

**DEVELOPMENT OF NOVEL ALGORITHMS FOR FUNCTIONAL ANNOTATION OF
PROTEIN SEQUENCES**

AMITA PATHAK



**DEPARTMENT OF CHEMISTRY
INDIAN INSTITUTE OF TECHNOLOGY DELHI
September 2021**

DEVELOPMENT OF NOVEL ALGORITHMS FOR FUNCTIONAL ANNOTATION OF
PROTEIN SEQUENCES

by

AMITA PATHAK

Department of Chemistry

Submitted

in fulfilment of requirements of the degree of Doctor of Philosophy

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

September 2021

Dedicated to the Almighty Power

Thanks for Taking Care of this Universe...

Certificate

This is to certify that the thesis entitled, “**DEVELOPMENT OF NOVEL ALGORITHMS FOR FUNCTIONAL ANNOTATION OF PROTEIN SEQUENCES**”, being submitted by **Ms. Amita Pathak** to the Indian Institute of Technology, Delhi, for the award of the degree of **Doctor of Philosophy** in Chemistry, is a record of bonafide research work carried out by her. Ms. Amita Pathak has worked under our guidance and supervision and has fulfilled the requirements for the submission of this thesis, which to our knowledge has reached the requisite standard. The results contained in this dissertation have not been submitted in part or full to any other University or Institute for the award of any degree or diploma.

Prof. B. Jayaram

Place: New Delhi

Date: 08-9-2021

Prof. Narayanan D Kurur

Department of Chemistry

Indian Institute of Technology Delhi

Acknowledgements

One often falls short of words while thanking people who crossed the paths of reaching the goal. Today, I am very pleased to take this opportunity to thank all the people for helping me throughout my PhD journey.

First and foremost, I would like to thank my supervisor, Prof. B. Jayaram for his constant support and guidance who not only helped me to become a better researcher but also inculcated the desire to pursue good quality research. I would like to thank him for trusting me by providing multiple opportunities to work independently. Without his encouragements, insights, and attention, I would never have been able to initiate or to end my thesis work without his guidance. His patience and encouragement have let me grow at my own pace. I again thank him wholeheartedly for providing a great work environment for active research.

I am extremely grateful to my co-supervisor Prof Narayanan D Kurur for his consistent moral support and tips on balancing life. His presence in the professional career not only helps me in growing and keeping myself calm during my PhD, also encouraged me for a healthy lifestyle, enlightens me for the life vision and long term goals.

I am thankful to all my SRC members (Prof. Siddharth Pandey, Prof. Shashank Deep and Prof. Tapan Chaudhuri) for their sound advice, support and keeping the PhD progress formalities process always smooth.

Special thanks to Prof Nalin Pant for his scientific inputs, encouragement and motivation for my research work.

I would like to thank my biggest encouragement source Ruby ma'am for showing me the path of success.

I am extremely grateful to my loving and supportive family members (Grands, parents and brothers) for making me believe in myself and standing by me in all the odds.

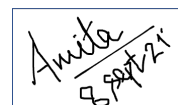
Not everyone is lucky to find a brother like figure in a friend. My spesh thanks to Pradeep Pant for his special love, unconditional support, and care. His contribution in my life (both personal & academic) is way beyond the pages of this thesis.

I am grateful to work with E. Avinash (a brilliant mind) for lifting my spirits and tried to fill me up with immense positive attitude. Without him this journey was almost impossible to dream of.

I am extremely glad to thank all my friends and colleagues for their immense support so far. Thanks to Akhilesh and Ruchika for keeping friendly lab environment. I would like to thank all my seniors Goutam, Avinash, Ashutosh, Anjali, Abhilash, Rahul, Ankita and Debarati (Sir/Ma'am) and staff members cum friends Manpreet Sir, Puneeta ma'am and Rao Sir. Special thanks to Shashank Sir for his always availability for any kind (technical or personal or proffessional advice) of help. At the same time I would like to acknowledge all my juniors (Sheenam, Arti, Naushad, Sahil, Vishvajeet, Toran, Bhumika, Smriti, Akshata, Devendra and Dinesh) for their immense love and respect.

