IPASJ International Journal of Computer Science (IIJCS)
Web Site: http://www.ipasj.org/IIJCS/IIJCS.htm
A Publisher for Research Motivation ..... Email: editoriijcs@ipasj.org
Volume 4, Issue 6, June 2016 ISSN 2321-5992

Prediction of Baldness of Upcoming Generation using Genetic Algorithm Crossover Mutation and Multiple Sequence Alignment

[1]Uma Anand, [2]Chain Singh

[1]Student M.Tech (4th sem) Department of Computer Science Engineering
Dronacharya College of Engineering, Gurgaon-123506, India

[2]Assistant Proffesor Department of Computer Science Engineering
Dronacharya College of Engineering, Gurgaon-123506, India

ABSTRACT

Human genetics is involved in identifying genes that lead to an increased risk of Genetic disease in certain individuals. The identification of disease susceptibility in the genes can help us to improve human health through the development of new prevention, diagnosis, and treatment strategies. In our proposed paper we are researching on baldness i.e. We are predicting that weather upcoming generation of given sequence would suffer from baldness or not. Multiple sequence alignment (MSA) is a ubiquitous downside in machine biology. Although it is np-hard to seek out associate degree best answer for associate degree capricious range of sequences, due to the importance of this problem researchers are attempting to push the boundaries of actual algorithms any. Since MSA can be solid as a classical path finding downside, it is attracting a growing number of AI researchers inquisitive about heuristic search algorithms as a challenge with actual sensible connection

KEYWORDS: Genetic Algorithm, Multiple Sequence Alignment, Chromosome Information, Sequence Data, Alignment.

1. INTRODUCTION

A multiple sequence alignment (MSA) arranges protein sequences into a rectangular array with the goal that residues {in a|during a|in associate degree exceedingly|in a very} given column square measure homologous (derived from one position in an ancestral sequence), superposable (in a rigid native structural alignment) or play a common useful role. Although these 3 criteria square measure primarily equivalent for closely connected proteins, sequence, structure and function diverge over biological process time and totally different|completely different} criteria might end in different alignments. Manually refined alignments continue to be superior to purely automatic methods; there's thus an eternal effort to boost the biological accuracy of MSA tools. Additionally, the high computational value of most naive algorithms motivates enhancements in speed and memory usage to accommodate the fast increase in on the market sequence knowledge. [1] . With advances in modern molecular biology, the DNA sequence of virtually any sequence is offered. This has naturally become the starting purpose for any genetic analysis. To formally introduce type two alpha-5-reductase, its genetic profile is presented. Although there is a another Isoform of this protein within the physique.

Multiple sequence alignment (MSA) are a vital and generally utilized computational methodology for natural succession examination in sub-atomic science, computational science, and bioinformatics. MSA are finished where homologous arrangements are contrasted all together with perform phylogenetic recreation, protein optional and tertiary structure examination, and protein capacity forecast investigation [2]. Organically great and precise arrangements can have huge importance, demonstrating connections and homology between various groupings, and can give helpful data, which can be utilized to promote recognize new individuals from protein families. The precision of MSA is of basic significance because of the way that numerous bioinformatics methods and methodology are subject to MSA results [3].

In this paper, we 1st review 2 previous, complementary lines of research. Based on hirschberg's algorithmic rule, dynamic programming needs o(knk−1) area to store each the search frontier and the nodes required to reconstruct the answer path, for k sequences of length n. Best first search, on the other hand, has the advantage of bounding the search space that has to be explored employing a heuristic. However, it is necessary to take care of all explored nodes up to the ultimate solution so as to stop the search from re-expanding them at higher value. Earlier approaches to reduce the closed list square measure either incompatible with pruning ways for the open list, or must retain at least the boundary of the closed list.

2. Proposed Model

Medical science has found that some disease like diabetes, heart problem, and baldness is hereditary, means this disease has been transferred from parents or maternal parents to child or upcoming generations.
Although male pattern baldness is a very common condition and it is even a harmless condition. It can occasionally be linked to metabolic syndrome that is changes in gene. This can lead to severe condition of obesity, diabetes, raised blood pressure and raised cholesterol. People with this syndrome have an increased risk that they will suffer from heart disease. This is most often seen in men who develop baldness at a relatively young age.
In our proposed approach we are researching on baldness i.e. We are predicting that weather upcoming generation of given sequence would suffer from baldness or not.

