Original Research Article   https://doi.org/10.20546/ijcmas.2018.704.358

Microbial Taxonomy Ontology for Agriculturally Important Microorganisms (AMO) Coupled with Sequence Alignment Reinforcement Options

Chandan Kumar Deb[1*], Saket Kumar Karn[1], Madhurima Das[2] and Sudeep Marwaha[1]

[1]*Indian Agricultural Statistics Research Institute, New Delhi-110012, India*
[2]*Indian Agricultural Research Institute, New Delhi-110012, India*

*Corresponding author

A B S T R A C T

Ontology is a knowledge representation technique, devised for the web based systems to provide the capability to deal with the semantics of the concepts in the specific knowledge domain. Alternatively, taxonomy describes the real world concepts in a well-defined hierarchy and exists in standard form for various domains in science. The present study dealt with the taxonomy of microorganisms. The Three Domain System taxonomy is most widely adopted taxonomy in this domain. It covers Bacteria, Archaea and Eukarya domains. In this research work a web based application has been developed using N-tier architecture which extended the previously developed Microbial Ontology covering Archaea domain up to the species level. Developed application easily identified new microorganisms by matching their characteristics. Domain experts can insert, delete and edit any new information about the microbial taxonomy. The web interface also provided search facility for finding information about the concepts and 16S rRNA sequences of various Archaea species. This software also facilitated name based search for microorganism's taxonomic terms. A sequence alignment tool is also developed in the system for aligning the query sequence with the existing sequence in the ontology. The use of ontologies to represent the taxonomic information and the ability of this software to provide this knowledge to other applications increases the utility of this work to a greater extent.

Introduction

Microbes, coherently indispensable for agriculture and crop productivity; apart from the catastrophic damage it results. Proper utilization of the microbe can only be achievable through its explicit knowledge of domains and capability of drawing inference from them for better utilization of that knowledge. It is only feasible through an efficient knowledge representation technique – Ontology. Ontology is used in agriculture in various ways like Gene Ontology (GO): Gene Ontology (GO) was developed by Gene Ontology Consortium (Ashburner *et al.,* 2000). AmiGO is an HTML based browser, which one can use to browse and search Gene Ontology (GO). Gene Ontology covers three domains Molecular Function, Biological Process and Cellular Component. Plant

Ontology (PO): Plant Ontology (PO) was developed by Plant Ontology Consortium, 2002 It deals with plant genome databases and plant systematics to describe phenotype and expression patterns of plant genes. Designing Ontology from Traditional Taxonomies (Bedi and Marwaha, 2004): proposed a methodology for the conversion of taxonomies to ontologies. The proposed methodology is tested and implemented for a pilot soil ontology using the IEEE standard Web Ontology Language (OWL) and protégé 2.1 OWL plugin. Ontology-based intelligent retrieval system for soil knowledge (Minz *et al.,* 2009): This system search the documents related to soils by using soil domain ontology. Classification information in soil domain ontology is displayed in a tree structure form, from the navigation database Building and Querying Soil Ontology for Agriculture (Das *et al.,* 2012). This deals with various aspects of development of web based software for the information regarding USDA Soil Taxonomy. This system describes only seven soil orders (Alfisols, Aridisols, Entisols, Inceptisols, Mollisols, Ultisols and Vertisols) seen in India. One can classify the newly found soil according to the USDA Soil Taxonomic Classification system up to Subgroup level (Deb *et al.,* 2015). It was the enhancement of work done by Das, 2010. It was extended up to the soil series level of existing 7 soil order and adding 5 soil order in to the soil ontology. It also provides the query interface for adding, deleting and updating information to the soil ontology. Ontology also facilitates sustainable agriculture techniques. Building and Querying Microbial Ontology (Biswas *et al.,* 2013) deals with various aspects of developing a web based software for the information regarding Three Domain System classification of microbial taxonomy for the microbes important in agricultural purpose. This system contains information mainly about the microorganisms (Bergey *et al.,* 1989) that are important in agriculture. This system contains information from Domain to Genus specifically.

In this work, an attempt has been made to conceptualize and develop ontology for agriculturally important microorganisms (Madigan *et al.,* 2006). Microbial Taxonomy mainly comprises of three parts: Classification, Nomenclature and Identification. Taxonomy can be defined as the science of classification, consisting of two parts: identification and nomenclature. 16S rRNA sequence data is an identifiable characteristic of Archaea. Microbial Ontology contains various classes, properties, restrictions and individuals related to Basic Characteristics, Ecology, Cell Structure, mode of respiration, type of nutrition, shape, Gram Staining etc. In this work the ontology is extended for Archaea from Domain to Species level.

The present study is proposed to extend the work carried out by Biswas, 2012 for the Archaea Domain. The extended system also aims to store and establish relationship between corresponding Archaea microorganism's upto Species level and its 16S rRNA sequence.

This research work includes three objectives: firstly, to perform requirement analysis for strengthening and enhancing microbial ontology, secondly, to develop and populate the microbial ontology, and thirdly, to develop a query interface for querying the ontology.

