A Weighted and Deterministic Annealing Solution for Interpolative Dimension Reduction

Authors Name/s per 1st Affiliation *(Author)*

line 1 (of *Affiliation*): dept. name of organization

line 2: name of organization, acronyms acceptable

line 3: City, Country

line 4: e-mail address if desired

Authors Name/s per 2nd Affiliation *(Author)*

line 1 (of *Affiliation*): dept. name of organization

line 2: name of organization, acronyms acceptable

line 3: City, Country

line 4: e-mail address if desired

*Abstract*— Advances in modern bio-sequencing techniques have led to a proliferation of raw genomic data that led to an unprecedented opportunity for data mining. To analyze such large volume and high-dimensional scientific data, we apply high performance dimension reduction and clustering algorithms. Among the known algorithms, we used Multidimensional Scaling (MDS) to reduce the dimension of original data and Pairwise Clustering, to classify the data. With a tradeoff of precision, an interpolative technique is applied on such massive data. It reduced memory complexity to linear and reduced the overall processing time cost. However, pairwise alignments may produce aligned sequence of very low quality. Many existing methods only process data where all distances are presented. In this paper, we proposed a robust MDS and interpolation algorithm using Deterministic Annealing technique, to solve the missing distance problem by adding weight function. We compared our method to three state-of-art techniques. By experimenting on three common types of sequence dataset, the results illustrate that the precision of our algorithms are better than other algorithms, and the weighted solutions has a lower time cost as well.

Keywords—Deterministic Annealing; Multidimensional Scaling

# Introduction

The speed of data collections by modern instruments in every scientific and technical field is accelerating rapidly with the advancement of science technologies. As this massive amount of data can easily overwhelm single computer, traditional data analysis technologies needs to be modified and upgrade to adapt to high performance computational environment for acceptable performance. Many data mining and machine learning algorithms have been developed to solve these big data problems. Among them, dimension reduction has been proved to be useful in data clustering and visualization field [1].

In our previous research, we proposed DACIDR, a pipeline which can generate robust clustering and visualization result on millions of sequences. It uses dimension reduction to enable scientists browse the sequence clustering result in a 3D space, which proved to be useful in many cases.

In DACIDR, we mainly use MDS since it is a reliable method and can be parallelized more efficiently compared to some existing dimension reduction methods. Among different techniques in MDS, such as Classical MDS, Newton Raphson method and Chi Square method, we chose to use Scaling by Majorizing a Complicate Function (SMACOF). This algorithm has been proved to run under quadratic time complexity, which is much faster than other methods. However, it also uses quadratic memory while doing the computation, which means with large dataset, this algorithm can easily overflow the memory on a single machine or even high performance computers. Therefore, an interpolation algorithm called Majorizing Iterative MDS has been proposed to solve the memory problem. By using a pre-configured mapping from sample data, this algorithm successfully eliminates the memory requirement for massive dataset. Furthermore, we proposed an hierarchical method in DACIDR to speedup to process and reduced the time complexity to sub-quadratic.

However, these methods can only generate robust result while input data has no missing values. In the case where the input data are sequences, the pairwise sequence alignment can easily generate sequence alignment with very low qualities which could cause inaccuracy in final result. Therefore, we proposed a robust solution for input data with missing values by adding a weight function to both SMACOF and interpolation. Furthermore, to improve the accuracy of interpolation, a deterministic annealing function has been add to MI-MDS.

The structure of the paper is organized as following: Section 2 is related work and explains the background, Section 3 introduces and explains Weighted DA SMACOF (WDA-SMACOF) and made comparison to other similar methods; Section 4 introduces the weighted DA-MI-MDS and shows the advantage of this scalable robust algorithm; In Section 5 we present the conclusion and future work.

