Generation of Genetic Maps Using the Travelling Salesman Problem (TSP) Algorithm

Divya Venkatesh[1], Tanmay Mishra[1], Vivekanand S Gogi[2]

Abstract— Genetic maps are the best guides available to traverse the genome of an organism. The challenge for geneticists is to generate the genetic maps from the huge amount of data, which has to be integrated in a concise and precise manner. The goal here is to obtain the best possible arrangement of genetic markers on the map and this necessitates the use of optimisation techniques. This paper elucidates the use of one such technique, Travelling Salesman problem (TSP) to generate genetic maps using recombination frequency values. Though softwares widely use this technique, not many research papers show the 'method' in detail, and this has been the motivation to write this paper. The traditional TSP algorithm yields multiple optimal solutions, whereas there can exist only one order of genes in a map and the programs need to resolve this matter as well. In this paper a possible constraint to achieve this has been explained. This paper also includes the validation of this technique with the criterion by using the crude data from already established genetic maps and mapping it back using our technique.

Index Terms— centiMorgan, genetic distance, genetic optimality criterion, genetic map, recombination frequency, marker, TSP.

1 Introduction

The global advancements in molecular markers and multiple genetic experiments being carried out, necessitate an increasing requirement of mapping their results into a single entity for a quick and comprehensive understanding of the genetic system. One of the most widely used and approved analysis tools is gene maps or ideograms. A genetic map is an ordered set of DNA markers derived from their inheritance patterns in an experimental (inbred or controlled) population [1].

Genetic maps come in handy to lower overall complexity and improve our understanding of the genome, its structure, organisation and evolutionary relationships. D.Mester et al [2] explain that these maps are "related to unidimensional ordering of many elements such as markers, clones, SNP sites, etc. With n such elements, the number of all possible orders will be n!/2, out of which only one is considered as the true order". Formation of genetic maps is based on determination of genetic distances between any two elements, using recombination data, and then placing them physically in a linear physical distance order according to the genetic distance between the markers. Genetic maps are essentially built in two steps: assigning markers into meaningful groups (analogous to chromosomes), and ordering the markers within groups to minimize the overall genetic map distance [1], [2].

To appreciate the biology in the application, we need to acquaint ourselves with the terminologies frequently used in the field:

Chromosome: The physical entity for a genetic map. It is a compact organisation of DNA into a visible rod-like structure.
DNA Marker: It is an aberration in the DNA, or a portion of the DNA which can be identified and pinpointed in the genome and can be utilised as the node for the formation of genetic maps. The markers can be of several types, the most common of them being SNP (Single Nucleotide Polymorphism), SSRs and RFLPs.
Physical distance: This is the distance (in number of base pairs) between two markers in a chromosome.
Genetic distance: It is an experimentally measurable quantity which is a function of the number of cross-overs in a recombination experiment with a fairly large sample size. It is measured in Morgans or centiMorgans (cM). Genetic distance is proportional to physical distance as the number of crossover will reduce as we reduce the physical distance between two markers.

The genetic distance, in cM can be calculated using two mapping functions, the Haldane map function and Kosambi Mapping function. For simplicity of description and calculation, we shall describe only the Haldane mapping function [3], which is as follows:

$$m = -\frac{\ln(1-2c)}{2}$$

Where, m is the genetic distance in Morgan scale and c is the recombination frequency. The cM scale can be obtained by multiplying 'm' value by 100.
Linkage: If two markers are very close to each other, the probability of cross overs between them reduces, and the markers are termed as 'linked'.[4] If the probability of a cross over between two markers is very high, then this indicates a large genetic (and hence physical) distance between the two mark-

- *Divya Venkatesh is completing her Bachelors degree program in Biotechnlogy in RV College of Engineering, India, divyadivd@gmail.com*
- *Tanmay Mishra is completing his Bachelors degree program in Biotechnlogy in RV College of Engineering, India, tanmaymishra_92@yahoo.co.in*
- *Vivekanand S Gogi is Assistant Professor in Department of Industrial Engineering and Management in RV College ofEngineering, India*

ers, thus 'unlinked'.

