Graduate School ETD Form 9
(Revised 12/07)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Sirisha Peyyeti

Entitled
Identification of Publications on Disordered Proteins from PubMed

For the degree of Master of Science

Is approved by the final examining committee:

Dr. Yuni Xia
Chair

Dr. Keith Dunker

Dr. Jake Chen

To the best of my knowledge and as understood by the student in the *Research Integrity and Copyright Disclaimer (Graduate School Form 20)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): Dr. Yuni Xia

Approved by: Dr. Shiaofen Fang 07/14/2011
Head of the Graduate Program Date

Graduate School Form 20
(Revised 9/10)

PURDUE UNIVERSITY
GRADUATE SCHOOL

Research Integrity and Copyright Disclaimer

Title of Thesis/Dissertation:

Identification of Publications on Disordered Proteins from PubMed

For the degree of Master of Science

I certify that in the preparation of this thesis, I have observed the provisions of *Purdue University Executive Memorandum No. C-22,* September 6, 1991, *Policy on Integrity in Research.**

Further, I certify that this work is free of plagiarism and all materials appearing in this thesis/dissertation have been properly quoted and attributed.

I certify that all copyrighted material incorporated into this thesis/dissertation is in compliance with the United States' copyright law and that I have received written permission from the copyright owners for my use of their work, which is beyond the scope of the law. I agree to indemnify and save harmless Purdue University from any and all claims that may be asserted or that may arise from any copyright violation.

Sirisha Peyyeti

Printed Name and Signature of Candidate

07/14/2011

Date (month/day/year)

*Located at http://www.purdue.edu/policies/pages/teach_res_outreach/c_22.html

IDENTIFICATION OF PUBLICATIONS ON DISORDERED PROTEINS FROM PUBMED

A Thesis

Submitted to the Faculty

of

Purdue University

by

Sirisha Peyyeti

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

August 2011

Purdue University

Indianapolis, Indiana

Dedicated to My Husband, In Laws, Parents and Sister.

ACKNOWLEDGEMENTS

I would like to convey my sincere thanks and gratitude to my committee chair and advisor, Dr. Yuni Xia, for her patience, continuous guidance and technical support through the course of my research work. I specially thank Dr. Keith Dunker and Dr. Jake Chen for their time, interest and support in introducing me to the world of bio-informatics and disordered proteins.

In addition, I would like to thank Dr. Robert W. Williams and Ms. Caron Morales. They both provided much encouragement and were great mentors and often provided much needed support and ideas.

I would also like to thank NSF for supporting my research and the architects of NLProt for sharing their protein search tool. Finally, I would like to thank the entire faculty and staff at Computer Science department and at the Center for Computational Biology and Bioinformatics for being helpful at all times.

TABLE OF CONTENTS

Page
LIST OF FIGURES vi

ABSTRACT viii

CHAPTER 1 INTRODUCTION 1
1.1 Introduction 1
1.2 Significance 4
1.3 Assumptions 5

CHAPTER 2 PROBLEM DISCUSSION AND LITERATURE REVIEW 6
2.1 Problem Discussion 6
2.2 Identifying Protein Names 7
2.2.1. Rule Based Systems 7
2.2.2. Machine Learning Systems 7
2.2.3 Dictionary Based Systems 8
2.3 Available Software Tools for Identifying Protein Names 9
2.3.1 Banner 9
2.3.2 ABNER 9
2.3.3 LingPipe 12
2.3.4 NLPROT 12
2.3.5 A Comparison of Existing Techniques to Identify Protein Names 12
2.4 Disorder Predictors 13

Page

CHAPTER 3 SYSTEM AND METHODS 15
3.1 Identifying Publications.....15
3.2 Datasets.....17
3.3 Tests and Results 18

CHAPTER 4 DISCUSSION.....28

CHAPTER 5 USING DISPROT 29
5.1 Work Flow Diagram 29
5.2 Step by Step Description 30

LIST OF REFERENCES.....37

APPENDIX.....42

LIST OF FIGURES

Figure Page

Figure 1.1 Number of publications retrieved from PubMed using keyword search.....3

Figure 2.1 Sample Result from Banner.....10

Figure 2.2 Sample Result from ABNER.....11

Figure 2.3 Sample Result from NLProt 14

Figure 3.1 A graph showing number of structured proteins having 25 consecutive disordered amino acids 20

Figure 3.2 Overall disorder percentages in the 100 structured proteins 21

Figure 3.3 A graph showing the total length of the protein 22

Figure 3.4 Score distribution for the test on 100 DisProt abstracts 23

Figure 3.5 Number of publications ranked as relevant 24

Figure 3.6 Number of true and false positives in identifying relevant abstracts 25

Figure 3.7 Number of true and false positives in identifying relevant abstracts 26

Figure 3.8 A comparative analysis of sensitivity, specificity and accuracy 27

Figure 5.1 Workflow for the algorithm.....29

Figure 5.2 A screen shot of abstracts upload mechanism 32

Figure 5.3 A screen shot of pre-processed abstracts.....33

Figure 5.4 A screen shot of NLProt output.....34

Figure 5.5 A screen shot of a abstract in the output.....35

Figure 5.6 A screen shot of final output 36

LIST OF ABBREVIATIONS

IDP	Intrinsically Disordered Protein
IDPs	Intrinsically Disordered Proteins
IDR	Intrinsically Disordered Region
IDRs	Intrinsically Disordered Regions

ABSTRACT

Sirisha, Peyyeti. M.S., Purdue University, August 2011. Identification of Publications on Disordered Proteins from PubMed. Major Professor: Yuni Xia.

The literature corresponding to disordered proteins has been on a rise. As the number of publications increase, the time and effort needed to manually identify the relevant publications and protein information to add to centralized repository (called DisProt) is becoming arduous and critical. Existing search facilities on PubMed can retrieve a seemingly large number of publications based on keywords and does not have any support for ranking them based on the probability of the protein names mentioned in a given abstract being added to DisProt. This thesis explores a novel system of using disorder predictors and context based dictionary methods to quickly identify publications on disordered proteins from the PubMed database.

NLProt, which is built around Support Vector Machines, is used to identify protein names and PONDR-FIT which is an Artificial Neural Network based meta-predictor is used for identifying protein disorder. The work done in this thesis is of immediate significance in identifying disordered protein names.

We have tested the new system on 100 abstracts from DisProt [these abstracts were found to be relevant to disordered proteins and were added to DisProt manually by the annotators.] This system had an accuracy of 87% on this test set. We then took another 100 recently added abstracts from PubMed and ran our algorithm on them. This time it had an accuracy of 68%. We suggested improvements to increase the accuracy and believe that this system can be applied for identifying disordered proteins from literature.

CHAPTER 1 INTRODUCTION

1.1 Introduction

The Experiments and predictors developed by numerous researchers have shown that many proteins lack rigid 3D structure under physiological conditions in vitro, existing instead as dynamic ensembles of inter converting structures that we are calling intrinsically disordered (ID) proteins [1, 2]. Indeed, the literature published on these ID proteins is virtually exploding (see Figure 1.1). This literature explosion is consistent with bio-informatics studies indicating that about 25 to 30% of eukaryotic proteins are mostly disordered [3], that more than half of eukaryotic proteins have long regions of disorder [3, 4], and that more than 70% of signaling proteins have long disordered regions [5].

DisProt is a database that is aimed at becoming a central repository of disorder related information [6, 7] and it makes a best effort in providing structure and function information about proteins that lack a fixed 3D structure under putatively native conditions, either in their entireties or in part. There are currently 643 disordered proteins and 1375 disordered regions in DisProt. The number of publications shown in Figure 1.1 indicates that there are even more disordered proteins than the numbers indicated in DisProt. Owing to the exponential rise in publications, it is a difficult, time consuming and resource intensive manual task to be abreast with the publications and to read them to identify the most relevant abstracts. Having an automated method to estimate relevance of a PubMed publication to be a new DisProt entry and extract the protein information would significantly contribute to increasing the number of entries in DisProt and reduce the amount of manual work required by annotators to add new proteins.

In this thesis we aimed to take the current state of art of identifying disordered protein names a step forward by applying the concepts of search relevance ranking, protein name extraction and disorder predictors.

As an exploratory study, we selected three key features to estimate relevance:

a. An expansive set of keywords that would describe the structure of a disordered protein.
b. Listing of the detection methods that are used for identifying disordered proteins.
c. PONDR-FIT disorder prediction score for the proteins mentioned in the publication.

We tested this idea on a set of 100 abstracts from DisProt and we could identify the abstracts related to disordered proteins with 87% accuracy. We repeated the test on a set of 100 abstracts from PubMed and had an accuracy of only 60% because of high amount false positives by the feature c. We studied the results of the test and made an observation that not all abstracts having a disordered protein present in the abstract, discuss about the structure or experimental methods of the disordered protein and one of the criterion for adding publications to DisProt is that the publication should be discussing about the structure of a disordered protein or an experimental result performed on a disordered protein.

So, we modified the algorithm to first identify papers that discuss about the structure or an experimental method for a disordered protein and then to check if the selected papers have a protein name. If they have a protein name, we try to determine the chance of that protein being disordered based on its proximity from the protein search terms, detection methods and the prediction results of PONDR. We tested this modified algorithm on the 100 abstracts from PubMed and had 70% accuracy.

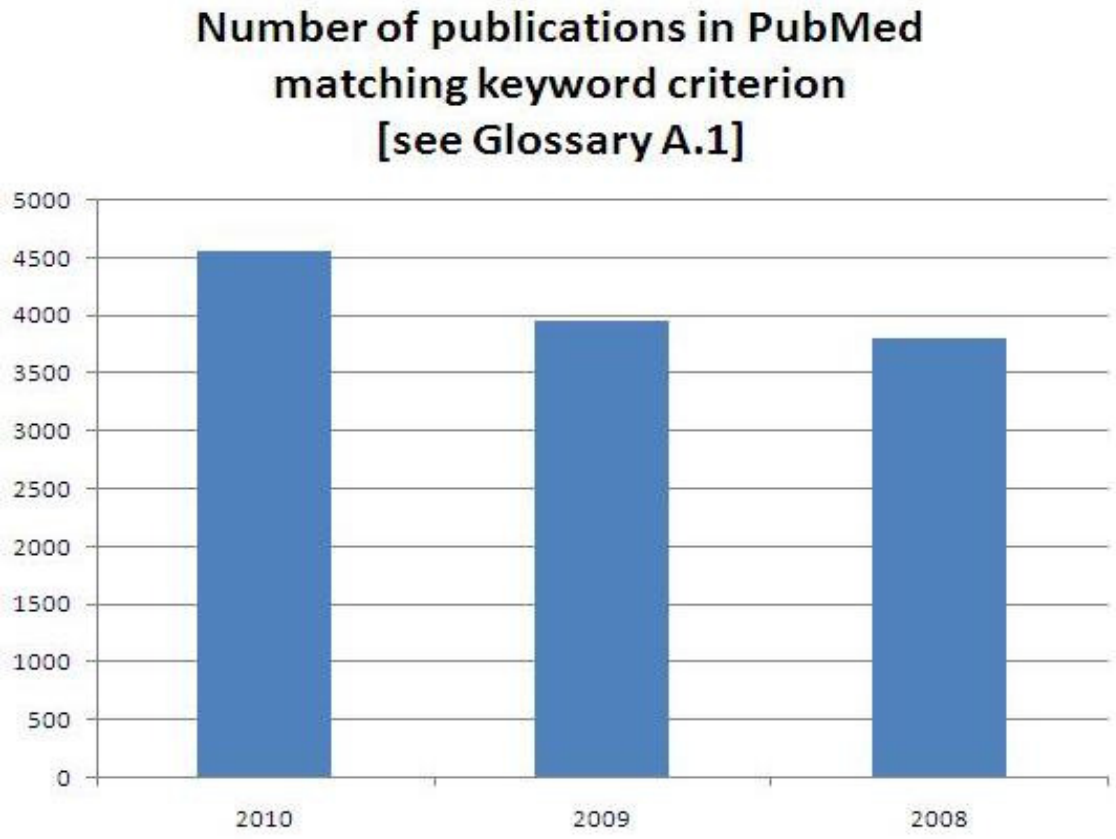

Figure 1.1 Number of publications retrieved from PubMed using keyword search.

