RESEARCH ARTICLE

A three-way approach for protein function classification

Hafeez Ur Rehman[1☯]*, **Nouman Azam**[1☯], **JingTao Yao**[2☯], **Alfredo Benso**[3☯]

1 Department of Computer Science, National University of Computer and Emerging Sciences, Peshawar Pakistan, **2** Department of Computer Science, University of Regina, Regina, SK S4S 0A2, Canada, **3** Department of Computer & Control Engineering, Politecnico di Torino, I-10129, Torino, Italy

☯ These authors contributed equally to this work.
* hafeez.urrehman@nu.edu.pk

Abstract

The knowledge of protein functions plays an essential role in understanding biological cells and has a significant impact on human life in areas such as personalized medicine, better crops and improved therapeutic interventions. Due to expense and inherent difficulty of biological experiments, intelligent methods are generally relied upon for automatic assignment of functions to proteins. The technological advancements in the field of biology are improving our understanding of biological processes and are regularly resulting in new features and characteristics that better describe the role of proteins. It is inevitable to neglect and overlook these anticipated features in designing more effective classification techniques. A key issue in this context, that is not being sufficiently addressed, is how to build effective classification models and approaches for protein function prediction by incorporating and taking advantage from the ever evolving biological information. In this article, we propose a three-way decision making approach which provides provisions for seeking and incorporating future information. We considered probabilistic rough sets based models such as Game-Theoretic Rough Sets (GTRS) and Information-Theoretic Rough Sets (ITRS) for inducing three-way decisions. An architecture of protein functions classification with probabilistic rough sets based three-way decisions is proposed and explained. Experiments are carried out on *Saccharomyces cerevisiae* species dataset obtained from *Uniprot* database with the corresponding functional classes extracted from the Gene Ontology (GO) database. The results indicate that as the level of biological information increases, the number of deferred cases are reduced while maintaining similar level of accuracy.

OPEN ACCESS

Citation: Ur Rehman H, Azam N, Yao J, Benso A (2017) A three-way approach for protein function classification. PLoS ONE 12(2): e0171702. doi:10.1371/journal.pone.0171702

Editor: Yong Deng, Southwest University, CHINA

Received: August 1, 2016

Accepted: January 6, 2017

Published: February 24, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The third author was partially supported by Discovery Grant from NSERC Canada. We are also thankful to the Higher Education Commission of Pakistan for providing student grant under the SRGP (Startup Research Grant Program) initiative. There was no additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

All living organisms are composed of cells, which are intricately arranged chemical factories that obtain matter from their environment and use this raw matter to generate copies of themselves [1]. Behind this miraculous functioning of the cells are the most important biochemical molecules called proteins. Due to their role in almost every biological activity, it is crucial to have a clear understanding of their respective functions. Moreover, the knowledge of protein functions is also essential for understanding how biological activities are performed at the molecular level. This is useful in developing personalized medicine, more effective therapeutic interventions as well as understanding biological entities as engineered systems [2–7]. On the other hand, when the number of sequenced genomes are growing, the overwhelming majority of new proteins with unknown functions continue to emerge at an exponential rate. Under these conditions, it is not feasible to manually identify and assign functions to proteins. Intelligent mechanisms are generally relied upon to automatically predict and assign functions to proteins [8–10].

Several methods have been proposed for characterization of protein functions. The early and conventional techniques were generally based on the most fundamental type of information about proteins i.e., their amino acid sequence, utilizing tools such as Basic Local Alignment Search Tool (BLAST) [11]. Sequence of a protein determines its different characteristics such as its sub-cellular localization, possible structural conformations as well as its functions [3]. Some of the prominent approaches in this category can be found in [12–14]. With the availability of data from massive high-throughput experiments, features based on different data such as genomic contextual data and Protein-Protein Interactions (PPIs) data, has also emerged. Recent and advanced computational methods utilized these and similar information in designing approaches for prediction task. For example, features based on genomic contextual data were utilized by [12, 15, 16], features based on protein-protein interaction data were used by [17–20], and features exploiting function structure relationship were reported in [21–23]. As we have access to more interesting information, we may expect more effective models and approaches for precise prediction of protein functions.

Due to technological advancements, our understanding of biological processes is improving and new features describing proteins are emerging on regular basis [3]. It is inevitable to ignore these anticipated features in designing more effective and efficient prediction techniques. An important issue that needs to be addressed in this context is how to develop effective models by incorporating and taking advantage from the ever evolving biological information that leads to new features and characteristics of proteins. This however has generally been overlooked and received little or no attention in the existing literature. A general assumption, although not explicitly stated, is that the information is being fixed (i.e., not dynamic and evolving) while developing classification approaches. This assumption may not be always useful, for instance, consider the classification of proteins whose functions may not be precisely identified due to lack of associated biological information (although we may anticipate it in future) thereby leading to compromised results. To address this issue, i.e., incorporating the anticipated future information into the predictive task, we propose a three-way decision making approach that includes a decision option of deferment. This option is exercised whenever we have inconclusive and insufficient evidence to reach confirmed or certain decisions. The deferred decision option provides provisions for incorporating future information which may be used in deciding the deferred cases. In particular, three types of decisions are used, i.e., accept, reject and deferment in order to classify functions of proteins.

There are different models for inducing three-way decisions. In this article, we investigate and examine probabilistic rough sets based three-way decision making approaches for protein functions classification [24]. The probabilistic rough sets can be used to induce three regions corresponding to a concept (represented in terms of a set), namely, positive, negative and boundary regions. The three regions lead to three-way decisions in the form of acceptance, rejection and deferment, respectively. The three regions and their respective decisions are defined and controlled by a pair of thresholds. There are different forms and models of probabilistic rough sets based on how these thresholds are obtained and interpreted. We consider two such models, i.e., Game-Theoretic Rough Sets (GTRS) [25–27] and Information-Theoretic

Rough Sets (ITRS) [28]. Moreover, we examine and define five three-way approaches based on the GTRS and ITRS by employing different measures and iterative methods. To incorporate and take benefit from these three-way approaches in real applications, we propose an architecture of protein functions classification. Lastly, we evaluated the three-way approaches on the dataset of *Saccharomyces cerevisiae* species proteins which is obtained from *Uniprot* database [29], with the corresponding functional classes extracted from the well known Gene Ontology (GO) database [30]. The experimental results indicate that by increasing the level of biological information associated with proteins, the number of deferred cases can be reduced while maintaining the same level of accuracy. We comprehensively benchmark our approaches under these settings and conclude that the classification becomes more crisp as the knowledge of associated biological information matures.

The code (Python/Bash/Matlab) and data files used in this work are available as a zip file ("Protein_Functions_TWD_data_code.zip") from http://tinyurl.com/jdpwkkq.

Background

Protein function classification

An important factor that impacts the performance of function prediction models is the type of biological information used to infer functional association among proteins. Until recently, many high throughput techniques have been developed to devise mechanisms leading to precise prediction of protein functions. These techniques utilize information derived from sequence similarity, protein 3D structure, phylogenetic profiles, protein complexes, PPIs, gene expression profiles [31–33]. The most prominent techniques utilize proteome-scale PPI networks that have been retrieved for several organisms including yeast and human. Protein-protein networks are graphs where each node represents a protein and edges between nodes represent an interaction. An interaction in the network is either a direct physical association between the proteins (typically retrieved via two hybrid analysis [34] or on the other hand if two interacting proteins are part of the same multi-protein complex, they are also considered as interacting proteins [35]. Thus from informatics point of view an interaction is not necessarily a direct physical association of proteins but sometimes it is mutual presence in the same protein complex depending on the experiment which reveals the interaction.

The most recent as well as renowned approaches in the field of protein function prediction use protein-protein interactions data in different ways [31–33]. A wide majority of these techniques are based on the fact that interacting proteins are likely to share common functions as they interact for an associated biological activity. Methods in this category assign annotations to protein under question, based on the functions of their neighboring proteins. The methods vary in the extent to which they employ global features of the interactome in annotating proteins, or the way they exploit the topological features of the interactome [17, 18]. In addition to that, the methods are based on quite varied underlying formulations and use well understood concepts from the fields of graph theory, graphical models, stochastic processes, probabilistic graphs or clustering [18, 19].

Another class of approaches are based on utilizing the GO structure into computational models by incorporating the semantic similarity offered by the Direct Acyclic Graph (DAG) architecture of gene ontology. The integration of multi-level gene ontology terms exploiting their relationships for protein function prediction was investigated in [2, 8, 9, 36]. These methods calculate different similarity measures by operating on GO term dependencies to define functional associations among proteins. A similar technique based on the Markov Random Field (MRF) properties of protein-protein networks, integrated the inter-species protein homolog information to construct MRF based graphs using the gene ontology terms was outlined in [8]. The authors report high precision when tested for a limited set of functional terms [8].

Another type of biological information that is frequently used for uncharacterized proteins is the number of motifs conserved in those proteins [9, 36]. Several functionally conserved proteins are found to have motifs that associate them to a particular molecular activity. For example, hypothetical protein YIL169C is conserved with Chemotaxis_Transduce_2 and T_SNARE motifs, and similar motifs in known proteins can be used to link functional information with the protein under investigation. Integrating heterogeneous information conserved across proteins of unknown function, with state of the art classification scheme may help to increase protein function prediction accuracy.

The existing computational approaches have significantly contributed in understanding and characterizing protein functions by investigating and utilizing different types of features. However, there is still a need for approaches to incorporate and integrate the ever evolving features of proteins for precise prediction of their functions. These new features, once known and available, will give better insight into biological activities thereby are expected to provide more precise characterization of proteins. In the later sections of this article, we present a three-way approach to address these issues.

Three-way decisions

In many real life decision making scenarios involving vague and uncertain information, the three-way decision making strategy including a delay, deferment or non-commitment decision option is a better and more useful approach [37–39]. To explain this, consider the following examples: 1) How do we make a purchase decision based on information gathered from blogs, reviews, friend suggestions and experiences? 2) How do doctors make diagnosis decisions based on the presence of some symptoms and tests? 3) How do military commanders decide to carry out military actions based on intelligence information? In all these and similar decision scenarios, the decision makers are faced with two types of situations. Either they have sufficient and convincing information necessary to make a decision or they are faced with vague and incomplete information which is insufficient to make a useful decision. In the former case, the decision makers can exercise immediate and certain decisions in the form of acceptance/rejection, yes/no or true/false. In the latter case, the decision makers may not be able to make certain decisions. For instance, the diagnosis tests are inconclusive or the intelligence information is vague or incomplete. A better and more useful choice in such uncertain and doubtful situations is to delay the decisions, assuming that future information will evolve which will make the decision making more obvious and evident. Three-way decisions is essentially the same approach to decision making. We make immediate accept or reject decisions if we have convincing and sufficient evidence based on the available information. On the other hand, we make a deferment decision whenever we lack sufficient evidence.

In fact, three-way decisions has been practiced over the years across different domains, including, medical decision making, psychology, social judgment theory, management sciences and machine learning [40–44]. These application domains suggest that three-way decisions enjoy a good history from usage and application perspective, however, it is surprising to note that from theoretical perspective, it lacks a unified formal description over the years [45]. This theoretical gap was first recognized in the rough sets community. In particular, Yao introduced a general theory of three-way decisions, motivated by the rough sets based three regions [45]. The essential notion in the theory adopted from rough sets is the division of the universe into three pair-wise disjoint regions. The theory however is not restricted to rough sets and goes beyond it by considering rough set theory as one of many possible ways to construct and induce the three regions [39, 46]. Three-way decisions may be formulated based on the theories such as rough sets, interval sets, shadowed sets, approximations of fuzzy sets, a threshold approach in medical and orthopairs [47–54].

An important consideration in formulating three-way decisions is the division of the universal set into three pair-wise disjoint regions. It is recently argued that an equally important consideration is the design of effective strategies for processing the three regions [39]. The realization of these two essential components, i.e., division and processing lead to the trisecting and acting framework of three-way decisions [39].

The trisecting and acting framework explains and presents three-way decisions as a two step process. In the first step, i.e., trisecting, the universe is divided into three pair-wise disjoint regions. This means that we seek tripartition of the universe. In the second step, i.e., acting, strategies are designed for processing the three regions to obtain three-way decisions. This framework aimed at introducing three-way decisions at a more generic level. Generally, the division of the universe is carried out based on an evaluation function and a pair of thresholds. The evaluation function assign an evaluation value to each object by employing some criteria. The objects whose evaluation values are at or above a certain threshold of acceptance makes up the POS region. The objects whose evaluation values are at or below a certain threshold of rejection make up the NEG region. The objects whose evaluation values are above the rejection threshold but below the acceptance threshold make up the BND region. A specific definition of evaluation based three-way decisions based on a single evaluation function (used for evaluating both acceptance and rejection) and totally order set is given in [39].

There are many issues and challenges for building and using three-way decision models. Some of these issues include the definition and construction of evaluation functions, the definition of the domains for the evaluation functions, the determination and interpretation of acceptance and rejection levels, the measurement of the quality of the three regions, generation of predictive rules from the three regions for making decisions on new objects, descriptive rules for describing the three regions and design of strategies and actions corresponding to the three regions [39, 45]. Based on how these issues are handled and interpreted, we may have different three-way decision making models and approaches. We focus on three-way decisions with probabilistic rough sets.