Steps involved in processing sequence of DNA:

1. **Start**

2. **Initialization**: Number of sequence is calculated after searching maximum number of gaps given with respect to the bigger number of sequence in the set of DNA sequences that needs to be summarized. For e.g. The summarised sequences DNA length is given by length, creates initial alignment by inserting required number of gaps given by, length-Sequence length (i). An initial population of several alignments is created in this way. Size of the initial population is set.

3. **Chromosome representation**: Encode the alignments of initial population into chromosomes using the representation scheme.

4. **Genetic operations**: Creates a newly created population using following steps repeatedly, until the smallest desired fitness value is not get or desired n generations are done:

- **Selection**: using selection schemes like elitism or random selection, few sequences are selected to perform crossover & mutation operations.
- **Crossover** operations are generally acts on the pairs of minimum fit chromosomes. Single point crossover, double point crossover and min-max crossover methods have been used.
- **Selection for next generation**: chromosomes with better fitness values among the lot are used for producing other fit chromosomes using crossover and mutation schemes. Here we have experimented with a simple scheme where the chromosomes produced whose fitness value is less than the parent chromosomes are discarded. I.e. The best 2 chromosomes of parent1, parent2, child1 and child2. One & two point crossover schemes are tried.
- **Mutation operation** is performed on selected chromosomes. Following mutations are performed- random gap shuffling, insertion and deletion of gaps.
- **Calculate overall alignment fitness value** of the obtained alignments from crossover & mutation operations.
- **Discard the chromosomes**, whose fitness value is not sufficient then the parent chromosomes. Save the alignment representation & its associated parameters.

5. **Result:** the best sequence alignment would be corresponding to the chromosome with highest fitness value after n generations are done or desired minimum acceptable score is obtained.

6. **End**

3. Computation Model

Androgenic alopecia (AGA), additionally alluded to as male example hair sparseness, is a typical dermatologic condition tormenting about half percent of Caucasian guys beyond 50 years old. AGA is less continuous in guys of Asian, American-Indian, or African tolerable, yet shows itself in a trademark and unsurprising route in all cases. The unmistakable example is truly 1 of 7 variations on the Hamilton-Norwood scale(see figure), which was produced in 1975 [4]. The example has subsequent to end up named a blend of both a retreating frontal hairline and incomplete or complete male pattern baldness in the vertex district of the scalp[5]. As a rule some hair stays to frame a ring around the transient, parietal, and occipital regions.....the great horseshoe design.- 5

IPASJ International Journal of Computer Science (IIJCS)
Web Site: http://www.ipasj.org/IIJCS/IIJCS.htm
A Publisher for Research Motivation ..... Email: editoriijcs@ipasj.org
Volume 4, Issue 6, June 2016 ISSN 2321-5992

Hamilton-Norwood Scale

Figure 1: Standards for classification of most common types of male pattern baldness

3.1. Genetic basis of male pattern baldness

There are two Iso Enzymes of 5-alpha reductase found in the human genome, sort 1 and sort 2. The principle activity of both isoenzymes is the transformation of testosterone into the more intense androgen dihydrotestosterone (DHT). SRD5A2 is the quality that codes for the protein item that is in charge of the pathogenesis of male example hairlessness. Ponders utilizing monoclonal antibodies have discovered sort 2 5AR confined in the vertex and frontal scalp, however not the occipital districts. Besides, these same areas have been appeared to be androgen touchy, which means the hair becomes constitutively without the hormone. Within the sight of rich DHT, the hair follicles get to be scaled down; bringing about fine, short hair that is inclined to fall out[6].

The legacy of male example hair sparseness may include more qualities and turns out to be more perplexing as more revelations are made (see quality data page). There is however much established proof that pinpoints the immediate inclusion of DHT in male example hair sparseness. Investigations of pseudohermaphrodites (lacking 5AR) indicated security from male example sparseness all through life. It has additionally been demonstrated that the presentation of exogenous testosterone into emasculated guys can prompt commonplace examples of male pattern baldness. At last, inhibitors of the protein have hindered the rate of balding, demonstrating a direct involvement[7].