1.2 Significance

One of the methods that investigators working on DisProt use to identify disordered proteins is literature search, specifically by searching the PubMed using the keywords. [See Appendix A.1]

This is one of the methods that have worked well so far. Nearly 50000 abstracts were found on PubMed by querying PubMed for the search terms mentioned in Appendix A.1 and manually reading each of these abstracts to identify disordered protein names would be a difficult and time consuming task. 5.7 Release of DisProt has 643 disordered protein entries and 1375 disordered regions. So, there is a good probability that this work would assist in identifying highly relevant abstracts and reduce the number of papers that will require reading by human experts. The work done in this thesis can be of immediate use to the annotators at DisProt and can assist in increasing the entries in DisProt, a widely used public database of protein disorder.

1.3 Assumptions

The following are the assumptions in the study:

a. Either the protein name or the search terms mentioned under significance sections or the detection methods used for finding a disordered protein occurs in the abstract.

b. The protein name that is closest to the disorder search term is the disordered protein that the author of the paper is referring to.

c. NLProt [8, 9] is used in this study to identify the protein names from abstract. We have found in our preliminary tests that it can identify protein names with 88.7% accuracy but while designing the disorder protein identification algorithm we assumed NLProt has accurately identified all the protein names and have built our algorithm on top of it.

d. PONDR-FIT [10] is used in this study to identify protein disorder. We tested the results of PONDR-FIT on 100 structured and 100 unstructured proteins and made the assumption that at least 25 consecutive segments in the protein sequence are predicted as disordered or segments comprising of at least 25% of overall protein length are predicted as disordered by PONDR-FIT.

CHAPTER 2 PROBLEM DISCUSSION AND LITERATURE REVIEW

2.1 Problem Discussion

The problem that we are addressing in this thesis is to identify the publications returned from PubMed search based on their relevance to a disordered protein. We proposed to solve this problem by assigning a higher score to a publication that has mention about disordered proteins. So, our problem is to identify disordered proteins from publications. We subdivided this problem into two problems:

a. Problem 1 Identifying protein names from publications.

b. Problem 2 Predicting if a protein is disordered.

Considerable amount of work has been done and number of approaches has been proposed on both the problems. A brief literature review is presented in Section 2.2.

2.2 Identifying Protein Names

A number of methods have been proposed for identifying protein names from scientific abstracts. They differ in their degree of reliance on dictionaries, Statistical or Knowledge Based approaches, and in the rule generation mechanism (manual vs. automatic). All methods can be roughly split into three categories: Dictionary Based approaches, Rule Based approaches, and Machine Learning approaches, and although some interesting mixed systems have also been described.

2.2.1. Rule Based Systems

Rule Based Systems rely on a set of expert derived rules, which may combine word alphanumerical composition, presence of special symbols, and capitalization with word syntactic and semantic properties, to initiate, extend, and terminate the chains of sentence tokens. Some systems can also use small dictionaries to improve precision and recall. Examples of Rule Based Systems are presented in:

a. Narayanaswamy et al. [11] (precision 96%, recall 62%)

b. Fukuda et al. [12] (precision 40%, recall 40%)

c. And Franzen et al. [13] (precision 68%, recall %).

d. Seki and Mostafa [14] used surface clues to anchor a protein name, but instead of syntactic features they used word first order transition probabilities learned from annotated test corpora used in the original match. The reported precision and recall rates are 60% and 66%, respectively.

2.2.2. Machine Learning Systems

Machine Learning approaches rely on the presence of an expert annotated training corpus to automatically derive the identification rules by means of various statistical algorithms. The features used in Machine Learning methods are mostly the same as those in Rule Based approaches: surface clues, parts of speech, and, sometimes, semantic word properties obtained from rough classification. Nobata et al. [15] used Bayesian classifier and decision tree algorithms to identify a noun phrase as a protein, based on its word composition. They report an F-score of 70% to 80% for protein detection. Collier et al. [16] used a first order hidden Markov model (HMM) trained on annotated corpus to detect the protein names in text and report a 76% F-score.

Kazama et al. [17] applied support vector machines to the same problem and achieved a 65% F-score. Burr Settles et al. [21] applied linear chain 1st order conditional random fields and achieved 72.46 F-score. Robert Leaman et al. [22] applied second order conditional random fields and achieved 81.96% F-score.

2.2.3 Dictionary Based Systems

Dictionary Based approaches utilize a provided list of protein terms to identify protein occurrences in a text, usually by means of various substring matching techniques. Proux et al. [19] used a Drosophila protein dictionary derived from a fly base for identification of proteins with 91% precision and 94% recall. However, they recognized only single word protein names. They also reported that precision of the system dropped from 91% to 70% when transferred from a corpus of sentences from fly base to a more general set of Medline articles. An interesting combination of the Dictionary Based approach with the Basic Local Alignment Search Tool (BLAST) based identification algorithm has been proposed by Krauthammer et al. [20]. The basic idea was to perform an approximate string match after converting both input text and a dictionary into the DNA sequence like strings. The authors reported 79% recall and 72% precision.

2.2.4 A Hybrid System

An interesting combination of a Machine Learning approach with hand crafted rules is reported in Tanabe and Wilbur [18]. As a first step, the transformation- based part of speech tagger has been trained on the corpus of Medline sentences with hand marked gene occurrences to induce the rules for tagging the text. Next, a complex set of manually derived contextual, morphologic, and Dictionary Based post processing rules have been applied. Reported precision and recall are 86% and 67%, respectively. Sven Mika, Burkhart Rost et al. [8, 9] developed a system based on support vector machines. Additionally filtering rules and protein name dictionary are used to improve performance. Reported precision and recall are 70% and 85% respectively.

2.3 Available Software Tools for Identifying Protein Names

Many Solutions have been implemented for protein name identification, among them NLProt [8, 9] produces good result by combining dictionary based method and support vector machines, Banner [22], ABNER [21], GENIA [36] also produce good results based on Conditional Random Fields. Few of these solutions are described and compared in the following subsections. These software tools are open sourced or licensed under GPL and are available freely for research and educational purposes.

2.3.1 Banner

It is an open-sourced, executable survey of advances in biomedical named entity recognition, intended to serve as a benchmark for the field. It is implemented in Java as a machine-learning system based on conditional random fields and includes a wide survey of the best techniques recently described in the literature. It is designed to maximize domain independence by not employing brittle semantic features or rule-based processing steps. The details of the system are described in this paper. A sample output is shown in Figure 2.1.

2.3.2 ABNER

ABNER is a software tool for molecular biology text analysis. It began as a user-friendly interface for a system developed as part of the NLPBA / BioNLP 2004 Shared Task challenge. The details of that system are described in (Settles, 2004) [21]. At ABNER's core is a statistical machine learning system using linear-chain conditional random fields (CRFs) with a variety of orthographic and contextual features. Version 1.5 includes two models trained on the NLPBA and Bio Creative corpora, for which performance is roughly state of the art (F1 scores of 70.5 and 69.9 respectively). The new version also includes a Java API allowing users to incorporate ABNER into their systems, as well as train and use models for other data. A sample output is shown in Figure 2.2.

```
p35|O /|O cdk5binds|B-GENE and|O
phosphorylates|B-GENE beta|I-GENE -|O
cateninand|O regulates|O beta|B-GENE -|I-GENE
catenin|I-GENE /|O presenilin|B-GENE -|I-GENE
1interaction|I-GENE .|O The|O neuronal|O
cyclin|B-GENE -|I-GENE dependent|I-GENE
kinase|I-GENE p35|I-GENE /|O cdk5comprises|B-
GENE a|O catalytic|O subunit|O {|O cdk5|B-
GENE }|O and|O an|O activator|O subunit|O {|O
p35|O }|O .|O To|O identify|O novel|O p35|O /|O
cdk5substrates|B-GENE ,|O we|O utilized|O the|O
yeast|O two|O -|O hybrid|O system|O to|O
screen|O for|O human|O p35binding|B-GENE
partners|O .|O From|O one|O such|O screen|O ,|O
we|O identified|O beta|B-GENE -|I-GENE
cateninas|I-GENE an|O interacting|O
protein|O .|O Confirmation|O that|O p35binds|B-
GENE to|O beta|B-GENE -|I-GENE cateninwas|I-GENE
obtained|O by|O using|O glutathione|B-GENE S|I-
GENE -|I-GENE transferase|I-GENE {|O GST|B-
GENE }|O -|O beta|O -|O cateninfusion|O
proteins|O that|O interacted|O with|O both|O
endogenous|O and|O transfected|O p35|O ,|O and|O
by|O showing|O that|O beta|O -|O cateninwas|O
present|O in|O p35immunoprecipitates|O .|O
p35and|B-GENE beta|
```

Figure 2.1 Sample Result from Banner.

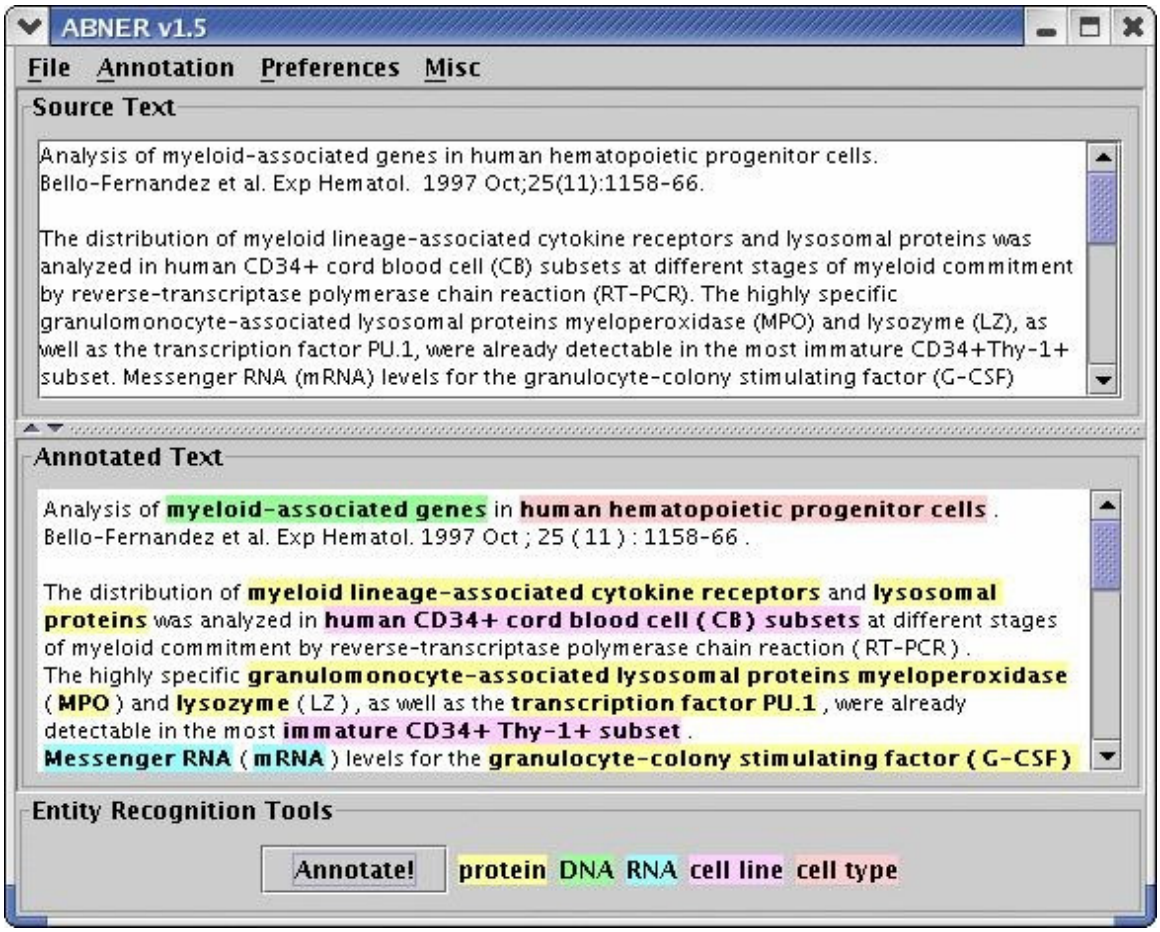

Figure 2.2 Sample Result from ABNER.

