ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com
Journal DOI: 10.21474/IJAR01

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

Multiple Sequence Alignment in Bioinformatics.

HarshitaBadlani, Abhinav Sinha, AbhinavMittal, AlokKumar(Asst. Prof., GCET)
Galgotias college of Engineering and Technology, Greater Noida.

Manuscript Info *Abstract*

Manuscript History:

Received:14 February 2016
Final Accepted: 18 March 2016
Published Online: April 2016

Key words:

Corresponding Author

HarshitaBadlani.
Abhinav Sinha
Abhinav Mittal

This paper proposes few genetic operators to obtain better alignments of multiple molecular sequences. All the proposed operators in the method have been implemented and validated within a self developed software tool which allows the user to select the various genetic operators for crossover, mutation, fitness calculation, population initialization. It guarantees the next generation of populations with better fitness value. Improvement in the overall population fitness is also calculated and evaluated. Survival of the fittest policy is followed to arrive at a better fitness in following generations. These fitness values then help to find heart and diabetes problems for that chromosome. Observations based on variable parameters have been recorded, analyzed & presented in the form of results. Results were also compared with few standard existing online tools to study the feasibility of the proposed operators.

Introduction:-

Multiple Sequence Alignment (MSA) is one of the most challenging and active ongoing research problems in the field of computational molecular biology. Multiple sequence alignment of DNA, RNA, or amino acids is essential for biologists to study similarity in sequences which often leads to similarity in function and provides valuable evolutionary information. The alignment enables us to infer the evolutionary history of the sequences.
Proposed algorithm guarantees that the next obtained generation of populations will have better fitness value as compared to their ancestors and therefore we can expect that the tool provides at least near to optimal alignments after some N number of generations. During the generation of the next population the tool ensures that only the fitter candidates (here alignments) are considered and weaker ones are ignored. The overall purpose remains to improve the alignment with each generation. On the basis of this fitness value, probability of heart and diabetes problems is also found for the next generation.

The proposed tool offers some of the advantages:
1. It always guarantees that the next obtained generation of populations will have better fitness value
2. Facility given to the users to set various GA parameters like crossover rate, mutation rate, no. of generations, selection of various implemented GA operations like selection, crossover or mutation schemes.
3. Probability of heart and diabetes problems in the next generation based on their fitness value.
The objective of this paper is to find the rules for finding fitness of any person based on his/her parents DNA and examining the probability of Heart or Diabetes diseases.

Related Work:-

Genetic algorithm (GA) refers to a model introduced and investigated by John Holland in 1975 for adaptation processes of nature. Generally stated, a GA is any population based model that uses selection and recombination operators to generate new sample points in a search space. GA computationally utilizes a natural evolutionary process similar to the processfirst described by Charles Darwin in his "The Origin of Species", to solve a given problem. GA is a global search procedure that searches from one population of points to another. GA is a

probabilistic search procedure, which is being frequently applied to difficult optimization and learning problems. There are two versions of the GA, namely the natural GA and the computational GA.

Genetic algorithms were inspired by the processes observed in natural evolution. They attempt to mimic these processes and utilize them for solving a widerange of optimization problems. In general, genetic algorithms perform directed random searches through a given set of alternatives with respect to the given criteria of goodness. These criteria are required to be expressed in terms of an objective function, which is usually referred to as a fitness function.

Genetic algorithms require that the set of alternatives to be searched through be finite. If we want to apply them to an optimization problem where this requirement is not satisfied, the set involved and select an appropriate finite subset. It is further required that the alternatives be coded in strings of some specific finite length which consist of symbols from some finite alphabet. These strings are called chromosomes, the symbols that form them are called genes, and their set is called a gene pool. Genetic algorithms search for the best alternative in the sense of a given fitness function through chromosomes evolution.

Genetic Algorithms search for the best alternative (in the sense of a given fitness function) through chromosomes' evolution. Basic steps in genetic algorithms figure. First, an initial population of chromosomes is randomly selected. Then each of the chromosomes in the population is evaluated in terms of its fitness (expressed by the fitness function).

