Optimization Techniques to Deeply Mine the Transcriptomic Profile of the Sub-Genomes in Hybrid Fish Lineage

Zhong Wan1*, Jiayi Tang1, Li Ren2, Yamei Xiao2 and Shaojun Liu2*

1 School of Mathematics and Statistics, Central South University, Changsha, China, 2 State Key Laboratory of Developmental Biology of Freshwater Fish, Hunan Normal University, Changsha, China

It has been shown that reciprocal cross alldiploid lineage with sub-genomes derived from the cross of Megalobrama amblycephala (BSB) × Culter alburnus (TC) generates the variations in phenotypes and genotypes, but it is still a challenge to deeply mine biological information in the transcriptomic profile of this lineage owing to its genomic complexity and lack of efficient data mining methods. In this paper, we establish an optimization model by non-negative matrix factorization approach for deeply mining the transcriptomic profile of the sub-genomes in hybrid fish lineage. A new so-called spectral conjugate gradient algorithm is developed to solve a sequence of large-scale subproblems such that the original complicated model can be efficiently solved. It is shown that the proposed method can provide a satisfactory result of taxonomy for the hybrid fish lineage such that their genetic characteristics are revealed, even for the samples with larger detection errors. Particularly, highly expressed shared genes are found for each class of the fish. The hybrid progeny of TC and BSB displays significant hybrid characteristics. The third generation of TC-BSB hybrid progeny (BT3F3 and TB3F3) shows larger trait separation.

Keywords: transcriptomic profile, distant hybridization, optimization model, algorithm, classification, hybrids of fish, nonnegative matrix factorization

INTRODUCTION

Taxonomy aims to define and name groups of biological organisms on the basis of their shared similarity in morphological structure and physiological functions (Tautz et al., 2002). It plays an important role in understanding the relationship and evolution between different groups (Tautz et al., 2003). From classical morphology to new achievements in modern molecular biology, taxonomy also involves the comprehensive application of biological multidisciplinary, which can be used as a basis for classification, such as chromosome-based cell taxonomy (or chromosomal taxonomy), serum taxonomy based on serum reaction, chemical composition-based chemical taxonomy, and DNA taxonomy, with the sequence analysis of a uniform target gene (Stoeckle, 2003).

In the past two decades, with an increasing number of genome-wide sequencing and fine mapping, extensive data on transcriptomics, proteomics and metabolomics are available in the literature (Liu et al., 2016; Ren et al., 2016; Ren et al., 2017a; Ren et al., 2017b; Flouriou-Servou et al., 2018; Li et al., 2018; Wang L. et al., 2018; Wang M. et al., 2018; Wang N. et al., 2018; Ye et al., 2018; Chen et al., 2019; Liu et al., 2019; Ning et al., 2019). To mine more and more biological information from these data, many computational models have been established to classify different species or examine their genetic relationships (Yang et al., 2015; Tan et al., 2019). For example, in (Wang L. et al., 2018; Wang M. et al., 2018; Wang N. et al., 2018; Yu et al., 2015; Wang et al., 2017; Hu et al., 2012), some statistical methods
and statistical software have been used for biological classification by analyzing the data of protein sequences. However, to our best knowledge, there exists no research result on classification of distant multi-generation hybrid fishes in virtue of transcriptome data and optimization techniques.

Distant hybridization is a hybrid between two different species (Lou and Li, 2006). For this interspecific hybridization, it may be a hybrid of different species of the same genus, or between different genus, between different subfamilies, between different families, and even between different individuals (Zhang et al., 2014). Since distant hybridization can transfer a set of genomes from one species to another, it can effectively change the genotype and phenotype of hybrid progeny (Liu et al., 2001). In terms of genotype, distant hybridization can lead to changes in the genomic level and sub-genome levels of the offspring, and the formation of these different hybrid progeny often depends on the genetic relationship of the parent. In terms of phenotype, the distant hybridization can integrate the genetic characteristics of the parents, which may make hybrid progeny show heterosis in aspects of shape, growth rate, survival rate and disease resistance (Hu et al., 2012). It has been shown that the distant hybridization occurs widely in fishes and has become an effective tool to integrate existing natural species and quickly cultivate more excellent traits in fisheries development. For more details, readers are referred to recently published article (Qin et al., 2014; Hu et al., 2019) and the references therein.

