ejpmr, 2016,3(1), 76-83

SJIF Impact Factor 2.026

EUROPEAN JOURNAL OF PHARMACEUTICAL AND MEDICAL RESEARCH
www.ejpmr.com

Review Article
ISSN 3294-3211
EJPMR

# SEQUENCE RETRIEVAL AND ANALYSIS: A USEFUL TOOL IN BIOINFORMATICS – A REVIEW

**Essien Okon*[1], Mary Esien Kooffreh[2], Anthony John Umoyen[2], Peter Nkachukwu Chukwurah[2], Ekeoma Malvina Nwankpa[2] and Nwagu Kingsley Ekene[2].**

[1]Dept. of Biological Science, Cross River University of Technology, Calabar, Nigeria.
[2]Dept. of Genetics & Biotechnology, Cross River University of Technology, Calabar, Nigeria.

***Correspondence for Author: Dr. Essien Okon**
Dept. of Biological Science, Cross River University of Technology, Calabar, Nigeria.

**Article Received on 10/11/2015** **Article Revised on 01/12/2015** **Article Accepted on 24/12/2015**

## ABSTRACT

Bioinformatics is an interdisciplinary scientific field that develops methods for storing, retrieving, organizing and analyzing biological data. A major activity in bioinformatics is to develop software tools to generate useful biological knowledge. Bioinformatics uses computers to better understand biology. Bioinformatics as a science can provide input to all previously mentioned scientific fields, as the recording and processing of detailed biological data is the first step towards doing something with them. Bioinformatics uses many areas of computer science, statistics, mathematics and engineering to process biological data. Complex machines are used to read in biological data at a much faster rate than before. Databases and information systems are used to store and organize biological data. Analyzing biological data may involve algorithms in artificial intelligence, soft computing, data mining, image processing, and simulation. The algorithms in turn depend on theoretical foundations such as discrete mathematics, control theory, system theory, information theory, and statistics. Commonly used software tools and technologies in the field include Java, C#, XML, Perl, C, C++, Python, R, SQL, CUDA, MATLAB, and spreadsheet applications. The aim and purpose of this present review is to discuss and expound on sequence retrieval and analysis as one of the basic and fundamental tools used in the study of Bioinformatics.

**KEYWORDS:** Bioinformatics, sequence alignment, sequence retrieval and analysis.

## 1.0 INTRODUCTION

Besides earning Alfred Sanger his first Noble Prize, the sequencing of insulin inaugurated the modern era of molecular and structural biology. Traditionally a *soft science*, biology got a taste of its first fundamental dataset: molecular sequences. In the early 1960s, known protein sequences accumulated slowly – given that the computers capable of analyzing them had not been developed. In this so-called pre-computer era, sequences were assembled, analyzed, and compared by (manually) writing them on pieces of paper, taping them side by side on laboratory walls, and /or moving them around for optimal alignment (now called *pattern matching*) (Claverie and Nortredame, 2007). As soon as the early computers became available the first computational biologists started to enter these manual algorithms into the memory banks. This practice was brand new – nobody before them had to manipulate and analyze molecular sequences as *texts*. Most methods had to be invented from scratch, and in the process, a new area of research – the analysis of protein sequences using computers – was generated. This was the genesis of bioinformatics. Over the years, the amount of DNA and protein sequence data generated in Biology has grown enormously and even so, continues to grow exponentially. As these sequences are discovered, they are typically placed in public online databases where researchers can readily access them. Expectedly, the available public DNA and protein sequence databases are huge. In bioinformatics therefore, sequence retrieval and analysis are very important.

The study of the process of biological information transfer is important for the understanding of the metabolic, pathological, physiological, anatomical and biochemical activities of living organisms. Living organisms as an assemblage of various organ-systems which are built up by primary precursors, called *cells*, vary from one species to another. The variations may be due to inherited (genetic) or acquired (environmental) factors. The genetic factor dominates the environmental factors. Various fields of science including molecular biology, biochemistry, and biotechnology have contributed immensely towards elucidating the causes, differences, and effects of genetic variations among organisms of the same and different species.