Next, a new population of chromosomes is selected from the given population by giving a greater change to select chromosomes with the high fitness. This iscalled *natural selection*.The new population may contain duplicates. If given stopping criteria (e.g., no change in the old and new population, specified computing time, etc.,) are not met, some specific, genetic – like operations are performed on chromosomes of the new population. These operations produce new chromosomes, called offspring's. The same steps of this process, evaluation and natural selection, are then applied to chromosomes of the resulting population. The whole process is repeated until given stopping criteria are met. The solution is expressed by the best chromosome in the final population.

There are many variations on these basic ideas of genetic algorithms. To describe a particular type a genetic algorithm is greater detail, let G denote the gene pool, and let n denote the length of strings of genes that form chromosome. That is, chromosomes are n - tuples in Gn. The size of the population ofchromosomes is usually kept constant during the execution of genetic algorithm. That is, when newmembers are added to the population, the corresponding the umber of old members are excluded. Let m denote this constant population size.
Since each population may contain duplicates of chromosomes, we express populations by m-tuples whose elements are n-tuples from the set Gn. Finally, let f denote the fitness function employed in the algorithm.

Proposed Work:-

The pseudo code is as follows:
1. **Start**
2. **Initialization: S**equence length is computed after finding maximum number of gaps allowed with respect to the longest sequence in the set of sequences that needs to be aligned. Let say the aligned sequences' length is given by ***length***, generate initial alignment by inserting required number of gaps given by, ***length- sequence_Length(i).*** An initial population of several alignments is created in this manner. Size of the initial population can be set by the user as well.
3. **Chromosome Representation:** Encode the alignments of initial population into chromosomes using the representation scheme described later in the section**.**
4. **Genetic Operations:** Create a new population using following steps repeatedly, until the minimum desired fitness value is not obtained or desired N generations are done :
Selection: Using selection schemes like **elitism** or **random selection**, few sequences are selected to perform crossover & mutation operations.
5. ***Crossover*** operations are performed on the pairs of less fit chromosomes. **Single point** crossover, **double point** crossover and **min-max** crossover methods have been used.

Selectionfornextgeneration:chromosomeswithbetterfitnessvaluesamongthelotareusedforproducingotherfitchromosomesusingcrossoverandmutationschemes.Herewehaveexperimentedwithasimpleschemewherethechromosomesproducedwhosefitnessvalueislessthantheparentchromosomesarediscarded.i.e.thebest2chromosomesofparenti,parentj,childiandchild j.One &twopointcrossoverschemesaretried.
Mutationoperationisperformedonselectedchromosomes**.**Followingmutationsareperformed-
randomgapshuffling,**insertionanddeletion**of gaps.
Calculateoverallalignmentfitnessvalueoftheobtainedalignmentsfromcrossover&mutationoperations.
Discard the chromosomes, whose fitness value is less than the parent chromosomes. Save the alignment and its associated parametes.
Result:-
Thebestsequencealignmentwouldbecorrespondingtothechromosomewithhighestfitnessvalueafter Ngenerationsaredone ordesiredminimumacceptablescoreisobtained.
End
ExampleDemonstration:INPUTSEQUENCE
>MMVHLTPMMKSAVTALWGKVNVDMVGGMALGRLLVVYPWTQRFFMSFGDLSTPDAVM
>MMGLSDGMWQLVLNVWGKVMADIPGHGQMVLIRLFKGHPMTLMKFDKFKHLKSMDMMKAS
>ALVMDNNAVAVSFSMMQMALVLKSWAILKKDSANIALRFFLKIFMVAPS
>MMRPMPMLIRQSWRAVSRSPLMHGTVLFARLFALMPDLLPLFQYNCRQFSSPMD

Initialization
Weinsertgapsintheinputsequencetomaketheinitialpopulationofsay10alignments.
Pseudocode:init_pop(*lmax:*lengthoflongestinputsequence,leni:lengthofithinputsequence,iPop:initialpopulationset)
{
Calculatethelengthofsequenceinalignmentusing*length(N)*=1.2**lmax*
Fork=1to ipop
{
Foreachsequencei=1to N
computeno.ofgapstobeinserted asgapi=*length*-leni.
Foreachsequencei=1to N
insertgap
inumberofgapsatrandompositions.
thisinitialalignmentisreferredusingseq-alk.
}
}
lmax=61,correspondingtothelongestsequence,***length(N)*****=**1.2***lmax*=**74,gap1=*length*-len1=17,gap2=*length*-len2=13,gap3=*length*-len3=25,gap4=length-len4=20**

InitialPopulation'sSingleAlignmentInstanceAfterinsertionofgaps,asstatedinthealgorithm:
>MMVHLT---PM---MKSAV-T-AL-WGKVNVDMVGGMALGR--LLV-VYPWTQ-R-FFMSF-GDLSTPDA--VM
>M-MGL--SDGM-WQ-LVL-N--VW-GKVM-ADIP-GHGQMVLIRLFKGHPMTL-MKFDKF-KHLKSMD-MMKAS
>A-LVMDNNA--VAV--S--FS--MM-Q--MA--LVL-KS-W-A-ILKKD---S-A-N-IALRFFLKIFM-VAPS
>MMRP-MPML-I-RQSWR--AVS-RS-P-LMHGT-VLF-ARLFALM--PDLLP--L---FQ-YNCRQF-SSP-MD

Reproduction/Selection:
Reproductionisusuallythefirstoperatorappliedonpopulation.Chromosomesareselectedfromthepopulationtobe parents tocrossoverandproduceoffspring.According toDarwin's Theoryof survivalof thefittest,thebest onesshouldsurviveand createnewoffspring**[4]**.That iswhyreproductionoperatoris sometimesknown astheselectionoperator.

Crossover
Crossoverisaprocessoftakingmorethanoneparent chromosomesandproducingachildsolutionfromthem.Inourtool,we haveimplementedthreetypesofcrossovers-singlepointcrossover,twopointcrossoverandmax-mincrossover.
Crossoverisperformedby
selectingtwoparentswithhigherfitnessvaluesasshowninexampleandthenselectingasinglecrossoverpointwhichmaybesomeformulabasedorrandomlydeterminedbasedonthelengthoftheparents.Eachsuchcrossoverresultsintwochildchromoso

mes.Asanexperimentalschemewehaverestoredonlythosechildchromosomeswhichhavebetterfitnessscoresthantheirparents.

Mutation

Mutationisageneticoperatorusedtomaintaingeneticdiversityfromonegenerationofapopulationofalgorithmchromosomestothenext.Mutationaltersoneormoregenevaluesinachromosomefromitsinitialstate.Aftercrossoverthebestsetofchromosomeswithnumberofchromosomesequalingtothesizeof**iPop**(initialpopulationset)areselectedandmutationisappliedupontheem.

GapShuffling

Gap-Shuffle-Mut(chrom-pop$_m$: chromosomegenerationofmchromosomes,mutP:mutationprobability)

{

fork=1tomrepresentingeachchromosome,performoperationsondecodedalignmentsequence,seq_al$_k$

{

forj=0tomutP *length(seq)

{

forallsequencesseq,

findanygapinseqandshuffleitrandomlywithintheseq.

}

Calculatethefitnessofthenewalignment.

Outofparentandchild,bestalignment'schromosomebecomespartofnextgeneration.

}

}

FitnessFunction

Thefitnessfunctiondetermineshow"good"analignmentis.Fitnessevaluationmethodsplayanimportantroleintheperformanceofevolutionaryalgorithms.Themostcommonstrategythatis used,albeitwithsignificantvariations,iscalledthe "Sum-Of-Pair" ObjectiveFunction.Inthis method,foreachlocationonthealignedsequences,one ofthreesituationswilloccur:match,mismatch oragap.Thefitnessofanalignmentis calculatedas**fitness=symReward–Pen(d,g)**wheresymRewardistheoverallrewardofallpairwisesymbolmatches.Duringthefitnessevaluation,allfully-gappedcolumnsinanalignmentareignored.

Acknowledgment:-

It is grateful opportunity for us to write this paper. At the time of preparing this paper, we have gone through different books and websites.

We acknowledge with gratitude to our assistant professor Alok Kumar, who has always been sincere and helpful in making us understanding different problems.

Conclusion:-

We'veusedvariousmethodsofcrossover,mutationandselectionschemesformultiplealignment.Theresultsofeachalignmenttendtoimprove,whichisbeingshownbytheincreasingfitnessvaluewithincreaseinnumberofiterations.

References:-

1. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA*, 74, 5463–5467.

2. Maxam,A.M. and Gilbert,W. (1977) A new method for sequencing DNA. *Proc. Natl Acad. Sci. USA*, 74, 560–564.

3. Sanger,F., Air,G.M., Barrell,B.G., Brown,N.L., Coulson,A.R., Fiddes,C.A., Hutchison,C.A., Slocombe,P.M. and Smith,M. (1977) *Nucleotide sequence of bacteriophage phi X174 DNA. Nature*, 265, 687–695.

4.Sanger,F., Coulson,A.R., Friedmann,T., Air,G.M., Barrell,B.G., Brown,N.L., Fiddes,J.C., Hutchison,C.A.III, Slocombe,P.M. *et al.* (1978) *The nucleotide sequence of bacteriophage phiX174. J. Mol. Biol.*, 125, 225–246.

5. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A. et al. (2001) *The sequence of the human genome. Science*, 291, 1304–1351.

6. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M. et al. (2001) *Initial sequencing and analysis of the human genome. Nature*, 409, 860–921.
7. Sanger,F. (1949) *The terminal peptides of insulin. Biochem. J.*, 45, 563–574.
8. Lecturenoteson'SequenceAlignment'byKun-MaoChao,DepartmentofComputerScienceandInformationEngineeringNationalTaiwanUniversity(2005)
9. Goldberg,D.E.(1989).*Geneticalgorithmsinsearch,optimization&machinelearning.*Reading,MA:Addison-WesleyPublishingCompany,Inc.
10. Hernandez,D,Grass,R.,andAppel,R,(2004).MoDEL:anefficientstrategyforungappedlocalmultiplealignment.*ComputationalBiologyand Chemistry,28,119-128.*
11. Horng,J.T.,Wu,L.C.,LinCM,and Yang,B.H.(2005).Ageneticalgorithmfor multiplesequencealignment.*SoftComputing,9,407-420.*
12. Wang,C, andLefkowitz,E.J. (2005).Genomicmultiplesequencealignments:Refinementusinga geneticalgorithm.*BMCBioinformatics,6:200.*
13. Shyu,C,Sheneman,L.,andFoster,J.A.(2004).Multiplesequencealignmentwithevolutionarycomputation.*GeneticProgrammingand EvolvableMachines,5,121-14.*
14. Buscema, M.(2004). Geneticdopingalgorithm(GenD): Theoryand applications.*Expert Systems, 21(2),*63-79.
15. NotredameC,&Higgins,D.G.(1996).SAGA: Sequencealignment bygeneticalgorithm.*NucleicAcidsResearch,24,8,1515-1524.*
16. SegunAFatumo,IbidapoOAkinyemi,andEzekielFAdebiyi:AligningMultipleSequenceswithGeneticAlgorithm,InternationalJournal of Computer Theoryand Engineering,Singapore,Vol.1No.2,[186-190],2009.
17. H.CarrilloandD.Lipman.Themultiplesequencealignmentprobleminbiology",Siam J.Appl.Math.,vol.48,no.5,pp.1073-1082,October 1988.
18. K.KosmasandH.K.Donald.GeneticAlgorithmsandtheMultipleSequenceAlignmentProbleminBiology(1996).Proceedingsofthe SecondAnnualMolecular Biologyand BiotechnologyConference,February,Baton Ronge,LA.
19. S.F.J.Altschul."GapCostsforMultipleSeqenceAlignment"(1989) Theoretical Biol.,Vol 138,pp297 – 309.
20. AmoudaNizam,BuvaneswariShanmugham,KuppuswamiSubburaya,Self-OrganizingGeneticAlgorithmforMultipleSequenceAlignment,GJCST(2011),Volume11,Issue 7,7-14
21. YangChen,JingluHu,Member,IEEE,KotaroHirasawa,Member,IEEE,SongnianYu.(2008).MultipleSequenceAlignment BasedonGeneticAlgorithmswith ReserveSelectionICNSC,pp 1511-1516.