INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

Invitation to MTech Thesis Defense of **Ankit Sharma**: May 10, 2018 (Thursday): 15.00-16.30 IST

In Partial Fulfillment of the Requirements for the Degree of
M.Tech CB

Ankit Sharma (MT16121)

Will defend his thesis

Title: "Protein Classification on the basis of Thermal Stability using Supervised Learning"

IIIT-D Faculty and Students are invited

Date: May 10, 2018 (Thursday)
Time: 15.00 – 16.30 IST
Place: A320 (Meeting Room, 3rd Floor, New Academic Building, IIITD)

Examiner:	**Internal:**	**Sriram K.**
	External/Internal:	**Lipi Thukral, CSIR-IGIB**
	Advisor:	**Debajyoti Bera,**
	Co-Advisor:	**Ganesh Bagler**

****

Abstract

Species evolve by adapting to variable thermal conditions. The differences in the thermal stability of hyperthermophilic, thermophilic and mesophilic proteins arise partly due to their structural variations. The goal of this thesis is to identify structural features responsible for these variations using machines learning techniques that use features derived from the residue interaction graphs (RIG) and the amino acid sequences of proteins. For the RIG network model, we studied the features linked to thermal stability which capture different notions of centrality, connection strength, weighted clustering coefficient and such. We evaluated them against a few features that were hitherto not studied in the context of thermal classification and demonstrated that the new features can significantly improve classification accuracy. We further improved the performance by using a histogram of centrality values as a feature vector instead of using a single statistic such as mean that has been the trend so far among researchers. We discovered that the histograms corresponding to edge betweenness centrality, current flow closeness centrality and 2-hop degree centrality lead to the best classification accuracy among the network-based features. We also investigated the state-of-the-art features based on amino acid sequences and proposed a new one using the amino acid trimers of a protein. For empirical evaluation, we investigated a set of 842 hyperthermophilic, 533 thermophilic and 2248 mesophilic proteins and compared our proposed features with the state-of-the-art features using commonly known classification techniques such as SVM, ANN and random forest. We obtained an overall accuracy greater than 90% which is significantly better than what has been reported so far.