In the last decade, the focus on genomics has increased to a large extent and softwares based on optimaisation techniques have been developed to handle the large amount of data needed for genetic analysis. In line with this, softwares for the generation of genetic maps using experimental data have also come into light.CONCORDE and SAS JMP genomics are two such softwares that use the TSP algorithm [5] as basis.

CONCORDE is short for 'combinatorial optimization and networked combinatorial optimization research and development environment' and is based on Traveling salesman problem algorithm. It is accepted to be a highly efficient mapping program which gives high MLE (Maximum Likelihood Estimation) and low OCB (Obligate Chromosome Breaks) [6], [7].

SAS JMP genomics is a modification of the well-known JMP statistical analysis software, which provides a user friendly program for genetic analysis. It is engineered for complex genetic analysis using the optimization technique as TSP, with an array of constraints to suit the need. It does not require much programming at the user end [1], [8], [9]

In this paper the method of creation of genetic maps using TSP has been examined and demonstrated in detail. There are a number of research articles on the incorporation of this technique in the genomics field, and a lot of modifications are tried, tested and validated, to suit the needs of genetic mapping. It has been understood that the recombination rates vary a lot over populations of the same organim due to effect of genotype age and environment, and this makes the genetic mapping a very difficult task [10]. But when the genetic map is based on markers, this problem does not surface at all. Various constraints are introduced to increase the efficiency of the program to provide more reliable results and also to sift between various markers and linkage sets. To understand the changes and the constraints incorporated in the basal algorithm, a very thorough understanding of the algorithm, its procedure and working, becomes necessary.

Here, we make an effort to simplify this aspect and provide a clear understanding of the basal algorithm and the incorporation of a constraint to achieve desired maps.

TSP is an optimisation algorithm, and thus has a defined problem definition, an iterative methodology and problem specific constraints. The original algorithm was based on permutations, but with development in computer systems, matrix based algorithms have also been developed. The methodology we describe here is based on the matrix based solution, which can easily be incorporated into a mapping program, and is far more efficient and fast compared to the permutation based program.

This algorithm was designed initially to optimise the distance travelled while traversing multiple cities or 'nodes'. Since the genetic maps almost exclusively deal with placement of DNA markers on the chromosomes of the organism, characterised by the genetic and physical distances between marker-marker pairs, TSP has been used as the optimisation technique to optimise the genetic distances to yield the shortest possible arrangement of the markers in a given chromosome. This is the basis of the TSP based genetic mapping softwares. The traditional TSP algorithm has to be modified for such use in genetics appropriately. The use of Genetic Algorithms [11] and GES (Genetic Evolution Strategy) [12], [13] for the same is one of the recent and prevalent techniques.

The genetic distance between marker pairs is experimentally obtained by carrying out mating of a large number of organisms in a sample set and checking for the recombination frequency (Rf) of every marker compared to the other. A large sample set is necessary since recombination frequency is a statistical data. Once the Rf values are obtained for the marker pairs of a single chromosome, the genetic distances can be calculated as per the Haldane mapping function. This converted data functions as the raw data for TSP algorithm.

Today any genetic analysis is accompanied by a program specifically designed for the purpose. For creation of genetic maps too there are such softwares and many of these use the TSP algorithm as basis [14]. As a geneticist, in order to use these softwares to the best advantage it is necessary to understand the working mechanism of them. The paper serves this very purpose and elucidates the method of TSP algorithm and how it can be adapted to suit the biological needs

2 METHODOLOGY

2.1 Problem Definition, Solution and Iteration

The basic data needed to formulate the genetic mapping problem as an assignment problem is the genetic distance between the markers chosen. All the markers on one particular chromosome are chosen at once. Among these, recombination is tested by test crosses in wet lab. Then the recombination frequency between the gene markers in pairs of two is calculated and recorded.

These Rf values fill the matrix, which can be solved by the TSP method. Here the symmetric TSP model has been used, with number of rows and columns equal to the number of markers present.

