International Journal of Oceans and Oceanography
ISSN 0973-2667 Volume 11, Number 2 (2017), pp. 147-158

http://www.ripublication.com

A Technique for Parameter Identification, classification and structuring of Marine Genomics and Oceanographic mutations

Soumya Mahalakshmi A[1], Dr. Shantha Rangaswamy[2], Karthik S P[3], Madhuri N Kirani[4], Himanshu Tanwar[5]

Department of Computer Science and Engineering,
R.V. College of Engineering, Bengaluru, India

Abstract

It is widely known that any population in the world is under the constant purview of evolution. In ocean organisms, mutations are more visibly noticed, and also of prime importance, as they are at the bottom of the food chain. The ocean ecosystem is also highly fragile, whose imbalance can lead to widespread destruction. New observations of mutations in organisms are usually presented by scientists as journal papers or news reports. The loophole in the system is that all the data about oceanic mutations in the world are textual and scattered. This fails to give a realistic picture about the impact that various human factors have had on oceanic mutation. The proposed solution involves a prior retrieval of textual data from various sources around the Web pertaining to ocean mutations, by using a Web Crawler to create a repository. Further, an algorithm has been developed using Natural Language tools in order to convert textual data into a relational database of relevant attributes. In order to obtain visualization that is more comprehensive and interactive, a Geographic Information System (GIS) centered on this mutation data was constructed and uploaded online to benefit biologists worldwide. Analysis on structured data was simpler, and the results show an alarming number of mutations recorded in the Pacific Ocean which account for almost 92% of the observed mutations. The analysis has also predicted sex changes that occur in 13% of the mutations in ocean organisms due to radiation and pollution, amongst several other observed oceanic mutations. Besides, oil spills and

micro plastics have resulted in fatal mutations, and hence, it calls for a more surgical approach in eliminating these causes, to preserve ocean ecosystems.

Keywords **-** Marine mutations, Oceanography, Unstructured data, Text mining, NLTK, Web Crawler, GIS

A. Introduction

In ocean organisms, mutations are more visibly noticed, and also of prime importance, as they are at the bottom of the food chain. The ocean ecosystem is also highly fragile, whose imbalance can lead to widespread destruction. Positive mutations shall cause no harm, but negative mutations are very dangerous. Also, Mutations are biologically complex and difficult to understand and experiment. Hence, new observations of mutations in organisms are usually presented by scientists as journal papers or news reports. In spite of the closing gap between computer scientists and biologists, this divide between the two sciences is one of the reasons why this problem arose in the first place. It is widely known that any population in the world is under the constant purview of evolution. Ocean organisms are no different. However, certain sudden spurts of growth may occur, resulting in mutations.

The loop hole in the system is that no real attempt has been made to create a database of observed mutations in the ocean. As a result, all the data about oceanic mutations in the world are textual and scattered. This fails to give a realistic picture about the impact that various human factors have had on oceanic mutation. For instance, in Fukushima in Japan, fish developed devilish features, and even changed their gender due to radiation in the water. When read as a news article, it generates a short-lived moment of fear, but when there is a unified database of all such mutations, the results are truly alarming.

In the paper “Ecosystem based Fishery Management (EBFM)”[1], spatial zoning of the marine environment is explored as one of the primary management tools. It reviews the single-species models that have been used to model spatial zoning, including current work on fishing effort reallocation after an area closure, and discusses how spatial management might bias assessment. The review of the available ecosystem-based models and metrics, and how they might account for spatial management is explored in this research. Metrics that could be derived from explicitly spatial approaches such as GIS-based ecosystem and fishery evaluations have also been explored as valid techniques to marine research. However some of the cons were, EBFM will require indicators of the effectiveness of spatial management, as well as an understanding of how indicators related to other management objectives, such as fisheries yield, will be influenced by spatial management. Paper “Shark Genetics and Breeding Biology”[2], explains that large sharks are highly mobile, there is growing evidence that population declines are often

remarkably localized. This results in the establishment of a series of largely independent subpopulations that dwell around nursery areas. This paper develops a comprehensive picture of how sharks are connected to their birthplace, known as their "natal area", throughout their life. However some of the disadvantages were, conservation would require management over wide areas and probably across multiple jurisdictions and more rigorous local management and monitoring efforts.