Endless hangouts, food hunting, chats and emotional support of dear friends make life joyable. Thanks to Anand for his consistent support in all my needs. I would like to thank all my friends (Sonika, Moshi, Ekta, Shashank, Atanu, Niranjana, Sintoo, Rakesh, Supriya, Garima, Annu, Karuna and Deepak) for their presence in my life.

I would like to thank the faculties and staff members of Department of Chemistry, Indian Institute of Technology Delhi for their help and support during my PhD program. I gratefully acknowledge MHRD and IIT Delhi for supporting my financial needs and enabling my participation in international and national conferences.

A rectangular box containing a handwritten signature in black ink. The signature appears to be 'Amita' followed by a stylized surname, possibly 'Pathak', with a horizontal line crossing through the middle.

Amita Pathak

Abstract

Protein molecules are the workhorses of cells. An in-depth knowledge of protein function is essential to understand the working of a healthy organism and for mitigation of diseases. The most reliable form of function annotation is through biological experiments. However, the exponential rise in the number of protein sequences is making the job of experimental function detection a formidable challenge. With numerous methodological advancements in genome sequencing projects, number of new protein sequences reported has been steadily increasing. There are a total of 184,998,855 sequences in the protein sequence database. However, only ~69,306 entries are verified functionally using experimental methods. In this scenario, computational biologists must aid the experimental community in high-throughput function prediction. It would be immensely useful to gain insights into the activity of these biomolecules for guiding further experiments. Given that most of the drug targets are proteins, one of the pressing goals of protein function prediction is to aid in the development of novel drug targets and new therapeutics. While there are a total of 70,076 human proteins in UniProt and 25,995 experimentally determined structures, there are only ~3021 human protein targets known to bind with FDA approved drugs. Thus, predicting the function of all proteins is crucial to drug target identification and to initiate drug discovery. This thesis work focuses on the development of a few new computational protocols for protein function prediction, starting from sequence.

Chemical information residing in the protein sequences determines their structures and structures determine their functions. Hence, knowledge of both protein sequence and structure can be exploited for accurate function prediction. This thesis work focuses on evolving a methodology for function prediction using physicochemical features (Chapter-II). A comparison of diverse textbook classifications of amino acids in relation to function prediction is described and it is demonstrated that a new classification of amino acids (NCL) does better in function prediction (Chapter-III). Enzymes are vital protein molecules from both function and malfunction perspectives especially in the context of drug-discovery. An application of the NCL integrated with mask BLAST for enzyme function is demonstrated (Chapter-IV) and further extended to all proteins (69,306) with known function (Chapter-V). In function prediction, due to the rich diverse ideas and algorithms already generated, it pays to take a consensus view. In this spirit, I integrated

NCL-mask BLAST method developed here with alternate approaches to create a metasever, S2F, for function prediction which captures the merits of various methods (Chapter-VI). Perspectives emerging from the thesis work on accurate function prediction of millions of sequences in growing sequence databases are summarized (Chapter-VII).

