International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

An Ontology based Online Assistant for Bioinformatics Tools

Dr. Archana Bachhav[1], Dr. Madhukar Shelar[2]

[1]MVP Samaj's KSKW Arts, Science and Commerce College, Nashik, Maharashtra, India.

[2]MVP Samaj's Commerce, Management and Computer Science (CMCS) College, Maharashtra, Nashik, India

Abstract: ***Algorithms are at the core of the whole field of Bioinformatics. The implementations of these algorithms in terms of computer programs are called as bioinformatics tools. Nowadays, biologists use biological data sources and tools to find relevant information for their research. However, with the explosion of the amount of online accessible data and tools, finding the relevant sources and retrieving the relevant information is not an easy task. Both novices and specialists need assistance in navigating the space of possible bioinformatics tools. This research presents the design of an ontology based Online Assistant which can enumerate valid tools for the Bioinformatics processes as well as their information. The Online Assistant for bioinformatics tools will act as a help function to obtain comprehensible information about them. The Online Assistant will also assist in navigating the space for bioinformatics tools The Online Assistant of bioinformatics tools can also provide ranking to the tools according to their comparative parameters i.e. selectivity, sensitivity and speed so as to help the users in the selection of the tool for their tasks.***

Keywords: Bioinformatics, Bioinformatics tools, Ontology, Online Assistant, Bioinformatics algorithms

1. Introduction

1) Bioinformatics

Gerstein et al.(2007) states that "Bioinformatics is conceptualizing biology in terms of molecules (in the sense of physical-chemistry) and then applying "informatics" techniques derived from disciplines such as applied Math, Computer Science and Statistics to understand and organize the information associated with these molecules on a large scale". Bioinformatics is Management Information System (MIS) for Molecular Biology information. The primary goal of bioinformatics is to increase our understanding of biological processes. Bioinformatics would not be possible without advances in computing hardware and software. Fast and high capacity storage media are essential even to maintain the archives. To retrieve and analyze information from those archives, there is need for computer programs. Application of bioinformatics can be looked at the following three levels.

a) To organize biological data to help the researchers to access information, add new information arising from experiments and modify existing information.
b) To develop tools and resources that aid in the analysis of data.
c) To use these tools for analyzing and interpretation of the results in a biologically meaningful manner.
d) The Bioinformatics tools are the software programs for saving, retrieving and analysis of biological data and extracting the information from them. Bioinformatics tools are categorized as follows.
e) Homology and similarity tools
f) Protein function analysis tools
g) Structural analysis tools
h) Sequence analysis tools

There are various tools available for each category. Various organizations implement their own tool for providing better performance by using some different algorithms and statistical methods so that it creates a large space of bioinformatics tools.

2) Ontology

The concept of ontology was first borrowed from Philosophy by Artificial Intelligence researchers and has since become a matter of interest to computer and information scientists, in general. According to Guarion "An ontology is generally regarded as an artifact consisting of a specific shared vocabulary used to describe entities in some domain of interest as well as the set of assumptions about the intended meanings of the terms in vocabulary" (Guarion, 1998). "An ontology is the description of the concepts and relationships that can exist for an agent or a community of agents." (Gruber and Olsen, 1994). Ontology can be used for knowledge sharing and reuse. Noy et al. describes that "Ontologies have become common on the World Wide Web. The ontologies from the web range from large taxonomies categorizing web sites (such as on Yahoo!) to categorization of products and their sale and features (such as on Amazon.com). An Ontology defines a common vocabulary for researchers who need to share information in a domain" (Noy and McGuinness, 2001). They also define the concept of ontology as "It is a formal explicit description of concepts in a domain of discourse (classes (sometimes called concepts)), properties of each concept describing various features and attributes of the concept (slots (sometimes called roles or properties)), and restrictions on slots (facets (sometimes called role restrictions))". Figure 1 shows the structure of an ontology. An ontology for a domain enumerates and gives semantic descriptions of concepts in the domain of discourse, defining domain-relevant attributes of concepts and various relationships among them.

