Bioinformatic Method for Fungi Identification

Disegha, G. C.[1] & Akani, N. P.[2]
Department of Microbiology,
Rivers State University
[1]gabrieldisegha2@gmail.com, [2]nedieakani@yahoo.com

Abstract
Bioinformatics applies several methods for solving biotechnological problems. Bioinformatics is an output of genetic research at the molecular level. Our concern in this paper is to apply BLAST to solve problem of identification of fungi. It is an algorithm that performs sequence analysis via pair-wise or multiple comparison of nucleic acids (NAs) and other sequences. This algorithm indicates some key bioinformatic significance. Bioinformatic methods can be performed on several open domain genetic or biotechnological databases such as the GenBank and the Yeastgenome. In the past, some methods for identification of fungi have depended mostly on conventional procedures. These procedures have their limitations due to multi-linking species phenotypic features closely related to other species. In medical mycology, the risk of false prescriptions from wrong diagnosis is highly dangerous. The technique in question uses genetic molecules, and it is useful in obtaining species-specific results in fungal identification and taxonomy. This review considered a preharvested complete genome sequence (for demonstrative purpose, instead of 18S rRNA) of a fungal NA from the GenBank and analysing it on the domain. The resulting hits correspond with Aspergillus niger after considering the Maximum score (4.093e+05), E value (0.0) and identification percentage (100%).With the development of more functional and precise algorithms, Bioinformatic analysis is thus advocated for more thorough analysis. As the mycological research community needs more authentic results, improved techniques for analysis are required to achieve more rapid and precise fungal identification, while dwindling cost of facilities are expected in the future.

Key Words*: Bioinformatics, fungal, mycology, Sequence analysis, diagnosis, identification.*

Introduction

Traditionally, specimens submitted to the mycology laboratory from samples collected from skin, tissue, sputum, and other body parts from patients suspected to have fungal infection are examined conventionally for the presence of fungal elements. This is aimed at providing diagnostic views prior to taking more serious actions (Larone, 2011). In routine laboratory and research practice, some methods of identifying fungi have relied on phenotypic methods such as colony morphology/direct microscopic observations, transparency tape method, microslide culture technique, wet-mounts, and biochemical / enzymatic tests. Identification methods still vary for dermatophytes, yeasts, and higher fungi (Baker *et al.*, 2007). These methods and techniques have their inherent disadvantages, and some times, accuracy may be subject to uncertainties. One unique limitation of phenotype-based method is that it is applicable to fungi which can only be cultured on mycological media (Blazewicz *et al.*, 2009). Furthermore, some biochemical expressions may not fit into any known phenomena used for identification of known species (Premier Biosoft, 2015). Fortunately, there are improved modern techniques which are more rapid and accurate than the conventional methods (Tshikhudo *et al.*, 2013). In genotypic-based methods, fungal elements can be obtained from clinical samples or from cultured environmental colonies, thus

enhancing the sensitivity and reducing the time budget for diagnoses and identification of fungi. Thermocyclers are useful in this respect as it utilizes primers designed with specifications to facilitate fungal identification at high level of specificity (Premier Biosoft, 2015). Genotypic-based methods of fungal identification are more sensitive, uniform and procedures similar for all types of fungi, since the bioinformatic method is similarly performed for all sequences of nucleic acids or protein sequences (Disegha and Jeapudoari, 2017). In the overall, these methods, however, are more demanding due to the type of facilities in question (Adzitey *et al*., 2013).

Sequence alignment techniques are applied to search for the best matching sequences (Disegha and Jeapudoari, 2017). Fungal identification is applicable in various aspects of study, namely; Environmental Studies, Microbial Forensics, Analytical studies even in Criminal investigations (Ortet, 2010).

Molecular techniques have become useful in overcoming some of the limitations in fungal identification. This is because non-culture based methods, such as bioinformatics techniques. Sequence alignment is one of such key aspects in handling fungal identification as implied in Kim and Lee (2008).

In Bioinformatics, "sequence alignment'' is a method used generally to analyze strands of DNA, RNA or proteins in order to determine the similarities between the strands which may depict a significance of functional, structural or evolutionary relationships between the sequences" (Mount, 2004).

Although there are other methods of fungal identification, including conventional methods, the present method under discussion lays focus on;

- **a)** How sequence alignment method is used for fungal identification (not involving a whole genome sequence, but a conserved sequence such as 18S rRNA).
- **b)** Tools and software used in sequence alignment,
- **c)** The pros and cons associated with sequence alignment technique of identifying fungi,
- **d)** Its general application in various areas or fields of study.
- **e)** And that the bio-scientist must be ICT compliant.

Methods of Sequence Alignment

There are various methods of sequence alignment but the most common are Pairwise, Multiple sequence and Structural alignments. Pairwise alignment is carried out when two sequences are used to carry out the alignment using appropriate tools. In this method of alignment, one of the sequences is inscribed over the other such that the overlapping element chains are then observed and noted (Agraval *et al*., 2008). BLAST, Mega, *etc*. are applications used for pairwise alignment.

