Available online at http://www.ajol.info/index.php/njbas/index
Nigerian Journal of Basic and Applied Science (March, 2013), 21(1): 27-32
DOI: http://dx.doi.org/10.4314/njbas.v21i1.4
ISSN 0794-5698

Full Length Research Article

Retrieval and Representation of Nucleotide Sequence of *Saccharomyces cerevisiae* Cystathionine Gamma-Lyase (CYS3) Gene in Five Formats

***R. A. Umar, H. Abdullahi and N. Lawal**
Department of Biochemistry, Usmanu Danfodiyo University, P.M.B. 2346, Sokoto
[*Corresponding Author, E-mail:rabiu34@yahoo.com; ☎: +2348036154828]

ABSTRACT: Educational programmes all over the world are facing increasing pressure to integrate information technology in the curriculum. Knowledge of bioinformatics is at infancy in Nigeria it is therefore imperative to develop and build the capacity for high-throughput determination and computational analysis of the nucleotide base sequences of the genomes of organisms. The present communication navigated the ENTREZ Web page and downloaded sequences of Cystathionine gamma- lyase gene from *Saccharomyces cerevisiae*. The sequence is then represented in the five best known database formats namely Plain, FASTA, EMBL, GCG and Genebank thereby making it more visible and available for other research applications such as comparative genomic analysis, evolutionary studies, searching for and identification of regulatory elements and scanning for mutations. The present study highlights data retrieval and representation. Data retrieval is important as it provides the opportunity to engage in data mining for discovery, a convenient alternative to traditional wet laboratories, providing biological insights, and proficiency to access and use the vast repository of computational and web-based resources which are the most available information in the world today.
Keywords: Nucleotide, Database, Genome, GenBank.

INTRODUCTION

Educational programmes all over the world are facing increasing pressure to imbibe the new information and communication technologies to teach students the knowledge and skills needed for participation in knowledge economies of the world (Tonukari, 2004). Molecular biology and bioinformatics have made the greatest contributions to biomedical research and in opening up new frontiers in the past two decades. Bioinformatics was previously defined as an interdisciplinary field involving biology, computer science, mathematics and statistics to analyze biological sequence data, genome content and arrangement and to predict the function and structure of macromolecules (Luscombe *et al*., 2001).

Over the past twelve years, the complete or nearly complete genome sequences of some of the most important metazoan organisms-*Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Anopheles gambiae*, *Rattus norvegicus*, *Saccharomyces cerevisiae*, *Oriza sativum*, *Artemisia annua*, *Pan troglodytes*, *Danio rerio*, *Gallus gallus* and *Ciona Savignyi* –have been determined and made publicly available (Ureta-vidal *et al*., 2003). Several other sequences would be released in the next coming years. Most of the organisms whose genomes have been sequenced are relevant to human health, agriculture and the environment in Africa (Masiga and Isokpehi, 2004). Various databases, algorithms, computational and statistical techniques have been developed to enable effective use of the sequence data.

To access and use the vast array of data resulting from sequencing activities in several laboratories all over the developed countries necessitates the development and delivery of several types and levels of training that will enhance the use of these computational resources as genome research tools by African scientists (Masiga and Isokpehi 2004).

Numerous works have dealt with the topic of sequence databases (Tateno *et al*., 1997; Bairoch and Apweiller, 2000; Baker *et al*., 2000a; Benson *et al*., 2000). These publications emphasize the great rate at which the databases have grown, and they suggest various ways of utilizing such vast biological resources. Such information is scarce in Nigerian journals. Even in the rest of Africa extremely wide disparities exist in human resources and infrastructure for access and utilization of genome data (Masiga and Isokpehi 2004).

GenBank is one of the repositories of DNA sequences of organisms. As early as 1999 it houses million of sequences of more than 18,000 organisms. In the

GenBank, information describing each sequence entry is given, including literature references, information about the function of the sequence, location of mRNA and coding regions, and position of important mutations (Benson *et al*., 2000). This information is organised in to fields, each with an identifier, shown as the first text on each line. In some entries, these identifiers may be abbreviated to two letters, e.g. RF for references, and some identifiers may have additional subfields.

The sequence includes number on each line so that sequence positions can be located by eye, because the sequence position count or a sequence checksum value may be used by the computer programmes to verify the sequence's composition. The GenBank sequence format often has to be changed for use with sequence analysis software (Pearson and Lipman, 1988).