In the paper "Adverse Eco-Feedbacks"[3], development of a seasonally and spatially improved thermal threshold for coral bleaching on the basis of a weekly climatology of sea surface temperatures extending from austral spring to late summer is elucidated, and the method is applied to two case-study sites. The application of thermal stress algorithms that reflect the long term mean pattern in seasonal variation allows coral bleaching to be forecast with higher precision. However some of the cons were, current satellite-derived warning systems were unable to detect severe bleaching conditions in the region because of their use of a constant thermal threshold (summer maximum monthly mean) and low spatial resolution (50 km). In the paper "Reversal of Undesirable Evolution in Fish"[4], it was estimated that if the largest fish were left unharnessed, full recovery of the silverside would take approximately twelve generations in a controlled situation such as the laboratory. Recovery for wild populations of those and other fish could be shorter or longer depending upon the species and environment. Dr. Conover's research has provided the first direct experimental evidence that the growth rate of fish, and therefore productivity, can rapidly evolve in response to the pressures of size-selective harvest and can be reversed if allowed to recover without interference. However some of the cons were, Fishery management plans fail to incorporate these evolutionary dynamics. Since this unwittingly promotes the evolution of fish to be smaller and less productive, it is essential for fishery management plans to be analyzed and modified to promote sustainable practices and healthy future fisheries.

Hence, the aim of this research was to convert the textual data that reports oceanic mutations from around the world, and convert this unstructured data into a structured database, by using text mining. This database will act as a precursor to several applications that are centred around this topic. For instance, a Geographic Information System about Oceanic mutations can be built, which will aid all future marine conservation projects. Focus can be laid on the prime causes of mutations to individual human factors and work towards eliminating them. If the database is mined for patterns, it can also aid lab simulations of the mutations. Therefore, this project will act as an essential precursor to host new possibilities.

There can be several causes of oceanic mutations, such as radiation, pollution, microplastics and so on. The following examples are indications of the impact that these effects can have on ocean organisms. Once in seawater, radiation can hurt ocean

animals in several ways—by killing them outright, creating "bizarre mutations" in their offspring, or passing radioactive material up the food chain, according to Joseph Rachlin, director of Lehman College's Laboratory for Marine and Estuarine Research in New York City.Marine organisms' eggs and larvae are highly sensitive to radiation, since radioactive atoms can replace other atoms in their bodies, resulting in radiation exposure that could alter their DNA. Most such deformed organisms don't survive, but some can pass abnormalities on to the next generation, Lehman College's Rachlin said. Either way, the radiation exposure could hurt the population's ability to survive long-term.In addition to its threats to reproduction, pockets of radioactive material can burn fish passing through, hitting them like a stream of searing water. Complicating matters is the fact that predator species in the Pacific such as tuna and sailfish are already stressed by overfishing.

Therefore, it becomes extremely important to understand mutations in the ocean in a structured manner, unlike today, where the information is scattered. A structured collection of all the data from available sources across the internet can guarantee a surgical approach to combat the root causes of harmful mutations in the organisms which are at the bottom of the food chain.

B. Methodology

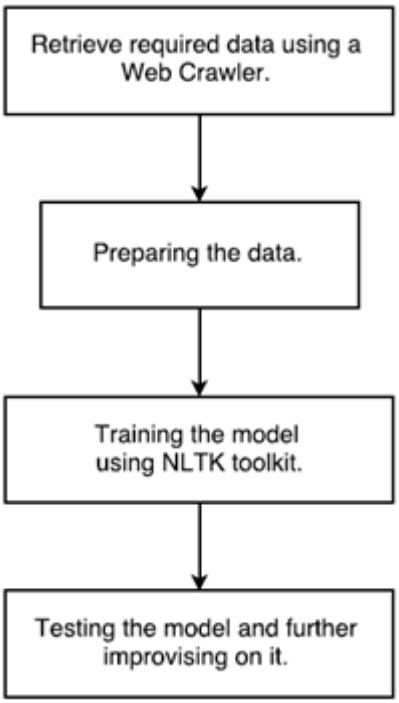

Figure 1: Methodology of research design

In figure 1 the steps to be followed have been explored. The first step involves the collection and retrieval of articles pertaining to oceanic mutations. Hence, a web crawler is used to crawl over the World Wide Web and pull the links of articles which are of relevance to marine mutations. From the obtained links of news articles of journal papers, the textual data is retrieved and stored in a repository.