Multiple sequence alignment operates similarly as the pairwise alignment, but it uses more than two sequences, that is, it uses three or more sequences. It can sometimes be presented in a tree-shaped form (Elias and Isaac, 2006). MUSCLE, T-coffee, ClustalW, *etc*. are some software used in multiple sequence alignment include

Alignment of structures otherwise called structural alignment compares shapes of two or more sequences thereby establishing homology between them. Structural alignment takes a look at the entire sequence as an individual unit, of which, pairwise and multiple alignment deals only with the elements within the arrangements. Structural alignment usually makes comparison in a three dimensional format (Bourne and Shindyalov, 2003). Some software used for structural alignment include Vector Alignment Search Tool (VAST), Local-Global alignment (LGA), FSA, Expresso, , MAFFT, POSA, , FATCAT, *etc*. (Omic Tools, 2015).

Types of Sequence Alignment

There are three types or methods of sequence alignment, viz Global, Local, and Semiglobal alignment methods.

Global alignment is a sequencing method that tries to align all the residues from the beginning to the end of the sequence to be able to find out the best possible alignment (Brudno *et al.,* 2003). The Global alignment method is very much appropriate for closely related sequences that are of the same length (Brudno *et al*., 2003). This technique is carried out using the Needleman-winsch algorithm, i.e, based simply on dynamic programming (Mount, 2004)

Local alignment is method of sequence alignment used for comparing sequences suspected to have a similarity or even dissimilarity by finding local regions on a sequence having high level of similarity. The local alignment method used is usually the Smith-Waterman algorithm, which just likes the global alignment method, is also based on dynamic programming (Polyanovsky *et al.,* 2011).

This is a Combined or hybrid method of sequence alignment developed from the combination of the global and local methods, which leads to the expression "semi-global". This method is used in finding the optimal alignment consisting of the initial and terminal of one sequence or the other. This method is best used when the downstream part of one of the sequences intersects with the upstream part of the other (Brudno, *et al*., 2003).

Global and Local alignments are clearly characterized by their different algorithms (that is, a dynamic programming approach), which aligns two different series of sequence by using scoring matrices (Polyanovsky *et al.,* 2011). The table below gives major differences between global and local sequence alignment (Major Differences, 2015).

Direct Bioinformatics to the exclusion of Thermocycler procedures

The stages in thermocycler operations are not necessary if sequences of fungal sample are already handy. This is also true if sequences are harvested from the data bank and used as part of bioinformatic exercise. However, PCR procedures are described vividly by Valones *et al.,* (2009).

The NCBI

NCBI (National Centre for Biotechnology Information, founded in 1988 and located in Bethesda, Maryland via the sponsored legislation by Senator Claude Pepper, is part of the United States National Library of Medicine (NML). The NCBI is a collection of numerous databases that have relevance to biotechnology and biomedicine. These databases also serve as an important source for bioinformatics algorithms and processes. Two examples of databases are GenBank for DNA sequences and PubMed - a database for biomedical literature. Others are the NCBI Epigenomics database, all available online via the Entrez search engine (Ostell, 2002; Disegha & Jeapudoari, 2019)

NCBI has had responsibility for making available the GenBank DNA sequence database since 1992 (NCBI Handbook). The GenBank coordinates with individual laboratories and other sequence databases such as those of the European Molecular Biology Laboratory (EMBL) and the DNA Date Bank of Japan (DDBJ). (Maglott *et al*. 2005, Altschul *et al*., 2009).

The NCBI also uses software tools and algorithms available online, via browsing to help researchers collate data and information that are useful for further research. An example of such tool is the Basic Local Alignment Tool developed by David Lipman and his co-

researchers. The BLAST tool is a program that searches for similarities that exists between sequences. BLAST is able to carry out sequence comparisons between sequences against the GenBank DNA database as quick as 15seconds. (Maglott et al., 2005). The NCBI contains other databases such as, the NCBI Bookshelf, Entrez, Gene, Protein database for protein resource, PubchemBioAssay database *etc*. (Disegha and Jeapudoari, 2017).

Blasting Via the Yeast genome Data base

BLASTing can also be performed via other open domain biological databases, one of which is the yeast genome database accessible via the universal resource locator https://www.yeastgenome.org/

The BLAST Algorithm

The Basic Local Alignment Search Tool can be used in sequence alignment in searching for similarity regions between sequences by comparing proteins or nucleotide sequences using databases and then calculates the statistical significance of the matches that occur. All these are carried out using the BLAST program (Ochmen and Baxter, 2013)

Relationship exist between sequences, therefore, BLAST becomes an appropriate tool for inference of these functional and evolutionary relationships, as well as aid in the identification of the members of gene families. BLAST is one of the most widely used bioinformatics programs for sequence searching the BLAST algorithm and the computer program that implements it were developed by Stephen Attschul, Warren Gish and David Lipman at the U.S. National Center for Biotechnology Information (NCBI), Web Miller at the Pennyslyvania State University and Gene Myers at the University of Arizona and it is available on the web on the NCBI website (Ochmen and Baxter, 2013).

