Analysis of Covariations of Sequence Physicochemical Properties

MOSHE A. GADISH, DAVID K.Y. CHIU
Department of Computing and Information Science
University of Guelph
Guelph, Ontario, N1G 2W1
CANADA

Abstract: - Sequence analysis often does not take the physicochemical properties into account. On the other hand, some of these properties could be useful in inferring the folding and functional attributes of the molecule when considered with the original sequence information. We evaluated here an analysis using multiple aligned sequences incorporating five physicochemical properties. In addition to site invariance information, we also consider the covariation or interdependence patterns between aligned sites using an information measure. We propose a method based on analyzing the expected mutual information between sites that is statistically significant with a confidence level. When summing the measured information along the aligned sites, we compare the pattern from the measure to the structural and active site of the molecule. In the experiments, the model enzyme molecule lysozyme is chosen. The aligned sequence data are evaluated based on the mapped physicochemical properties of the amino acid residues. Analysis between the original and the transformed sequence data incorporating the physicochemical properties are then compared, subtracted and visualized. From the comparisons, the plots show that some of the selected physicochemical properties in the analysis correlate to the locations of active sites and certain folding structure such as helices. The experiments generally support the useful role of incorporating additional physicochemical properties into sequence analysis, when significance of the statistical variations is taken into account.

Key-Words: - protein sequence analysis, physicochemical properties, expected mutual information, statistical significance, lysozyme

1 Introduction

The effect of various physicochemical properties of amino acids on the protein structure and function is well known. For example, by considering the conserved physicochemical properties in addition to the amino acid types of the sequences, a meaningful alignment may be obtained. Thus, the classifier using PHYSEAN (PHYsical Sequence Analysis) adds position-specific physicochemical information for protein classification [1]. PHYSEAN predicts protein classes with highly variable sequences on the basis of their physical, chemical and biological characteristics (such as hydrophobicity). PHYSEAN produces reasonably accurate predictions, indicating the importance of incorporating the physicochemical properties into protein sequence analysis. Hydrophobicity plots have also been used in protein sequence analysis for the purpose of discovering hydrophobic cores and resolving some of the problems in protein folding. Other successes of incorporating physicochemical properties include the use of amino acid scales and physicochemical properties in predicting secondary structure propensity (alpha helix, beta sheet, turn, etc.) [2]. This paper evaluates further how these physicochemical properties can be used to analyze multiple aligned sequences of a protein family.

Many speculations on why physicochemical properties in protein analysis are useful can be made. Proteins have remarkable range of functions from the many distinctive three-dimensional structures given their sequences [3]. Sequence analysis may determine how the amino acids specify the conformations of their structure. An important step in analyzing the sequences then involves finding recurrent patterns in the sequence that may not be obvious. From these patterns, relationship to patterns of the function of the protein can then be analyzed [4]. Information measures such as the Shannon entropy function or mutual information are mathematical measures that are general and may reveal implicit statistical relationships even though the exact properties that are involved may be unknown. Crooks and Brenner [5] have used entropy densities and local inter-sequence mutual information density to study the effect of primary and secondary protein structure. A transformation score is mapped from each amino acid into the three secondary structure classes of extended beta sheets, helices, and loops. Their study supports the view that these information measures may capture the cooperative processes where secondary and tertiary structure can then form.

This paper develops a method by analyzing the statistical significance of expected mutual information

based on the physicochemical properties. It further sums all such mutual information at each position and compares it to that without taking the physicochemical properties into account. The plots showing the differences are then evaluated.

In the experiments, the model enzyme molecule lysozyme c is chosen. The aligned sequence data are evaluated based on the mapped physicochemical properties of the amino acid residues. Analysis between the original and the transformed sequence data incorporating the physicochemical properties are then compared and visualized. From the comparisons, the plots show that some of the selected physicochemical properties correlate to the locations of active sites and certain folding structure. The experiments generally support the interesting role of these physicochemical properties when their statistical variations are taken into account.