Sequence analysis refers to the process of subjecting a DNA, RNA or peptide sequence to any analytical procedure aimed at understanding its features, function, structure, or evolution. A number of methodologies are available for carrying out sequence analysis. They include sequence alignment, searches against databases, and others. Sequence analysis is especially important because the rate of addition of new gene and protein sequences to the databases has increased exponentially. However, by comparing these new sequences to those with known functions, the understanding of the biology of the organism from which the new sequence came can be achieved. Sequence analysis, therefore, can be used to assign functions to genes and proteins by the study of the similarities between the compared sequences. Many tools and techniques are available lately to provide sequence comparisons (sequence alignment) and analyze the alignment product to understand its biology.

The development of sequence analysis and retrieval methods has depended on the contributions of many individuals (scientists) from varied scientific backgrounds. It is important to be able to retrieve quickly and easily any number of sequences matching any given specific criteria from the huge amount of data that can be found in the database.

The central dogma of molecular biology shows the expression of genes from DNA to RNA and then to protein. It involves the process of transfer of genetic information from **DNA to DNA (replication)**, **DNA to RNA (transcription)**, and **RNA to protein (translation)**. These biomolecules contain sequences of nucleotides (nucleobases) which need to be studied and analysed in order to decipher the precise order of the nucleobases in them in a process of *sequence retrieval and analysis.* When these sequences are retrieved and analysed, they form a viable tool in the study of heredity and variation, diagnosis and treatment of diseases and production of drugs.

## 2.0 SEQUENCE RETRIEVAL SYSTEM

This is the operation of accessing the precise order of gene in a DNA, RNA and protein. Sequence information comes from many sources in which some are reliable than others in different aspects of sequence curation. Many methods have been employed in determining the genomic sequence including the Sangers, manual and automated methods but the recent knowledge of bioinformatics proffered an efficient and effective way of accessing the sequence using computer. When the name of a gene or its ID number is given, it is possible to find and retrieve its DNA sequence. Sequence of genes are given an accession number as identification for database processing. Each sequence has its own unique accession number, but there may be some sequences that have more than one accession number. The most consistent source of sequence data comes from sequencing centres. The foundation of online computer database for storing and distributing sequence data has made bioinformatics an invaluable tool for sequence retrieval.

### 2.1 Types of Sequence Retrieval Databases

There are different types of sequence retrieval database. The main resources for sequence retrieval are three large databases called ***global nucleotide sequence storage.*** They include the following:

i. National Centre for Biotechnology Information (NCBI) database – (www.ncbi.nlm.nih.gov/)
ii. European Molecular Biology Laboratory (EMBL) database – (www.ebi.ac.uk/embl/)
iii. DNA Database of Japan (DDBJ) database – (www.ddbj.nig.ac.jp/)

They collect all publicly available DNA, RNA and protein sequence data and make it available for free. Due to their daily exchange of data, they contain essentially the same data.
Other sequence database are ***genome centered database*** and ***protein database.***

**Genome centered database includes the following.**

i. NCBI genomes: Entrez Life Sciences Search Engine (US National Institutes of Health)- www.ncbi.nlm.nih.gov/sites/gquery
ii. Ensemble genome browser( European Bioinformatics Institute )- www. ensemble.org
iii. UCSC genome bioinformatics site (University of California at Santa Cruz) - www.genome.ucsc.edu.

**Protein database includes**

i. Swiss-Prot
ii. TrEMBL
iii. PDB

### 2.2 Sequence retrieval in a database - NCBI

The National Centre for Biotechnology Information (NCBI) is U.S. government- funded national resource for molecular biology information. It advances science and health by providing access to biomedical and genomic information. NCBI database contains essentially the same data as in the EMBL/DDBJ databases. Sequences in the NCBI Sequence Database (or EMBL/DDBJ) are identified by an accession number. This is a unique number that is only associated with one sequence. For example, the accession number NC_001477 is for the DEN-1 Dengue virus genome sequence. The accession number is what identifies the sequence. It is reported in scientific papers describing that sequence. As well as the sequence itself, for each sequence the NCBI database (or EMBL/DDBJ databases) also stores some additional annotation data, such as the name of the species it comes from, references to publications describing that sequence, etc. Some of this annotation data was added by the person who sequenced the sequence and submitted it to the NCBI database, while some may have been added later by a human curator working for NCBI. The NCBI database contains several sub-databases, the most important of which are the "*NCBI Nucleotide database"* which contains DNA and RNA sequences, the "*NCBI Protein database"* which contains protein sequences, "*EST*" contains ESTs (expressed sequence tags), which are short sequences derived from mRNAs , the "*NCBI Genome database"*which contains DNA sequences for whole genomes. *PubMed* contains data on scientific publications.

