International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-5S3, July 2019

Design of a Rule Based Bio Medical Entity Extractor

G. Suganya, R. Porkodi

Abstract: ***The field of Biomedical Entity Extraction/ Identification plays a vital role in Bioinformatics and rapidly growing to meet the needs of different text mining tasks. Many biomedical entity extraction tools have been developed so far. This research work has focused to develop a Rule based Biomedical Entity Extraction and tested with PubMed Medline abstracts of Colon cancer and Alzheimer disease categories. The proposed Biomedical Entity Extractor gives promising result when compared with existing tools. The proposed method is incorporating of two phases such as preprocessing the input text document using NLP techniques and create the rules to find out the biomedical entities using regular expression. The results of Rule based Biomedical Entity Extractor are validated with the well-known Biomedical Genia tagger and Genecards Database. The method proposed in this paper almost good as genia tagger. The evaluation results on Colon cancer and Alzheimer disease abstracts corpus of Biomedical Entity Extraction achieve an accuracy of 92% and 88% respectively which identifies more number of entities compared to other existing tools.***

Index Term: ***Medline Abstracts, Genia tagger, Pubtator, BCC-NER, Biomedical text mining***

I. BACKGROUND STUDY

Text mining (TM) is an automatic and used to identify the important known and unknown knowledges from the different types of given input text document. It can also be called as Knowledge discovery from text (KDT) [1]. It used to identify and extract the useful and important patterns from large number of text documents for improving the entities extraction range. It applies the basic knowledge of the analytical functions which included in the data mining techniques and also used the basic concepts of natural language functions which includes the information retrieval and information extraction techniques. The main aim of TM is used to identify the interesting patterns and accomplish the retrieval, extraction and etc. TM is an interdisciplinary field that includes the several techniques such as data mining, web mining, information retrieval, information extraction, computational linguistics and natural language processing [2]. Text databases are useful in the increasing of the information. The large databases are maintained in to the various sectors such as publications, documents, e-mail, and the world wide web. At the present scenario, all the data are stored into the databases of electronical format. Finally, from this analysis all the data is in the form of unstructured or structured [3] but mostly the data will be in semi-structured data. TM in bioinformatics field which is the subfield of the text data and it includes the various and different fields such as biology, medicine and chemistry. The biomedicalarticles are not a homogeneous realm [4]. The generic workflow of the TM is shown in fig 1.

TM process starts with the initial stages of collecting the input text document from different types of format. TM tool is used for retrieving the useful information from the input document. The next phase is text analysis. It is used to retrieve the high dimensionality of the data. There are various techniques are available to find out those entities. In that some of the techniques are repeated applied because to find out the correct entities and needed entities. Finally, the desired output is obtained.

Burr Settles [5] has developed a tool named as “ABNER” to identify the entity names from the abstracts like Protein ,DNA, RNA, Cell line and Cell type. The identification of entities is based on the regular expressions which are manually created. The tool achieved better results. The tool was tested with two corpora namely NLPBA and Biocreative.

Gurusamy Murugesan et al [6] had described the BCC-NER approach is a combination of forward and backward models with Conditional Random Fields (CRF) technique. For the experimental study Biocreative II GM corpus are considered. Totally 15000 training sentences and 5000 testing sentences are taken for the experimental study. MIRA algorithm was applied to integrate both models for constructing the new model.

Raja et al [7] found the model to identify and tag the biomedical entities of protein/ gene from the biomedical articles which used the concept of named entity recognition with manual rule based post processing method.

Leaman et al[8] proposed the new concept for the identification of entities by using the CRF with post processing techniques.

Campos et al [9] proposed a concept for the biomedical entities. It is the hybrid method which incorporates the machine learning algorithm and lexicon-based approaches.

Chith-Hsuan Wei et al [10] developed a tool named as GNormPlus for the identification of end to gene protein extraction as well as the identification. It includes the various advanced biomedical natural language processing techniques such as GenNorm, SR4GN, Ab3P and SimConcept. For the experimental, two datasets are taken which are named as BioCreative II GN and NLM citation GIA test data. From this analysis, to conclude that the proposed tool getting more results than the existing one. This is an open source tool. This concept is tested with the PubMed databased and the results are stored in the Pubtator tool.