Different from protein (DNA) sequences, the transcriptome of a cell or a tissue is the collection of RNAs transcribed in it, and is often dynamic and a good representative of the cellular state (Carnes et al., 2018). Ease of genome-wide profiling using sequencing technologies further makes the transcriptome analysis an important research tool of bioinformatics, where the information content of an organism is recorded in the DNA of its genome and expressed through transcription (Kaletsky et al., 2018). Therefore, full-length transcriptome analysis of distant multi-generation hybrid fishes seems to be a more useful tools to provide a more profound explanation for the biological performance of distant multi-generation hybrid fishes. However, on the one hand, cultivating new generation of hybrid fishes often needs more than one and a half years, hence collection of the relevant experimental data is difficult, such that only the small-size sample inference can be made (Rogoza, 2019). On the other hand, owing to a lack of effective classic statistical methods to analyze the small-size and full-length transcriptome sample data, genomic research on similarity of this species and its descendants based on optimization models is unavailable in the literature. Actually, since the full-length transcriptome data is associated with expressed levels of ten thousands genes, classification of small-size sample data becomes impossible by using existing statistical methods. In this paper, combining the RNA sequencing group data of distant hybrid progeny and parental types, we intend to develop a new method for the genetic regulation of the whole transcriptome to statistically analyze the distant hybrid progeny and its excellent germplasm selection.

Basically, our new research method originates from optimization techniques, called a nonnegative matrix factorization method (NMF). By this method, we attempt to approximately factorize the small-size and full-length transcriptome sample data of the distant multi-generation hybrid fishes such that their classification and the gene-expression characteristic of each class can be revealed. As a result, it is associated with solution of large-scale optimization problems with nonnegativity constraints. Therefore, we also aim to develop an efficient algorithm for solving this large-scale optimization problem.

Clearly, one of the challenges in this research lies in making statistical inference from the small-size samples. We have collected 24 samples (liver tissues) of the distant multi-generation hybrid fishes, which constitutes three different groups corresponding to the three sampling periods. Each group consist of 8 samples with 20093 features. As stated in (El-Shagi, 2017; Ristic-Djurovic et al., 2018; Rogoza, 2019), if the size of samples is small, it is difficult to believe that the classical statistical methods can give good prediction accuracy owing to bias of small-size samples. For the small-size samples, the existing main inference methods include: the probabilistic index models (Amorim et al., 2018), the bootstrapping U-statistics method (Jiang and Kalbfleisch, 2012), the Jackknife empirical likelihood inference (Zhao et al., 2015), the SVM-based methods (Cong et al., 2016), the grey-theory-based methods (Meng et al., 2017), and the neural network (Zhu et al., 2019). However, for the small-size samples with more than ten thousand features, such as the full-length transcriptome sample data of the distant multi-generation hybrid fishes, it is desirable to study new statistical inference methods to mine their statistical information.

The NMF has been regarded as a useful tool of unsupervised machine learning to classify the small-size samples with large-scale features (Pauca et al., 2006; Wan et al., 2018). It can integrate the functions of k-mean clustering method and PCA. However, the performance of NMF depends significantly on the development of efficient algorithms to solve the generated large-scale optimization problem such that the deviation of nonnegative matrix (sample data) factorization is minimized. Especially, if we need to classify 8 full-length transcriptome data of distant multi-generation hybrid fishes, it is necessary to factorize a matrix in $R^{20093 \times 8}$. Suppose that there are $r$ classes of fishes, then the number of design variables is $20093 \times r + 8$. For solving such a large-scale optimization model, it is still a challenge to develop an efficient algorithm. In this research, we intend to modify the spectral conjugate algorithm in (Deng et al., 2013) to solve the generated large-scale optimization problems. Our goal is to reveal the relationship between multi-generation hybrid fishes on the basis of their gene expression profile described by their transcriptome data.

