An Analysis of Breast Cancer Gene Sequences using Differential Evaluation

K. Lohitha Lakshmi, P. Bhargavi, S. Jyothi

Department of Computer Science, Sri PadmavatiMahilaVisvavidyalayam, Tirupati, India.

***Abstract* - In the present age of innovation the medicinal field has emerged as a most vital areas of research with cancer being a significant topic, where real treatment has not been found yet. Cancer diseases must be diagnosed at an early stage to increase survival rate. Breast cancer is a leading cause of death mostly among women worldwide. Soft Computing and artificial intelligence provide methodologies for the early detection of breast cancer tumors due to their capabilities to handle complex, large and noisy proteomic and genomic data sets. Differential Evolution (DE) optimization algorithms are proposed here to determine the optimal treatment set based on available data. Worst treatment results are removed from each optimization stage and randomly generated new treatment routes are added in each step to find an approximate optimal solution to the given data set. In the present paper this methodology is implemented on breast cancer and normal breast genomic data sets to generate best average and best value for each generation in the optimization algorithm. These values can be used in further diagnosis and analysis.**

Keywords - *Differential Evolution, Soft Computing Techniques, Breast Cancer, Diagnosis, Analysis, Prediction.*

I. INTRODUCTION

Soft computing is computing which is not hard. The term Soft Computing (SC) was coined byLotfi A Zadeh, a pioneer in this field. Soft computing has various perspectives due to its methodological traits, problem solving abilities and ability of satisfying some strong constraints. SC provides techniques with capacity to solve some class of problems for which other conventional techniques are found to be inadequate. The principle components of soft computing include fuzzy systems, roughset theory, neural networks, probabilistic reasoning and evolutionary search strategies including genetic algorithms, simulating annealing, ant colony optimization, particle swarm optimization, differential evolution etc. Data in real life problems is full of uncomfortable features with vague perceptions.SC techniques are developed to deal data with certain characteristics such as flexible, adjustable, random, vague, inexact/approximate, imprecise, perceivable, porous, and non-deterministic. SC is a family of highly interactive and complementary techniques. A complex optimization problem requires more advanced techniques such as evolutionary algorithms to obtain solutions within reasonable time frame [1].

