(NSAZ-2022)

Science and Engineering Research Board (SERB) Sponsored
National Symposium
On
Applied Zoology, Profitable Animal Production, and Health: Current Status and Future Progress (NSAZ-2022) 23rd & 24th September- 2022

Recent Trends in Applied Zoology

Dr.D.S.Rathod
Editor

Associate Editors
Dr. K.S.Raut
Mr.Datta Nalle

National Edited Book

PRABHAKAR PUBLICATION

(ISBN: 978-93-92464-55-3)

Recent Trends in Applied Zoology

Edited by: Dr.D.S.Rathod

ISBN: 978-93-92464-55-3 (First printing edition, May, 2023)

Published by: **Prabhakar publication:**

Address: 304 Siddhivinayak apartments, sutmill road, Latur - 413512

Phone: +91 8600881127, +91 8149153018

Website: www.prabhakarpublication.com

Science and Engineering Research Board (SERB) Sponsored
National Symposium
On

Applied Zoology, Profitable Animal Production, and Health: Current Status and Future Progress (NSAZ-2022) 23rd & 24th September- 2022

Recent Trends in Applied Zoology

Editor
Dr.D.S.Rathod
Head
Department of Zoology and Fishery Science,
Rajarshi Shahu Mahavidyalaya (Autonomous),

Latur- 413531, Maharashtra

Associate Editors
Dr. K.S.Raut
Mr.Datta Nalle
Department of Zoology and Fishery Science, Rajarshi Shahu Mahavidyalaya (Autonomous),
Latur- 413531, Maharashtra

Index

Chapter-20

Bioinformatics Tools for DNA Barcoding

Dnyaneshwar S. Rathod and Dhanshree M. Jagtap

Department of Zoology and Fishery Science, Rajarshi Shahu Mahavidyalaya,
[Autonomous] Latur 413512 Maharashtra.

Abstact:

The primary sciences involved in the interdisciplinary field of bioinformatics are computer science, mathematics, statistics, molecular biology, and genetics and other fields of science. Large-scale, data-intensive biological issues are handled from a computational perspective. Modelling molecular biological processes and drawing conclusions from gathered data are the two most frequent issues. Sequence and structural alignment, database design and data mining, macromolecular geometry, the creation of phylogenetic trees, the prediction of protein structure and function, gene discovery, and clustering of expression data are just a few of the computational methods used in bioinformatics.

Analyses in bioinformatics predominantly focus on three types of large datasets available in molecular biology: macromolecular structures, genome sequences, and the results of functional genomics experiments (e.g. expression data). Additional information includes the text of scientific papers and "relationship data" from metabolic pathways, taxonomy trees, and protein-protein interaction networksThis chapter gives a brief introduction to bioinformatics tools and techniques for the welfare of animal and consults research.

Key words: **BLAST, DNA barcoding**

Introduction: Bioinformatics Tools:

Software programmes called bioinformatics tools are created with the purpose of sifting through the vast amount of molecular biology and biological databases to find the relevant data and perform sequence or structural analysis.

Both visualisation tools and data-mining software are available for the analysis and retrieval of information from proteomic databases and genomic sequence databases, respectively. These fall under several categories, including homology and similarity tools, tools for protein functional analysis, tools for sequence analysis, and other tools.

In order to distinguish between species, DNA barcoding is used. It functions by examining a certain DNA region. The DNA barcode is the name of this area. After that, the DNA barcode's sequence is compared to a reference library that contains data about several species connected to their barcodes.