Materials and Methods

Software development

Tools and technique used to develop microbial taxonomy ontology

Microbial Taxonomy Ontology is a web based software which follows the N-tier architecture.

Figure 1 describes the block diagram of the software. The client side interface layer (CSIL) is in front layer, made to communicate with the user i.e. to take the user query and respond to it. The CSIL layer is made up of HTML, CSS and Java scripts. The server side application layer (SSAL) is made up of java server pages (JSP) and build up on J2EE platform. The SSAL layer handles the user query and process it to get the information from the back end of the software. The back end is made up with database layer (DBL) and knowledgebase layer (KBL). DBL is built up by the RDBMS (Relational Database Management System) SQL server 2008 and on the other hand, KBL is built up of protégé which follows the standard OWL (Web Ontology Language). KBL also enabled to deal with OWL Lite, OWL DL and OWL Full. KBL and semantic web framework layer (SWFL) made the system semantically enabled and it can handle the complex semantic query and decision making hurdle. SWL consists of JENA; a programming framework to handle Resource Description Framework (RDF), Resource Description Framework Schema (RDFS) and Web Ontology Language (OWL). It contains the implementation of SPARQL specifications. SPARQL (Clark, 2008) is a query language which obtains information from RDF graph. Jena is used to store and retrieving data information from Ontology. Additionally this layer uses OWL Protégé, OWL syntax etc. Java API is used to edit the Ontology through Java. The sequence alignment in this software is done by integration of BioJava in the system.

Sequence diagram of microbial taxonomy ontology

To develop a software, the designing and visualization of the output is a very important aspect. In accordance with this, the most important tool to visualize the output is the sequence diagram. We have designed a sequence diagram, to visualize the step by step output and also the interaction with the software (Fig. 2).

Results and Discussion

The result of our study can be divided into two sections. Firstly, we have developed a back end of our web base software and secondly, we have developed a front end to extract, manipulate and process the stored information in the back end. In the ontology development process, we have used Protégé OWL editor from Domain to Species level and a query interface has been developed that will help a detailed study of classification of microorganisms, microbial taxonomy. In this research work, we have enhanced the Microbial Ontology, developed by Biswas *et al.,* 2013. The existing ontology was populated with the information of bacteria up to the genus level. The Microbial Ontology has been extended to the Species for bacteria and also added the information of Archea up to the species level (Domain → Phylum → Class → Order→ Family → Genus → Species).

Creating classes, individuals and their properties

The building block of ontology development is the classes, individuals and the properties of the domains. Figure 3 depicts some snaps of the ontology class which has been developed in the Microbial Taxonomy Ontology.

In the hierarchy, the Class Microbial Taxonomy is created as topmost class. Therefore it is created as the subclass of the class owl: Thing. The class Microbial Taxonomy has three subclasses-Archaea, Bacteria and Eukarya. Two Phylum classes Crenarchaeota and Euryarchaeota (as given in the Bergey's Manual of Systematic

Bacteriology) created as child classes of the Archaea class. Then child classes of both the phylum classes were created. Likewise all the hierarchical data has been incorporated to the ontology (e.g. Phylum-Crenarchaeota has one subclass Class-*Thermoprotei*) and class Thermoprotei have three subclasses (i.e. Order-*Desulfurococcales, Sulfobales* and *Thermoproteales*). After creating the hierarchy class of both the phylum, the classes representing the properties of the microbes such as Basic_Characteristics, Nutrition_type, Other_Characteristics, and Shape were created.

After creating the class hierarchy of the Microbial Taxonomy, the next step is to populate the classes with their respective individuals. Individuals are the instance of classes. In protégé individuals are created in the INDIVIDUAL EDITOR. Individuals of all the subclasses of Microbial Taxonomy must be with the same name of their respective classes, and names have been written in small letters. In Ontology, more than one resource cannot exist with same name.

After creation of classes and their individuals the next step is creation of Properties in Ontology (Figure 5 and 6). In protégé OWL plug-in, Property Browser is used for creating properties. Properties are of two types- Object and Data type. For each property, Domain and Range have to be specific and clearly defined. Domain is the class where the property has to be applied and Range is the class from which the property values have been taken; e.g. the property has16SrRNAsequenceis an object property and its Domain is Archaea and Range is Basic_Characteristics.

Querying the microbial ontology

For retrieving the knowledge from the ontology, Protégé provides a query interface known as Open SPARQL; a query panel where one can write query to find out particular knowledge from the Ontology.

A query interface for querying the Microbial Ontology has been developed. This information is used for extraction from the OWL Ontology layer. Framework layer is implemented by using Jena. The system authorizes three types of users viz. Normal User, Domain Experts (the user has detail knowledge about Microbial Taxonomy) and Administrator. Domain Experts are those users who can insert, delete and update knowledge in the knowledge base. Administrator is the owner of the system and has privilege to add / delete / modify the rights of various users. After verification of sign in all tabs (options) are available to normal users, except "Edit Ontology". "Edit Ontology" option is available only to Domain Experts and Administrator. Home Page of the Software is as shown in Figure 7. Figure 8 depicts the steps involved in the detailed study of Microbial Taxonomy with the help of "Taxonomy" tap after log in of every user.