2.3.3 LingPipe

Lingpipe is open source Natural Language Processing software that is developed by Alias-I, incorporated (http://www.alias-i.com/). LingPipe is regarded as "a suite of Java tools designed to perform linguistic analysis on natural language data." LingPipe provides linguistic analysis functions such as sentence boundary detection and named entity detection using first order hidden markov models.

2.3.4 NLPROT

NLProt is a novel system that combines Dictionary and Rule Based filtering with several support vector machines (SVMs) to tag protein names in PubMed abstracts. When considering partially tagged names as errors, NLProt still reached a precision of 70% at a recall of 85%. By many criteria this system outperformed other tagging methods significantly; in particular, it proved very reliable even for novel names. Input can be PubMed or MEDLINE identifiers, authors, titles and journals, as well as collections of abstracts or entire papers. A sample output of NLProt is shown in Figure 2.3.

2.3.5 A Comparison of Existing Techniques to Identify Protein Names

Compared with Rule Based approaches, Dictionary Based protein identification systems are more accurate, and their performance is in direct correlation with the quality and completeness of the provided protein dictionaries. Development and maintenance of comprehensive protein name dictionaries is not simple task because new proteins are constantly being identified. However, both Machine Learning and Rule Based approaches require significant amounts of expert work for creation of rules and manual tagging of the training corpus respectively.

2.4 Disorder Predictors

One approach that has been important in the study if IDPs and IDRs is the use of disorder predictors. This method is extremely powerful in terms of time and cost of study of disordered proteins compared to traditional experimental methods [23–26]. More than 50 predictors have been developed by now [27]. These include the early PONDR series [28–30], DisEMBL [31], DISOPRED [32], POODLE [33], DISPro [34], IUPRED [35] and PONDR-FIT [10]. PONDR-FIT was assembled by combining PONDR-VLXT, PONDR-VSL2 and PONDR-VL3 and the authors of PONDR-FIT have reported an increase in accuracy in the aggregate as compared to the individual component predictors.

ID: 631

Intrinsically disordered proteins can form highly dynamic complexes with partner proteins. One such dynamic complex involves the intrinsi
its partner Cdc4 in regulation of yeast cell cycle progression. Phosphorylation of six N-terminal Sic1 sites leads to equilibrium engagem
site with the primary binding pocket in Cdc4, the substrate recognition subunit of a ubiquitin ligase. ENSEMBLE calculations using experi
resonance and small-angle X-ray scattering data reveal significant transient structure in both phosphorylation states of the isolated ens
modulates their electrostatic potential, suggesting a structural basis for the proposed strong contribution of electrostatics to binding.
dynamic pSic1-Cdc4 complex demonstrates the spatial arrangements in the ubiquitin ligase complex. These results provide a physical pictur
predominantly disordered in both its free and bound states, enabling aspects of its structure/function relationship to be elucidated.

The following protein names could be found by NLProt:

Organism

SVM score

UniProt Cross Ref

NAME	ORGANISM	TXT-POS	SCORE	METHOD	DB-ID(S)
Cdc4	yeast	25	0.461	SVM	cc4_yeast (100%)
Cdc4	yeast	54	1.061	SVM	cc4_yeast (100%)
ubiquitin	yeast	62	0.234	SVM	ubiq_yeast (100%)
pSic1	yeast	92	0.881	SVM	SIC1_YEAST (1%)
ubiquitin	yeast	128	0.686	SVM	ubiq_yeast (100%)

Figure 2.3 Sample Result from NLProt.

CHAPTER 3 SYSTEM AND METHODS

3.1 Identifying Publications

Three different features are used to rank the publications returned from PubMed search:

1. feature 1 - keywords that would describe the structure or property of a disordered proteins

   To benefit from the advantages of the frequently occurring words occurring in the context of describing disordered proteins, we compiled a list of keywords. These words were compiled under the guidance of an annotator who is experienced in manually reading the publications and identifying proteins. [Refer to Appendix A.2 for a listing of the keywords used.]

2. feature 2 - keywords that would describe the detection methods that are commonly used for identifying disordered proteins

   These words were compiled by using a combination of the detection methods that are currently present in DisProt database and from the listing of detection methods by Uversky. [Refer to Appendix A.3 for a listing of the keywords used.]

3. feature 3 - prediction result of a disorder predictor

   Using disorder predictors has been powerful in terms of time and cost to the study of disordered proteins compared to traditional experimental methods [23–26]. So, we are using NLProt [8, 9] to extract protein names and their SwissProt ids, We use the SwissProt ID to get the protein sequence in fasta format from UniProt and use this sequence as an input to PONDR-FIT [10]. PONDR-FIT returns a score for each amino acid in the sequence and we use the following criterion to make the decision of whether the protein is disordered or structured.

a) Criterion a PONDR-FIT has predicted at least 25 consecutive amino acids of a protein as disordered.

b) Criterion b PONDR-FIT has predicted 25% of the complete sequence of a protein to be disordered.

c) Criterion c The protein that is closest(in terms of number of words between protein name and search terms) to the search terms found by feature 1 or feature 2. A score is assigned to each publication by using a scoring mechanism based on features 1, 2 and 3 as follows:

Score = 1, if feature 1 is present
= 1, if feature 2 is present
= 1, if feature 3 is identified
= 2, if any two of features 1, 2 and 3 are present
= 3, if all the three features are present

One of the motivations for this work is to add new proteins to DisProt. If all the protein names found by NLProt are already in DisProt, then we assign a score of -1 and put these into a classification that is different from the ones used in this section, so that the annotator can focus on new protein names first and then come back and check these later. Publications are ranked based on the score in descending order, i.e., annotator would first see publications with score 3 and then those with scores 2 and 1.

3.2 Datasets

Three different datasets are used in this study:

Dataset-1 consisted of 100 abstracts that are cited as references in DisProt. These 100 abstracts were found to be relevant to disordered proteins by annotators and were manually added to DisProt.

Dataset-2 consisted of 100 results from PubMed keyword search. (keywords mentioned in section A1 in Appendix are used for this search).

Dataset-3 consisted of 100 completely ordered sequences. This dataset had the names and sequences of 100 completely structured proteins.

3.3 Tests and Results

Test1: To test the correctness of the thresholds used on PONDR score. We have tested PONDR-FIT output by using the thresholds [mentioned in Criterion a and Criterion b of Section 3.1] on 100 completely structured proteins and found that PONDR-FIT predicted 92 of the 100 proteins as structured. The graphs in Figures 3.1, 3.2, 3.3 show the distribution of consecutive disordered amino acid lengths and the overall disorder score for these 100 structured proteins.

Test2: To test the correctness of the features selected for identifying publications to be added to DisProt. We first tested for the correctness of each of the individual features on 100 abstracts from DisProt and the results were as follows:

a. 63 abstracts were ranked with scores 1 or higher when we used just feature1.
b. 39 abstracts were ranked with scores 1 or higher when we used just feature2.
c. 81 abstracts were ranked with scores 1 or higher when we used just feature3.

We then tested the algorithm mentioned in Section 3.1 on 100 abstracts from DisProt. We used features 1, 2 and 3 for testing. 87 out of these 100 abstracts with scores 1 and more. 29 abstracts were ranked with score 3, 38 were ranked with score 2 and 20 were ranked with score 1. [Refer to scoring criterion mentioned in Section 3.1]. Figure 3.4 shows the distribution of scores for these abstracts. The Venn diagram in Figure 3.5 shows the results from Tests 2a, 2b, 2c and their overlap.

Test3: We tested the algorithm on 100 most recently added publications from PubMed [We used all the three features in this test]. An annotator at DisProt manually read this abstracts and identified 42 of them as the ones that may potentially be added to DisProt. The algorithm reported 72 abstracts with non-zero score and had an accuracy of 60.6%.

We discussed the results with our annotator and found that some of the false positives are due to the fact that the algorithm is scoring abstracts that have a disordered protein name (identified by as a result of using NLProt and PONDR-FIT) and not a disorder structure or experiment related term. The distribution of true positives and false positives for each of the feature is shown in Figure 3.6.

Test4: We modified the algorithm to first identify disorder structure or experiment related terms, only if these terms are found in the publication, the algorithm would look for protein names and run PONDR-FIT predictor. This modification helped us reduce the number of false positives and the algorithm has an accuracy of nearly 67.89%. The number of true positives and false positives for features 1, 2 and their combinations is shown in Figure 3.7. A comparative analysis of sensitivity, specificity and accuracy is shown in Figure 3.8.

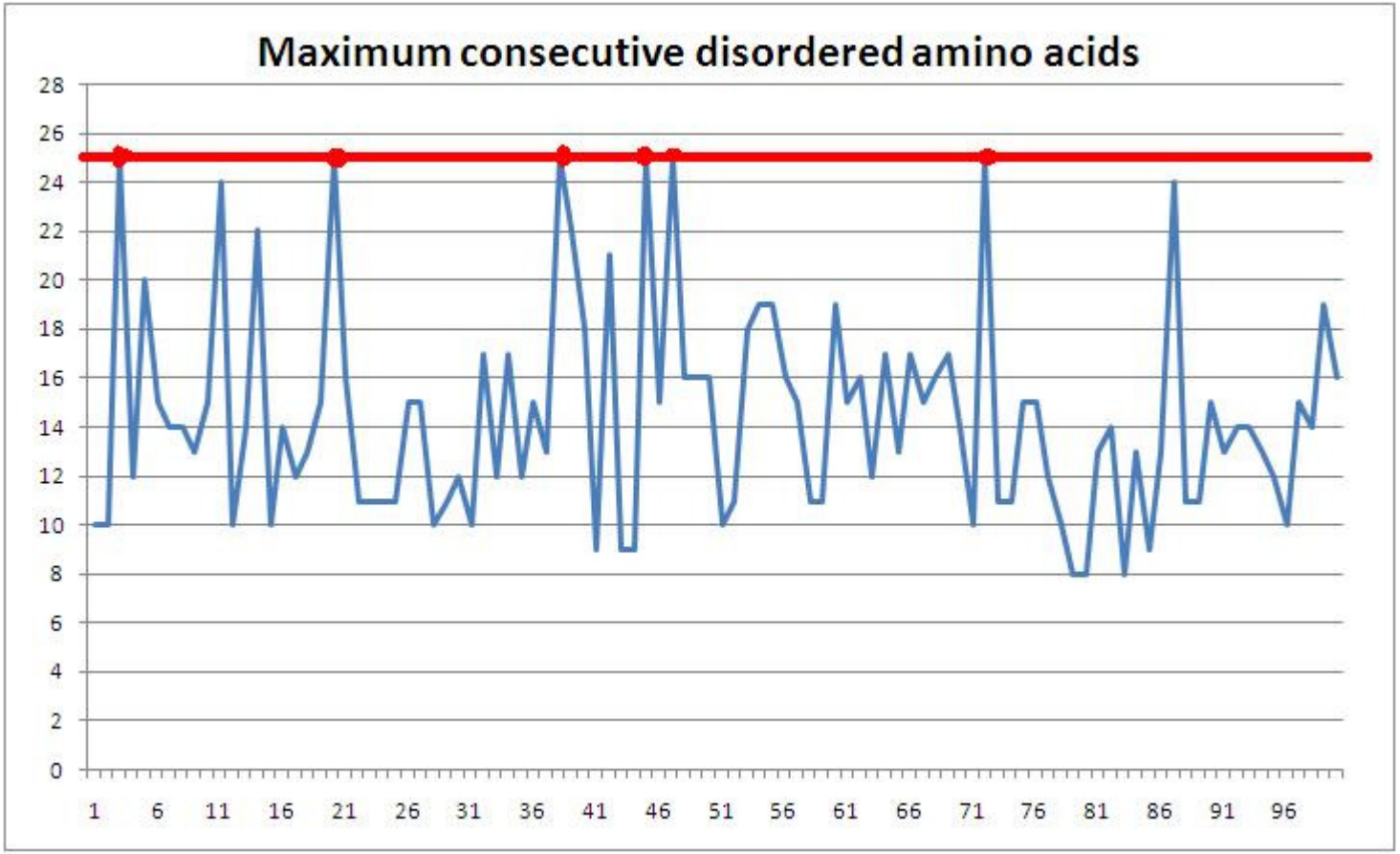

Figure 3.1 A graph showing number of structured proteins having 25 consecutive disordered amino acids.