The sequence retrieval system is a homogeneous interface to over 80 biological databases that had been developed at the European Bioinformatics Institute (EBI) at Hinxton, UK. It includes databases of sequences, metabolic pathways, transcription factors, application results (like BLAST, SSEARCH, FASTA), protein 3-D structures, genomes, mappings, mutations, and locus specific mutations (Biaroch et al., 1997; Londin et al., 2013). The foundation of online computer database for storing and distributing sequence data has made bioinformatics an invaluable tool for sequence retrieval.

## 3.0 SEQUENCE ANALYSIS

### 3.1 Sequence Alignment

In bioinformatics, a **sequence alignment** is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Sequence alignments are also used for non-biological sequences, such as those present in natural language or in financial data. In bioinformatics, alignment of two or more DNA or protein sequences is a very common task. This is because an important goal of genomics is to determine if a particular sequence is like another sequence. Sequence alignment can be therefore done to compare a sequenced gene of unknown function with similar sequences from a database so as to identify the gene function. It also provides information on the degree of similarity between two sequences and an estimate of the length of time within which they diverged (Johnson and Doolittle, 1986; Sanger et al., 2007). Sequence alignment is also important as a common first step in other processes like phylogeny building.

The alignment process will uncover those regions that are identical or closely similar and those regions with little (or any) similarity. Conserved regions might represent motifs that are essential for function ( Pearson, 1990; Bowie and Eisenberg, 1991; Tosto and Reitz, 2013 and Veron, 2014) . Regions with little similarity could be less essential to function. In a sense, these alignments are used to determine if a database contains a potential homologous sequence to the newly derived sequence. Further, phylogenetic studies are necessary to determine the orthologous/paralogous nature of the two aligned sequences ( Saitou and Nei, 1987; Carvajal-Rodriguez, 2012 and Aston et al., 2014).

### 3.1.1 Types of Alignments

Sequence alignment may involve aligning two sequences at a time, called ***pairwise sequence alignment***, or more than two sequences at a time, called ***multiple sequence alignment***. An alignment basically is used for building phylogenetic trees, looking for sites of interest/conservation within a gene (motifs, binding sites, etc., identifying positive/negative selection and references for short read analysis. (Gibrat, 1996). Pairwise alignment is a tool designed for performing sequence alignments in a wide variety of combinations. It implements sequence to sequence, sequence to profile and profile to profile alignments with optional support of secondary structure.

A multiple sequence alignment is one of more than two sequences. Homologous sites between sequences are aligned. This is achieved by inserting gaps. Alignments allow for identification of regions of similarity between sequences. They identify indels (insertion and deletions) caused by DNA/RNA replication. Multiple sequence alignment is progressive. Usually, a pairwise alignment is done and a clustering method is used to create a guide tree. The guide tree is used to create a succession of pairwise alignments starting with the two closest sequences and ending with the most distant from these.
Pairwise alignment methods are important largely in the context of a database search but for the analysis of individual protein families, ***multiple alignment*** methods are critical. The main principle underlying popular algorithms for multiple alignments is hierarchical clustering that roughly approximates the phylogenetic tree and guides the alignment (Nussinov and Jacobson, 1980; Zuker and Stiegler,1981; Fleischmann et al., 1995). The sequences are first compared using a fast method (e.g. FASTA) and clustered by similarity scores to produce a guide tree. Sequences are then aligned step-by-step in a bottom-up succession, starting from terminal clusters in the tree and proceeding to the internal nodes until the root is reached. Once two sequences are aligned, their alignment is fixed and treated essentially as a single sequence with a modification of dynamic programming. Thus, the hierarchical algorithms essentially reduce the $O\ (n^{k})$ multiple alignment problem to a series of $O\ (n^{2})$ problems, which makes the algorithm feasible but potentially at the price of alignment quality. The hierarchical algorithms attempt to minimize this problem by starting with most similar sequences where the likelihood of incorrect alignment is minimal, in the hope that the increased weight of correctly aligned positions precludes errors even on the subsequent steps. The most commonly used method for hierarchical multiple alignment is Clustal, which is currently used in the ClustalW or ClustalX variants and available at.
http://www.ebi.ac.uk/clustal,http://clustalw.genome.ad.jp,andhttp://www.bork.emblheidelberg.de/Alignment/alignment.html).