By using this tab, user can study in detail about microbial taxonomy of Bacteria and Archaea up to the Species level, for ten Phyla, twenty two Classes, thirty six Orders, fifty five Families, one hundred thirty five Genera, and sixty six Species as given in the "Brocks Biology of Microorganisms" and "Bergey's Manual of Systematic Bacteriology". It will show basic characteristics, cell structure, ecology, shape, nutrition type, respiration mode, 16S rRNA sequence etc.

This software also facilitates name based search for all the Microbial Taxonomic terms as shown in Figure 9. This is the result of the search term thermoprotus. Software gives all the hierarchy from the domain up to the genus where the term thermoprotus resides. All red ellipse is clickable and it navigates to the corresponding page of the term.
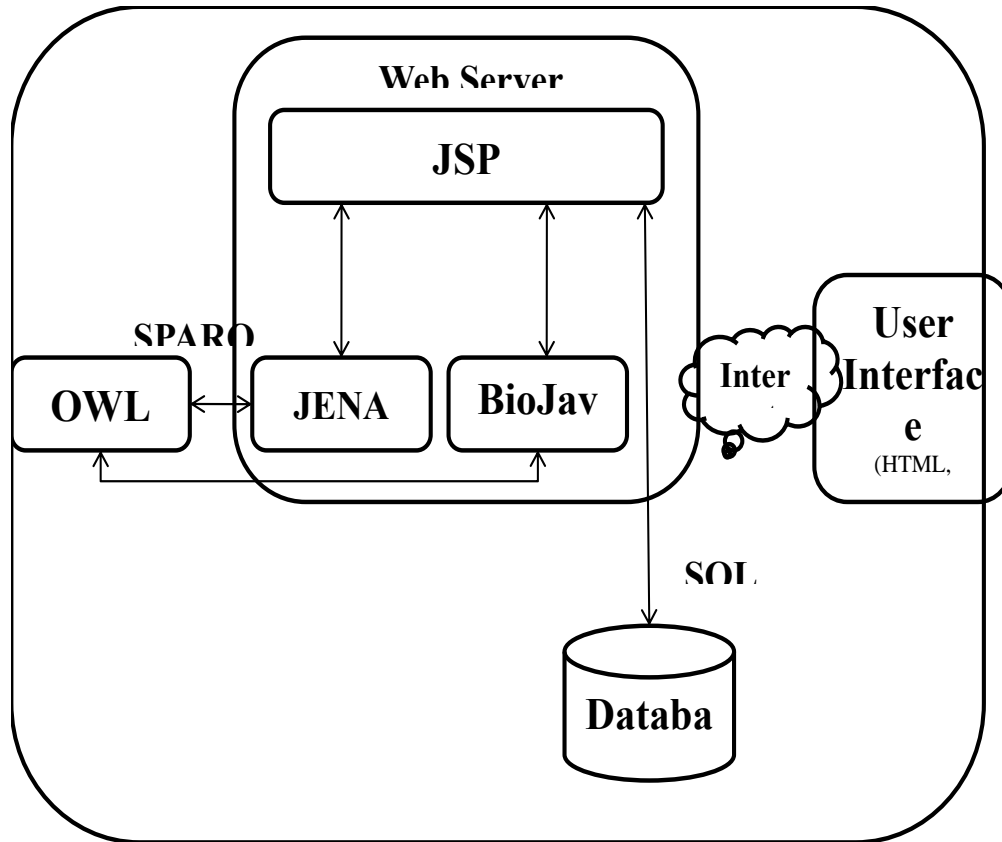
Fig.1 Block diagram representation of the software

