International Journal for Pharmaceutical Research Scholars (IJPRS)

V-2, I-4, 2013 ISSN No: 2277 - 7873

RESEARCH ARTICLE

Proposed Enhanced Proteins Classification Databases

Kumar M*

**Ph.d Scholar, Shri Venkateshwara University, Uttar Pradesh, India.*

Manuscript No: IJPRS/V2/I4/00213, Received On: 25/11/2013, Accepted On: 02/12/2013

ABSTRACT

Proteins are classified according to both sequence and structure. The classes of protein include – Class α, class β, class α/β, and class α+β. Other classes include multi-domain (α and β) proteins and membrane and cell-surface proteins. Protein data bank is used to organize proteins into databases that include – SCOP, FSSP, DALI, Pfam, CATH, and MMDB. Databases promote keyword search, sequence search, navigation, hierarchy classification, and external online links. These databases are not consistent in determining which classes of proteins belong to the same family. Some proteins have been put in the same class despite the fact they have less robust relationship. It is essential for the available classification system to be compared and examine the classes to determine which proteins remain in the same family. In this study, different databases and signature types would be combined (more than 10 databases) in order to produce a powerful protein classification tool that would facilitate accurate prediction of protein function.

KEYWORDS

Hidden Markov Models, Fingerprints, Phylogenetics, Phylogenomic

INTRODUCTION

Proteins can be classified according to both sequence and structure[1]. Proteins have been classified based on conserved amino acid patterns[2]. There are two important considerations for protein classification and they include the following. First, different proteins from different evolutionary origins can fold into similar structure. Second, different proteins with some degree of similar sequences also share evolutionary origin and/or some structural features[3]. The extent of similarity between two proteins sequences can be based on the percentage of sequence identity and/or conversation[4]. For example, the blast programs are used to query sequences and scores where higher scores represent greater degree of similarity. Levitt and Chotia (1976) presented four principal classes of protein structures based on type and arrangement of secondary structural elements. These classes include – Class α (domains consisting of α-helices), class β (domains consisting of ß-sheets), class α/β (β sheets intervening α helices), class α+β (segregated α helices and anti-parallel β sheets). Other classes include multi-domain (α and β) proteins and membrane and cell-surface proteins, which exclude proteins of the immune system. Protein data bank is used to organize or classify proteins into database that promote keyword search, sequence search, navigation, hierarchy classification, and external online links[5]. These features allow databases to retrieval of structure-oriental information easily. Some of the identified databases include the following – SCOP, FSSP, DALI, Pfam, CATH,

***Address for Correspondence:**
Manish Kumar
Phd Scholar,
Shri Venkateshwara University,
Uttar Pradesh, India.
E-Mail Id: bioinfomoney@gmail.com

and MMDB. However, these databases are not consistent in determining which classes of proteins belong to the same family. It is important to develop a tool that ensure consistency and structural and sequence accuracy since some have been classified in families, which they have less robust relationships.

SCOP (Structural Classification of Proteins) database use hierarchical level classification system[6]. The protein structures are classified according to number of both evolutionary and structural relationships[6]. The evolutionary hierarchical level starts with family, super-family, then fold whereas the structural classes are Alpha (α), Beta (β), Alpha+Beta (α+ β), Alpha/Beta (α/β) and miscellaneous ‘small proteins’. FSSP (fold classification based on structure-structure alignment of proteins) database classifies proteins based on their pair-wise combinations i.e. structural alignment in the Brookhaven structural database. DALI (Distance matrix ALIgnment) database can identify similar folding patterns. It uses screening program to examine the entire PDB and identify similar structures to the newly analyzed structure[7]. It also classifies protein domain structures using all-against-all comparison mechanism. Pfam (protein families) database is a large collection of multiple protein sequences alignment and Profile hidden Markov models[8]. The latest version (Ver 27.0) has 14831 families consisting of 69% of proteins in SWISS-PROT 39 and TrEMBL 14 structural domains and predictions of the non-domain regions. This database is available on the World Wide Web and search tools supports taxonomy and domain search. Additionally, structural data are supported through multiple sequence alignment.

CATH (Classification by class, architecture, topology, and homology) database classifies proteins according to architecture, fold, family, and super-family[9]. It uses hierarchical levels similar to SCOP. However, CATH database groups Alpha/Beta (α/β) and Alpha+Beta (α+β) proteins in one class hence its fourth class consists of proteins with few secondary structures. MMDB (Molecular modeling database) utilizes Brookhaven PDB and VAST (Vector Alignment Search Tool) program to classify proteins of known structure into structural related groups[10]. The role of the VAST program is to align 3-dimension according to secondary structural elements and thus facilitate rapid identification of PDB structures that are not statistical.

Nevertheless, mutations such as gene duplication and rearrangement may give rise to new genes that transcribe newer proteins with different structure and functions[11]. In addition, proteins with different amino acid sequences may fold to form an active site around a substrate. An active site occurs within the tertiary (3-dimensional) or quaternary protein structure as a localized combination of amino acid side groups. The ideology of protein function (cascade of inference) flow from (order of) sequence, structure, and then function. Similar sequences of amino acids produce similar protein though the relationship between structures and function is more complex[12]. This is because proteins with similar sequence and structure can have different functions whereas proteins with different structure and sequence can have similar functions[11]. For instance, unrelated proteins with different folds can perform similar functions that related folds cannot[11]. The evolution of protein may lead to retention of function and specificity, retention of function only but alter specificity, alteration of the metabolic context of related function, or completely change the unrelated function.