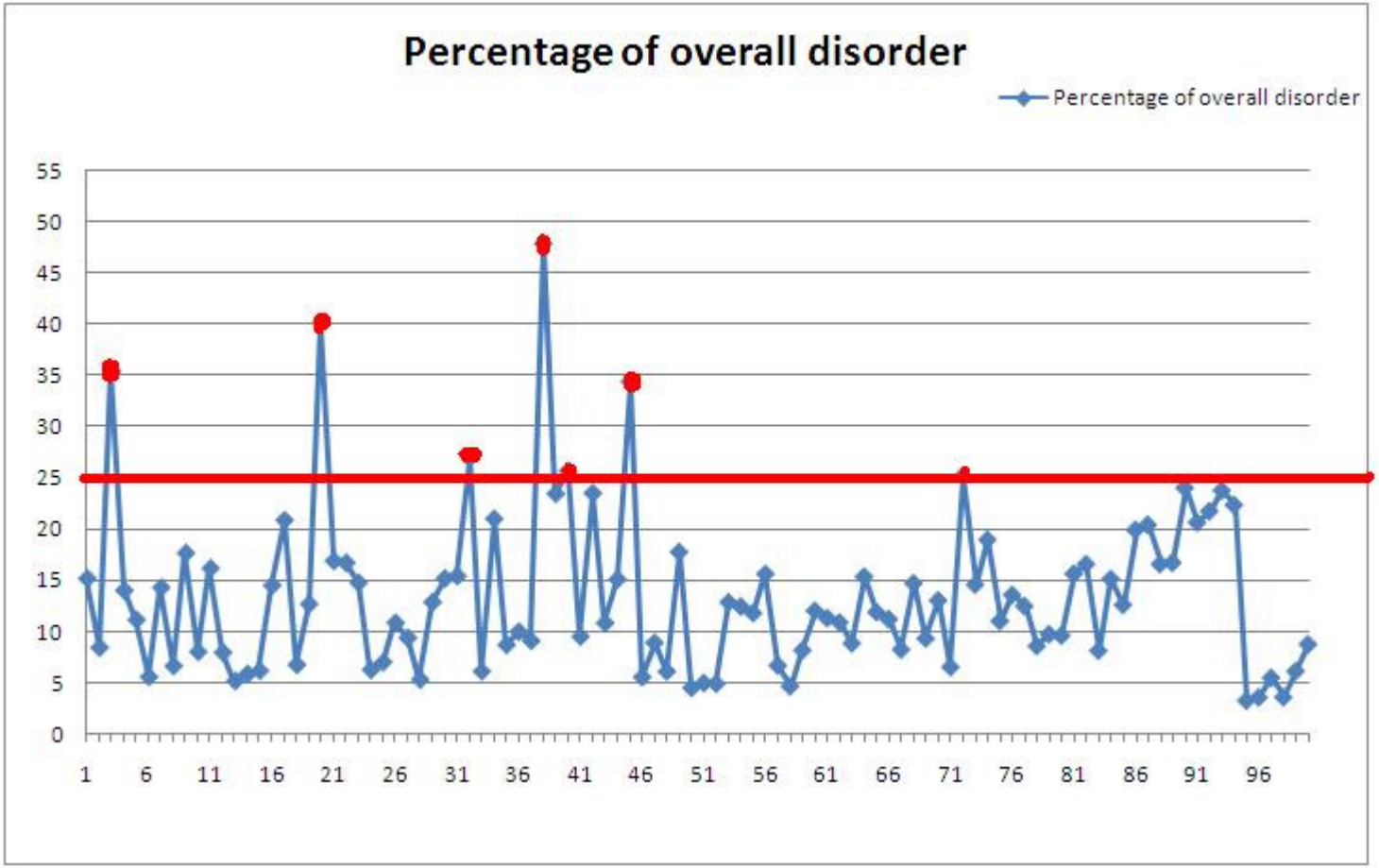

Figure 3.2 A graph showing overall disorder percentage in the 100 structured protein.

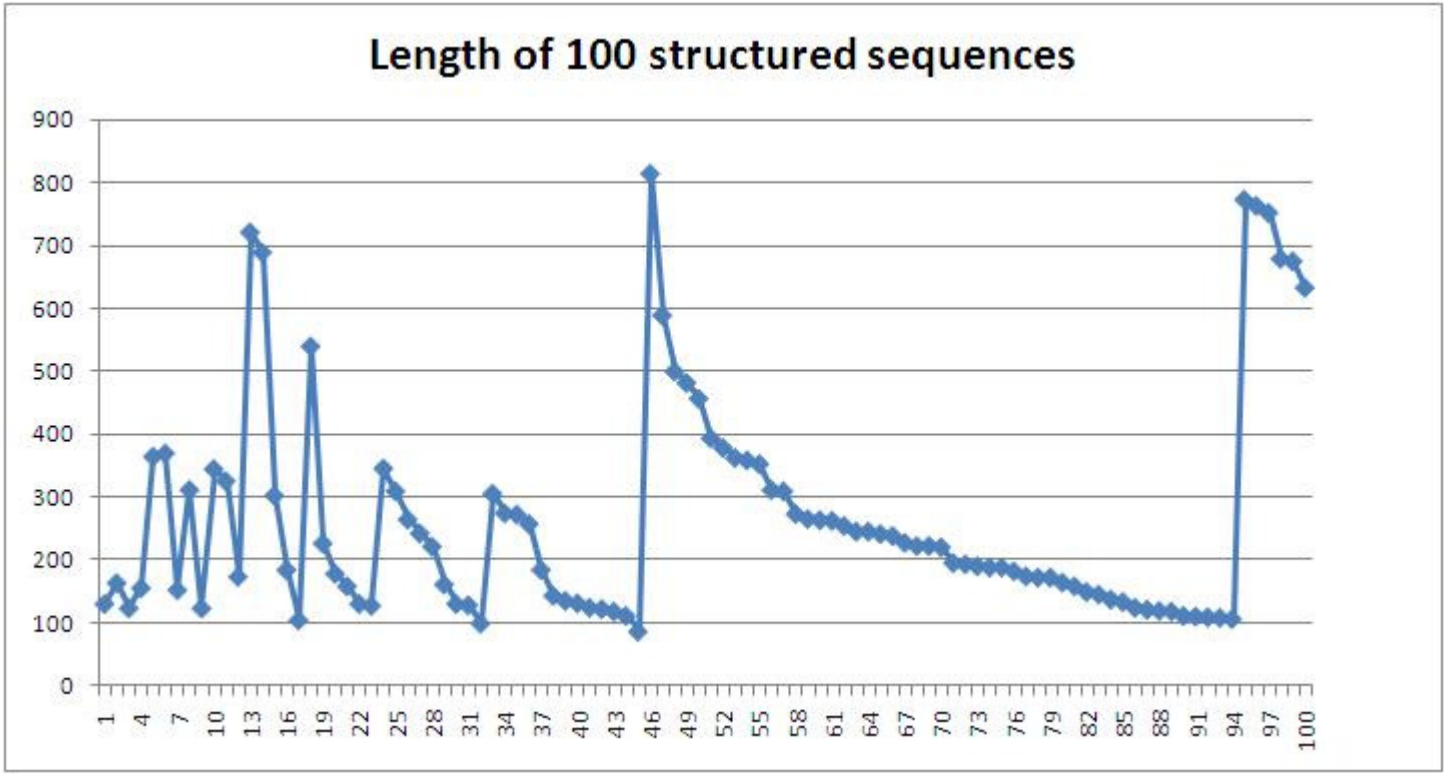

Figure 3.3 A graph showing the total length of the protein.

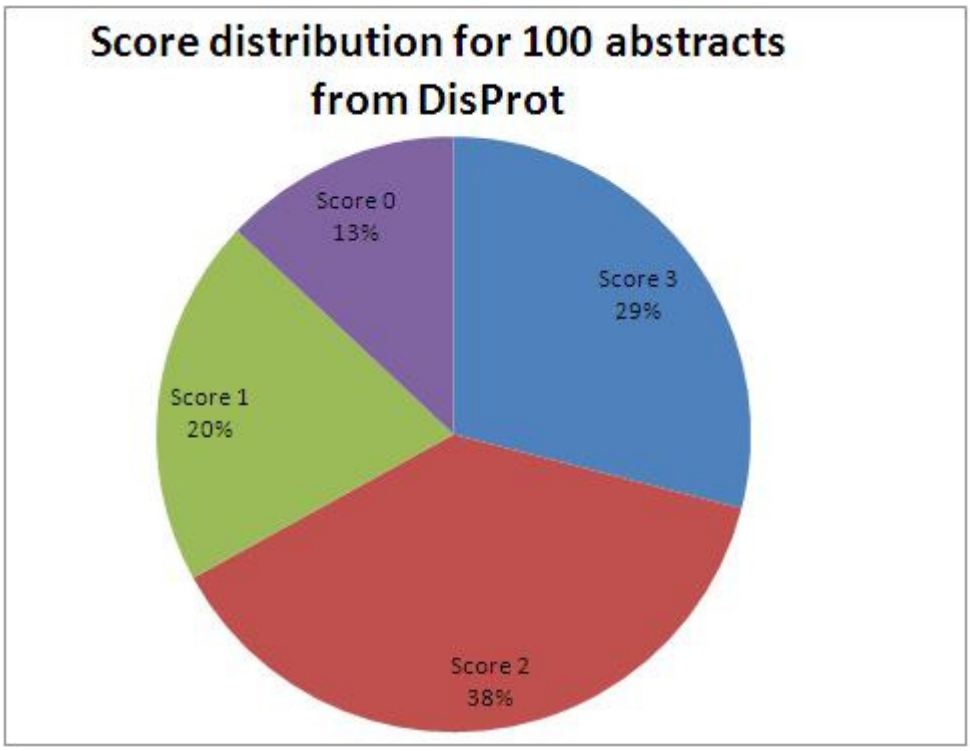

Figure 3.4 A graph showing the score distribution for the test on 100 DisProt abstracts, [Publications with score greater than 0 are considered as relevant].

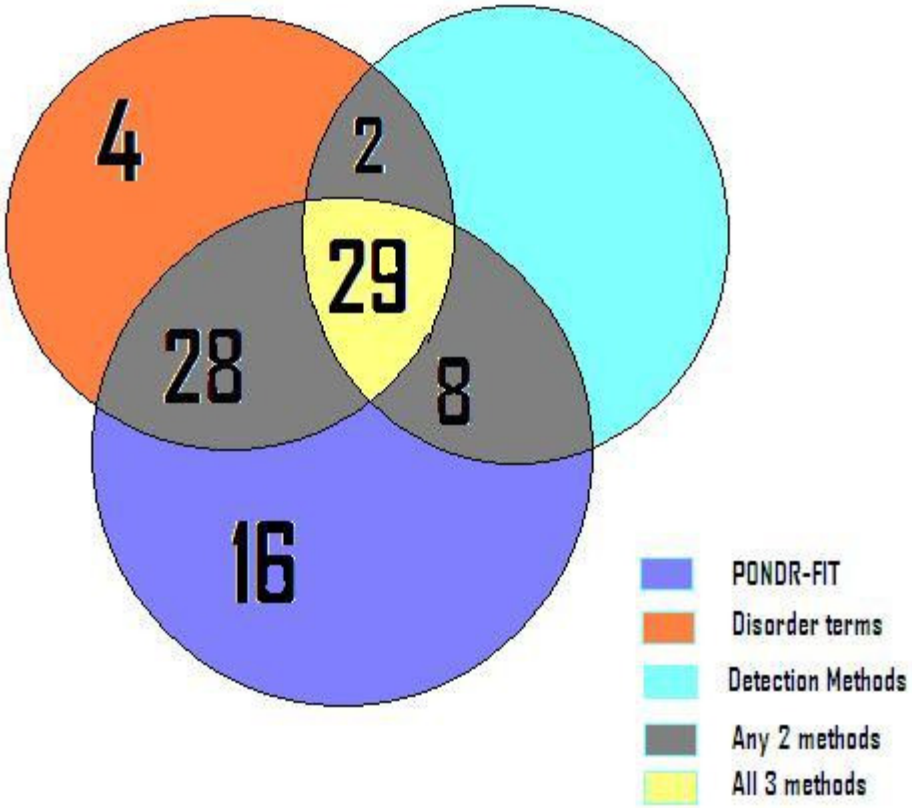

Figure 3.5 A Venn diagram showing the number of publications ranked as relevant.