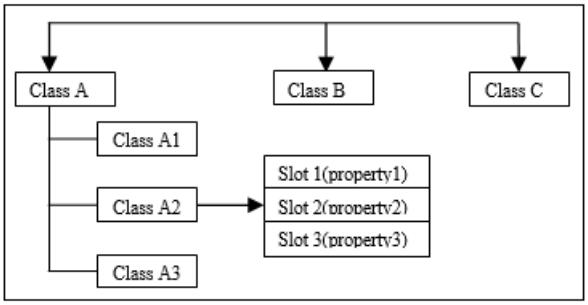

Figure 1: Structure of Ontology

Why should not there be an ontology of bioinformatics tools which will act as a comprehensible help function to navigate the huge space of bioinformatics tools? The concept of semantic web, ontologies of bioinformatics tools make users aware of all the bioinformatics tools and databases and their related information. The vision of a semantic Web alleviates these difficulties. Lambrix P. has stated that "The semantic Web is an extension of the current Web in which information is given a well-defined meaning by annotating Web content with ontology terms" (Lambrix, 2005). So it is necessary to build ontology of bioinformatics tools which will help the biologist as well as novices to navigate the space of bioinformatics tools.

An Ontology is an important technology needed for semantic web vision. An Ontology can be used for knowledge sharing and reuse. There are various concepts related to ontology. Using these concepts one can build the ontology. Some of the reasons for creating the ontologies are as below.

a) To share common understanding of the structure of information among people or Software agents
b) To enable reuse of domain knowledge
c) To make domain assumptions explicit
d) To separate domain knowledge from the operational knowledge
e) To analyze the domain knowledge

The ontology of bioinformatics, **Online Assistant for Bioinformatics Tools** will enumerate valid tools according to their requirements. Database search methods provide a tradeoff between sensitivity, selectivity and speed (Mount, 2004). Sensitivity is picking up even very distant relationships. Selectivity defines all the relationships that reported are true.

From the enumerated bioinformatics tools, the decision of which are to be selected is again critical. The comparative study of pair wise sequence alignment methods (Essoussi and Fayech, 2007), comparison of multiple sequence alignment programs (Diamantis and Anna, 2005) and comparative study performance of protein structure prediction algorithms (Helles, 2008) can be used to provide ranking to these enumerated tools so that user can easily select the bioinformatics tool to meet his objectives.

In this manner, the Online Assistant of Bioinformatics tools can be used to assist the users for navigating the huge space of bioinformatics tools. It will help them for selection of the appropriate tool for their task.

2. Literature Survey

The literatures based on Bioinformatics algorithms are extensively researched and the comparative study is presented according to their process categories. Protein sequence alignment has become an essential task in modern molecular biology research. A number of alignment techniques are available with their corresponding tools freeware. The choice and use of these tools is not trivial for end users with limited skill in Bioinformatics. The initial algorithms were sluggish but produced optimal alignment since they were based on a method called dynamic programming (Reddy, 2020). Pair wise alignment is performed for randomly selected set of sequences for one data set to sequences in other data set using four algorithms – Smith and Waterman, Needleman and Wunsch, FASTA, BLAST (Essoussi and Fayech, 2007). Their performance in terms of execution time is measured and came to the conclusion that BLAST is faster than FASTA. FASTA is faster than Smith and Waterman. Smith and Waterman algorithm is faster than Needleman and Wunsch. That means the heuristics methods are faster than dynamic programming methods (Essoussi and Fayech, 2007).

Similarity searching is used to identify homologies between a query sequence and sequences in a database to elucidate the function of the former by considering the latter. The sensitivity of the search is a measure of how well an algorithm can locate all related or matching sequences in the database.