The European Molecular Biology Laboratory (EMBL) maintains DNA and protein sequences database. The EMBL sequence format is similar to the GenBank format. The main differences are in the use of the term ORIGIN in the GenBank format to indicate the start of sequence; also the EMBL entry does not include the sequence of any translation product, which is shown instead as different entry in the database (Baker *et al*., 2000b). This EMBL sequence format often has to be changed for use with sequence analysis software. The FASTA sequence format consists of three parts: a comment line identified by a > character in the first column, the sequence in standard one letter symbols and an optional * indicating the end of the sequence that may or may not be present.

The FASTA format is the sequence most often used by sequence analysis software. This format provides a very convenient way to copy just the sequence part from one window to another because there are no numbers or other non sequence character within the sequence. This is the only acceptable format for submission of sequences to servers for sequence alignment and comparison (Altschul *et al*., 1990).

Genetic Computer Group (GCG) Sequence Format
GCG programmes require a unique sequence format and include programmes that convert other sequence format in to the GCG format (Benson *et al*., 2000). Letter versions of GCG accept several sequence formats. Information about the sequence in GenBank entry is first included, followed by a line of information about the sequence and checksum value. The value is provided as a check on the accuracy of the sequence by the addition of the ASCII values of the sequence. If the sequence has not been changed, this value should stay the same. If one or more sequence characters become changed through error, a programme reading the sequence will be able to determine that the change has occurred because the checksum value in the sequence entry will no longer be correct. Lines of information are terminated by two periods, which mark the end of information and the start of the sequence on the next line, the rest of the text in the entry is treated as sequence (Benson *et al*., 2000). Note the presence of line numbers because there is no symbol to indicate the end of the sequence. No text other than sequence should be altered except by programmes that will also adjust the checksum score for the sequence. The GCG sequence format may have to be changed for use with other sequence analysis software. GCG also includes programmes for reformatting sequence file (Benson *et al*., 2000).

Cystathionine gamma lyase gene of *S.cerevisiae* is a single open reading frame consisting of 1182 bp (394 amino acid residues) located on chromosome XII (Bork *et al*.,1992). Its gene symbol is MET 17 or MET 17 P. It encodes o-acetyl-L-serine o-acetyl –L-homoserine sulfhydrylase (EC 4.2.99.10) enzyme which catalyses the *de novo* synthesis of L-cysteine and o-alkyl-L-homoserine in some microorganisms. It recycles the methylthio group of methionine (Yagamata *et al*., 1994). Disruption of the gene leads to cysteine auxotrophy.

METHOD

ENTREZ is a window compatible and user friendly site accessible through the National Centre for Biotechnology Information under the auspices of the National Institute of Health, USA. It integrates data from the major DNA and protein sequence databases along with taxonomy, genome, mapping and protein structure information. Also available is biomedical literature via Pub Med .After connecting to the Internet through the Internet Explorer the following Universal Research Locator (URL) was typed :http://www.ncbi.n/m.nih.gov/entrez. The web page was opened and navigation was continued by selecting the specific text and clicking the Entrez web pages. The nucleotide database was selected and clicked. The complete nucleotide (DNA) sequence of cystathionine gamma lyase gene from *Saccharomyces cerevisiae,* was downloaded.

The gene sequence was then represented according to the five database file formats in compliance with the specific rules for each format; Plain, GCG, EMBL Gene bank and FASTA.

RESULTS

The results are presented in Tables 1-5. Table 1 presents the sequence of the gene in plain format. The sequence conforms fully with the features of the format. Table 2 presents the sequence of the cystathionine gamma lyase gene according to EMBL file format included are additional information such as identification (ID) number, accession (AC) number, description (DE) and sequence (SQ). The sequence of the cystathionine gamma lyase gene according to GenBank file format is depicted in Table 3. The file format has unique features such as locus, and origin of the sequence and others such as general identification (gi) number, accession (AC) number, definition (DE) and sequence (SQ). Table 4 depicts the sequence in FASTA format. The FASTA format is used in a variety of molecular biology software suites. The ''greater than'' character (>) designates the beginning of a new file. An identifier (L04459) is followed by the DNA sequence in uppercase letters (it is allowed to be written in lowercase), usually with 60 characters per line. Users and databases can then, if they wish, add a certain degree of complexity to this format. For example, without breaking any of the rules just outlined, one could add more information to the FASTA definition line, making the simple format a little more informative.