**MATERIALS AND METHODS**

**Samples and Transcriptome Sequencing**

The *Megalobrama amblycephala* or Bluntnose black bream (BSB, $2n = 48$) and *Culture Alburnus* or Topmouth culter (TC, $2n = 48$) at sexual maturity in natural waters of the Yangtze River in China were collected for creating the allodiploids BT (BSB ($\Phi$) × TC ($\delta$)).
and TB (TC (♀) × BSB (♂)) F1 individuals through intergeneric reciprocal crosses of BSB and TC, respectively. Then, the allopolyploid F2 - F1 (2n = 48) hybrid offspring were obtained by self-mating of F1 - F2 populations, respectively. The chimeric offspring was identified based on 45S rDNA sequencing characteristics (Xiao et al., 2016), had been used in our study.

**Transcriptome Sequencing and Gene Expression Profiles**

To sequence the transcriptomes of reciprocal cross hybrids and their inbred parents, total RNA was isolated and purified from the liver by a TRIzol extraction method (Río et al., 2010). RNA concentration was measured using Nanodrop technology. Total RNA samples were treated with DNase I (Invitrogen) to remove any contaminating genomic DNA. The purified RNA was quantified using a 2100 Bioanalyzer system (Agilent, Santa Clara, CA, USA). After the isolation of 1 μg mRNA using the beads with oligo (dT) Poly (A), fragmentation buffer was added for interrupting mRNA to short fragments. The resulting short fragments were reverse transcribed and amplified to produce cDNA. An Illumina RNA-seq library was prepared according to a standard high-throughput method ephigh-throughput method (Dillies et al., 2013). The cDNA library concentration and quality were assessed by the Agilent Bioanalyzer 2100 system, after which the library was sequenced with paired-end setting using the Illumina HiSeq 2000/4000 platform. Then, the raw reads containing adapters, ploy-N and low quality were removed using in-house perl scripts. The high quality reads were used in our analysis. The transcriptome data was obtained from the NCBI database.

All Illumina reads of *M. Amblycephala* and *C. albimnus* were aligned to the *M. Amblycephala* and *C. albimnus* genome using Star (v 2.4.0) with the default parameters (Bennett et al., 2001), respectively. The other RNA-seq reads of reciprocal cross hybrids were aligned to the two reference genomes of *M. Amblycephala* and *C. albimnus*, respectively. The numbers of mapping counts in each gene were calculated with in-house perl scripts. Consequently, the two mapping results of aligning to two reference genomes were obtained in hybrid offspring, and the total expression value was normalized based on ratio of the number of mapped reads at each gene to the total number of mapped reads for the entire genome.

**Data Download**

The collected data of 24 samples (liver tissues) of the distant multigeneration hybrid fishes in this research have been uploaded to https://github.com/TJY0622/TJY and can be downloaded freely such that the numerical experiments in this paper can be repeated by anyone. The last upload time is 07-20-2019 (File name as 2019_7_8 Copy.xlsx).

**An Optimization Model for Classifying the Hybrids Fishes**

We first propose an optimization model for classifying the hybrids fishes on the basis of NMF. Mathematically, NMF is stated as follows. For a given matrix $A \in \mathbb{R}^{n \times m}$, we need to decompose $A$ into two nonnegative matrices $W$ and $H$, i.e.

$$A = WH$$

(2.1)

where $W \in \mathbb{R}^{n \times r}$ and $H \in \mathbb{R}^{r \times m}$. In particular, if the matrix $A$ in (2.1) is the full-length transcriptome data of the distant multigeneration hybrid fishes, and $A = WH$, then $r$ can represent the number of classes for this classification of fishes in the case that each column of $H$ has only a unique element 1, while the other elements are zeros. Clearly, in this ideal case, the $k$-th column of $W$ stands for the gene expression level of the $k$-th class of fishes, and its elements show the expression levels of different genes for each class. Therefore, $W$ in Model (2.3) is called a base matrix in view of its practical meanings, while $H$ is called a coordinate matrix.