Fig.2 Representation of the Sequence diagram of Microbial Taxonomy Ontology
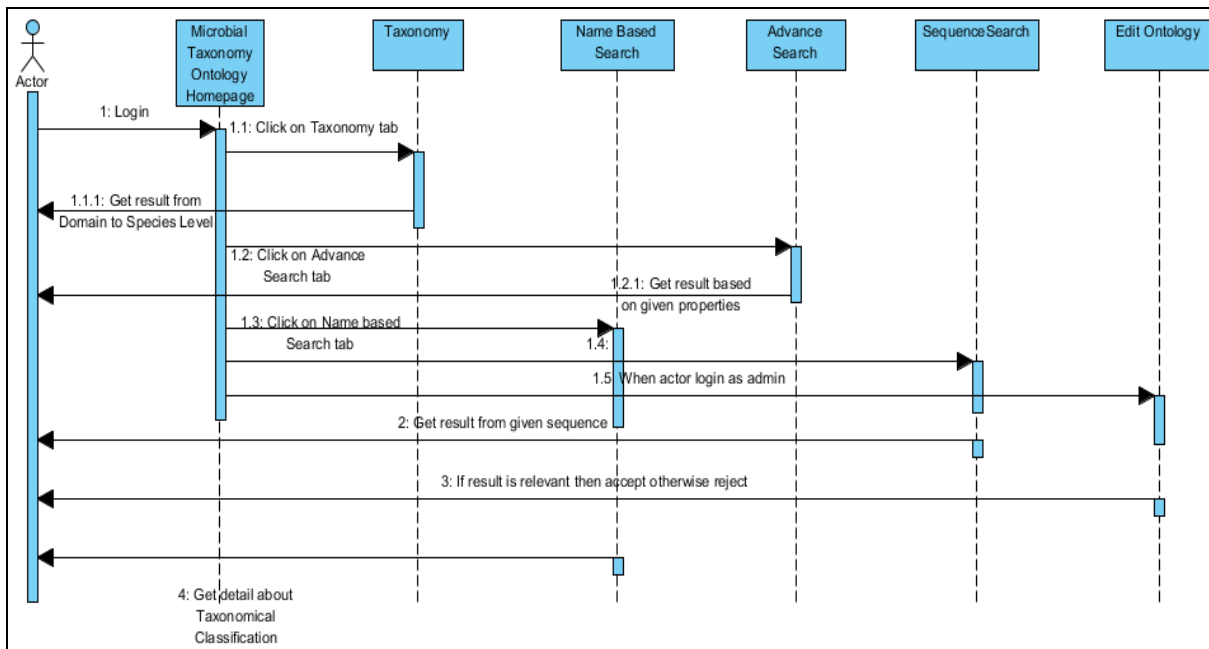
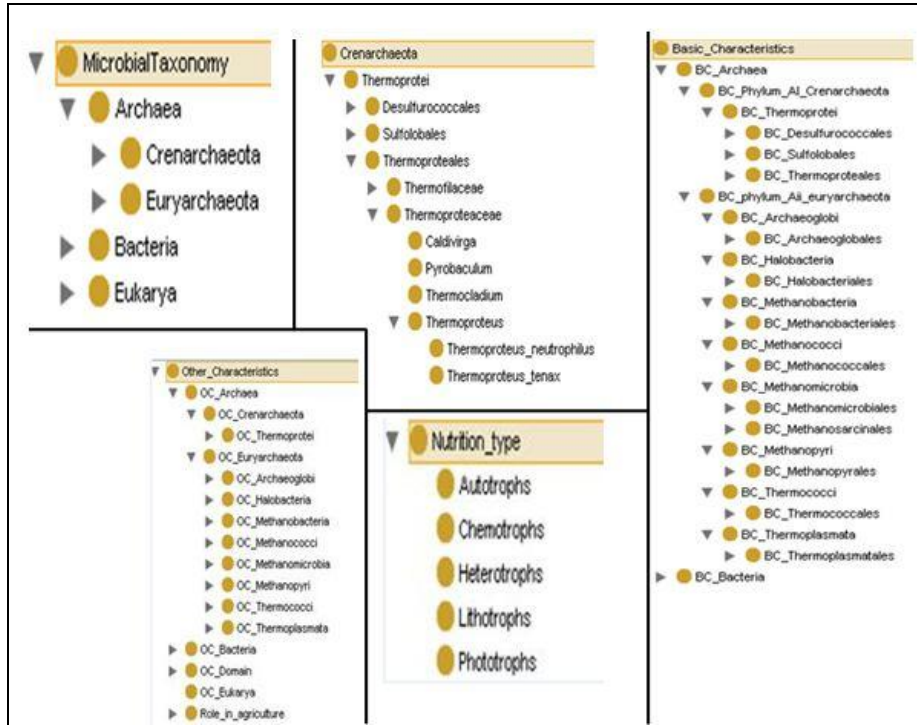**Fig.3** Representation of the entire class hierarchy of Microbial Taxonomy

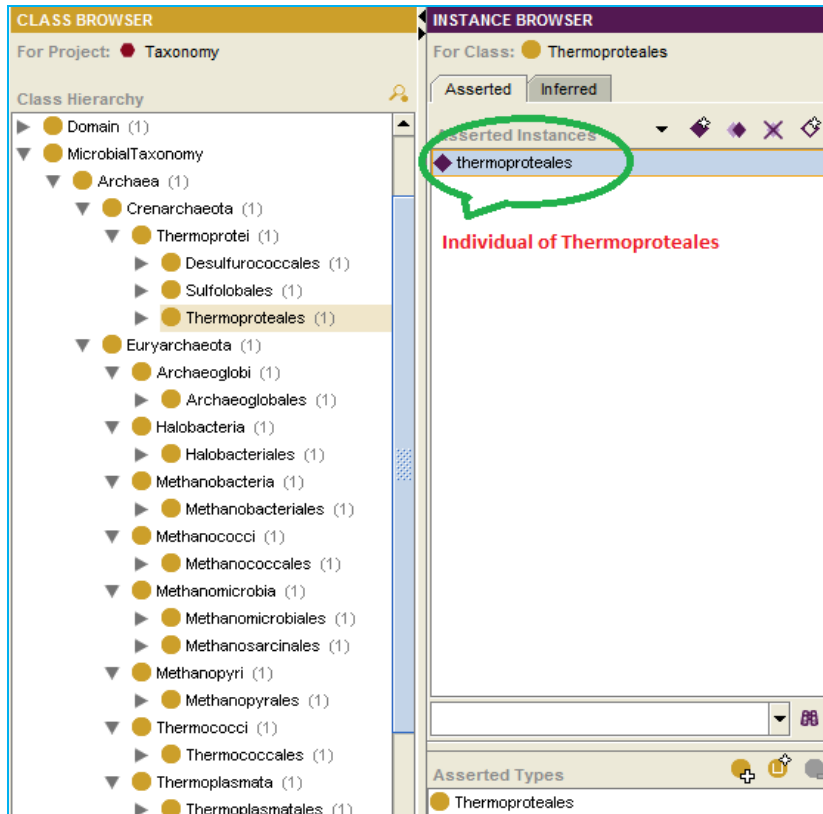**Fig.4** Representation of the individuals of a class

