International Journal of Engineering & Technology, 7 (4.7) (2018) 335-338

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



An Analysis of Breast Cancer DNA Sequences Using Particle Swam Optimization

K. Lohitha Lakshmi[1], P.Bhargavi[2], S.Jyothi[3]

[1] *Research Scholar, Department of Computer Science, Sri Padmavati MahilaVisvavidyalayam, Tirupati-517 502*
[2] *Assistant professor, Department of Computer Science, Sri Padmavati MahilaVisvavidyalayam, Tirupati-517 502*
[3] *Professor, Department of Computer Science, Sri Padmavati MahilaVisvavidyalayam, Tirupati-517 502*

Abstract

Conceptual Breast tumour conclusion, examination, and visualization are essential research challenges in Bioinformatics. Bosom tumour analysis incorporates recognizing of malignancy bumps and ordinary tissue. Investigation incorporates the present phase of the malignancy tissue and anticipation incorporates expectation of repeat of the bosom tumour in future ages in light of structure and game plan of the individual DNA succession. This paper investigations bosom disease DNA succession to anticipate event of bosom tumour utilizing Particle Swarm Optimization (PSO).PSO procedure is a populace based pursuit calculation that mirrors the social conduct of swam. As the piece of investigation of bosom disease in human, the DNA arrangements of ordinary bosom tissue are contrasted and DNA groupings of bosom tumour tissue utilizing PSO... The distinction between the ordinary and breast cancer disease DNA sequences are broke down in view of the summarized values generated by applying PSO algorithm.

Keywords: *PSO, Soft Computing Techniques, Breast Cancer, Diagnosis, Analysis, Prognosis.*

1. Introduction

Soft Computing (SC) is a noteworthy research zone in programmed control building. Soft Computing mirrors the lead of human cerebrum to defeat the downsides of customary processing to deal with a perplexing constant issue. To take care of a Real time issue needs to work with perplexing, dubious, halfway, uncertain, and inexact information. SC strategies are urbanized to manage such kind of information and these procedures give more resistance to such sort of information. The utilizations of SC systems at present end up being more risen in the region of Bio-Informatics because of its resistance towards complex hereditary information. Bio-informatics is a developing territory which works with natural issues with the guide of data innovation to tackle hereditary related complex ongoing problems.SC incorporates differing regions like Fuzzy Logic, Genetic Algorithms, Artificial Neural Networks, Machine Learning, and Expert Systems [1].

As indicated by National Breast Cancer Organizations (NABCO), bosom malignancy is a typical type of growth in ladies and ends up basic infections which condemned towards death except if it is determined and treated having in time. That why it is dealt with as critical malady to do inquire about for conceivable location and fix in beginning time. As per NABCO growth is an ailment that happens when cells wind up irregular and gap without control prompts frame a tumour [2]. In breast cancer such tumour forms in breast of women in common and rarely in men [3].

To diagnose the cause of breast cancer researchers are concentrating on analysing gene sequences of cancerous cells for detecting any abnormal changes. Soft Computing is identified as a developing field for genetic related data experiments through its various techniques.

2. SC Techniques for Gene Sequence Analysis

SC Techniques for Gene Sequence Analysis are enlivened by nature and can give answers for true issues by utilizing some improvement systems. These procedures are useful in transforming knowledge from normal framework to computational framework [4]. Science is a piece of nature. The calculations or methods created in SC can manage quality successions which are parallel, irregular, random, dynamic, asynchronous, partial, and complex information which is enlivened by nature. Gene sequence analysis is helpful in identifying structural similarities of genes and to predict some critical and hereditary diseases [5].

Proteins are considered as building squares of life constructed using a chain of simpler molecules called amino acids linked by polypeptide bonds. DNA conveys encoded data necessary to build protein in an organism.DNA is a sequence of organic molecules called nucleotides. Every nucleotide is made up of sugars and bases. The four bases are Adenine (A), Thymine (T), Guanine (G), and Cystone (C). These bases match up in presence of hydrogen bonds. A *gene* is a portion of DNA. The entire DNA content of a cell is known as *genome*. Gene Sequence Analysis or Gene Prediction is the issue of distinguishing the portions of DNA sequence that are biologically functional.

Exploratory gene sequence analysis is moderate and tedious. Data included in such type of genetic experiments is very complicated. Such experiments need some advanced techniques which can deal with dynamic data and are incorporated with computational intelligence to support real time systems [6]. A definitive objective of SC techniques is to imitate human mind to manage genuine issues with imprecision, vulnerability, estimated, vigour, and minimal effort arrangements [7] [8]. The one of a kind of methodologies developed in soft computing is the capacity of gaining from ex-

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ploratory information which makes it appropriate for hereditary trials. Numerous trial issues identified with hereditary qualities require Fuzzy Logic, Neural Networks and Genetic Algorithms related techniques which is the part of SC for getting optimal solution.

3. Need of Breast Cancer Gene Sequence Analysis

Cancer is a hereditary disease that is caused by changes in DNA succession. These changes sometimes may be inherited and sometimes arise randomly during life time. Each kind of tumour has some one of a kind blend of hereditary changes. To identify these unique changes in DNA sequence analysis will be help full. Sometimes as indicated by the hereditary modifications in cancer can help in determining the treatment plan. At times as the part of cancer diagnosis tumour DNA sequences will be contrasted with normal healthy DNA sequences. Based on the tumours unique genetic alterations treatment design will be chosen [9].

Breast cancer is a standout among the most broadly perceived diseases happens in individuals and the rate among woman is most when compared to men (women account for 99%).Breast cancer frequently spread through lymphatic system and is a perilous sickness. The gene research has been rapidly developed on the breast cancer due to the reason that changes or erasures are related with numerous qualities in breast cancer. The susceptibility of most commonly mutant genes are under study for example TP53, PTEN, RUNX1 (being detected in 7% to 15% cases), CCND3, PTPN22 (low frequency mutation genes) [10].

A gene that is usually found in breast tumour cells is the BRCA gene. In another theory developed by molecular biologist Massaki Hemaguchi noticed that a gene called DCB2 is completely absent in cancerous cells. They also observed that while mRNA is missing 58% in breast cancer tissue. Other than BRCA1/2, PALB2 is also treated as a second most much of the time mutated gene in breast cancer as indicated by research [11]. In such way extraordinary examinations uncover that numerous qualities are related with the event of breast cancer. Incidence of many other factors also leads to the cause of breast cancer. Different facts related to mutant genes leads to perform experiments on genetic sequences [12].

4. Particle Swarm Optimization Algorithm for Breast Cancer Sequence Analysis

PSO and Ant Colony Optimization [ACO] are population based evolutionary behavioural algorithms. Normal for every individual gene decides the properties of population [13].Particle swarm optimization [PSO] are a population based stochastic search algorithm that emulates the capability (social behaviour or similarity) of flocking birds. A molecule flies with the speed which is powerfully balanced in light of its own and its colleagues past conduct. Every molecule position speaks to an answer for the problem. Particles have an affinity to fly toward better hunt regions over the way of the search procedure [14], [15]. Different topology structures can be adapted for PSO with diverse systems to share data for every molecule in the swarm... Due to this sort of behaviour in finding the similarity PSO is implemented in finding similarity and analysing genetic sequences.PSO is producing good results progressively in real time applications [16]. Analysis gene sequence is a discrete optimization issue for this a discrete optimization algorithm is needed.PSO fulfils such need and the intensity of PSO is its straightforward calculations and data will be imparted to in the calculation which is gotten from the social conduct of individuals(particles) in the swarm. A particle flown through the multidimensional inquiry space and each particle provides possible optimal solution to the multi dimensional issue. All individual solution fitness is estimated based on performance function which gives an optimal solution [17] and this process will be repeated until the joined arrangement is obtained for particular solution.

5. Implementation

In the proposed strategy PSO is implemented using python. Python is a perfect language for applications for life science study and development. Python language has self instruction tool which is convenient reference to solve real life programming problems. Python is an interactive language every unit of python is interactive means each command produces instant results. This feature is suitable for handling dynamic data for real world applications. Python is more flexible language for bio informatics applications [18]. Due this influential nature of Python language it is opted for PSO implementation in proposed algorithm. Python programming for bioinformatics applications is Bio Python helpful for true critical thinking. A, C, G and T letters are used to represent bases in DNA sequence [19].DNA and RNA both carry genetic information. There lies a slight contrast between them in structure and functionality. The functionality of DNA is capacity and transmission of genetic data to make different cells and new organisms. RNA utilized to transmit hereditary information.RNA has simpler structure and is needed for DNA to function [20].Messenger person RNA (mRNA) is an extensive group of RNA particles that pass on hereditary data from DNA to the ribosome [21]. mRNA is a solitary stranded copy of the gene and afterwards translated in to a protein atom [22]. In the present paper mRNA sequences are considered for PSO implementation to lessen time complexity and space complexity and due to its similar functionality with DNA sequences.

5.1 Python Implementation of PSO

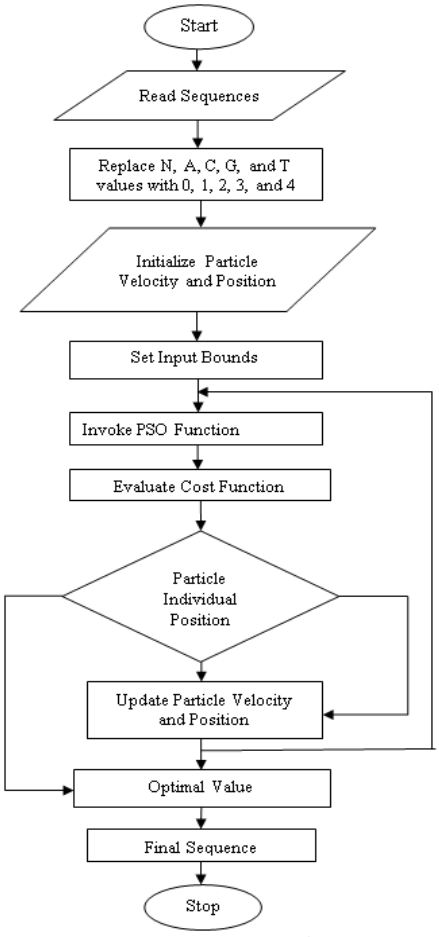


Fig 5.1 Flow chart for PSO implementation

1. Read the mRNA sequences from NCBI site
2. Replace N, A, C, G, and T values with 0, 1, 2, 3, and 4
3. Initialize the starting location, velocity of the particle and individual best position to zero. Individual best error value and individual error value with -1.
4. Set the input bounds to some value which specifies the beginning position of the particle movement and closure position of the molecule movement.
5. Invoke the PSO function and best error for group and best position for group are initialized.
6. Establish the swarm and evaluate the current fitness value using cost function. Cost function is a user defined function with required mathematical equation.
7. Check the current position is individually best or not. There are three unique forces working on each particle. They are particles initial velocity, distance from the individual particle that is cognitive force and separation from the swarm's best known position called social force.
8. Refresh the speed and position for each particle in all emphases through swarm.
9. Steps from 6-8 are rehashed until the improved esteem is produced. The produced last ideal value determines that the present particle or succession is the best solution (universally) for current issue.
10. This process is rehashed for a population of Gene Sequences.
11. The optimized value of gene sequences is used for analysis of breast cancer.

The Details of input sequences with Accession Numbers are:

Breast Cancer Data Sequence Analysis

Table 5.1 List of breast cancer input sequences

SNO	ACCESSION NO & DESCRIPTION
SEQ 1	NM_139163.3 -Homo sapiens amyotrophic lateral sclerosis 2 chromosome region 12 (ALS2CR12), transcript variant 1
SEQ 2	NM_001289993.1-Homo sapiens amyotrophic lateral sclerosis 2 chromosome region 12 (ALS2CR12), transcript variant 3
SEQ 3	NM_001127391.2- Homo sapiens amyotrophic lateral sclerosis 2 chromosome region 12 (ALS2CR12), transcript variant 2
SEQ 4	NM_025059.3 Homo sapiens coiled-coil domain containing 170 (CCDC170)
SEQ 5	NM_001024957.1 Homo sapiens breast cancer metastasis suppressor 1 (BRMS1), transcript variant 2
SEQ 6	NM_015399.3 Homo sapiens breast cancer metastasis suppressor 1 (BRMS1), transcript variant 1
SEQ 7	NM_017909.3 Homo sapiens required for meiotic atomic division 1 homolog (RMND1), transcript variant 1
SEQ 8	NM_001271937.1 Homo sapiens required for meiotic nuclear division 1 homolog (RMND1), transcript variant 2

(i) Non Breast Cancer Data Sequence Analysis

Table 5.2: List of non-breast cancer input sequences

SNO	ACCESSION NO & DESCRIPTION
SEQ 1	NM_052997.2 Homo sapiens ankyrin repeat domain 30A (ANKRD30A)
SEQ 2	NM_001174088.1 Homo sapiens nuclear receptor coactivator 3 (NCOA3), transcript variant 4
SEQ 3	NM_001174087.1 Homo sapiens nuclear receptor coactivator 3 (NCOA3), transcript variant 3
SEQ 4	NM_006534.3 Homo sapiens nuclear receptor coactivator 3 (NCOA3), transcript variant 2
SEQ 5	NM_181659.2 Homo sapiens nuclear receptor coactivator 3 (NCOA3), transcript variant 1
SEQ 6	NM_178863.4 Homo sapiens potassium channel tetramerization domain containing 13 (KCTD13), transcript variant 1
SEQ 7	NM_001278580.1 Homo sapiens activin A receptor type 2A (ACVR2A), transcript variant 3
SEQ 8	NM_001616.4 Homo sapiens activin A receptor type 2A (ACVR2A), transcript variant 2
SEQ 9	NM_001278579.1 Homo sapiens activin A receptor type 2A (ACVR2A), transcript variant 1

PSO algorithm Result for Breast cancer sequences

Table 5.3 PSO result for breast cancer input sequences

Seq No	Seq Acc.No	PSO Result
SEQ 1	NM_139163.3	15210
SEQ 2	NM_001289993.1	11940
SEQ 3	NM_001127391.2	14726
SEQ 4	NM_025059.3	37705
SEQ 5	NM_001024957.1	9181
SEQ 6	NM_015399.3	9824
SEQ 7	NM_017909.3	14178
SEQ 8	NM_001271937.1	10561

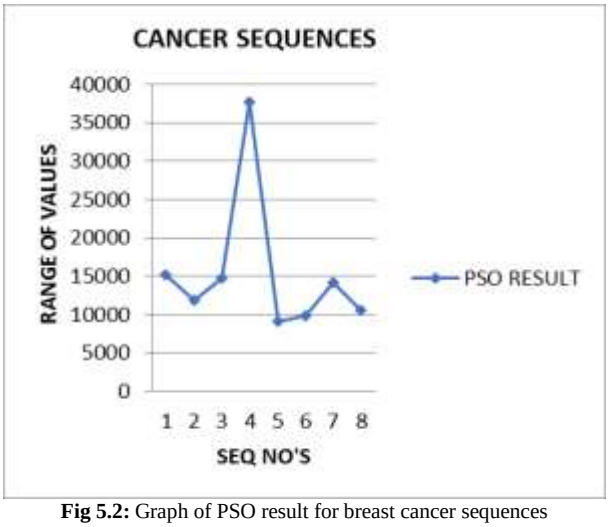


Fig 5.2: Graph of PSO result for breast cancer sequences

PSO algorithm Result for normal or healthy sequences

Table 5.4 PSO result for normal input sequences

Seq No	Seq Acc.No	PSO Result
SEQ 1	NM_052997.2	28989
SEQ 2	NM_001174088.1	60373
SEQ 3	NM_001174087.1	60633
SEQ 4	NM_006534.3	60564
SEQ 5	NM_181659.2	60554
SEQ 6	NM_178863.4	12296
SEQ 7	NM_001278580.1	40386
SEQ 8	NM_001616.4	41557
SEQ 9	NM_001278579.1	41974

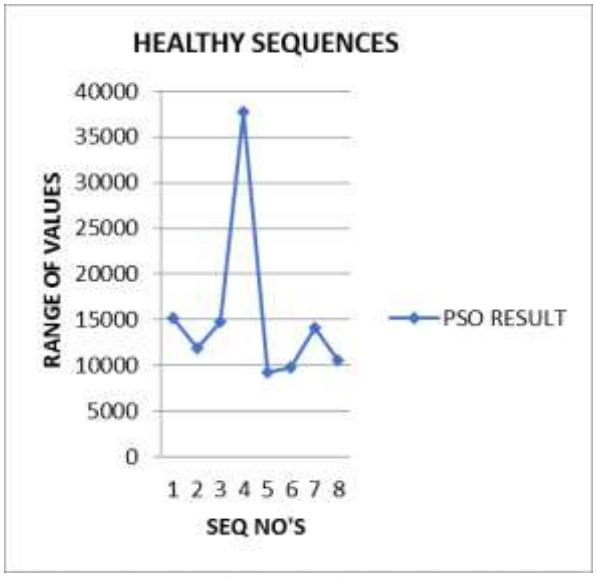


Fig 5.3: Graph of PSO result for healthy sequences

6. Result Analysis

In the proposed technique the ideal value or resultant value generated after implementation of PSO algorithm on individual breast cancer sequence, the value is in the range of 9000-15000 for 95% of sequences. Only 5% values are deviating from this scope of qualities to some extent. These values are represented in table 5.3. Similarly table 5.4 is representing the values when PSO algorithm implemented on individual non breast cancer or healthy breast sequences. The optimal values generated in each implementation lies in between the range of 25000-61000 for 95% of sequences. When observed only 5% range of values is deviating from these ranges in case of non breast cancer sequences like as cancer sequences.

It can be observed that PSO generated optimal values for breast input sequences are less than the PSO generated optimal values for normal cancer sequences. In Fig 5.2 it is graphically represented that maximum probability of PSO generated values for breast cancer sequences are in range of 9000 to 16000 except one value i.e. approximately 38000. In case of non breast cancer sequences these values ranges approximately from 28000 to 61000 except one value i.e. 12000 which is graphically represented in Fig 5.3. In the proposed technique PSO is implemented on 8 test sequences from breast cancer category and 9 test sequences are taken from non breast cancer category.

Based on these optimal values generated after PSO implementation on breast cancer and non breast cancer sequences the proposed method provides a scope to identify the difference between those sequences. This methodology can be used on any test sequences to identify whether the sequence is similar to breast cancer sequence based on the PSO value in most of the cases with high probability. These results can be used for further diagnosis.

7. Conclusion

In the proposed technique when PSO values of both breast cancer sequences and non breast cancer sequences are compared. The range of values of breast cancer sequences are less when compared to non breast cancer sequences PSO values at most maximum probability except less than 5 percent values. Based on the PSO result values the proposed method is useful to find the difference between cancer and non cancer gene data sequences. In future the proposed method will be implemented with other algorithms from Soft Computing techniques. The last outcomes which are obtained with the implementation of different algorithms will be compared to justify which algorithm is producing more accurate results in such type of genetic sequence analysis.

References

[1] Dogan Ibrahim, "An overview of soft computing", 12th International Conference on Application of Fuzzy Systems and Soft Computing, ICAFS 2016, 29-30 August 2016, Vienna, Austria, Procedia Computer Science 102 (2016) 34 – 38, Science Direct.

[2] Analyn Salaria," Breast Cancer", Malabog National High Schoolsalvacion, Daraga, Albay, A RESEARCH PAPER IN ENGLISH 1V.