सार

प्रोटीन अणु कोशिकाओं के वर्कहॉर्स होते हैं। एक स्वस्थ जीव के कार्य को समझने और रोगों के शमन के लिए प्रोटीन कार्य का गहन ज्ञान आवश्यक है। फ्रंक्शन एनोटेशन का सबसे विश्वसनीय रूप जैविक प्रयोगों के माध्यम से है। हालांकि, प्रोटीन अनुक्रमों की संख्या में घातीय वृद्धि प्रयोगात्मक कार्य का पता लगाने के काम को एक कठिन चुनौती बना रही है। जीनोम अनुक्रमण परियोजनाओं में कई पद्धतिगत प्रगति के साथ, रिपोर्ट किए गए नए प्रोटीन अनुक्रमों की संख्या लगातार बढ़ रही है। प्रोटीन अनुक्रम डेटाबेस में कुल 184,998,855 हैं। हालांकि, प्रयोगात्मक विधियों का उपयोग करके केवल 69,306 प्रविष्टियों को कार्यात्मक रूप से सत्यापित किया जाता है। इस परिदृश्य में, कम्प्यूटेशनल जीवविज्ञानी को उच्च-श्रुपट फ्रंक्शन भविष्यवाणी में प्रयोगात्मक समुदाय की सहायता करनी चाहिए। आगे के प्रयोगों का मार्गदर्शन करने के लिए इन जैव अणुओं की गतिविधि में अंतर्दृष्टि प्राप्त करना बेहद उपयोगी होगा। यह देखते हुए कि अधिकांश दवाओं की खोज के लिए लक्ष्य प्रोटीन हैं, प्रोटीन फ्रंक्शन भविष्यवाणी के दबाव लक्ष्यों में से एक उपन्यास दवा लक्ष्य और नए चिकित्सीय के विकास में सहायता करना है। जबकि यूनीप्रोट में कुल 70,076 मानव प्रोटीन और 25,995 प्रयोगात्मक रूप से निर्धारित संरचनाएं हैं, केवल 3021 मानव प्रोटीन लक्ष्य हैं जिन्हें FDA द्वारा अनुमोदित दवाओं के साथ बाँधने के लिए जाना जाता है। इस प्रकार, दवा लक्ष्य की पहचान और दवा की खोज शुरू करने के लिए सभी प्रोटीनों के कार्य की भविष्यवाणी करना महत्वपूर्ण है। यह थीसिस कार्य अनुक्रम से शुरू होने वाले प्रोटीन फ्रंक्शन भविष्यवाणी के लिए कुछ नए कम्प्यूटेशनल प्रोटोकॉल के विकास पर केंद्रित है।

प्रोटीन अनुक्रमों में रहने वाली रासायनिक जानकारी उनकी संरचनाओं को निर्धारित करती है और संरचनाएं उनके कार्यों को निर्धारित करती हैं। इसलिए, सटीक कार्य भविष्यवाणी के लिए प्रोटीन अनुक्रम और संरचना दोनों के ज्ञान का फायदा उठाया जा सकता है। यह थीसिस कार्य भौतिक रासायनिक विशेषताओं (अध्याय-II) का उपयोग करके कार्य भविष्यवाणी के लिए एक पद्धति विकसित करने पर केंद्रित है। कार्य भविष्यवाणी के संबंध में अमीनो एसिड के विविध पाठ्यपुस्तक वर्गीकरण की तुलना का वर्णन किया गया है और यह प्रदर्शित किया गया है कि अमीनो एसिड (एनसीएल) का एक नया वर्गीकरण फ्रंक्शन भविष्यवाणी (अध्याय-III) में बेहतर करता है। एंजाइम विशेष रूप से दवा-खोज के संदर्भ में कार्य और खराबी दोनों दृष्टिकोणों से महत्वपूर्ण प्रोटीन अणु हैं। एंजाइम फ्रंक्शन के लिए मास्क ब्लास्ट के साथ एकीकृत एनसीएल का एक अनुप्रयोग प्रदर्शित किया गया है (अध्याय-IV) और आगे ज्ञात फ्रंक्शन (अध्याय-V) के साथ सभी प्रोटीन (69,306) तक बढ़ाया गया है। फ्रंक्शन भविष्यवाणी में, समृद्ध विविध विचारों और पहले से उत्पन्न एल्गोरिदम के कारण, यह सर्वसम्मति से देखने के लिए भुगतान करता है। इस भावना में, मैंने फ्रंक्शन प्रेडिक्शन के लिए मेटासर्वर, S2F बनाने के लिए वैकल्पिक दृष्टिकोणों के साथ यहां विकसित