For real sample data, it is often difficult to obtain the above ideal result of factorization. Therefore, we relax $A = WH$ by $A \approx WH$. In this case, each column of the matrix $A$ is approximately equal to the linear combination of all column vectors of the matrix $W$, and the combination coefficients are given by the corresponding column vector of the matrix $H$. We express the $j$-th column of the matrix $A$, $W_{j}$ stands for the $j$-th column of the matrix $W$, and $h_{j}$ represents the element of the $j$-th row and the $j$-th column in the matrix $H$. In other words, $A=[A_{1},...,A_{m}] \in \mathbb{R}^{n \times m}$, $W=[W_{1},...,W_{r}] \in \mathbb{R}^{r \times m}$, and $H = [h_{1},...,h_{m}] \in \mathbb{R}^{r \times m}$.

Thus, if we define a membership matrix $R \in \mathbb{R}^{r \times m}$:

$$R_{j,i} = \frac{1}{m} \sum_{k=1}^{m} \frac{W_{j,k} \cdot h_{i,k}}{\sum_{k=1}^{r} W_{j,k} \cdot h_{i,k}}, i=1,...,r; j=1,...,m.$$  

(2.2)

Clearly, the $j$-th column of $R$ represents the membership degrees of the $j$-th sample being affiliated all the different classes. Therefore, for all the samples, distinct differences of all the elements in each column of $R$ imply an approximate classification result. By definition, the matrix $R$ shows the result of classification in term of membership degrees, while each column of the matrix $H$ exactly stands for the coordinate of each sample in the $r$-dimensional space linearly expanded by the $r$ columns of $W$. In the case that all the $r$ elements in each row of $W$ have the same orders of magnitude, the classification results by $H$ or $R$ are same.

Unfortunately, it is very difficult to solve Problem (2.1) when $n$ is very large, let alone the requirement of finding the unknown optimal number of classes $r$. To solve Problem (2.1), we first transform (2.1) into the following optimization model:

$$\min_{W,H} F(W,H) = \frac{1}{2} \| A - WH \|_F^2$$

(2.3)

s.t. $W,H \geq 0$.

where $\| \cdot \|_F$ is the Frobenius norm. It has been shown that (2.3) is non-convex and NP-hard (Vavasis, 2009). Then, similar to the technique of alternating non-negative least squares (ANLS) in (Chu et al., 2004), we solve (2.3) by finding the optimal solutions of the following two convex sub-problems:

$$W_{(k+1)} = \arg \min_{W \geq 0} F(W,H_{(k)}),$$

(2.4)

$$H_{(k+1)} = \arg \min_{H \geq 0} F(W_{(k+1)}, H).$$

(2.5)
It is noted that the above model of NMF was first proposed in (Paatero and Tapper, 1994). Summarily, there are two types of algorithms to solve Model (2.3) (Lin, 2007): the multiplicative update (MU) method (Cai et al., 2010; Shang et al., 2012; Huang et al., 2018; Deng et al., 2019) and the technique of alternating non-negative least squares (ANLS) (Chu et al., 2004). For the ANLS, a main focus is on development of efficient algorithms to solve the subproblems (2.4) and (2.5). For example, the projected gradient (PG) method (Lin, 2007), the projected Newton method (Gong and Zhang, 2012), and the projected quasi-Newton method (Zdunek and Cichocki, 2006) have been reported to be efficient for solving the large-scale optimization model (2.3), although no one method has overwhelming advantage compared with the others.

