International Research Journal of Biochemistry and Bioinformatics (ISSN-2250-9941) Vol. 1(9) pp. 215-221, October, 2011
Available online http://www.interesjournals.org/IRJBB

Review

Bioinformatics and applications in developing countries. Case study: Nigeria

***Coker D O, Odofin T W, Aladele S E, Atoyebi O J**

National Centre for Genetic Resources and Biotechnology, PMB 5382 Moor Plantation, Apata, Ibadan.

Accepted 05 October, 2011.

Biology is defined as the study of living things. Biologists collect and interpret data, in the course of the study. Now that we are equipped with sophisticated laboratory technology that allows us to collect data faster than we can interpret it. We have vast volumes of DNA sequence data at our fingertips. Bioinformatics determines which parts of DNA control the various chemical processes of life? How do we get and integrate (incorporate) a desired gene into a recipient? How do we get DNA sequences? The function and structure of some proteins are known, but how do we determine the function of new proteins? And how do we predict what a protein will look like, based on knowledge of its sequence? We understand the relatively simple code that translates DNA into protein. But how do we find meaningful new words in the code and add them to the DNA-protein dictionary? *Bioinformatics (computational biology)* is the tool that answers the questions above and other biologically related questions. This new technology makes it possible to give overwhelming interpretation to data and assign meaning where none really exists. Once we understanding and become an intelligent user of bioinformatics methods, the speed at which research progresses in Nigeria and other developing countries can be truly amazing resolving our problems on Food security-production, drug design etc. These are the more reason why the field of study (bioinformatics) must be fully embraced through capacity building of researchers. This is a window of opportunity for Nigeria to make headway- a major breakthrough in Africa in Scientific endeavors by her commitment. However, Bioinformatics and its potential seems cloudy to many researchers. Here we present a simple, easy to understand explanation of bioinformatics and its potentials in molecular biology.

Keywords: Computational biology, bioinformatic, protein prediction, DNA, sequence; interpretation, Sequence

INTRODUCTION

The Human genome project which started in 1990 with the release of the draft in 2001 and its completion in 2003 has generated biological data at a masssive rate. Other related projects that study gene expression, determine the protein structure encoded by genes and explain interactions between these products leading to explosion of sequence data.

Bioinformatics is coined from the combination of Biology, Computer Science and Information Technology. It involves storage, retrieval and analysis of data of biological origin including nucleic acid and protein sequence, function, structure and pathways so as to have a good biological view. And to make this happen, applied Mathematics, Statistics and Computer Science are all essential. It must be acknowledged that bioinformatics has recorded major successes in the field of medicine. The National Centre for Genetic Resources and Biotechnology, Ibadan; with her Conversation mandate and varietal release program; will be able to properly identify germplasm using bioinformatics tools. The newly commissioned molecular biology laboratory will be an important boost towards fulfilling this national mandate and research objective.

Definition

(Molecular) bio – informatics: bioinformatics is conceptualising biology in terms of molecules (in the sense of physical chemistry) and applying "*informatics*

*Corresponding author email:ceedos@yahoo.com

techniques" (derived from disciplines such as applied maths,computer science and statistics) to *understand* and *organise* the *information* associated with these molecules, on a ***large scale***. In short, bioinformatics is a management information system for molecular biology and has many *practical applications*.

Bioinformatics - a definition

As submitted to the Oxford English Dictionary

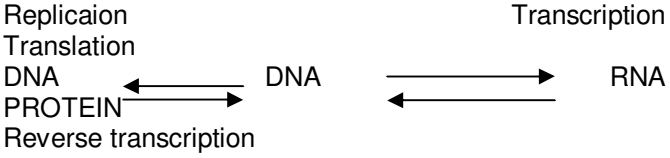
Bioinformatics operates on this threefold platform

The organization of the biological data in a way that allows researchers to access, existing data, modify and even add new entries as they are generated.
The development of bioinformatics tool(software tools) in the analysis of the acquired data.
The utilization of the software tools to analyze the data and interpretation of the results.