एनसीएल-मास्क ब्लास्ट विधि को एकीकृत किया, जो विभिन्न तरीकों (अध्याय-VI) के गुणों को पकड़ता है। थीसिस से उभरने वाले परिप्रेक्ष्य बढ़ते अनुक्रम डेटाबेस में लाखों अनुक्रमों की सटीक कार्य भविष्यवाणी पर संक्षेप में हैं (अध्याय-VII)।

Contents

Certificate.....	i
Acknowledgements.....	ii
Abstract.....	v
Contents	vii
List of Figures.....	x
List of Tables	xiv
Chapter I Introduction.....	1
1.1 PROTEIN SEQUENCE DATABASE	4
1.2 PROTEIN FUNCTION	5
1.3 SCHEMES FOR PROTEIN FUNCTION CLASSIFICATION.....	5
1.3.1 Gene Ontology (GO) Classification.....	6
1.4 PROTEIN SEQUENCE BASED FUNCTION PREDICTION.....	12
1.4.1 Homology based annotation transfer	14
1.4.2 Alternative approaches for protein annotation.....	15
1.5 PROTEIN FUNCTION DIVERGENCE AND CONVERGENCE	18
1.5.1 Functional divergence (Evolutionarily related enzyme superfamilies with different functions)	18
1.5.2 Functional convergence (Similarity in function between non homologous proteins) .	19
1.6 MOTIVATION AND SCOPE OF THE THESIS	20
REFERENCES	21
Chapter II A new maskBlast methodology for predicting function.....	30
2.1 MOTIVATION	31
2.2 The maskBLAST METHODOLOGY FOR PREDICTING FUNCTION	39
2.2.1 Implementation of New Chemical Logic (NCL) with BLASTp algorithm.....	40
2.3 DISCUSSION	44
REFERENCES	44
Chapter-III What is fundamental to amino acids?	50
3.1 MOTIVATION	51
3.2 METHODOLOGY	52
3.3 RESULTS AND DISCUSSION.....	54
3.3.1 Dataset.....	54
3.3.2 Comparison with BLAST	54

3.3.3 Comparison with some popular amino acid classifications	58
3.4 DISCUSSION	61
REFERENCES	62
Appendix I	64
Chapter IV <i>Seq2Enz</i>: An application of maskBLAST methodology with a new chemical logic of amino acids for improved enzyme function prediction.....	73
4.1 INTRODUCTION	74
4.2 MATERIALS AND METHODS.....	76
4.3 RESULTS AND DISCUSSION.....	77
4.3.1 Comparison with BLAST	77
4.3.2 Performance comparison with ECPred.....	79
4.4 A DESCRIPTION OF THE <i>Seq2Enz</i> WEB-SERVER	81
4.5 DISCUSSION AND CONCLUSION	81
REFERENCES	82
Chapter V A comparison of <i>Seq2Func</i> with other software on Protein Function Prediction.....	85
5.1 MOTIVATION	86
5.2 METHODOLOGY	87
5.3 RESULTS AND DISCUSSION	87
5.3.1 A comparison of <i>Seq2Func</i> with other protein function prediction software.....	90
5.3.2 WEBSERVER AVAILABILITY	93
5.3.3 DISCUSSION	95
REFERENCES	96
Chapter VI <i>S2F</i>: A meta server for sequence to function prediction.....	99
6.1 MOTIVATION	100
6.2 METHODOLOGY	102
6.2.1 Dataset.....	102
6.2.2 Sequence similarity based function prediction	103
6.2.3 Pattern/motif/domain-based function prediction	103
6.2.4 Phylogeny Based Methods.....	103
6.2.5 Network based Methods	103
6.2.6 Cluster-based function prediction	104
6.2.7 Chemistry-based function prediction.....	104
6.2.7 S2F method	104

6.3 RESULTS AND DISCUSSION	105
6.4 Webserver availability	108
6.5 DISCUSSION	108
REFERENCES	110
Chapter VII Conclusions and perspectives	112
Curriculum Vitae	115

List of Figures

Figure 1.1. There is a huge gap between the total reviewed protein sequences present in UniProt database and the number of proteins having experimentally known functions.....03