Methodology

The algorithm BLAST is chosen and activated as a menu via the open public domain https://www.ncbi.nlm.nih.gov/ (Figure 1)

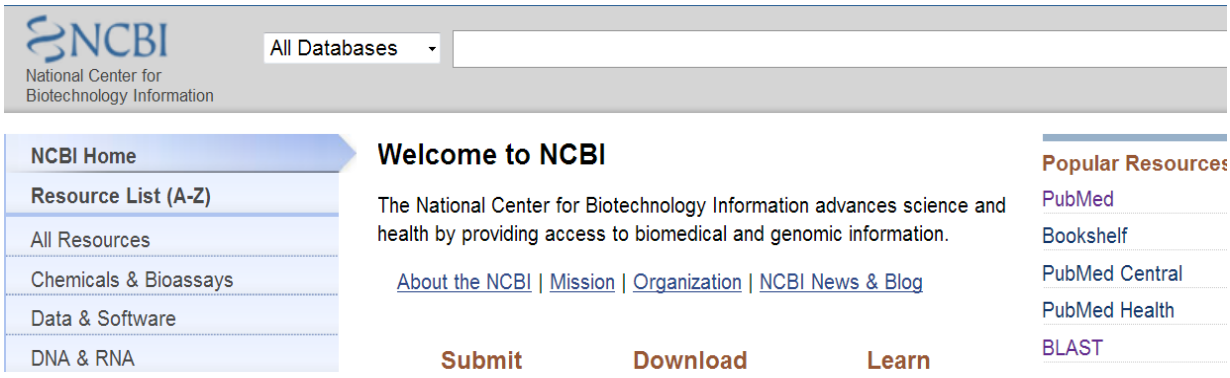

Figure 1: Direct print screen output of NCBI open Public domain (BLAST algorithm command on bottom right extreme pane).

BLAST Procedure for Fungal Sequence Alignment

The procedure for using the BLAST tool in sequence alignment for identifying fungi includes five different steps;

1. **Select the BLAST program (if you are already on NCBI website.) (Figure 1)**

The type of BLAST program is usually specified since they are more than one, by selecting from the data base like; BLASTp, BLASTn, BLASTx, tBLASTn, tBLASTx (Li *et al.* 2009; Disegha and Jeapudoari, 2017) (Figure 2)

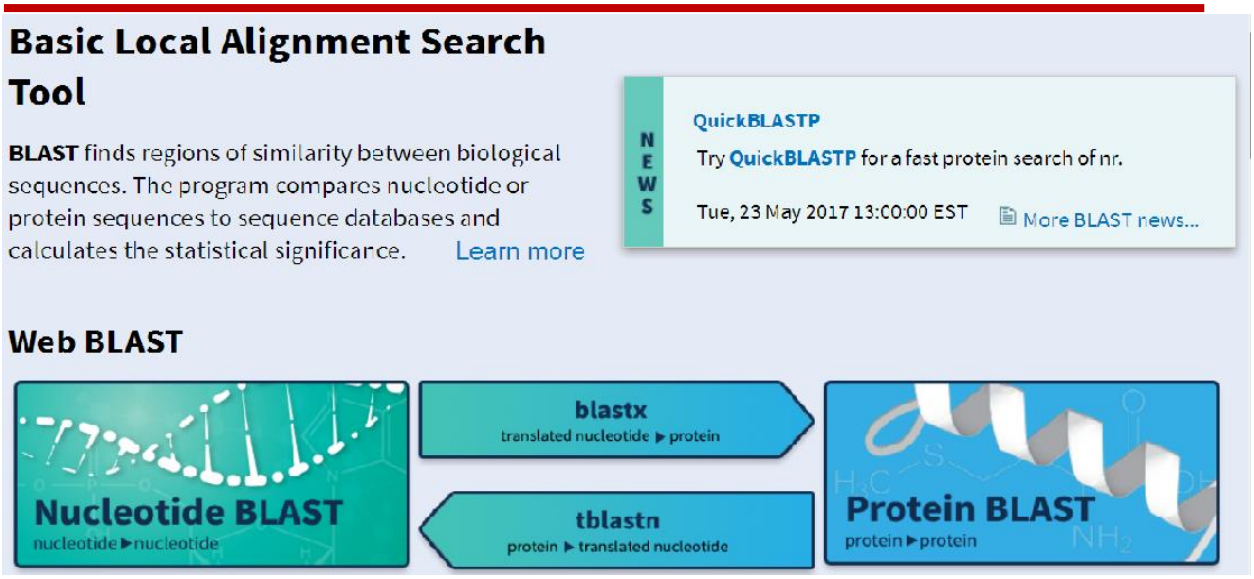

Figure 2: BLAST options dialogue.

