

EFFICIENTLY SCREENING MONOMERS OF PROTEINS USING NAÏVE BAYES CLASSIFIER

Pralay Mitra, Debnath Pal

*Bioinformatics Centre, Supercomputer Education and Research Center,
Indian Institute of Science, Bangalore-560012, India*

pralay@pallab.serc.iisc.ernet.in
dpal@serc.iisc.ernet.in

The Protein Data Bank (PDB) provides annotation on the quaternary structure for each protein in the repository, but is at times inaccurate. To fill in the lacunae several other databases¹ and programs^{2,3} have attempted to screen for correct quaternary structure of proteins. Here we present a new method to identify protein monomers from crystal contacts. The task is challenging because in lattice protein-protein contacts at times are substantially large, but not to the extent to confer it a definitive biological relevance. Our method match closely with PiQSi¹, a curated protein quaternary structure database, compared to prediction by PISA³, a popularly used web server to annotate protein quaternary structure.

We have designed a naïve Bayes classifier to discriminate contacts in the crystal lattice using ten physico-chemical properties at the protein interface. The model has been built on a known dataset of 609 protein complexes, consisting of biological and non-biological contacts, where PiQSi and PISA agreed on the quaternary structures. The accuracy of the classifier with 10 fold cross-validation is 91.95% and receiver-operator-curve (ROC) area is 0.963. The kappa statistics ($\kappa=0.83$) indicates that the predicted output class is in perfect agreement with the actual class⁴. The model is tested with a test set of 113 protein complexes with maximum interface area $<1500 \text{ \AA}^2$, where PiQSi and PISA contradict each on the question whether a monomer exists or not. Our classifier is able to identify true protein monomers in 78.6% cases, which are indeed monomer as reported by PiQSi but not by PISA.

References

1. Levy, E. D. (2007). PiQSi: protein quaternary structure investigation. *Structure* **15**, 1364-7.
2. Henrick, K. & Thornton, J. M. (1998). PQS: a protein quaternary structure file server. *Trends Biochem Sci* **23**, 358-61.
3. Krissinel, E. & Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J Mol Biol* **372**, 774-97.
4. Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**, 159-74.