Manuscript published on 30 July 2019.
* Correspondence Author (s)
G Suganya, Ph.D Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore.
R. Porkodi, Associate Professor, Department of Computer Science, Bharathiar University, Coimbatore.

Fig. 1 Text Mining Generic Workflow

C.-H. Wei et al [11] developed a methodology to fetch the data which is in the form of composite entities where single entity named into multiple concept in different fields. Most of the studies ignored this problem. The create and developed is named as SimConcept. It is mainly useful to tag the entities which are in the form of composite named entities.

S.Sohn et al [12] designed and developed a tool named as Ab3P and it is an open source tool. It is useful to retrieve the data in the form of both long and short form of the abbreviation formats. When both the form of representation is different than long form of abbreviation is marked and changed into short form for increasing the performance of the given input text document.

The motivation of this research paper work is to identify and extract the entities from Colon cancer and Alzheimer disease abstracts using rule based approach. The paper is organized into 4 phases. The phase 2 discusses methodology framework, Phase 3 focuses on results and discussion. The work is concluded in section 4.

II.METHODS

A. Dataset Description

The Colon cancer and Alzheimer disease abstracts are used for the experimental purpose and both have more than 30 abstracts. These abstracts are collected from NCBI, Pubmed database. For each dataset, accuracy is measured at different levels at various abstracts size. Overall average performance is calculated for each text mining tools for the given dataset. Table I shows the overall abstracts for each text mining tools under all levels of significance. In fig.2 represents the comparative performance evaluation process.

B. Pre- Processing

Three pro-processing techniques are applied to experimental abstracts such as tokenization, stop word removal and stemming. The Tokenization is the technique to split the sequence of strings into meaningful words . the splitted words are in the different format such as keywords, phrases, symbols and etc. here, the special characters are also removed. This is the main reason to perform this task. There are many more stop words are available. No universal list for the stop words. It removes the unwanted words from the text document which are not useful to predict the correct entities. Some of the stop words are the, those, many, high, low and etc. the stop word removal is useful in the reducing of the number of data and also improves the system performance. Stemming is the process to convert into the single unique format. The words of accept, accepts ,accepting, accepted are represented into the common word called accept. This is the main technique in the preprocessing techniques for both information retrieval(IR) and information extraction (IE) [13].

C. Tools for Identifying / Extracting Biomedical Entities

There are enormous entity extraction tools were developed and some of the tools are Genia tagger, BCC-NER, Pubtator, tmvar, BANNER, Gimli, GAPSCORE and Abner. Though there are number of tools Genia tagger, BCC-NER and Pubtator are focused more in literatures and discussed in below.

1. Genia Tagger

Genia tagger is used to analyses the many subfield of English sentences, part of speech tags (POS tag) and named entity tagger (NER). The POS tag is identified and created for the Medline abstracts and also extracted the meaningful and important information from given input biomedical documents. This is a one of the preprocessing technique[15]. The Genia tagger is useful in different kinds of tasks. Some of the important functionalities are identifies the needed words and also extracts the biomedical entities such as proteins, cell lines and cell types. It also identifies the noun phrases. The shallow parsing is easy for extracting the noun phrases from the sentences [16].

2.BCC-NER

BCC-NER is the hybrid named entity recognition system which uses the benchmarking datasets of BioCreative II GM corpus. It composed into three modules such as the first model is for NLP preprocessing, feature extraction and selection. PCA algorithm is applied to reduce the high number of features associated with it. The second module is a combinations of two models such as bidirectional CRF for learning and MIRA. Features set with the bidirectional CRF is nothing but including both backward and forward directions. For the performance analysis, it used the precision, recall and f-scores. The third module is post processing module which includes contextual clues and abbreviation identification algorithm. Finally the tagger identifies the entities and marked with the GENE tag (<gene>) [6].

3. Pubtator

Pubtator is a web-based system and it does not need any installation part and not restricted to any specific computer platforms. It is the one step system service provider for searching the articles as well as to annotate the searched or important or selected articles. For selecting the articles as the input, it may be from the search results or from the whole PubMed articles databases. It designed like an interface which some of that are found similar and it required only few numbers of training articles are required.