The second step involves cleaning the data and making it suitable for execution of the algorithm in the most optimal manner. The next stage involves the execution of the conversion algorithm which converts the textual data into a structured data table with four attributes, at present. Depending upon the context of the data, each of the values under the corresponding attributes is added. Ultimately, a comma separated values file (.CSV File) is obtained, that can be imported to Microsoft Excel to carry out data analysis and GIS map generation.

The Web Crawler is a separate module which merely returns relevant links to papers and articles related to oceanic mutations from all around the web, which would have been highly difficult to trace manually. The textual data from these documents are retrieved manually, and processed by NLTK toolset in Python. It works on creating similar context, and retrieving features, queried for from the textual data. The retrieved data is in the form of a .CSV file, which can be exported to Microsoft Excel. This data can be used to get further insights, one such instance being illustrated as the Geographic Information System (GIS) of mutations in the ocean.

One of the problems encountered was the recognition of multiple words as the parameter value. This was an issue because there was no specific pattern in the occurrence of single or multiple words and so specific rules were difficult to be laid down for the identification of the multi word parameters. The solution to this problem was to consider all the words following the first word as part of the parameter as long as a punctuation doesn't break the sentence or one of the words from the list of prepared repository of the words that do not form part of the parameter appear in the sentence. Example of this repository is [“of”, ”and”, ”the”, ”but”, ”or”, ”also”, ”is”, ”was”, ”a”, ”an”, ”at”, ”under”, ”above”, ”then”]. Another major issue was in writing in the excel file from the python source in a tabular form. Several implementations were considered from the internet but most of them were very complex and without much customization possible. The solution for this problem was to first convert the data gathered into comma separated values and storing into a csv file which was then converted into a excel file. This solution was chosen because it is very easy to convert the data into comma separated values, then writing to .csv file, and a .csv file can be very easily converted into an excel file.

C. Experimental Analysis

The problem of marine mutations and their study has far reaching consequences, and are worth being watched. In the process of providing a mechanism for the structured organization and study of marine mutations, it was necessary to automate the conversion of the large repository of scattered and textual data about marine mutations into a structured and tabular representation, which can act as the foundation for further study and analysis.

It has been established that the input data is in the form of news articles, blog posts and research papers, the output has been obtained in the form of a .csv file. This has been imported into Microsoft Excel to perform further graphical analysis. The Microsoft Excel File consisted of 37 tuples, each of which has 4 attributes which are Species Name, Mutated Feature, Location, Ocean and Cause as shown in Table 1.1.

Table 1.1 Structured data table of some oceanic mutations

Species	**Feature**	**Location**	**Ocean**	**Cause**
Sea Anemone	Sex Changes	Fukushima	Pacific Ocean	Radiation
Fish	Visible Tumors	Gulf Coast	Atlantic Ocean	Oil Spill
Fish	Visible Tumors	France	Atlantic Ocean	Pollution
Shrimp	No eyes	Fukushima	Pacific Ocean	Radiation
Crabs	No eyes	Fukushima	Pacific Ocean	Radiation
Fish	Irradiated Features	United States	Pacific Ocean	Contamination
Bull Shark	Two Jaws	Florida	Atlantic Ocean	Pollution
Fish	Three Eyes	United States	Pacific Ocean	Microplastics
Fish	Intestinal Injury	California	Pacific Ocean	Microplastics
Sea Birds	Small Stomachs	California	Pacific Ocean	Microplastics

D. Result Analysis

In order to obtain visualization, that is more comprehensive and interactive; a Geographic Information System (GIS) is constructed centred on this mutation data. The following map in figure 2 has been obtained from a GIS tool called BatchGeo, which has been used to map the occurrence of mutations to the corresponding locations of occurrence.