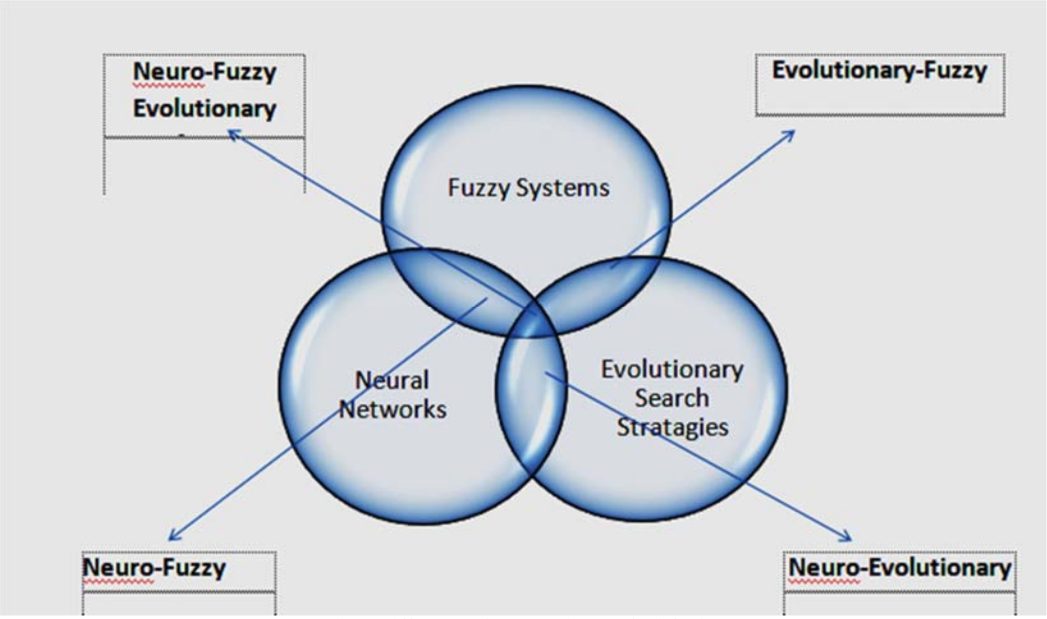

Fig 1: Different Soft Computing Methodologies

II. THE NEED OF BREAST CANCER GENE SEQUENCE ANALYSIS USING SOFT COMPUTING TECHNIQUES

Breast cancer is the second deadliest infection in women Overall, the hazard expanding with the age [2]. Cancer is an anomalous cell to growth happening in any part of the human body which leads to affect any organs. This anomalous cell division leads to form a tumour. Such tumours undergo meiotic division and spread over the body through blood circulatory system [3]. Early diagnosis for breastcancer is viewed as a vital issue as it is one of the fundamental elements for its treatment. So there is a need to create present day and successful techniques to manage this infection and the speed of its prevention. Soft Computing provides various techniques for medical data sequence analysis by analysing gene (DNA and RNA) sequences to diagnose some endemic diseases [4].

III. PROPOSED METHODOLOGY

A. Evolutionary Computing Techniques of SC for Genetic Data Sequence Analysis

Evolutionary computation is a sub field of soft computing comprises of meta-heuristic optimization algorithms which are termed as evolutionary algorithms or techniques. These mechanisms are inspired by biological evolution, finding candidate solutions to the optimization problem and on the cost function or fitness function based on the environment [5].

An evolutionary computation technique involves family of population-basedalgorithms for global optimization. These algorithms produce initial set of solutions to the problem and updated in each cycle of iteration stochastically reducing undesired solutions and producing highly optimizedsolutions with small random changes. This wide range capability of problem setting makes these techniques popular in computer science. Evolutionary computing techniques are classified in to exact methods and meta-heuristic methods. Exact methods granted solution to the problem directly but not suitable for solving real life problems due to the complex nature included in real time applications. Another type of methods are meta-heuristic methods, which are suitable to successfully applied on large and complex problems over the years has provided fruitful results. In meta-heuristic methods some techniques or algorithms are developed under population based. Differentialevolution algorithm categorized under population based meta-heuristic optimization algorithm [6].

B. Role of Differential Evolution in Breast Cancer Gene Data Sequence Analysis

Differential Evolution (DE) is a strategy that streamlines a problem by iteratively attempting to enhance an applicant solution as to a given proportion of condition and value.DE is a population-based optimization method that works on real number coded individuals.DE is quite robust, fast and effective with global optimization ability.DE does not require differential objective function and it works well with noisy data.The essential thought behind implementation of DE is to produce trail parameter vectors.DE generates new vector in each generation by adding weighted difference vector to the present generation vector. If the newly generated vector produces lower objective value than previous population the newly generated value will be considered to perform further computation for future generations. The best parameter vector is generated for eachgeneration duringminimization process to find optimal parameter value based on some specific stopping condition. Different variants of DE are proposed for various conditions of genetic experiments to yield optimal value based on conditional requirement [7].

IV. IMPLEMENTATION OF DIFFERENTIAL EVOLUTION ALGORITHM FOR BREAST CANCER GENE DATA SEQUENCES

In the proposed methodology breast cancer data sequence is analysed using differential evaluation. Python is a more avant-garde and better organized language. Python is a language with great capabilities for bio informatics applications by consisting bioinformatics libraries, indicated as bio-python. The packages which are included in these libraries for bioinformatics are moderately useful. Python has turned into a well-known programming language in the biosciences, to a great extent in light of the fact that (i) due to its clear syntax and clean semantics (ii) it is expressive and well suited to data in real world applications like genetic data sequences (iii) the numerous accessible libraries and outsider toolboxes broaden the usefulness of the language, because it now pervades virtually each biological domainfor data sequence and structure analysis, phylogenomics for molecular evolution and to perform statistical computations on dynamic data produced by real world applications and beyond[8].due to its great capabilities python programming language has been selected for implementation of differential evolution to find near optimal solution for the breast cancer and normal breast data sequence analysis.

Differential Evolution will be proposed to reduce the pressure to handle data which includes properties like random, un-deterministic, self-adaptive and dynamic [9]. Dealing with genetic related dynamic data and to choose a best approach for biological data is not always an easy thing [10]. Due to the large DNA sequences available in the referred NCBI and other sites toreduce computational complexity RNA shotgun sequences are taken into consideration due to its small size.