Multiple competition text mining approach have been integrated into for identify the important and needed entities automatically. It searching options incudes both keyword searches and semantic searches based on the bio-entities. It contains five options named as Pubmed, Gene, Disease, PMID List and Chemical. It returns the obtained results in the reverse
sequential order. From that only few results are visible to the user because of the page range. For example, only 20 articles are visible instead of all searched results. Pubtator can be used for annotating relationships between entities. It allows curators to specify what the relations obtained from the given abstracts. Pubtator improved both manual curation accuracy and user curation tasks. It provides practical benefits to biocurators in their curation work [17].

4. Proposed Method: Rule Based Biomedical Entity Extractor

The rules are constructed using regular expression. The Biomedical Entity extractor is containing the two phases. In that first phase is used for preprocessing by using basic Natural language preprocessing (NLP) techniques such as Tokenization, stop word removal and Stemming. The second phase is to build rules to identify only 2 biomedical entities such as Gene and Protein names and not including the Disease, Species, Chemical and Mutation. There are 193 rules are framed using regular expressions and few rules are listed in Table II. The proposed method includes all the gene nomenclature which are not included in the existing text mining tools namely BCC-NER, Pubtator and Genia tagger. The results of Rule based Biomedical Entity Extractor are validated with the well-known Biomedical Genia tagger and Genecards Database. The method proposed in this paper almost good as Genia tagger.

Some interesting rules are discussed which are not included in the Pubtator and BCC-NER. In the Rule 1, "*The word of the letter matches the full capital letter of each word in the full name*" identifies the entities like "BRAF". Rule 2 "*The word of the letter matches the two words with special character; first word full capital letters and second word full capital letters followed by Arabic numerals*" identifies the entities like NVP-LDE225 and Rule 3, "*The word of the letter matches the first letter capital and followed by numbers and last letter is capital letter*" identifies the entities like F691L.

III. RESULTS AND DISCUSSION

The proposed method tested and evaluated with the two groups Medline abstracts Colon cancer and Alzheimer disease and these abstracts are considered as four levels of size 5,10,15 and 30 in each group as presented in Table I.

Table I Level of Significance for Colon cancer and Alzheimer disease

Levels		1	2	3	4
Colon Cancer	No. of Abstracts	5	10	15	30
	Size (KB)	21	30	40	95
Alzheimer Disease	No. of Abstracts	5	10	15	30
	Size (KB)	13	20	31	70

The found entities accuracy of the biomedical entity extractor has been compared with the other techniques. Biomedical Entity Extractor is based on the result of the Geina tagger as a threshold presented in Table III (a) and 3 (b). For the Colon cancer abstracts the proposed method identified 21,62,79 and 142 gene names Biomedical entities from 4 different levels of abstracts of size 5,10,15 and 30 respectively where as Pubtator and BCC-NER identified 11,24,27, 52 and 8,14,20,34 gene names as biomedical entities respectively from the level of abstracts as mentioned in Table I. Thus, the proposed method achieves good average performance accuracy 92% where as the Pubtator and BCC-NER achieves 41% and 28% average accuracy respectively.

For the Alzheimer disease abstracts the proposed method identified 19,32,51 and 137 gene names Biomedical entities from 4 different levels of abstracts of size 5,10,15 and 30 respectively where as Pubtator and BCC-NER identified 10,10,14,28 and 7,12,15,40 gene names as biomedical entities respectively from the level of abstracts as mentioned in Table I. Thus, the proposed method achieves good average performance accuracy 88% where as the Pubtator and BCC-NER achieves 28% and 29% average accuracy respectively.

The execution time taken by the already existing tools and proposed method are calculated for the execution platform with a laptop with the intel inside core i5, 3 GHz and 1TB RAM.