2 Detecting Significant Interactions between Sites.

2.1 Representation of Aligned Sequence Data

Multiple biological sequences (of a protein family) can be aligned to form a sequence ensemble. For example, each amino acid site in the protein sequence can be considered as a variable where the corresponding amino acid of a sequence is the outcome. This can be represented as $X = (X_1, X_2 \ldots X_m)$ where m is the number of variables, indicating the length of the alignment. An instance of X is a realization denoted as $x = (x_1, x_2 \ldots x_3)$. Each $x_j (1 \le j \le m)$ can take up an attribute value denoted as $x_j = a_{jq}$. An attribute value a_{jq} is a value taken from an attribute value set $\Gamma_j = \{a_{jq} | q = 1, 2, \ldots, L_j\}$ where L_j is the number of possible values for the variable X_j, or the cardinality of the set.

2.2 Expected Mutual Information

Expected mutual information is a measure of the statistical interdependence between two variables. The stronger the interdependence between the two variables, the larger is the expected mutual information between them. If the two variables are statistically independent, then the expected mutual information between them is zero [6, 7, 8].

Expected mutual information, denoted as $I(X_i, X_k)$, is a measure that can calculate the deviation from independence between two discrete valued variables X_i and X_k. It is defined as,

$$I(X_i, X_k) = \sum_{i=1}^{L_i} \sum_{k=1}^{L_k} P(X_i^j, X_k^h) \log \frac{P(x_i^j, x_k^h)}{P(x_i^j) P(x_k^h)} \quad (1)$$

2.3 Testing for Statistical Interdependence

It is important when calculating statistical interdependence to take into consideration their statistical significance, so that their correspondence is not due to chance, otherwise considerable error can be accumulated. This is especially important in case when information from multiple variables is summed. Evidence of statistical interdependence can be evaluated by comparing the two competing hypothesis between the independence and interdependence assumptions. Since expected mutual information has an asymptotic chi-square distribution [8], a statistical test that is based on the chi-square statistics can be used.

When comparing the statistical independence between two outcome values of the distinct variables, we use the following method based on evaluating the standard residual [8]. Let us denote a joint outcome of the two variables X_i and X_k as $e_{ik}^{jh} = (a_i^j, a_k^h)$, where e_{ik}^{jh} represents the joint observation of $X_i = a_i^j$ and $X_k = a_k^h$. The standard residual is defined as:

$$z(e_{ik}^{jh}) = \frac{obs(e_{ik}^{jh}) - exp(e_{ik}^{jh})}{\sqrt{exp(e_{ik}^{jh})}} \quad (2)$$

Here, $obs(e_{ik}^{jh})$ is the observed frequency and $exp(e_{ik}^{jh})$ is the expected frequency for the joint observation e_{ik}^{jh} in the samples. Given M as the total number of samples, the adjusted residual is defined as [8]:

$$d(e_{ik}^{jh}) = \frac{z(e_{ik}^{jh})}{\sqrt{v(e_{ik}^{jh})}} \quad (3)$$

Where,

$$v(e_{ik}^{jh}) = 1 - P(X_i = a_{ij}) P(X_k = a_{kh}) / M \quad (4)$$

The adjusted residual, $d(e_{ik}^{jh})$, has an asymptotic normal distribution. Hence, by convention, a statistical significance level of either 95% or 99% can be chosen. Using a 2-tailed test, the corresponding tabulated threshold values are 1.96 and 2.58, respectively. A statistically significant event, that is, the two values being statistically interdependent is,

$$d(e_{ik}^{jh}) > N_\alpha \quad (5)$$

Where N_α is the threshold value with a statistical significance level α.