DNA-A digital information device

An organism's hereditary and functional information is stored as DNA, RNA, and proteins, all of which are linear chains composed of smaller molecules. A nucleotide is made up of a pentose sugar, nitrogenous bases (adenine, thymine, cytosine, and guanine) and a polynucleotide is a manifold repeat of nucleotides. DNA consist of two polynucleotides chains (Taylor et al., 1998). RNA is made up from the four nitrogenous bases (adenine, uracil, cytosine, and guanine), and proteins are made from the 20 amino acids. Because these macromolecules are linear chains of defined components, they can be represented as sequences of symbols. These sequences can then be compared to find similarities that suggest the molecules are related by form or function. It must be stated tha the physiology and existence of every organism is considerably determined by the gene which is known as the digital information.

Collection of Data



-DNA fragments are obtained by a literal fragmentation of plasmid or amplified by PCR.
-Then denature the fragments into single strands and hybridized to an oligonucleotide primer
-And submitted for sequencing to synthesize new strand from the end of the primer using heat resistant DNA polymerase enzym .
The DNA polymerase synthesizes DNA complimentary to the DNA fragments.
-The dideoxyadenosinetriphosphate/ deoxyribose Adenine Nucleotide Triphosphate introduced causes the termination of the chain synthesis.
-When the resulting mixture is subjected to electrophoresis, the fragments get separated on the basis of size.
-Then a laser beam is used to excite the fluorescent labels that can be recorded using a detector. This data is then fed into the Computer and a program is used to determine the probable order of the band to predict the sequence (Sonika and Dubey, (2006).
Nucleotide and protein sequences obtained from experiment procedures are submitted to the databank. At this stage, the sequence records are reviewed, updated and accession number is given for publishing and reference for the sequence.

Below are some of the data analysed in bioinformatics

Raw DNA sequence: -Separating coding and non-coding regions
-Identification of introns and exons
-Gene product prediction
-Forensic analysis

Protein sequence: **-**Sequence comparison algorithms
-Multiple sequence alignments algorithms
-Identification of conserved sequence motifs

Macromolecular : **-**3D structural alignment algorithms
structure -Protein geometry measurements
-Secondary, tertiary structure prediction
-Surface and volume shape calculations
-Intermolecular interactions
-Molecular simulations (force-field calculations, molecular movements , docking predictions) structure

Genomes: **-** Characterisation of repeats
-Structural assignments to genes
-Phylogenetic analysis
-Genomic-scale censuses
-Linkage analysis relating specific genes to diseases

Table 1. The Table below shows the range of the sources of the information studied (raw DNA sequences, protein sequences, macromolecular structures, genome sequences) the databases into which they are being conducted using transcription regulatory system.

Database	Websites
Protein sequence (primary)	
SWISS-PROT (Both search protein sequence,function)	www.expasy.ch/sprot/sprot-top.html
PIR-International (and domain structure)	www.mips.biochem.mpg.de/proj/protseqdb
Protein sequence (composite)	www.bioinf.man.ac.uk/dbbrowser/OWL
OWL (Both filter sequence data to produce a)	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein
NRDB (complete non redundant sets)	
	www.expasy.ch/prosite
Protein sequence (secondary)	www.bioinf.man.ac.uk/dbbrowser/PRINTS/PRINTS.html
PROSITE	www.sanger.ac.uk/Pfam/
PRINTS	
Pfam	
Macromolecular structures	www.rcsb.org/pdb
Protein Data Bank (PDB)	ndbserver.rutgers.edu/
Nucleic Acids Database (NDB)	www.ncifcrf.gov/CRYS/HIVdb/NEW_DATABASE
HIV Protease Database	www2.ebi.ac.uk:8081/home.html
ReLiBase	www.biochem.ucl.ac.uk/bsm/pdbsum
PDBsum	www.biochem.ucl.ac.uk/bsm/cath
CATH (Identify proteins by their structures so as to identify their structural and evolutionary relationship)	scop.mrc-lmb.cam.ac.uk/scop
SCOP	www2.embl-ebi.ac.uk/dali/fssp
FSSP	
Nucleotide sequences	www.ncbi.nlm.nih.gov/Genbank
GenBank	www.ebi.ac.uk/embl
EMBL	www.ddbj.nig.ac.jp
DDBJ	
	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome
Genome sequences	bioinfo.mbb.yale.edu/genome
Entrez genomes	www.ncbi.nlm.nih.gov/COG
GeneCensus	
COGs	
Integrated databases	www.ebi.ac.uk/interpro
InterPro	www.expasy.ch/srs5
Sequence retrieval system (SRS)	www.ncbi.nlm.nih.gov/Entrez
Entrez	

Gene expression: -Correlating expression patterns

-Mapping expression data to sequence, structural and biochemical data

Bioinformatic Tools

Tools for this newly evolving technologies for genomics and proteomics are available for sequences and structure comparison, protein finding and prediction, structure prediction,gene expression analysis, functional characterization of proteins, drug design and pathway modeling.