# Background and Related work

Existing dimension reduction algorithms, to name a few, such as Principal Component Analysis (PCA), Generative Topographic Mapping (GTM) [3], and Self-Organizing Maps (SOM) [4], were focused on using the vector information in original dimension to construct a configuration in lower dimension space. However, in our case, we use data generated from sequencing technology, which is difficult to be considered as vectors since original sequences have various lengths. One possible way to do that is through Multiple Sequence Alignment (MSA). However, it is well known that MSA only works on small amount of data, and we are processing large-scale data. In contrast, MDS focused on using pairwise dissimilarities. Therefore, it is more suitable to use it for sequence clustering. Among the various MDS techniques, SMACOF has been proved to use only quadratic time complexity compared to other methods. Simulated Annealing, genetic algorithms has also been used to optimize the result from SMACOF. However, due to their Monte Carlo type of sampling, the time cost of their methods is much higher than the original method. The DA-SMACOF has eliminated the local optima problem as well, and compared to the previous methods, it is much faster. However, the main drawback of DA-SMACOF is that it can only deal with pairwise dissimilarities with no missing values. In our data, the missing values could be as much as to 90%. Therefore, it is critical for us to add weighted function to DA-SMACOF in order to process these data. However, this function only works with moderate amount of data processing because MDS requires quadratic memory usage.

To address this issue, many algorithms has been developed to extend the capability of various dimension reduction algorithms by embedding new points with respect to previously configured points, or known as out-of-sample problem. A generalized out-of-sample solution has been provided by S. Xiang by using coordinate propagation for non-linear multidimensional scaling. An out-of-sample extension for the algorithms based on the latent variable model has been provided by M. Carreira-Perpi˜nÅLasn and Z. Lu. This problem could also be considered as unfolding problem in MDS since only pairwise distance between in-sample sequences and out-of-sample sequences are observed. [2] and [20] proposed out-of-sample extension for the classical multidimensional scaling (CMDS). In contrast to that, Seung-Hee Bae proposed an EM-like optimization solution, called MI-MDS to solve the problem with STRESS criteria in Equation (1), which finds embeddings of approximating to the distance rather than the inner product as in CMDS. In addition to that, Yang Ruan proposed a heuristic method, called HE-MI, to lower the time cost of MI-MDS. An oct-tree structure called Sample Sequence Partition Tree is used in HE-MI to partition the in-sample space in the lower dimension space, and then interpolated the out-of-sample data hierarchically to avoid additional time cost. HE-MI enables DACIDR to process millions of sequences within sub-quadratic time complexity. As the proposed MI-MDS and HE-MI are also EM-like optimization algorithms, they could suffer from local optima problem as SMACOF. So in this paper, we proposed a DA solution to solve this issue, called DA-MI-MDS. And we also added weight function support for our new algorithm.

# weighted solution for da-SMACOF

In this section, we proposed a weighted solution for DA-SMACOF, a DA and weighted solution for MI-MDS. MDS and DA will be briefly discussed first, followed by WDA-SMACOF and WDA-MI-MDS.

## Multidimensional Scaling

Multidimensional Scaling is a set of statistic techniques used in dimension reduction. It is a general term for these techniques to apply on original high dimensional data and reduce their dimensions to target dimension space while preserving the correlations, which is usually Euclidean distance calculated from the original dimension space from the dataset, between each pair of data points as much as possible. It is a non-linear optimization problem in terms of reducing the difference between the mapping of original dimension space and target dimension space. After bioinformatics data visualization, each sequence in the original dataset will be a point in both original and target dimension space. The dissimilarity between each pair of sequences is considered as original distance used in MDS.

In formal definition, given a data set of *N* points in original space, a pairwise distance matrix *Δ* can be given from these data points (*Δ* = [*δij*]) where *N* is the number of data points and *δij* is the dissimilarity between point *i* and point *j* in original dimension space which follows the rules: (1) Similarity: . (2) Positivity: . (3) Zero Diagnosal: . Given a target dimension *L*, the mapping of points in target dimension can be given by an *N* \* *L* matrix *X*, where each point xi from original space is represented as ith row in *X*.

The well-known objective function, so called STRESS value, which evaluates the precision of MDS result, is given in equation 1:

(1)

where wij denotes the possible weight from each pair of points, dij denotes the Euclidean distance between point i and j in target dimension. Scaling by Majorizing a Complicated Object Function (SMACOF) is an algorithm designed to reduce the stress value in (1), which will be explained in detail in sub section C.

By adding a computational temperature to the SMACOF function, Deterministic Annealing SMACOF (DA-SMACOF) has been proved to be reliable, fast and always generate robust result, so we choose this method to generate the L-dimension mapping of our sample data set.

