Bioinformation
www.bioinformation.net

open access
Hypothesis

Nearest-neighbor classifier as a tool for classification of protein families

Mona Chaurasiya[1*], Gohel Bakul Chandulal[1], Krishna Misra[1*], Vivek Kumar Chaurasiya[2]

[1] Indian Institute of Information Technology, Allahabad, India; [2] Indian Institute of Technology, Roorkee, India; Krishna Misra - Email: kkmisra@yahoo.com, [*]Corresponding authors

Received December 07, 2009; Revised January 30, 2010; accepted November 13, 2010; published March 31, 2010

Abstract:
Knowledge about protein function is essential in understanding the biological processes. A specific class or family of protein shares common structural and chemical properties amongst its member sequences. The set of properties that display its unique characteristics for clearly classifying a protein sequence into its corresponding protein family needs to be studied. Our study of these important properties conducted on four major classes of proteins namely Globins, Homeoboxes, Heat Shock proteins (HSP) and Kinase have shown that frequency of twenty naturally occurring amino acids, hydrophobic content of protein, molecular weight of protein, isoelectric point of protein, secondary structure composition of amino acid residues as helices, coils and sheets and the composition of helices, coils and sheets in the secondary structure topology plays a significant role in correctly classifying the protein into its corresponding class or family as indicated by the overall efficiency of Nearest Neighbor Classifier as 84.92%.

Keywords: proteins, family, classification, classifier

Background:
There is a vast gap between the between the amount of sequence information and functional characterization of the protein. Hence, fast computational methods are required for the correct characterization of its function **[1, 2].** The present classification system of proteins is based either on Sequence-Sequence similarity or Sequence-Structure similarity **[3, 4].** These two methods play a critical role in predicting a possible function for a new sequence. But these methods do not function properly when clear sequence or structural similarities do not exist as in case of distantly related proteins. As we know that not all homologous proteins have analogous functions **[5]**. Proteins have many shared domains, need not necessarily perform the same function e.g. SH2, WD40 etc., are known to have different functions **[6].** Proteins of specific functional class share common structural and chemical features essential for performing similar functions **[7].** It is of interest to consider protein functional family classification as a method for facilitating protein function prediction, which is expected to be particularly useful in the cases, described above and may thus be used as a protein function prediction tool to complement sequence alignment methods**.**

It has been reported that physical and chemical properties of the protein's primary sequence play an important role in determining the protein's function **[8]**. We have tried to explore new physiochemical properties along with the secondary structure information of the proteins for the correct characterization of its family. Instead of direct comparison of the sequences, Nearest Neighbor classifier **[9]** was used to cluster the physiochemical properties that were generated from the protein primary sequence. Samples of the protein known to be in a functional class are used to train the Nearest Neighbor system to recognize specific features and to classify the protein sequences. Such an approach may be applied to closely related proteins as well as distantly related proteins.

Methodology:
The physical parameters of the protein are very important in defining an unknown protein into a specific class. The most important physical parameters are the hydrophobic and the polar residues. It has been found in research of transmembrane proteins that the discriminatory features are observed in the intermediate steps when the patch of the hydrophobic residues followed by neutral amino acids and the same is observed when the polar residue string is followed by neutral amino acids **[4].** Thus on the basis of physiochemical properties and structural properties, protein family can be classified using derived parameters from primary sequence.

Our dataset comprised of sequences of protein that were randomly chosen from the Swissprot database **[10].** Every protein sequence was represented by a specific feature vector assembled from encoded representations of tabulated residue properties including the composition of twenty naturally occurring amino acids, hydrophobic content of the protein, molecular weight of protein, isoelectric point of protein, secondary structure composition of amino acid residues as helices, coils and sheets and the composition of helices, coils and sheets in the secondary structure topology. Each of above mentioned features are defined as follow.

(1) Amino acid composition (**aa**): $\mathbf{aa_i = a_i /N}$, where a_i is frequency of amino acid **'i'** in the protein and **i** = twenty different amino acid $\mathbf{aa_i}$ represents the proportion of amino acid **'i'** in the protein **N =** total number of residue in protein sequence;

(2) Hydrophobic content (**H**): **H = h/N**, where **h** = no. of hydrophobic residue in the protein sequence, **N =** total number of residue in protein sequence.

(3) Isoelectric point of protein ($\mathbf{I_p}$)
The Isoelectric point is the pH at which the protein has no net charge. The net charge of a protein was calculated as the sum of the number of positively charged residues (protonated lysine, arginine, histidine), minus the number of negatively charged residues (deprotonated tyrosine, cysteine, glutamate, aspartate), plus the number of protonated amino termini, minus the number of deprotonated carboxyl termini. The net charge calculation does not take into account any electrostatic interactions within the protein that may perturb ionization. For each amino acid of interest, the number of protonated residues is determined by the following equation: $N_{(p)} = N_{(t)} [H^{(+)}] / ([H^{(+)}] + K_{(N)})$, where $N_{(p)}$ = number of protonated residues, $N_{(t)}$ = total number of residues of a specific amino acid, $[H^{(+)}]$ = hydrogen ion concentration, $K_{(N)}$ = dissociation constant for the amino acid of interest that is equal to the following: $_{(10)}{}^{(-pK)}N$

(4) Secondary structure feature (S and SS)
We predicted the secondary structure of protein sequence using SSPro4.0 **[11]**. We then calculated the following features; Proportion of Amino acid falling within particular secondary structure, $\mathbf{S_j = a_j/N}$, where $\mathbf{a_j}$ is no. of amino acid that fall in the **'j'** and **j** = (helix, sheet, loop) **N=** total number of residue in protein sequence. Secondary structure proportion composition $\mathbf{ss_j = n_j/M,}$ **w**here $\mathbf{n_j}$ is the no. of secondary structure **'j'** in topology and **j =** (helix, sheet, loop) **M** = total number of secondary structure (helix, sheet, loop) in sequence

topology.