An architecture of protein function classification with three-way decisions

To make effective use of three-way decisions, we propose an architecture for supporting protein functions classification decisions. The architecture may be utilized in building systems to provide decision support capabilities for deciding protein functions. Fig 1 shows the logical view of the architecture and highlights its intended applications. The architecture supports user queries in the form of protein IDs (also called Uniprot IDs) which are mapped to functional classes by making use of three-way decisions. Fig 2 shows the physical view of the architecture along-with its various components. These components have different capabilities and functionalities ranging from supporting interaction of end user using interface to storing, collecting and manipulating the data for providing decision support.

The clients or end users will interact with the system through interfaces. Considering a typical client-server model over the Internet platform where the system is deployed over the servers and provides services by responding to client queries. The interfaces may be presented to the end users through Web browsers. The users can send their queries on proteins and see the resulting functions returned by the system. The interfaces should be carefully designed and has to be clear, complete, consistent and should provide guidance to users for correctly using

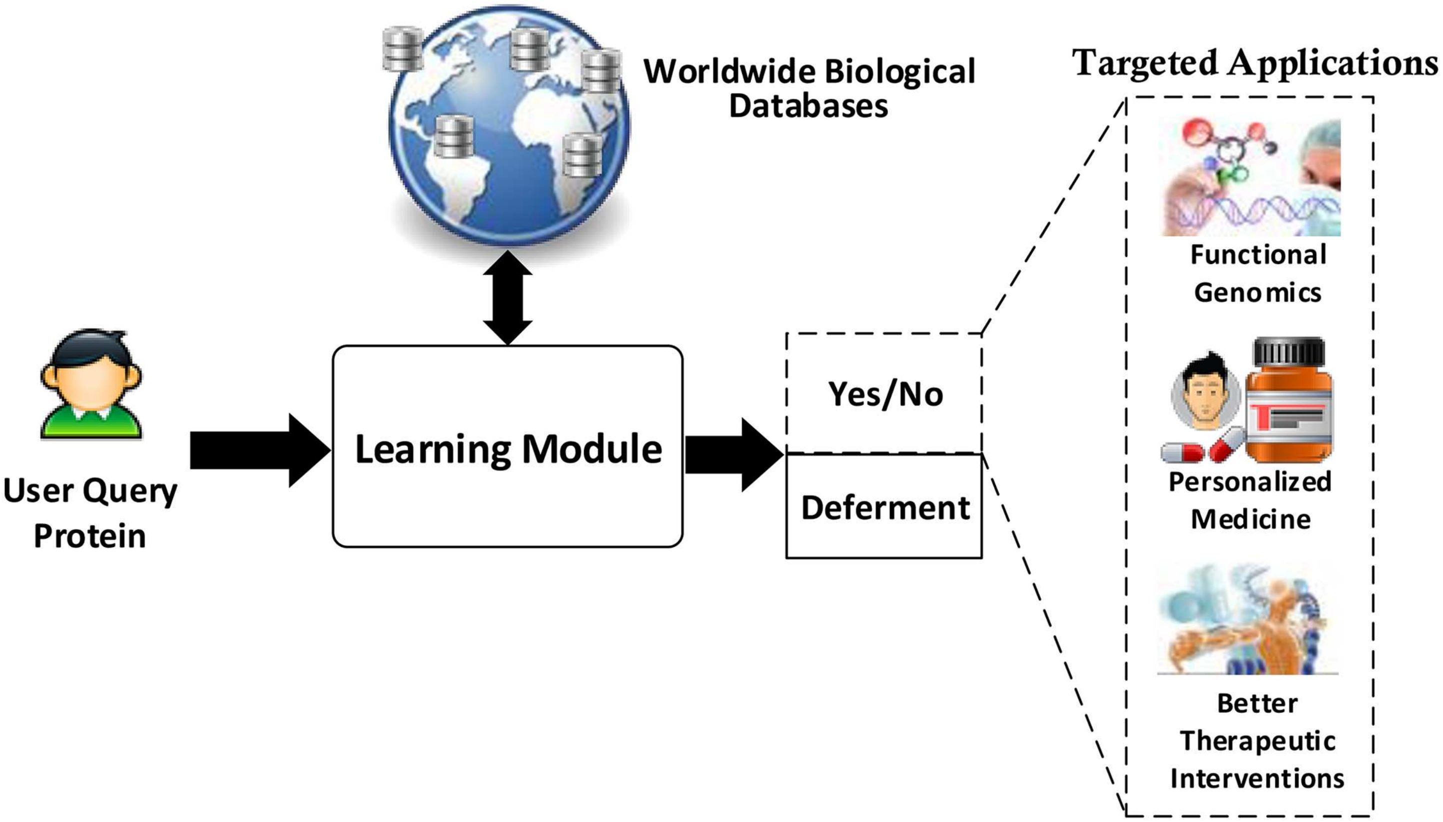

Fig 1. Logical view of the architecture with three-way decisions for protein function classification.

doi:10.1371/journal.pone.0171702.g001

the systems. In some cases, it may also provide auto correction facility. Besides interface, there are various other components at the core of the architecture namely Knowledge Discovery, Feature Extractor, Learning Module, Knowledge Base and Control Functions. We now explains each of them briefly with their intended functionalities.

Knowledge Discovery/Information Retrieval: The Knowledge Discovery module interacts with both the worldwide biological databases and feature extractor module. As new features are evolving, the feature extractor module may require different type of biological information to compute feature values. On one hand, it will provide querying and searching facilities for extracting information from relevant biological databases and on the other side, it is responsible for passing them to the feature extractor module.

Feature Extractor: The features describing proteins are computed based on relevant data extracted from biological databases which are spread around the world. The feature extractor module request or query the information retrieval component for providing relevant information necessary for computing a feature value. The information retrieval component extracts the required information from the world wide biological databases. For example, for getting one of the features namely, protein interaction networks (PIN), this module will ask for relevant information, i.e., number of interactors corresponding to a protein. The information retrieval identifies and searches the relevant databases such as STRING and IntAct databases and will pass the respective information to the feature extractor. The feature extractor module

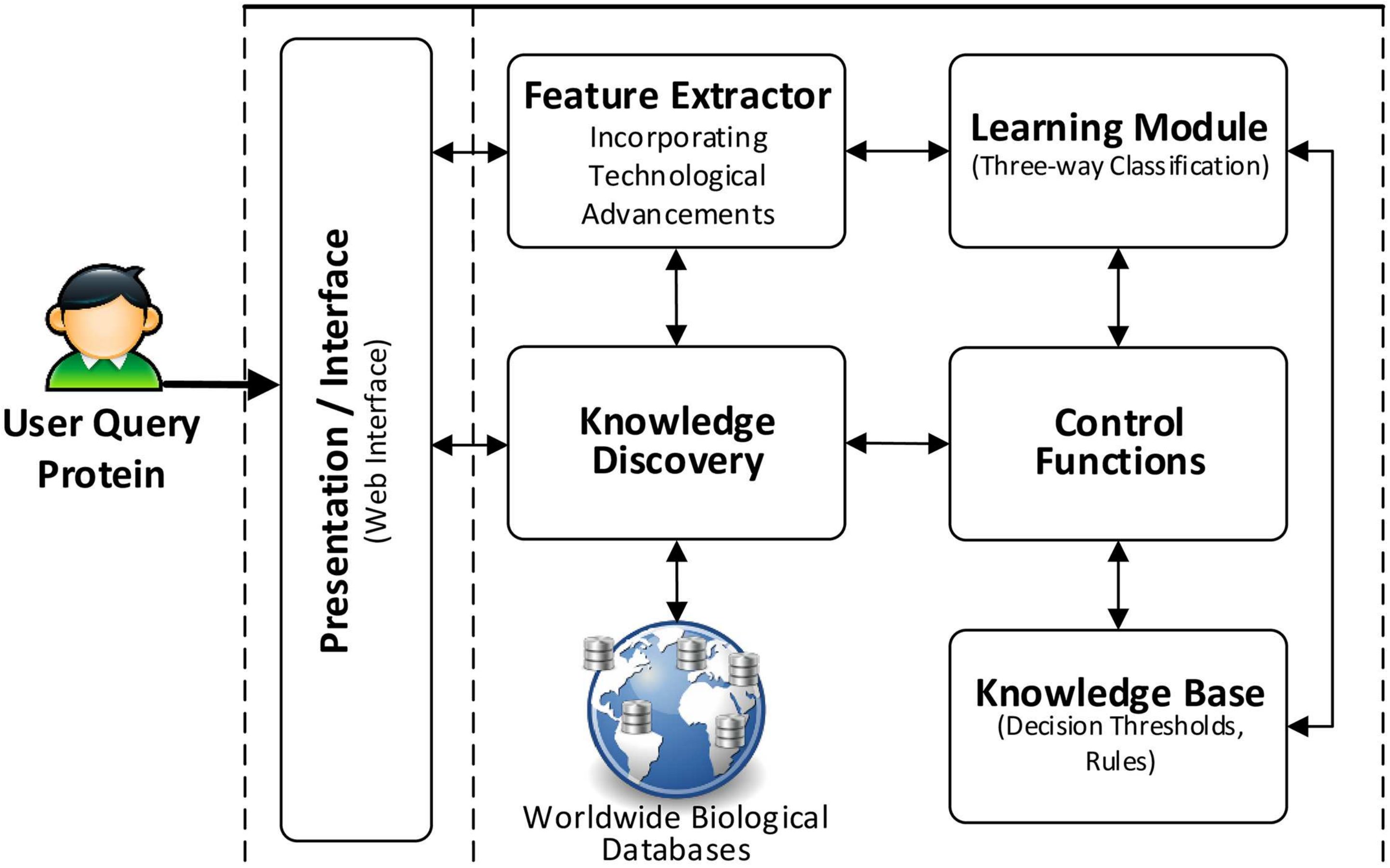

Fig 2. Physical view of the architecture with three-way decisions.

doi:10.1371/journal.pone.0171702.g002

is then responsible for further processing in order to calculate the feature value, e.g. the number of interactors present in both the databases. As new features become available due to technological advancements, the feature extractor module will ask for new type of information from the information retrieval component and do the relevant computation and processing to calculate feature values. In section **Data Preparation**, we describe different types of features that have evolved over the time and explain the types of data that is required to computed them.

Learning Module: The learning module interacts with the feature extractor and knowledge base. This module will incorporate intelligent techniques to make useful inferences from the data to reach effective classification decisions. In this article, we suggest a classification mechanism based on three-way decisions as one of the possibility. An important output of the learning module will be a set of functions that are being performed by a protein.

Knowledge Base: The knowledge base contains necessary information that is learned and made available by the learning module. The information, such as, decision thresholds and rules for classifying proteins may be stored in the knowledge base for future use.

Control Functions: This module is included to ensure security and protect the system from attacks and unauthorized usage. It should provide functionalities such as access rights and permissions.

Realization of rough sets based three-way decisions

Three-way decisions is a better and useful choice in applications with evolving information. In this section, we explain this phenomena using rough sets based three-way decisions. For the sake of completion, we review the main notions of rough sets.

Three-way decisions using rough sets

Three-way decisions using rough sets are defined by considering an information table S which is defined as a tuple.

$$S = (U, At, \{V_a | a \in At\}, \{I_a | a \in At\}), \tag{1}$$

where U is a finite set of objects also known as the universe, At is a finite set of attributes, V_a is the domain of attribute $a \in At$ and I_a is an information function which provides a mapping from $U \rightarrow V_a$. In particular, the information function I_a assigns to each object $x \in U$ a value in V_a i.e., $I_a(x) \in V_a$. A major concern in rough set theory is how to discern objects. The equivalence relation defined on U is used for this purpose. For a set of attributes $A \subset At$, the equivalence relation, namely, E_A is defined as,

$$E_A = \{(x, y) \in U \times U | \forall a \in A, I_a(x) = I_a(y)\}. \tag{2}$$

This means that any two objects x and y in U are equivalent or in other words indiscernible based on attribute set $A \in At$ if they share the same values on all attributes in A.

The equivalence relation may be used to create equivalence classes which induces a partition of U denoted by U/E. An equivalence class with an object x is given by $[x] = \{y \in U|xEy\}$. The fundamental notion of rough set theory, i.e., approximations and the three regions are defined using equivalence classes as follows.

$$\underline{apr}(C) \quad = \quad \{x \in U \ | \ [x] \subseteq C\}, \tag{3}$$

$$\overline{apr}(C) \quad = \quad \{x \in U \ | \ [x] \cap C \neq \oslash\}. \tag{4}$$

The lower and upper approximations are used to define the positive, negative and boundary regions (which leads to three-way decisions, already discussed in the section **Three-way Decisions**) given by [24, 55],

$$\begin{aligned} \text{POS}(C) &= \underline{apr}(C) \\ &= \{x \in U \ | \ [x] \subseteq C\}, \end{aligned} \tag{5}$$

$$\begin{aligned} \text{NEG}(C) &= (\overline{apr}(C))^c \\ &= \{x \in U \ | \ [x] \cap C = \oslash\}, \end{aligned} \tag{6}$$

$$\begin{aligned} \text{BND}(C) &= \overline{apr}(C) - \underline{apr}(C) \\ &= \{x \in U \ | \ [x] \nsubseteq C, [x] \cap C \neq \oslash\}. \end{aligned} \tag{7}$$

The three regions has a simple but very meaningful interpretation. We accept an object as belonging to the concept if it is in the positive region. We reject an object as belonging to the concept if it is in the negative region. We defer the decision for an object as belonging to the concept if it is in the boundary region. The three regions representation of rough sets defined in Eqs (5)–(7) has lead to the introduction of the theory of three-way decisions [39]. In fact,

Table 1. An information table for proteins.

