examined with biochemical properties to give an explanation on the similarity between amino acids that it contains. To validate our examination, a decision tree based machine learning system is used to characterize the AA clusters obtained with each binary codes. From this experimentation, it comes out that one of the AI based codes allows to obtain clusters that have significant biochemical properties. As a consequence, it appears that even if attributes of binary codes generated with AI methods, do not separately correspond to a biochemical property, they can be significant in the whole. Conversely binary codes based on biochemical properties can be insignificant when forming a whole.

Keywords: Bioinformatics and AI, Amino acids, Classification, Clustering

1. Introduction

More and more methods and techniques of Artificial Intelligence (AI) are applied to solve problems of molecular biology such as protein secondary or tertiary structures prediction. Such techniques are generally based on AA mutation matrices. However there are a lot of symbolic AI techniques based on binary representations that could be applied in this domain. These works used single representation which does not catch many biochemical properties of AA. If the biochemical properties of AA can be perfectly described with certain binary codes, it will improve performance accuracy on the prediction of protein structure and function from its AA sequences [1]. Hence in this paper we will focus on binary representation for AA.

Expressing binary rules is more understandable for human-expert and could be helpful for providing efficient results explanation to the expert. Many symbolic AI systems deal with binary representations. They are unable to treat numerical values as that encodes in AA mutation matrices. If several research works are being devoted to AA indices or mutation matrices, our investigation of the literature gives rise only to four works on binary representation of AA:

Dickerson & Geis (DG) [2], Marlière & Saurin (MS) [3], De la Maza (DM) [1], and finally Gracy & Mephu (GM) [4].

In this study, we compare and analyse these 4 methods of AA binary representation. In order to search for a global significance of each binary code, firstly, we use a clustering algorithm to group AA. Then a machine learning system is used to explain the significances of the clusters using biochemical data. If the clusters perfectly correspond to certain biochemical properties of Amino acids, we consider this binary code as a good binary representation. In order to validate and explain the clusters of AA, the decision tree C4.5 is used to characterize the AA clusters.

The paper is organized as follows. The four AA binary representations are presented in next section. In the third section, we present hierarchical clustering to generate clusters of binary representations, and discuss the results obtained. And then a decision tree system C4.5 is used to validate the clusters of representations of AA in the fourth section.

2. Binary representations of amino acids

Binary representation of AA is a table of twenty rows and different columns. Each column is a property which can correspond to a biochemical property. Each row corresponds to an AA.

2.1 Binary Codes based on biochemical properties

Two representations based on biochemical properties of AA are described by: DG and MS.

DG's binary representation considers following properties of AA: aliphatic, aromatic, charged, polar, size of AA and hydrophobic. The table of the properties of AA could be easily transform to a binary representation.

DG make an analysis of some protein sequences of the heavy and light chains of immunoglobulines to create this representation. A problem with this representation is that some AA have exactly the same physical and chemical properties in their classification, so the binary representations can't be distinguished. A way to solve this problem could be to add additional properties that allows to distinguish them.

MS propose to represent AA with 8 biochemical properties [3]. On the basis of these 8 biochemical properties, each AA can be represented by a bit string like with the DG's code. For example, we use 00010100 to represent the amino acid I (Isoleucine).

This coding is a topologic description of AA.The coding appears sharply particular choices to certain types of studies, because certain criteria such as hydrophobicity in particular are debatable. This coding can be spread with the addition of other properties such as hydrophobicity, hydrophilic, etc. With such coding, it is also necessary to explicitly or implicitly add negation of properties in order to avoid inclusion between two AA codes.

2.2 Binary codes based on AI methods

DM's [1] and GM's [4] codes apply AI algorithms to generate binary representations. They propose a complete system to generate and test the binary representations of AA.

They use different techniques, but the whole structure is the same. In order to generate best binary representations, they use searching algorithms to find the best solution for representation of AA, from some data of AA or protein.

DM used the primary and secondary structure of proteins to create AA representations that facilitate secondary structure prediction. A genetic algorithm searches the space of AA representations. The quality of each representation is quantified by training a neural network to predict secondary structure using that representation. The genetic algorithm then uses the per-

formance accuracy of the representation to guide its search and to create AA representations (see an example in [1]) that improve the performance accuracy.

DM [1] describes a system that synthesizes regularity exposing attributes from large protein databases. After processing primary and secondary structure data, this system discovers an AA representation that captures what are thought to be the three most important AA characteristics (size, charge, and hydrophobicity) for tertiary structure prediction.

GM's method uses the Dayhoff matrix and simulated annealing algorithms to generate the binary code of representation of 20 AA. Using this method, we can get different representations of AA by changing its parameters.

3. Clustering analysis of binary representations of AA

In the previous section, four methods to represent AA with binary codes are briefly presented. However, we face some questions: What's the significance of each binary representation? Which is a good AA representation? To answer these questions, we use a clustering algorithm and decision tree system to validate the AA binary representations.