ClustalW is fast and tends to produce reasonable alignments, even for protein families with limited sequence conservation, provided the compared proteins do not differ in length too much. A combination of length differences and low sequence conservation tends to result in gross distortions of the alignment. The T-Coffee program is a modification of ClustalW that incorporates heuristics partially solving these problems (Gibrat, 1996).

### 3.1.2 Alignment Methods

Computational approaches to sequence alignment generally fall into two categories: *global alignments* and *local alignments*. Calculating a global alignment is a form of global optimization that "forces" the alignment to span the entire length of all query sequences. By contrast, local alignments identify regions of similarity within long sequences that are often widely divergent overall. Local alignments are often preferable, but can be more difficult to calculate because of the additional challenge of identifying the regions of similarity. Local alignments use heuristic programming methods that are better suited to successfully search very large databases, but they do not necessarily give the most optimum solution. Even given this limitation, local alignments are very important to the field of genomics because they can uncover regions of homology that are related by descent between two otherwise diverse sequences (Abagyan and Batalov, 1997; Altschul and Gish, 1996; Adams et al., 2000).The global approach is useful when one is comparing a small group of sequences, but becomes become computationally expensive as the number of sequences in the comparison increases.

### 3.2 Dot plots

In bioinformatics, the use of a similarity matrix, called a dot plot, is one way to visualize the similarity between two nucleic acid or protein sequences. A dot plot can be defined as a graphical method that allows the comparison of two biological sequences so as to identify regions of close similarity between them. Simply put, a dot plot is a matrix that is used to visually detect alignments. It was introduced by Gibbs and McIntyre (1970) and is one of the oldest but straightforward ways of comparing two sequences.

The dot plot method involves constructing a matrix with one of the sequences to be compared running horizontally across the bottom, and the other running vertically along the left-hand side. In other words, the dot plots compare two sequences by organizing one sequence on the x-axis and another on the y-axis of a plot. Orientation however does not matter as the second sequence can be put on the bottom and the first one on the right side. Each entry of the matrix is a measure of similarity of those two residues on the horizontal and vertical sequence. In the Gibbs and McIntyre paper, they used the simplest scoring system which distinguishes only between identical (dots) and non-identical (blank) residues. However, one can also use graded measures that give chemically similar pairs of bases, higher similarity scores such as the BLOSUM and PAM matrices and enter a dot whenever the similarity exceeds a prescribed value. Similar sequences tend to have many identical or chemically related residues along the main diagonal; hence conspicuous diagonal runs of dots signal regions of similarity. Some idea of the similarity of the two sequences can be gleaned from the number and length of matching segments shown in the matrix. Identical sequences will obviously have a diagonal line in the center of the matrix. Insertions and deletions between sequences give rise to disruptions in this diagonal. Regions of local similarity or repetitive sequences give rise to further diagonal matches in addition to the central diagonal.

Simple as it is, dot matrix analysis is still a popular tool for researchers to visually inspect the similarity between two sequences. It is often used as a first examination. From its output, the researcher can pick out regions from the two sequences on which more detailed alignment will be performed. However, doing alignments by hand, as in dot plots, only works for very short sequences.