Objects	Localization available at t_0	Interacting proteins available at t_1	No. of Domains available at t_2	Function
P_1	Mitochondria	0	0	Yes
P_2	Mitochondria	0	1	No
P_3	CytoPlasm	1	0	Yes
P_4	CytoPlasm	2	0	No
P_5	CytoPlasm	2	0	No
P_6	CytoPlasm	0	0	Yes
P_7	CytoPlasm	2	1	No
P_8	Mitochondria	0	1	No

doi:10.1371/journal.pone.0171702.t001

the major notion of three-way decisions, i.e., the division of universal set into three regions is borrowed from rough sets.

The deferment decision option which is exercised based on the boundary region is useful in at least two aspects. Firstly, it provides hints for seeking and incorporating anticipated future information in the decision making model for making decisions on the deferred cases. It is hoped that as information matures, the number of deferred cases will reduce thereby leading to more precise decisions. Secondly, the deferred cases which are typically associated with high levels of uncertainty and therefore, no obvious immediate decisions, the deferment decision option may help avoiding some false decisions. The former aspect is of particular interest from protein functions classification perspective We further elaborate this in the next section.

Three-way decisions and evolving information

The information describing the functions of proteins are evolving. An interesting issue is how to build effective decision making model for taking advantage of evolving information of protein functions. In this section, we elaborate the role of rough sets based three-way decisions as one of the possibility. In particular, we explain, how evolving information leading to new features can be effectively utilized in three-way decision making based on rough sets. We consider a demonstrative example for this purpose based on an information table of Table 1.

The rows of the Table 1 represent the proteins labeled as P_i's and the columns describe the feature or characteristics of proteins. The last column labeled as "Function" represents the decision attribute. The Function = Yes means that a protein performs the function and Function = No means that the protein does not perform the function. Let us assume three instances in time, i.e., t_0, t_1 and t_2 with $t_0 < t_1 < t_2$. Further assume that at time instance t_0, we only have information about the "Localization" of the proteins. At time instance t_1, we have additional information of "interacting proteins" of the proteins and at time instance t_2, we have more additional information about the "No. of Domains" of the proteins. Using the information available at time instance t_0, i.e., "Localization", we have the following equivalence classes.

$$\{O_1, O_2, O_8\}, \{O_3, O_4, O_5, O_6, O_7\}, \tag{8}$$

Using the above equivalence classes, we can compute the positive, negative and boundary

Table 2. Property of the three regions with evolving information.

	Localization time t_0	Interacting proteins time t_1	No. of Domains time t_2
POS(C)	$\emptyset$	$\{O_3, O_6\}$	$\{O_1, O_3, O_6\}$
NEG(C)	$\emptyset$	$\{O_4, O_5, O_7\}$	$\{O_2, O_4, O_5, O_7, O_8\}$
BND(C)	$\{O_1, O_2, \ldots, O_8\}$	$\{O_1, O_2, O_8\}$	$\emptyset$

doi:10.1371/journal.pone.0171702.t002

regions using Eqs (5)–(7). The three regions are given by,

$$\begin{aligned} \mathrm{POS(C)} &= \emptyset, \\ \mathrm{NEG(C)} &= \emptyset, \\ \mathrm{BND(C)} &= \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8\}, \end{aligned} \tag{9}$$

In the same way we can compute the three regions at time instances t_1 and t_2, when we have additional information in the form of "interacting proteins" and "No. of Domains". Table 2 summarizes the three regions corresponding to the information available at the three instances of time. Looking at the three regions for the different time instances, we may note that objects in the boundary region are decreasing and are becoming part of the positive or negative regions as more information is available at time instances of t_1 and t_2. In other words, the additional information about the "interacting proteins" and "No. of Domains" of the objects has increased the size of the positive and negative regions. This means that we can make more decisions in the form of acceptance or rejection when the level of available information increases. In this article, we argue that this property of three-way decision making can be quite useful for making decisions on protein functions classification.

In order to see the same phenomena visually, we include Fig 3. In each sub figure, the green, red and orange colours represents the positive, negative and boundary regions. The circle represents a certain concept and the small rectangles depict the equivalence classes. From Fig 3(a), we have least information and in Fig 3(c), we have most information. In Fig 3(b), we have moderate level information. We may note that as information matures, we have finer

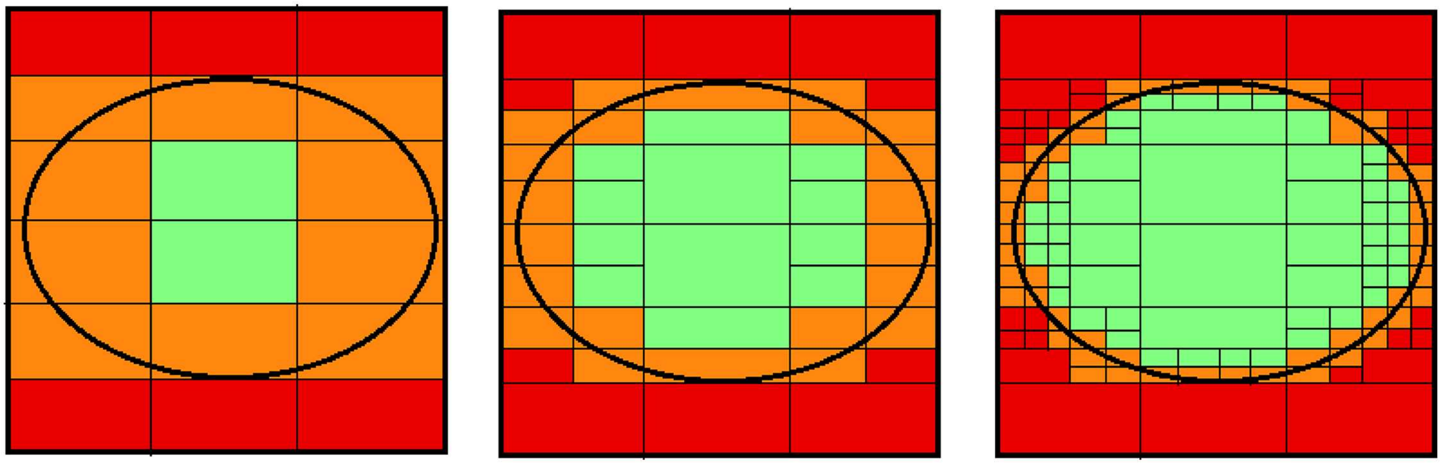

Fig 3. The three regions with evolving information. The sub-figures from left to right should be read as a, b and c respectively.

doi:10.1371/journal.pone.0171702.g003

level details leading to refined partitions. This is shown by the smaller sized boxes in Fig 3(b) and 3(c). The finer level details due to additional information enables us to move some of the equivalence classes from boundary to either positive or negative regions thereby increasing their respective sizes. This leads to fine tuning of positive and negative regions and we gradually converge to the concept, i.e., the circle (Fig 3(c)).

Three-way decisions using probabilistic rough sets

The conventional rough set model (outlined in Section **Three-way Decisions using Rough Sets**) provide a useful approach for inducing and making three-way decisions. However, this conventional rough sets model which is also sometimes referred to as Pawlak rough set model suffers from a key limitation. Researchers argue that the conditions in the upper and lower approximations may be too strict when it comes to real applications. Specifically the conditions in Eqs (3) and (4), i.e., $[x] \subseteq C$ and $[x] \cap C \neq \emptyset$ reflecting whether $[x]$ is fully contained in C and whether $[x]$ has some overlap with C, respectively, may be too restricted in the sense that they ignore the degree of an overlap between a set and a concept. To overcome this difficulty, many researchers proposed different extensions of rough sets. The probabilistic rough sets represent one class of such extensions and include decision-theoretic rough sets, variable precision rough sets, 0.5-probabilistic rough sets, Bayesian rough sets, information-theoretic rough sets and game-theoretic rough sets.

The general form of probabilistic rough sets resulted from the studies on decision-theoretic rough sets [56, 57]. The probabilistic lower and upper approximations for a concept C are defined using a pair of thresholds (α, β) as [58],

$$\underline{apr}_{(\alpha,\beta)}(C) = \{x \in U \mid P(C|[x]) \geq \alpha\}, \tag{10}$$

$$\overline{apr}_{(\alpha,\beta)}(C) = \{x \in U \mid P(C|[x]) > \beta\}, \tag{11}$$

where $P(C|[x])$ denotes the conditional probability of a concept C with an equivalence class $[x]$. Given that an object $x \in [x]$, the conditional probability highlights the evaluation of an object x to be in C. The three rough set regions based on lower and upper approximations are defined as,

$$\mathrm{POS}_{(\alpha,\beta)}(C) = \{x \in U|P(C|[x]) \geq \alpha\}, \tag{12}$$

$$\mathrm{NEG}_{(\alpha,\beta)}(C) = \{x \in U|P(C|[x]) \leq \beta\}, \tag{13}$$

$$\mathrm{BND}_{(\alpha,\beta)}(C) = \{x \in U|\beta < P(C|[x]) < \alpha\}. \tag{14}$$

The $\mathrm{POS}_{(\alpha,\beta)}(C)$, $\mathrm{NEG}_{(\alpha,\beta)}(C)$ and $\mathrm{BND}_{(\alpha,\beta)}(C)$ in Eqs (12)–(14) are referred to as positive, negative and boundary regions, respectively. Based on how these thresholds are determined and interpreted we have different probabilistic rough set models.

To demonstrate the use of three-way decisions for proteins functions classification, we focus on two probabilistic rough set models, namely, GTRS [25–27] and ITRS [59]. These two models have at least two advantages over other models.

- Firstly, compared to some of the earlier probabilistic models, such as, 0.5-probabilistic rough set model and $(0.5, \beta)$ model, where due to restricted pairs of thresholds, the determination and interpretation of thresholds are ignored, the GTRS and ITRS allows for investigation and examination of thresholds based on different aspects.

- Secondly, unlike other models that require user intervention to set the thresholds, such as, decision-theoretic rough sets and variable precision rough sets, the GTRS and ITRS can be used to learn and set the thresholds automatically when combined with some typical search mechanism [25].

For the sake of being complete, we briefly explain and discuss the GTRS and ITRS models.

Three-way decisions using game-theoretic rough sets. The game-theoretic rough sets or GTRS utilizes game-theoretic formulation to determine thresholds of probabilistic rough sets [26, 27]. In particular, the thresholds are interpreted based on a tradeooff solution between multiple criteria employed in a game setting for analyzing rough sets [25–27]. A typical game in GTRS has three essential components, i.e., game players, strategies and payoff or utility functions. These components are generally represented as a tuple $\{P, S, u\}$. We now explain each of them.

Game players: The players in the game are denoted by a set P. Generally, there can be n players in a game. However, for the sake of simplicity, a two player game is commonly considered in GTRS. Based on the overall game objective and goals, we may have different types of game players. For instance, in a previous game for analyzing region uncertainty, the players were defined as the uncertainty of the immediate and deferred decision regions and in another game that seek for a balanced rough set model, the players of accuracy and generality were used [25, 60]. In general, the players in the game are selected to reflect the overall purpose of the game. In GTRS, the players are defined as different aspects and properties of rough sets based classification and decision making such as accuracy, generality, precision and uncertainty.

Strategies: Each player in the game participate by playing different strategies. The set of strategies available to player i is denoted by S_i. The Cartesian product of all possible strategy sets is denoted by $S = S_1 \times S_2 \times \ldots \times S_n$, where S contains ordered pairs of the form $(s_1, s_2, \ldots, s_n)$ such that $s_1 \in S_1$, $s_2 \in S_2$ and $s_n \in S_n$. Each order pair in S is called a strategy profile and represents a certain situation encountered in a game.

The strategies in GTRS are realized as different changes and modifications in the (α, β) thresholds. Depending on the initial values of thresholds, we may have different types of strategies. For instance, if the initial values of (α, β) are set to $(1, 0.5)$, then the strategies may be formulated as decreasing levels of α and β. Alternatively, when the initial values of (α, β) are set to $(1, 0)$, then the strategies may be formulated as decreasing levels of α and increasing levels of β. Please note that in order to keep the regions disjoint, it is assumed that $0 \leq \beta < \alpha \leq 1.0$. The strategies of the players in a game lead to effective modification of the thresholds which ultimately determines the final configuration of the thresholds.