3.1 Hierarchical Clustering

We use different hierarchical clustering methods available inside the SAS datamining package: Ward's method, Average Linkage method, and Centroid method. For example,using Ward's method with the 24 bits representation of GM's method, when the number of clusters is 5, the result obtained is: Cluster 1: Asp, Glu, His, Lys, Asn, Pro, Gln, Arg. Cluster 2: Gly, Met, Val. Cluster 3: Cys, Ser, Thr. Cluster 4: Ala, Ile, Leu. Cluster 5: Phe, Trp, Tyr.

3.2 Examination of clusters

The AA biochemical properties are at the basis of the interpretation of clusters of binary representations. We modify and extend the representation of chemical properties of AA proposed by DM. Modifications and extensions come from discussion reported in recent publications on biochemistry. We add some properties such as the mass, the number of atoms, and the hydrophobicity scale.

As an example, the result of 5 clusters with the 24 bits representation of GM, is well-adapted to biochemical conditions and these clusters of AA have a certain logic of biochemical affinity:

Phenylalanine, Tryptophane and Tyrosine : aromatic AA with cyclic side chain and no charged. Alanine, Isoleucine and Leucine : AA with side chains aliphatic. Cysteine, Serine and Threonine : the Threonine differs of Serine by a grouping methyl and the Cysteine differs of serine by the presence of one atom of sulfur in the place of the atom of oxygen. Asp, Glu, His, Lys, Asn, Pro, Gln, Arg: are more hydrophilic. Gly, Met, Val are hydrophobic.

From the examination of all the sets of clusters obtained, it appears very often that there were always two or three clusters with debatable similarity, except for the previous one: the set of 5 clusters of 24-bits representation of GM. With the coding of DM, we didn't obtain the same clusters as that reported in [1]. This may be due to the fact that DM uses the Cobweb clustering algorithm which is different from the Ward's method. With the coding based on biochemical properties, we were unable to find a set with good clusters.

From this first observation, it appears that coding based on AI method can have a good global significance, whenever coding based on biochemical properties could not be significant in a whole. This global significance arises from the fact that similarity between AA is expressed inside the Dayhoff matrix in the case of GM, or inside protein sequence alignment and protein structure in the case of DM.

A second observation is made on properties of AI based coding. For the binary representation based on biochemical properties, each column corresponds to one biochemical property. For representations based on AI methods, we find that properties of binary codes do not separately correspond to biochemical properties.

4. Interpretation of clusters

In order to verify the clusters of binary representations of AA using biochemical data, we use a public domain of decision tree system C4.5 to predict the cluster of an AA.

From the results of decision tree, the representation based on AI methods can be shown that it corresponds to some biochemical properties of AA in varying degrees. It validates the accuracy of the clusters of AA representations.

However, even with the clusters and biochemical properties reported in [1], we were unable to find the explanation tree obtained with the decision tree system as mentioned in the paper.

The results show that the clustering results of the 24-bits representation produced by GM's method are correct and can be characterize with biochemical properties. This is in concordance with the results of our examinations by hand. This shows that the 24-bits representation produced by GM's method is one of the best binary representations of AA. This corroborates previous results reported in [4], where this representation allows to find good alignment when dealing with weakly homologous protein sequences.

For biochemical based representations, the decision tree can't give an understandable explanation of the clusters. This is in concordance with our analysis by hand. Thus biochemical properties based representations need to be preprocessed before being used, in order to express a whole significance.

From the results of clustering and decision tree process, we know that the methods of DM, and of GM based on AI methods can generate good binary representations of AA. These representations are significant in a whole, and allow to take into account biochemical properties when dealing with protein primary structure.

5. Conclusion

This paper reviews two kinds of AA binary representations respectively based on biochemical properties, and on searching methods. A comparative study of these codes is described, and provides some significant results.

Good AA representations can facilitate the prediction of proteins secondary or tertiary structure, can allow to find good alignment of proteins primary sequences. This work could allow to improve results of AI methods when dealing with protein folding problem as it is well-established that data representation is one of the keys of success of AI methods.

References

[1] De la Maza M. Generate, Test and Explain: Synthesizing Regularity Exposing Attributes in Large Protein DataBases. In *Proc. of (HICSS)*, pages 123–129, Hawaï, USA, 1994.

[2] Dickerson R.E. and Geis I. The structure and actions of proteins. *Harper & Row Publishers,New York, NY*, pages 16 – 17, 1969.

[3] Sallantin J., Marlière P., and Saurin W. Description logique des contextes spatiaux dans les protéines: application à la conception de polypeptides artificiels. In *Actes des journées Point Curie*, pages 141–153, Paris, France, 1984. Institut Curie.

[4] E. Mephu Nguifo. *Concevoir une abstraction à partir de ressemblances*. Thèse de doctorat d'université, Université de Montpellier II, Mai 1993. 276 pages.