A complete list of bioinformatic search and analysis tools are now available and compiled here:*http://bioinformatics.ubc.ca/resources/links-directory/narweb2006/categorized.php*. (Sonika and

Dubey, 2006)

Organisation of Data

- For raw DNA sequences, investigations to be carried out involves the identification and separation of introns and exons and identification of promoters
- For protein sequences, analysis includes developing algorithm for sequence comparison (Miller et al., 1999) and methods for producing multiple sequence alignments and searching for functional domains from conserved motifs in such alignments (Gonnet et al., 2000).
- For structural data, investigation includes prediction of secondary and tertiary protein strucrures, generating methods for 3D structural alignments (Orengo and Taylor, 1996;Orengo, 1999) examining protein geometry and calculation of surface and volume, shapes and analysis of protein interaction with other subunit, DNA,RNA and smaller molecules. These studies have led to molecular simulation topics in which structural data are being used to determine the energetics involved in stabilizing macromolecules and computing the energies involved in molecular docking (Luscombe et al., 2001).

In addition to finding the relationship between different proteins, much of bioinformatics involves the analysis of one type of data to infer and understand the observations for another type of data example 1-The use of sequence and structural data to predict the secondary and tertiary structure of new protein sequences.This method is based on statistical rules derived structures such as the propensity for certain amino acids to produce different secondary structural elements. example 2-The use of structural data to understand protein folds and their functions and analysed similarities between different binding sites in the absence of homology.

Sequence Alignment

Sequence comparison is possibly the most useful computational tool to molecular biologists .And ,the internet has made it possible for a single public database of genome sequence data to provide services through a uniform interface to a worldwide community of users. With a commonly used computer program called fsBLAST, molecular biologist can compare an uncharacterized DNA sequence to the entire publicly held collection of DNA sequences. Below is an example of how sequence comparison using the BLAST program can help to gain insight into a real disease. Sequence alignment also help to determine similarities in two or more macromolecules, isolate the desired sequence and integrates into a recepient for expression.

The Eye of the Fly

Fruit flies (*Drosophila melanogaster*) are a popular model system for the study of development of animals from embryo to adult. Fruit flies have a gene called *eyeless*, which, if it's "knocked out results in fruit flies with no eyes. It's obvious that the *eyeless* gene plays a role in eye development. Researchers have identified a human gene responsible for a condition called *aniridia*. In humans who are missing this gene (or in whom the gene has mutated just enough for its protein product to stop functioning properly), the eyes develop without irises. If the gene for *aniridia* is inserted into an eyeless drosophila "knock out," it causes the production of normal drosophila eyes. It's an interesting coincidence. Could there be some similarity in how *eyeless* and *aniridia* function, even though flies and humans are vastly different organisms? Possibly. To gain insight into how *eyeless* and *aniridia* work together, we can compare their sequences (Cynthia and Per, 2001).

Few years ago, the search for similarities between DNA sequences was a difficult task. So, what most ost scientists did was to compare the respective gene sequences by hand-aligning them one under the other in a word processor and looking for matches character by character. That was pretty tedious and at the same time consuming.

Pair wise comparison of biological sequences is the foundation of most widely used bioinformatics techniques. Many tools that are widely available to the biology community--including everything from multiple alignment, phylogenetic analysis, motif identification, and homology-modeling software, to web-based database search services--rely on pairwise sequence-comparison algorithms as a core element of their function.

Once a nucleotide or protein has been sequenced, the most common step is to compare it with the known sequences present in the database. Alignment of protein or nucleotide sequences are very important when it comes to analyzing sequences. They provide information about conserved sequence region or its domain. Alignment is also very useful in predicting the structure of proteins, identifying new members of protein families and inferring the evolutionary history of the sequences caused by mutation during evolution creating differences between various families of species. Most of these differences are due to local mutation.

A high-quality sequence match between two full-length sequences may suggest the hypothesis that their functions are similar, although it's important to remember that the identification is only tentative until it's been experimentally verified.