Payoff functions: The payoff functions for the players are represented by a set $u = (u_1, \ldots, u_n)$. Each u_i is a real valued utility function for player i and it maps the strategy profiles to real values, i.e., $u_i: S \mapsto \Re$. In particular, the payoffs reflect the utilities of performing or selecting a certain strategy. Recall the game players in GTRS which are represent different aspects or properties of rough sets, the payoff function for a certain player is based on particular measure employed for evaluating its respective property.

In a game setting, every player wants to perform a strategy that will maximize its payoff. The selected strategies of the players however affect their opponents payoffs. The game solution is used to choose a balanced and trade off point based on the utilities of all the players. The game solution of Nash equilibrium is commonly used in GTRS for this purpose.

Considering a strategy profile $s_{-i} = (s_1, s_2, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$, which means a strategy profile without player i strategy. Moreover, the strategy profile $(s_1, s_2, \ldots, s_n)$ may be denoted in revised notation as (s_i, s_{-i}). The strategy profile $(s_1, s_2, \ldots, s_n) = (s_i, s_{-i})$ is a Nash equilibrium if

[61],

$$\forall i, \forall s_i' \in S_i, \quad u_i(s_i, s_{-i}) \geq u_i(s_i', s_{-i}), \quad \text{where}(s_i' \neq s_i) \tag{15}$$

This means that for all players i, their respective strategies, i.e., s_i is the best response to s_{-i}. In other words, a strategy profile constitutes a Nash equilibrium when no player is benefited from changing his strategy alone.

The above game description is used in GTRS to formulate a game. However, with a single one time and non-repeated game, we may not be able to reach effective thresholds that fulfill the demands of the underlying applications. We need to repeat the game. The essential idea is to repeatedly modify and refine the thresholds, until we achieve certain performance criteria. By formulating a game and utilizing the notions such as game solution and repetitive games, the GTRS seek for an effective configuration of the threshold levels that are employed in the probabilistic rough sets framework to induce three-way decisions.

Three-way decisions using information-theoretic rough sets. The Information-theoretic rough sets (or ITRS) approach the threshold determination issue from the viewpoint of minimizing the information uncertainty of the probabilistic rough set regions [59]. Let $\Delta_P(\alpha, \beta)$, $\Delta_N(\alpha, \beta)$ and $\Delta_B(\alpha, \beta)$ denote the overall uncertainties of the probabilistic positive, negative and boundary regions respectively. The ITRS is based on minimization or optimization of the following problem.

$$\begin{aligned} &\arg \min_{(\alpha,\beta)} \Delta(\alpha, \beta), \qquad \text{where,} \\ &\Delta(\alpha, \beta) = \Delta_P(\alpha, \beta) + \Delta_N(\alpha, \beta) + \Delta_B(\alpha, \beta) \end{aligned} \tag{16}$$

Please be noted that we used slightly modified notations that were reported in [25]. Eq (16) suggests that we seek thresholds (α, β) that will minimize the uncertainty of the three regions.

The overall uncertainty in Eq (16) is typically considered as an average uncertainty of the three regions [59].

$$\Delta_P(\alpha, \beta) = P(\text{POS}_{(\alpha,\beta)}(C))\delta_P(\alpha, \beta), \tag{17}$$

$$\Delta_N(\alpha, \beta) = P(\text{NEG}_{(\alpha,\beta)}(C))\delta_N(\alpha, \beta), \tag{18}$$

$$\Delta_B(\alpha, \beta) = P(\text{BND}_{(\alpha,\beta)}(C))\delta_B(\alpha, \beta), \tag{19}$$

where $\delta_P(\alpha, \beta)$, $\delta_N(\alpha, \beta)$ and $\delta_B(\alpha, \beta)$ are the uncertainties of the three regions which may be computed and interpreted using different measures of uncertainties. Moreover, $P(\text{POS}_{(\alpha,\beta)}(C))$, $(\text{POS}_{(\alpha,\beta)}(C))$ and $P(\text{POS}_{(\alpha,\beta)}(C))$ are the probabilities of the three regions. Two measures, i.e., Shannon entropy and gini coefficient are being previously employed for interpreting and measuring the uncertainties of the three regions, i.e., $\delta_P(\alpha, \beta)$, $\delta_N(\alpha, \beta)$ and $\delta_B(\alpha, \beta)$. We now define each of them.

Consider a partition based on a concept C, given by, $\pi_C = \{C, C^c\}$ and another partition with respect to the thresholds (α, β), given by, $\pi_{(\alpha,\beta)} = \{\text{POS}_{(\alpha,\beta)}(C), \text{NEG}_{(\alpha,\beta)}(C), \text{BND}_{(\alpha,\beta)}(C)\}$. The uncertainty in π_C with respect to the three probabilistic regions based on Shannon entropy is given by, [59],

$$\begin{aligned}\delta_P(\alpha,\beta) &= H(\pi_C|\mathrm{POS}_{(\alpha,\beta)}(C)) = -P(C|\mathrm{POS}_{(\alpha,\beta)}(C))\ \log\ P(C|\mathrm{POS}_{(\alpha,\beta)}(C))\\ &\quad -P(C^c|\mathrm{POS}_{(\alpha,\beta)}(C))\ \log\ P(C^c|\mathrm{POS}_{(\alpha,\beta)}(C)),\end{aligned} \tag{20}$$

$$\begin{aligned}\delta_P(\alpha,\beta) &= H(\pi_C|\mathrm{NEG}_{(\alpha,\beta)}(C)) = -P(C|\mathrm{NEG}_{(\alpha,\beta)}(C))\ \log\ P(C|\mathrm{NEG}_{(\alpha,\beta)}(C))\\ &\quad -P(C^c|\mathrm{NEG}_{(\alpha,\beta)}(C))\ \log\ P(C^c|\mathrm{NEG}_{(\alpha,\beta)}(C)),\end{aligned} \tag{21}$$

$$\begin{aligned}\delta_P(\alpha,\beta) &= H(\pi_C|\mathrm{BND}_{(\alpha,\beta)}(C)) = -P(C|\mathrm{BND}_{(\alpha,\beta)}(C))\ \log\ P(C|\mathrm{BND}_{(\alpha,\beta)}(C))\\ &\quad -P(C^c|\mathrm{BND}_{(\alpha,\beta)}(C))\ \log\ P(C^c|\mathrm{BND}_{(\alpha,\beta)}(C)).\end{aligned} \tag{22}$$

Where we used the additional notations $H(\pi_C|\mathrm{POS}_{(\alpha,\beta)}(C))$, $H(\pi_C|\mathrm{POS}_{(\alpha,\beta)}(C))$ and $H(\pi_C|\mathrm{POS}_{(\alpha,\beta)}(C))$ to be consistent with the earlier notations [59]. The measure of gini coefficient is also used in the same way to determine the uncertainties of the three regions [62]. The uncertainties of the three regions are computed as [62],

$$\begin{aligned}\delta_P(\alpha,\beta) &= G(\pi_C|\mathrm{POS}_{(\alpha,\beta)}(C)) = 1 - P(C|\mathrm{POS}_{(\alpha,\beta)}(C))^2\\ &\quad -P(C^c|\mathrm{POS}_{(\alpha,\beta)}(C))^2,\end{aligned} \tag{23}$$

$$\begin{aligned}\delta_N(\alpha,\beta) &= G(\pi_C|\mathrm{NEG}_{(\alpha,\beta)}(C)) = 1 - P(C|\mathrm{NEG}_{(\alpha,\beta)}(C))^2\\ &\quad -P(C^c|\mathrm{NEG}_{(\alpha,\beta)}(C))^2,\end{aligned} \tag{24}$$

$$\begin{aligned}\delta_B(\alpha,\beta) &= G(\pi_C|\mathrm{BND}_{(\alpha,\beta)}(C)) = 1 - P(C|\mathrm{BND}_{(\alpha,\beta)}(C))^2\\ &\quad -P(C^c|\mathrm{BND}_{(\alpha,\beta)}(C))^2.\end{aligned} \tag{25}$$

Please note again that the notation $G(\pi_C|\mathrm{BND}_{(\alpha,\beta)}(C))$, $G(\pi_C|\mathrm{BND}_{(\alpha,\beta)}(C))$ and $G(\pi_C|\mathrm{BND}_{(\alpha,\beta)}(C))$ are being used for the sake of being consistent with the previous notations [62].

The ITRS is generally combined with a searching mechanism to determine effective thresholds. In particular, the minimization of overall uncertainty in Eq (16), is used to guide the search towards optimal thresholds. Recently, the gradient descent approach was suggested in this regards [59].

Three-way decision algorithm for classifying protein

In this section, we look at three-way decision approach from implementation perspective. Algorithm 1 is presented for this purpose. The algorithm explains how three-way decisions can be used in classifying proteins with evolving information.

Algorithm 1 Iterative Three-way decision making algorithm

```
Input: An information table containing a new feature and $\mathrm{POS}_{(\alpha,\beta)}(C)$, $\mathrm{NEG}_{(\alpha,\beta)}(C)$, and $\mathrm{BND}_{(\alpha,\beta)}(C)$ based on information from previous features
Output: Updated regions, $\mathrm{POS}_{(\alpha,\beta)}(C)$, $\mathrm{NEG}_{(\alpha,\beta)}(C)$ and $\mathrm{BND}_{(\alpha,\beta)}(C)$
if $Q_P(\alpha, \beta) \geq c_1$ and $Q_N(\alpha, \beta) \geq c_2$ then
    Determine thresholds $(\alpha', \beta')$ using GTRS and ITRS for information table with $U = \mathrm{BND}_{(\alpha,\beta)}(C)$
    $\mathrm{POS}_{(\alpha',\beta')}(C) = \{x \in \mathrm{BND}_{(\alpha,\beta)}(C) \mid P(C|[x]) \geq \alpha'\}$
    $\mathrm{NEG}_{(\alpha',\beta')}(C) = \{x \in \mathrm{BND}_{(\alpha,\beta)}(C)|P(C|[x]) \leq \beta'\}$
```

```
5:   $\mathrm{BND}_{(\alpha',\beta')}(C) = \{x \in \mathrm{BND}_{(\alpha,\beta)}(C) \mid \beta' < P(C|[x]) < \alpha'\}$
6:   $\mathrm{POS}_{(\alpha,\beta)}(C) = \mathrm{POS}_{(\alpha',\beta')}(C) \bigcup \mathrm{POS}_{(\alpha,\beta)}(C)$
7:   $\mathrm{NEG}_{(\alpha,\beta)}(C) = \mathrm{NEG}_{(\alpha',\beta')}(C) \bigcup \mathrm{NEG}_{(\alpha,\beta)}(C)$
8:   $\mathrm{BND}_{(\alpha,\beta)}(C) = \mathrm{BND}_{(\alpha',\beta')}(C) - \mathrm{BND}_{(\alpha,\beta)}(C)$
9: else
10:   Determine thresholds $(\alpha, \beta)$ using GTRS and ITRS.
11:   $\mathrm{POS}_{(\alpha,\beta)}(C) = \{x \in U \mid P(C|[x]) \geq \alpha\}$
12:   $\mathrm{NEG}_{(\alpha,\beta)}(C) = \{x \in U \mid P(C|[x]) \leq \beta\}$
13:   $\mathrm{BND}_{(\alpha,\beta)}(C) = \{x \in U \mid \beta < P(C|[x]) < \alpha\}$
14: end if
15: return $\mathrm{POS}_{(\alpha,\beta)}(C)$, $\mathrm{NEG}_{(\alpha,\beta)}(C)$, $\mathrm{BND}_{(\alpha,\beta)}(C)$
```

The algorithm accepts information table containing information about a new feature and the three regions based on the previous features, i.e., positive, negative and boundary regions denoted as $\mathrm{POS}_{(\alpha,\beta)}(C)$, $\mathrm{NEG}_{(\alpha,\beta)}(C)$, and $\mathrm{BND}_{(\alpha,\beta)}(C)$, respectively. In line 1, the algorithm evaluates the positive and negative regions by employing some quality criteria denoted as $Q_{POS}(\alpha,\beta)$ and $Q_{NEG}(\alpha,\beta)$ (representing some quality related aspect of the positive and negative regions, respectively). These notations are introduced to represent the general notion of any criteria that is employed for evaluating the three regions. They may be interpreted in terms of cost, risks, uncertainty, accuracy or precision. The quality of the regions may be measured based on the notions such as risks, cost, uncertainty, accuracy or precisions. As discussed in the previous subsection titled **Three-way Decisions and Evolving Information**, when the features evolve, the positive region gradually converges to the concept C (i.e., more precisely reflect the region representing the concept) and the negative region gradually converges to the complement of the concept C^c (i.e., more precisely reflect the region not in the concept), respectively. As a result, the quality of the two regions improves. As improvement in quality is a gradual process in this case, at the current level of information, the quality of the positive and negative regions may or may not be effective (please be noted that the term effective here may have different interpretation based on the underlying applications). We deal with these two cases separately.