Table III (a) No. of entities Retrieved for Colon Cancer abstracts

Tool / Significance Level (%)	1	2	3	4	Avg
Pubtator	52	39	34	38	41
BCC-NER	38	23	25	24	28
Proposed Method	86	89	95	98	92

Table III (b) No. of entities Retrieved for Alzheimer Disease abstracts

Tool / Significance Level (%)	1	2	3	4	Avg
Pubtator	40	29	24	19	28
BCC-NER	28	35	25	28	29
Proposed Method	76	94	86	94	88

The Table IV, for the Colon cancer abstracts the proposed method identified gene names Biomedical entities in 8, 15, 20 and 35 seconds from 4 different levels of abstracts of size 5,10,15 and 30 respectively where as Pubtator and BCC-NER identified 5,10,15 and 30 seconds from the level of abstracts as mentioned in Table I. Thus, the proposed method achieves good average execution time 20 seconds where as the Pubtator and BCC-NER achieves 15 seconds of both.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

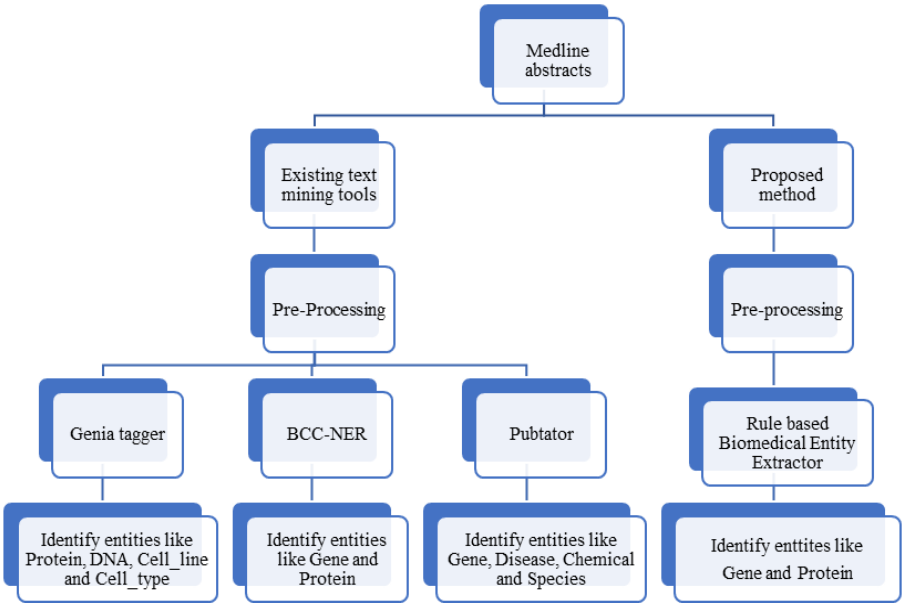

Fig 2. Comparative analysis of text mining tools with the proposed method

The Table V, for the Alzheimer disease abstracts the proposed method identified gene names Biomedical entities in 8, 12, 18 and 25 seconds from 4 different levels of abstracts of size 5,10,15 and 30 respectively where as Pubtator and BCC-NER identified 5,10,15 and 30 seconds from the level of abstracts as mentioned in Table I. Thus, the proposed method achieves good average execution time 16 seconds where as the Pubtator and BCC-NER achieves 15 seconds of both.