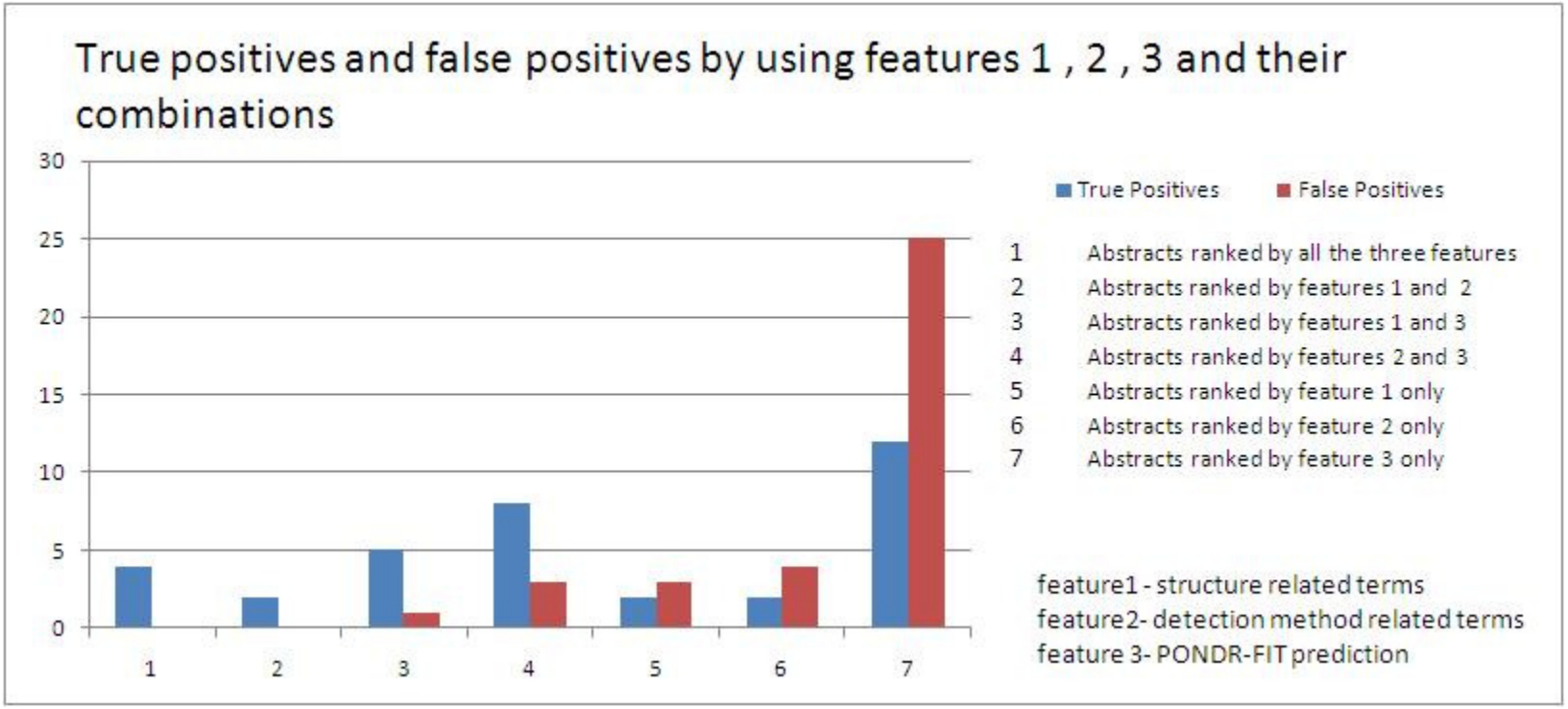

Figure 3.6 A graph showing the number of true positives and false positives in identifying abstracts relevant to DisProt. We used features 1, 2 and 3 in this test.

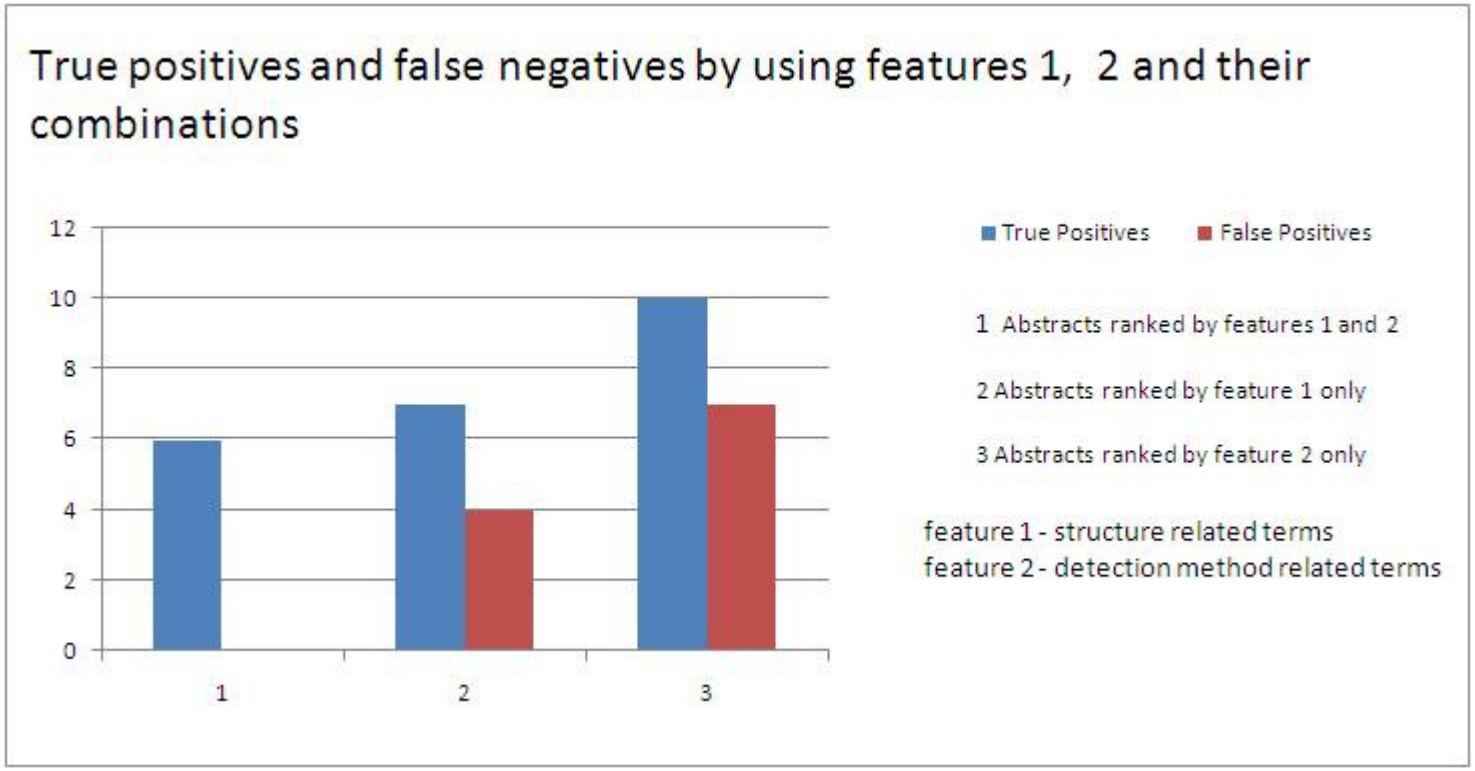

Figure 3.7 A graph showing the number of true positives and false positives in identifying abstracts relevant to DisProt. We used features 1 2 in this test.

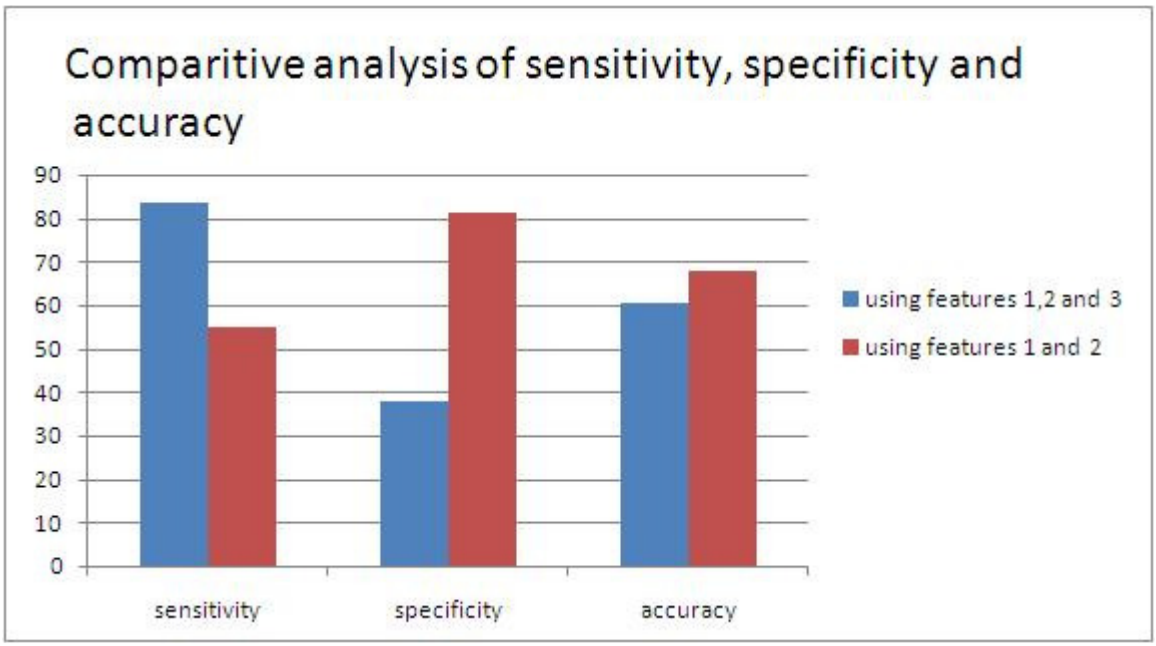

Figure 3.8 A comparative analysis of sensitivity, specificity and accuracy.

CHAPTER 4 DISCUSSION

Automatic identification of publications related to disordered proteins will prove to be useful to the study of disordered proteins and to the process of adding entries to DisProt. In this exploratory study we used three features to identify the relevant publications but in reality we would need more than three features as the process of identifying publications is complicated and requires the algorithm to learn the patterns in the entire publication instead of focusing on a single entity or a term. The results of Test2 and Test3 indicate that tagging a publication just based on the presence of a disordered protein or a detection method can lead to a considerable number of false positives.

Theoretically speaking, in order to reduce the number of false positives, we may consider the results returned by more than one feature. However by manually reading few of the abstracts we found that there may be publications which do not explicitly mention an experimental method or structure related term but are yet relevant to be added to DisProt. Since there is no one way to describe a disordered protein, it may be useful to search for the either the terms in feature1 or their synonyms.

In my opinion, the ideal way to improve the accuracy and identify disorder relation publications is to take a large set of publications having a fair distribution of publications that are disordered and those that are not disordered and to iteratively test this algorithm on smaller sets, analyze the false positive false negatives, make required changes and repeat the test. This process has to go on, till we have a reliable dictionary and we can then use a Rule Based algorithm like decision tree to build rules that determine the decision of whether to classify a given publication as relevant or not.

CHAPTER 5 USING DISPROT

5.1 Work Flow Diagram

A high level work flow diagram for the entire process can be seen in the Figure 5.1.