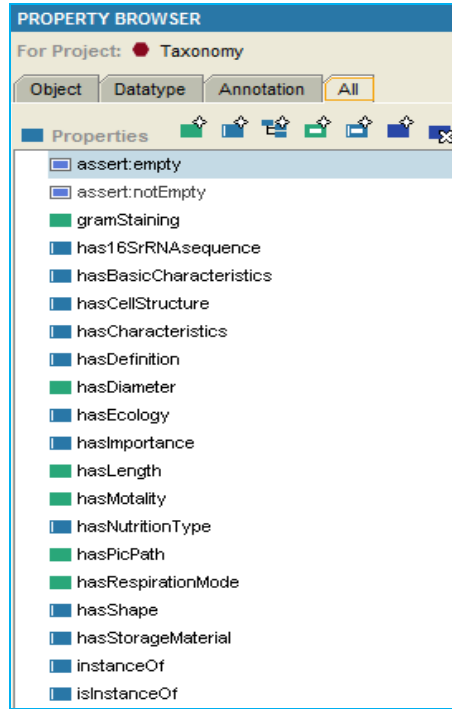**Fig.5** Representation of the list of all the properties of microbial ontology

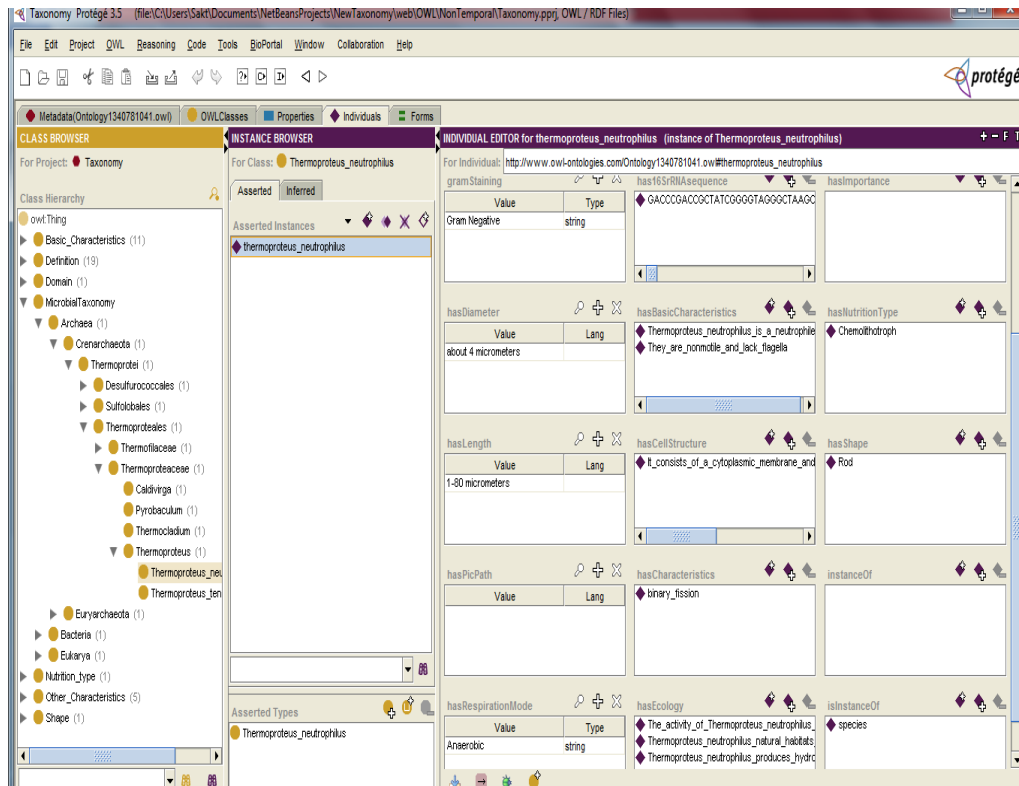**Fig.6** Representation of *Thermoproteus neutrophilus* class with its individuals and their properties