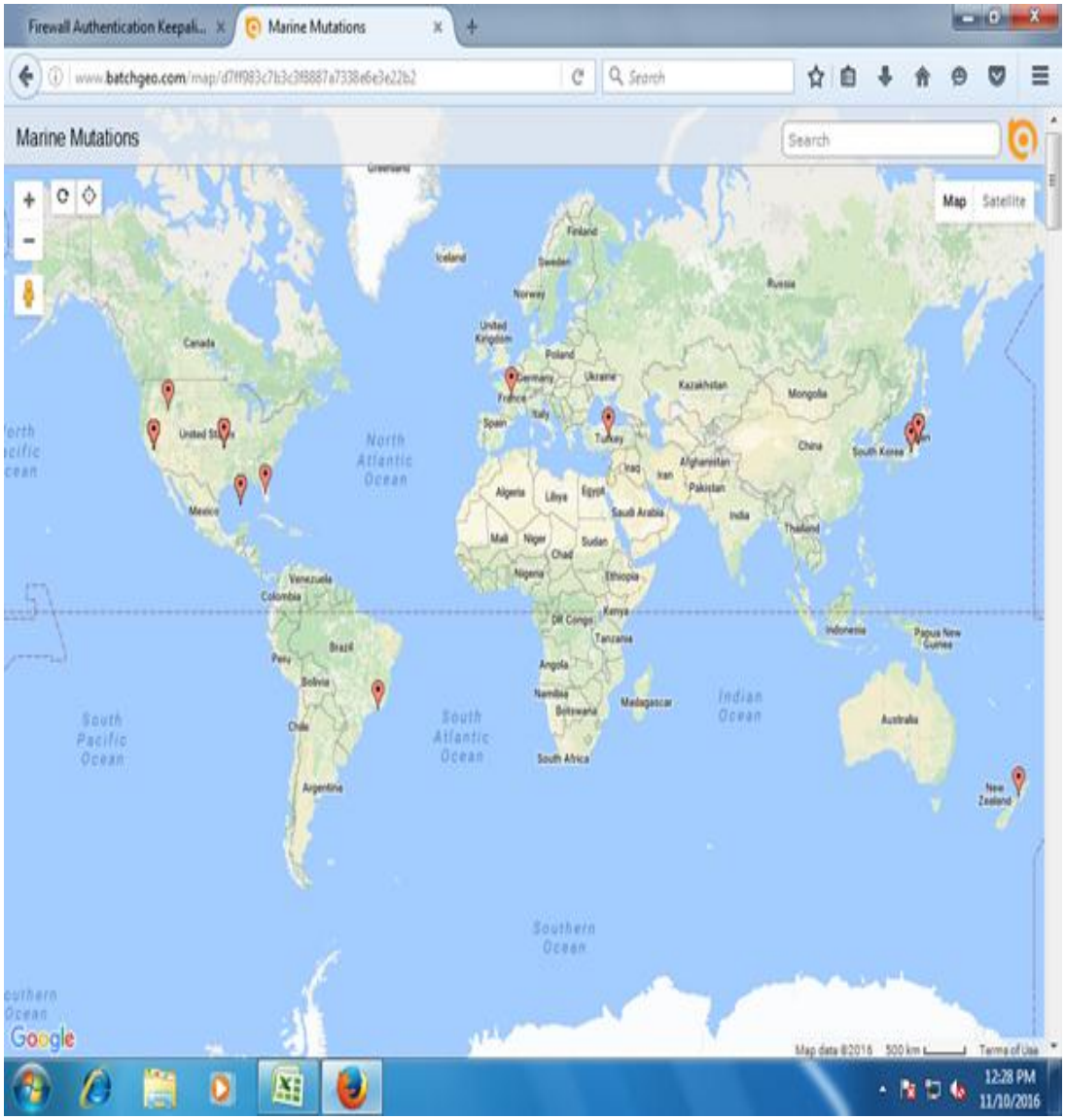

Figure 2: Map using GIS tool called BatchGeo

The experimental data has also been used to construct the following graphs which provide a deeper understanding of mutations in a statistical sense, so that this quantification can aid a more surgical approach towards environment conservation, and precautionary and recovery measures. In figure 2 it has been observed that a majority of the mutations are concentrated in the Pacific Ocean. While the vastness of the ocean itself might be a contributing factor to an increased chance of finding mutations, this next graph in figure 3, provides a better perspective on the reasons behind this concentration of mutations in the Pacific Ocean.