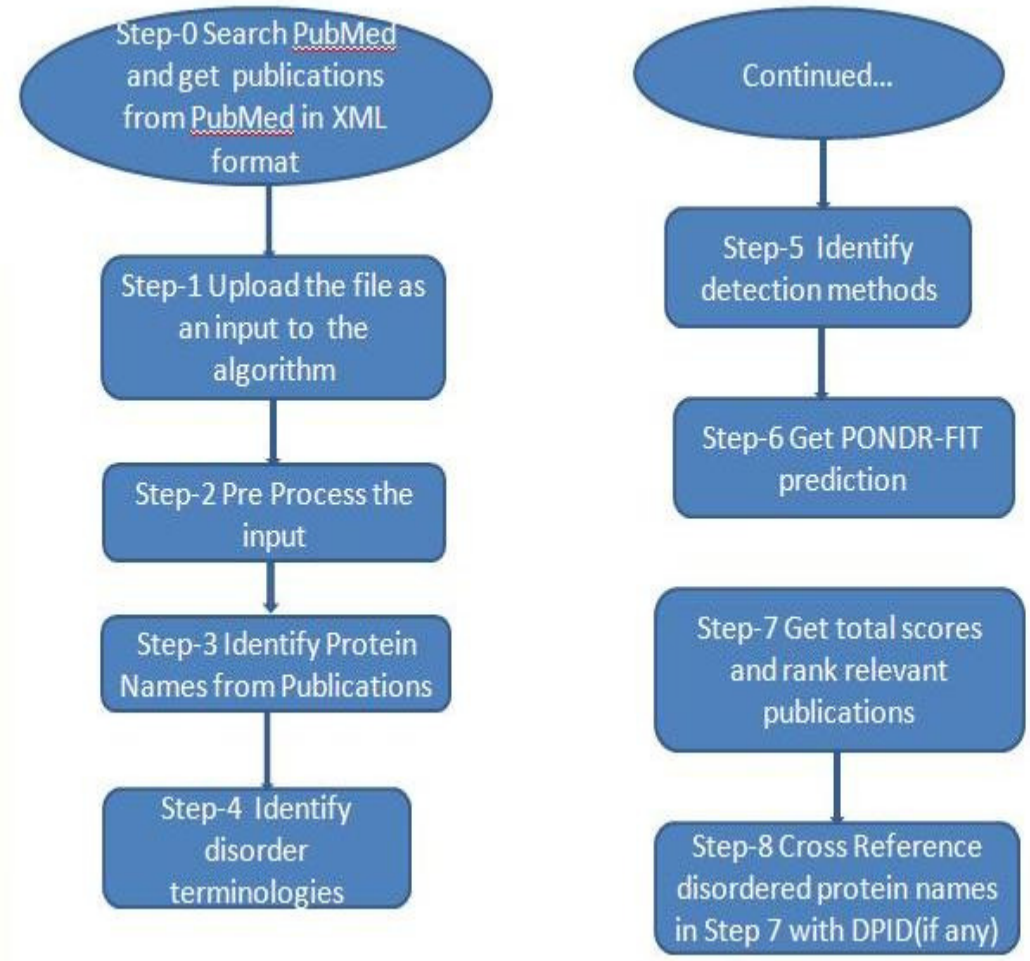

Figure 5.1 Workflow for the algorithm.

5.2 Step by Step Description

Step 0 - Search PubMed

The first step in the process of getting publications is to search PubMed using a set of keywords [See Appendix A.1]. The publications from the search results are downloaded in XML format.

Step 1 - Upload abstracts as an input to the algorithm

Input to the algorithm consists of abstracts from PubMed. Once the XML abstract file from Step 0 is ready it can be uploaded to the server. A screen shot of the file upload page is shown in Figure 3.1.

Step 2 - Pre-process the abstracts

NLProt requires the abstracts to be in the following format: Protein id followed by a greater than symbol followed by the abstract text. A line break separates one abstract from the other. Once the XML abstracts file from PubMed is uploaded to the server [Step 1], we pre process the file to extract the abstracts and PubMed ID and convert them into the format that is required by NLProt to identify protein names. Shown in Figure 3.2 is a screen shot of the pre processed abstracts.

Step 3 - Identify protein names

Once the pre processed abstracts file is available from Step 2, it is used as an input to our algorithm. We run NLProt which identifies the protein names and the corresponding SwissProt ID. A sample output from NLProt is shown in Figure 3.3.

Step 4 - Identify disorder related terminology

While NLProt was reading the abstracts and searching for protein names, our algorithm uses the same abstract that was read into memory by NLProt and searches for disorder related terms [see Appendix A.1]. If any of these terms are found in the abstract, we add one to the score of the publication.

Step 8 - Extract protein name

Out of all the protein names that are identified in the abstract by NLProt, select the proteins that are identified as disordered in Step7 and display those protein names as the relevant protein names.

Step 9 - Cross reference the protein names with DisProt ids

DisProt currently has 643 disordered proteins and 1375 disordered regions. It is possible that the disordered proteins identified by our algorithm in steps 1 to 9, may have been already present in DisProt. So, we search for the protein name and cross reference it with the corresponding DisProt id. If all the protein names found by our algorithm exist in DisProt then the publication is given a score of -1 indicating it as a subset of publications that the annotator may want to refer at a later point of time and see if the publications talk about any new region to an existing protein in DisProt.

Step 10 - Prepare a report indicating just the proteins that are predicted to be disordered by and are new to DisProt

This report can be used by the annotator to go through the abstracts and add the protein to DisProt.

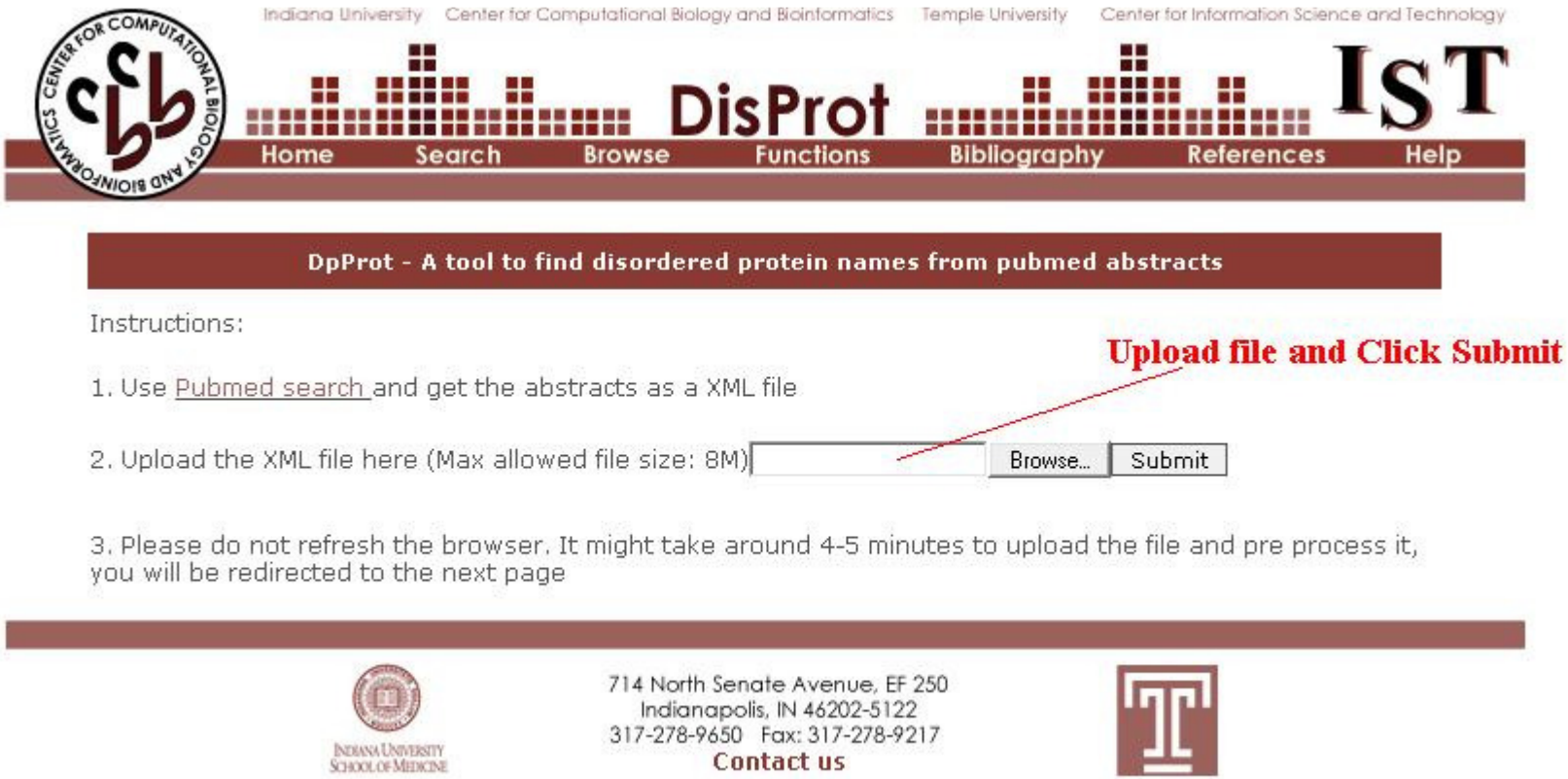

Figure 5.2 A screen shot of abstracts upload mechanism.

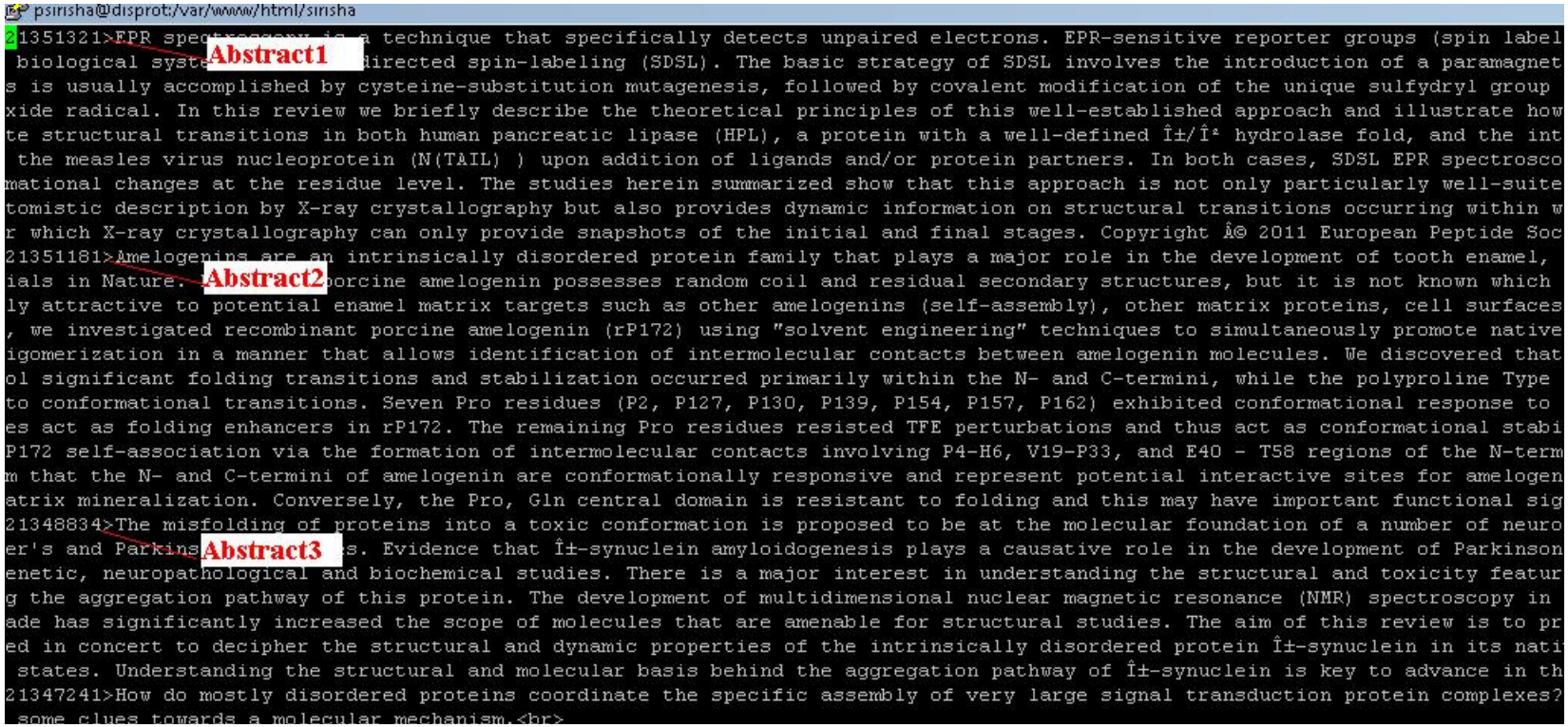

Figure 5.3 A screen shot of pre-processed abstracts.