Fig.7 Representation of the home page of the software
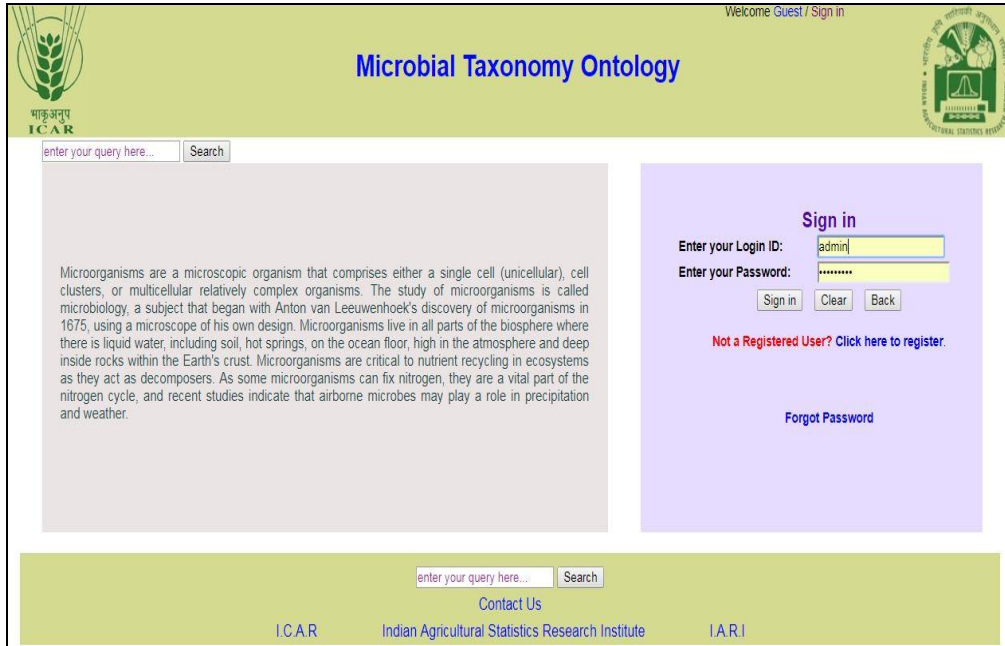
Fig.8 Representation of a detailed study of microbial taxonomy
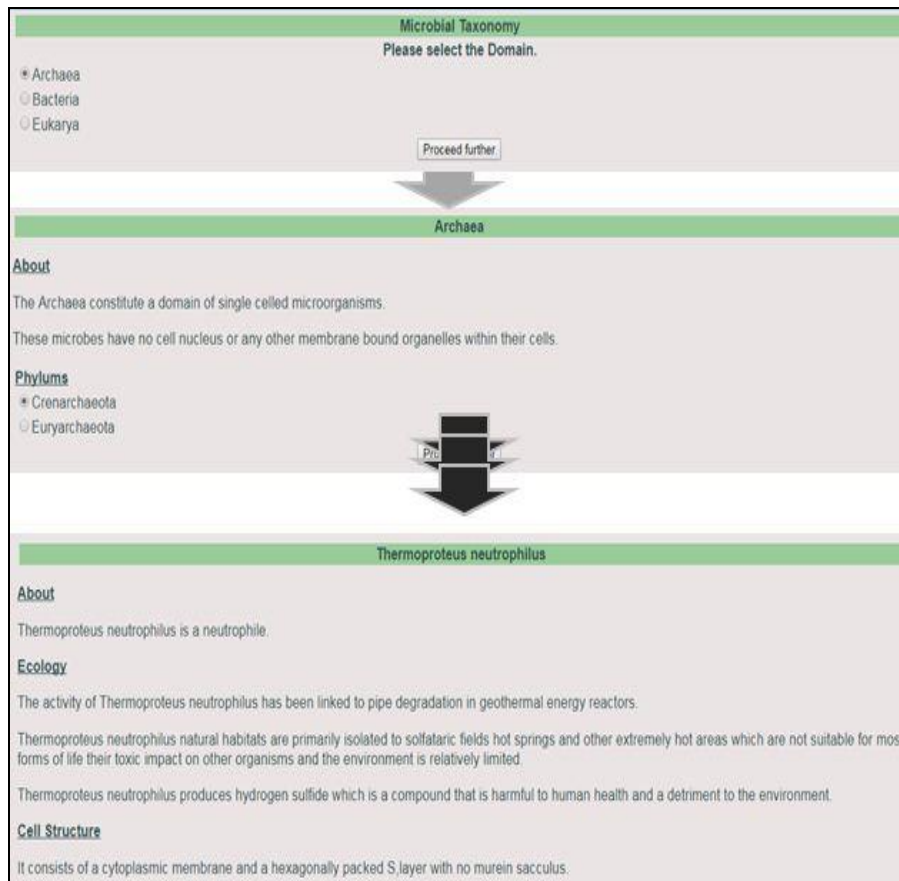
Fig.9 Representation of the results of name based search