Recently, Deng et al. (2013) proposed an efficient algorithm to solve general large-scale unconstrained optimizations, and they demonstrated that the numerical performance of this algorithm outperforms the similar ones available in the literature. In this paper, we intend to extend it into solution of the subproblems (2.4) and (2.5), which are two large-scale optimization problems with nonnegativity constraints.

**Development of Algorithm**

We are now in a position to present an efficient algorithm to solve the subproblems (2.4) and (2.5). Since both of them are large scale (the size of the problem is over 80000), we will extend the spectral conjugate gradient algorithm in (Deng et al., 2013) to solve the subproblems (2.4) and (2.5). Actually, in our previous research, this algorithm has been implemented to solve more than 700 large-scale benchmark test problems, and has been shown that its numerical performance outperforms the similar ones available in the literature.

In need of modifying the developed algorithm in (Deng et al., 2013) such that it can be used to solve Model (2.3), we first define the gradients of $W$ and $H$, respectively:

$$\nabla_W F(W, H) = -AH^T + WHH^T, \quad \nabla_H F(W, H) = -W^TA + W^TWH.$$  (2.6)  (2.7)

Then, we denote the following two matrices the gradients of $F(W, H)$ with respect to the matrices $W$ and $H$, respectively:

$$\nabla_W F(W, H) = -AH^T + WHH^T, \quad \nabla_H F(W, H) = -W^TA + W^TWH.$$  (2.8)

For two given matrices $S$ and $T$ with the same size, we define their inner product by

$$\langle S, T \rangle = \sum_{i,j} S_{ij} \times T_{ij}.$$  (2.9)

Then, for $k = 0$, a search direction of $F$ at a given initial point $W^{(0)}$ is

$$D_0 = -\nabla_W F(W^{(0)}, H^{(0)}) = A(H^{(0)})^T - W^{(0)}H^{(0)}(H^{(0)})^T.$$  (2.10)

And for $k \geq 1$, we define four matrices:

$$s_k = W^{(k)} - W^{(k-1)}, \quad G_k = \nabla_W F(W^{(k)}, H^{(k)}), \quad y_k = G_k - G_k^{(k-1)}, \quad y_k = y_k - y_k^{(k-1)}.$$

where $H^{(k)}$, $W^{(k)}$ and $W^{(k-1)}$ are two given matrices. Similar to (Deng et al., 2013), we compute the spectral parameter and conjugate parameter by

$$\theta_k = \begin{cases} \frac{\langle D_k, y_k \rangle - \langle D_k, G_k \rangle \langle G_k, s_k \rangle}{\|G_k\|^2}, & \text{if } \langle D_k, y_k \rangle > \eta \|G_k\|^2, \\ \frac{\langle D_k, s_k \rangle}{\|G_k\|^2}, & \text{otherwise}, \end{cases}$$  (2.11)

And

$$\beta_k = \begin{cases} \frac{\langle G_k, y_k \rangle - \langle G_k, s_k \rangle \langle s_k, y_k \rangle}{\|G_k\|^2}, & \text{if } \langle D_k, y_k \rangle > \eta \|G_k\|^2, \\ \frac{\langle G_k, s_k \rangle}{\|G_k\|^2}, & \text{otherwise}, \end{cases}$$  (2.12)

where $D_k$ is the search direction at $W^{(k-1)}$, determined by

$$D_k = \begin{cases} D_0, & \text{if } k = 0, \\ -\theta_k G_k + \beta_k D_{k-1}, & \text{if } k > 0. \end{cases}$$  (2.13)

The following algorithm is developed to solve the subproblem (2.4) with the given $H^{(k)}$.

**Algorithm 1** (Modified Spectral Conjugate Gradient Algorithm)

**Step 0 (Initialization).** Given constants $0 < \delta_1, \eta, \rho < 1$, $0 < \delta_2, \epsilon$. Choose an initial matrix $W^{(0)} \in R^{m \times n}$. Set $k = 0$.

**Step 1 (Search direction).** If $\|G_k\| \leq \epsilon$, then the algorithm stops. Otherwise, compute $D_k$ by (2.9) and (2.13).