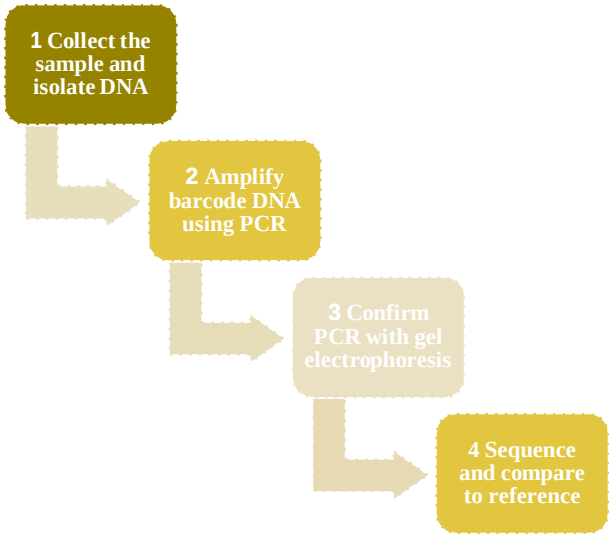

Fig., 1-The initial step is to gather a sample for the investigation. After that, DNA is extracted from the sample using a variable technique **(1),** and the target DNA region is amplified using PCR **(2),** creating several copies of the DNA barcode. The success of the PCR is assessed by visualizing the amplified DNA by gel electrophoresis **(3).** The DNA barcode is finally sequenced **(4).** By comparing the sequenced DNA barcode to reference databases, the species can be determined.

BLAST:

An alignment between a nucleotide or protein sequence, known as a "query," and nucleotide or protein sequences within a database, known as "subject" sequences, is created using a programme suite known as BLAST, which stands for Basic Local Alignment Search Tool. A protein "query" sequence was employed by the first BLAST programme to search a database of protein sequences. Soon after, a nucleotide sequence database and a version that used nucleotide query sequences were released. Cross-comparisons between nucleotide and protein sequences are made possible by the addition of an intermediary layer in which nucleotide sequences are translated into the appropriate protein sequences in accordance with a given genetic code. Large-scale nucleotide database searches and the creation of alignments between sequences are both made possible by specialised variants of BLAST.

Query and Database Sequence

A defining line that starts with the symbol ">" and contains identifiers and descriptive information is provided before the BLAST "query" sequences, which are character strings of single-letter nucleotide or amino acid codes. The name of this format is FASTA. Using the "formatdb" programme, which creates a mixture of binary- and ascii-encoded files containing

the sequences and indexing data used during the BLAST search, BLAST databases are created from concatenated FASTA formatted sequences.

Alignments and Substitution Matrices

Each letter in one sequence is paired with, or "aligned to," either a letter or a gap in the other in a process known as BLAST alignment. Each pair of letters that are aligned is given a value, and the alignment score is calculated by adding together all of these values over the alignment's length. Scores are provided for every potential pair of amino acid letters in a "substitution matrix" for protein sequence alignments, where likely substitutions have positive values and unlikely replacements have negative values. While other members of the PAM series are also available, BLAST by default employs the "blosum62" matrix, a member of the most popular series of replacement matrix . The reward for nucleotide alignments in BLAST is +2**.**

Algorithm

In order to start a search, BLAST first indexes all character strings of a specific length contained within the "query" by their initial location within the query. User-configurable "wordsize" refers to the size of the string to index. According to the BLAST programme being used, the "wordsize" has a range that varies; typical values are 3 for protein-to-protein sequence searches and 11 for nucleotide-to-nucleotide searches.

Then, BLAST searches the database for strings that match the "words" indexed in the "query" and those discovered in the database sequences. These matches must be accurate for nucleotide-to-nucleotide searches and must meet a minimum score requirement for protein-to-protein searches based on the results of a substitution matrix.

BLAST tries to stretch both forward and backward from the match to generate an alignment when a word match—or two neighbouring terms in the case of protein searches—is discovered. The alignment score will be extended by BLAST until it reaches a critical level due to the negative scores provided by mismatches, at which point it will stop increasing. The "dropoff" refers to this vital quantity.

Statistical Significance (BLAST):

As previously mentioned, the alignments discovered by BLAST during a search are scored and given a statistical value termed the "Expect Value." The "Expect Value" is the likelihood that, given the size of the database searched, a BLAST alignment as excellent or better will be discovered by chance.

Spotlightingting alignments are reported depends on the user-defined "Expect Value" threshold. The BLAST default of "10" is created to make sure that no physiologically significant alignment is overlooked. A higher "Expect Value" threshold is less restrictive.