3.2. Chromosome Information

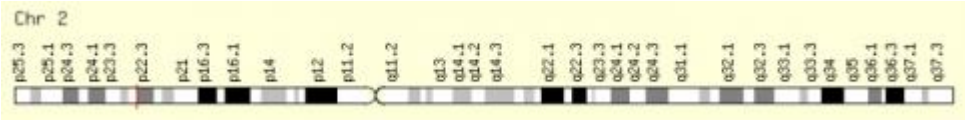

Figure 2 : The SRD5A2 gene is localized to chromosome 2. The cytogenetic band is 2p23.1

Arabidopsis Phenotypes

In Human Expression Profiling despite the fact that not straightforwardly examining the going bald condition, this expression profile from GEO on colon disease tumor backslide is somewhat fascinating.

IPASJ International Journal of Computer Science (IIJCS)
Web Site: http://www.ipasj.org/IIJCS/IIJCS.htm
A Publisher for Research Motivation ..... Email: editoriijcs@ipasj.org
Volume 4, Issue 6, June 2016 ISSN 2321-5992

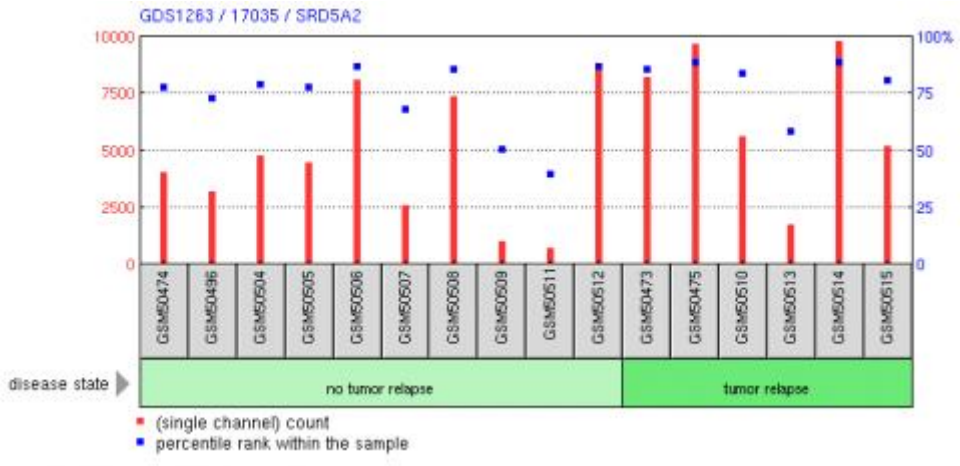

Figure 3: Arabidopsis Phenotypes

The *SRD5A2* gene provides information that is used for making an enzyme called steroid 5-alpha reductase 2. This enzyme is involved in processing androgens. Androgens are hormones that help in male sexual development. Specifically, the function of this enzyme is to converts the hormone testosterone to a more potent androgen, dihydrotestosterone (DHT), in male reproductive tissues.

4. Multiple sequence alignment Using GA

Hereditary calculations have been utilized for MSA creation as a part of an endeavour to extensively mimic the speculated developmental procedure that offered ascend to the uniqueness in the inquiry set. The strategy works by breaking a progression of conceivable MSAs into pieces and over and over revamping those parts with the presentation of crevices at different positions. A general target capacity is streamlined amid the recreation, most by and large the "total of sets" augmentation capacity presented in element programming-based MSA techniques. A procedure for protein groupings has been executed in the product program SAGA (Sequence Alignment by Genetic Algorithm) [9] and its proportional in RNA is called RAGA [8]. The strategy of reproduced toughening, by which a current MSA delivered by another technique is refined by a progression of adjustments intended to discover more ideal districts of arrangement space than the one the info arrangement as of now involves. Like the hereditary calculation technique, mimicked tempering expands a target capacity like the whole of-sets capacity. Recreated strengthening utilizes a figurative "temperature consider" that decides the rate at which improvements continue and the probability of every reworking; run of the mill utilization interchanges times of high revision rates with generally low probability (to investigate more removed areas of arrangement space) with times of lower rates and higher probabilities to all the more altogether investigate nearby minima close to the recently "colonized" locales. This methodology has been executed in the project MSASA (Multiple Sequence Alignment by Simulated Annealing) [10].

4.1 Outline of Genetic Algorithm

i)Choose initial population

ii)Evaluate the fitness of each individual in the population

iii)Repeat

a)Select best-ranking individuals to reproduce

b)Breed new generation through genetic operations (crossover and mutation) and give birth to offspring

c)Evaluate the individual fitness of the offspring

d)Replace worst ranked part of population with offspring

iv)Until a solution is found that satisfies minimum criteria or a fixed number of generations reached.

IPASJ International Journal of Computer Science (IIJCS)
Web Site: http://www.ipasj.org/IIJCS/IIJCS.htm
A Publisher for Research Motivation ..... Email: editoriijcs@ipasj.org
Volume 4, Issue 6, June 2016 ISSN 2321-5992

4.2 Applying GA to MSA

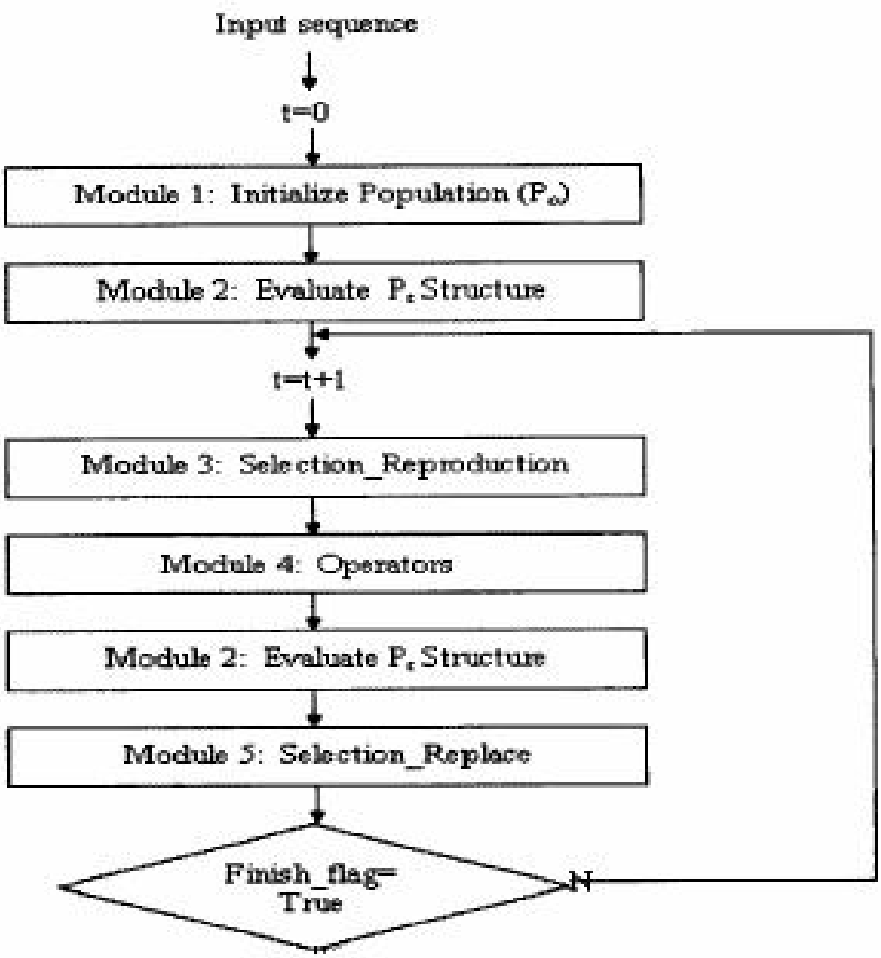

Figure 4: Applying GA to MSA

Step1: Initialize Population

The population is represented as an array of sequence where each sequence was encoded as an array of character over the alphabet. The symbol "-" will refer to the gap in the alignment which represent an insertion or a deletion of an amino acid residue.

Step2 :Evaluate P_t structure

Individual that scores a high fitness function (F) will survive for the next iteration. The Scoring function is as follows:

$$\cos t = \sum_{i=1}^{l} \sigma(S'[i], T'[i])$$

where $l=|S'| = |T'|$, $\sigma(x,x) = 1$ and $\sigma(x,y) = \sigma(-,y) = \sigma(x,-) = 0$

Step3 : Selection Reproduction

The selection probability for each individual is proportional to the fitness function value. In this case, the fitter the individual, the more likely it will be chosen. Compute selection probabilities for the current population based on fitness value. Select two individual randomly based on the selection probabilities to obtain clones which may then be subjected to mutation or recombination.

IPASJ International Journal of Computer Science (IIJCS)
Web Site: http://www.ipasj.org/IIJCS/IIJCS.htm
A Publisher for Research Motivation ..... Email: editoriijcs@ipasj.org
Volume 4, Issue 6, June 2016 ISSN 2321-5992

Step 4: Crossover Mutation

Crossover: The crossover operator will use point-to-point crossover. This operator takes two alignment sequences from the population. Hence it randomly selects a fully matched (no gap) column. After crossover, it evaluates Child 1 and Child 2. The fittest offspring will survive in the next iteration. [11]

Mutation: The mutation operator picks a random amino acid from a randomly chosen row (sequence) in the alignment. It then checks whether one of its neighbours has a gap. If this is the case, the algorithms swaps (2-opt) the selected amino acid with a gap neighbour. If both neighbours are gaps, one of them will be picked randomly.[12]

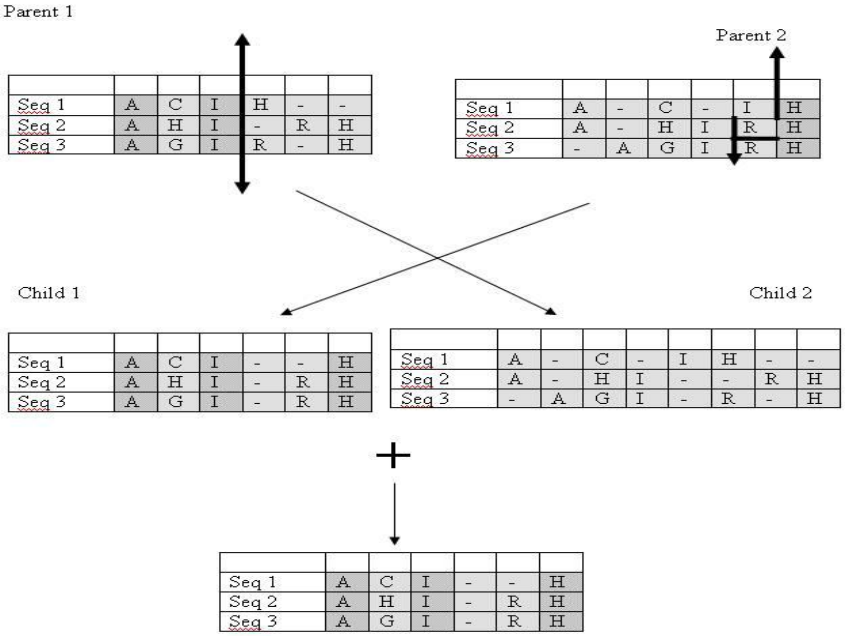

Figure 5: Cross Over and Mutation

Step 5: Selection Procedure

This Step will replace old individual having less fitness function values. It will also insert new offspring to the population.

5. Hereditary Testing

On the off chance that you knew you would have been uncovered in ten years, would you adjust your life know? Well perhaps you haven't contemplated it, yet it is currently a reality to know whether you harbor the hereditary qualities of male example sparseness. HairDX will test an example of your DNA for a variations of the androgen receptor quality and figured out whether you are liable to end up uncovered.