## 4.0 SEQUENCE DATABASE SEARCH ALGORITHMS

Initial characterization of any new DNA or protein sequence starts with a database search aimed at finding out whether homologs of this gene (protein) are already available, and if they are, what is known about them (Ewing and Green, 1998; Abu-Jamous et al., 2013a). Clearly, looking for **exactly** the same sequence is quite straightforward. One can just take the first letter of the query sequence, search for its first occurrence in the database, and then check if the second letter of the query is the same in the subject. If it is indeed the same, the program could check the third letter, then the fourth, and continue this comparison to the end of the query. If the second letter in the subject is different from the second letter in the query, the program should search for another occurrence of the first letter, and so on. This will identify all the sequences in the database that are identical to the query sequence (or include it). This approach, however, is primitive computation-wise, and there are sophisticated algorithms for text matching that handle it much more efficiently.

Practically, these sequence comparisons programs search for ***high-scoring segment pairs (HSPs)***, instead of looking for perfect matches. HSPs are fragments of the alignment of two sequences whose similarity score cannot be improved by adding or trimming any letters. Once a set of HSPs is found, different methods, such as Smith-Waterman, FASTA, or BLAST, deal with them in different fashions.

### 4.1 Smith-Waterman algorithm

Any pairwise sequence alignment method in principle can be used for database search in a straightforward manner. All that needs to be done is to construct alignments of the query with each sequence in the database, one by one, rank the results by sequence similarity, and estimate statistical significance (Gibrat et al., 1996; Felsenstein, 1988 and Smith and Waterman, 1981).

The classic Smith-Waterman algorithm is a natural choice for such an application, and it has been implemented in several database search programs. The most popular one is SSEARCH which was written by

William Pearson and distributed as part of the FASTA package. It is currently available on numerous servers around the world. The major problem preventing SSEARCH and other implementations of the Smith-Waterman algorithm from becoming the standard choice for routine database searches is the computational cost, which is an order of magnitude greater than it is for the heuristic FASTA and BLAST methods. Since extensive comparisons of the performance of these methods in detecting structurally relevant relationships between proteins failed to show a decisive advantage of SSEARCH, the fast heuristic methods dominate the field. Nevertheless, on a case-by-case basis, it is certainly advisable to revert to full Smith-Waterman search when other methods do not reveal a satisfactory picture of homologous relationship for a protein of interest. A modified, much faster version of the Smith-Waterman algorithm has been implemented in the MPSRCH program, which is available at the EBI web site (http://www.ebi.ac.uk/MPsrch).

### 4.2 FASTA

FASTA, introduced in 1988 by William Pearson and David Lipman, was the first database search program that achieved search sensitivity comparable to that of Smith-Waterman but was much faster. FASTA looks for biologically relevant global alignments by first scanning the sequence for short exact matches called “words”; a word search is extremely fast. The idea is that almost any pair of homologous sequences is expected to have at least one short word in common. Under this assumption, the great majority of the sequences in the database that do not have common words with the query can be skipped without further examination with a minimal waste of computer time. The sensitivity and speed of the database search with FASTA are inversely related and depend on the “k-tuple” variable, which specifies the word size; typically, searches are run with $k$ = 3, but, if high sensitivity at the expense of speed is desired, one may switch to $k$ = 2. Subsequently, Pearson introduced several improvements to the FASTA algorithm, which are implemented in the FASTA3 program available on the EBI server at http://www2.ebi.ac.uk/fasta3. A useful FASTA-based tool for comparing two sequences, LALIGN, is available at http://fasta.bioch.virginia.edu/fasta/lalign2.htm.

A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line (defline) is distinguished from the sequence data by a greater-than (">") symbol at the beginning. It is recommended that all lines of text be shorter than 80 characters in length.

**An example sequence in FASTA format is.**

```
>gi|129295|sp|P01013|OVAX_CHICK GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFN
AEDTREMPFHVTKQESKPVQMMCMNNSFNVATL
PAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINF
EKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAF
MELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP
```

Blank lines are not allowed in the middle of FASTA input. Sequences are expected to be represented in the standard IUB/IUPAC amino acid and nucleic acid codes, with these exceptions: lower-case letters are accepted and are mapped into upper-case; a single hyphen or dash can be used to represent a gap of indeterminate length; and in amino acid sequences, U and * are acceptable letters (see below). Before submitting a request, any numerical digits in the query sequence should either be removed or replaced by appropriate letter codes (e.g., N for unknown nucleic acid residue or X for unknown amino acid residue). The nucleic acid codes supported are.