ID: 631
Intrinsically disordered proteins can form highly dynamic complexes with partner proteins. One such dynamic complex involves the intrinsi
its partner Cdc4 in regulation of yeast cell cycle progression. Phosphorylation of six N-terminal Sic1 sites leads to equilibrium engagem
site with the primary binding pocket in Cdc4, the substrate recognition subunit of a ubiquitin ligase. ENSEMBLE calculations using experi
resonance and small-angle X-ray scattering data reveal significant transient structure in both phosphorylation states of the isolated ens
modulates their electrostatic potential, suggesting a structural basis for the proposed strong contribution of electrostatics to binding.
dynamic pSic1-Cdc4 complex demonstrates the spatial arrangements in the ubiquitin ligase complex. These results provide a physical pictur
predominantly disordered in both its free and bound states, enabling aspects of its structure/function relationship to be elucidated.

The following protein names could be found by NLProt:

Organism

SVM score

UniProt Cross Ref

NAME	ORGANISM	TXT-POS	SCORE	METHOD	DB-ID(S)
Cdc4	yeast	25	0.461	SVM	cc4_yeast (100%)
Cdc4	yeast	54	1.061	SVM	cc4_yeast (100%)
ubiquitin	yeast	62	0.234	SVM	ubiq_yeast (100%)
pSic1	yeast	92	0.881	SVM	SIC1_YEAST (1%)
ubiquitin	yeast	128	0.686	SVM	ubiq_yeast (100%)

Figure 5.4 A screen shot of NLProt output.

ID: 21336827
Dehydrins are a class of stress proteins that belong to the family of Late Embryogenesis Abundant (LEA) proteins in plants, so named because they are highly expressed in late stages of seed formation. In somatic cells, their expression is very low under normal conditions, but increases critically upon dehydration elicited by water stress, high salinity or cold. Dehydrins are thought to be intrinsically disordered proteins, which represents a challenge in understanding their structure-function relationship. Herein we present the backbone (1)H, (15)N and (13)C NMR assignment of the 185 amino acid long ERD14 (Early Response to Dehydration 14), which is a K(3)S-type, typical dehydrin of A. thaliana. Secondary chemical shifts as well as NMR relaxation data show that ERD14 is fully disordered under near native conditions, with short regions of somewhat restricted motion and 5-25% helical propensity. These results suggest that ERD14 may have partially preformed elements for functional interaction with its partner(s) and set the stage for further detailed structural and functional studies of ERD14 both in vitro and in vivo.

Figure 5.5 A screen shot of a abstract in the output.

ID: 21297620
We combined rapid microfluidic mixing with single-molecule fluorescence resonance energy transfer to study the folding kinetics of the intrinsically disordered human protein Î±-synuclein. The time-resolution of 0.2 ms revealed initial collapse of the unfolded protein induced by binding with lipid mimics and subsequent rapid formation of transient structures in the encounter complex. The method also enabled analysis of rapid dissociation and unfolding of weakly bound complexes triggered by massive dilution.

The following disordered protein names could be found by DPProt:

>

NAME	ORGANISM	TXT-POS	SCORE	METHOD	DB-ID(S)	DisProt-ID(S)
Î±-synuclein	human	23	0.710	SVM	SYUG_HUMAN (75%)	DP00630

Figure 5.6 A screen shot of final output.

LIST OF REFERENCES

LIST OF REFERENCES

[1] Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol. 1999; 293:321331

[2] Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z. Intrinsically disordered protein. J Mol Graph Model. 2001; 19:2659

[3] Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, Dunker AK. Coupled folding and binding with alpha-helix-forming molecular recognition elements. Biochemistry. 2005; 44:1245412470

[4] Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. Genome Inform Ser Workshop Genome Inform. 2000; 11:161171

[5] Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK Intrinsic disorder in cell-signaling and cancer-associated proteins. J Molecular Biology. 2002; 323(3):573-84

[6] Vucetic, S., et al. DisProt: a database of protein disorder. Bioinformatics. 2005; 21(1): p. 137-40

[7] Sickmeier, M., Hamilton, J. A., LeGall, T., Vacic, V., Cortese, M. S., Tantos, A., Szabo, B., et al. DisProt: the Database of Disordered Proteins. Nucleic Acids Research. (2007); 35(Database issue), D786-D793. Oxford University Press

[8] Sven Mika, Burkhard Rost. Protein names peeled precisely off free text. Bioinformatics. 2004; 20 (Supplement 1), I241-I247

[9] Sven Mika, Burkhard Rost. NLProt: extracting protein names and sequences from papers. Nucleic Acids Research. 2004; 32 (Supplement 2), W634-W637

[10] Xue B, DunBrack RL, Williams RW, Dunker AK, Uversky VN. PONDR-Fit: A meta-predictor of intrinsically disordered amino acids. Biochim. Biophys 2010;

[11] Narayanaswamy M, Ravikumar KE, Vijay Shanker K. A biological named entity recognizer. Proc Pacific Symp Biocomput. 2003;8:42738

[12] Fukuda K, Tsunode T, Tamura A, Takagi T. Toward information extraction: identifying protein names from biological papers. Pac Symp Biocomput.1998;3:7078

[13] Franzen K, Eriksson G, Olsson F, Asker L, Linden P, Coster J. Protein names and how to find them. Int J Med Inf. 2002;67:4961

[14] Seki K, Mostafa J. A Probabilistic Model for Identifying Protein Names and Their Name Boundaries. Stanford, CA: IEEE Computer Society Bioinformatics Conference. 2003;

[15] Nobata C, Collier N, Tsujii J. Automatic term identification and classification in biology texts. Proc Natural Language Pacific Rim Symposium. 1999;36975

[16] Collier N, Nobata C, Tsujii J. Extracting the Names of Genes and Gene Products with a Hidden Markov Model. Proc Intl Conf Comput Linguistics. 2000;18:2017

[17] Kazama J, Makino T, Ohta Y, Tsujii J. Tuning support vector machines for biomedical named entity recognition. In: Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain. 2002;18

[18] Tanabe L, Wilbur J. Tagging gene and protein names in biomedical text. Bioinformatics. 2002;18:112432

[19] Proux D, Rechenmann F, Julliard L, Pillet VV, Jacq B. Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. Genome Inform. 1998;9:7280

[20] Krauthammer M, Rzhetsky A, Morozov P, Friedman C. Using BLAST for identifying gene and protein names in journal articles. GENE. 2001; 259:24552

[21] B. Settles. ABNER: an open source tool for automatically tagging genes, proteins, and other entity names in text.

[22] Leaman R, Gonzalez G. BANNER: An executable survey of advances in biomedical named entity recognition. Pacific Symposium on Bio computing 2008;13:652-663

[23] Cortese MS, Baird JP, Uversky VN, Dunker AK. Uncovering the unfoldome: enriching cell extracts for unstructured proteins by acid treatment. J Proteome Res. 2005;4:16101618

[24] Csizmok V, Dosztanyi Z, Simon I, Tompa P. Towards proteomic approaches for the identification of structural disorder. Curr Protein Pept Sci. 2007;8:173179

[25] Midic U, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN. Protein disorder in the human diseasome: unfoldomics of human genetic diseases. BM Genomics. 2009;10 Supplement 1:S12

[26] Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK. Intrinsic disorder and functional proteomics. Biophys J. 2007;92:14391456

[27] He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK. Predicting intrinsic disorder in proteins: an overview. Cell Res. 2009;19:929949

[28] Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. Proteins. 2001;42:3848

[29] Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK, Obradovic Z. Optimizing long intrinsic disorder predictors with protein evolutionary information. J Bioinform Comput Biol. 2005;3:3560

[30] Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. Bmc Bioinformatics. 2006;7:208

[31] Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. Structure. 2003;11:14531459

[32] Jones DT, Ward JJ. Prediction of disordered regions in proteins from position specific score matrices. Proteins. 2003;53 Supplement 6:573578

[33] Shimizu K, Hirose S, and Noguchi T. POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. Bioinformatics. 2007;23:23372338

[34] Cheng J, Sweredoski MJ, Baldi P. Accurate prediction of protein disordered regions by mining protein structure data. Data Mining Knowl. Disc. 2005;11:213222

[35] Dosztanyi Z, Csizmok V, Tompa P, Simon I. The pair wise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol. 2005;347:827839

[36] Kulick S, Bies A, Liberman M, Mandel M, McDonald R, Palmer M, Schein A, Ungar L. Integrated Annotation for Biomedical Information Extraction. HLT/NAACL 2004 Workshop: Biolink 2004; pp. 61-68

APPENDIX

APPENDIX

A.1 Keywords used in PubMed Search for finding publications

(Protein OR peptide) AND (“backbone flexibility” [All Fields]) OR (“collapsed coil” [All Fields]) OR (“collapsed coils” [All Fields]) OR (“collapsed disorder” [All Fields]) OR (“conformational change” [All Fields]) OR (“conformational changes” [All Fields]) OR (“conformational disorder” [All Fields]) OR (“conformational extension” [All Fields]) OR (“conformational extensions” [All Fields]) OR (“conformational flexibility” [All Fields]) OR (“conformational fluctuation” [All Fields]) OR (“conformational fluctuations” [All Fields]) OR (“conformational isomer” [All Fields]) OR (“conformational isomers” [All Fields]) OR (“conformational mobility” [All Fields]) OR (“conformational variability” [All Fields]) OR (“conformationally disordered” [All Fields]) OR (“conformationally dynamic” [All Fields]) OR (“conformationally extended” [All Fields]) OR (“conformationally flexible” [All Fields]) OR (“conformationally mobile” [All Fields]) OR (“conformationally random” [All Fields]) OR (“conformationally unfolded” [All Fields]) OR (“conformationally unstructured” [All Fields]) OR (“disorder to order transition” [All Fields]) OR (“disorder to order transitions” [All Fields]) OR (“disorder-order transition” [All Fields]) OR (“disorder-order transitions” [All Fields]) OR (“disordered C-terminal” [All Fields]) OR (“disordered C-terminus” [All Fields]) OR (“disordered coil” [All Fields]) OR (“disordered coils” [All Fields]) OR (“disordered conformation” [All Fields]) OR (“disordered conformations” [All Fields]) OR (“disordered domain” [All Fields]) OR (“disordered domains” [All Fields]) OR (“disordered extension” [All Fields]) OR (“disordered extensions” [All Fields]) OR (“disordered linker” [All Fields]) OR (“disordered linkers” [All Fields]) OR (“disordered loop” [All Fields]) OR (“disordered loops”