Table 1: Sequence Arrangement of *Saccharomyces cerevisiae* Cystathionine Gamma-Lyase (CYS3) Gene in Plain Format

GCAGCGCACGACAGCTGTGCTATCCCGGCGAGCCCGTGGCAGAGGACCTCGCTTGCGAAAGCATCGAGTAC
CGCTACAGAGCCAACCCGGTGGACAAACTCGAAGTCATTGTGGACCGAATGAGGCTCAATAACGAGATTAGC
GACCTCGAAGGCCTGCGCAAATATTTCCACTCCTTCCCGGGTGCTCCTGAGTTGAACCCGCTTAGAGACTCCG
AAATCAACGACGACTTCCACCAGTGGGCCCAGTGTGACCGCCACACTGGACCCCATACCACTTCTTTTTGTTA
TTCTTAAATATGTTGTAACGCTATGTAATTCCACCCTTCATTACTAATAATTAGCCATTCACGTGATCTCAGCCA
GTTGTGGCGCCACACTTTTTTTTCCATAAAAATCCTCGAGGAAAAGAAAAGAAAAAAATATTTCAGTTATTTAA
AGCATAAGATGCCAGGTAGATGGAACTTGTGCCGTGCCAGATTGAATTTTGAAAGTACAATTGAGGCCTATAC
ACATAGACATTTGCACCTTATACATATAC

Table 2: Sequence Arrangement of *Saccharomyces cerevisiae* Cystathionine Gamma-Lyase (CYS3) Gene, in EMBL Format

ID gi|171361|gb|L04459|YSCCYS3A
AC gi|171361|gb|L04459|YSCCYS3A
DE *Saccharomyces cerevisiae* cystathionine (CYS3) gene, complete cds.
SQ Sequence 541bp
GCAGCGCACGACAGCTGTGCTATCCCGGCGAGCCCGTGGCAGAGGACCTCGCTTGCGAAAGCATCGAGTAC
CGCTACAGAGCCAACCCGGTGGACAAACTCGAAGTCATTGTGGACCGAATGAGGCTCAATAACGAGATTAGC
GACCTCGAAGGCCTGCGCAAATATTTCCACTCCTTCCCGGGTGCTCCTGAGTTGAACCCGCTTAGAGACTCCG
AAATCAACGACGACTTCCACCAGTGGGCCCAGTGTGACCGCCACACTGGACCCCATACCACTTCTTTTTGTTA
TTCTTAAATATGTTGTAACGCTATGTAATTCCACCCTTCATTACTAATAATTAGCCATTCACGTGATCTCAGCCA
GTTGTGGCGCCACACTTTTTTTTCCATAAAAATCCTCGAGGAAAAGAAAAGAAAAAAATATTTCAGTTATTTAA
AGCATAAGATGCCAGGTAGATGGAACTTGTGCCGTGCCAGATTGAATTTTGAAAGTACAATTGAGGCCTATAC
ACATAGACATTTGCACCTTATACATATAC*//*

Table 3: Sequence Arrangement of *Saccharomyces cerevisiae* Cystathionine Gamma-Lyase (CYS3) Gene, in **GenBank** Format

LOCUS gi|171361|gb|L04459|YSCCYS3A 541bp DNA 26-Aug- 2010
DEFINITION *Saccharomyces cerevisiae* cystathionine gamma-lyase (CYS3)gene, complete cds.
ACCESSION gi|171361|gb|L04459|YSCCYS3A
ORIGIN
GCAGCGCACGACAGCTGTGCTATCCCGGCGAGCCCGTGGCAGAGGACCTCGCTTGCGAAAGCATCGAGTAC
CGCTACAGAGCCAACCCGGTGGACAAACTCGAAGTCATTGTGGACCGAATGAGGCTCAATAACGAGATTAGC
GACCTCGAAGGCCTGCGCAAATATTTCCACTCCTTCCCGGGTGCTCCTGAGTTGAACCCGCTTAGAGACTCCG
AAATCAACGACGACTTCCACCAGTGGGCCCAGTGTGACCGCCACACTGGACCCCATACCACTTCTTTTTGTTA
TTCTTAAATATGTTGTAACGCTATGTAATTCCACCCTTCATTACTAATAATTAGCCATTCACGTGATCTCAGCCA
GTTGTGGCGCCACACTTTTTTTTCCATAAAAATCCTCGAGGAAAAGAAAAGAAAAAAATATTTCAGTTATTTAA
AGCATAAGATGCCAGGTAGATGGAACTTGTGCCGTGCCAGATTGAATTTTGAAAGTACAATTGAGGCCTATAC
ACATAGACATTTGCACCTTATACATATAC//

Table 4: Sequence Arrangement of *Saccharomyces cerevisiae* Cystathionine Gamma-Lyase (CYS3) Gene, in **FASTA** Format