Figure 1.2. DAG (directed acyclic graph) flowchart shows the GO structure in three disjoint classes (namely molecular function, biological process and cellular component). The figure shows portions of all the three functional class structures.....07

Figure 1.3. Molecular function sub-levels present in UniProt showing number of proteins for each individual GO class.....10

Figure 1.4. Biological process sub-levels present in UniProt showing number of proteins for each individual GO class.....11

Figure 1.5. Cellular component sub-levels present in UniProt showing number of proteins for each individual GO class.....12

Figure 1.6. Total number of proteins available in the UniProt database for each organism versus the protein count with known functions. The light green bars represent the number of proteins with experimentally known functions and the light red bars show the total protein sequences of the particular organism.....13

Figure 1.7. Human protein count distribution in terms of total number of computationally predicted and experimentally determined protein sequences (Total Proteins/TrEMBL proteins), experimentally determined (SwissProt Proteins), and protein sequences with experimentally known functions (Annotated Proteins). The distribution of annotated proteins coded with experimental evidence codes is zoomed and plotted.....14

Figure 2.1. Generic structure of amino acids and peptide bonds. (a).C α is the central carbon atom that is covalently linked with the amino group (-NH₂), carboxylic group (-COOH), hydrogen (H), and side chain (R) group. R group is different for all the 20 amino acids. (b). Peptide bond

formation between two amino acids which is accompanied by the release one water (H₂O) molecule.....31

Figure 2.2. Amino acid classifications given by Lehninger and Voet and Voet [2, 3].....34

Figure 2.3. Amino acid classification based on New Chemical Logic (NCL). Properties of each amino acid are highlighted in cyan colour for g, d and s; and square bracket for l property; explaining the categorization of the class (g, d, s, l).....36

Figure 2.4. All the 20 amino acids representing four properties and their occurrence given by the New Chemical Logic (NCL) classification (Table 2.1).....39

Figure 2.5. Steps involved in implementing NCL+maskBLAST methodology for homolog detection leading to function prediction. (a) Creation of chemically fortified masked BLAST databases with annotated protein sequences. (b) Screening of modified query sequence against the databases. (c) Homolog detection and transfer of functions. (d) Compilation of predicted functions defining all levels of Gene Ontology (GO) terms and (e) Function prediction.....41

Figure 2.6. Stepwise commands for the implementation of NCL+maskBLAST methodology with NCL properties for protein function prediction.....44

Figure 3.1. Implementation of NCL based mask BLAST.....53

Figure 3.2. A comparison of NCL+maskBLAST methodology with BLAST at different sequence identities for function prediction for all the 69,306 experimentally annotated data available in the SwissProt database. Light blue (for Molecular function), light green (for biological process) and light pink (for cellular component) are shown by BLAST whereas dark blue (for Molecular function), dark green (for biological process) and dark red (for cellular component) are the improved Fmax values of NCL+maskBLAST methodology.55

Figure 3.3. A comparison of NCL+maskBLAST methodology with BLAST with the filter of E-value of 0.001 for function prediction for all the 69,306 experimentally annotated data available in the SwissProt database.....56

Figure 3.4. A comparison of NCL+maskBLAST methodology with BLAST with the filter of E-value-100 for function prediction for all the 69,306 experimentally annotated data available in the SwissProt database.....57

Figure 4.1. A comparison of enzyme function prediction with standard BLAST and *Seq2Enz* on all (267,362) experimentally characterized protein sequences available in the UniProt/SwissProt database, in terms of precision achieved for all the levels of enzyme functions (Level 0: enzyme or non-enzyme, Level 1: main class, Level 2: subclass, Level 3: sub-subclass and Level 4: substrate classes). All the solid bars shows precision, recall and Fmax values (orange, blue and green respectively) of BLAST results. Checked bars denote precision, recall and Fmax values (orange, blue and green respectively) of *Seq2Enz* results. Performance of *Seq2Enz* is consistently better for all the enzyme levels in terms of precision and Fmax.....79