| | | | | | |
|---|---|---|---|---|---|
| **A** | **adenosine** | **C** | **cytidine** | **G** | **guanine** |
| **T** | **thymidine** | **N** | **A/G/C/T (any)** | **U** | **uridine** |
| ***K*** | ***G/T (keto)*** | ***S*** | ***G/C (strong)*** | ***Y*** | ***T/C (pyrimidine)*** |
| ***M*** | ***A/C (amino)*** | ***W*** | ***A/T (weak)*** | ***R*** | ***G/A (purine)*** |
| ***B*** | ***G/T/C*** | ***D*** | ***G/A/T*** | ***H*** | ***A/C/T*** |
| ***V*** | ***G/C/A*** | ***-*** | ***gap of indeterminate length*** | | |

**For those programs that use amino acid query sequences (BLASTp and tBLASTn), the accepted amino acid codes are.**

| | | | |
|---|---|---|---|
| A | alanine | P | proline |
| B | aspartate/asparagine | Q | glutamine |
| C | cystine | R | arginine |
| D | aspartate | S | serine |
| E | glutamate | T | threonine |
| F | phenylalanine | *U* | *selenocysteine* |
| G | glycine | V | valine |
| H | histidine | W | tryptophan |
| I | isoleucine | Y | tyrosine |
| K | lysine | Z | glutamate/glutamine |
| L | leucine | X | any |
| M | methionine | * | translation stop |
| N | asparagine | - | *gap of indeterminate length* |

**NOTE**

1. The degenerate nucleotide codes in red are treated as mismatches in nucleotide alignment. Too many such degenerate codes within an input nucleotide query will cause *blast.cgi* to reject the input. For protein queries, too many nucleotide-like code (A,C,G,T,N) may also cause similar rejection.
2. For protein code, U is replaced by X first before the search since it is not specified in any scoring matrices.
3. *blast.cgi* will not take "-" in the query. To represent gaps, use a string of N or X instead.

### 4.3 Basic Local Alignment Search Tool (BLAST)

Basic Local Alignment Search Tool (BLAST) is the most common and widely used local alignment tool developed by Altschul *et al.* (1990). It is the most widely used method for sequence similarity search; it is also the fastest one and the only one that relies on a complete, rigorous statistical theory. It is a pairwise alignment program that compares nucleotide or protein sequences to sequence databases with the primary purpose of finding regions of local similarity between them. The program also calculates the statistical significance of matches. It is quite a bit more complicated than the simple scoring algorithm (dot plots). Like FASTA, and in contrast to the Smith-Waterman algorithm, BLAST employs the word search heuristics to quickly eliminate irrelevant sequences, which greatly reduces the search time.

In very general terms, BLAST takes one of the sequences and splits it up into words, or tuples. One can decide how big to make the words. The default for nucleotides is a size of 11. For amino acids, the default word size is 3. For example, the amino acid sequence, GEQPM, can be divided into 3 tuples each with size 3: GEQ, EQP, and QPM. Each word is then slid against the other sequence. Each alignment is given a score. If the score is above a specified threshold, the alignment is retained. Each local alignment is then extended and measured using a ***high-scoring segment pair*** (HSP) score. If there are two or more HSP regions, they will be combined, if possible. If it is not possible, one may end up with multiple HSPs. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families (Altschul et al., 1997).

The BLAST is a set of algorithms that attempt to find a short fragment of a ***query*** sequence that aligns perfectly with a fragment of a ***subject*** sequence found in a ***database***. Put differently, BLAST looks for short sequences in the query that matches short sequences found in the database. That initial alignment must be greater than a ***neighborhood score threshold (T)***. For the original BLAST algorithm, the fragment is then used as a seed to extend the alignment in both directions. The alignment is extended in both directions until the T score for the aligned segment does not continue to increase.

The first step of the BLAST algorithm is to break the query into short ***words*** of a specific length. A word is a series of characters from the query sequences. The default length of the search is three (3) characters. The words are constructed by using a sliding window of three characters. For example, twelve amino acids near the amino terminal of the *Aradbidopsis thaliana* protein phosphoglucomutase sequence are.