2. Query entering or file upload

A query sequence is entered and this is performed by pasting a sequence in the query field. Or, the file containing the sequence for similarity search is uploaded in a FASTA format. (Li *et al.* 2009). Accession number for the sequence can also be used in the query sequence field. (Figure 3)

Figure 3: A dialog box for entering a query sequence or uploading file containing sequence. (Source; NCBI, 2016)

3. Data base selection to search

Using the query sequence uploaded, the user, searches for similar sequences from databases. But before this search, the data bases available and the type of sequences in those databases must be known (Madden, 2002).

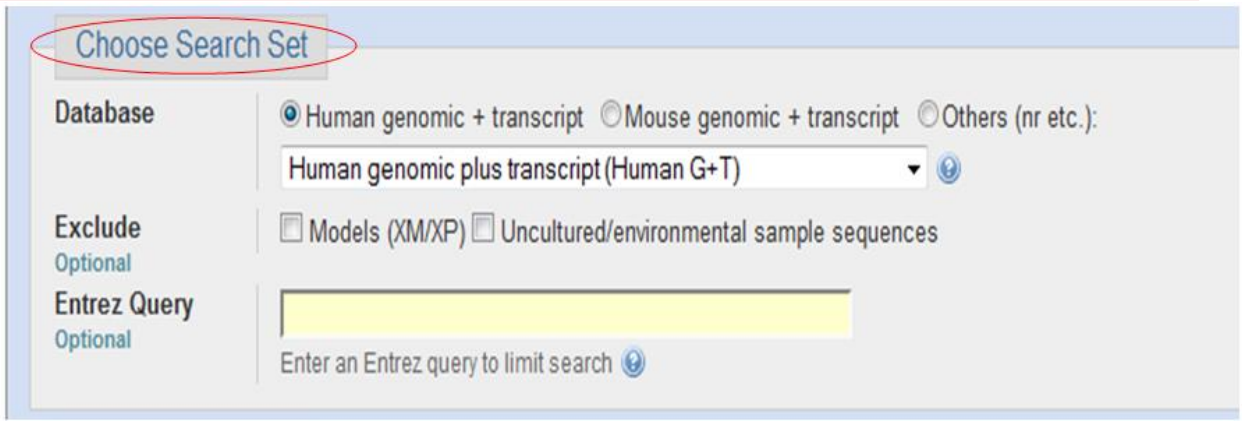

Figure 4. A dialog box for selecting database to search
(Source; NCBI, 2016)

Algorithm selection and parameters of the algorithm

A unique algorithm must be specified for the BLAST program to use, since different BLAST programs have different algorithms. In this case Nucleotide BLAST is selected as the preferred BLAST algorithm (see Figure 2). Algorithm parameters are set according to knowledge and preferences, and they include Target sequences, word size, query range, short queries, E-value, filters (filter and mask) and scoring parameters. All these are required in order to run BLAST programs. Most of the time, default values are provided, the user can make necessary adjustments to suit his desired search criteria (Rastogi *et al*., 2013). MetaBLAST option may be selected if highly similar sequences are targeted.

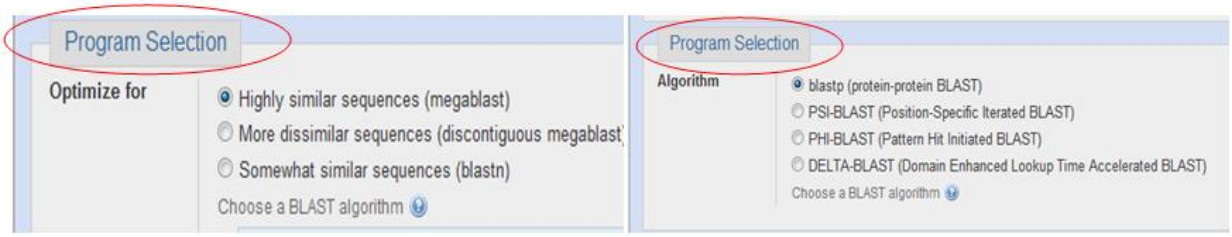

Figure 5: A dialog box for Algorithm and the parameters
(Source; NCBI, 2016)

Running the BLAST program

After inputting the required values, the program is submitted or in this case, can be run by clicking the "BLAST" button at the end section of the page. Following the submission, a result page will pop-up, displaying the information such a: Query id, Description, length of sequence, etc., and BLAST program, also displaying the detected domains resulting from he search. (Madden, 2002).

BLASTing

After selecting desired and appropriate options, BLAST is executed by simply clicking BLAST command. The command will execute the actual type of BLAST selected prior to this state, e.g. Nucleotide BLAST (Zheng *et al*., 2000).