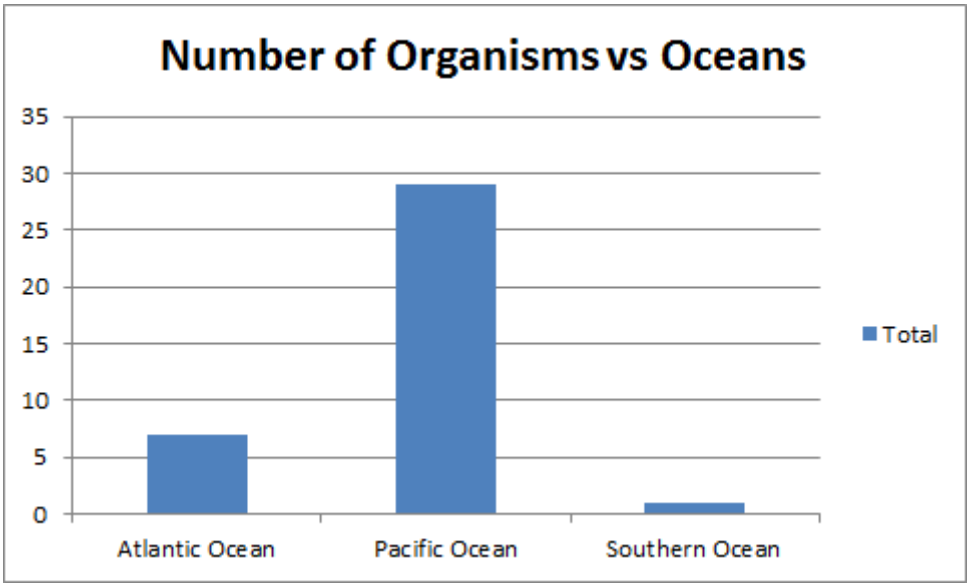

Figure 1.2 Number of Organisms vs Ocean

It has been observed that radiation and pollution are the causes that have been anticipated to cause largest number of mutations in ocean organisms. In the next graph in fig. 1.4, it can be seen that out of all the oceans, the Pacific Ocean has recorded the largest number of mutations due to radiations and pollution, as compared to other oceans. Hence, there are a large number of mutations concentrated in the Pacific Ocean.

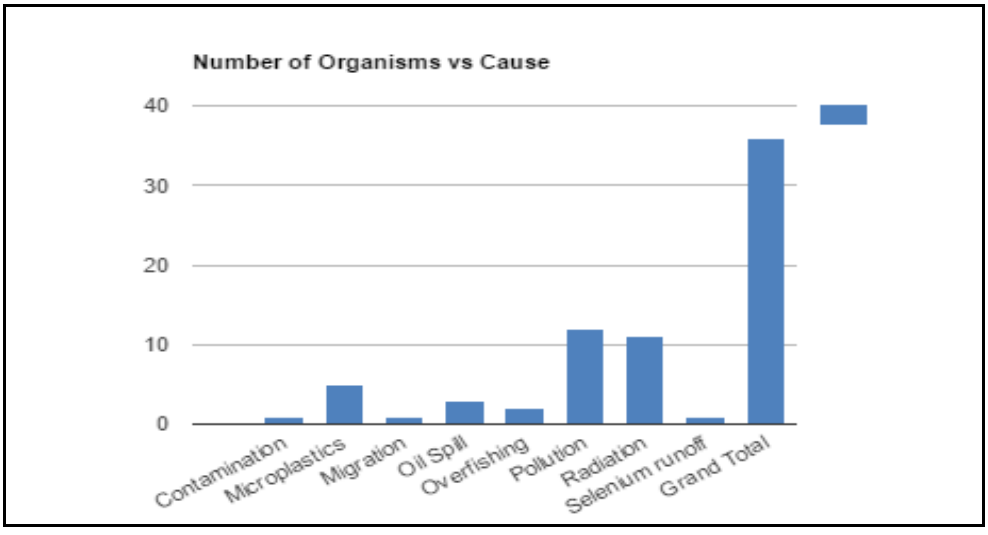

Figure 1.3 Number of Organisms vs Cause

Also, it would help geneticists if the causes that are responsible for each of the mutated features in the organisms could be understood. This not only enables correlation of probability, but also, possible simulation of cause conditions for further study of mutations, or to understand how various environmental conditions affect gene structure. It has been illustrated in fig. 1.5.