FASTA:

One of the first and most used database similarity search methods is FASTA. FASTA (or FastA), which stands for "Fast-All," is a sequence alignment programme that compares input nucleotide or protein sequences to databases already in existence. Since its initial creation in 1985 by David J. Lipman and William R. Pearson, it has been improved upon and tailored for a variety of uses.

The FASTA programme gave rise to the text-based file format that is presently accepted in bioinformatics for representing nucleotide or protein sequences.

Mechanism of FASTA

1. Identifying Regions-

By building an index for the query sequence, the initial step is to find regions with high similarity. It's also known as the hashing step. The query sequence is divided into smaller words called k-tuples (ktup) before being used to build the lookup table.

The quantity of background word hits decreases as the ktup value rises. The algorithm may concentrate on the most pertinent hits and increase search speed by cutting down on the quantity of these background word hits. Protein k-tuples are typically 2 while nucleotide sequence k-tuples are 6.In order to find matches between the k-tuples in the query sequence and the database sequences, the lookup table must first be established. In a two-dimensional matrix, similar sections are shown as diagonals.

2. Re-Scoring-

Using appropriate scoring matrices, the top ten diagonals are rescored. The identity matrix is utilised for DNA sequences, while the BLOSUM50 or PAM matrix is used for proteins. For each of the rescanned diagonal regions, the subregion with the highest score is determined. Initial regions are those parts of the diagonals with high scores.

3. Joining Threshold-

Applying a scoring cutoff or joining threshold eliminates segments that are unlikely to be included in the final alignment. The rankings of the library sequences are determined by their starting scores. The regions that have initial scores over the predetermined level are chosen, and their viability as a unit is examined. Gap penalties are applied while gaps are established between the diagonals in this stage.

4. Final Alignment-

In order to create the final alignment, the gapped alignment is improved. The optimal score (opt) for alignment is calculated via the banded Smith-Waterman algorithm, a dynamic programming algorithm, in order to achieve this. Statistics calculations are made using this score.

Statistical Significance (FASTA):

The statistical significance of each discovered alignment is also estimated using FASTA. The E-value, which gauges how likely it is that a sequence alignment score will be obtained by random, is used to evaluate it. The alignment becomes increasingly significant as the E-value decreases. Statistical parameters are not limited to the e-value. In addition to using the bit score and similarity score based on the scoring matrix and gap penalties, FASTA also employs other statistical methods to assess the relevance of sequence alignments.

The Z-score, a statistical metric that expresses the number of standard deviations from the mean score of the database search, is also included in the FASTA output. A stronger match is denoted by a higher Z-score value.

References:

1. An Overview of Bioinformatics Tools for DNA Meta-Barcoding Analysis of Microbial Communities of Bioaerosols: Digest for Microbiologists HamzaMbareche , Nathan Dumont-Leblond ,Guillaume J. Bilodeau and Caroline Duchaine .
2. Madden TL, McGinnis S. Blast: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 2004;32:W20–W25.
3. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. ProcNatlAcadSci USA. 1992;89:10,915–10,919.
4. Biomed Res Found. Washington, DC; 1978. Atlas of Protein Sequence and Structure, chapter Matrices for detecting distant relationships.
5. Altschul SF, Gish W. Local alignment statistics. Methods Enzymol. 1996;266:460–480.
6. Madden TL, Schaffer AA, Zhang J, et al. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402.
7. Madden T. The NCBI Handbook. The blast sequence analysis tool.
8. Barton, G. J. (1996). Protein sequence alignment and database scanning. Protein structure prediction: A practical approach, 31-63.
9. FASTA Programs (virginia.edu)
10. FASTA/SSEARCH/GGSEARCH/GLSEARCH < Sequence Similarity Searching < EMBL-EBI
11. Lloyd, Andrew. (2001). Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins (Methods of Biochemical Analysis, 43). Briefings in Bioinformatics. 2. 10.1093/bib/2.4.407.
12. Pearson, W. R. (2005). FASTA Algorithm. Encyclopedia of Life Sciences.