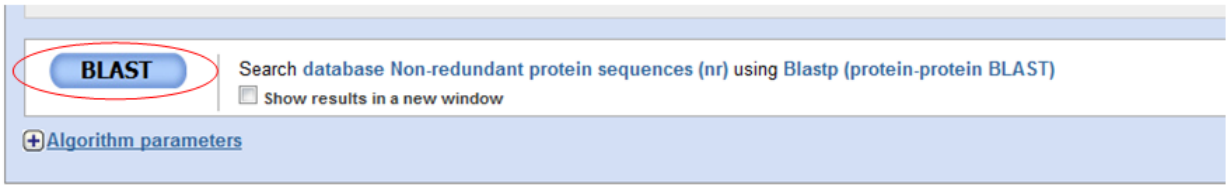

Figure 6: A dialog box for running the BLAST program
(Source: NCBI, 2016)

BLAST Search Output:
This is the final result of the BLAST exercise. The first matching sequence with Max Score and 100% identity is taken as the identified fungal organism. User can find more descriptions about these alignments, by dragging the mouse to the each colored bar which is shown in Figure 7 (Madden, 2002; Wheeler and Bhagwat, 2007).

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident
Aspergillus niger contig An12c0160, genomic contig	4.093e+05	4.296e+05	8%	0.0	100%
Aspergillus niger contig An12c0060, genomic contig	3.654e+05	3.735e+05	7%	0.0	100%
Aspergillus niger contig An12c0030, genomic contig	3.472e+05	3.472e+05	7%	0.0	100%
Aspergillus niger contig An12c0340, genomic contig	3.139e+05	3.157e+05	6%	0.0	100%
Aspergillus niger contig An12c0070, genomic contig	2.602e+05	2.767e+05	5%	0.0	100%
Aspergillus niger contig An12c0110, genomic contig	2.578e+05	2.581e+05	5%	0.0	100%
Aspergillus niger contig An12c0130, genomic contig	2.453e+05	2.458e+05	5%	0.0	100%

Figure 7: A dialog box showing the BLAST result (obtained by direct BLASTing on NCBI) using an unknown nucleotide sequence of fungal element.

The alignment is preceded by the sequence identities, along with the definition line, length of the matched sequence, followed by the score and E-value. The line also contains the information about the identical residues in alignment (identities), number of positivity's, number of gaps used in the alignment. Finally, it shows the actual alignment, along with the query sequence on the top and database sequence below the query. The number on either sides of the alignment indicates the position of amino acids/nucleotides in sequence (Madden, 2002; Wheeler and Bhagwat, 2007).

Results and Discussion
From the BLAST output, the organism suspected is *Aspergillus niger* based on the top three correlations and matches. This was obtained from a total of 200 Blast hits derived from 100 subject sequences.

Aspergillus niger contig An12c0160, genomic contig	4.093e+05	4.296e+05	8%	0.0	100%
Aspergillus niger contig An12c0060, genomic contig	3.654e+05	3.735e+05	7%	0.0	100%
Aspergillus niger contig An12c0030, genomic contig	3.472e+05	3.472e+05	7%	0.0	100%

Advantages of Sequence Alignment

There are advantages or merits of sequence alignment in identifying fungi and other organisms. This includes;

1. Sequenced RNA, such as expressed sequence tags and full-length mRNA's, can be aligned to a sequenced genome to find where there are genes and get information about alternative splicing and RNA editing. (Kim and Lee, 2008).
2. Sequence alignment is a part of genome assembly, where which sequences are often aligned to find overlap so that contigs which are long stretches of sequences can be generated. (Blazewicz, *et al*., 2009; Wheeler and Bhagwat, (2007).
3. Multiple sequence alignments can be used to create a phylogenetic tree. This is because functional domains known in annotated sequences can be used in carrying out alignment in non-annotated sequences. Also, the regions that are conserved and are known to have functional importance can be found (Budd and Aidan, 2009).
4. Sequence alignment, especially multiple sequence alignment methods can be used to identify sites that are functionally important like; active sites and binding sites by locating or finding conserved domains.
5. Sequence alignment uses similarity in sequences to find common ancestry to species level. This has played a very vital role.
6. Sequence alignment far outweighs the traditional methods of identification since it is faster, more accurate, precise, and more reliable (Disegha and Jeapudoari, 2017).

Disadvantages of Sequence Alignment

Due to the advantages enumerated above and the high demand of molecular or genomic approaches to fungal identification, sequence alignment, has over recent times, become in vogue in as part of molecular identification method**.** However, several current disadvantages pose a limitation on the use of genomics and sequence alignment for fungi identification (Sentausa and Fournier, 2013). Klenk and Göker (2010) reported that completely sequenced genomes for many of the major lineages of prokaryotes (fungi inclusive) are not available. Because of the vast number of fungi yet to be discovered open domains of databases cannot possibly have storage of nucleic acid sequences of such fungi. According to Sentausa and Fournier (2013), "the currently available genome sequences have been obtained mostly from three phyla (Proteofungi, Firmicutes, and Actinofungi). Thus, many phyla are poorly represented in genomics (http://www.genomesonline.org/cgi-bin/GOLD/index.cgi)". Furthermore, Klenk and Göker (2010) noted that, "even if the genome sequences of the species of interest are available, in many cases they are not type strains, and, therefore must be used with caution, as prokaryote taxonomy is based on type strains only" (Tindall *et al.*, 2010).