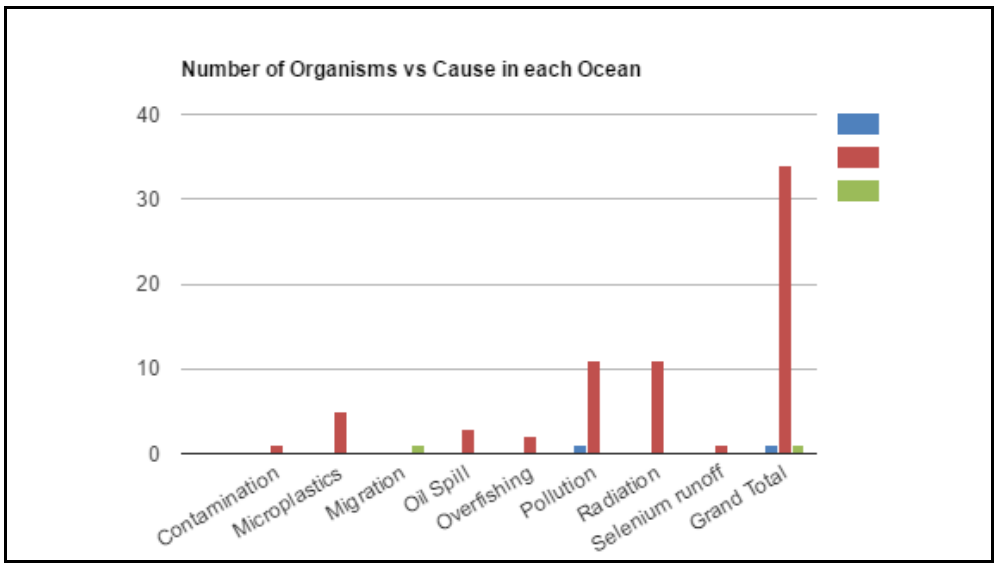

Figure 1.4 Number of Organisms vs Cause in each Ocean

In order to understand the number of organisms corresponding to each mutated feature, the following graph in fig. 1.6 was constructed. It is seen that a maximum number of organisms have been affected with sex changes. From the previous graph in fig. 1.5, it can be observed that sex changes have been attributed to majorly radiation and overfishing.