>gi|171361|gb|L04459|YSCCYS3A *Saccharomyces cerevisiae* cystathionine gamma-lyase|len (CYS3) gene, complete cds.541bp
GCAGCGCACGACAGCTGTGCTATCCCGGCGAGCCCGTGGCAGAGGACCTCGCTTGCGAAAGCATCGAGTAC
CGCTACAGAGCCAACCCGGTGGACAAACTCGAAGTCATTGTGGACCGAATGAGGCTCAATAACGAGATTAGC
GACCTCGAAGGCCTGCGCAAATATTTCCACTCCTTCCCGGGTGCTCCTGAGTTGAACCCGCTTAGAGACTCCG
AAATCAACGACGACTTCCACCAGTGGGCCCAGTGTGACCGCCACACTGGACCCCATACCACTTCTTTTTGTTA
TTCTTAAATATGTTGTAACGCTATGTAATTCCACCCTTCATTACTAATAATTAGCCATTCACGTGATCTCAGCCA
GTTGTGGCGCCACACTTTTTTTTCCATAAAAATCCTCGAGGAAAAGAAAAGAAAAAAATATTTCAGTTATTTAA
AGCATAAGATGCCAGGTAGATGGAACTTGTGCCGTGCCAGATTGAATTTTGAAAGTACAATTGAGGCCTATAC
ACATAGACATTTGCACCTTATACATATAC//

Table 5: Sequence Arrangement of *Saccharomyces cerevisiae* Cystathionine Gamma-Lyase (CYS3) Gene, in **GCG** Format

ID gi|171361|gb|L04459|YSCCYS3A
AC gi| sequence 171361|gb|L04459|YSCCYS3A;
DE *Saccharomyces cerevisiae* cystathionine gamma-lyase (CYS3)gene, complete cds.
SQ Sequence
gi|171361|gb|L04459|YSCCYS3A Length 541: check:..
541bp
GCAGCGCACGACAGCTGTGCTATCCCGGCGAGCCCGTGGCAGAGGACCTCGCTTGCGAAAGCATCGAGTAC
CGCTACAGAGCCAACCCGGTGGACAAACTCGAAGTCATTGTGGACCGAATGAGGCTCAATAACGAGATTAGC
GACCTCGAAGGCCTGCGCAAATATTTCCACTCCTTCCCGGGTGCTCCTGAGTTGAACCCGCTTAGAGACTCCG
AAATCAACGACGACTTCCACCAGTGGGCCCAGTGTGACCGCCACACTGGACCCCATACCACTTCTTTTTGTTA
TTCTTAAATATGTTGTAACGCTATGTAATTCCACCCTTCATTACTAATAATTAGCCATTCACGTGATCTCAGCCA
GTTGTGGCGCCACACTTTTTTTTCCATAAAAATCCTCGAGGAAAAGAAAAGAAAAAAATATTTCAGTTATTTAA
AGCATAAGATGCCAGGTAGATGGAACTTGTGCCGTGCCAGATTGAATTTTGAAAGTACAATTGAGGCCTATAC
ACATAGACATTTGCACCTTATACATATAC//

DISCUSSION

The most widely available information in today's world is in textual form in the World Wide Web and proficiency in information retrieval is needed for its productive use. The information is in sequences either of DNA or amino acids. These sequences contain a wealth of information hidden within them, including protein structure, disease mechanisms and drug target sites. The challenges facing researchers in biomedical and pharmaceutical disciplines are how to extract biologically useful information from millions of sequences. This is what bioinformatics tries to address, using a multidisciplinary approach which combines computer science, information theory and molecular biology. Hence bioinformatics is enhancing the use of genome data and the associated computational resources in basic and applied research in biotechnology and biomedical sciences.

There are two main reasons for putting data on a computer: retrieval and discovery. Retrieval is basically being able to get back out what was put in. Amassing sequence information without providing a way to retrieve it makes the sequence information, in essence, useless. Although this is important, it is even more valuable to be able to get back from the system more knowledge than was put in to begin with—that is, to be able to use the information to make biological discoveries.

Most biologists are familiar with the use of animal models to study human diseases. Although a disease that occurs in humans may not be found in exactly the same form in animals, often an animal disease shares enough attributes with a human counterpart to allow data gathered on the animal disease to be used to make inferences about the process in humans. Mathematical models describing the forces involved in musculoskeletal motions can be built by imagining that muscles are combinations of springs and hydraulic pistons and bones are lever arms, and, often times,such models allow meaningful predictions to be made and tested about the obviously much more complex biological system under consideration. The more closely and elegantly a model follows a real phenomenon, the more useful it is in predicting or understanding the natural phenomenon it is intended to mimic. In the same vein, the National Center for Biotechnology Information (NCBI) introduced a new model for sequence-related information. This new and more powerful model made possible the rapid development of software and the integration of databases that underlie the popular Entrez retrieval system and on which the GenBank database is now built. The advantages of the model (e.g., the ability to move effortlessly from the published literature to DNA sequences to the proteins they encode, to chromosome maps of the genes, and to the three-dimensional structures of the proteins) have been apparent for years to biologists using Entrez, but very few biologists understand the foundation on which this model is built. As genome information becomes richer and more complex, more of the real, underlying data model is appearing in common representations such as GenBank files, EMBL, GCG etc.

One of the important task for molecular biologists and bioinformaticians is to decipher biological information from DNA sequences (Mount,2004). As more DNA sequences become available in the late 1970s interest in developing computer programmes grew to analyze the sequences in various ways and to store them in databases. Plain sequence format is a computer file that includes only the sequence with no other accessory information. The sequence must be further formatted to be used for most other sequence analysis programmes.

Initially researchers used straightforward approaches to compare genomes directly in terms of sequence. These methods searched for: (i) homologues, motifs (eg regulatory or DNA binding), and common oligonucleotide and oligopeptide words (ii) orthologs (see for instance the COGS database);(iii) gene duplications; and (iv) the occurrence of conserved families in several different genomes. Several semi- and fully-automated methods have also been developed for comparing whole genome sequences against multiple databases.

CONCLUSION

This paper down loaded sequences of *Saccharomyces cerevisiae* cystathionine gamma-lyase (CYS3)gene by sending queries through the Entrez search engine. The retrieved sequences were then represented in five known formats for easy access, analysis, annotations etc. The knowledge of sequences will be of incalculable value to medicine and human biology. It has already resulted in the identification of the genes associated with many hereditary disorders and reveal the existence of a genetic basis or component for many other diseases not previously known to have one.

REFERENCES

Altschul, M.A., Brown, N.P., Leroy, C., Hoersch S., de Daruvar, A., Reich, C. (1990). Automated genome sequence analysis and annotation. *Bioinformatics,* **15:** 391-412.

Bairoch, A. and Apweiler, R. (2000). The SWISS-PROT protein sequence data blank and its supplements TrEMBL. *Nucleic Acids Research*, **28:** 45-48

Baker, W. C., Garavelli, J. S., Huang, H., McGarvey, P. B., Orcutt, B. C., Srinivasarao, G. Y., Xiao, and C., Yeh. (2000a) *Biochimie,* **71(11-12):** 1125-1143

Baker, W., van den Broek, A., Camon, E., Hingamp, P., Sterk, P., Stoesser, G. and Tuli, M. A. (2000b). The EMBL Nucleotide Sequence Database. *Nucleic Acids Research,* **28:** 19-23.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Rapp, B.A. and Wheeler, D.L. (2000). GenBank. *Nucleic Acids Research,* **25:** 1-6.

Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., & Sonnhammer, E. (1992). What's in a genome? *Nature*, **358:** 287-287.

Luscombe, N.M., Greenbaum, D. and Gerstein, M. (2001). What is bioinformatics? A proposed definition and Overview of field. *Method informatics in Medicine,* **40:**346-358

Masiga, D.K. and Isokpehi, R.D. (2004). Opportunities in Africa for training in genome Science *African journal of biotechnology,* **3(2):** 117-122.

Mount, D.W. (2004). *Sequence and Genome analysis*, 2nd edition. New York, U.S.A, Pp. 56-60.

Pearson, W.R and Lipman, D.J. (1988). Improved tools for boil. Sequence analysis. *Protocol Natural Academic Science,***85:** 2444 – 2448.

Tateno, Y., Miyazaki, S., Ota, M., Sugawara, H., and Gojobori, T. (1997). DNA databank of Japan (DDBJ) in collaboration with mass sequencing teams. *Nucleic Acids Research,* **28:** 24-26.

Tonukari, N.J. (2004). Application of computer in Biochemistry Education: The African challenge. *African Journal of Biomedical Research,* **9(2):** 11 – 14.

Ureta – vidal, A., Lsttwiller, L and Birney, E. (2003). Cooperative Genomics: genome – wide analysis in metazoan eukaryotes. *Nature Reviews Genetics,* **4:** 251 – 262.