Fig.10 Representation of an advanced search module

Fig.11 Representation of sequence search module
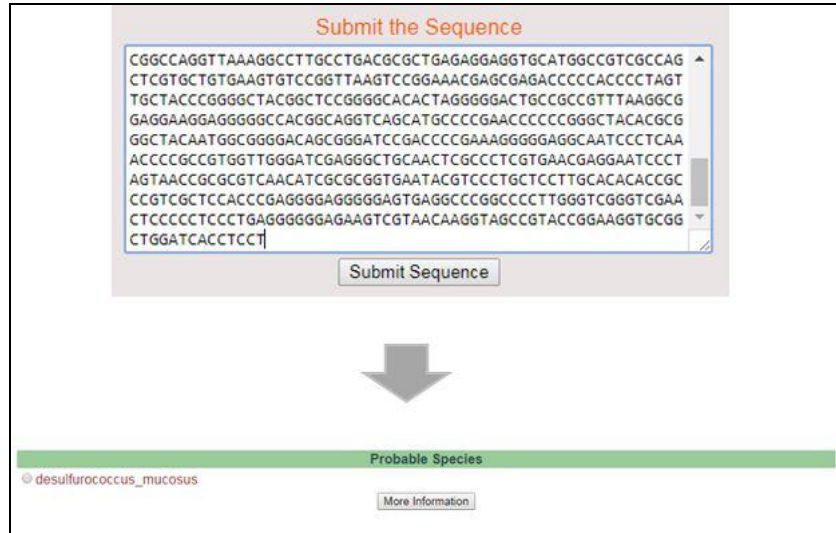
Fig.12 Representation of alignment of the query sequence with the existing sequence and the identification of the species as Archaea
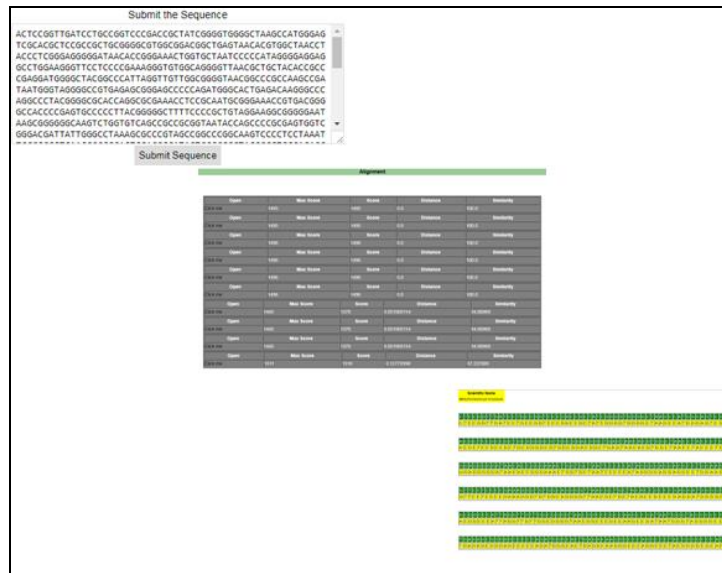
Table.1 Table showing some of the query results of the Domain

	Functional Attributes	Cell Structure Attributes	Results
1.	Nitrogen Fixation	Cell nucleus not present	
	Organic Matter Decomposition	Histone protein present	Archaea
		Ribosome sedimentation value is 70S	
2.	Nitrogen Fixation	Cell nucleus not present	
	Organic Matter Decomposition	Histone protein not present	Bacteria
		Ribosome sedimentation value is 70S	

Table.2 Table showing some of the query results of the Phylum

Sl No.	Functional Attributes	Shape	Other Characteristics	Results
1.	Ammonia Oxidizing Archaea	Coccus, Rod, Filamentous	Anaerobic thermophilic and fermentative	Crenarchaeota
	Photosynthesis		Gram negative	
2.	Halophilic archaea	Bacillus, Coccus, Disc, Filamentous, Rod	All major nutritional types	Euryarchaeota
	Thermophilic archaea		These are mainly halophiles and methanogens	
3.	Bio remediation	Bacillus	Low G C DNA composition	Proteobacteria
	Acetic acid bacteria		Gram negative	

Table.3 Table showing some of the query results of the Family

Sl No.	Functional Attributes	Shape	Other Characteristics	Results
1.	Ammonia Oxidizing Archaea	Rod	Cells are gram positive	Thermoproteaceae
	Nitrogen fixation		Nutrition type Chemolithotrophic	
2.	Nitrogen fixation	Spiral	Nitrite oxidizing bacteria	Nitrospiraceae
	Chemolithotrophic		Important for healthy marine ecosystems	

Table.4 Table showing some of the query results of Genus

Sl No.	Functional Attributes	Shape	Other Characteristics	Results
1.	Sulfer reduction	Rod	Chemoorganotrophs	Thermoproteus
	Thermophilic archaea		Gram negative cells	
			Rod shaped	
2.	Photosynthesis	Coccus	Anaerobic respiration	Pyrococcus
	Thermophilic archaea		Gram negative cells	
			Photosynthetically helpful	
3.	Chemoorganotrophic	Filamentous, Spherical	Aerobic respiration	Thermoplasma
	Thermophilic archaea		Gram negative cells	
			Unicellular organism	
4.	Green sulfur bacteria	Spherical	Anaerobic respiration	Chlorobium
	Photosynthesis		Cell division by fission	
			Gram negative cells	