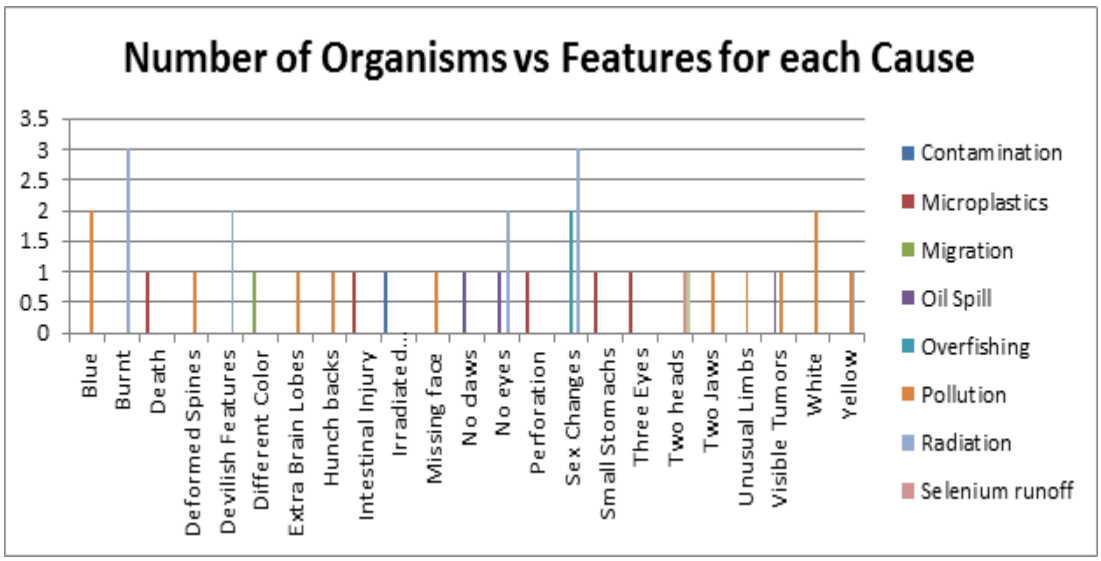

Figure 1.5 Number of Organisms vs Features for each Cause

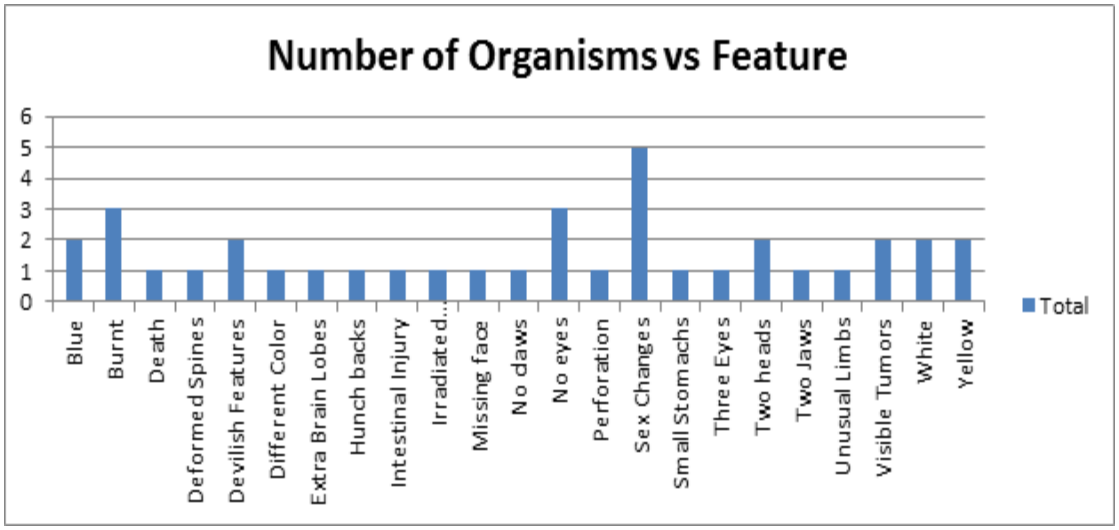

Figure 1.6 Number of Organisms vs Feature

Also, from the graph below in fig. 1.7, it is observed that sex changes are observed in the Pacific Ocean mostly. This graph in fig. 1.7 also illustrates the maximum number of organisms having each of the observed mutated features with respect to each ocean.

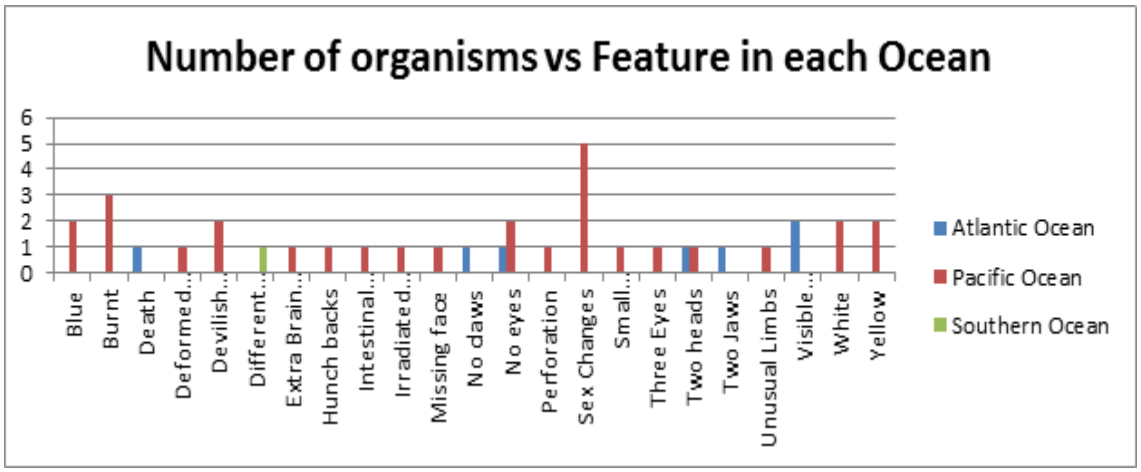

Figure 1.7 Number of organisms v/s Feature in each Ocean

From all the results illustrated in the previous section, following observations of the trends are made:

(i) The Ocean that has been affected the most with oceanic mutations is the Pacific Ocean, with the Atlantic Ocean following at a close second.