A. Implementation of DE in Python

1. Initialize the population. Read mRNA shotgun sequences of both breast cancer and breast cancer suppressor gene sequences from NCBI site.
2. Convert character formatted sequence to numeric format by replacing N, A, C, G, T values with 0,1,2,3,4.
3. Start mutation process by selecting three random vectors x1, x2, x3 with indexed positions excluding current vector.
4. Find the difference between x3 and x2 and create a new vector x $_{\text{diff}}$ with resultant difference values.
5. Multiply difference vector with mutation factor and add to x1 vector then it generates current generation donor vector.
6. Recombination step- In this step crossover is performed on donor vector which generates third vector.
7. Greedy selection step-Apply the cost function on trial vector and target vector.
8. If trail vector score is less than target vector score then considers current trail vector as population vector for next generation. If trail vector score is greater than target vector appends target score to generation score.
9. Find the generation average and generation best values and best solution vector for current generation for each iteration.
10. Repeat steps 3-9 to find the optimal solution for each generation.
11. Display the result.

B. Flow Chart for DE in Python

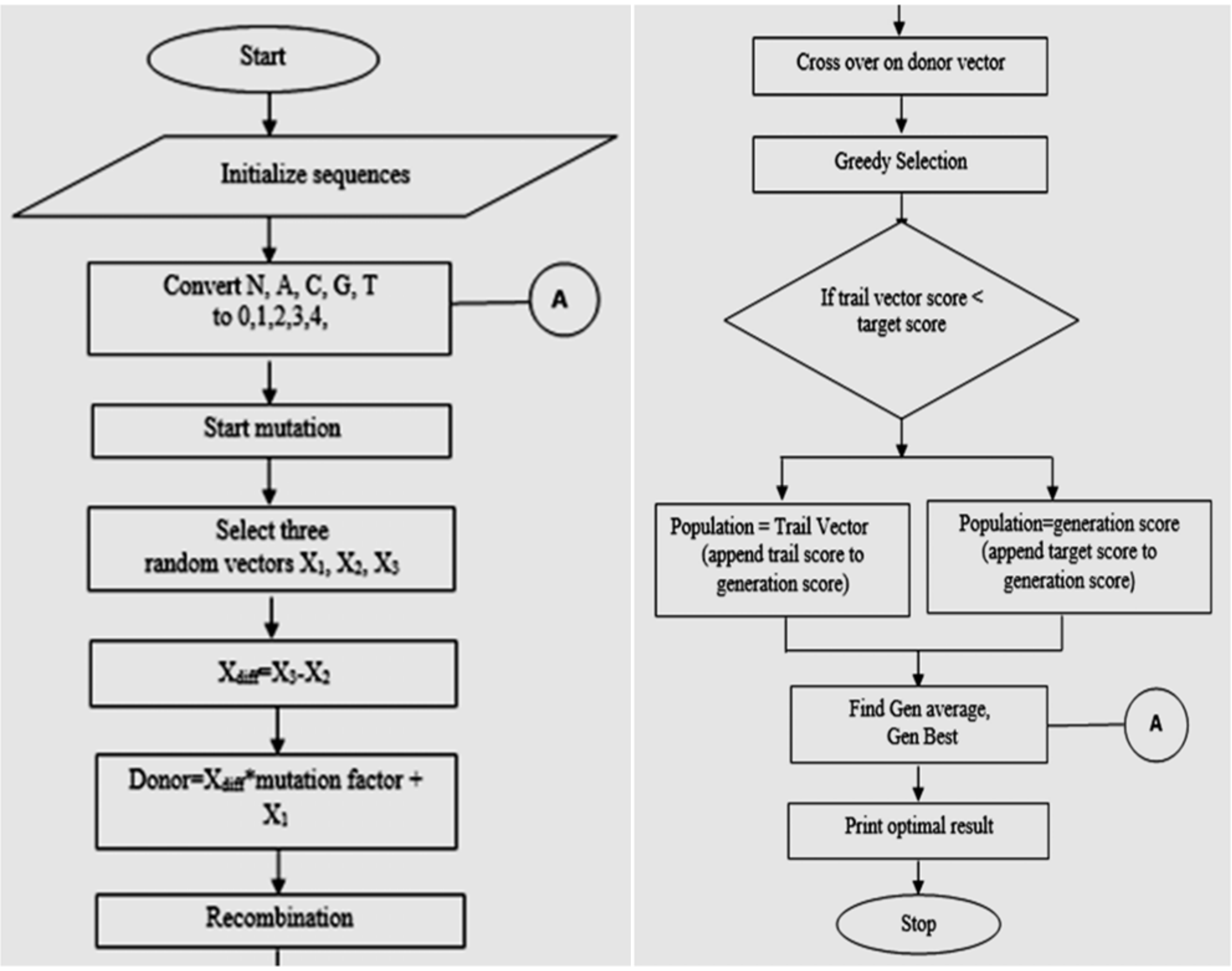
Start
Initialize sequences
Convert N, A, C, G, T to 0,1,2,3,4,
A
Start mutation
Select three random vectors X1, X2, X3
Xdiff=X3-X2
Donor=Xdiff*mutation factor + X1
Recombination
Cross over on donor vector
Greedy Selection
If trail vector score < target score
Population = Trail Vector (append trail score to generation score)
Population=generation score (append target score to generation score)
Find Gen average, Gen Best
A
Print optimal result
Stop

Fig.2 Flow chart for DE implementation in python.

The input sequences are collected from the popular and trusted web based medical resources i.e. NCBI. The details of input sequences with Accession numbers and description are mentioned below.

C. Breast Cancer Gene Input Sequences

TABLE I. LIST OF BREAST CANCER DATA INPUT SEQUENCES

NO	ACCESSION NO & DESCRIPTION
SEQ 1	AF284812- Homo sapiens BRCAI (BRCA1) gene, exon 20 and partial cds.
SEQ 2	AF507075- Homo sapiens breast cancer susceptibility protein (BRCA1) gene, exon 20 and partial sequence.
SEQ 3	AF507076- Homo sapiens IRCHS11B breast and ovarian cancer susceptibility protein (BRCA1) gene, exon 20 and partial cds.
SEQ 4	AF507077-Homo sapiens IRCHS6A breast and ovarian cancer susceptibility protein (BRCA1) gene, exon 20 and partial cds.
SEQ 5	AF507078-Homo sapiens IRCHS6B breast and ovarian cancer susceptibility protein (BRCA1) gene, exon 20 and partial cds.
SEQ 6	AY093484-Homo sapiens isolate IRCHS8A breast and ovarian cancer susceptibility protein (BRCA1) gene, partial cds.
SEQ 7	AY093485-Homo sapiens isolate IRCHS8B breast and ovarian cancer susceptibility protein (BRCA1) gene, partial cds.
SEQ 8	AY093486- Homo sapiens isolate IRCHS7A breast and ovarian cancer susceptibility protein (BRCA1) gene, partial cds.
SEQ 9	AY093487- Homo sapiens isolate IRCHS7B breast and ovarian cancer susceptibility protein (BRCA1) gene, partial cds.
SEQ 10	AY093488-Homo sapiens isolate IRCHS16A breast and ovarian cancer susceptibility like protein (BRCA1) gene, partial sequence.
SEQ 11	AY093489-Homo sapiens isolate IRCHS16B breast and ovarian cancer susceptibility protein (BRCA1) gene, partial cds.
SEQ 12	AY093490-Homo sapiens isolate IRCHF4A breast and ovarian cancer susceptibility protein (BRCA1) gene, partial cds.
SEQ 13	AY093491-Homo sapiens isolate IRCHF4B breast and ovarian cancer susceptibility protein (BRCA1) gene, partial cds.
SEQ 14	AY093492-Homo sapiens isolate IRCHS4A breast and ovarian cancer susceptibility protein (BRCA1) gene, partial cds.
SEQ 15	AY093493- Homo sapiens isolate IRCHS4B breast and ovarian cancer susceptibility protein (BRCA1) gene, partial cds.
SEQ 16	AY144588-Homo sapiens truncated breast and ovarian cancer susceptibility protein (BRCA1) gene, partial cds.
SEQ 17	AY150865- Homo sapiens truncated breast and ovarian cancer susceptibility protein (BRCA1) gene, exon 12 and partial cds.