**Step 2 (Step length).** Determine a step length $\alpha_k = \max\{\alpha_k | \alpha_k = \rho^j, j = 0, 1, 2, \ldots\}$ such that $\alpha_k$ satisfies the following inequality:

$$F(W^{(k+1)}, \alpha_k D_k, H^{(k)}) \leq F(W^{(k)}, H^{(k)}) + \delta_2 \alpha_k \langle G_k, D_k \rangle - \delta_1 \alpha_k^2 \|G_k\|^2.$$  (2.14)

where

$$\|G_k\| = \sum_{i=1}^m \sum_{j=1}^n |G_{ij}|.$$  (2.15)

**Step 3 (Update).** Set $W^{(k+1)} := W^{(k)} + \alpha_k D_k$ and $k := k + 1$. Return to Step 1.
Similarly, to solve the subproblem (2.5), we only need replace $W$ and $H$ by $h$ and $w$ in Algorithm 1, respectively. Particularly, we need to compute

$$s_{k+1} = H^{(k)} - H^{(k-1)},$$

$$G_k^{(k)} = \nabla H F(W^{(k)}(H^{(k)}),$$

$$y_{k+1} = G_k^{(k)} - G_k^{(k-1)},$$

$$\bar{y}_{k+1} = y_{k+1} - G_k^{(k)} H^{(k+1)},$$

and

$$D_k = \begin{cases} D_k, & \text{if } k = 0, \\ -\theta_k G_k^{(k)} + \beta_k D_{k-1}, & \text{if } k > 0, \end{cases}$$

where

$$\theta_k = \begin{cases} \frac{\langle D_{k-1}, y_{k+1} \rangle - \langle D_{k-1}, y_{k+1} \rangle}{\|D_{k-1}\|} & \text{if } \langle D_{k-1}, y_{k+1} \rangle > \eta \|G_{k-1}^{(k-1)}\|, \\ \frac{\langle D_{k-1}, y_{k+1} \rangle - \langle D_{k-1}, y_{k+1} \rangle}{\|D_{k-1}\|} & \text{otherwise}, \end{cases}$$

and

$$\beta_k = \begin{cases} \frac{\langle G_k^{(k)} y_{k+1} \rangle - \langle G_k^{(k)} y_{k+1} \rangle}{\|G_k^{(k)} y_{k+1}\|} & \text{if } \langle D_{k-1}, y_{k+1} \rangle > \eta \|G_{k-1}^{(k)}\|, \\ \frac{\langle G_k^{(k)} y_{k+1} \rangle - \langle G_k^{(k)} y_{k+1} \rangle}{\|G_k^{(k)} y_{k+1}\|} & \text{otherwise}, \end{cases}$$

With the above preparation, we now develop an overall algorithm to solve Model (2.3) in the end of this section.

**ALGORITHM 2** | Step 0 (Initialization). Randomly generate two initial non-negative matrices $W_0 \in R^{m \times m}$ and $H_0 \in R^{n \times n}$. Take constants $S_0$, $S_1$, $S_2$, $S_3$, $S_4$, $S_5$ in the interval (0,1). Choose $\theta < S_0$, $S_1$, $S_2$, $S_3$, $S_4$, $S_5$. Then, set $k = 0$.

Step 1 (Judgement). If KKT $(W_0, H_0)$ does not satisfy KKT $(W_0, H_0)$, where KKT denotes the KKT conditions of Problem (2.1), and KKT $(W_0, H_0)$ denotes the value of KKT at the matrix $W$ and $H$. Then, this algorithm stops.

Step 2 (Solution of Subproblem (2.4)). Solve the subproblem (2.4) with $H = H_0$ by Algorithm 1; its optimal solution is referred to as $W^{(k+1)}$.