## Deterministic Annealing

Deterministic Annealing (DA) was an annealing process used to find global optima of an optimization process instead of local optima. A computational temperature has been added to the target object function. By lowering the temperature during the annealing process, the problem space gradually revealed to the original object function. Different from Simulated Annealing, which is based on Metropolis algorithm for atomic simulations, it does not rely on the random sampling process and random decisions based on current state. DA uses an effective energy function, which is derived through expectation and is deterministically optimized at successively reduced temperatures.

In DA-SMACOF, the stress function in (1) is used as object function. We denote the as the cost function for SMACOF, and as a simple Gaussian distribution:

(2)

(3)

where is the average of simple Gaussian distribution of ith point in target dimension *L*. Also, the probability distribution and free energy are defined as following:

(4)

(5)

where X is the N \* L matrix, where each line is a vector representing a point in the target dimension, and T is the computational temperature used in DA.

## Weighted DA-SMACOF

The goal of DA in SMACOF is to minimize with respect to parameters is independent of so the problem can be simplified to minimize if we ignore the terms independent of . By differiating (5), we can get

(6)

where is the ith point in the target dimension L.

Take (6) into (2), finally the became

(7)

(8)

As the original cost function and target dimension configuration gradually changed when the computational temperature changes, we denote as the target dimensional configuration and as the dissimilarities of each pair of sequences under temperature T. So the updated stress function of DA-SMACOF becomes

(9)

where is defined as

(10)

Note that if the distance between point I and point j is missing from, then. There won't any difference between and since both of the distances are considered missing values. This was not proposed in the original DA-SMACOF where all weights for all distances in are set to 1.

By expanding (9), updated stress value can be defined as

(11)

Equation (11) has three parts, first part only dependent on fixed weights and temperature, so it is constant. Then to obtain the majorization algorithm for and , they can be defined as following:

(12)

= (13)

where V is defined as that is simply a matrix with And is defined as following:

(14)

Finally, to find the Majorizing function for (9), we apply (12) and (13) to (11), By using *Cauchy-Schwarz* inequality, the majorization inequality for the stress function can be obtained in (16).

(15)

(16)

By setting the derivatives of to zero, we finally get the formula of the WDA-SMAOCF,

(17)

where is the pseudo-inverse of V, and is the estimated from previous iteration. This formula is also called Guttman transform by De Leeuw and Heiser. (More details about how the formula evolves can be found in Modern Multidimensional Scaling).

Since T represents the portion of entropy to the free energy . The DA starts the temperature T in the high value, and lower it as times goes on, like physical annealing process. In each temperature, the is updated and an iterative majorizing algorithm is used to calculate the best mapping for X in current temperature by (17). The initial temperature is critical in WDA-SMACOF that a flat initial configuration (all distance in equals to zero) needs to be avoided. So the is calculated based on maximum value of weight times distance.

As un-weighted DA-SMACOF consider all weights equals to 1, so observed from (10) and (14), it will take longer time to process the data without avoiding missing distance values. Note that in original SMACOF, the weight function has been proposed. But it could not guarantee to find global optima, so by adding weighted function to DA-SMACOF, we finally get a robust algorithm which can give accurate result on *Δ* with missing values.

## Parallezation of WDA-SMACOF

As WDA-SMAOCF is an iterative optimization algorithm, we use an iterative MapReduce runtime, called Twister, to parallelize it for maximum performance. Note that different from DA-SMACOF, a weighted matrix W which contains all the weights for every distance needs to be included during the computation, so the memory usage of WDA-SMAOCF will be higher compared to DA-SMAOCF. However, since W is also an N \* N matrix, both of the algorithms have a memory (space) complexity of *O(N2)*.

The parallelized WDA-SMACOF uses three MapReduce computations in one iteration. Two of them are matrix multiplication and one of them is stress calculation. In detail, one MapReduce job calculate and another one calculate times the product matrix of previous job. The reducer is used to combined the updated X from each mapper, as shown in Figure 1.