The aim is to optimise the chromosome's genetic distance, which is essentially the total Rf between the first and last marker in the so determined order of the markers by TSP method.

Extending this method to all chromosomes of the genome of any organism the entire genetic map of the organism can be generated using optimization techniques.

Algorithm: Selective TSP [15]

1. Ensure there is atleast one zero in every row of the matrix. To do so, chose the smallest element of each row and subtract this value from all other values of the matrix.
2. Repeat the same procedure and ensure every column has atleast one zero.
3. For every zero calculate the penalty. The penalty is equal to the sum of the smallest element in the row of the zero and the smallest element in the column of the zero, excluding this ze-

ro.
4. Now choose the zero with the maximum penalty and cross out its row and column.
5. The row marker to column marker of this strike will be one particular path.
6. Now delete the struck out row and column. Repeat all the steps from 1, iteratively till you end up with a 1x 1 matrix. Note down the paths every time you strike. The last path will be from the row to column of the 1x 1 matrix obtained.
7. Now collect all the marker pair paths obtained.
8. Beginning the path from the reference marker (mentioned as part of data), arrange all other paths such that the path finally returns back to the reference marker itself.
9. The path so obtained gives the optimal path with least total genetic distance between the reference marker and the last marker.
10. This path has to be modified to suit our genetic map, as explained in the solution and finally the order of markers is determined.
11. The Rf values are then converted to cM values and the genetic map is constructed in the order of markers as determined by the TSP method.
12. If at step 4 more than one zero exists with the same maximum penalty, then multiple optimal solutions arise, which diverge at the step where the multiple max penalty zeroes arise. Here we need all the optimal paths for the TSP, so for every zero with max penalty all steps after 4 are repeated individually and all optimal paths are collected.

One of the major confusions that arise while using TSP to decide the order of markers is when multiple optimal paths exist. According to TSP as long as the distance is minimised the path in itself referring to the order loses importance. But since our goal here is genetic optimality, marker order takes prime importance. So the distance we are aiming to optimise excludes the last distance which indicates travel from last marker to the first. Optimising this subtracted path (as we term it) would in turn optimise the order primarily as opposed to concentrating only on the total distance like in general TSP.

We need to use what we term as; Genetic optimality criterion as opposed to the traditional TSP criterion is due the chromosome being able to have only one particular pattern of markers in reality whereas general TSP offers multiple possible solutions in many cases. So the best order is to be arrived at for the genetic map to accurately represent the real chromosome.

The paper confirms this technique of using an alternative optimality criterion by obtaining data for which genetic maps have already been constructed by various methods, and applying TSP with our criterion to validate that the results so obtained match the already proved ones.
13. Sample DATA: The Table 1 represents two point crosses of markers on one chromosome in an organism [4]. Here, Y is the reference marker. Keeping in mind that Rf of Y-W is same as W-Y, and so on for all marker combination the following matrix is filled.

Note that the Rf values of Y-Y, V-V, R-R, W-W, and M-M in Table 2 are all ∞ as those are forbidden paths, since a marker can't follow itself.

This matrix is now in the form of a TSP and hence can be solved using optimization techniques.

TABLE 1
Sample Data

Marker pair	Rf %
Y-W	0.011
Y-V	0.33
Y-M	0.343
Y-R	0.429
W-V	0.321
W-M	0.328
W-R	0.421
V-M	0.04
V-R	0.241
M-R	0.178

TABLE 2
Initial working matrix

	Y	V	R	W	M
Y	∞	33	42.9	1.1	34.3
V	33	∞	24.1	32.1	4
R	42.9	24.1	∞	42.1	17.8
W	1.1	32.1	42.1	∞	32.8
M	34.3	4	17.8	32.8	∞

2.2 Data Collection and Computer-based Validation

To evaluate whether our methodology works in all cases, this technique is verified using a large amount of data. Since the raw data of the Rf values is largely a product of wet lab analysis, the best available source was the already established gene maps shown in Fig. 1, Fig. 2, Fig.3 and Fig. 4. Backtracking from these gene maps to obtain the data in the form needed and later using our technique, it was verified that the order of markers so obtained, matched that of the existing gene map selected.

Two programs were used to process the large amount of data:

1. Data program

Algorithm: Take the number of markers and their centiMorgan values as input. Using the centiMorgan distances calculate the distance matrix for all the markers (subtract one distance from the other). Convert the centiMorgan matrix values to Rf values using the formula mentioned in the introduction. The matrix of Rf values is the data we need. This has to be done for each chromosome's gene map

2. Permutation program

This program doesn't use the algorithm that has been discussed above, but a more basic approach of finding all possible paths given a set of markers, using permutation.
Algorithm: Take the matrix obtained above (for every chromo-

some individually) as input. Also take the reference marker for all the paths to begin at. Use permutations to obtain all possible paths starting with the reference marker, and involving all other markers. Select those paths with least total distance (multiple ones if more than one has the same least total). Amongst them, the order for the path with least subtracted distance will be reported as genetically optimal order.

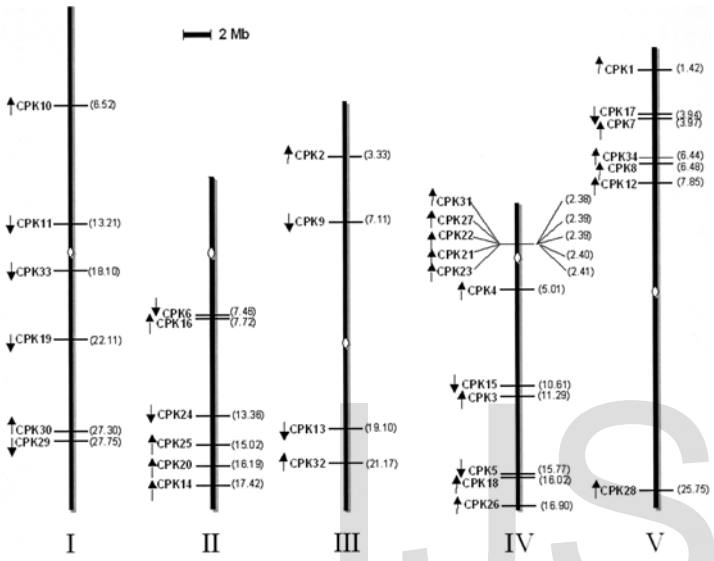

Fig. 1. Ideogram of Arabidopsis [16]