Figure 4.2. A comparison of *Seq2Enz* and ECPred software on 7,062 proteins deposited between 1st July 2017 and 30th Jan. 2020 in the UniProt/SwissProt database in terms of Precision, Recall, and Fmax achieved for enzyme function prediction at all levels (Level0: enzyme or non-enzyme, Level 1: main class, Level 2: subclass, Level 3: sub-subclass and Level 4: substrate classes)....80

Figure 4.3. Front-end of the *Seq2Enz* developed for enzyme class identification and a snapshot of the results provided by *Seq2Enz* server in terms of enzyme class number. Details of the enzyme class can be retrieved by clicking on the result. The user is redirected to the <https://enzyme.expasy.org/EC/> [20] page for further details on the enzyme.....81

Figure 5.1. A comparison of *Seq2Func* and PFP in terms of Precision, Recall and Fscore for all the level of functions i.e., Molecular Function, Biological Process and Cellular Component....93

Figure 5.2. Front-end of the web page of *Seq2Func* software. The box takes the input protein sequence and submit button allows users to fetch the results.....94

Figure 5.3. A snapshot of the result page of the *Seq2Func*. The output is in the form of GO terms. Each GO term is linked with the descriptive explanation and allows user to get the insights of the function of the input protein sequence. The green highlighted level shows the associated GO terms. Click on the left mouse button facilitates visualization of the details such as MOL or BIO or CEL of GO terms.95

Figure 5.4. A snapshot of the redirected page of the results of all the levels of function. The details of the gene ontology identifiers can be obtained by clicking over the ids that will redirect to the AmiGo page provided by the GO consortium. This way user will be able to go into the differentially expressed (DE) gene description form.....95

Figure 6.1. Front-end of the *S2F*metaserver developed for consensus function prediction of proteins.....108

List of Tables

Table 2.1. Representation of unique classification of the side chain of amino acids based on the New Chemical Logic.....	37
---	----

Table 3.1. A comparison of all the combinations of NCL properties, i.e., DSL_MOL (Non_G), GDL_MOL (Non_S), GDS_MOL (Non_L), GSL_MOL (Non_D), GDSL_MOL (all four together); leaving one out each time for function prediction for all the experimentally annotated data available in the SwissProt database.....	58
--	----

Table 3.2. A comparison of BLAST and masked BLAST results with Lehninger, Voet and Voet, and NCL classification on all the protein sequences (69,306) with known experimental annotation available in SwissProt.....	59
---	----

Table 3.3. A comparison of BLAST and masked BLAST results with Lehninger, Voet and Voet, and NCL classifications at 40% sequence identity (33,003) of all the protein sequences with known experimental annotation available in SwissProt.....	60
---	----

Table 4.1. An overview of the annotated (i.e. experimentally verified) and unannotated (provisionally annotated through computational methods) protein sequences with the EC Class Division [6].....	75
---	----

Table 5.1. A comparison of <i>Seq2Func</i> and other software for all the available 69,306 proteins with experimentally known functions for Molecular Function (MOL).....	91
--	----

Table 5.2. A comparison of <i>Seq2Func</i> and other software for all the available 69,306 proteins with experimentally known functions for Biological Process (BIO).....	91
--	----

Table 5.3. A comparison of <i>Seq2Func</i> and other software for all the available 69,306 proteins with experimentally known functions for Cellular Component (CEL).....	91
--	----

Table 6.1. A list of different software for predicting protein function. Information on availability of access to the software (Web server or open-source codes availability) is given in 2nd and 3rd columns. All the softwares are arranged in chronological order of their respective publication date.....101

Table 6.2. A comparison of *S2F* with other software for all the available (69,306)proteins with known functions in terms of molecular function (MOL).....106

Table 6.3. A comparison of *S2F* with other software for all the available (69,306)proteins with known functions in terms of biological process (BIO).....107

Table 6.4. A comparison of *S2F* with other software for all the available (69,306)proteins with known functions in terms of cellular component (CEL).....107