Comparing eyeless and aniridia with BLAST

When the two sequences are compared using BLAST, you'll find that *eyeless* is a partial match for *aniridia*. The text that follows is the raw data that's returned from this BLAST search:

pir||A41644 homeotic protein aniridia - human
Length = 447
Score = 256 bits (647), Expect = 5e-67
Identities = 128/146 (87%), Positives = 134/146 (91%), Gaps = 1/146 (0%)
Query:24
IERLPSLEDMAHKGHSGVNQLGGVFVGGRPLPDSTR QKIVELAHSGARPCDISRILQVSN 83
I R P+ M + HSGVNQLGGVFV GRPLPDSTRQKIVELAHSGARPCDISRILQVSN
Sbjct:17
IPRPPARASMQNS-HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPC DISRILQVSN 75
Query:84
GCVSKILGRYYETGSIRPRAIGGSKPRVATAEVVSKIS QYKRECPSIFAWEIRDRLLQEN 143
GCVSKILGRYYETGSIRPRAIGGSKPRVAT EVVSKI+QYKRECPSIFAWEIRDRLL E
Sbjct:76
GCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIA QYKRECPSIFAWEIRDRLLSEG 135
Query: 144
VCTNDNIPSVSSINRVLRNLAAQKEQ 169
VCTNDNIPSVSSINRVLRNLA++K+Q
Sbjct: 136 VCTNDNIPSVSSINRVLRNLASEKQQ 161
Score = 142 bits (354), Expect = 1e-32
Identities = 68/80 (85%), Positives = 74/80 (92%)
Query:398
TEDDQARLILKRKLQRNRTSFTNDQIDSLEKEFERTHY PDVFARERLAGKIGLPEARIQV 457
+++ Q RL LKRKLQRNRTSFT +QI++LEKEFERTHYPDVFARERLA KI LPEARIQV
bjct:222
SDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTH YPDVFARERLAAKIDLPEARIQV 281
Query: 458
WFSNRRAKWRREEKLRNQRR 477
WFSNRRAKWRREEKLRNQRR
Sbjct: 282 WFSNRRAKWRREEKLRNQRR 301

The above is the output of the BLAST search. In each set of three lines, the query sequence (the *eyeless* sequence that was submitted to the BLAST server) is on the top line, and the *aniridia* sequence is on the bottom line. The middle line shows where the two sequences match. If there is a letter on the middle line, the sequences match exactly at that position. If there is a plus sign on the middle line, the two sequences are different at that position, but there is some chemical similarity between the amino acids (example, D and E, aspartic and glutamic acid). If there is nothing on the middle line, the two sequences don't match at that position. (Cynthia and Per 2001).

Homologous

The traits are similar due to common ancestry. And the fact that two sequences share a stretch of nearly identical nucleotide or amino acid does not mean homology.As a rule, 25% identity over a stretch 100 amino acids can be considered a good evidence of common ancestry of the two sequences (Russell et al., 1997).

Analogous

The traits are similar due to convergent evolution.

Orthology

The traits are homologous with conserved functions.

Paralogous

The traits are homologous due to divergent functions.

Xenologous

The sequence identity due to horizontal transfer.

Pairwise Sequence Alignment

is the process of comparing two sequences globally or locally.

Multiple Sequences Alignment

is the process of comparing many sequences to identify and visualize conserved regions in a family of homologous protein. Phylogenetic methods depend on Multiple Sequence Alignment to infer evolutionary history of the sequence.Multiple Sequence Alignment is one of the frequent activities of a bioinformatician.

Phylogenetic Analysis

This technique involves establishing an evolutionary history between organisms using different DNA patterns. This is based on the assertion of a common ancestral DNA and evolution of genomes by slow accumulation of mutations. That implies that genomes with fewer differences will have a recently shared common ancestor. Phylogenetic analysis can be used to map the genes in two organisms that may have