The analysis of the databases reveals that there is no consistency among major databases such as CATH, SCOP, and Pfam among others. Additionally, new techniques for analysis and classification of proteins need to be tested for consistency before adoption. It is essential for the available classification system to be compared. It would also be important to examine the classes and determine which groups of proteins remain in the same family because some proteins have been classified in the same class despite the fact they have less robust relationship.

MATERIALS AND METHODS

In this study, different databases and signature types would be combined based on individual strengths in order to produce a powerful protein classification tool that would facilitate accurate prediction of protein function. A single searchable resource with high output accuracy and consistency would be created by bringing together or combining Hidden Markov Models (HMMs), fingerprints, profiles, patterns, and InterPro. The Gene3D is a HMM database that would provide structural domains. PIRSF, SMART, and Pfam databases are HMMs that would be incorporated to provide functional annotation of families and/or domains. The non-HMMs databases for functional annotation of families and/or domains would include Finger-Prints (PRINTS), Profiles (ProSite/ProDom). Protein features (site) would be provided by ProSite database. The system would program to facilitate Phylogenomic analysis.

RESULTS

The new alternative classification scheme for predicting protein sequences and structures would incorporate physical properties, new descriptors, contact orders, coordination numbers, and solvent accessibility in order to facilitate efficiency and consistency. This would make it a powerful classification tools since it would combines multiple databases and signatures, simplify and rationalize the process of analyzing protein sequences, and remove redundancy.

DISCUSSION

The proposed database would become a powerful classification since it combines multiple databases and signatures. The combined or integrated database would simplify and rationalize the process of analyzing protein sequences and remove redundancy. This would be made possible by combining and organizing information consistently. Additionally, matching of protein databases and signatures would provide extensive annotation and useful links. Phylogenemic analysis would enhance accuracy because of its high-accuracy functional annotation though requires technical expertise due to its complexity. Additionally, the phylogenetics tool identifies orthologs as the basis of annotation transfer.

ACKNOWLEDGEMENT

The authors are thankful to Shree P.M.Patel Institute of P.G. Studies and Research in Science (APMS) for providing research facilities. We are thankful to The Director, SICART, Vallabh Vidyanagar for analytical facilities. Authors are also grateful to The Director, SAIF-RSIC, Punjab University, and Chandigarh for NMR analysis. We are articulate our appreciation to IIT-Bombay for TG and DTG scan analysis facility.

CONCLUSION

The proposed integrated resource for protein motifs, families, and domains would provide a single and consistent database. The combination of different databases and signature types would produce a powerful protein classification tool and facilitate accurate prediction of protein function. Based on its description, the protein signature interface would use different method for protein signature derivation since databases such as CATH, SCOP, FSSP, DALI, Pfam, and MMDB.

REFERENCES

1. Patronov A, Dimitrov I, Flower DR, Doytchinova I, Peptide binding prediction for the human class II MHC allele HLA-DP2: a molecular docking approach. BMC Structural Biology, 2011, 11, 32, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3146810/

2. Greenbaum J, Sidney J, Chung J, Brander C, Peters B, Sette A, Functional classification of class II human leukocyte antigen (HLA) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes. Immunogenetics, 2011, 63(6), 325–335.

3. Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C et al., “SCOP database in 2004: refinements integrate

structure and sequence family data", Nucleic Acids Research, 2004, 32, D226–D229: http://www.ncbi.nlm.nih.gov/pubmed/14681400.

4. Saha I, Maulik U, Bandyopadhyay S, Plewczynski D, "Fuzzy clustering of physicochemical and biochemical properties of amino acids", Amino Acid, 2011, 43(2), 583–594
5. Plewczynski D, Basu S, Saha I, AMS 4.0, "Consensus prediction of post-translational modifications in protein sequences", Amino Acid, 2012, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3397139/
6. Andreeva A, Prlic A, Hubbard TJP, Murzin AG, "SISYPHUS – structural alignments for proteins with non-trivial relationships", Nucleic Acids Research, 2007, 35, D253–D259: http://www.ncbi.nlm.nih.gov/pubmed/17068077.
7. Holm L, Rosenstrom, P, "Dali server: conservation mapping in 3D", Nucleic Acids Research, 2010, 38, W545–W549: http://www.ncbi.nlm.nih.gov/pubmed/20457744.
8. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ et al., The Pfam protein families' database. Nucleic Acids Research, 2008, 36, D281-D288: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2238907/
9. Cuff AL, Ian S, Tony L, Clegg AB, Robert R, et al., "Extending CATH: increasing coverage of the protein structure universe and linking structure with function", Nucleic Acids Research, 2011, 39, D420–D426: http://www.ncbi.nlm.nih.gov/pubmed/21097779.
10. Madej T, Addess KJ, Bryant SH, MMDB: 3D structures and macromolecular interactions. Nucleic Acids Research, 2012, 40 (D1), D461-D464.

    http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3245041/
11. Lavelle DT, Pearson WR, Globally, "Unrelated protein sequences appear random", Bioinformatics, 2010, 26, 310–318.
12. Fisher MA, McKinley KL, Bradley LH, Viola SR, Hecht MH, "De Novo Designed Proteins from a Library of Artificial Sequences Function in Escherichia Coli and Enable Cell Growth", Plos One, 2011. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3014984/