(5) Molecular weight of protein (mw)
mw = $^{N}\Sigma_{i=1}$ w_i; w_i = molecular weight of amino acid, **N** = total number of residue in protein sequence combining all above mentioned features, Finally feature vector is **F = [aa *H Ip s ss mw*]** that becomes the input for Nearest Neighbor classifier.

Classification of proteins
For the classification of various classes of protein based on mentioned physiochemical properties (**F**) using nearest neighbor classifier (NN). NN is quite popular amongst pattern recognition community mainly due to its simplicity and good performance. Consider the problem classifying unknown object into P classes with the training data set which is formulated as **X= {(x_i, c_1) … (x_N, c_N)}** where **x** is the feature vector, **{c_i|i = 1 to *P, P = no. of classes*}** and **N** = no. of objects. According to NN rule, a new unclassified object **x** is assigned a class c_i of x_i nearest to **x.** In our case, there is four class **c** = {globin, homeobox, heat shock protein, kinase} and **N** is 2000. We used Euclidean distance as a measure of similarity.

Results and Discussion:
The study was conducted on four major classes of proteins namely Globins, Homeoboxes, Heat Shock Proteins and Kinase. Sequences of all four protein class were downloaded from Swissprot database. Then proposed physiochemical features were evaluated using Nearest Neighbor Classification technique. The results obtained from Nearest Neighbor Classifier fed with training set of five hundred sequences of each protein family against test set of three hundred sequences of each protein family, chosen randomly, are as follows:

The results show that the properties used namely the frequency of twenty naturally occurring amino acids, hydrophobic content of protein, molecular weight of protein, isoelectric point of protein, secondary structure composition of amino acid residues as helices, coils and sheets and the composition of helices, coils and sheets in the secondary structure topology plays a significant role in correctly classifying the protein into its corresponding class or family as indicated by the overall efficiency of Nearest Neighbor 84.92% **(Table 1 in supplementary material).** On excluding the class Kinase the overall efficiency is 97.67% **(Table 2 in supplementary material)** suggesting that the protein family "Kinase" is more widely distributed than the Globin and the Homeobox protein families. Our further analysis on the individual protein family showed that Kinase and Heat Shock protein are closely related with respect to the above defined parameters.

Conclusion
The study suggests that the above taken feature set of proteins has the potential for classifying the proteins into protein families. Further addition of property / properties will help in discrimination of protein families that are widely distributed.

References
[1] M Pellegrini *et al*., *Curr. Opin. Chem. Biol.* **5**:46–50 (2001) [PMID: 11166647]
[2] S. A. Teichman & G. Mitchison, *Nat. Biotechnol.*, **18**:27 (2000) [PMID: 10625385]
[3] P. Bork & E.V Koonin, *Nature Genet.*, **18**:313–318 (1998) [PMID: 9537411]
[4] S.A Teichmann *et al*., *Curr. Opin. Struct. Biol.*, **11**:354–363 (2001) [PMID: 11406387]
[5] S.A Benner *et al.*, *Res Microbiol*., **151**:97–106 (2000) [PMID: 10865954]
[6] E.M Marcotte *et al. Science*, **285**:751–753. (1999) [PMID: 10427000]
[7] Istav Ladunga *et al.*, *Bioinformatics*, (1999) **15**:1028-1038.[PMID: 10745993]
[8] C.Z. Cai, L.Y. Han, Z.L. Ji, X. Chen *et al*., *Nucleic Acids Research*, **31**:3692-3697 (2003) [PMID: 12824396]
[9] Hans Peter Keck and Thomas Wetter *et al*., *In Silico Biology* **3**:0023 (2003) [PMID: 12954089]
[10] http://www.uniprot.org/downloads
[11] J. Cheng, A. Randall *et al*., *Nucleic Acids Research*, **33**:72-76 (2005). [PMID: 15980571]

Edited by P. Kangueane
Citation: Chaurasiya *et al.*, Bioinformation 4(9): 396-398 (2010)

Supplementary material:

Table 1: Result set obtained by Nearest Neighbor classifier

Families	**Globin**	**Homeobox**	**HSP**	**Kinase**	**Efficiency**
Globin	299	0	1	0	99.67%
Homeobox	0	277	2	21	92.33%
Heat Shock protein	10	26	203	61	67.67%
Kinase	3	2	55	240	80.00%
Efficiency	**84.92 %**				

After excluding kinase protein family we have the following results.

Table 2: Result set obtained by Nearest Neighbor classifier excluding kinase protein family

Families	**Globin**	**Homeobox**	**HSP**	**Efficiency**
Globin	297	2	1	99.00%
Homeobox	1	299	0	99.67%
Heat Shock protein	8	9	283	94.34%
Efficiency	**97.67%**			