The BLAST heuristic is probably the most widely used sequence matching method today due primarily to its availability on public servers with graphical interfaces (such as the one at NCBI) and its speed. Many commercial versions are available that are accelerated in some manner. The FASTA heuristic is also used although it is slower than BLAST because it is more sensitive. Both of these methods are based on approximations that aggregate the sequence into tokens prior to the search to reduce the computational complexity (i.e., decrease the time to search). Heuristic techniques like FASTA and BLAST may not always produce the most accurate results, they can nonetheless provide respectably good results quickly (Salomon, 2020). The Smith-Waterman algorithm is an exhaustive search based on Bellman's dynamic programming algorithm and is therefore the most sensitive (and historically slowest) of the three (Muratet, 2002).

Database search methods provide a tradeoff between sensitivity, selectivity and speed (Mount, 2004). So the bioinformatics tools which are implementations of the above homology and sequence similarity algorithms can be ranked with respect to their sensitivity, selectivity and speed.

The speed, selectivity and sensitivity parameters are well defined for popular homology (Jain 2018, Soh 2020) and similarity tools – BLAST, FASTA and MPSrch. So, the

Online Assistant can rank them accordingly as shown in table 1 and comparative study is represented in figure 2

Table 1: Ranking of Homology and Similarity Tools (1-Low, 3-High)

Tool	Speed	Sensitivity	Selectivity
BLAST	3	1	3
FASTA	2	2	2
MPSrch	1	3	1

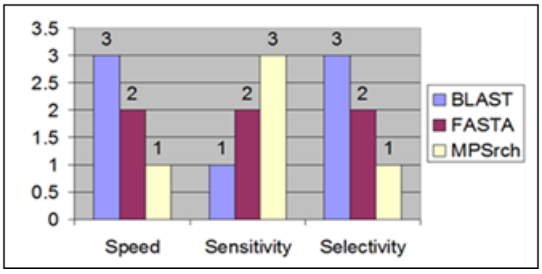

Figure 2: Comparative study of Homology and Sequence Similarity sample tools on various parameters

Multiple sequence alignment (MSA) algorithms are compared in terms of accuracy and speed. Exact algorithms are high quality heuristics but require more time and memory space. Progressive algorithms are widely used as they can align multiple sequences in little time and with less memory. Iterative algorithms are comparatively slower than progressive algorithms. Programs ClustalW and T-coffee which implement the progressive algorithms are widely used tools.

T-Cofee gives more accurate results than ClustalW but it is slower than ClustalW (Lambert, 2019. ClustalW performs well for difficult sequence sets also while T-coffee performs well with sequences with great similarity (Diamantis and Anna, 2005). Here the author has made comparison of 15 well known MSA programs for some selected sets of sequences. Figure 3 shows the comparison of three widely used MSA programs i.e ClustalW, T-coffee and Dalign –T.

Multiple Sequence Alignment programs are the currently available in (Edgar and Batzoglou 2006, Zhang 2022). Here the authors have tested some of the well-known Muli-sequence alignment program using benchmark data set such as BALiBase, PREFAB, OXBENCH, SABmark, IRMBase of reference alignments.

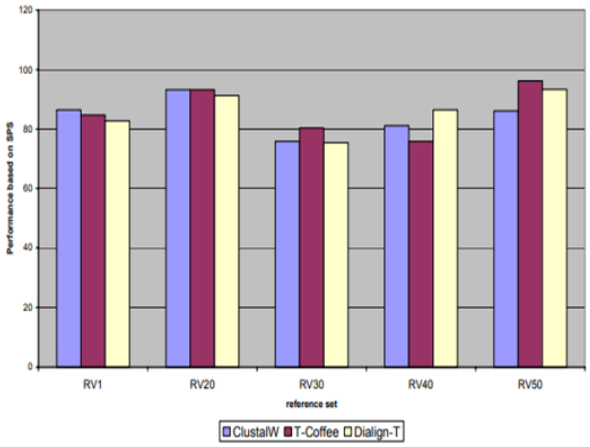

Figure 3: Performance based on SPS, constructed from the benchmark data (Diamantis, 2005)

"Protein structure prediction is one of the major challenges in bioinformatics today. Throughout the past five decades, many different algorithmic approaches have been attempted, and although progress has been made the problem remains unsolvable even for many small proteins. While the general objective is to predict the three-dimensional structure from primary sequence, our current knowledge and computational power are simply insufficient to solve a problem of such high complexity" discussed by G. Helles (2008).

Some prediction algorithms do, however, appear to perform better than others, although it is not always obvious which ones they are and it is perhaps even less obvious why that is (Helles, 2008). Here the author had tested 18 protein prediction algorithms by providing test sequence data sets as input and inferred that different parameters can influence the running time of structure prediction algorithms that are representation of protein, dihedral angle space, energy function, folding strategy and test sets. However, I-TASSER algorithm performed well.

In CASP VII (Critical Assessment for Protein Structure Prediction) competition many protein structure prediction algorithms competed but result is not yet published. Various teams are working yet on comparative study of the protein structure prediction tools and protein function analysis tools.

From the above discussion in this research, the researchers got motivated for the creation of a ranker in the Online Assistant for bioinformatics tools to provide ranks to the enumerated tools which will overcome the confusion of selecting the appropriate bioinformatics tool for the desired task to the user. But as the performance of the homology and similarity searching tools is well defined only and performance for other categories is not well defined yet, so the researchers have given ranking to homology and similarity searching tools only. Here static ranking is given to the tools with respect to the sensitivity, selectivity and speed. But the ranking should be flexible i.e. dynamic means if new tool is added then ranker should consider it while ranking.

Motivation

In this research the researchers had surveyed some of the bioinformatics tools according to their parameters by studying various books. Then the researchers had visited the Bioinformatics department of a college for collection of data related to them by using questionnaires and came to know that most of users in the college do not know many of the tools which are used widely in bioinformatics for various processes.

As mentioned earlier, there are four categories of Bioinformatics Tools and numerous databases and tools are available for each category. Various organizations implement their own tool for providing better performance by using some different algorithms and statistical methods so as it creates a large space of bioinformatics tools in the biomedical field. For example: BLAST (Basic Local Alignment Serach Tool). There are various implementations of this algorithm as follows.

1) BLAST network service on ExPasy
2) BLAST at EMBnet-CH/SIB (Switzerland)
3) BLAST at NCBI
4) WU-BLAST at EBI
5) BLAST at PBIL (Lyon)

These are available on the World Wide Web and again BLAST has various services for different tasks as NBLAST, MEGABLAST, PBLAST, PHI-BLAST, PSI-BLAST. TurboBlast, the parallel implementation of Blast is available to speed up the execution of Blast. The reason behind the unawareness of these tools is that the information about them is very scattered on the Web so there is difficulty in finding them.

Biologists use these data sources and tools to find relevant information for their research.

However, "Successful development of future bioinformatics applications will depend on an appropriately formalized representation of domain of knowledge" (Baldock et al., 2008). With the explosion of the amount of online accessible data and tools, finding the relevant sources and retrieving the relevant information is not an easy task. Further, often, information from different sources needs to be integrated. Though there are large numbers of online accessible tools of bioinformatics available, both novices and experts do not know all of them. They always try to work with traditional and widely used tools.

Many users may ignore most of the bioinformatics tools because they do not have access to the tools easily or because they have complicated installations and execution procedures. They are not aware of the details of those tools. There must be some comprehensive helping function to navigate this huge space of bioinformatics tools which will help the novices as well as experts i.e. biologists. In this research work, the researchers have designed the ontology based Online Assistant which can enumerate valid tools for the bioinformatics processes as well as provide their detail information. The Online Assistant will also assist in navigating the space for bioinformatics tools. From the enumerated bioinformatics tools by the Online Assistant which one is to select is another problem of the users so that it will meet their objectives appropriately. The Online Assistant of bioinformatics tools can provide ranking to the tools according to their comparative parameters i.e selectivity, sensitivity, speed so as to help the users for selection of the tool for their tasks.

Design of Online Assistant

There are various process categories of Bioinformatics tools. Each process imitates different methods or algorithms. These algorithms or methods are implemented through numerous tools creating the large space of them. To search any relevant tool both biologists and novices need some comprehensible help function i.e. Ontology of Bioinformatics Tools. Figure 4 shows the structural view of prototype ontology of Bioinformatics Tools. In this ontology, Bioinformatics process represents a root node; actual tools are shown as leaves.

Figure 5 represents simplified elements of Bioinformatics ontology where process categories such as Homology & Similarity Search, Sequence analysis, protein structure prediction, Function analysis represent classes. Different tools under each process category represent subclasses of process categories. Each tool can be characterized on the basis of their properties or attributes i.e. input format, output format, scoring matrices, availability, URL etc, which represents their slots or properties.

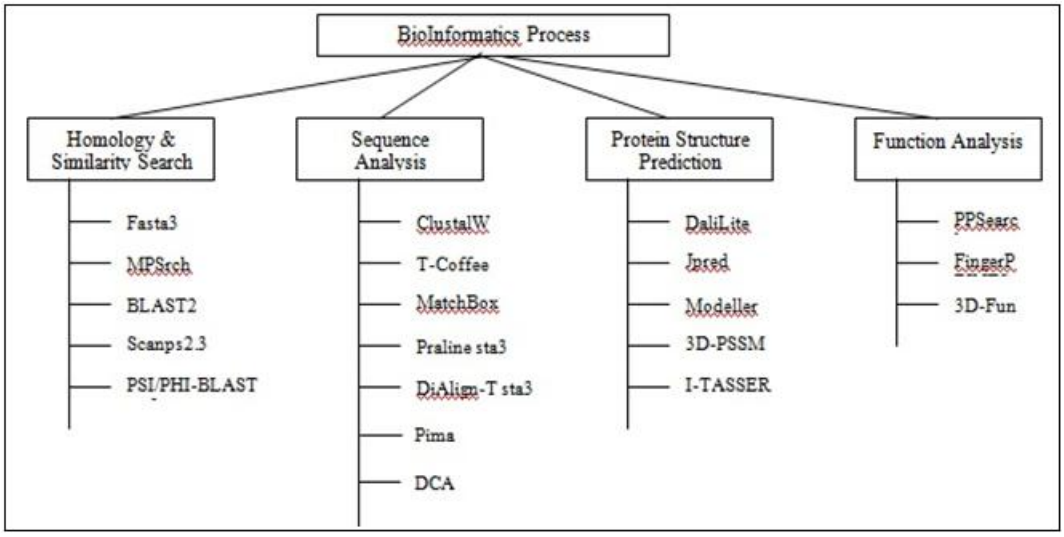

Figure 4: Structural view of prototype

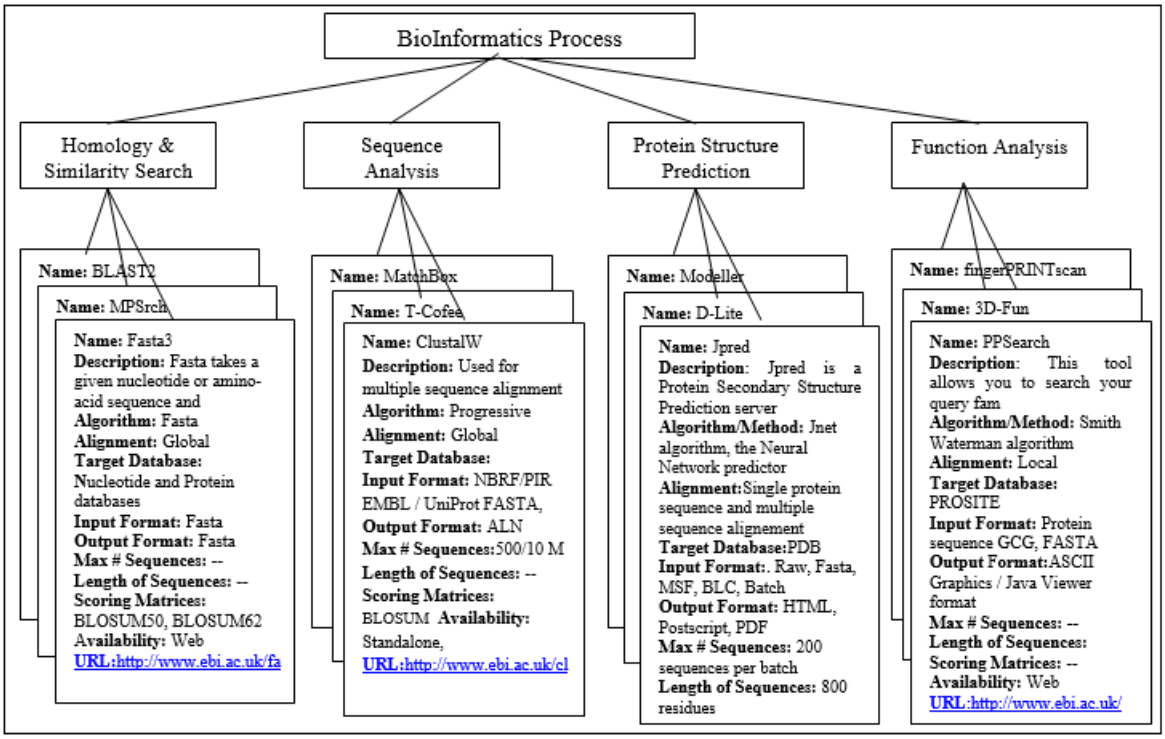

Figure 5: Simplified elements of Bioinformatics tools ontology

3. Limitations and Future Work

In this research, the researchers have designed an Ontology based Online Assistant to navigate the huge space of Bioinformatics tools. It enumerates various appropriate tools as per the requirement of the user. The online assistant for bioinformatics tools also provides the rankings to the enumerated tools with respect to the parameters such as selectivity, sensitivity and speed so that it becomes easy to the users for selection of appropriate one to meet their objectives.

The Online Assistant for Bioinformatics Tools will act as a comprehensible help function to both novices and experts i.e. biologists. The ranking is given to the Bioinformatics tools manually by this Online Assistant and this is a limitation. The ranker is flexible or dynamic i.e. it will take into account, the ranking of the newly added tools through verification by the administrator which is the laborious job. To overcome this problem, the ranker should be autonomous and intelligent. It should provide routine which will calculate sensitivity, selectivity and speed of newly added tools by providing a benchmark dataset as an input in order to provide rank to it.

4. Conclusion

Both novices and specialists of bioinformatics need assistance in navigating the space of possible bioinformatics tools. This Research has characterized many Bioinformatics tools based

on various characteristics such as input and output formats, target database, scoring matrices, maximum length of sequence, number of sequences etc. which is made available on World Wide Web for users through "Online Assistant for Bioinformatics tools". An Ontology based Online Assistant, has been presented, arguing that it can provide information of valid Bioinformatics tools as per the need of users. Further the research has also shown that, how ranking can be provided for the tools with respective to the parameters such as selectivity, sensitivity and speed for Homology and Similarity tools, where values for these parameters for each tool is well defined. The Ranker will help the users to select appropriate tool from the enumerated space by the Online Assistant. Finally, we argue that the Ontology based Online Assistant can be particularly useful as a knowledge sharing environment, creating the network effects as the tool becomes more valuable as it gets more and more users.

After development of Online Assistant for bioinformatics tools, it has been demonstrated to the biotechnology students and teachers and taken feedback from them using questionnaires. They found the Online Assistant is very useful for their studies. Using the Online Assistant, they find searching of any tool's information is very easy. They realized that this ontology based Online Assistant can be used to share knowledge among them.

If any new tool is discovered and add it to Ontology, it can be come to known to all of the users, who only try with traditional tools. Though many users ignore most of the tools because they do not have easy access to them or because of complicated installations and execution procedures, but they will be aware of all the tools with their information through the Online Assistant. The Online Assistant will act as a help function to the users for choosing the appropriate tools for their application.

References

[1] Baldock, R., Burger, A., & Davidson, D. (2008). Anatomy Ontologies for Bioinformatics. Springer.
[2] Diamantis, S., & Anna, C. (2005). Comparison of multiple sequence alignment programs. National and Kapodistrian university of Athens.
[3] Edgar, R. C., & Batzoglou, S. (2006). Multiple sequence alignment. Current opinion in structural biology, 16(3), 368-373.
[4] Essoussi, N., & Fayech, S. (2007). A comparison of four pair-wise sequence alignment methods. Bioinformation, 2(3), 166-168.
[5] Gerstein, M., Greenbaum, D., Cheung, K., & Miller, P. L. (2007). An interdepartmental Ph. D. program in computational biology and bioinformatics: the Yale perspective. Journal of biomedical informatics, 40(1), 73-79.
[6] Gruber, T. R., & Olsen, G. R. (1994, January). An ontology for engineering mathematics. In *Principles of Knowledge Representation and Reasoning* (pp. 258-269). Morgan Kaufmann.
[7] Guarion, N. I. C. O. A. L. (1998). Formal ontology and information systems. In Proceedings of International Conf on Formal Ontology in Information Systems (FOIS'98) (pp. 3-15). Trento, Italy Amsterdam: IOS Press.
[8] Helles, G. (2008). A comparative study of the reported performance of ab initio protein structure prediction algorithms. Journal of the royal society interface, 5(21), 387-396.
[9] Jain, C., Koren, S., Dilthey, A., Phillippy, A. M., & Aluru, S. (2018). A fast adaptive algorithm for computing whole-genome homology maps. Bioinformatics, 34(17), i748-i756.
[10] Lambert, M. È., Arsenault, J., Delisle, B., Audet, P., Poljak, Z., & D'Allaire, S. (2019). Impact of alignment algorithm on the estimation of pairwise genetic similarity of porcine reproductive and respiratory syndrome virus (PRRSV). BMC veterinary research, 15, 1-10.
[11] Lambrix, P. (2005, June). Towards a semantic web for bioinformatics using ontology-based annotation. In 14th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise (WETICE'05) (pp. 3-7). IEEE.
[12] Mount, D. M. Bioinformatics: sequence and genome analysis 2004 2 Cold Spring Harbor.
[13] Muratet, M. A. (2002). Comparing the speed and accuracy of the Smith and Waterman algorithm as implemented by MPSRCH with the BLAST and FASTA heuristics for sequence similarity searching. TheScientificWorldJOURNAL, 2, 21-22.
[14] Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology.
[15] Reddy, B., & Fields, R. (2020, March). Multiple Anchor Staged Alignment Algorithm–Sensitive (MASAA–S). In 2020 3rd International Conference on Information and Computer Technologies (ICICT) (pp. 361-365). IEEE.
[16] Salomon, M. (2020). Blast Compatible Non-heuristic Biological Sequence Alignment on Hetereogenous Systems. California State University, Fresno.
[17] Soh, J. H., Chan, H. M., & Ying, J. Y. (2020). Strategies for developing sensitive and specific nanoparticle-based lateral flow assays as point-of-care diagnostic device. Nano Today, 30, 100831.
[18] Subramanian, A. R., Weyer-Menkhoff, J., Kaufmann, M., & Morgenstern, B. (2005). DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. BMC bioinformatics, 6(1), 1-13.
[19] Zhang, H., Hao, M., Wu, H., Ting, H. F., Tang, Y., Xi, W., & Wei, Y. (2022). Protein residue contact prediction based on deep learning and massive statistical features from multi-sequence alignment. Tsinghua Science and Technology, 27(5), 843-854