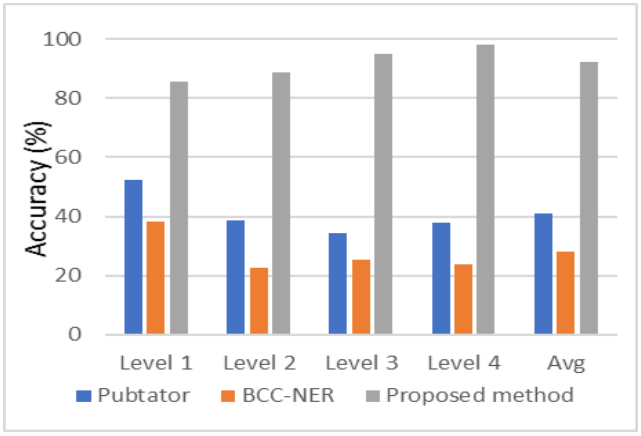

Fig 4. Accuracy for the Colon cancer

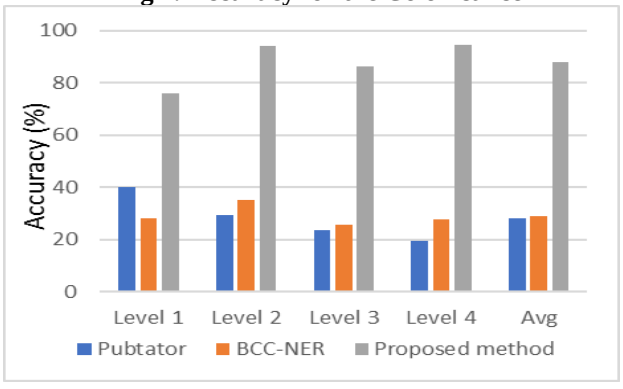

Fig 5. Accuracy for the Alzheimer disease

Table IV Execution time for Colon cancer

Tool / Significance Level (sec)	1	2	3	4	Avg
Pubtator	5	10	15	30	15
BCC-NER	5	10	15	30	15
Proposed Method	8	15	20	35	20

Table V Execution time for Alzheimer disease

Tool / Significance Level (sec)	1	2	3	4	Avg
Pubtator	5	10	15	30	15
BCC-NER	5	10	15	30	15
Proposed Method	8	12	18	25	16

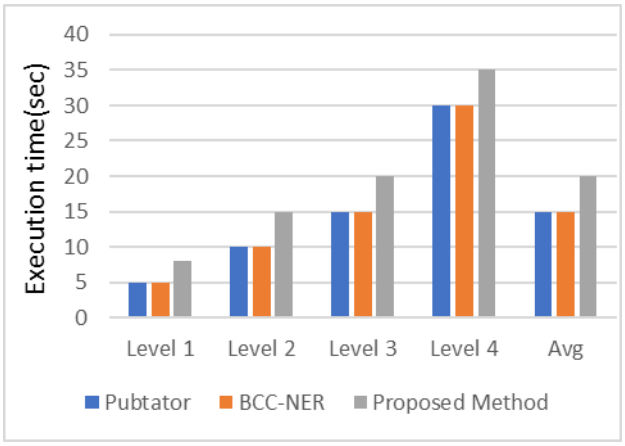

Fig 6. Execution time for the Colon cancer

Retrieval Number: E10550785S319/2019©BEIESP
DOI:10.35940/ijeat.E1055.0785S319
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

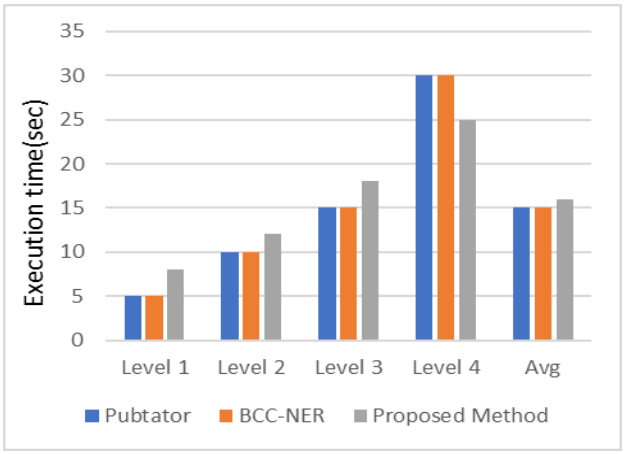

Fig 7. Execution time for the Alzheimer disease

IV. CONCLUSION

The proposed Rule based Biomedical Entity Extractor is deployed with two phases such as Text preprocessing and Identification of Biomedical entities. The proposed method is experimented with the two different kinds of diseases such as Colon Cancer and Alzheimer disease abstracts from Medline, to detect the gene names which belongs to BioMedical entities. The different text mining tools like Pubtator and BCC-NER are employed and obtained the entities from the 30 abstracts which is selected and compared with the Rule based Biomedical Entity Extractor. This paper introduces the comparative study for the text mining tools with the proposed method. The results were compared based on the accuracy and execution time in finding the entities and shows the different levels of significance. The rule based biomedical entity extractor achieves the high accuracy for the abstracts of colon cancer and Alzheimer disease is 92% and 88% respectively. The high accuracy is achieved due to the inclusion of the gene names which are not considered in Pubtator and BCC-NER. The entities obtained from the proposed method has been verified with the benchmark database 'GeneCards database'. This method is found to be suitable, reliable and accuracy for all Medline abstracts. In future, the large number of input abstracts are considered from various databases along with the Medline and the novel or hybridized fast approach will be considered further.