If the quality of the regions are above some acceptable levels c_1 and c_2, we will only examine the objects in the boundary region and will not further investigate the positive and negative regions. The boundary is expected to shrink further as we have access to new features. In any other case, we will examine the full information table to obtain the three regions. In other words, we are not satisfied with the quality of the positive and negative regions (they are below the levels c_1 and c_2) and we expect that additional information may improve their respective quality levels. We first deal with the former case. In line 3, we determine thresholds based on the reduced information table with $U = \mathrm{BND}_{(\alpha,\beta)}(C)$. As new information becomes available in the form of a new feature, we may be able to confidently classify further objects in the boundary. This is shown in line 4-7 where we further divide the objects in the boundary region. In line 6-8, we update the three regions based on further examination of the boundary. From line 10-13 we examine the case when the positive and negative regions based on the previous knowledge were not of acceptable quality. We therefore examine the full information table and update the three regions accordingly. The Fig 4 represents the essential ideas of the algorithm 1 in diagrammatic form.

It may be noted that the constants c_1 and c_2 may be defined in different ways depending on the application needs and requirements. For instance, if we want to reduce the processing overload, we may define them moderately. On the other hand, if processing overload is not an issue and we are more concerned about the accuracy, then we may define them more strictly. Other ways in which they may be defined are by making comparison with the quality of the

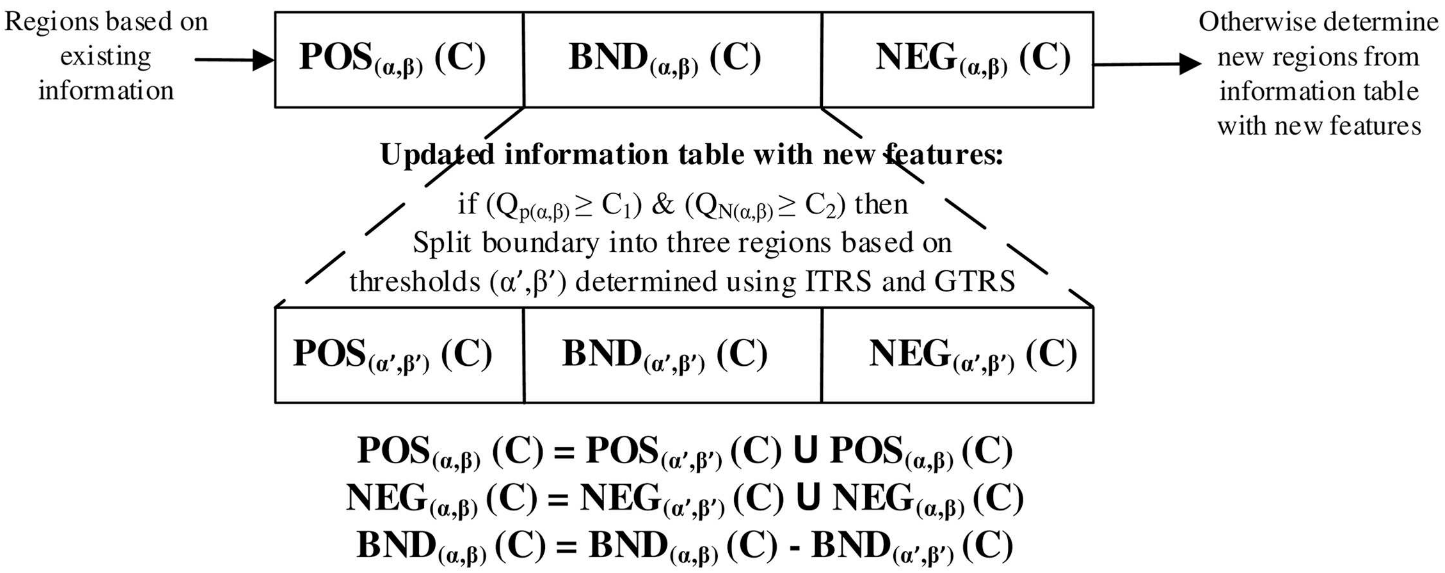
Regions based on existing information
$POS_{(\alpha,\beta)}$ (C)
$BND_{(\alpha,\beta)}$ (C)
$NEG_{(\alpha,\beta)}$ (C)
Otherwise determine new regions from information table with new features
Updated information table with new features:
if ($Q_{p(\alpha,\beta)} \geq C_1$) & ($Q_{N(\alpha,\beta)} \geq C_2$) then
Split boundary into three regions based on
thresholds (α′,β′) determined using ITRS and GTRS
$POS_{(\alpha',\beta')}$ (C)
$BND_{(\alpha',\beta')}$ (C)
$NEG_{(\alpha',\beta')}$ (C)
$POS_{(\alpha,\beta)}(C) = POS_{(\alpha',\beta')}(C) \cup POS_{(\alpha,\beta)}(C)$
$NEG_{(\alpha,\beta)}(C) = NEG_{(\alpha',\beta')}(C) \cup NEG_{(\alpha,\beta)}(C)$
$BND_{(\alpha,\beta)}(C) = BND_{(\alpha,\beta)}(C) - BND_{(\alpha',\beta')}(C)$

Fig 4. Visualization of iterative three-way decision making algorithm.

doi:10.1371/journal.pone.0171702.g004

regions obtained with the standard Pawlak models or other known models in the domains or by considering the improvement in quality based on the new features.

Experiment setup

Data preparation

To evaluate the use of three-way approach, we examine the application of three-way decisions on well studied *Saccharomyces cerevisiae* species proteins [63, 64], obtained from most widely used *Uniport* database [29]. From various classification schemes developed to standardize the descriptions of protein functions, we chose the state of the art Gene Ontology (GO) [30] classification scheme. The gene ontology is a structured, controlled vocabulary of protein functions also called terms. GO terms provide consistency in annotating protein roles in the cellular context. It is arranged in a DAG (please refer to Section Background) structure in which each node of the graph represents a unique functional term and each term is arranged in a parent child relationship with other terms. The child term either *is a* special case of the parent or is a *part of* the parent process i.e., a sub-process or component. For the evaluation of our methodology we operate on *molecular function* category of gene ontology. To reveal the evolving nature of biological information, we present features in the order in which they are evolved over the time i.e., most basic type of information is presented first and so on [3]. For classifying a protein into one or more molecular function terms of gene ontology, we retrieve ten different types of features from varied biological databases. Each feature is helpful in characterizing one or more functional categories and is represented by the symbol *Fi*.

Protein Sequence Length (F_1): In every cell, genes are converted into proteins via the processes of transcription and translation also called the central dogma of molecular biology. The end product of these processes is a sequence built from twenty amino acids, and is commonly known as the primary structure of a protein. The amino acid sequence is the most basic type of information available about a protein, as it can provide concrete evidence about different characteristics of a protein such as its binding sites, sub-cellular localization, structure and function. To quantify these biological aspects of a protein, we use feature (namely F_1), as the length of protein sequence which is extracted from *Uniprot* database [29].

Protein Localizations (F_2): The location of a protein in the cell can also be associated with its function. Co-localized proteins are more likely to be part of same molecular activity. Likewise, proteins localizing in many different locations can be part of diverse activities. To capture this aspect, we calculate feature F_2 as the number of locations a protein can localize. The protein localization data is retrieved from the *Uniprot* database [29].

Biological Processes (F_3): A biological process refers to the series of events performed by one or more assemblies of molecular functions with a defined beginning and end. A protein participating in many biological processes is more likely to have many molecular level roles. Thus the number of biological processes of a protein can also be used to capture the molecular level activities of a protein. As a third feature (F_3), we count the number of biological processes of a protein. It is obtained by retrieving counts of Biological Process ontology terms from the Gene Ontology database [30].

Number of Interacting Proteins (F_4): For calculating the fourth feature of our method, namely *F4*, we use genome wide protein-protein interactions (PPI) data to predict proteins function. In a living cell, protein-protein interactions are amongst the most ubiquitous types of interactions and their precise knowledge helps in understanding the activities performed by a protein as well as the processes it is part of. A protein having many different interacting partners can be said to be part of many different functions. Thus the number of interactions (F_4) of a protein can be linked to the wide variety of activities it performs. We obtain PPI data from most widely used PPI databases, namely, *IntAct* [65] and *STRING* [66]. Since protein-protein interactions databases are noisy, we only consider interactions that are experimentally verified and are supported by at least two experiments.

Number of Domains (F_5): Protein domains are the sequential units that fold in a particular shape, making independent structures in different proteins. Several classification schemes have been proposed e.g., [67] to define and demarcate different domains of which some based on clustering conserved subsequences into related domain families, others on known distinct structural classes [68]. One of the most famous and widely used domain classification schema is the *Interpro* database [69]. *InterPro* database contains diagnostic signatures of protein sequences consisting of models e.g., regular expressions models, Hidden Markov Models etc., which describe protein domains found within sequence. Domains are the most important feature among relevant sequence features of a protein that associate it to a particular kind of functions. To integrate domain relevance we also use as a feature (namely F_5) the number of *Interpro* conserved domains within a query protein sequence.

Number of Conserved Motifs (F_6): A motif is a conserved amino acid sequence pattern in a protein sequence that may be associated to a specific function. These subsequences may often contain small “gaps” of fixed or variable lengths among amino acids of the subsequence. The knowledge of exact patterns of motifs and their functions is helpful in the understanding of structure and function of related proteins in which such motifs may appear. For example, if a motif of a certain family is present in a protein sequence then it will make it highly probable to functionally associate that protein with the functions of that motif i.e., we can associate proteins with functions by merely checking the presence of certain motifs. Thus in our technique, as sixth feature (F_6) we count the number of conserved motifs in a protein sequence using *Prosite* motif database [70].

Number of Protein Structures (F_7): A protein’s primary structure consists of sequence of amino acids. These amino acids due to their varied physical and chemical properties as well as the presence of different participant cellular forces, assumes a unique configuration in three-dimensional space. This stable configuration of proteins is also called the tertiary structure of proteins. This final configuration or structure of a protein is strongly correlated to its function, because in many biological processes, the interacting proteins have to come into physical contact in order to accomplish the desired function. The structure of a protein also determines many of its functional characteristics, for example its inter-facial binding sites, the specific ligands it binds to, cellular localizations, as well as other proteins it can interact with. Among all the structural databases PDB (Protein Data Bank) [71] is by far the most reliable, wide-ranging as well as popular repository for experimentally derived protein 3D structures. We query the PDB database to obtain the number of experimentally determined structures associated to a protein under investigation and use this information as a feature (namely F_7) to characterize its function.

Molecular Weight of Protein (F_8): Although weight of a protein is not strongly related to its function but in some cases it can be used to generally group them into broader functional categories. We retrieve Molecular weight rounded to the nearest mass unit (Dalton) from Uniprot Database [29] and use it as a feature (namely F_8) for our 3-way classifier.

Number of Interfacing Residues in Protein Structure (F_9): Many proteins bind together and form multi-protein complexes. Different proteins in the complex perform different functions. These functions are associated with the number of residues on a protein's interface that enables it to stabilize, bind and form complexes. Owing to the significance of interfacing residues we utilize a structural feature i.e., the number of residues on the protein's interface to characterize function of a protein. The interfacing residues can vary for various functional activities. To capture this aspect we used *PDBe PISA* server [72], to retrieve the number of predicted interfacing residues and use it as feature (namely F_9) for our 3-way classifier.

Binding sites in the Predicted Interface (F_{10}): A protein's physical interaction with other molecules, determines its biological activities. For example antibody proteins selectively bind to viruses or bacteria to choose them for destruction, the hexokinase protein binds to ATP molecule as well as with glucose molecule in order to catalyze their chemical reaction, and so on. Without any doubt almost all proteins stick, or bind, to other molecules in order to perform their activities at molecular level. Some proteins bind very tightly while others bind for a short period of time depending on their specificity as well as the molecular task they have to perform. Each protein can usually bind to one or few other molecules determined by the nature of binding residues (also called binding sites) at its surface. To determine the specificity of a protein for binding and performing wide variety of functions we calculate a feature (namely (F_{10}), which is the number of binding sites on its surface that are predicted using *PDBeFold* Server [73].

The above features namely, F_1 to F_{10}, are extracted using the Feature Extractor module (already described in Section **An Architecture of Protein Function Classification with Three-way Decisions**), from the world wide biological databases using the knowledge discovery module. The Feature Extractor module also has the capability to incorporate any new feature, say F_{11} in the predictive task. To imitate the ever evolving nature of biological information, we selected and ranked features from most basic type to the latest type i.e, F_1 namely, sequence similarity, is a basic type of feature and F_{10} namely, number of binding sites on a protein interface, is a specific feature known after information evolution.

Three-way approaches used in the experiments

We performed experiments with five three-way decision making approaches based on GTRS and ITRS. Specifically three of these approaches are based on GTRS and two of them are based on ITRS.

The three approaches with GTRS are based on different games that are formulated based on description in Section **Three-way Decisions using GTRS**. The essential difference in these games are the consideration of different types of game players. Two of these games are based on examining a balance between the uncertainties of probabilistic rough set regions. These games are based on two players, namely, immediate decision region, denoted as *I* and deferred decision region, denoted as *D*. The player *I* reflects the collective uncertainty in probabilistic positive and negative regions and the player *D* denotes the uncertainty in the probabilistic boundary region. By realizing changes in thresholds as game strategies, the players in a game compete in a game by selecting appropriate changes in the thresholds which are used in determining the final settings of the thresholds. Two games are constructed with these game players, i.e., player *I* and *D* by realizing different interpretation and computation of uncertainty. In one game, the uncertainty is measured with the Shannon entropy and in another game it is measured with gini coefficient. These two games will be referred to as GTRS_E and GTRS_G, respectively. These game were previously examined in the context of text categorization and medical decision making [25, 62, 74]. The third game in GTRS is based on determining a trade off between two aspects of rough sets based classification, namely, accuracy and generality. This game was previously examined in the context of recommender systems in [60]. We will refer to this game as $\text{GTRS}_{(A,G)}$.