2.4 Significant Expected Mutual Information

A measure of expected mutual information involving only the significant events in the variable-pair can be denoted as $I^* = (X_i X_k)$. Expected mutual information $I(X_i, X_k)$ as defined in Equation (1) subjected to the

selections from the statistical test, as derived in Equation (2), can be denoted as (6):

$$I^*(X_i,X_k) = I(X_i,X_k) \mid d\left(e_{ik}^{jh}\right) > N_\alpha \quad (6)$$

This measure of expected mutual information then calculates the significant expected mutual information of events only if they are selected to be statistically significant.

Significant expected mutual information, $I^*(X_i, X_k)$, can be normalized to produce values between 0 and 1 by dividing it to Shannon entropy involving only those events. Shannon entropy involving the significant selected events can be denoted as:

$$H^*(X_i,X_k) = -\sum_j^{L_j}\sum_k^{L_k} P\left(X_i^j,X_k^h\right)\log P\left(X_i^j,X_k^h\right) \mid d\left(e_{ik}^{jh}\right) > N_\alpha \quad (7)$$

The normalized expected mutual information based on the selected significant events, can now be defined as:

$$R^*(X_i,X_k) = \frac{I^*(X_i,X_k)}{H^*(X_i,X_k)} \quad (8)$$

To evaluate the total amount of interdependency expressed on a given variable (or site on the aligned sequences) induced by the detection of R^*, it can be calculated as:

$$MR(X_i) = \sum_k^{L_k} R^*(X_i,X_k) \quad (9)$$

2.5 Significant Expected Mutual Information

To incorporate amino acid properties into protein sequence analysis, we substitute identified physicochemical properties into the corresponding amino acids. The aligned sequences are then transformed, and discretized into different pre-defined intervals for evaluation. The transformed sequences are then analyzed for their statistical interdependency from these discretized physicochemical properties. This method allows analysis on discrete and continuous physicochemical properties. It can also handle patterns due to non-linear and linear dependency. The physicochemical properties of different amino acid types sharing similar characteristics can then be compared and analyzed.

For physicochemical properties that have continuous value (such as molecular weight here), a scheme is developed to discretize the property. After discretization, each amino acid is substituted with its corresponding calculated label for that property. Each continuous physicochemical property is divided into n equal intervals,

$$Interval=(max-min)/n \quad (10)$$

Where *max* and *min* are the maximum and minimum values respectively an amino acid has for that property, and *Interval* is the interval size. Each property then falls into one of the predefined *n* intervals. Amino acids that share similar physicochemical values fall into the same interval are assigned identical discrete values. This process is repeated for each physicochemical property, producing a transformed sequence ensemble for each property, with a specified accuracy of discretization.

Significant expected mutual information can be compared between that from the original sequence ensemble and the sequences transformed from the physicochemical properties. The difference can be visualized along the aligned position of the sequences that reflects the summation from all positions. Each generated plot is visualized in two ways. First, the plot shows the value of significant expected mutual information (normalized) between every pair of sites in the sequence. Second, the cumulative plot visualizes the total significant expected mutual information at that position.

From the plot, a high score reflects strong interdependencies between sites. Furthermore, clusters (regions with similar characteristics) can also be observed. Because of the transformation and analysis of the differences between the original and the transformed sequences, the strong interactions can be attributed to the physicochemical property being displayed.

3 Experiments

3.1 Experimental Data

The sequence ensemble consists of 75 complete lysozymes c sequences. The aligned protein sequences have 130 residues. Lysozyme c was chosen because of its qualities as a model protein. It is classified as a monomer (or protein with a single amino acid chain). This simplifies the analysis by eliminating interactions among amino acids of different peptide chains as in more complex polymeric proteins. In addition, lysozyme c lacks any cofactors or prosthetic groups, thus eliminates interactions due to these groups. Lysozyme c has been well studied and its structure and function is reasonably understood [9].