Another disadvantage is that existing genomic sequences vary greatly in their finished quality, often being available only as unfinished draft assemblies that may be less informative than finished whole genome sequences (Klassen *et al*. 2012; Ricker *et al.*, 2012). These inadequacies have lead to giving conditions for minimal sequencing quality which are required for genomes to be incorporated in taxonomic analyses. For example, the guidelines developed by the Next-generation Sequencing: "Standardization of Clinical Testing work

group" intended for use in open domains such as genbank (Gargis *et al.,* (2010).

Another drawback in sequence based identification is that results obtained through sequence analysis often do not correspond with existing taxonomic categories and related levels. This is based on the fact that prokaryotes do not have one universal method classification for prokaryotes (AI-Ozen and Vesth, 2012).

Sequence alignment generally does not yield single-specific results, rather cumbersome alignment scores are given for the analyst to chose what is best fitted with respect to maximum scores, total scores, query content, E-values , percentage identify, and the total matches displayed

Yet another disadvantage of sequence alignment hinges on the fact that, determining phylogeny of an organism based on sequences, poses difficulty in aligning distantly related sequences using pairwise, alignments without errors creeping in.

Sequence alignment is also a very expensive method of identifying fungi. This is so because the series of steps, processes, and equipment required prior to sequence alignment proper are high-priced. Equipments such as PCRs or thermocyclers, gene analyzers or gene sequencing machines are sold at very high prices. In addition, there are various tools that are best described as molecular laboratory apparatus that add more to the general cost of molecular identification. And if the skill is lacking, it becomes an additional cost.

Conclusion

Some methods of identifying fungi have relied on conventional routine methods via direct examination of specimens, cultural practices, and biochemical methods although with some limitations such as giving unreliable results. Sequence analytical techniques are useful in overcoming some of these limitations in fungi identification. This affords the researcher the opportunity of utilizing online tools to facilitate processes leading to identification, characterization, phylogenetic analysis and prediction of elements for future advances, while bearing in mind the challenge of cost in the use of this technology.

References

Adzitey, F., Huda, N., Gulam, R. R., and Ali , G. E. E. (2013). Molecular techniques for detecting and typing of fungi, advantages and application to foodborne pathogens isolated from ducks. Journal of Biotechnology. https://www.ncbi.nlm.nih.gov/pmc/articles/

Agrawal, A., Brendel, V.P., and Huang, X. (2008). Comparative analysis of Bioinformatics. International Journal of Computational Biology and Drug Design. 1(4):347-367.

AI-Ozen, T. and Vesth, D.W. (2012). Ussery from genome sequence to taxonomy—a skeptic's view. E Rosenberg, E.F. DeLong, E. Stackebrandt, S. Lory, F Thompson (Eds.), The prokaryotes (4th edn.), Springer: Berlin (2012)

Altschul, S.F., Gish, W., Miller, W., and Myers E.W., Lipman, D.J. (1990) Basic Local Alignment Search Tool. Journal of Molecular Biology, 215:403–410.

Baker, F. J. Silverton, R. E., and Pallister, C.J. (2007). Baker and Silverton's Introduction to Medical Laboratory Technology (7th Ed.). Lagos: Bounty Press.

Blazewicz, J., Bryja, M., and Figlerowicz, M. (2009). Whole genome assembly from 454 sequencing output via modified DNA graph concept. Journal of Computational Biology Chemistry,33(3)224-30.

Bourne P.E and Shindyalov, I. N. (2003): Structure Comparison and Alignment. In: Bourne, P.E., Weissig, H. (Eds): Structural Bioinformatics. Hoboken: Wiley-Liss.

Brudno, M., Malde, S., and Poliakov, A. (2003). The Protein Structure. Journal of Bioinformatics. 19 (9): 151-62.

Budd, I. and Aidan, D. (2009). Multiple Sequence alignment exercises and demonstrations. European Molecular Biology Laboratory. 1(1) 24-30.

Disegha, G.C. and Jeapudoari, T. F. (2017). Sequence Alignment as a Method of Bacterial Identification. Current Studies in Comparative Education, Science and Technology, 4(1): 221-238.

Elias, T., and Isaac, C. (2006).Whole genome assembly. Journal of Computational Biology. 1323-1339.

Fournier, P.E., Drancourt, M., Colson, P., J Rolain, J.M., La Scola, B., and Raoult, D. (2013) Modern clinical microbiology: new challenges and solutions Nature Reviews Microbiology, 11, 574–585.