**NYLENFVQATFN**

This sequence is broken down into three character words by selecting the first amino acid characters, moving over one character, selecting the next three amino acid characters, and so on to create the following words.

**NYL YLE LEN ENF NFV FVQ VQA QAT ATF TFN**

These words are then compared against a sequence in a database, and this search is performed for all words. For the original BLAST search, those words whose T value was greater than 18 were used as seeds to extend the alignment.

The T value is derived by using a scoring matrix. The BLOSUM 62 matrix is the default for protein searches. The alignment is extended in both directions until the alignment score decreases in value. Those alignments whose T score does not decrease are then compared with scores obtained by random searches. Those alignments whose score is above the cutoff are called a ***High Scoring Segment Pair*** (HSP). Once this alignment process is completed for a query and each subject sequence in the database, a report is generated. This report provides a list of those alignments (default size of 50) with a value greater than the S cutoff value.

For each alignment reported, an ***Expectation (e) Value*** is reported. This value is a function of the S value and the database size. An e value of 1 means that one alignment using a query of this size will by chance produce a S score of this value in a database of this size. As one can imagine an e value of –10 (=1x10$^{-10}$) means that it is much more unlikely that random chance lead to this current alignment compared to an alignment with an e-value of 1.

The expectation value is often considered to be a probability. In other words, the probability of achieving a score of this value using a sequence of this length against a database of this size is equal to the expectation value. Therefore, a lower e-value means that alignment is significant at a specific probability level.

It is important to note that the expectation value is specific to a database of a certain size. This means, that if one performs the BLAST alignment at a later date, the e-value might change because the size of the database has changed.

In general, if one obtains an e value of –30, one can be assured that the sequence is homologous to the sequence to which aligned in this database. Furthermore, e values of –5 are often considered significant enough when annotating a genome.

Alignments are also possible between a nucleotide query and a nucleotide database. The entire BLAST process described above is the same for nucleotide searches except the default word size is eleven and a different scoring matrix is applied. Scoring matrices are used to obtain the S value. For nucleotides, these are simple; each identical match is given the same score, and all mismatches are given a penalty (negative) score.

As already mentioned, BLAST is actually a collection of algorithms. So when a BLAST search is done, it is important to specify the type of search that being performed. Outlined below is a summary of the available BLAST programs for the different sequence similarity searches.

- **blastp** compares an amino acid query sequence against a protein sequence database
- **blastn** compares a nucleotide query sequence against a nucleotide sequence database
- **blastx** compares a nucleotide query sequence translated in all reading frames against a protein sequence database
- **tblastn** compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames
- **tblastx** compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

**The following table outlines each algorithm and the nature of the query and database used.**

| Search | Query | Database |
|---|---|---|
| **blastn** | nucleotide | Nucleotide |
| **blastx** | translated nucleotide in all six frames | Protein |
| **blastp** | protein | Protein |
| **tblastx** | translated nucleotide in all six frames | translated nucleotide in all six frames |

The **blastx**, **tblastn**, and **tblastx** programs are used when either the query or the database or both are uncharacterized sequences and the location of protein-coding regions is not known. These programs translate the nucleotide sequence of the query in all six possible frames. BLAST is the first choice program in any situation when a sequence similarity search is required, due to its speed, high selectivity, and flexibility. Importantly, this method is used most often as the basis for genome annotation.

## 5.0 CONCLUSION

Sequence retrieval and analysis is a very useful tool and probably the bedrock of the study of bioinformatics. With this tool the various DNA and protein sequences in different organisms can be studied and analyzed through a number of methodologies ranging from Smith-Waterman algorithm ( SSEARCH), FASTA, BLAST CLUSTALW, COBALT, MUSCLE, ProbCons and others. When these sequences are retrieved and analysed they form a viable tool in the study of heredity and variation, diagnosis and treatment of diseases and production of drugs.