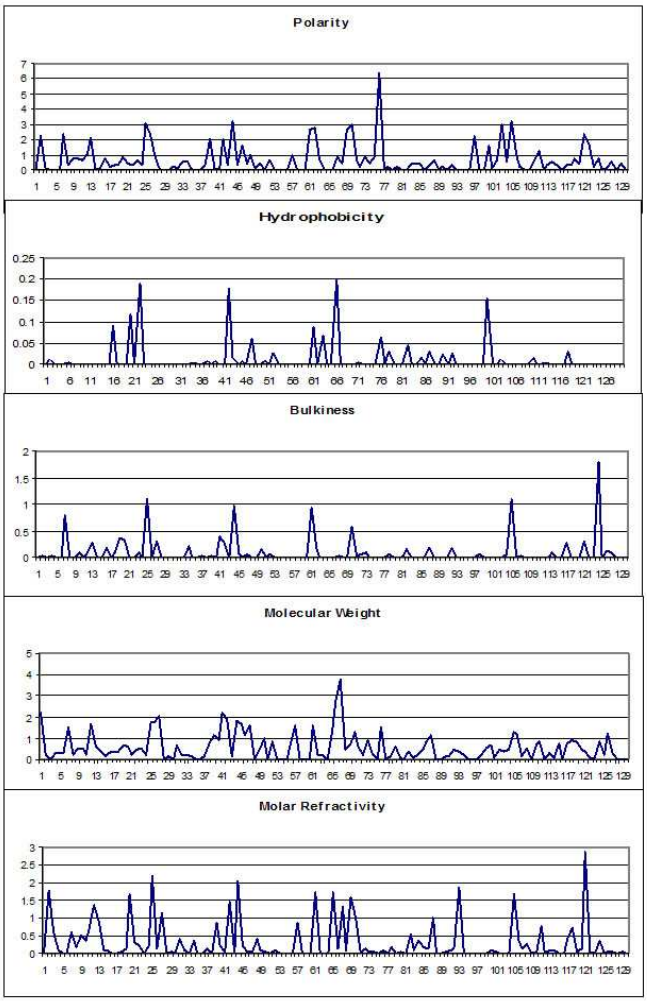

Fig. 1 Plots of $MR(X)$ calculated as the differences from that of the original sequences among the 5 physicochemical properties (X-axis indicates the site number; Y-axis is the calculated $MR(X)$).

Five physicochemical properties were chosen here: polarity, hydrophobicity [10], molecular weight [11], molar refractivity [2] and bulkiness [12]. The polarity property is represented as discrete values. It can be broken down into six distinct values [13,14]. Hydrophobicity, molecular weight, molar refractivity and bulkiness are all continuous values. They are each discretized using different number of intervals, n=4, 5, 6, 7, 8, for evaluation.

3.2 Experimental Method

Experiment 1 plots the value of significant expected mutual information on the original aligned sequences between sites. Since the alignment has 130 sites, it forms a 130 by 130 matrix of the $R^*(X_i, X_k)$ value. Next, these values calculated from the original sequence are compared to those from each of the five physicochemical properties based on their different discretized labels. The calculated interdependency between all positions is summarized in Table 1. Three classes of patterns were discovered in the analysis, labeled as: gap, peak, and cluster. Gap in the plots refers to a region in the sequence with low cumulative significant value. (In the plot, they are identified as horizontal and vertical white color coded bands.) The positions that are located within the gaps are often conserved with respect to the property considered. A peak in the plots reflects positions that have high cumulative significant values (indicative of strong interdependency.) A cluster is an area in the plot that represents at least one position having strong value with another region (with a length of more than one site).

3.3 Experimental Results

A comparison of different physicochemical effects on the sequence can be visualized (Fig. 1). Bulkiness and hydrophobicity have the weakest effect; while polarity has the clearest, peaked at position P76. Hydrophobicity and to a lesser extent bulkiness, display clear gaps across the sequence. Gaps are indicative of lack of the effect due to the physicochemical property. Some gaps overlap between plots (Table 2). Overlapping indicates a combined effect of the physicochemical properties at these positions. Gaps that overlap among properties correspond to regions in the aligned sequences that are not affected by the physicochemical properties (e.g. hydrophobicity and bulkiness). Some of the gaps include amino acids that are located in the secondary structural regions. The gap at positions P53 – P59 includes amino acids that line up with the active site cleft, in positions P57 – P59 [9]. The gap at positions P79 – P81 includes amino acids that form part of a 3_{10} helix (a helix structure that is characterized by shorter turns then found in an α-helix) [15]. Molecular weight, molar refractivity and polarity are more irregular than hydrophobicity and bulkiness in terms of distribution of peaks and gaps in their cumulative plots.