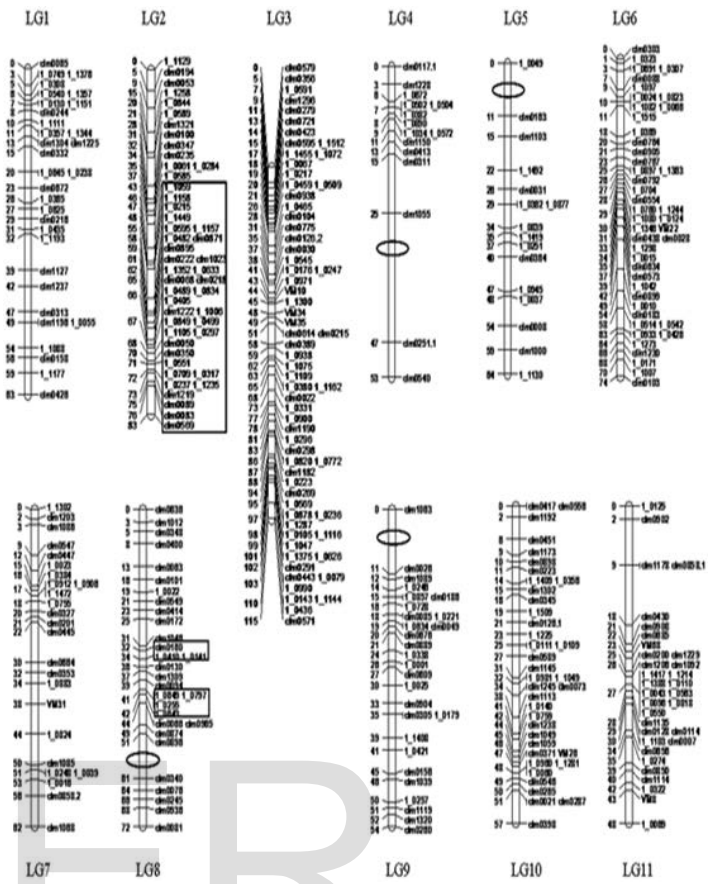

Fig. 2. Ideogram of Asparagus [17]

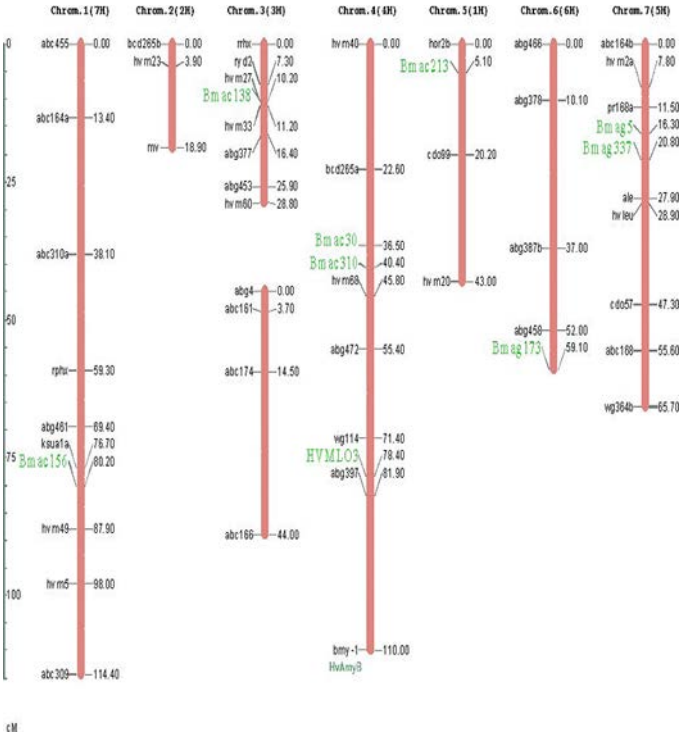

Fig. 3. Ideogram of Barley [18]

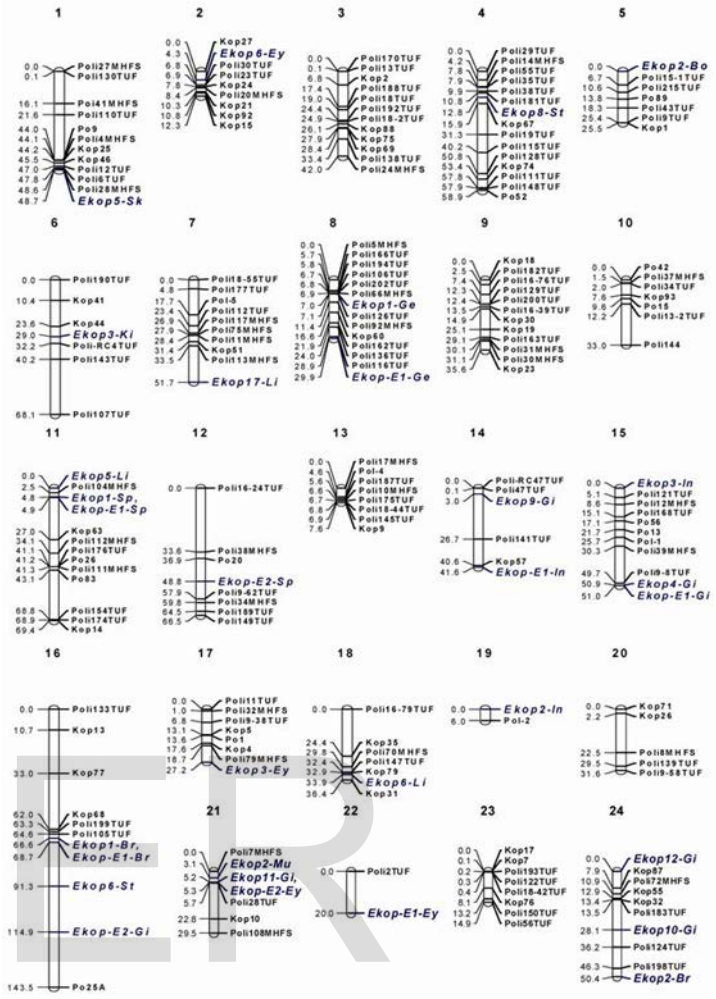

Fig. 4. Ideogram of Olive Flounder [19]

3 RESULTS

Using travelling salesman method the order with least genetic distance is calculated for the sample data, and multiple optimal paths are obtained: Y->W->V->M->R->Y and Y->R->M->V->W->Y. Both paths have the same reference point. And both will show the same optimal overall distance from Y to Y (97.9) as shown in Table 3 and 4.

Now to decide which amongst these two give the desired order we are going to use the genetic optimality condition.

Here though the total distance remains the same, 97.9, between both the paths, since the subtracted distance of 55 is our optimal genetic distance, the order we desire is Path 1: Y->W->V->M->R**.** This is the *genetically optimal path.*

In a genetic map the distances are depicted using cM as discussed before. So we convert the RFs into cM values as per the Haldane mapping function formula discussed earlier. The converted values are shown in Table 5.

The genetic map can now be depicted as in Fig. 5, where

the letters on the right side indicate the names of the genetic markers and the values on the left side indicate the cM values of each marker from the reference marker(Y). The thick line in the centre represents the chromosome on which these marker's loci are present.

The genetic distances are not additive [20], so distances of all markers are calculated between one another and not from the reference marker. The latter may fail to match the former in cases where there is underestimation or over estimation of the degree of recombination (due to double and triple crossing over). For example to represent marker V, Y-W is 1.1 cM and W-V is 51cM, so Y-V is now calculated as 52.1cM, but using the data, Y-V is given as 54cM (an overestimation).

TABLE 3
First path Rf calculation

TABLE 4
Second path Rf calculation

TABLE 5
cM to Rf conversions

Marker pair	Rf	Morgan	cM
Y-W	0.011	0.011	1.1
Y-V	0.33	0.54	54
Y-M	0.343	0.579	57.9
Y-R	0.429	0.97	97
W-V	0.321	0.51	51
W-M	0.328	0.53	53
W-R	0.421	0.92	92
V-M	0.04	0.04	4
V-R	0.241	0.33	33
M-R	0.178	0.22	22

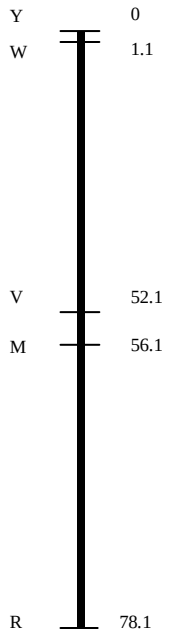

Fig. 5. Final genetic map from data processing using TSP, showing the arrangement of the markers according to their genetic distances.

As mentioned under data collection we wanted to verify the veracity of our contraints for the standard aforementioned organisms' idiograms. For the same we needed multiple optimal solutions (distance wise) upon which when our constraint is used, the right result would be arrived at.

Since there were no programs based on the algorithm we elucidated in the paper, which yield multiple optimal

PATH 1	RF VALUE
Y->W	1.1
W->V	32.1
V->M	4
M->R	17.8
Subtracted distance	55
Y->W->V->M->R: distance = 55 Y->W->V->M->R->Y: distance = 97.9	

PATH 2	RF VALUE
Y->R	42.9
R->M	17.8
M->V	4
V->W	32.1
TOTAL	96.8
Y->R->M->V->W:distance = 96.8 Y->W->V->M->R>Y:distance = 97.9	

, we had to instead develop a program that was based on permutations that yield the same multiple optimal results as with our algorithm, although with much more processing involved. (The algorithms are previously mentioned).

When our genetic optimality criterion was used on such multiple optimal results for established data, we were able to obtain the desired single solution. We were thereby successful in reproducing the maps using our constraint.

The drawback with this approach is, when trying to run data for more than 14 markers, it takes huge amount of time to process the many permutations and render the results. This has limited our ability to check our technique for bigger maps with larger amount of data. But since the main aim was to establish the veracity of our genetic optimality criterion, this limitation has not hampered our verification. This thereby validates the working, efficacy and necessity of the genetic optimality criterion.

This also further validates the merit of using this algorithm

instead of permutations to arrive at the optimal path for the TSP, so as to process large amount of data with accuracy and speed.

4 Conclusions and Future Prospects

Today any genetic analysis is accompanied by a software program specifically designed for the purpose. For creation of genetic maps too there are such softwares and many of these use the TSP algorithm as basis.

The softwares consider recombinant frequencies for solving the TSP, and only later convert them to cM values for map depiction. The reason behind this was explored in this study and proved that much more accurate results are obtained using this method when compared to conversion first to cM values, followed by solving the TSP using these converted values.

The study also elaborates on the use of genetic optimality criterion to zero in on the best solution out of the multiple optimal ones generated. This criterion can be one of the ways that the softwares arrive at the single optimal solution as the traditional TSP method yield multiple optimal solutions. We propose that the traditional TSP needs to be modified to suit for the applications in genetics. This constraint is one of the ways to modify it and the method was also verified using standard idiograms as reference.

Though in most cases TSP is completely successful, since it's a mathematical approach at best there may be more than a single final path at times, even with use of constraints and interpretation of the result could be erroneous. Each program would then use its own method to choose one solution in such a case, and we are yet to explore these methods. Also on a larger scale, repetition of RF values need to be taken into account and more data will be required to segregate and map such markers, thereby indicating more modifications that maybe required to the TSP method.

This paper gives a clear overview of the working principle of the softwares based on TSP, used to create genetic maps. The softwares would use complex programs based on the algorithm explained here or derivatives of the same. The program would also incorporate multiple constraints as mentioned above to finally order the markers such that they map the natural organisation of chromosomes.

5 Acknowledgment

The Author thanks the valuable insights provided by Dr K.N. Subramanya, Dr H.G Ashok Kumar and Dr Nagashree N Rao. They have been instrumental in shaping our progress to a better conclusion.

Efforts by L.P. Ashwanth, Agneev Ghosh and Sathvik M Ashok, also should be acknowledged for their assistance in the development of the software programs for data validation.

References

[1] Miclaus, Kelci, Rob Pratt, and Matthew Galati. "The Traveling Salesman Traverses the Genome: Using SAS® Optimization in JMP® Genomics to build Genetic Maps.", SAS global forum, 2012

[2] Mester, David I., Yefim I. Ronin, Eviatar Nevo, and Abraham B. Korol. " Constructing Large-Scale Genetic Maps Using an Evolutionary Strategy Algorithm," Genetics 165, 2003, pp: 2269–2282

[3] Haldane, J.B.S. "The combination of linkage values, and the calculation of distance between linked factors" . J. Genet., Vol 8, 1919, pp: 299-309.

[4] Robert Brooker, Ch. chapter 6 : Linkage and Genetic Mapping in Eukaryotes Genetics: Analysis and Principles, 2/e, University of Minnesota (Book)

[5] Lin, S., and B. Kernighan, "An effective heuristic algorithm for the TSP." Oper. Res. Vol. 21, 1973, pp: 498–516.

[6] C. Hitte, T. D. Lorentzen, R. Guyon, L. Kim, E. Cadieu, H. G. Parker, P. Quignon, J. K. Lowe, B. Gelfenbeyn, C. Andre, E. A. Ostrander, And F. Galibert, "Comparison of MultiMap and TSP/CONCORDE for constructing radiation hybrid maps," Journal of Heredity, Vol 94(1), 2003, pp: 9-13.

[7] Richa Agarwala, David L. Applegate, Donna Maglott, Gregory D. Schuler, and Alejandro A. Scha¨ffer, "A Fast and Scalable Radiation Hybrid Map Construction and Integration Strategy," Genome Res. Vol.10, 2000, pp: 350-364

[8] Oded Maimon, Lior Rokach "Data Mining And Knowledge Discovery Handbook", Springer Science and Business Media, 2010 (Book)

[9] Zhiwu Zhang, Edward S. Buckler, Terry M.Casstevens and Peter J. Bradbury, "Software engineering the mixed model for genome-wide association studies on large samples," Briefings In Bioinformatics. Vol.10. no.6, 2009, pp: 664-675

[10] Mester, D.I., Ronin, Y.I., Korostishevsky, M.A., Pikus, V. L., Glazman, A.E. and Korol, A.B. , "Multilocus consensus genetic maps (MCGM): Formulation, algorithms, and results," Computational Biology and Chemistry, Vol. 30(1), 2006, pp: 12-20.

[11] Zakir H. Ahmed, "Genetic Algorithm for the Traveling Salesman Problem using Sequential Constructive Crossover Operator", International Journal of Biometrics & Bioinformatics (IJBB) Vol.3(6), 2003, pp. 96-105.

[12] Mester, D. and Braysy, O., "Active guided evolution strategies for large-scale vehicle routing problems with time windows," Computers & Operation Research, Vol. 32(6), 2005, pp: 1593-1614.

[13] Yefim Ronin, David Mester, Dina Minkov, Abraham Korol, "Building reliable genetic maps: different mapping strategies may result in different maps", Natural Science Vol.2, No.6, 2010, pp: 576-589

[14] Y. Ronin, D. Mester, D. Minkov, R. Belotserkovski, B. N. Jackson, P. S. Schnable, S. Aluru and Korol, "Two-Phase Analysis in Consensus Genetic Mapping", g3 Journal, vol. 2, no. 5, 2012, pp.537-549

[15] J.K. Sharma(2011), Ch. Assignment Problem, Operations Research, theory and applications, 4t edn, Macmillan, India, pp 313-342.(Book)

[16] Shu-Hua Cheng, Matthew R. Willmann, Huei-Chi Chen and Jen Sheen ", Calcium Signalling through Protein Kinases. The Arabidopsis Calcium-Dependent Protein Kinase Gene Family", Plant physiol. Pubmed, 2002, pp: 469-85.

[17] Pei Xu, Xiaohua Wu, Baogen Wang, Yonghua Liu, Jeffery D. Ehlers, Timothy J. Close, Philip A. Roberts, Ndeye-Ndack Diop,Dehui Qin, Tingting Hu,Zhongfu Lu, Guojing Li, (2011), "A SNP and SSR Based

Genetic Map of Asparagus Bean (Vigna. unguiculata ssp. sesquipedialis) and Comparison with the Broader Species", PloS one 6, no. 1, 2011, e15952.

[18] Ariel Castro, Patrick M. Hayes, Tanya Fillichkin, and Carlos Rossi, " Update of barley stripe rust resistance QTL in the Calicuchima-sib x Bowman mapping population.", Barley Genetics newsletter, Vol. 32, 2002, pp: 1-12

[19] Jung-Ha Kang, Woo-Jin Kim, Woo-Jai Lee,, "Genetic Linkage Map of Olive Flounder, Paralichthys olivaceus", International Journal of Biological Sciences, 2008, pp: 143-149.

[20] A. Touré, B.I.G. Haussmann, N. Jones, H. Thomas , and H. Ougham, "Construction of a genetic map, mapping of major genes, and QTL analysis", HaussmannManual, Sorghum Millet, Cornell University. 2007.