Table.5 Table showing some of the results of Species

Sl No.	Functional Attributes	Shape	Other Characteristics	Results
1.	Chemolithotrophic	Rod	Binary fission	*Thermoproteus neutrophilus*
	Nitrogen fixation		Gram negative cell	
			Nonmotile and lack flagella	
2.	Chemolithotrophic	Rod	Gram negative cell	*Thermoproteus tenax*
	Thermophilicarchaea		Nonmotile and lack flagella	
			Mode of respiration Anaerobic	
3.	Chemoorganotrophic	Disc	Gram positive cell	*Haloferax volcanii*
	Halophilicarchaea		Mode of respiration Aerobic	
			Nonmotile and lack flagella	

Property based search or advanced search

The Advance search module is as shown in Figure 10. This is an advanced version of the name based search described in the next section. This module comes under the tab of "Advanced Search". This search is dedicated to all the hierarchy of microbial taxonomy from domain to species level. Every level of hierarchy has some typical attribute that separates one from the another. We tried to capture those characteristics of the microbial taxonomy.

Figure 10 describes the Probable Domain; which is Archaea. User can study in detail about the Domain Archaea, on click More Information. Similarly one can search for Phylum, Family, Genus and Species by Advanced Search module (Table 1, 2, 3, 4 and 5).

Sequence search and alignment

Microbial ontology consists of 16S rRNA sequence of Archaea. This information not only used for showing purpose we take it in one step ahead. If users have unknown sequence, then user can know by Sequence Search tab to unknown sequence corresponds to exact match with the existing sequences, Figure 11. Otherwise if the users have partial sequence data then they can align their sequence with the existing sequence and identify the probable species Archaea, Figure 12.

Provision for editing knowledge base (Ontology) by domain experts

Microbial Taxonomy Ontology knowledge base can be edited by domain experts; if there is any wrong entry by the system developer or any new information is available regarding the particular microorganisms. It is done by using Edit Ontology tab. On clicking this tab, an interface will guide domain expert for the editing purpose. After proper review of the changes made by domain experts the final change may be committed in the ontology. Microbial Taxonomy Ontology is a rich repository of information of agriculturally important microorganisms- Bacteria and Archea. This system will be beneficiary for the community of microbiologists and agriculturalist worldwide. The taxonomic description of microbial taxonomy will help in the detailed study of agriculturally important Bacteria and Archaea. Apart from the sequential study of the taxonomy, the system enables us to randomly search term related to the microbial taxonomy called the name based search. The term based search or the name based search is not sufficient, hence, Microbial Taxonomy Ontology provides the advanced search module or the property based search module. This module provides the selection facility of the special characteristics of a particular hierarchy (e.g. Domain, Phylum etc.). On the basis of the

property combination, it can give the probable hierarchy that matched with the particular set of characteristics. As we discussed earlier, the system provides the sequence search and sequence alignment concurrently. Both types of sequence search can be of utmost importance to the microbiologist as well as the experts in the field of bioinformatics. The system also has the secure login facility to maintain the user privileges.

Ontology is applied in several research area, including database design and integration, information retrieval and extraction, software engineering and natural language processing. Knowledge base of this software can be enriched by the information of all the microorganisms based on the Taxonomic Classification to classify any recognized microorganisms. There is a scope of enhancement of information of the knowledge base up to strains level, therefore making it as a tool for the other usage of microbiological areas; such as-Industrial microbiology, Marine microbiology, Medical microbiology etc.

Acknowledgments

References

Ashburner M, Ball C A, Blake J A, Botstein D, Butler H, Cherry J M, and Harris M A., 2000. Gene Ontology: tool for the unification of biology. Nature genetics *25*: 25.

Bedi P, Marwaha S., 2004. Designing Ontologies from Traditional Taxonomies. Proceedings of International Conference on Cognitive Science 324-329.

Bergey D H, Harrison FC, Breed RS, Hammer BW, Huntoon FM., 1989. Bergey's Manual of Systematic Bacteriology 3.

Biswas S, Marwaha S, Malhotra P K, Wahi S D, Dhar D W, and Singh R., 2013. Building and querying microbial ontology. Procedia Technology, 10:13-19.

Clark K., 2008. SPARQL protocol for RDF, W3C Recommendation, http://www.w3.org/TR/rdf-sparql-protocol/.

Das M, Malhotra PK, Marwaha S, Pandey RN., 2012. Building and Querying Soil Ontology. Journal of the Indian Society of Agricultural Statistics. 66(3): 459-464.

Deb C K, Marwaha S, Malhotra P K, Wahi S D, Pandey R N., 2015. Strengthening soil taxonomy ontology software for description and classification of USDA soil taxonomy up to soil series. Proceedings of International Conference on Computing for Sustainable Global Development (INDIACom) IEEE. 1180-1184.

Madigan MT, Martinko JM, Parker J., 2006. Brock Biology of Microorganisms Eleventh Edition. USA: Pearson Prentice Hall.

Ming Z, Qingling Z, Dong T, Ping Q, Xiaoshuan Z., 2009. Ontology-based intelligent retrieval system for soil knowledge. Wseas transactions on information science and applications 6: 7.

Plant Ontology Consortium. 2002. The Plant Ontology consortium and plant Ontologies. International Journal Plant Genomics. 3: 137-142.