Two approaches are considered with the ITRS. These two approaches are ITRS based on Shannon entropy and ITRS based on Gini coefficient as discussed in Section **Three-way Decisions using ITRS**. We denote these approaches as ITRS_E and ITRS_G, respectively. Both of these measures interpret the uncertainty in a different way and therefore will lead to different thresholds.

In all experiments, we considered the top five most frequent protein functions in the database. For each protein function (recall that each protein function is considered as a category), we learn the probabilistic thresholds (α, β) and performed three-way decisions using the five approaches discussed above in this section. We considered four feature sets. In each feature set, we consider the features whose relevant information was previously available or which emerged roughly at the same time. In particular, the first feature set comprise of F_1, F_2 and F_3 (please refer to Section **Data Preparation** for their details). We denote the first feature set as FS_1. The second feature set denoted as FS_2, is given by $FS_1 \cup F_4$. The third and fourth feature sets, denoted as FS_3 and FS_4 are given by $FS_3 = FS_2 \cup \{F_5, F_6\}$ and $FS_4 = FS_3 \cup \{F_7, F_8, F_9, F_{10}\}$, respectively. Please be noted that the FS_1 contains the oldest available information about proteins while FS_4 is the represents the most recent information comprising the previous knowledge and newly evolved information. Finally, all the results are based on 10 folds cross validation.

Results and discussion

Experimental results

We report the results of accuracy and generality for the considered three-way approaches. The accuracy and generality may be defined as [24],

$$Accuracy(\alpha, \beta) = \frac{|(\text{POS}_{(\alpha,\beta)}(C) \cap C) \bigcup (\text{NEG}_{(\alpha,\beta)}(C) \cap C^c)|}{|\text{POS}_{(\alpha,\beta)}(C) \bigcup \text{NEG}_{(\alpha,\beta)}(C)|}, \tag{26}$$

$$Generality(\alpha, \beta) = \frac{|\text{POS}_{(\alpha,\beta)}(C) \bigcup \text{NEG}_{(\alpha,\beta)}(C)|}{|U|}, \tag{27}$$

Table 3. Results of accuracy and generality for GTRS.

Features	$GTRS_{(A,G)}$		$GTRS_E$		$GTRS_G$	
	Accuracy	Generality	Accuracy	Generality	Accuracy	Generality
FS_1	0.8031	0.2377	**0.7938**	0.2276	0.7969	0.2913
FS_2	**0.8058**	0.3108	0.8077	0.2888	**0.8071**	0.3467
FS_3	0.7808	0.6654	0.7853	0.6737	0.7807	0.6724
FS_4	0.7815	**0.6875**	0.7797	**0.6851**	0.7807	**0.6891**

doi:10.1371/journal.pone.0171702.t003

The accuracy highlights the relative number of correct classification decisions for the objects in the universal set and the generality reflects the relative number of objects for whom classification decisions can be made. Table 3 shows the results obtained with the GTRS based approaches. The rows of the table correspond to the results obtained with a particular set of features and the columns correspond to results of accuracy and generality for different approaches. The best results for accuracy and generality against each approach is presented in bold. We may note that the best results for the generality for the three approaches are against the highest feature set size. Moreover, the generality of the three approaches improve as the feature set size is increased. In particular, the generality of $GTRS_{(A,G)}$ with lowest feature size is 23.77% and the highest feature set size is 68.75%. This represents a total increase of 44.98% in generality. For the other two approaches, i.e., $GTRS_E$ and $GTRS_G$, similar increases in generality with values 45.75% and 39.78% are noted based on the lowest feature set size and highest feature set size. Since the features represent the available level of information for predicting protein functions. We may conclude from these results that as the level of information improves (i.e., as we include more features), we are able to make classification decisions for more proteins.

Let us now look at the results of accuracy in Table 3. We may observe that in general, the values of accuracy decrease slightly as we move from lower to higher feature set sizes. However, compared to the generality, we do not have significant different between these values. For the three approaches, i.e., $GTRS_{(A,G)}$, $GTRS_E$ and $GTRS_G$, the differences between the values of accuracy for the lowest and highest feature set sizes are 2.16%, 1.41% and 1.62%, respectively. From the results of accuracy and generality, we may notice that by increasing the number of features or the level of information, we are able to make more decisions while mainlining the same or similar level of accuracy.

Table 4 shows the results obtained with the ITRS based approaches. The increase in generality for the two approaches, i.e., $ITRS_E$ and $ITRS_G$ between the lowest feature set size and highest feature set sizes are 14.36% and 18.29%, respectively. The accuracy values for the $ITRS_E$ and

Table 4. Results of accuracy and generality for ITRS.

Features	$ITRS_E$		$ITRS_G$	
	Accuracy	Generality	Accuracy	Generality
FS_1	**0.8247**	0.6008	**0.8139**	0.5972
FS_2	0.8043	0.6296	0.8101	0.631
FS_3	0.7927	0.7394	0.7878	0.7411
FS_4	0.791	**0.7444**	0.7865	**0.7801**

doi:10.1371/journal.pone.0171702.t004

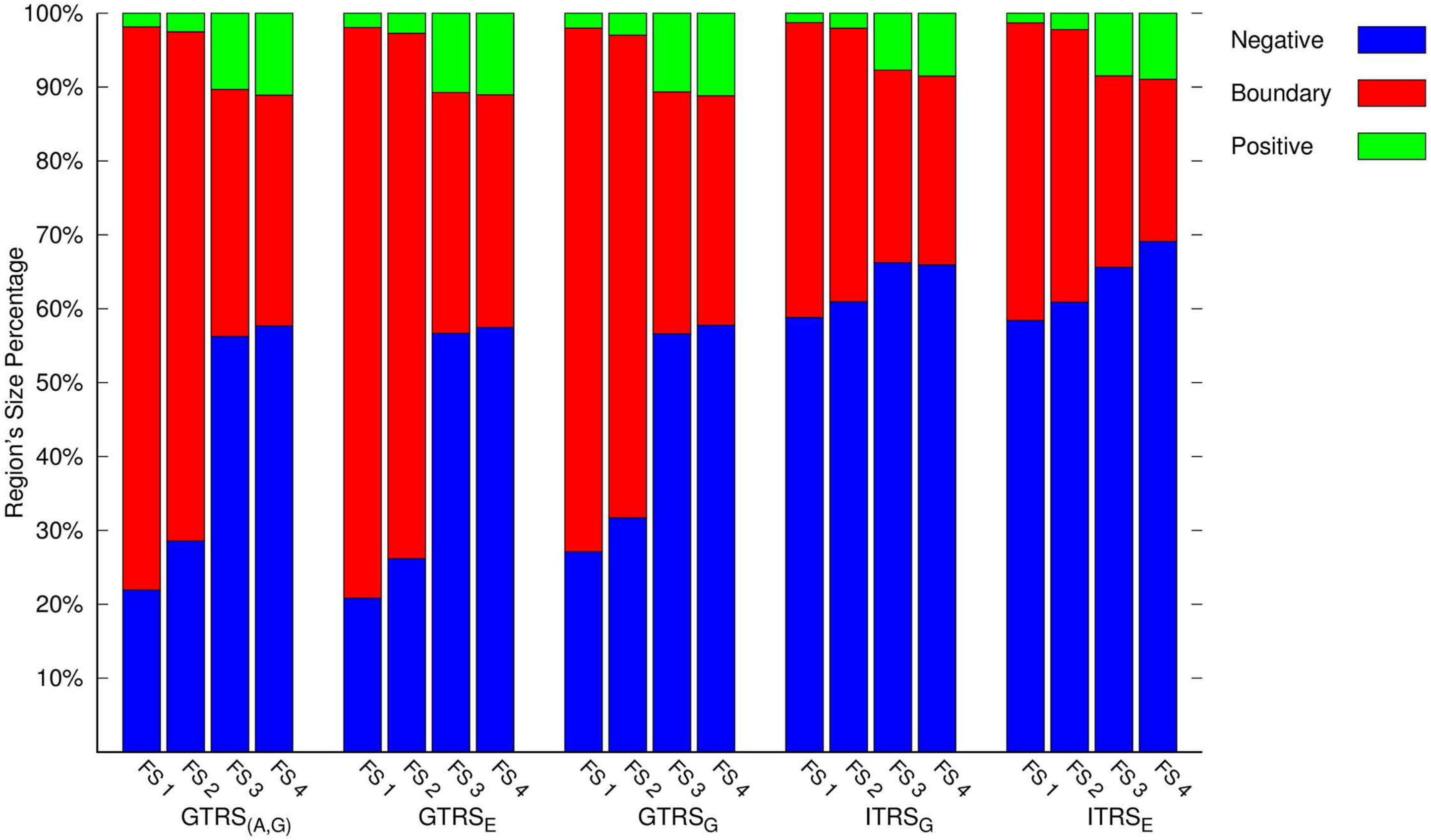

Fig 5. Results of the positive, negative and boundary regions.

doi:10.1371/journal.pone.0171702.g005

ITRS$_G$ approaches are decreased by a small 3.37% and 2.74%, respectively as we increase the feature set size. comparing these results with the GTRS based approaches, we may note that the generality values of the ITRS approaches are significantly better than those obtained with the GTRS based approaches. However, for accuracy there is no significant different between ITRS and GTRS based approaches as both of them are around 80%. Despite some differences in the results with the two approaches, the key observation noted earlier in the discussion of GTRS based results holds for the results in the case of ITRS as well. We may notice again that increasing the number of features or the level of information lead to better generality (which implies more classification decisions) while maintaining the same or similar level of accuracy. In order to highlight this observation, we constructed two figures.

Fig 5 shows the results of the positive, negative and boundary regions based on the GTRS and ITRS based approaches. Each bar in the figure is split into three parts, representing the positive, negative and boundary regions respectively. Each set of four bars corresponds to a particular approach and is separated by a large space. The four bars are placed in increasing order of feature set sizes. In each set of four bars, the leftmost bar corresponds to the least feature set size and the rightmost bar corresponds to the highest feature set size. We may note in Fig 5, that as we increase the feature set sizes, the positive and negative regions grow in size while the boundary regions shrinks. According to the definition of generality in Eq (27), the union of the positive and negative regions represents the generality. This figure highlights the

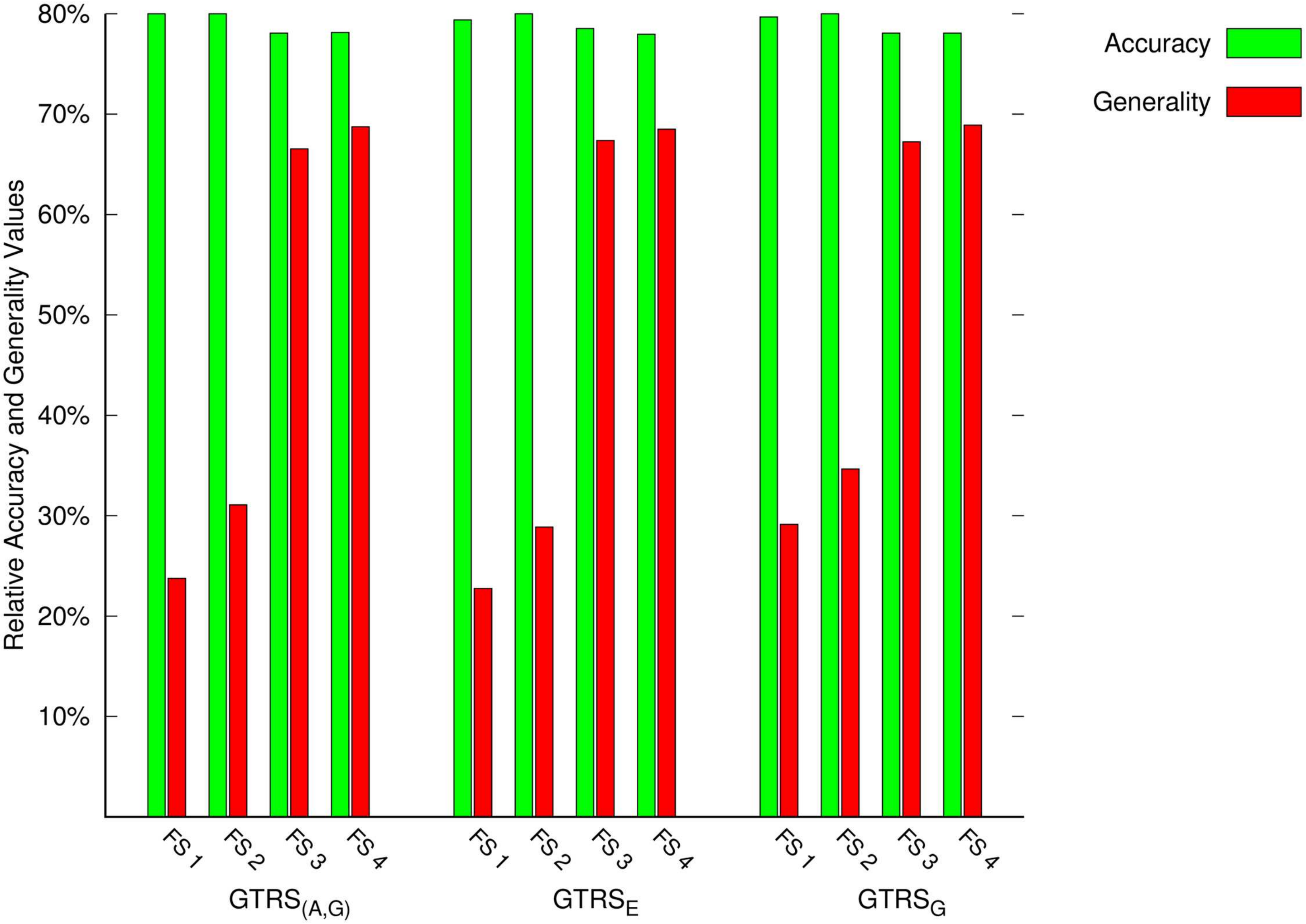

Fig 6. Accuracy and generality results of the GTRS based approaches.

doi:10.1371/journal.pone.0171702.g006

same fact, noted earlier in the previous discussion, i.e., we are able to make more classification decisions for proteins as the level of information increases (or number of features increases). Please be noted that in probabilistic rough sets, it is not always necessary that the addition of features will increase the positive and negative regions. However, we want to emphasize the fact that it will result in improvement in the quality of the regions.