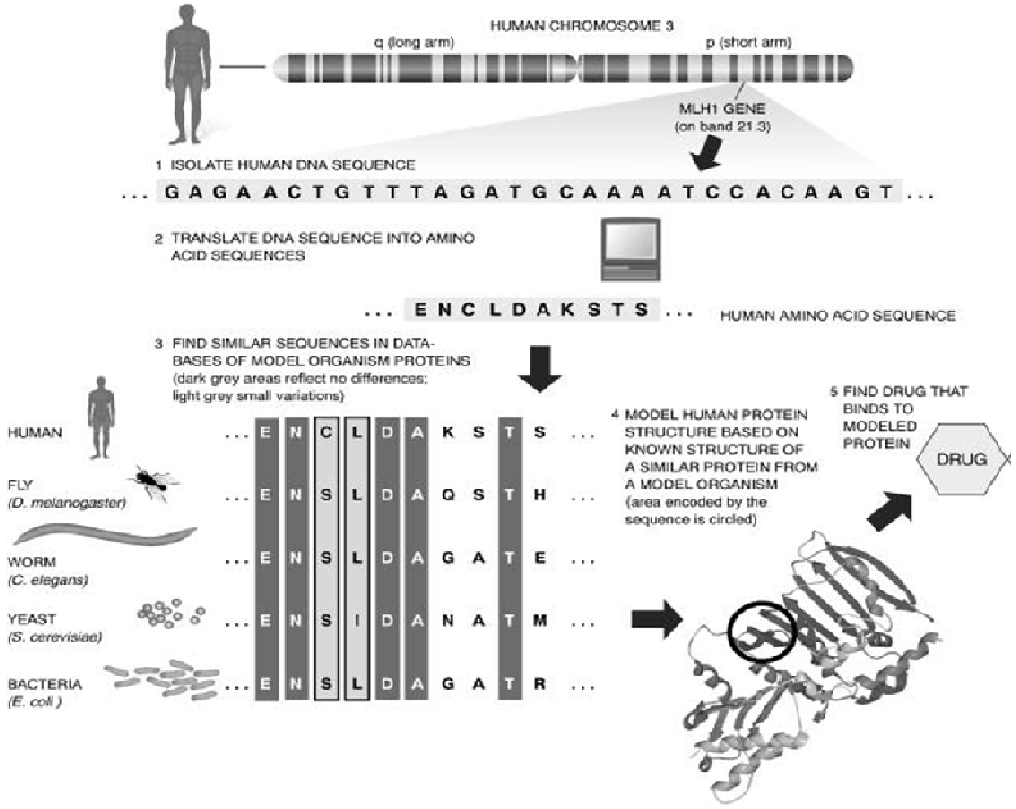


Figure 1

similar functions and to map the changes in a rapidly changing genomes. PAUP and PHYLIP are some of the commonly applied phylogenetic analysis programs. The The predictions can be made using either of these methods for DNA or protein sequences.

There are common terms to describe the relationship between pairs of proteins or the genes from which they are derived: analogous proteins have related folds, but unrelated sequences, while homologous proteins are both sequentially and structurally similar.

The two categories can sometimes be difficult to distinguish especially if the relationship between the two proteins is remote (Russell et al., 1997;1998). Among homologues, it is useful to distinguish between orthologues, proteins in different species that have evolved from a common ancestral gene, and paralogues, proteins that are related by gene duplication within a genome (Fitch, 1970). Normally, orthologues retain the same function while paralogues evolve distinct, but related functions (Tatusov et al., 1997).

Drug Discovery and Design

Most of the tissues are made up of proteins having well arranged amino acids as their primary building block. As we have known, each amino acid is coded for by a codon-a process well under the control of the genes. The genes code for all the amino acid-protein in all living organisms.

However, there can be mutation, an alteration in the nucleotide sequence of a DNA thereby coding for an entirely different amino acid-a possible disease factor.

Bioinformatics is being used in aiding rational drug design.

Taking gene product of MLH1 as an example drug target.

MLH1 is a human gene encoding a mismatch repair protein(mmr) situated on the short arm of chromosome 3(Kok et al.,1997).This gene has been implicated in the nonpolyposis colorectal cancer (Syngal et al 2000).Given the nucleotide sequence,the likely amino acid sequence of the encoded protein can be determined using translational software. And,based on sequence similarity, it is possible to model the structure that is found in human.

Finally,docking algorithms could be used to design molecules that could bind the model structure,paving the way to biochemical assays to test their biological potency on the actual protein.
Figure1.

How to Configure A Pc for Bioinformatics Research

Most computer users are familiar with the following versions of Microsoft operating systems: Windows 95;

Windows 98; Windows; 2000 Professional Windows; XP Windows; Vista and the recently released Windows 7.Most computer manufacturing companies like HP,Toshiba,Dell, have the above mentioned operating systems as their default setting.