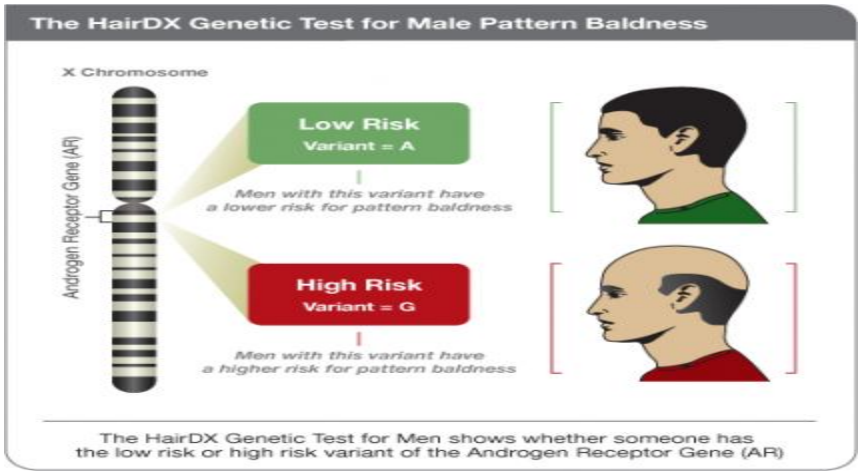

Figure 6: The HairDx Genetic Test For Male Pattern Baldness

The company also offers a test for female pattern baldness. A female DNA sample is a analyzed and a CAG repeat score is computed. A lower CAG score is indicative of higher risk for balding. The website claims that only 2.3% of women with a CAG score of 15 or lower do NOT show signs of hair loss.

6. Conclusion

There for the most part exist three classes of streamlining calculation for various arrangement; careful, dynamic and iterative. Various MSA programs have been connected utilizing numerous strategies and calculations. Most normally utilized methods are dynamic and iterative strategies. The precise strategy experiences vague grouping arrangement and lead to a forceful examination on dynamic and iterative calculations.

The precise strategy can adjust up to ten firmly related successions. However, when the quantity of groupings gets to be bigger, the space and time multifaceted nature is tremendous. Dynamic arrangement constitutes one of the least difficult approaches to adjust grouping. This methodology has the benefits of pace and straightforwardness. However the significant issue with dynamic arrangement strategy is that blunders in the underlying arrangements are the most firmly related succession spread to the different arrangement. Iterative arrangement strategies rely on upon calculations that can create an arrangement and to refine through a progression of cycles until no more change can be made. Iterative techniques can be deterministic or stochastic, contingent upon the procedure used to enhance the arrangement. The least complex iterative techniques are deterministic. It includes extricating the grouping one by one from various arrangements and realigning them to the remaining successions. This methodology is ended when no change can be made (union). Stochastic iterative strategies incorporate Hidden Markov Model (HMM), re-enacted strengthening and transformative calculation, for example, hereditary calculations (GAs) and developmental programming. Their fundamental preference is to take into consideration a decent detachment between the streamlining procedure and assessment criteria. It is the target work that characterizes the point of any advancement methodology.

REFERENCES

[1]. Yi, W., Ross, J.M., Zarkower, D. 2000. Mab-3 is a direct tra-1 target gene regulating diverse aspects of C. elegans male sexual development and behavior.
[2]. Wallace IM, Blackshields G, Higgins DG (2005) Multiple sequence alignments. Curr Opin Struct Biol 15: 261–266.
[3]. Gotoh O (1999) Multiple sequence alignment: Algorithms and applications. Adv Biophys 36: 159–206.
[4]. Blackshields G, Wallace IM, Larkin M, Higgins DG (2006) Analysis and comparison of benchmarks for multiple sequence alignment. In Silico Biol 6: 321–339.
[5]. Hogeweg P, Hesper B (1984) The alignment of sets of sequences and the construction of phylogenetic trees. An integrated method. J Mol Evol 20: 175–186.
[6]. Edgar RC (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.
[7]. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res 15: 330–340.
[8]. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: Improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511–518.
[9]. Simossis VA, Heringa J (2005) PRALINE: A multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. Nucleic Acids Res 33: W289–W294.
[10]. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302: 205–217.
[11]. O'Sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C (2004) 3DCoffee: Combining protein sequences and structures within multiple sequence alignments. J Mol Biol 340: 385–395.
[12]. Wallace IM, O'Sullivan O, Higgins DG, Notredame C (2006) M-Coffee: Combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res 34: 1692–1699.

AUTHOR

Uma Anand received the B.Tech. and M.Tech. degrees in Computer Science and Engineering from Dronacharya College Of Engineering in 2013 and 2016, respectively. She has done various researches in the same field.