Gargis, A.S., Kalman, L., and Berry, M.W. (2010). Assuring the quality of next-generation sequencing in clinical laboratory practice. Journal of Natural Biotechnology, 30 (12): 1033-1036.

Kim, N., and Lee, C. (2008). Bioinformatics detection of alternative splicing. Journal on Methods in Molecular Biology. 452:179-97.

Klassen, J. and Currie, C. (2012). Gene fragmentation in fungal draft genomes: extent, consequences and mitigation. Journal of BMC Genomics, 13 (2012):14.

Klenk, H. P. and Göker, M. (2010). En route to a genome-based classification of Archaea and Fungi?Syst Journal of Applied Microbiology, 33 (10), 175-182.

Larone, D. H. (2011). Medically important fungi. Washington, DC: ASM Press.

Li, J.B., Levanon, E.Y., and Yoon, JK.(2009). Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science 342 (5932):1210-3.

Madden, T. (2002). The NCBI Handbook 2nd edition Chapter 16. The BLAST sequence Analysis Tool. www.ncbi.nlm.nih.org.

Maglott , D.1., Ostel,l J., Pruitt, K.D., and Tatusova, T. (2005). Nucleic Acids Research, 1;33.

Major Differences (2016). Diffeences between global and local sequence alignment. www.majordifferences.com

Mount, D.M. (2004). Genome Analysis. Journal of Bioinformatics Sequence. 2(1)54-60.

NCBI (2016). Different BLAST programs. www.ncbi.nlm.nih.org

Oehmen, C.S., and Baxter, D.J. (2013). Multiple Sequence alignment exercises. Journal of Bioinformatics. 29(6):797-798.

Omic Tools (2015). Multiple structure alignment software tools and Protein data analysis. https://omictools.com/multiple-protein-structure-alignment-category

Ortet, P., and Bastien, O. (2010). Bioinformatics detection. Journal of Evolutionary Bioinformatics. 6:159-187.

Ostell, J. (2002. The Entrez search and retrieval system. IN The NCBI handbook [Internet], National Library of Medicine (US), National Center for Biotechnology Information, Bethesda, MD, Chapter 14. Available from the Entrez Books database

Polyanovsky,V.O., Roytberg, M.A., and Tumanyan, V.G. (2011). Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences. Journal of Algorithms for Molecular Biology.6(1): 25.

Premier Biosoft (2015) Identifying fungi. http://www.premierbiosoft.com/tech

Rastogi, S.C. Mendiratta, N. and Rastogi, P.(2013). Bioinformatics – Methods and applications: Genomics, Proteomics and drug discovery. Delhi: PHI Learning.

Ricker, N, Qian, H., and Fulthorpe, R.R. (2012). The limitations of draft assemblies for understanding prokaryotic adaptation and evolutionGenomics, 100 (12), 167-175.

Sentausa, E. and Fournier, P.-E. (2013). Advantages and limitations of genomics in prokaryotic taxonomy. Journal of Clinical Microbiology and Infection, 19 (9), 790-795.

Tindall, B.J., Rosselló-Móra, R. Busse, H.J. Ludwig, W. and Kämpfer, P.(2010). Notes on the characterization of prokaryote strains for taxonomic purposes International Journal of Systematic Evolution in Microbiology, 60 (10), 249-266.

Valones, M,A.A., Guimarães, R.L., Brandão, L.A. C., de Souza, P. R. E., Carvalho, A.A. T., and Crovel, S. (2009) Principles and applications of polymerase chain reaction in medical diagnostic fields: a review. Brazilian Journal of Microbiology, 40:1-11

Wheeler D, and Bhagwat M. (2007). BLAST QuickStart: Example-Driven Web-Based BLAST Tutorial. In: N.H, Bergman (ed.).. Comparative genomics: Volumes 1 and 2. Totowa: Humana Press. Available from: https://www.ncbi.nlm.nih.gov.

Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller (2000), A greedy algorithm for aligning DNA sequences, Journal of Computational Biology, 7(1-2):203-14.

APPENDIX:

Appendix 1: Sequences producing significant alignments

Description	Max score	Total score	Query cover	E value	Ident	Accession
1. *Aspergillus niger* contig An12c0160, genomic contig	4.093e+05	4.296e+05	8%	0.0	100%	AM270272.1
2. *Aspergillus niger* contig An12c0060, genomic contig	3.654e+05	3.735e+05	7%	0.0	100%	AM270263.1
3. *Aspergillus niger* contig An12c0030, genomic contig	3.472e+05	3.472e+05	7%	0.0	100%	AM270260.1
4. *Aspergillus* niger contig An12c0340, genomic contig	3.139e+05	3.157e+05	6%	0.0	100%	AM270290.1
5. Aspergillus niger contig An12c0070, genomic contig	2.602e+05	2.767e+05	5%	0.0	100%	AM270264.1
6. Aspergillus niger contig An12c0110, genomic contig	2.578e+05	2.581e+05	5%	0.0	100%	AM270267.1
7. Aspergillus niger contig An12c0130, genomic contig	2.453e+05	2.458e+05	5%	0.0	100%	AM270269.1
8. Aspergillus niger contig An12c0220, genomic contig	2.181e+05	2.308e+05	4%	0.0	100%	AM270278.1
9. Aspergillus niger contig An12c0090, genomic contig	2.138e+05	2.180e+05	4%	0.0	100%	AM270266.1
10. Aspergillus niger contig An12c0380, genomic contig	1.696e+05	1.698e+05	3%	0.0	100%	AM270294.1
11. Aspergillus niger contig An12c0080, genomic contig	1.646e+05	1.650e+05	3%	0.0	100%	AM270265.1
12. Aspergillus niger contig An12c0290, genomic contig	1.531e+05	1.531e+05	3%	0.0	100%	AM270285.1
13. Aspergillus niger contig An12c0020, genomic contig	1.358e+05	1.361e+05	2%	0.0	100%	AM270259.1
14. Aspergillus niger contig An12c0010, genomic contig	1.271e+05	1.271e+05	2%	0.0	100%	AM270258.1
15. Aspergillus niger contig An12c0190, genomic contig	1.256e+05	1.343e+05	2%	0.0	100%	AM270275.1

16.	Aspergillus niger contig An12c0280, genomic contig	1.209e+05	1.209e+05	2%	0.0	100%	AM270284.1
17.	Aspergillus niger contig An12c0270, genomic contig	96758	96758	2%	0.0	100%	AM270283.1
18.	Aspergillus niger contig An12c0230, genomic contig	96078	96078	2%	0.0	100%	AM270279.1
19.	Aspergillus niger contig An12c0330, genomic contig	91949	91949	1%	0.0	100%	AM270289.1
20.	Aspergillus niger contig An12c0170, genomic contig	91436	95357	2%	0.0	100%	AM270273.1
21.	Aspergillus niger contig An12c0180, genomic contig	87561	93623	2%	0.0	100%	AM270274.1
22.	Aspergillus niger contig An12c0310, genomic contig	80426	80426	1%	0.0	100%	AM270287.1
23.	Aspergillus niger contig An12c0260, genomic contig	79316	79316	1%	0.0	100%	AM270282.1
24.	Aspergillus niger contig An12c0120, genomic contig	60567	62018	1%	0.0	100%	AM270268.1
25.	Aspergillus niger contig An12c0200, genomic contig	57577	59301	1%	0.0	100%	AM270276.1
26.	Aspergillus niger contig An12c0360, genomic contig	52451	52451	1%	0.0	100%	AM270292.1
27.	Aspergillus niger clone AXAS121-C12, complete sequence	49328	69504	1%	0.0	99%	AC253895.1
28.	Aspergillus niger contig An12c0240, genomic contig	48660	48660	1%	0.0	100%	AM270280.1
29.	Aspergillus niger contig An12c0040, genomic contig	48239	48367	1%	0.0	100%	AM270261.1
30.	Aspergillus niger contig An12c0300, genomic contig	46898	46898	0%	0.0	100%	AM270286.1
31.	Aspergillus niger contig An12c0350, genomic contig	42474	42474	0%	0.0	100%	AM270291.1
32.	Aspergillus niger contig An12c0370, genomic contig	34948	37213	0%	0.0	100%	AM270293.1

33. Aspergillus niger clone AXAS125-B01, complete sequence	31196	67877	1%	0.0	99%	AC253951.1
34. Aspergillus niger contig An12c0050, genomic contig	22552	22552	0%	0.0	100%	AM270262.1
35. Aspergillus niger CBS 513.88 hypothetical protein, mRNA	21455	22150	0%	0.0	99%	XM_001395739.1
36. Aspergillus niger contig An12c0320, genomic contig	19621	19621	0%	0.0	100%	AM270288.1
37. Aspergillus oryzae RIB40 DNA, SC113	16866	48583	1%	0.0	99%	AP007166.1
38. Aspergillus oryzae RIB40 DNA, SC023	16862	44633	0%	0.0	99%	AP007157.1
39. Aspergillus niger accA gene for acetyl-CoA carboxylase, exons 1-3	16294	16294	0%	0.0	99%	AJ748127.1
40. Aspergillus niger contig An12c0250, genomic contig	15640	15640	0%	0.0	100%	AM270281.1
41. Aspergillus niger CBS 513.88 hypothetical protein, mRNA	15080	28272	0%	0.0	99%	XM_001395326.2
42. Aspergillus niger contig An12c0210, genomic contig	13566	13971	0%	0.0	100%	AM270277.1
43. Aspergillus niger CBS 513.88 acetyl-CoA carboxylase, mRNA	12829	13434	0%	0.0	100%	XM_0013