[3] Web reference: https://www.customwritings.com/blog/sample-research-papers/research-paper-breast-cancer.html

[4] Nature Inspired Computation Techniques and Its Applications in Soft Computing: Survey K. Himabindu1 , S. Jyothi2 1, 2Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam (Women's University), Tirupati, INDIA, International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887 Volume 5 Issue VII, July 2017- Available at www.ijraset.com

[5] K. Bhargavi* and S. Jyothi," Classification of DNA Sequence Using Soft Computing Techniques: A Survey", Indian Journal of Science and Technology, Vol 9(47), DOI: 10.17485/ijst/2016/v9i47/89343, December 2016 ISSN (Print): 0974-6846 ISSN (Online): 0974-5645.

[6] Neelam Goel, Shailendra Singh, and Trilok Chand Aseri, "A Review of Soft Computing Techniques for Gene Prediction", ISRN Genomics, Volume 2013, Article ID 191206, 8 pages, http://dx.doi.org/10.1155/2013/19120666.

[7] A. B. Kurhe, S. S. Satonkar, P. B. Khanale, and S. Ashok, "Soft computing and its applications," BIOINFO Soft Computing, vol. 1, pp. 5–7, 2011. View at Google Scholar.

[8] S. Rajasekaran and G. A. V. Pai, Neural Network, Fuzzy Logic and Genetic Algorithms- Synthesis and Applications, Prentice-Hall, 2005.

[9] Webreference:https://www.cancer.gov/about-cancer/treatment/types/precision-medicine/ tumour-dna-sequencing.

[10] RongMaJianpingGongXiaoweiJiang, "Novel applications of next-generation sequencing in breast cancer research", Genes & Diseases, volume 4, Issue 3, September 2017, Pages 149-153,open access.

[11] Kokichi Sugano,Seigo Nakamura ,Jiro Ando ," cross-sectional analysis of germ line BRCA1 and BRCA2 mutations in Japanese patients suspected to have hereditary breast/ovarian cancer", 23oct2008.

[12] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in Proceedings of the Sixth International Symposium on Micro Machine and Human Science, 1995, pp. 39–43.

[13] K. LohithaLakshmi, P.Bhargavi, S.Jyothi," Soft Computing Techniques for Gene Annotation", IJLEMR, *ISSN:*2455-4847,Volume 3-Issue 2,February 2018,PP 26-34.

[14] J. Kennedy and R. Eberhart, "Particle swarm optimization," in Proceedings of IEEE International Conference on Neural Networks (ICNN), 1995, pp. 1942–1948.

[15] R. Eberhart and Y. Shi, "Particle swarm optimization: Developments, applications and resources," in Proceedings of the 2001 Congress on Evolutionary Computation (CEC2001), 2001, pp. 81–86.

[16] S. Cheng, Y. Shi, and Q. Qin, "Population diversity based study on search information propagation in particle swarm optimization", in Proceedings of 2012 IEEE Congress on Evolutionary Computation, (CEC 2012). Brisbane, Australia: IEEE, 2012, pp. 1272–1279.

[17] Vijaylakshmi S and Priyadarshini J," An Analysis of Particle Swarm Optimization Technique for Breast Cancer Dataset", I J C T A, 9(3), 2016, pp. 297-308, © International Science Press.

[18] Ravi Shankar Verma, Vikas Singh, Sanjay Kumar," DNA Sequence Assembly using Particle Swarm Optimization", International Journal of Computer Applications (0975 – 8887), Volume 28– No.10, August 2011.

[19] Web ref: https://www.freelancinggig.com/blog/2017/07/19/best-programming-languages-bioinformatics/

[20] Hans Petter Langtangen, Geir Kjetil Sandve", Illustrating Python via Bioinformatics Examples", Center for Biomedical Computing, Simula Research Laboratory, Department of Informatics, University of Oslo, Mar 22, 2015.

[21] Web ref: https://www.thoughtco.com / dna-versus-rna-608191

[22] Web ref: https://en.wikipedia.org/wiki/ Messenger_RNA.

[23] Suzanne Clancy, Ph.D. & William Brown, Ph.D. (Write Science Right) © 2008 Nature Education

[24] Citation: Clancy, S. & Brown, W. (2008) Translation: DNA to mRNA to Protein. Nature Education 1(1): 101.