D. BreastCancer Suppressor GeneInput Sequences

TABLE II. LIST OF BREAST CANCER SUPPRESSOR GENE INPUT SEQUENCES

NO	ACCESSION NO & DESCRIPTION
SEQ 1	AB118156.1 Homo sapiens p53 gene for P53, exon 5, partial cds.
SEQ 2	AB699004-Homo sapiens gene for P53 protein, partial cds and exon 4.
SEQ 3	AF066082.1 Homo sapiens mutant p53 transformation suppressor gene, exon 6 and partial cds
SEQ 4	AF209128.1 Homo sapiens tumour suppressor p53 (TP53) gene, partial cds
SEQ 5	AF209129.1 Homo sapiens tumour suppressor p53 (TP53) gene, partial cds
SEQ 6	AF209130.1 Homo sapiens tumour suppressor p53 (TP53) gene, partial cds.
SEQ 7	AF209131.1 Homo sapiens tumour suppressor p53 (TP53) gene, partial cds
SEQ 8	AF209133.1 Homo sapiens cell-line A431 tumour suppressor p53 (TP53) gene, partial cds
SEQ 9	AF209136.1 Homo sapiens cell-line HN5 tumour suppressor p53 (TP53) gene, partial cds
SEQ 10	AF209137.1 Homo sapiens tumour suppressor p53 (TP53) gene, partial cds.
SEQ 11	AF209138.1 Homo sapiens tumour suppressor p53 (TP53) gene, partial cds
SEQ 12	AF209139.1 Homo sapiens tumour suppressor p53 (TP53) gene, partial cds
SEQ 13	AF209140.1 Homo sapiens tumour suppressor p53 (TP53) gene, partial cds
SEQ 14	AF209141.1 Homo sapiens cell-line Molt4 tumour suppressor p53 (TP53) gene, partial cds
SEQ 15	AF209142.1 Homo sapiens cell-line A431 tumour suppressor p53 (TP53) gene, partial cds
SEQ 16	AF209143.1 Homo sapiens cell-line HT29 tumour suppressor p53 (TP53) gene, partial
SEQ 17	AY304556- Homo sapiens breast and ovarian cancer susceptibility protein (BRCA1) gene, exon 11 and partial cds.

In order to study one whole genome, scientists use an easy strategy known as shotgun sequencing. The long DNA/RNA sequence is assembled to from a series of shorter overlapping sequences. Special machines, called sequencing machines are used to extract shotgun random genome sequences from a specific genome to get a target genome to work out [11].

V. EXPERIMENTAL RESULTS

Advances in technology have made vast quantity of data available for conducting practical evaluation on sequences for testing or comparing results on genetic variants associated with the risk of breast cancer.

In the present paper approximately 17 sets of gene data sequences are taken from each category i.e. breast cancer and breast cancer suppressor gene sequences to compare the results.DE algorithm is implemented on these sequences using python. In the obtained results it is observed that, the resultant or optimal value which is obtained after implementation of DE on each gene sequence is termed as Generation Best (GB) value for each generation and for all individual iterations. Here in this case Generation Best value is considered to perform comparison on different categories of individuals.

If the GB values are observed in case of breast cancer gene data sequences the range of values approximately lies in between 9.01 to 18.03 for maximum individuals except 1 or 2 individuals. Least percentage of individuals GB value is deviating from this range i.e. 0.92 to 3.71.

A. DE Algorithm Result for Breast Cancer Input Sequence

TABLE III DE RESULT FOR BREAST CANCER INPUT SEQUENCES

Seq No	**Acc.No**	**DEResult for I gen**	**DEResult for II gen**	**DEResult for III gen**
SEQ 1	AF284812	12.11	9.01	9.16
SEQ 2	AF507075	12.07	10.90	15.97
SEQ 3	AF507076	11.59	15.81	11.19
SEQ 4	AF507077	16.86	16.82	14.02
SEQ 5	AF507078	15.86	13.28	16.04
SEQ 6	AY093484	15.66	17.16	12.41
SEQ 7	AY093485	12.57	14.72	12.13
SEQ 8	AY093486	12.95	12.69	17.71
SEQ 9	AY093487	21.81	15.05	16.25
SEQ10	AY093488	18.29	17.66	18.71
SEQ11	AY093489	14.54	15.68	12.85
SEQ12	AY093490	13.90	13.80	16.66
SEQ13	AY093491	11.49	15.81	15.23
SEQ14	AY093492	13.65	14.87	15.88
SEQ15	AY093493	12.33	18.03	14.02
SEQ16	AY144588	0.94	1.10	1.91
SEQ17	AY150865	2.24	2.05	3.71

When the DE is implemented on approximately 17 sets of breast cancer suppressor gene sequences, it is observed that the resultant valuesrange lies in between minimum 1.7 to maximum 8.6. One individual is deviating from these range of values with minimum value 11.03 and maximum value 13.94.