(ii) The reason for mutation in the Southern Ocean seems to be migration, which is not a negative cause.

(iii) It has been observed that Radiation and Pollution are the factors that majorly result in mutations in ocean organisms.

(iv) It has also been observed that most mutations in organisms result in sex changes, which are less fatal as opposed to mutations that result in burning or loss of organs.

(v) It has also been observed that oil spills and micro plastics are the causes for mutations that are fatal such as loss of organs or death.

E. Summary and Conclusion

This project was initiated with the objective of removing a loophole in the way mutations are studied in the current scenario, by structuring mutation data about the oceans into a database to enable further study. While it has been a rigorous but successful computational challenge, it has also given greater insights on the environment, and the numbers predict an alarming degradation of ocean ecosystems which pleads for immediate action.

A system was conceived to convert the unstructured oceanic mutation data into structured database tables to study mutations better. Also, a mechanism to retrieve all the articles of interest to our study was built by using a Web Crawler. Further, an algorithm was built using Natural Language Tools such as NLTK and Python to convert the textual data to a relational table, by identifying contexts in the data.

The structured data was used to build a GIS of Ocean Mutations, with better visualization of the data which has been uploaded online to benefit biologists and scientists worldwide. Analysis on the structured data has called for immediate action in the Pacific Ocean in areas near California and Japan, which have recorded highest number of mutations, due to radiation and pollution. It has also predicted an increase in sex changes in ocean organisms due to radiation and pollution, amongst several other observed oceanic mutations. Apart from radiation and pollution, oil spills and micro plastics have resulted in fatal mutations, and hence, it calls for a more surgical approach in eliminating these causes, to preserve ocean ecosystems.

F. Future Enhancement

In the future, it has been proposed to make the database more comprehensive, by adding attributes that carry the biological name of the organism and so on, so as to avoid ambiguity. These are regarded as derived attributes, which can take the obtained data as inputs to make a PHP query to obtain the new derived attributes. Also, since the algorithms works based on identification of context, it is difficult to identify texts which contain multiple organism names attributed to various features in a single line. There are possibilities of retrieving articles of lower credibility which might lower the

authenticity of the database. Therefore, this database is still regarded as a primary database, which needs conversion to secondary databases.

REFERENCES

[1] Babcock E A, E K Pikitch, M K McAllister, P Apostolaki, C Santora (2005) "Ecosystem based Fishery Management (EBFM)", ICES Journal of Marine Science. 62: 469-476.

[2] Charles Sutton, Andrew McCallum (2011) "An Introduction to Conditional Random Fields" by Foundations and Trends in Machine Learning Vol. 4, No. 4, 267–373 c 2012 C. Sutton and A. McCallum DOI: 10.1561/2200000013.

[3] Conover D.O., Munch, S.B., Arnott S.A.(2009) "Reversal of Undesirable Evolution in Fish", Proceedings of the Royal Society B, doi: 10.1098/rspb.2009.0003,pp 1-6.

[4] Feldheim K.A, Chapman D.D, Snowden D, Fitzpatrick S, Prodohl (2010) "Shark Genetics and Breeding Biology", Journal of Heredity.

[5] G. Andrew and J. Gao (2007)"Scalable training of L1-regularized log-linear models", International Conference on Machine Learning (ICML), Corvallis.

[6] S.J, Anthony, Bakun A, Feldman G.C, Hoegh-Guldberg (2006) "Adverse feedback sequences in exploited marine systems: Are deliberate interruptive actions warranted? Fish and Fisheries, 7 4: 316-333, doi:10.1111/j.1467-2979.2006.00229.x

[7] Srinivas M. Aji and R. J. McEliece (2000) "The generalized distributive law," IEEE Transactions on Information Theory, vol. 46, no. 2, 0018–9448/(00)01679-5,pp. 325–343.

[8] Y. Altun, I. Tsochantaridis, and T. Hofmann (2003) "Hidden Markov support vector machines," Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC.