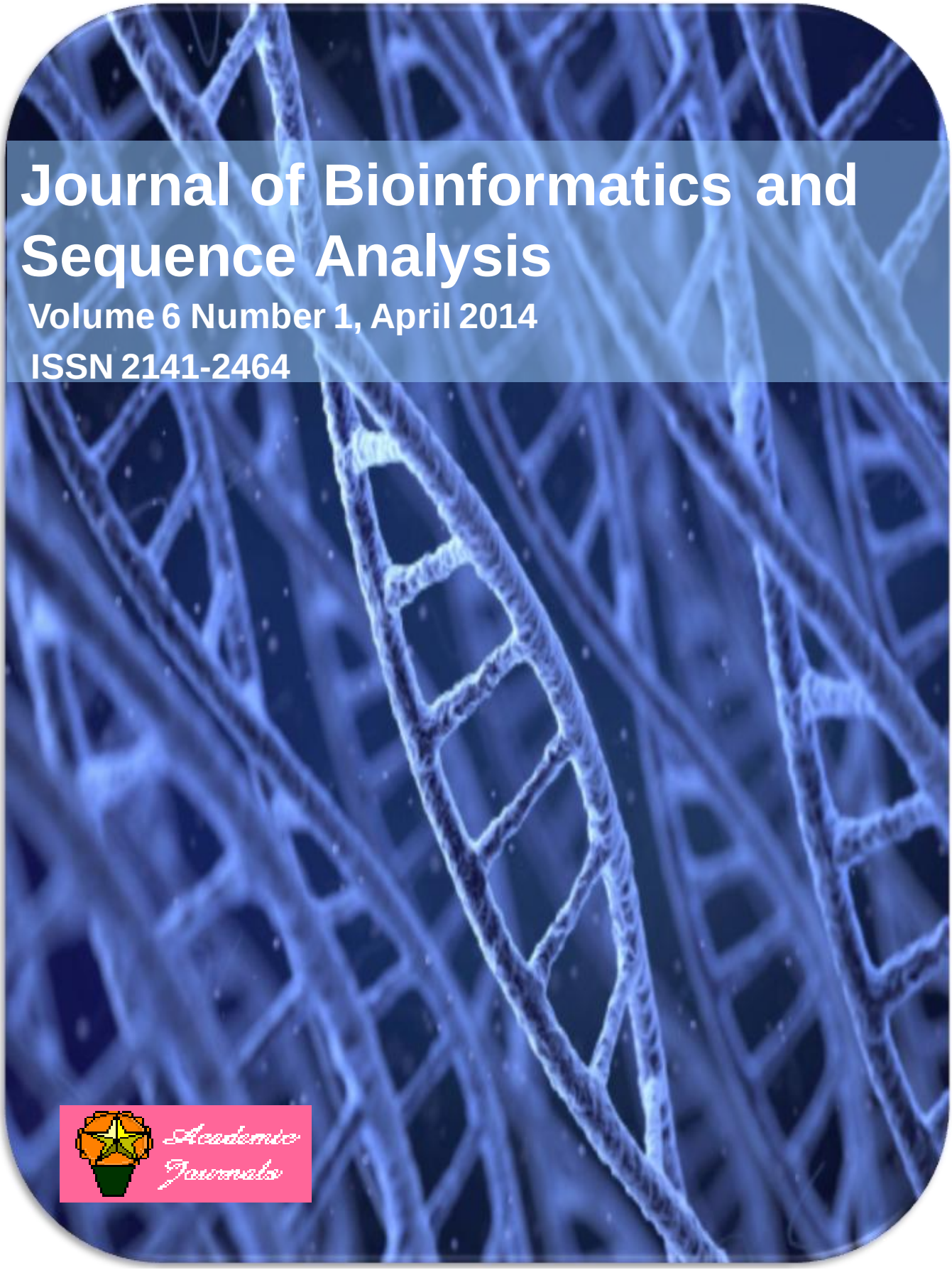


Journal of Bioinformatics and Sequence Analysis

Volume 6 Number 1, April 2014
ISSN 2141-2464

Academic Journals

ABOUT JBSA

The **Journal of Bioinformatics and Sequence Analysis (JBSA)** (ISSN 2141-2464) is published per article (one volume per year) by Academic Journals.

Journal of Bioinformatics and Sequence Analysis (JBSA) provides rapid publication (Per Article) of articles in all areas of the subject such as In silico effective inhibition of galtifloxacin on built Mtb-DNA gyrase, In silico sequence specific analysis of ERBB2 RTK alterations responsible for neuroectodermal tumors of Homo sapiens, Structural characterization of pfSerine hydroxymethyltransferase, Information fusion and multiple classifiers for haplotype assembly problem from SNP fragments and related genotype etc.

The Journal welcomes the submission of manuscripts that meet the general criteria of significance and scientific excellence. Papers will be published shortly after acceptance. All articles published in JBSA are peer-reviewed.

Submission of Manuscript

Please read the **Instructions for Authors** before submitting your manuscript. The manuscript files should be given the last name of the first author

Click here to Submit manuscripts online

If you have any difficulty using the online submission system, kindly submit via this email jbsa@academicjournals.org.

With questions or concerns, please contact the Editorial Office at jbsa@academicjournals.org.

Editorial Board

Instructions for Author

Electronic submission of manuscripts is strongly encouraged, provided that the text, tables, and figures are included in a single Microsoft Word file (preferably in Arial font).

The **cover letter** should include the corresponding author's full address and telephone/fax numbers and should be in an e-mail message sent to the Editor, with the file, whose name should begin with the first author's surname, as an attachment.

Article Types
Three types of manuscripts may be submitted:

Regular articles: These should describe new and carefully confirmed findings, and experimental procedures should be given in sufficient detail for others to verify the work. The length of a full paper should be the minimum required to describe and interpret the work clearly.
Short Communications: A Short Communication is suitable for recording the results of complete small investigations or giving details of new models or hypotheses, innovative methods, techniques or apparatus. The style of main sections need not conform to that of full-length papers. Short communications are 2 to 4 printed pages (about 6 to 12 manuscript pages) in length.

Reviews: Submissions of reviews and perspectives covering topics of current interest are welcome and encouraged. Reviews should be concise and no longer than 4-6 printed pages (about 12 to 18 manuscript pages). Reviews are also peer-reviewed.

Review Process

All manuscripts are reviewed by an editor and members of the Editorial Board or qualified outside reviewers. Authors cannot nominate reviewers. Only reviewers randomly selected from our database with specialization in the subject area will be contacted to evaluate the manuscripts. The process will be blind review.
Decisions will be made as rapidly as possible, and the journal strives to return reviewers' comments to authors as fast as possible. The editorial board will re-review manuscripts that are accepted pending revision. It is the goal of the AJFS to publish manuscripts within weeks after submission.

Regular articles

All portions of the manuscript must be typed double-spaced and all pages numbered starting from the title page.

The Title should be a brief phrase describing the contents of the paper. The Title Page should include the authors' full names and affiliations, the name of the corresponding author along with phone, fax and E-mail information. Present addresses of authors should appear as a footnote.

The Abstract should be informative and completely self-explanatory, briefly present the topic, state the scope of the experiments, indicate significant data, and point out major findings and conclusions. The Abstract should be 100 to 200 words in length.. Complete sentences, active verbs, and the third person should be used, and the abstract should be written in the past tense. Standard nomenclature should be used and abbreviations should be avoided. No literature should be cited.
Following the abstract, about 3 to 10 key words that will provide indexing references should be listed.

A list of non-standard **Abbreviations** should be added. In general, non-standard abbreviations should be used only when the full term is very long and used often. Each abbreviation should be spelled out and introduced in parentheses the first time it is used in the text. Only recommended SI units should be used. Authors should use the solidus presentation (mg/ml). Standard abbreviations (such as ATP and DNA) need not be defined.

The Introduction should provide a clear statement of the problem, the relevant literature on the subject, and the proposed approach or solution. It should be understandable to colleagues from a broad range of scientific disciplines.

Materials and methods should be complete enough to allow experiments to be reproduced. However, only truly new procedures should be described in detail; previously published procedures should be cited, and important modifications of published procedures should be mentioned briefly. Capitalize trade names and include the manufacturer's name and address. Subheadings should be used. Methods in general use need not be described in detail.

Results should be presented with clarity and precision. The results should be written in the past tense when describing findings in the authors' experiments. Previously published findings should be written in the present tense. Results should be explained, but largely without referring to the literature. Discussion, speculation and detailed interpretation of data should not be included in the Results but should be put into the Discussion section.

The Discussion should interpret the findings in view of the results obtained in this and in past studies on this topic. State the conclusions in a few sentences at the end of the paper. The Results and Discussion sections can include subheadings, and when appropriate, both sections can be combined.

The Acknowledgments of people, grants, funds, etc should be brief.

Tables should be kept to a minimum and be designed to be as simple as possible. Tables are to be typed double-spaced throughout, including headings and footnotes. Each table should be on a separate page, numbered consecutively in Arabic numerals and supplied with a heading and a legend. Tables should be self-explanatory without reference to the text. The details of the methods used in the experiments should preferably be described in the legend instead of in the text. The same data should not be presented in both table and graph form or repeated in the text.

Figure legends should be typed in numerical order on a separate sheet. Graphics should be prepared using applications capable of generating high resolution GIF, TIFF, JPEG or Powerpoint before pasting in the Microsoft Word manuscript file. Tables should be prepared in Microsoft Word. Use Arabic numerals to designate figures and upper case letters for their parts (Figure 1). Begin each legend with a title and include sufficient description so that the figure is understandable without reading the text of the manuscript. Information given in legends should not be repeated in the text.

References: In the text, a reference identified by means of an author‘s name should be followed by the date of the reference in parentheses. When there are more than two authors, only the first author‘s name should be mentioned, followed by ’et al‘. In the event that an author cited has had two or more works published during the same year, the reference, both in the text and in the reference list, should be identified by a lower case letter like ’a‘ and ’b‘ after the date to distinguish the works.

Examples:

Abayomi (2000), Agindotan et al. (2003), (Kelebeni, 1983), (Usman and Smith, 1992), (Chege, 1998; 1987a,b; Tijani, 1993,1995), (Kumasi et al., 2001)
References should be listed at the end of the paper in alphabetical order. Articles in preparation or articles submitted for publication, unpublished observations, personal communications, etc. should not be included in the reference list but should only be mentioned in the article text (e.g., A. Kingori, University of Nairobi, Kenya, personal communication). Journal names are abbreviated according to Chemical Abstracts. Authors are fully responsible for the accuracy of the references.

Examples:

Chikere CB, Omoni VT and Chikere BO (2008). Distribution of potential nosocomial pathogens in a hospital environment. Afr. J. Biotechnol. 7: 3535-3539.

Moran GJ, Amii RN, Abrahamian FM, Talan DA (2005). Methicillinresistant Staphylococcus aureus in community-acquired skin infections. Emerg. Infect. Dis. 11: 928-930.

Pitout JDD, Church DL, Gregson DB, Chow BL, McCracken M, Mulvey M, Laupland KB (2007). Molecular epidemiology of CTXM-producing Escherichia coli in the Calgary Health Region: emergence of CTX-M-15-producing isolates. Antimicrob. Agents Chemother. 51: 1281-1286.

Pelczar JR, Harley JP, Klein DA (1993). Microbiology: Concepts and Applications. McGraw-Hill Inc., New York, pp. 591-603.

Short Communications

Short Communications are limited to a maximum of two figures and one table. They should present a complete study that is more limited in scope than is found in full-length papers. The items of manuscript preparation listed above apply to Short Communications with the following differences: (1) Abstracts are limited to 100 words; (2) instead of a separate Materials and Methods section, experimental procedures may be incorporated into Figure Legends and Table footnotes; (3) Results and Discussion should be combined into a single section.
Proofs and Reprints: Electronic proofs will be sent (e-mail attachment) to the corresponding author as a PDF file. Page proofs are considered to be the final version of the manuscript. With the exception of typographical or minor clerical errors, no changes will be made in the manuscript at the proof stage.

Fees and Charges: Authors are required to pay a $550 handling fee. Publication of an article in the Journal of Bioinformatics and Sequence Analysis is not contingent upon the author's ability to pay the charges. Neither is acceptance to pay the handling fee a guarantee that the paper will be accepted for publication. Authors may still request (in advance) that the editorial office waive some of the handling fee under special circumstances

Submission of a manuscript implies: that the work described has not been published before (except in the form of an abstract or as part of a published lecture, or thesis) that it is not under consideration for publication elsewhere; that if and when the manuscript is accepted for publication, the authors agree to automatic transfer of the copyright to the publisher.

Journal of Bioinformatics and Sequence Analysis

Table of Contents: Volume 6 Number 1 April, 2014

ARTICLES

academicJournals

Vol. 6(1), pp. 1-6, April 2014
DOI: 10.5897/IJBC2013.0086
Article Number: 093849744377
ISSN 2141-2464

http://www.academicjournals.org/JBSA

Journal of Bioinformatics and Sequence Analysis

Review

Bioinformatics with basic local alignment search tool (BLAST) and fast alignment (FASTA)

Eric S. Donkor[1]*, Nicholas T. K. D. Dayie[1,2] and Theophilus K Adiku[1]

[1]Department of Microbiology, University of Ghana Medical School, Accra, Ghana.
[2]Department of Veterinary Disease Biology, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark.

Received 27 April, 2013: Accepted 25 April, 2014

Following advances in DNA and protein sequencing, the application of computational approaches in analysing biological data has become a very important aspect of biology. Evaluating similarities between biological sequences is crucial to our understanding of evolutionary biology, and this can be achieved by basic local alignment search tool (BLAST) and fast alignment (FASTA). BLAST and FASTA have become fundamental tools of biology and it is essential to know how they operate, the task they can accomplish and how to accurately interpret their output. This paper provides an analysis of BLAST and FASTA in sequence analysis. Both BLAST and FASTA algorithms are appropriate for determining highly similar sequences. However, BLAST appears to be faster and also more accurate than FASTA. Both BLAST and FASTA are limited in sensitivity and may not be able to capture highly divergent sequences in some cases. Consequently, evolutionarily diverse members of a family of proteins may be missed out in a BLAST or FASTA search.

Key words: Bioinformatics, basic local alignment search tool (BLAST), fast alignment (FASTA), sequence alignment, prokaryotes.

INTRODUCTION

The term bioinformatics was coined by Paulien Hogeweg of Utrecht University in 1979 for the study of informatic processes in biotic systems, but the field of bioinformatics did not become recognized until the 1990s (Hogeweg, 1978; Luscombe et al., 2001). Currently, bioinformatics is defined in many ways and there is no consensus definition. Perhaps one of the most appropriate definitions is that proposed by the National Institute of Health, USA which states that bioinformatics refers to "research, development or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data including those to acquire, store, organize, archive, analyse or visualize such data" (National Institutes of Health, 2010). This definition identifies three-fold objectives of bioinformatics. Firstly, bioinformatics organizes data in a way that permits researchers to access existing information and also to make submission of new entries to databases

*Corresponding author. E-mail: esampane-donkor@chs.ug.edu.gh.

such as Protein Data Bank. Secondly, bioinformatics seeks to develop tools for analysis of data, and thirdly, to use these tools to analyze and interpret data results, which is the focus of the current review.

In recent times, application of computational approaches to biological data has become a vital part of biology, particularly in the analysis of protein and DNA sequences. Evaluating similarities between biological sequences is probably the major means by which bioinformatics contributes to our understanding of biology (Pearson and Lipman, 1988; Bansal and Meyer, 2002). The most common bioinformatic tools for executing this purpose are basic local alignment search tool (BLAST) and fast alignment (FASTA), which perform comparisons between pairs of sequences, based on regions of local similarity (Altschul et al., 1990; Pearson and Lipman, 1988; Luscombe et al., 2001). BLAST and FASTA have therefore become fundamental tools of biology and it is essential to know how they operate, the tasks they can accomplish and how to accurately interpret their output.

Though there is considerable amount of literature on sequence analysis and database searching tools, generally the literature is unsuitable for beginners in the field, as the style of communication is highly advanced and rather targets expert bioinformaticians. With the increasing usage of BLAST and FASTA by non-bioinformaticians, there is the need for more basic review articles on the subject. This paper provides an analysis of the use of BLAST and FASTA in sequence analysis, and it is particularly targeted at beginners in the field of bioinformatics.

DNA AND PROTEIN SEQUENCES: PRIMARY STRUCTURE

DNA has a primary structure that arises from the directional polymerisation of single nucleotide units. Each nucleotide unit of a DNA molecule comprises a deoxyribose sugar, a nitrogenous base (adenine, guanine, cytosine, thymine) and a phosphate group. The nucleotides units are linked by phosphoester bonds which occur between 5' and 3' carbon atoms. DNA sequencing entails determining the precise order of nucleotides in a DNA molecule, and therefore the primary structure. DNA sequencing started with the basic sequencing methods developed by Maxam and Gilbert (1977) and also Sanger et al. (1977). Currently, DNA sequencing has attained high level of technological advancement with the so called next generation sequencing technologies which are high throughput (Mardis, 2008).

The primary structure of a protein is related to its sequence of amino acids linked through peptide bonds that form the covalent backbone of the proteins, and it includes disulphide bonds, if they are present (Orengo et al., 1999). The sequence of amino acids is read from the N-terminal amino acid to the C-terminal amino acid. There are twenty known amino acids and a polypeptide chain comprises a number of certain types of amino acids arranged in a definite sequence. This indicates that they could be a great diversity of possible protein sequences. In general, the primary structure of a protein contains all the necessary information required for the manifestation of higher, three-dimensional levels of structure and function (Orengo et al., 1999). Traditionally, amino acid sequences of proteins have been determined directly by the Edman degradation reaction (Niall, 1973). The other major direct method by which the sequence of a protein can be determined is mass spectrometry (Dhaunta et al., 2010). The amino acid sequence of a protein can also be determined indirectly from the DNA sequence of prokaryotes, but in the case of eukaryotes it may be complicated due to the presence of introns in the genome (Alberts, 2002; Lynch, 2002). Practically, primary structure of a protein is more easily determined by interpreting a gene sequence of nucleotides (with reference to the genetic code), than directly from a purified protein itself.

SEQUENCE ALIGNMENT AND ITS SIGNIFICANCE

Sequence alignment is the process of comparing different sequences by searching for a series of individual characters or character patterns that have the same arrangement in both sequences (Pearson and Lipman, 1988). There are three main types of sequence alignments: pairwise sequence alignment, multiple sequence alignment and structural sequence alignment (Pearson and Lipman, 1988; Luscombe et al., 2001). Pairwise sequence alignment can only be used between two sequences at a time. Multiple sequence alignment is an extension of pairwise alignment incorporating more than two sequences at a time. A structural sequence alignment analyzes the whole structure of a protein strand, unlike pairwise and multiple sequence alignments, and must be visualized three-dimensionally. Sequence alignments are either local or global. While local alignments finds the best match between two sequences, global alignments find the best match over the total lengths of the different sequences involved in the alignment. Most sequence alignments done are of pairwise alignments and are based on local alignments (Pearson and Lipman, 1988; Altschul et al., 1997). There are three primary methods of producing such pairwise alignments, and they are the dot-matrix method, dynamic programming and word methods (Mount, 2004).

To achieve the best possible alignment for two sequences, it is essential to include gaps in sequence alignments and use gap penalties (Mount, 2008). For a given alignment, a gap refers to any maximal, uninterrupted run of spaces in a single sequence. The concept of gaps in sequence alignments is essential,

since gaps account for indels that may appear in related DNA or protein sequences. Gaps are normally penalized by means of a linear gap function which assigns an initial penalty for a gap opening, and also an extra penalty for gap extensions that widen the gap length (Mount, 2008).

The objective of a sequence alignment is to identify similarity of the aligned sequences which may be a result of structural, functional or evolutionary relationships between the sequences (Pearson and Lipman, 1988; Mount, 2004). Similarity finding based on conserved sequence motif can be utilized in conjunction with mechanistic and structural information to identify catalytic site of enzymes (Altschul et al., 1997; Luscombe et al., 2001). Pertsemlidis and Fondon (2010) distinguished among three terminologies commonly used in sequence alignment, which have been abused in their usage. These terminologies include sequence identity, similarity and homology. Sequence identity refers to exactly the same position distribution of nucleotide or amino acid in aligned sequences. Sequence similarity takes approximate matches into consideration, and for this to be meaningful, there should be some scoring of such substitutions, with conservative substitutions assigned better scores than non-conservative ones. Homology strictly refers to the situation where nucleotide or amino acid sequences are similar because they have a common evolutionary origin. The term is often used loosely to indicate that sequences are very similar. Pertsemlidis and Fondon (2010) further indicate that although the comparison of two sequences is often presented as a percentage sequence homology, that usage is inaccurate as the value actually reflects identity and/or similarity, and does not necessarily indicate an evolutionary relationship.

SEQUENCE ALIGNMENT AND DATABASE SEARCHING TOOLS

Basic local alignment search tool (BLAST)

Background and types of the BLAST programme

BLAST is an algorithm used for comparison of amino acid sequences of different proteins or the nucleotides sequences of nucleic acid. BLAST was invented in 1990 and has since then become the defacto standard in search and alignment tools (Altschul et al., 1990). Through a BLAST search, one can compare a query sequence with a database of sequences, and thereby identify library sequences that share resemblance with the query sequence above a certain threshold. Based on such comparison, BLAST can be used to achieve several objectives including species identification, locating domains, DNA mapping and annotation (Altschul et al., 1990). There are several different types of BLAST programs available, and the choice of a BLAST programme depends on one's objective and the type of sequences being investigated. The most commonly used BLAST programmes are shown in Table 1.

Table 1. BLAST programmes that are commonly used.

Program	Query sequence type	Target sequence type
BLASTP	Protein	Protein
BLASTN	Nucleotide	Nucleotide
BLASTX	Nucleotide (translated)	Protein
TBLASTN	Protein	Nucleotide (translated)
TBLASTX	Nucleotide (translated)	Nucleotide (translated)

Source: http://www.ncbi.nlm.nih.gov/blast.

Apart from these BLAST programmes, relatively more recent BLAST programmes such as position-specific iterative (PSI) BLAST have been developed with improved sensitivity (Altschul et al., 1997; http://www.ncbi.nlm.nih.gov/blast).

How BLAST works

BLAST is based on a heuristic algorithm (Altschul et al., 1990). A heuristic algorithm is an algorithm that provides almost the correct answer or a solution for some instances of the problem. Through a heuristic approach, BLAST identifies homologous sequences by locating short matches between the two sequences being compared. This process is referred to as seeding, and it is after this initial match that BLAST begins to make local alignments. During the process of seeding, BLAST tries to locate all common three-letter words between the sequence of interest and the hit sequence, or sequences, from the database. In this context, a word is simply defined as a number of letters. For example, for blastp, the default word size is 3 W=3. If a query sequence has ABCDE, the searched words are ABC, BCD, CDE. After synthesizing words for a given sequence of interest, neighborhood words are also assembled. Once both words and neighborhood words are organized, they are compared with the database sequences in order to find matches. The alignment, which is normally 3 residues long, is extended in either direction by the BLAST algorithm. Each extension increases or decreases the score of the alignment, and should the score be higher than a pre-determined threshold, the alignment will be included in the results given by BLAST. However, should this score fall below the pre-determined threshold, the alignment will stop extending, thereby blocking areas of poor alignment to be included in the BLAST results. The detailed statistical aspects involved in the BLAST algorithm are described in a publication by Mount (2004).

A BLAST output can easily be generated by submitting a query sequence at the NCBI site http://blast.ncbi.nlm.nih.gov/. The output from a BLAST search consists of four parts. The first is the header

which is about the descriptions of the BLAST program used, the query sequence and the target database; the second part of the output is a list of sequences showing significant alignments, along with both normalized scores and expect (E) values; the third part summarizes the alignments and related statistical information, including the raw and bit scores, E value, and identity level, for each high-scoring alignment. The fourth part displays all of the parameters used in the BLAST search.

E values represent better statistical indicators of how significant a particular match is. By definition, the E value is equivalent to the number of sequences occurring in the database, that is expected to match a given query sequence at least as well as the listed sequence does, if the relationship between the sequences was random.

FASTA

Background

FASTA was invented in 1995 based on an improvement in FASTP, another sequence alignment tool invented in 1985 (Lipman and Pearson, 1985; Pearson, 1990). FASTP was used for protein similarity searching, however, its improvement in FASTA empowered it to execute DNA:DNA searches, translated protein:DNA searches, and also provided a more robust program for evaluating statistical significance (Pearson and Lipman, 1988, Mount, 2004). There are several different types of FASTA including TFASTAX, TFASTAY, FASTAX and FASTAY (http://www.ebi.ac.uk/Tools/fasta/index.html). TFASTA and TFASTAY handles query protein against a DNA library in all reading frames. FASTAX and FASTAY handle DNA query in all reading frames against a protein database.

How FASTA works

Like BLAST, The FASTA program is based on a heuristic algorithm (Pearson and Lipman, 1988). FASTA sets a certain size k for k-tuple subwords (ordered set of k values). The program then searches for diagonals in the comparison matrix of the query and search sequence along which many k-tuples match. It then re-scores the highest scoring regions with the aid of a replacement matrix (a matrix that shows the rate at which a particular character in a sequence changes to other character states over time) such as the PAM250. After this, it attempts to join together the high scoring diagonals, allowing for gaps. Finally, it makes an optimal local alignment around the regions it has discovered based on the Smith-Waterman algorithm.

An alternative way FASTA works is described as follows (Pearson and Lipman, 1988). In a first step FASTA tries to identify regions shared by the two sequences that have the highest density of single residue identities (ktup=1) or two-consecutive identities (ktup=2). In a second step, it then re-scans the best regions identified in the first step using the PAM-250 matrix. After that it determines if gaps can be used to join the regions identified in second step. If so, a similarity score for the gapped alignment is evaluated. Finally, it constructs an optimal alignment of the query sequence and the library sequence based on Smith-Waterman algorithm. The detailed statistical aspects involved in the FASTA algorithm have been described by Lipman and Pearson (1985).

A FASTA output can be generated easily be submitting a query sequence (in a FASTA format) at a database such as UniProt (http://www.uniprot.org/). The output from FASTA is divided into four parts. The first part has some information on the database searched and the query sequence submitted; the second part is a histogram display, which reports graphically the score distribution; the third is a list of matched sequences, and related statistical information, and fourth, the alignments themselves are displayed.

BLAST vs. FASTA in terms of precision, accuracy and algorithm runtime complexity

BLAST and FASTA are similar in that both programs are based on a common assumption that true matches are likely to have at least some short stretches of high-scoring similarity, but whereas FASTA targets exactly matching 'words' (strings of residues), BLAST employs a scoring matrix - BLOSUM62 for amino-acid sequences (Pearson and Lipman, 1988; Altschul et al., 1990; Pertsemlidis and Fondon, 2010).

It is worthwhile defining the terminologies precision, accuracy and algorithm runtime complexity in the context of bioinformatics software applications, as they form the basis of comparison of BLAST and FASTA in this section. Algorithm runtime complexity refers to how fast the bioinformatics tool performs and produces a results output. Accuracy is the degree of closeness of a measurement to its actual or true value. The precision of a measurement system is the extent to which repeated measurements under unvaried conditions yield the same results. It is important to note that accuracy of a bioinformatics tool can be evaluated by its sensitivity (ability to identify or recognize distantly related sequences), and then accuracy and precision may be used interchangeably.

In terms of algorithm runtime complexity, BLAST is faster than FASTA by searching for only the more significant patterns in the sequences. The sensitivity (or accuracy) of BLAST and FASTA tends to be different for nucleic acid and protein sequences (http://www.bioinfo.se/kurser/swell/blasta-fasta.shtml).

For protein alignments, the BLAST heuristic algorithm is more sensitive than that of FASTA, although BLAST employs a word size of three for proteins while FASTA works with a word size of two. In the case of nucleic acid sequences, BLAST employs a long word size of eleven. However, the heuristic algorithm that improves the sensitivity for protein sequences does not work as well for nucleic acid (Li et al., 2004), and FASTA is more sensitive than BLAST for nucleic acid sequences. BLAST and FASTA differ is the statistical evaluation of their output which would most likely affect their relative accuracy. FASTA produces an E-value that shows relatively accurate estimation for found matches and the expectancy to find them by chance (Brenner et al., 1998). On the other hand, BLAST calculates expectancy by removing the results scored lower than the threshold value (Brenner et al., 1998; Sansom, 2000). This results in elimination of statistically not significant alignments increasing the accuracy of BLAST over FASTA (Sansom, 2000). Despite the high numbers of citations of BLAST and FASTA in literature, there is hardly any quantitative comparison of the two tools in terms of speed, precision and accuracy. A comparative analysis on the algorithm runtime complexity and precision of BLAST and FASTA showed that BLAST was over six times faster for searching structural classification of proteins (SCOP) than FASTA (Chattaraj et al., 1999). However, the average precision of FASTA was about 2% higher than that of BLAST (Chattaraj et al., 1999). In the same study, it was also observed that Psi-BLAST is almost an order of magnitude slower than BLAST but over 3% more accurate in average precision (Chattaraj et al., 1999). It is important to note that both FASTA and BLAST allow gaps at some point in their mechanism of operation and for this reason, both methods have the potential to miss significant similarities present in the database. Another setback of both search tools is that many proteins are multifunctional multi-domain proteins, and thus a high hit to one domain does not necessarily define function.

CONCLUSIONS

Though BLAST and FASTA algorithms are suitable for determining highly similar sequences, BLAST is known to have a relatively greater speed (shorter algorithm runtime complexity) than FASTA and should therefore be the program of choice in initial database searches. However, due to the relatively higher sensitivity of FASTA in some cases, it should be included in an advanced search allowing for a comparison of results. In some cases, BLAST and FASTA are not sensitive enough to capture highly divergent sequences. Thus, evolutionarily diverse members of a family of proteins may be missed out in a BLAST or FASTA search. More studies are needed to compare BLAST and FASTA.

ACKNOWLEDGEMENT

We thank the authors of the various papers cited in this review article.

Conflict of interest

None declared.

REFERENCES

Alberts B (2002). Molecular biology of the cell. New York: Garland Science. p.760.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic alignment search tools. J. Mol. Biol. 215:403-410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389-3402.

Bansal AK, Meyer TE (2002). Evolutionary analysis by whole genome comparisons. J. Bact. 184(8):2260-2272.

Brenner SE, Chothia C, Hubbard TJP (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc. Natl Acad. Sci. USA, 95:6073-6078.

Chattaraj A, Williams HE, Cannane A (1999). Fast Homology Search using Categorization Profiles. RMIT University, Melbourne. http://www.jsbi.org/pdfs/journal1/GIW04/GIW04P085.pdf accessed on 15/04/2013.

Dhaunta N, Fatima U, Guptasarma P (2010). N-Terminal sequencing by mass spectrometry through specific fluorescamine labelling of α-amino groups before tryptic digestion. Anal. Biochem. 408(2):263-268.

Hogeweg P (1978). Simulating the growth of cellular forms. *Simulation* 31:90-96.

Lipman DJ, Pearson WR (1985). Rapid and sensitive protein similarity searches. *Science,* 227:1435–1441.

Li M, Ma B, Kisman D, Tromp J (2004). PatternHunter II: Highly sensitive and fast homology search. J. Bioinform. Comput. Biol. 2(3):417-439.

Luscombe NM, Greenbaum D, Gerstein M (2001). What is bioinformatics? A proposed definition and overview of the field. Methods Inf. Med. 40(4):346-358.

Lynch M (2002). Intron evolution as a population-genetic process. Proc Natl Acad Sci USA 99: 6118–6123.

Mardis ER (2008). Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. *9*:387–402.

Maxam AM, Gilbert W (1977). A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA 74:560-564.

Mount DW (2004). Alignment of pairs of sequences. In *Bioinformatics: Sequence and Genome Analysis*, 2nd edition, by David W. Mount. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA.

Mount DW (2008). Using gaps and gap penalties to optimize pairwise sequence alignments. *CSH Protoc*, 2008: pdb top40.

National Institutes of Health. (2010). NIH working definition of bioinformatics and computational biology. Bethesda, USA. http://www.bisti.nih.gov/docs/CompuBioDef.pdf.

Niall HD (1973). Automated Edman degradation: the protein sequenator. Meth. Enzymol. 27:942–1010.

Orengo CA, Todd AE, Thornton JM (1999). From protein structure to function. Curr. Opin. Struct. Biol. 9:374-382.

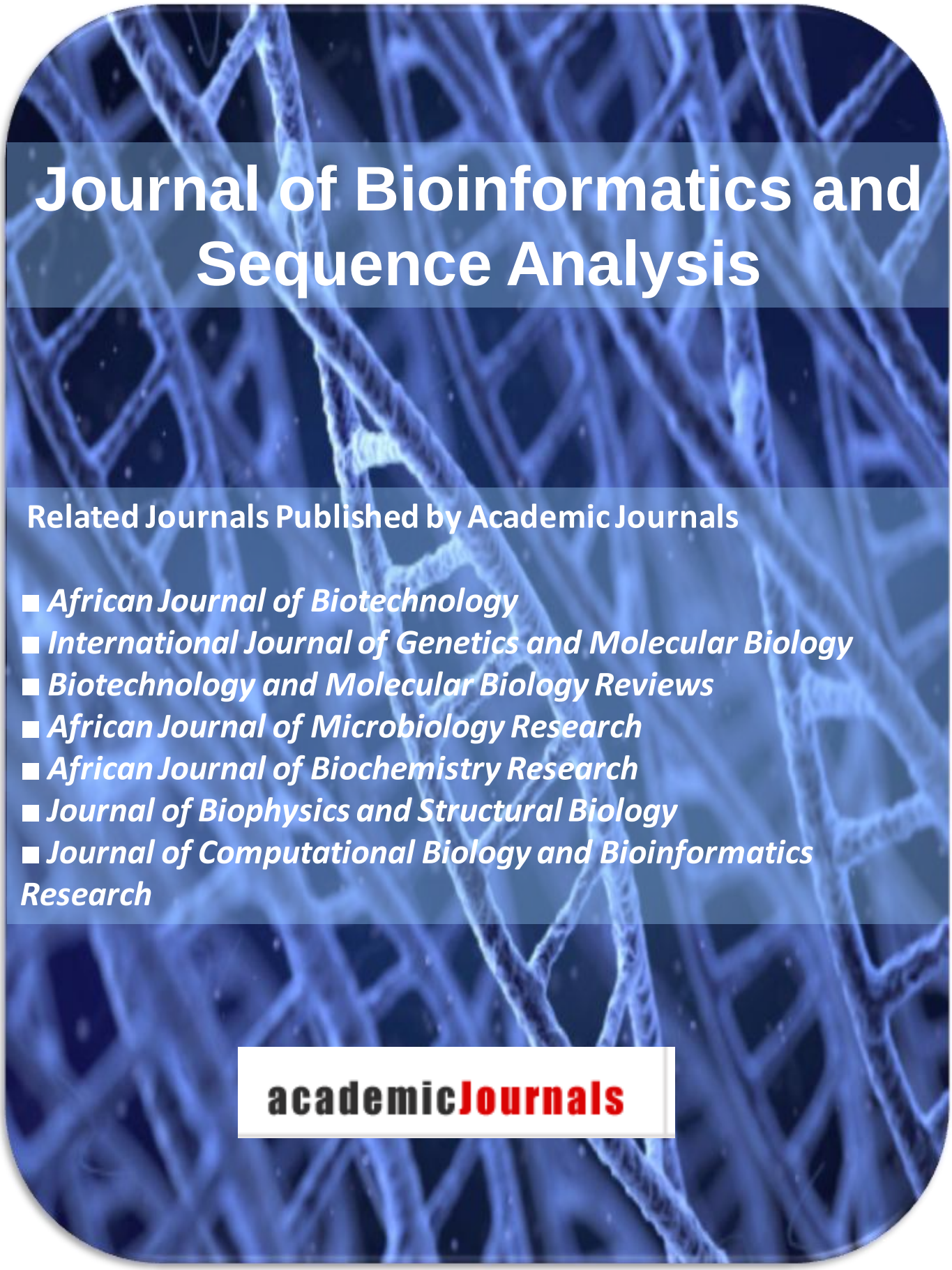
Pearson WR (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol*. 183:63-98.

Pearson WR, Lipman DJ (1988). Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85(8):2444-2448.

Pertsemlidis A, Fondon JW (2001). Having a BLAST with bioinformatics (and avoiding BLASTphemy). Genome Biol. 2(10): REVIEWS2002.

Sanger F, Nicklen S, Coulsen AR (1977). DNA sequencing with chain-

terminating inhibitors. Proc Natl Acad Sci USA, 74:5463-5467.
Sansom C (2000). Database Searching with DNA and Protein Sequences: An Introduction. *Brief Bioinform,* pp.22-32.
Weinstock GM, Smajs D, Hardham J, Norris SJ (2000). From microbial sequence to applications. Res Microbiol. 15:151-158.



Journal of Bioinformatics and Sequence Analysis

Related Journals Published by Academic Journals

- *African Journal of Biotechnology*
- *International Journal of Genetics and Molecular Biology*
- *Biotechnology and Molecular Biology Reviews*
- *African Journal of Microbiology Research*
- *African Journal of Biochemistry Research*
- *Journal of Biophysics and Structural Biology*
- *Journal of Computational Biology and Bioinformatics Research*

academicJournals