Additionally, many peaks overlap between bulkiness, molecular weight and molar refractivity. These overlapping positions show strong interactions with respect to the property being displayed. There are two possible explanations. It is possible that at these positions all three properties interact together, in synergy. Alternatively, the similarity between these plots can be possibly attributed to the close relationship between these three properties. For instance, molecular weight, bulkiness and molar refractivity are all alternate measures of amino acid size. The peak at position P44 is the active site, while peaks at positions P41, P65 – P67 are close to it [9]. Locations that are close to active site may be accounted from the shape of the catalytic site. Positions P105 is observed to interact with several other positions in the sequence (P25, P41) with respect

to bulkiness and molecular weight. Positions P105 and P41 are located in the active site, while positions P105 and P25 are spatial neighbors in the 3-Dimensional model.

3.3.1 Physicochemically Invariant Patterns

It is generally agreed that the amino acid sequence of protein when considering the physicochemical properties hold important information about the protein [16]. Lysozyme, when considering all the occurrences by a wide variety of organisms, provides a unique opportunity to examine the common relationship between its sequences and the other relevant information of the molecule such as structure, folding characteristics and evolutionary relationships. Lysozyme c sequences highly vary with respect to sequence similarity. For example, human and chicken lysozyme, show differences in 51 sites. However, they are structurally similar [17]. Although they exhibit differences in their amino acid values, many of the variant sites are actually invariant with respect to some physicochemical properties. Many of these invariant patterns are identified by the gaps in this study.

Table 1 Patterns from the plots after subtracting the value of R^* from the original sequences (extracted from Fig.1).

Site	Pol	Hydro	Bulk	MW	MR
P10-12					Cluster
P25	Peak		Peak	Peak, Cluster	Peak
P26	Peak			Peak, Cluster	
P27				Peak, Cluster	
P28			Gap	Peak	
P29-31			Gap		
P41			Peak	Cluster	
P42		Peak	Peak	Peak, Cluster	Peak
P43				Peak, Cluster	
P44			Peak	Cluster	
P45-47				Cluster	
P53-56			Gap	Gap	Gap
P57			Gap		Gap
P58-60			Gap	Gap	Gap
P61	Peak		Peak		
P62	Peak				
P65-67				Peak, Cluster	
P69					Peak
P76	Peak				
P86-87					Peak, Cluster
P95-97				Gap	Gap
P98-99					Gap
P100		Peak			
P101-4					Gap
P105	Peak		Peak		Peak
P121					Peak
P124			Peak		

Table 2 Consistent patterns observed from the selected physicochemical properties.

Physicochemical Properties	Pattern Type	Site	Characteristics of the molecule
Hydrophobicity, Bulkiness	Gap	P26 – P33	Inside α-helix at positions P25 – P36.
Hydrophobicity, Bulkiness	Gap	P53 – P59	Inside the active site cleft P57 – P59.
Hydrophobicity, Bulkiness	Gap	P79 – P81	Part of a single-turn 3_{10} helix P80 – P83, and half of the disulfine bridge between P64-P80.
Hydrophobicity, Bulkiness	Gap	P95 – P97	Positions are inside α-helix in P89 – P100.
Hydrophobicity, Bulkiness	Gap	P104 – P109	Overlap P104, P108 – P109 active site cleft.
Bulkiness, Molecular weight, molar refractivity	Peak	P25	Inside and start of α-helix P25 – P36.
Bulkiness, Molecular weight, molar refractivity	Peak	P41, P44	P44 is in the active site.
Molecular weight, molar refractivity	Peak	P65 – P67	Near positions P63 – P64 that are in the active site. Also next to P64 which is part of a disulfide bridge (P64-P80). Possible stability role.
Bulkiness, Molecular weight, molar refractivity	Peak	P105	In the active site.