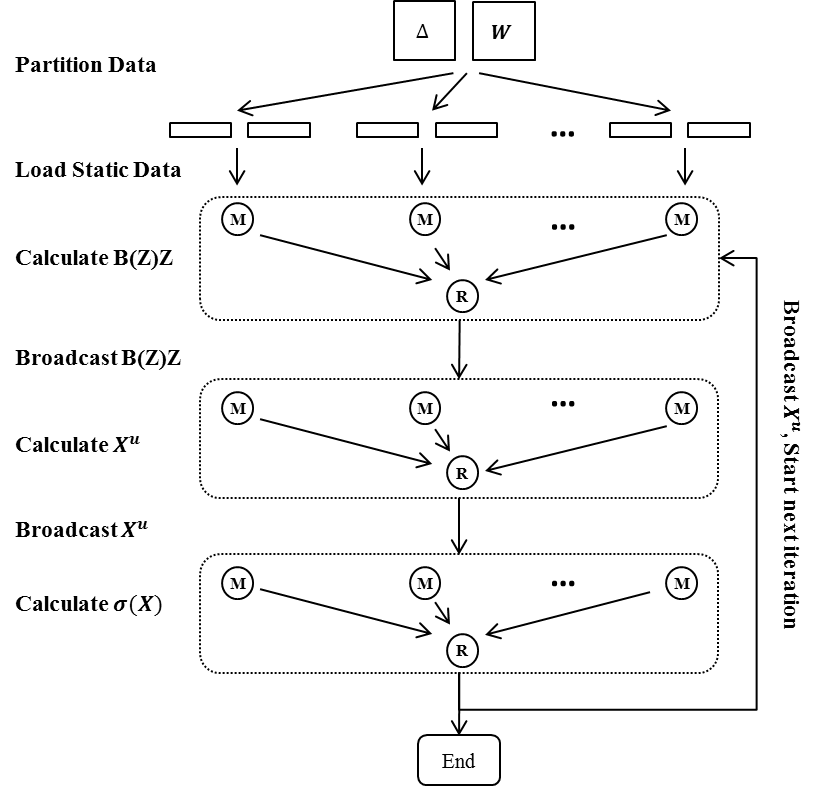


Figure 1 The flowchart of parallel WDA-SMACOF on an iterative MapReduce runtime

From Figure 1 we see that Z is broadcast to every mapper 3 times in one iteration. There are two reasons for this design: First, every point in the target dimension space is affected by every other point in it. So it is a presquites for each mapper to be updated to the same Z or same B(Z) in every iteration in order to get accurate result; Second, compared of sending updated B(Z), which is an N \* N matrix, Z is an N \* L matrix, where L is often set to two or three for visualization purpose. It is a much less communication cost than sending an N \*N matrix.

In original parallelized SMACOF or DA-SMACOF, the matrix can be partitioned evenly since all weights equals to one. Therefore the computations on different mappers are evenly distributed according to the size of partitioned. However, to ensure the load balance of parallelized WDA-SMACOF, different from previous partition strategy, and W are partitioned by the number of presented distance values, but not the size of matrix.

# Weighted DA solution for MI-MDS

In this section, we proposed a weighted deterministic solution for MI-MDS. First, we briefly discuss out-of-sample problem and MI-MDS, then we will describe Weighted DA-MI-MDS (WDA-MI-MDS) in detail.

## Out-of-Sample Problem and MI-MDS

The in-sample and out-of-sample problem was brought up in data clustering and visualization to solve the large-scale data problem. In DACIDR, WDA-MDS is used to solve the in-sample problem, where a relatively smaller size of data is selected to construct a low dimension configuration space. And remaining out-of-sample data can be interpolated to this space without the usage of extra memory. This particular problem can also be considered as unfolding model in MDS, where only dissimilarities between out-of-sample points and in-sample points are observed.

In formal definition, suppose we have a dataset contains size of N in-sample data, denoted as , and size of M out-of-sample points, denoted as , where in-sample points were already mapped into an L-dimension space, and the out-of-sample data needs to be interpolated to an L-dimension space, defined as , where and . Note that only one point at a time is interpolated to the sample space. So the problem can be simplified to interpolate a point to L-dimension with the distance observed to in-sample points. The stress function for is given by

(18)

where denotes the possible weight, is the distance from to in-sample point i in target dimension, and is the original dissimilarity between and point i. If all weights equals to 1, equation (18) is transformed to

(19)

MI-MDS is an iterative majorization algorithm proposed by Seung-Hee Bae to minimize the stress value in (19), where all weights are assumed to be 1. It will find k nearest neighbors from in-sample points of a given out-of-sample point at first, denoted as . Then by using a majorizing function , where its minimum can be obtained analytically by setting the its derivative to zero.

## Weighted DA solution for Majorizing Interpolation MDS

MI-MDS has been proved to be efficient when deal with large-scale data. However, there are two disadvantages with this method. First, it assumes that all weights equal to one, where it didn't state that how to deal with missing values and different weights. Secondly, this method is an EM-like optimization algorithm, which could be trapped in local optima as the EM-SMACOF.

Therefore, we proposed WDA-MI-MDS to solve these issues. To solve the weighted out-of-sample problem, we need to find an optimization function for (18). By expanding (18), we have

(19)

(20)

where is a constant irrelevant to X, so same as in SMACOF, only and needs to be considered to obtain the majorization function. can be deployed to

(21)

(22)

where where L is the target dimension. The Cauchy-Schwarz inequality can be applied on in to establish the majorization function, which is given as

**Algorithm 1 WDA-MI-MDS algorithm**

Input: , , , and

Generate random initial mapping .

for each in do

Compute and

Compute

;

while ≥ do

Update to using (10)

Initialize random mapping for , called

while

Update using (30)

end while

Cool down computational temperature

end while

end for

return

(23)

(24)

where z is a vector of length L which contains , and . By applying (24) to , we will have

(25)

(26)

where C is a constant irrelevant from . After applying (22) and (26) to (20), we will have

(27)

As both and are constants, equation (27) is a majorization function of the stress that is quadratic in X. The minimum of it can be obtained by setting the derivatives of to zero, that is

(28)

(29)

where z is the previous estimated x. Although this algorithm so far can guarantee to generate a serious of non-increasing stress value for various weight from original data, it still could be trapped into local optima. Therefore, to add a deterministic annealing solution into that, we apply (10) to (29), and finally we have the iterative majorization equation for WDA-MI-MDS in (30), and the algorithm is illustrated in algorithm 1.

(30)

where can be obtained using (10).

## Parallelization of WDA-MI-MDS

Different from WDA-SMACOF, one out-of-sample point's dimension reduction result only depends on the in-sample points. In another word, every out-of-sample point is independent from any other out-of-sample points. Therefore, WDA-MI-MDS can be pleasingly paralleled, which is illustrated in Figure 2.

# Experiments

The experiments were carried out on FutureGrid XRay Cluster, which has 168 CPUs and 1324 cores. We used 3 different bioinformatics dataset, include Metagenomics DNA, hmp16SrRNA and COG proteins, to test Weighted Deterministic Annealing (WDA), which is proposed in this paper, Un-weighted Deterministic Annealing (UDA), Weighted Expectation Maximization (WEM), and Un-weighted Expectation Maximization (UEM) solutions of SMACOF and interpolation.

## Metagenomics DNA Data

In this experiment, we used a metagenomics DNA dataset, which includes 4640 unique DNA sequences, with an average length of 300. 2000 of these sequences were selected as in-sample data, and rest 2640 sequences were considered as out-of-sample. As this dataset is relatively small, we tested the sequential version of both SMACOF and interpolation. A normalized stress value is calculated by

(31)

where is given by PID distance from pairwise sequence alignment using Smith-Waterman algorithm. As (31) gives the differences between the mapped distance after dimension reduction and original distance, so it can be considered as an error value, where lower Normalized Stress means better performance. 10.775% of the original distances values were considered as missing while calculating the original distance because of the low alignment quality.

The result is shown in Figure 3, both of the weighted solutions outperforms the un-weighted solution. And WDA is clearly more robust from the result. And DA took a longer time to compute because of its robustness for finding the global optima. Note that in SMACOF, the distances are calculated beforehand so the computation time cost is lower. In contrast, interpolation's computation time is mainly dominated by original distance calculation (the sequence alignment) because the distances are calculated as needed while the out-of-samples are interpolated, where it took 1455.889 seconds to compute 2640 \* 2000 sequence alignments on a single core. So the time cost difference between DA and EM solutions for interpolation is small.

## hmp16SrRNA Data

Blah, Blah, Blah

## COG Protein Data

Blah, Blah, Blah

# Conclusion and Future work

Blah, Blah, Blah

##### Acknowledgment *(Heading 5)*

Blah, Blah, Blah

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.