UNIX is another operating system that comes in a number of flavors,the three most popular being BSD , SunOs and Linux. Most scientific software is developed on Unix machines, and researchers prefer programs that can be run only on Unix. As it has grown popular in the mass market, Linux has retained the power of Unix systems for developing, compiling and running programs, networking and managing jobs for multiple users. And, also providing the standard trimmings of a desktop PC, including word processors, graphics programs, and even visual programming tools. For many of the specific bioinformatics tools, Unix is the most practical choice.

Macintosh computer operates on its own operating system (Mac O X)- a version of Unix, as it default operating system. It is recommended that the research Centre run their program on Macintosh computer system.

However, any system with Unix operating system installed can be used for bioinformatics work.

Advantages of ***Unix*** over ***Windows***

- Unix is more flexible and can be installed on many different types of machines, including main-frame computers, supercomputers and micro-computers.
- Unix is more stable and does not go down as often as Windows does, therefore requires less administration and maintenance.
- Unix has greater built-in security and permissions features than Windows.
- Unix possesses much greater processing power than Windows.
- Unix is the leader in serving the Web. About 90% of the Internet relies on Unix operating systems running Apache, the world's most widely used Web server.
- Software upgrades from Microsoft often require the user to purchase new or more hardware or prerequisite software. That is not the case with Unix.

CONCLUSION

Bioinformatics now encompasses a widerange of subject areas including structural biology, genomics and gene expression studies, proteomics and drug design by providing a more global perspective in design of experiments. It allows us to transfer information to new systems from other, well- characterized organisms and experiments using an integrative approach and analysis of biological sequences. The application of bioinformatics tools to genetic resources conservation program in Nigeria is to assist in proper identification and classification of our germplasm collections in the genebank.It will also assist the centre the gene search blast aspect of a molecular biology research as well as cataloging of her germplasm database.

REFERENCES

Taylor DJ, Green NPO, Stout GW (1998). Biological Science.:3(108).

Sonika B, Dubey AK (2006). Recombinant DNA Technology And Biotechnology Netaji Subhas Institute of Technology Dwarka, New Delhi.

Sonika B, Dubey AK, (2006). Recombinant DNA Technology And Biotechnology Netaji Subhas Institute of Technology Dwarka, New Delhi.

Miller C, Gurd J, Brass A (1999). A rapid algorithm for sequence database comparisons: application to the identification of vector contamination in the EMBL databases. Bioinformatics. 15(2): 111-21.

Gonnet GH, Korostensky C, Benner S (2000). Evaluation measures of multiple sequence alignments [In Process Citation]. J. Comput. Biol. 7(1-2):261-76

Orengo CA, Taylor WR (1996). SSAP: sequential structure alignment program for protein structure comparison. Methods Enzymol. (266): 617-35.

Orengo CA (1999). CORA—topological fingerprints for protein structural families.

Protein. Sci. 8 (4): 699-715

Luscombe NM, Greenbaum D, Gerstein M (2001) What is bioinformatics? An

introduction and overview.Rev. Med. Informatic ;88

Cynthia G, Per J (2001). Developing Bioinformatics Computer Skills. .1 (1): 83-99

10. Cynthia G, Per J (2001). Developing Bioinformatics Computer Skills. 1(1): 83-99.

Sonika Bhatnagar and A.K.Dubey.Recombinant DNA Technology And Biotechnology Netaji Subhas Institute of Technology Dwarka, New Delhi.2006

Russell RB, Saqi MA, Sayle RA, Bates PA, Sternberg MJ (1997). Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J. Mol. Biol. 269(3):423-39.

Russell RB, Saqi MA, Bates PA, Sayle RA, Sternberg MJ (1998). Recognition of analogous and homologous protein folds—assessment of prediction success and associated alignment accuracy using empirical substitution matrices. Protein Eng. 11 (1): 1-9.

Fitch WM (1970). Distinguishing homologous from analogous proteins. Syst Zool. 19:99-110.

15. Tatusov RL, Koonin EV, Lipman DJ (1997). A genomic perspective on protein families.Science. 278 (5338): 631-7.

Kok K, Naylor SL, Buys CH (1997). Deletions of the short arm of chromosome 3 in solid tumors and the search for suppressor genes. Adv. Cancer Res. 71:27-92.

Syngal S, Fox EA, Eng C, Kolodner RD, Garber JE (2000). Sensitivity and specificity of clinical criteria for hereditary nonpolyposis colorectal cancer associated mutations in MSH2 and MLH1. J. Med. Gen. 37(9):641-645.