4 Conclusion

The experiments showed that the selected physicochemical properties have an effect on the biosequence and can be measured using the proposed significant expected mutual information. This information measure reflects an underlying pattern of interactions. Some of these patterns are located at the active sites while others are located in the secondary structural elements like helices. Many of the identified patterns are spatial neighbors that congregate sequentially. The research shows the importance of eliminating statistical variations that are not significant and focusing on events that are, thus resulting in a more accurate calculation in very noisy sequence data.

Acknowledgements. The research is supported by the Discovery Grant of the National Science and Engineering Research Council of Canada and the Korea Research Foundation Grant (KRF-2004-042-C00020)..

References:

[1] Ladunga, I., PHYSEAN:PHYsical Sequence Analysis for the identification of protein domains on the basis of physical and chemical properties of amino acids. Bioinformatics. Vol.15, No.12, 1999, pp.1028-1038.

[2] Jones. D.D., Amino acid properties and side-chain orientation in proteins: a cross correlation approach. J Theor Biol. Vol.50, No.1, 1975, pp.167-83.

[3] Branden C., Toolze J., Introduction to Protein Structure. Garland Publishing; 2nd edition, 1999.

[4] Stolorz P, Lapedes A., Xia Y., Predicting protein secondary structure using neural net and statistical methods. J Mol Biol. Vol.225, No.2, 1992, pp.363-77.

[5] Crooks, G.E., Brenner, S.E., Protein secondary structure: entropy, correlations and prediction. Bioinformatics. Vol.20, No.10, 2004, pp.1603-11.

[6] Haberman, S.J., The analysis of residuals in cross-classified tables. Biometrics, Vo.29, 1990, pp.205-220.

[7] Li, W., Mutual Information Functions Versus Correlation Functions. Journal of Statistical Physics, Vol.60, No.5-6, 1990, pp.823-837.

[8] Wong, A.K.C., Wang Y., High-order pattern discovery from discrete-valued data. IEEE Trans. Knowledge and Data Eng., Vol.9, No.6, 1997, pp.877-893.

[9] Jolles, P., Lysozymes: Model Enzymes in Biochemistry and Biology. 1996.

[10] Eisenberg D., Schwarz E., Komarony M., Wall R., Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J Mol Biol. Vol.179, No.1, 1984, pp.125-42.

[11] Amino Acids Database. Frontiers in Bioscience website. [http://www.bioscience.org/urllists/aminacid.htm]. Retrieved in Oct 15, 2004.

[12] Zimmerman J.M., Eliezer N., Simha R., The characterization of amino acid sequences in proteins by statistical methods. J Theor Biol. Vol.21, No.2, 1968, pp.170-201.

[13] Darnell, J., Lodish, H., Baltimore, D., Molecular Cell Biology. Scientific American Books, 1990.

[14] Lesk M. A., Introduction to Protein Architecture: The Structural Biology of Proteins. Garland Publishing; 2nd edition, 1999.

[15] Lakshmanan K. Iyer & P.K. Qasba,Molecular dynamics simulation of a-Lactalbumin and calcium binding c-type lysozyme. *Protein Engineering*, 1999, Vol.12, No.2 , pp.129-139.

[16] Phillips, D., The Hen-White Lysozyme Molecule. *Proceedings of the National Academy of Sciences of the United States of America*. 1967, Vol57, pp.483-495.

[17] Hooke, S.D., Radford, S.E., Dobson, C.M., The Refolding of Human Lysozyme: A Comparison with the Structurally Homologous Hen. Biochemistry. 1994, Vol.33, No.19, pp.5867-76.