Poster M-15

A Numeric Coding Scheme for Amino Acids

Authors:
Siwo Geoffrey Henry (*Dept. Biochemistry, Egerton University*)

Short Abstract: This study presents a novel approach of numerically representing amino acids taking into account their biological properties as encoded in the PAM-250 matrix. Representation of amino acids in numerical form could make protein sequence analysis malleable in mathematics, data mining and artificial intelligence.

Long Abstract:
Numeric coding is the assignment of a meaningful numeric value to a non-numerical symbol. In protein sequences, amino acids are normally represented with one letter codes which are non-numerical. The representation of protein sequences as a string of letters has had a profound impact on biology in enhancing bioinformatics analysis of proteins.
With the rapid availability of vast amounts of protein sequence data, data mining holds the key to extracting meaningful information from these sequences.
Unfortunately, protein sequences are presented in a nominal form. A major requirement of data mining techniques, on the other hand, is that data should be presented in a form that facilitates representation and analysis mathematically. Thus, once nominal data is encoded in a numerical form it can be applied in data mining.
Some workers have chosen to encode amino acids in a 20-bit format where an amino acid is represented by 1 and nineteen 0s. Others use binary coding but of lesser bits such as 5. These have proved useful in neural networks for backtranslating amino acids, an important process in Degenerate PCR cloning, and identification of human gene coding sequences from open reading frames in DNA databases. However, these methods do not take into account the biological relationships between various amino acids and the numerical codes are assigned randomly.
This analysis presents a simple way of representing amino acids in a numeric form that takes into account biological properties of amino acids. An amino acid is represented as either a numerical score or a 20-dimensional feature vector.

Methods.
A given amino acid can be considered as appoint in a multidimensional space whose coordinates are various physical and chemical parameters.
However, there are numerous physical and chemical properties inherent in amino acids and we cannot identify all of them let alone measure their values.
To circumvent this challenge, we can employ a method that indirectly takes into account the various physicochemical parameters in amino acids.
So for simplicity, we can envision amino acids as points in a 20-dimensional space.
The reason for considering amino acids as points in a 20-dimensional space is because there are 20 naturally occurring amino acids in proteins. These 20 amino acids espouse a set of physical and chemical properties that are capable of enabling proteins to perform a diverse array of functions. Each of these amino acids is different from the other due to a set

of physicochemical properties it has. We can not explicitly state these physical and chemical properties but an amino acid espouses them all. A numerical measure of the relationship between amino acids is reflected in substitution matrices such as PAM-250 and BLOSUM-62. PAM-250 is constructed from alignments of closely related proteins hence it espouses most of the relationships between amino acids as compared to other matrices.

An amino acid can therefore be considered as a point in a 20-dimensional space whose coordinates are the substitution scores to each amino acid in PAM-250.

As a consequence we can represent each amino acid as a feature vector with 20-features:
(A,C, D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,.W, Y)

Where A is the substitution score of the particular amino acid to Alanine, C the substitution score to Cysteine etc. This was done for each of the 20 amino acids. Each dimension of the resulting vectors was standardized.

Generation of Numerical Scores for each Amino Acid:

To facilitate the derivation of a numerical score for each amino acid, a hypothetical amino acid was envisioned. This amino acid, which we may call a super amino acid, has all the physico-chemical properties of the 20 amino acids. The super amino acid can not exist in nature, but here it serves as a useful reference against which all amino acids can be compared. It is not biased towards any amino acid. Its feature vector consists of the highest possible score in each dimension.

Standard mathematical procedures were then employed in a perl script to determine the Euclidean distance of each amino acid to the hypothetical amino acid. The calculated Euclidean distance of a given amino acid to the reference amino acid was then taken as the numerical score for that amino acid. A self organizing map was then used to study natural clusters of amino acids based on numerical scores.

Results and Conclusion.

The numerical scores generated show that tryptophan, cytosine and phenylalanine have the highest scores in that order. The amino acids with least scores are serine, threonine and histidine. The implication of this is that of all amino acids, tryptophan is the least related to the rest while serine appears to be the most related to all the others. The numerical score generated agrees with some known facts about relationships between certain amino acids. For instance,serine and threonine are closely related. Their numerical scores too are closer to each other than to those of other amino acids. On the basis of Kohonen-self organizing map, the numerical scores can be used to classify amino acids into three natural clusters: V,A,R,N,H,I,M,S,T,Q occur in one cluster, W,Y,F,L,C in another and D,E,G,K,P in the other.

When each amino acid was represented as a 20 dimensional feature vector and then a self organizing map applied, three natural clusters slightly different to the ones above were obtained. In addition a supervised learning algorithm could correctly group amino acids into known groups in the charge and structural alphabets when a 20-dimensional vector is used for each amino acid.

The numerical codes for amino acid generated here may be applied in various areas of protein sequence analysis including artificial intelligence based techniques for antigen, allergen and protein function prediction. Protein sequences can be represented as feature vectors in space with each amino acid being replaced by the numerical score or a 20-dimensional vector.