B. DE Algorithm Result for Breast Cancer Suppressor Gene Input Sequences

TABLE IV. DE RESULT FOR BREAST CANCER SUPPRESSOR GENE SEQUENCES

Seq No	**Acc.No**	**DEResult for I gen**	**DEResult for IIen**	**DEResult forIIIen**
SEQ 1	AB118156	6.71	6.05	6.70
SEQ 2	AF066082	1.75	4.63	4.21
SEQ 3	AF209128	5.7	8.4	6.19
SEQ 4	AF209129	5.6	9.3	4.3
SEQ 5	AF209130	5.9	3.14	6.83
SEQ 6	AF209131	8.48	6.37	3.95
SEQ 7	AF209133	7.07	8.79	5.12
SEQ8	AF209136	9.36	3.87	6.86
SEQ9	AF209137	5.17	3.84	7.71
SEQ10	AF209138	5.73	5.34	6.09
SEQ11	AF209139	5.10	3.42	3.07
SEQ12	AF209140	3.66	2.57	4.17
SEQ13	AF209141	4.13	2.37	4.04
SEQ14	AF209142	4.08	5.48	4.38
SEQ15	AF209143	3.83	3.05	3.59
SEQ16	AF209556	1.50	1.56	5.73
SEQ17	AB699004	11.13	11.84	13.94

In view of these ideal qualities created after execution of DE algorithm on breast cancer and non-breast disease groupings the proposed strategy helps to propose on sort of degree to recognize the distinction between those arrangements. This technique can be utilized on any test successions to distinguish whether the arrangement is similar to suspect breast cancer growth sequencing dependent on the GB values after implementation of DE algorithm in greater part of the cases. These outcomes can be utilized for further finding.

VI. CONCLUSION

In the proposed strategy when DE estimations of both breast cancer sequences and non-breast cancer sequences are considered. The scope of estimations of breast cancer successions are less when contrasted with non-breast cancer arrangements. In light of the DE result the proposed strategy is valuable to discover the distinction among cancer and non-cancer disease. In future the proposed strategy will be actualized with different calculations from Soft Computing strategies. The last results which are acquired with the usage of various calculations will be contrasted with legitimize which calculation is delivering more exact outcomes in such kind of hereditary arrangement investigation. The fact that these experiments may not provide exact, appropriate, and desired results but make it sure that these experiments pave the way for generating some innovative ideas and imminent research.

REFERENCES

[1] Soft Computing: Neuro-Fuzzy and Genetic Algorithms Kindle Edition by SamirRoy (Author), Udit Chakraborty (Author)

[2] "An Analysis Of The Methods Employed For Breast Cancer Diagnosis", Mahjabeen Mirza Beg, Monika Jain, April 2012, DOI: 10.7815/ijorcs.23.2012.025 License CC BY-NC-ND 4.0.

[3] Web ref: http://www.cancer8.com/

[4] Classification_of_Breast_Cancer_Using_Softcomputing_Techniques [accessed Aug 30 2018].

[5] Net ref: https://en.wikipedia.org/wiki/Evolutionary_computationn

[6] shodhganga.inflibnet.ac.in/bitstream/10603/10161/11/11_chapter%203.pd, chapter-3 soft computing techniques – Shodhganga

[7] Evavolna," Introduction to soft Computing",1st edition@2013 Eva Volna & bookboon.com, ISBN 978-87-403-0573-9.

[8] Berk Ekmekci ,Charles E. McAnany ,Cameron Mura ," An Introduction to Programming for Bioscientists: A Python-Based Primer", Published: June 7, 2016, https://doi.org/10.1371/journal.pcbi.1004867

[9] Differential Evolution fundementals and applications in Electrical Engineering by Anyong Quing

[10] Rute R.da FonsecaaAndersAlbrechtsenaGonçalo Espregueira ThemudocJazmínRamos-MadrigalbJonas AndreasSibbesenaLasseMarettyaM. LisandraZepeda-MendozabPaula F.CamposbdRasmusHelleraRicardo J.Pereirab," Next-generation biology: Sequencing and data analysis approaches for non-model organisms", Marine Genomics, Volume 30, December 2016, Pages 3-13.

[11] Vijini Mallawaarachchi," DNA Sequence Data Analysis - Starting off in Bioinformatics", PhD Student at Australian National University | Loves Bioinformatics, Data Science, Music & Astronomy Aug 31, 2017.

.