Figs 6 and 7 summarizes the results of accuracy and generality for the considered approaches. The green colour in these figures represent the accuracy and the red colour indicate the generality. The values of accuracy and generality are reported for all four feature sets described in the previous section. It may be noted that for all the approaches, the generality improves as we use higher number of features. However, on the other hand the accuracy is not affected significantly. This means that by increasing the features, we are able to improve the generality while maintaining the similar level of accuracy.

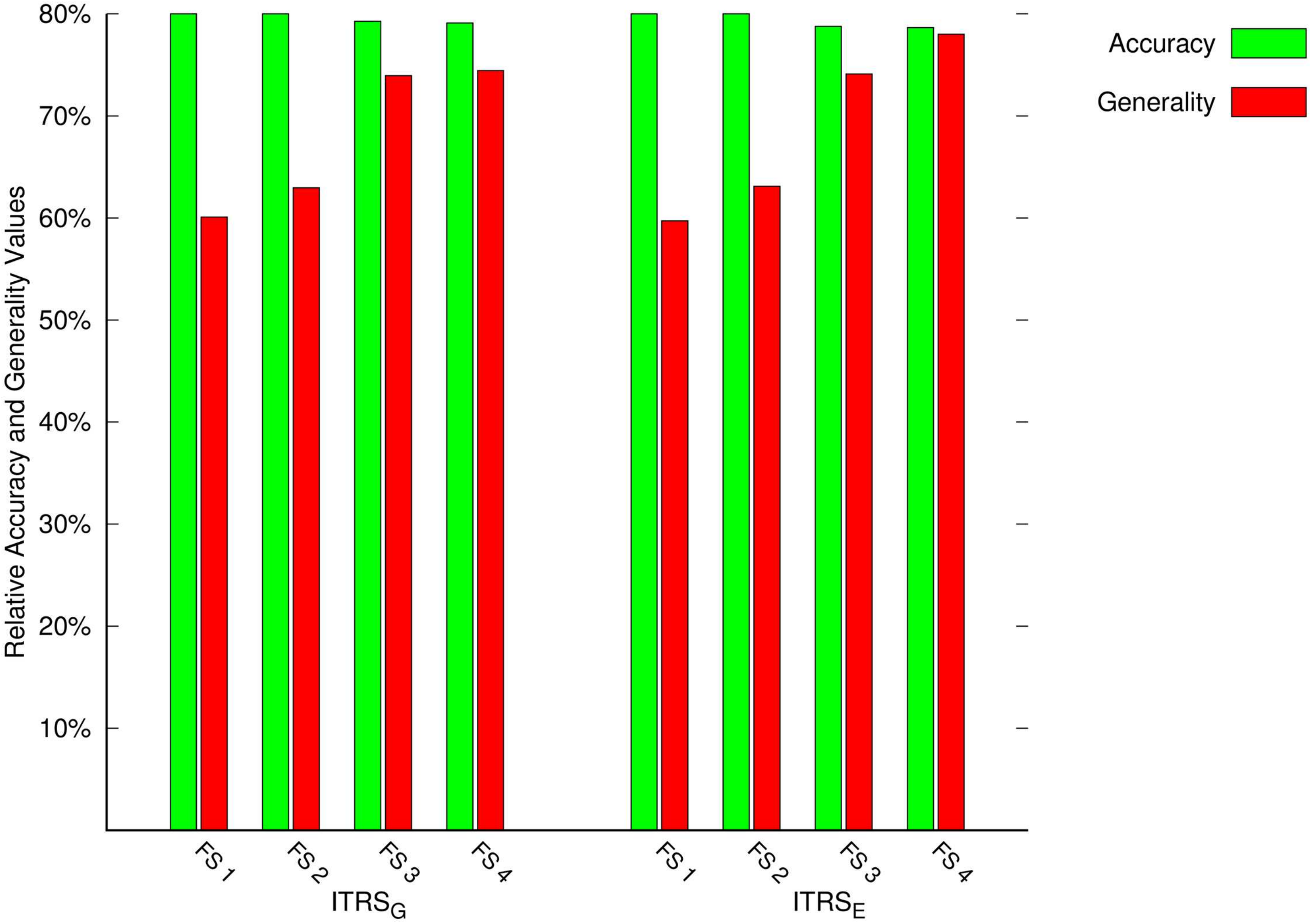

Fig 7. Accuracy and generality results of the ITRS based approaches.

doi:10.1371/journal.pone.0171702.g007

Comparison with other approaches

In this section, we compare our method to the most widely used group of function prediction techniques. The first choice of comparison is a most recent function prediction tool, namely INGA (Interaction Network GO Annotator) [75], which is a state of the art tool to predict protein functions. INGA is based on a consensus strategy that maximizes the F-score by utilizing protein interaction networks, sequence similarity as well as domain information for protein function prediction. Since, information such as protein domains, protein interaction networks, sequence similarity are directly related to protein function and are at the heart of overwhelming majority of methods that integrate it for protein function prediction, we chose to compare our method against this function prediction tool.

To attain the most accurate comparative results, we run our algorithm for wider range of target classes of the molecular function ontology in a ten-fold cross-validation setting. We present the comparative results (i.e., Accuracy and Generality as defined in Eq (26)) of INGA and our method in Table 5 when tried on *Saccharomyces cerevisiae* proteins. Clearly, our

Table 5. Comparison of the proposed three way classification method with top performing methods of the field. The target classes comprise of broader gene ontology terms for *Saccharomyces cerevisiae* species proteins.

Method's Name	Generality	Accuracy
Three way decision using GTRS	68%	78.40%
Three way decision using ITRS	**74**%	**79.2**%
INGA (Interaction Network GO Annotator) tool [75]	60%	57%
Jones-UCL [76]	62%	59.5%
Argot [76]	61%	59.4%
BLAST Annotation Transfer (baseline method) [76]	78%	38%

doi:10.1371/journal.pone.0171702.t005

method outperforms IGNA in both aspects. The strength of our method mainly comes from the fact that it is able to defer instances (i.e., proteins) for which there is less characterization evidence at present, thus improving prediction accuracy.

In order to get further insight into the relative performance, we also consider some results that were reported on similar problem. The most recent and well known schemes proposed for the prediction of protein functions are evaluated in the CAFA (Critical Assessment of protein Function Annotation) challenge [76]. The CAFA challenge is conducted after every two years to have comparative evaluation of top schemes for the prediction of protein functions. One of the best sequence alignment algorithm (i.e., BLAST), which is also used as a baseline scheme for annotation transfer, achieved an accuracy of 38% during the CAFA challenge, when tried on molecular function category of GO [76]. Likewise, the top schemes of the challenge have reported to have an accuracy of 59.5% and 59.4%, respectively [76], when tested against heterogeneous ontology classes. On the other hand, our method have achieved an overall accuracy of 80% when tried for the same target classes, depicting a significant gain in terms of prediction accuracy.

Another method, that is more recently proposed, by Mitrofanova et al. in [8], combines inter-species homology data for protein function prediction and reported an accuracy of 97.7% when tried on *Saccharomyces cerevisiae* proteins. However, the results of this method cannot be directly compared with our approach for a number of reasons. Firstly, this method operates on fixed ontology sizes thus giving results for only 16 GO terms (target classes) out of more than 30,000 GO terms. Secondly, the fixed GO terms chosen by the authors limit pertinence of their method to proteins directly annotated to those GO terms, hence limiting the applicability of their algorithm to only a small number of proteins (hence, significant reduction in generality). On the other hand, our algorithm has much wider GO coverage and results presented include all the yeast proteins.

As a final remark, it is pertinent to mention here that although three way classification achieved far better results than the earlier proposed schemes but the main purpose of this study was not the optimization of performance (in terms of precision or accuracy) of the earlier schemes. But rather it should be looked at from the perspective of an examination, feasibility and appropriateness of considering three way classification schemes based on evolving biological information for the task of protein function predictions.

In conclusion, the three way approaches considered in this study achieve an average accuracy of 80%. The incorporation of future information is useful as it improves the generality or applicability of the models while maintaining similar level of accuracy. In particular, there is an increase in generality by more than 40% for the GTRS based approaches and more than

14% for the ITRS based approaches. From the general trend in the results, it is suggested that as more information becomes available, the generality may improve further. These results advocate for the use of three-way approaches for protein functions classification.

Conclusion

Proteins are involved in almost every biological phenomena and the precise knowledge of their functions plays an essential role in understanding biological processes. Intelligent mechanisms are generally employed to assign and predict functions of proteins. The technological advancements are continuously resulting in new information and features describing protein functions which in turn can be utilized for improving the quality of protein function predictions. An important issue in this context is to develop effective classification schemes and models for classifying protein functions by incorporating evolving information. We propose a three-way decision making approach to address this issue. The approach includes a deferment decision option which is practiced in situations characterized by insufficient and incomplete information. In particular, we considered probabilistic rough sets based models i.e., game-theoretic rough sets and information-theoretic rough sets for inducing and making three-way decisions. An architecture of protein function classification with three-way decisions is also proposed and explained. Experimental results on dataset from *Uniprot* database indicate that as the level of biological information increases, the number of deferred cases are reduced while maintaining similar level of accuracy. In particular, an average accuracy of 80% (±%2) was reported for the considered approaches with an average generality improvement of 33% (±%5) as we increase features.

We investigated the probabilistic rough sets which is one possible way for inducing three-way decisions. Other approaches such as shadowed sets, statistical testing, interval sets and ortho-pairs may also be examined to investigate the potential benefits of three-way approach to protein function classification. Moreover, the three-way approach for protein function classification may further be evaluated and extended by incorporating new features resulting from next generation sequencing data or from other high throughput experiments.

Supporting information

S1 File. "S1_File.zip". The code (Python/Bash/Matlab) and data files along with instructions are provided as a zip file.
(ZIP)

Acknowledgments

The third author was partially supported by Discovery Grant from NSERC Canada. We are also thankful to the Higher Education Commission of Pakistan for providing student grant under the SRGP (Startup Research Grant Program) initiative. There was no additional external funding received for this study.

Author Contributions

Conceptualization: HU NA JY AB.

Formal analysis: HU NA JY AB.

Investigation: HU NA JY AB.

Methodology: HU NA JY AB.

Project administration: HU NA JY AB.

Software: HU NA.

Supervision: JY AB.

Validation: HU NA.

Writing – original draft: HU NA JY AB.

Writing – review & editing: HU NA JY AB.

References

1. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular Biology of the Cell, 5th edition. Anderson M, Granum S, editors. Garland Science; 2008.
2. Benso A, Di Carlo S, Ur Rehman H, Politano G, Savino A, Suravajhala P. A Combined Approach for Genome Wide Protein Function Annotation/Prediction. PROTEOME SCIENCE. 2013;Vol. 11(No. S1):1–12. doi: 10.1186/1477-5956-11-S1-S1
3. Panday G, Kumar V, Steinbach M. Computational Approaches for Protein Function Prediction: A Survey; 2006.
4. Jiang Y, Ronnen Oron T, Ur Rehman H, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biology. 2016; 17(184). doi: 10.1186/s13059-016-1037-6 PMID: 27604469
5. Ur Rehman H, Zafar U, Benso A, Islam N. A Structure based Approach for Accurate Prediction of Protein Interactions Networks. In: Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies. vol. 3. BIOINFORMATICS 2016. ScitePress; 2016. p. 237–244.
6. Benso A, Di Carlo S, Ur Rehman H, Politano G, Savino A, Vasciaveo A. Using Boolean networks to model post-transcriptional regulation in gene regulatory networks. Journal of Computational Science. 2014; 5(3):332–344. doi: 10.1016/j.jocs.2013.10.005
7. Benso A, Di Carlo S, Ur Rehman H, Politano G, Savino A, Vasciaveo A. Accounting for post-transcriptional regulation in boolean networks based regulatory models. International Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO 2013. Copicentro Editorial; 2013. p. 397–404.
8. Mitrofanova A, Pavlovic V, Mishra B. Prediction of Protein Functions with Gene Ontology and Interspecies Protein Homology Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2011; 8 no. 3:775–784. doi: 10.1109/TCBB.2010.15 PMID: 21393654
9. Ur Rehman H, Benso A, Di Carlo S, Politano G, Savino A, Suravajhala P. Combining Homolog and Motif Similarity Data with Gene Ontology Relationships for Protein Function Prediction. In: Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference; 2012. p. 1–4.
10. Zhang H, Chen Z, Liu Z, Zhu Y, Wu C. Location Prediction Based on Transition Probability Matrices Constructing from Sequential Rules for Spatial-Temporal K-Anonymity Dataset. PloS one. 2016; 11(8): e0160629. doi: 10.1371/journal.pone.0160629 PMID: 27508502
11. Stephen F, Gish W, Miller W, Eugene W, David J. Basic local alignment search tool (BLAST). Molecular Biology. 1990; 215:403–410. doi: 10.1016/S0022-2836(05)80360-2
12. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. Science. 1999; 285:751–753. doi: 10.1126/science.285.5428.751 PMID: 10427000
13. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences of the United States of America. 1988; 85(8):2444–8. doi: 10.1073/pnas.85.8.2444 PMID: 3162770
14. Watson J, Laskowski R, Thornton J. Predicting protein function from sequence and structural data. Current Opinion in Structural Biology. 2005; 15:275–284. doi: 10.1016/j.sbi.2005.04.003 PMID: 15963890
15. Gaudet P, Livstone MS, Lewis SE, Thomas PD. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Briefings in Bioinformatics. 2011; 12:449–462. doi: 10.1093/bib/bbr042 PMID: 21873635
16. Pellegrini M, Marcotte E, Thompson M, D E, TO Y. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proceedings of the National Academy of Sciences of the United States of America. 1999; p. 4285–4288.

17. Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of protein function using protein-protein interaction data. Journal of Computational Biology. 2003; 10:947–960. doi: 10.1089/106652703322756168 PMID: 14980019

18. Letovsky S, Kasif S. Predicting protein function from protein-protein interaction data: a probabilistic approach. Bioinformatics. 2003; 19(suppl. 1):i197–i204. doi: 10.1093/bioinformatics/btg1026 PMID: 12855458

19. Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics. 2005; 21((suppl. 1)):i302–i310. doi: 10.1093/bioinformatics/bti1054 PMID: 15961472

20. Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein-protein interaction networks. Nature Biotechnology. 2003; 21:697–700. doi: 10.1038/nbt825 PMID: 12740586

21. Laskowski RA, Watson JD, Thornton JM. Protein function prediction using local 3D templates. Journal of Molecular Biology. 2005; 351:614–626. doi: 10.1016/j.jmb.2005.05.067 PMID: 16019027

22. Pal D, Eisenberg D. Inference of protein function from protein structure. Structure. 2005; 13:121–130. doi: 10.1016/j.str.2004.10.015 PMID: 15642267

23. Pazos F, Sternberg MJ. Automated prediction of protein function and detection of functional sites from structure. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101:14754–14759. doi: 10.1073/pnas.0404569101 PMID: 15456910

24. Yao YY. Probabilistic rough set approximations. International Journal of Approximate Reasoning. 2008; 49(2):255–271. doi: 10.1016/j.ijar.2007.05.019

25. Azam N, Yao JT. Analyzing Uncertainties of Probabilistic Rough Set Regions with Game-theoretic Rough Sets. International journal of approximate reasoning. 2014; 55(1):142–155. doi: 10.1016/j.ijar.2013.03.015

26. Herbert JP, Yao JT. Game-theoretic Rough Sets. Fundamenta Informaticae. 2011; 108(3-4):267–286.

27. Yao JT, Herbert JP. A Game-theoretic Perspective on Rough Set Analysis. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition). 2008; 20(3):291–298.

28. Deng XF, Yao YY. A Multifaceted Analysis of Probabilistic Three-way Decisions. Fundamenta Informaticae. 2014; 132:291–313.

29. The UniProt Consortium. UniProt: a hub for protein information; 2015. Nucleic Acids Res. 43: D204–D212. doi: 10.1093/nar/gku989 PMID: 25348405

30. GO. The Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Research. 2015; 43(Database Issue):D1049–D1056. doi: 10.1093/nar/gku1179 PMID: 25428369

31. Francois E, Karsten S, Claverie J. Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. BMC Bioinformatics. 2005; 6(1):247. doi: 10.1186/1471-2105-6-247

32. Pellegrini M, Marcotte E, Thompson M, Eisenberg D, Yeates T. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proceedings of the National Academy of Sciences. 1999; 96(8):4285–4288. doi: 10.1073/pnas.96.8.4285

33. The NCBI handbook [Internet] Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. Chapter 18, The Reference Sequence (RefSeq) Project.; 2002.

34. Fields S, Song O. A Novel Genetic System to Detect Protein-Protein Interactions. Nature. 1989; 340:245–246. doi: 10.1038/340245a0 PMID: 2547163

35. Andreas Bauer BK. Affinity Purification-Mass Spectrometry. European Journal of Biochemistry. 2003; 270:570–578. doi: 10.1046/j.1432-1033.2003.03428.x PMID: 12581197

36. Ur Rehman H, Benso A, Di Carlo S, Politano G, Savino A, Suravajhala P. Using gnome wide data for protein function prediction by exploiting gene ontology relationships. IEEE Press; 2012. p. 497–502.

37. Liang DC, Liu D, Kobina A. Three-way group decisions with decision-theoretic rough sets. Information Sciences. 2016; 345:46–64. doi: 10.1016/j.ins.2016.01.065

38. Peters J, Ramanna S. Proximal three-way decisions: theory and applications in social networks. Knowledge-Based Systems. 2016; 91:4–15. doi: 10.1016/j.knosys.2015.07.021

39. Yao YY. Rough Sets and Three-Way Decisions. In: Proceedings of 10th International Conference on Rough Sets and Knowledge Technology (RSKT'15), Lecture Notes in Computer Science 9436; 2015. p. 62–73.

40. Baram Y. Partial Classification: The Benefit of Deferred Decision. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998; 20(8):769–776. doi: 10.1109/34.709564

41. Goudey R. Do statistical inferences allowing three alternative decisions give better feedback for environmentally precautionary decision-making? Journal of Environmental Management. 2007; 85 (2):338–344. doi: 10.1016/j.jenvman.2006.10.012 PMID: 17129664

42. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. The New England Journal of Medicine. 1980; 302(20):1109–1117. doi: 10.1056/NEJM198005153022003 PMID: 7366635

43. Sherif M, Hovland CI. Social judgment: Assimilation and contrast effects in communication and attitude change. 1961;.

44. Tversky A, Shafir E. Choice under conflict: The dynamics of deferred decision. Psychological science. 1992; 3(6):358–361. doi: 10.1111/j.1467-9280.1992.tb00047.x

45. Yao YY. An Outline of a Theory of Three-way Decisions. In: Proceedings of Rough Sets and Current Trends in Computing (RSCTC'12), Lecture Notes in Computer Science 7413; 2012. p. 1–17.

46. Yao JT, Zhang Y. A scientometrics study of rough sets in three decades. In: Proceedings of 8th International Conference on Rough Sets and Knowledge Technology (RSKT'13), Lecture Notes in Computer Science 8171; 2013. p. 28–40.

47. Deng XF, Yao YY. Decision-theoretic three-way approximations of fuzzy sets. Information Sciences. 2014; 279:702–715. doi: 10.1016/j.ins.2014.04.022

48. Liang DC, Liu D, Pedrycz W, Hu P. Triangular fuzzy decision-theoretic rough sets. International Journal of Approximate Reasoning. 2013; 54(8):1087–1106. doi: 10.1016/j.ijar.2013.03.014

49. Liang DC, Liu D. Systematic studies on three-way decisions with interval-valued decision-theoretic rough sets. Information Sciences. 2014; 276:186–203. doi: 10.1016/j.ins.2014.02.054

50. Liang DC, Xu Z, Liu D. Three-way decisions with intuitionistic fuzzy decision-theoretic rough sets based on point operators. Information Sciences. 2017; 375:183–201. doi: 10.1016/j.ins.2016.09.039

51. Liu D, Li TR, Li HX. Interval-valued decision-theoretic rough sets. Computer Science. 2012; 39(7):178–181.

52. Pedrycz W. Shadowed sets: representing and processing fuzzy sets. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics. 1998; 28(1):103–109. doi: 10.1109/3477.658584

53. Wald A. Sequential tests of statistical hypotheses. The Annals of Mathematical Statistics. 1945; 16 (2):117–186. doi: 10.1214/aoms/1177731118

54. Pawlak Z. Rough Sets. International Journal of Computer and Information Sciences. 1982; 11:241–256. doi: 10.1007/BF01001956

55. Pawlak Z. Rough sets: theoretical aspects of reasoning about data. Kluwer Academic; 1991.

56. Yao YY, Greco S, Slowinski R. Probabilistic rough sets. In: Handbook of Computational Intelligence, Projektorganisation und Management im Software Engineering; 2015. p. 315–339.

57. Yao YY, Wong SKM, Lingrass P. A decision-theoretic rough set model. Methodologies for Intelligent Systems. 1990; 35:17–24.

58. Yao YY. Two Semantic Issues in a Probabilistic Rough Set Model. Fundamenta Informaticae. 2011; 108(3-4):249–265.

59. Deng XF, Yao YY. An Information-theoretic interpretation of thresholds in probabilistic rough sets. In: Proceedings of Rough Sets and Current Trends in Computing (RSCTC'12), Lecture Notes in Computer Science 7413; 2012. p. 232–241.

60. Azam N, Yao JT. Game-theoretic rough sets for recommender systems. Knowledge-Based Systems. 2014; 72:96–107. doi: 10.1016/j.knosys.2014.08.030

61. Leyton-Brown K, Shoham Y. Essentials of Game Theory: A Concise Multidisciplinary Introduction. Morgan & Claypool Publishers; 2008.

62. Zhang Y. Optimizing Gini coefficient of probabilistic rough set regions using Game-Theoretic Rough Sets. In: Proceedings of 26th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE'13); 2013. p. 699–702.

63. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, et al. Functional profiling of the Saccharomyces cerevisiae genome. nature. 2002; 418(6896):387–391. doi: 10.1038/nature00935 PMID: 12140549

64. DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM. Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic acids research. 2013; p. gkt135.

65. Kerrien S, et al. The IntAct molecular interaction database in 2012. Nucleic Acids Research. 2012; 40 (Database issue):D841–D846. doi: 10.1093/nar/gkr1088 PMID: 22121220

66. Jensen L, Kuhn M, Stark M, et al. STRING 8–a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Research. 2009; 37(Database Issue):D412–6. doi: 10.1093/nar/gkn760 PMID: 18940858

67. Mulder NJ, et al. New developments in the InterPro database. Nucleic Acids Research. 2007; 35:D224–D228. doi: 10.1093/nar/gkl841 PMID: 17202162

68. Murzin AG, et al. SCOP: a structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology. 1995; 247:536–540. doi: 10.1016/S0022-2836(05)80134-2 PMID: 7723011

69. Mitchell A, Chang HY, Daugherty L, et al. InterPro protein families database: the classification resource after 15 years. Nucleic Acids Research. 2015; 43(Database Issue):D213–21. doi: 10.1093/nar/gku1243 PMID: 25428371

70. Hulo N, Bairoch A, Bulliard V, Cerutti L, et al. The PROSITE Database. Nucleic Acids Research. 2006; 34:D227–230. doi: 10.1093/nar/gkj063 PMID: 16381852

71. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Databank. Nucleic Acids Research. 2000; 28:235–242. doi: 10.1093/nar/28.1.235 PMID: 10592235

72. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. Journal of Molecular Biology. 2007; 372:774–797. doi: 10.1016/j.jmb.2007.05.022 PMID: 17681537

73. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallographica Section D. 2004; D60:2256–2268. doi: 10.1107/S0907444904026460 PMID: 15572779

74. Yao JT, Azam N. Three-way Decision Making in Web-based Medical Decision Support Systems with Game-theoretic Rough Sets. IEEE Transactions on Fuzzy Systems. 2014; 23(1):3–15. doi: 10.1109/TFUZZ.2014.2360548

75. Piovesan D, Giollo M, Leonardi E, Ferrari C, C E Tosatto S. INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity. Nucleic acids research. 2015; 43(W1):W134–40. doi: 10.1093/nar/gkv523 PMID: 26019177

76. Radivojac P, Clark WT, Oron TR, Schnoes, et al. A large-scale evaluation of computational protein function prediction. Nature methods. 2013; 10(3):221–227. doi: 10.1038/nmeth.2340 PMID: 23353650