REFERENCES

1. Sakthi Murugan R, P. Shanthi Bala, G. Aghila, "Ontology based information retrieval- an analysis", International journal of advanced research in computer science and software engineering, Vol 3, Issue 10, pp 486-493, 2013.
2. Saurav Sahay, Baoli Li, ernesst V.Garcia, Eugene Agichtein, Ashwin ram, " Domain ontology construction from biomedical text", International Conference on Artificial Intelligence (ICAI'07), Las Vegas, Nevada, USA, 2007.
3. Aarti Singh, Poonam Anand, "Automatic domain ontology construction mechanism", IEEE Recent Advances in Intelligent Computational Systems (RAICS) pp 304-309, 2013.
4. Annalakshmi V, Bhuvaneswari V, Aruna L, "Dictionary Based Approaches in Protein Name Recognition", International Research Journal of Engineering and Technology (IRJET), Vol 04, Issue 02, Feb -2017, pp 94-98, 2014.
5. [5]. Burr Settles, "ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text", *Bioinformatics*, Vol 21, Issue 14, pp 3191–3192, 2005.
6. Gurusamy Murugesan, Sabenabanu Abdulkadhar , Balu Bhasuran and Jeyakumar Natarajan, "BCC-NER: bidirectional, contextual clues named entity tagger for gene/protein mention recognition", EURASIP Journal on Bioinformatics and systems biology, Vol 7, pp 1-8, 2017.
7. Raja K, Subramani S, Natarajan J. "A hybrid named entity tagger for tagging human proteins/genes", International Journal of Data Mining Bioinformatics, Vol 10, Issue 3, 2014.
8. Robert Leaman , Graciela Gonzalez , "Banner: An Executable Survey of Advances in Biomedical Named Entity Recognition ", Pacific Symposium on Biocomputing, Vol 13, pp 652-663, 2008.
9. D Campos, s Matos, JL Oliveira, "Gimli:open source and high-performance biomedical name recognition", BMC Bioinformatics, 14, 2013.
10. Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu "GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains", BioMed Research International Vol 1, 2015.
11. C-H Wei, R Leaman, Z Lu, "SimConcept: A Hybrid Approach for Simplifying Composite Named Entities in Biomedicine", Proceedings of the ACM Conference on Bioinformatics Computational Biology and Health Informatics, Newport Beach, CA, p138-146, 2014.
12. Sohn S., Comeau D.C., Kim W., "Abbreviation definition identification based on automatic precision estimates", BMC Bioinformatics, 9, 402, 2008.
13. R Porkodi and B L Shivakumar, "Design and Development of Integrated Biomedical ontology for information extraction from Medline abstract", International Journal of Engineering Research and Development, Vol 1, Issue 11, pp 01-10, July 2012.
14. Xu Wang, Chen Yang, Renchu Guan, "A Comparative study for biomedical named entity recognition", International Journal of Machine Learning and Cybersecurity, Sep 2015.
15. Tsuruoka Y, "GENIA tagger: Part-of-speech tagging, shallow parsing, and named entity recognition for biomedical text", 2006.
16. J D Kim, T Ohta, Y Tateisi and Tsujii, "GENIA corpus- a semantically annotated corpus for bio-textmining", Bioinformatics, Vol 19, Suppl 1,2003.
17. Chih-Husan Wei, Hung-Yu Kao and Zhiyong Lu, "Pubtator: a web-based text mining tool for assisting biocuration", Nucleic Acids Research, Vol 41, 2013.

AUTHORS PROFILE

G Suganya, received gold madal for her Master degree Computer Science and pursuing PhD in Bharathiar University. Her research interests include Text mining and Bioinformatics.

Dr. **R. Porkodi**, received MCA degree and pursued Ph.D in Bharathiar University. She received UGC grant for her research study. She is the member of many academic bodies. She is the life member in computer society in India and member in IAENG and IACSIT. She acted as a committee member/ resource person/coordinator for various research conferences/events/Faculty development programmes. She published many articles in various reputed journals. Her research interests include Data mining, NLP, Image mining, CBMIR, Hyperspectral remote sensing and Bioinformatics