[All Fields]) OR ("disordered N-terminal" [All Fields]) OR ("disordered N-terminus" [All Fields]) OR ("disordered peptide" [All Fields]) OR ("disordered peptides" [All Fields]) OR ("disordered polypeptide" [All Fields]) OR ("disordered polypeptides" [All Fields]) OR ("disordered protein" [All Fields]) OR ("disordered proteins" [All Fields]) OR ("disordered region" [All Fields]) OR ("disordered regions" [All Fields]) OR ("disordered segment" [All Fields]) OR ("disordered segments" [All Fields]) OR ("disordered state" [All Fields]) OR ("disordered states" [All Fields]) OR ("disordered structure" [All Fields]) OR ("disordered structures" [All Fields]) OR ("disordered tether" [All Fields]) OR ("disordered tethers" [All Fields]) OR ("dynamic C-terminal" [All Fields]) OR ("dynamic C-terminus" [All Fields]) OR ("dynamic coil" [All Fields]) OR ("dynamic coils" [All Fields]) OR ("dynamic conformation" [All Fields]) OR ("dynamic conformations" [All Fields]) OR ("dynamic domain" [All Fields]) OR ("dynamic domains" [All Fields]) OR ("dynamic extension" [All Fields]) OR ("dynamic extensions" [All Fields]) OR ("dynamic linker" [All Fields]) OR ("dynamic linkers" [All Fields]) OR ("dynamic loop" [All Fields]) OR ("dynamic loops" [All Fields]) OR ("dynamic N-terminal" [All Fields]) OR ("dynamic N-terminus" [All Fields]) OR ("dynamic peptide" [All Fields]) OR ("dynamic peptides" [All Fields]) OR ("dynamic polypeptide" [All Fields]) OR ("dynamic polypeptides" [All Fields]) OR ("dynamic protein" [All Fields]) OR ("dynamic proteins" [All Fields]) OR ("dynamic region" [All Fields]) OR ("dynamic regions" [All Fields]) OR ("dynamic segment" [All Fields]) OR ("dynamic segments" [All Fields]) OR ("dynamic state" [All Fields]) OR ("dynamic states" [All Fields]) OR ("dynamic structure" [All Fields]) OR ("dynamic structures" [All Fields]) OR ("dynamic tether" [All Fields]) OR ("dynamic tethers" [All Fields]) OR ("extended C-terminal" [All Fields]) OR ("extended C-terminus" [All Fields]) OR ("extended coil" [All Fields]) OR ("extended coils" [All Fields]) OR ("extended conformation" [All Fields]) OR ("extended conformations" [All Fields]) OR ("extended domain" [All Fields]) OR ("extended domains" [All Fields]) OR ("extended linker" [All Fields]) OR ("extended linkers" [All Fields]) OR ("extended loop" [All Fields]) OR ("extended loops" [All Fields]) OR ("extended N-terminal"

[All Fields]) OR (“extended N-terminus” [All Fields]) OR (“extended peptide” [All Fields]) OR (“extended peptides” [All Fields]) OR (“extended polypeptide” [All Fields]) OR (“extended polypeptides” [All Fields]) OR (“extended protein” [All Fields]) OR (“extended proteins” [All Fields]) OR (“extended region” [All Fields]) OR (“extended regions” [All Fields]) OR (“extended segment” [All Fields]) OR (“extended segments” [All Fields]) OR (“extended state” [All Fields]) OR (“extended states” [All Fields]) OR (“extended structure” [All Fields]) OR (“extended structures” [All Fields]) OR (“extended tether” [All Fields]) OR (“extended tethers” [All Fields]) OR (“flexible C-terminal” [All Fields]) OR (“flexible C-terminus” [All Fields]) OR (“flexible coil” [All Fields]) OR (“flexible coils” [All Fields]) OR (“flexible conformation” [All Fields]) OR (“flexible conformations” [All Fields]) OR (“flexible domain” [All Fields]) OR (“flexible domains” [All Fields]) OR (“flexible extension” [All Fields]) OR (“flexible extensions” [All Fields]) OR (“flexible linker” [All Fields]) OR (“flexible linkers” [All Fields]) OR (“flexible loop” [All Fields]) OR (“flexible loops” [All Fields]) OR (“flexible N-terminal” [All Fields]) OR (“flexible N-terminus” [All Fields]) OR (“flexible peptide” [All Fields]) OR (“flexible peptides” [All Fields]) OR (“flexible polypeptide” [All Fields]) OR (“flexible polypeptides” [All Fields]) OR (“flexible protein” [All Fields]) OR (“flexible proteins” [All Fields]) OR (“flexible region” [All Fields]) OR (“flexible regions” [All Fields]) OR (“flexible segment” [All Fields]) OR (“flexible segments” [All Fields]) OR (“flexible state” [All Fields]) OR (“flexible states” [All Fields]) OR (“flexible structure” [All Fields]) OR (“flexible structures” [All Fields]) OR (“flexible tether” [All Fields]) OR (“flexible tethers” [All Fields]) OR (“intrinsic disorder” [All Fields]) OR (“intrinsic extension” [All Fields]) OR (“intrinsic extensions” [All Fields]) OR (“intrinsic flexibility” [All Fields]) OR (“intrinsic mobility” [All Fields]) OR (“intrinsically disordered” [All Fields]) OR (“intrinsically dynamic” [All Fields]) OR (“intrinsically extended” [All Fields]) OR (“intrinsically flexible” [All Fields]) OR (“intrinsically mobile” [All Fields]) OR (“intrinsically random” [All Fields]) OR (“intrinsically unfolded” [All Fields]) OR (“Intrinsically unstructured” [All Fields]) OR (“mobile C-terminal” [All Fields])

OR (“mobile C-terminus” [All Fields]) OR (“mobile coil” [All Fields]) OR (“mobile coils” [All Fields]) OR (“mobile conformation” [All Fields]) OR (“mobile conformations” [All Fields]) OR (“mobile domain” [All Fields]) OR (“mobile domains” [All Fields]) OR (“mobile extension” [All Fields]) OR (“mobile extensions” [All Fields]) OR (“mobile linker” [All Fields]) OR (“mobile linkers” [All Fields]) OR (“mobile loop” [All Fields]) OR (“mobile loops” [All Fields]) OR (“mobile N-terminal” [All Fields]) OR (“mobile N-terminus” [All Fields]) OR (“mobile peptide” [All Fields]) OR (“mobile peptides” [All Fields]) OR (“mobile polypeptide” [All Fields]) OR (“mobile polypeptides” [All Fields]) OR (“mobile protein” [All Fields]) OR (“mobile proteins” [All Fields]) OR (“mobile region” [All Fields]) OR (“mobile regions” [All Fields]) OR (“mobile segment” [All Fields]) OR (“mobile segments” [All Fields]) OR (“mobile state” [All Fields]) OR (“mobile states” [All Fields]) OR (“mobile structure” [All Fields]) OR (“mobile structures” [All Fields]) OR (“mobile tether” [All Fields]) OR (“mobile tethers” [All Fields]) OR (“molten globule” [All Fields]) OR (“molten globules” [All Fields]) OR (“native disorder” [All Fields]) OR (“native extension” [All Fields]) OR (“native flexibility” [All Fields]) OR (“native mobility” [All Fields]) OR (“natively disordered” [All Fields]) OR (“natively dynamic” [All Fields]) OR (“natively extended” [All Fields]) OR (“natively flexible” [All Fields]) OR (“natively mobile” [All Fields]) OR (“natively random” [All Fields]) OR (“natively unfolded” [All Fields]) OR (“natively unstructured” [All Fields]) OR (“partially folded” [All Fields]) OR (“Partially unfolded” [All Fields]) OR (“partly folded” [All Fields]) OR (“partly unfolded” [All Fields]) OR (“random C-terminal” [All Fields]) OR (“random C-terminus” [All Fields]) OR (“random coil” [All Fields]) OR (“random coils” [All Fields]) OR (“random conformation” [All Fields]) OR (“random conformations” [All Fields]) OR (“random domain” [All Fields]) OR (“random domains” [All Fields]) OR (“random extension” [All Fields]) OR (“random extensions” [All Fields]) OR (“random linker” [All Fields]) OR (“random linkers” [All Fields]) OR (“random loop” [All Fields]) OR (“random loops” [All Fields]) OR (“random N-terminal” [All Fields]) OR (“random N-terminus” [All Fields]) OR (“random

peptide" [All Fields]) OR ("random peptides" [All Fields]) OR ("random polypeptide" [All Fields]) OR ("random polypeptides" [All Fields]) OR ("random protein" [All Fields]) OR ("random region" [All Fields]) OR ("random regions" [All Fields]) OR ("random segment" [All Fields]) OR ("random segments" [All Fields]) OR ("random state" [All Fields]) OR ("random states" [All Fields]) OR ("random structure" [All Fields]) OR ("random structures" [All Fields]) OR ("random tether" [All Fields]) OR ("random tethers" [All Fields]) OR ("structural disorder" [All Fields]) OR ("structural extension" [All Fields]) OR ("structural extensions" [All Fields]) OR ("structural flexibility" [All Fields]) OR ("structural mobility" [All Fields]) OR ("structurally disordered" [All Fields]) OR ("structurally dynamic" [All Fields]) OR ("structurally extended" [All Fields]) OR ("structurally flexible" [All Fields]) OR ("structurally mobile" [All Fields]) OR ("structurally random" [All Fields]) OR ("structurally unfolded" [All Fields]) OR ("tether" [All Fields]) OR ("tethers" [All Fields]) OR ("unfolded C-terminal" [All Fields]) OR ("unfolded C-terminus" [All Fields]) OR ("unfolded coil" [All Fields]) OR ("unfolded coils" [All Fields]) OR ("unfolded conformation" [All Fields]) OR ("unfolded conformations" [All Fields]) OR ("unfolded domain" [All Fields]) OR ("unfolded domains" [All Fields]) OR ("unfolded extension" [All Fields]) OR ("unfolded extensions" [All Fields]) OR ("unfolded linker" [All Fields]) OR ("unfolded linkers" [All Fields]) OR ("unfolded loop" [All Fields]) OR ("unfolded loops" [All Fields]) OR ("unfolded N-terminal" [All Fields]) OR ("unfolded N-terminus" [All Fields]) OR ("unfolded peptide" [All Fields]) OR ("unfolded peptides" [All Fields]) OR ("unfolded polypeptide" [All Fields]) OR ("unfolded polypeptides" [All Fields]) OR ("unfolded region" [All Fields]) OR ("unfolded regions" [All Fields]) OR ("unfolded segment" [All Fields]) OR ("unfolded segments" [All Fields]) OR ("unfolded state" [All Fields]) OR ("unfolded states" [All Fields]) OR ("unfolded structure" [All Fields]) OR ("unfolded structures" [All Fields]) OR ("unfolded tether" [All Fields]) OR ("unfolded tethers" [All Fields]) OR ("unstructured C-terminal" [All Fields]) OR ("unstructured C-terminus" [All Fields]) OR ("unstructured coil" [All Fields]) OR

(“unstructured coils” [All Fields]) OR (“unstructured conformation” [All Fields]) OR (“unstructured conformations” [All Fields]) OR (“unstructured domain” [All Fields]) OR (“unstructured domains” [All Fields]) OR (“unstructured extension” [All Fields]) OR (“unstructured extensions” [All Fields]) OR (“unstructured linker” [All Fields]) OR (“unstructured linkers” [All Fields]) OR (“unstructured loop” [All Fields]) OR (“unstructured loops” [All Fields]) OR (“unstructured N-terminal” [All Fields]) OR (“unstructured N-terminus” [All Fields]) OR (“unstructured peptide” [All Fields]) OR (“unstructured peptides” [All Fields]) OR (“unstructured polypeptide” [All Fields]) OR (“unstructured polypeptides” [All Fields]) OR (“unstructured protein” [All Fields]) OR (“unstructured proteins” [All Fields]) OR (“unstructured region” [All Fields]) OR (“unstructured regions” [All Fields]) OR (“unstructured segment” [All Fields]) OR (“unstructured segments” [All Fields]) OR (“unstructured state” [All Fields]) OR (“unstructured states” [All Fields]) OR (“unstructured tether” [All Fields]) OR (“unstructured tethers” [All Fields])

Table A.1

Set of structure / function related terms used as feature 1

crystal structure	x-ray characterization	structural characterization
nmr structure	solution structure	resonance assignment
transition from * to *	transition to * from *	experimental data
conformational study	Conformational studies	experimentally determined

Table A.2

Set of detection methods used as feature 2

analytical ultracentrifugation	circular dichroism	electron microscopy	spectroscopy
nmr	x-ray crystallography	spectrometry	light scattering
nuclear magnetic resonance	proteolysis	proteomic	spectroscopic
crystallographic	viscometric	dichromatic	electrophoresis
microcalorimetry	hydrodynamic	neutron scattering	fluorescence
uv	fluorescent	ESR	microspectroscopy