## REFERENCES

1. Abagyan, R. A. and Batalov, S. Do aligned sequences share the same fold? *Journal of Molecular Biology*, 1997; 273: 355 – 368.
2. Abu-Jamous, B, Fa, R, Roberts, D.J. and Nandi A. K. Paradigm of Tunable Clustering Using Binarization of Consensus Partition Matrices (Bi-CoPaM) for Gene Discovery. *PLoS ONE*, 2013; 8(2).
3. Abu-Jamous, B. Fa, R, Roberts, D. J. and Nandi, A.K. Yeast gene CMR1/YDL156W is consistently co-expressed with genes participating In DNA-metabolic processes in a variety of stringent clustering experiments". *J R Soc. Interface*, 2013; 10(81).
4. Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W. Hoskins, R.A. and Galle, R.F. The genome sequence of Drosophila melanogaster. *Science*, 2000; 287: 2185–2195.
5. Altschul, S. F. and Gish, G. Local alignment statistics. *Methods of Enzymology*, 1996; 226: 460 – 480.
6. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.*, 1990; 215: 403–410.
7. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang , J., Zhang, Z., Miller, W., and Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.*, 1997; 25: 3389–3402.
8. Aston, K.I. Genetic susceptibility to male infertility: News from genome-wide association studies. *Andrology*, 2014; 2(3): 315–21.
9. Bairoch, A., Bucher, P. and Hofmann, K. The PROSITE database, its status in *Nucleic Acids Res.*, 1997; 25: 217–221.
10. Bowie, J.U., Luthy, R. and Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 1991; 253: 164–170.
11. Carvajal-Rodríguez, A. Simulation of Genes and Genomes Forward in Time. *Current Genomics*, 2012; 11(1): 58–61.
12. Claverie, J-M. and Nortredame, C. A brief history of sequence analysis. *In* Bioinformatic for dummies. Wiley Publishing, Inc, 2nd Edition. Indiana, USA, 2007: 12.
13. Ewing, B. and Green, P. Base-calling of automated sequence traces using phred. II. Error probabilities. *Genome Res.*, 1998; 8: 186–194.
14. Felsenstein, J. Phylogenies from molecular sequences: Inferences and reliability. *Annu. Rev. Genet.*, 1988; 22: 521–565.
15. Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A. and Merrick, J.M. Whole-genome random sequencing and assembly of Haemophilus influenzae. *Science*, 1995; 269(5223): 496–512.

16. Gibrat, J.F., Madej T. and Bryant, S.H. Surprising similarity in structure comparison. *Curr. Opin. Struct. Biol*., 1996; 6: 377–385.
17. Henikoff, S. and Henikoff, J.G. Amino acid substitution matrices from protein blocks. Proc. *Natl. Acad. Sci*., 1992; 89: 10915–10919.
18. Johnson, M.S. and Doolittle, R.F. A method for the simultaneous alignment of three or more amino acid sequences. *J. Mol. Evol.*, 1986; 23: 267–268.
19. Londin, E., Yadav, P., Surrey, S., Kricka, L.J., and Fortina, P. Use of Linkage Analysis, Genome-Wide Association Studies and Next-Generation Sequencing in the Identification of Disease-Causing Mutations. *Pharmacogenomics*, 2013; 1015: 127–46.
20. Nussinov, R. and Jacobson, A.B. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci.*, 1980; 77: 6903–6913.
21. Pearson, W.R. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzy- mol.*, 1990; 183: 63–98.
22. Saitou, N. and Nei, M. The neighbour-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*., 1987; 4: 406–425.
23. Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M. and Smith, M. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 2007; 265(5596): 687–95.
24. Smith, T.F. and Waterman, M.S. Identification of common molecular subsequence’s. *J. Mol. Biol*., 1981; 147: 195–197.
25. Zuker, M. and Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*., 1981; 9: 133–148.
26. Tosto, G. and Reitz, C. Genome-wide association studies in Alzheimer's disease: A review. *Current Neurology andNeuroscience Reports*, 2013; 13(10): 381.
27. Véron, A., Blein, S. and Cox, D.G. Genome-wide association studies and the clinic: A focus on breast cancer". *Biomarkers in Medicine*, 2014; 8(2): 287–96.
http://www.ebi.ac.uk/clustal,http://clustalw.genome.ad.jp,andhttp://www.bork.emblheidelberg.de/Alignment/alignment.html.