Step 3 (Projection of $W$). Replace $W^{(k+1)}$ by

$$W_{k+1} = \begin{cases} 0, & \text{if } W_{k+1} < 0, \\ W_{k+1}, & \text{otherwise}, \end{cases}$$

Step 4 (Solution of Subproblem (2.5)). Solve the subproblem (2.5) with $W = W^{(k+1)}$ by Algorithm 1; the optimal solution is referred to as $H^{(k+1)}$.

Step 5 (Projection of $H$). Replace $H^{(k+1)}$ by

$$H_{k+1} = \begin{cases} 0, & \text{if } H_{k+1} < 0, \\ H_{k+1}, & \text{otherwise}, \end{cases}$$

Step 6 (Update). Set $k := k + 1$. Go to Step 1.

**Remark 1** Compared with the similar algorithms available in the literature (Li and Wan, 2019), Algorithms 1 and 2 present a different computational procedure to solve Problem (2.1). Since the existing nonnegative matrix factorization methods depend on development of efficient solution algorithms, one of our contributions in this paper lies in developing Algorithms 1 and 2 to solve a sequence of subproblems like (2.4) and (2.5). Especially, in the section of result, we will implement them to solve the classification problem of distant multi-generation hybrid fishes based on their transcriptome profiles.

**Remark 2** In order to improve efficiency of Algorithm 2, before factorization of $A$, we conduct normalization of the sample data of fishes as follows.

$$b_i = \max_{\text{batch}} A_{ij} i = 1,...,n$$

$$a_i = \min_{\text{batch}} A_{ij} i = 1,...,n$$

$$A_{ij} = \frac{A_{ij} - b_i}{a_i - b_i} i = 1,...,n$$

where $A \in R^{n \times m}, A_j$ denotes the element of the $j$-th row and the $i$-th column in the matrix $A, AA_j$ denotes all the elements of the $j$-th row of the matrix $A$.

**Remark 3** In Algorithm 2, since it is possible that the sequences $\{W^{(k)}\}$ and $\{H^{(k)}\}$ are trapped near a curved valley, we take $\text{KKT}(W^{(k)}, H^{(k)})$ as the termination condition, rather than $\text{KKT}(W^{(k)}, H^{(k)}) < \epsilon$.

**RESULTS**

In this section, in virtue of Model (2.3) and Algorithm 2, we present the results on classification of the distant multi-generation hybrid fishes based on their transcriptome data.

**Result Of Classification**

With the given transcriptome data of the distant multi-generation hybrid fishes, we easily get Model (2.3). Then, we implement Algorithm 2 to solve this model by choosing the same values of algorithmic parameters as in (Deng et al., 2013):

$$\epsilon = 10^{-7}, \delta^W = \delta^H = 0.4, \eta^W = \eta^H = 0.001, \delta_W = \delta_H = 0.001, \rho^W = \rho^H = 0.65.$$
In Table 1, we report all the numerical results corresponding to the different class numbers.

Table 1 shows that when \( r = 6 \), all the samples are clearly classified owing to existence of greater deviation of elements in the same column of \( H \). In contrast, when \( r \) is equal to the other values, there are at least one sample that can not be clearly classified.

As \( r = 6 \), Table 1 indicates that the eight fishes can be categorized into 6 classes: BSB\(_{i}^{L} \), TC\(_{i}^{L} \), BT\(_{i}^{L} \) and BT\(_{i}^{H} \) belong to different four classes, respectively. BT\(_{i}^{L} \) and BT\(_{i}^{H} \) consist in another class. BT\(_{i}^{L} \) and BT\(_{i}^{H} \) are the same class.

For the sake of better understanding the above classification result, we use \( r = 6 \) as the number of classes to calculate the membership matrix \( R \) defined by (2.2). The numerical results are listed in Table 1, while Figure 1 more intuitively describe the biological similarity for the fish of each class.

Table 2 and Figure 1 further indicate that by membership matrices, the same classification result is obtained as that by coordinate matrices: \{BSB\(_{i}^{L} \}; \{TC\(_{i}^{L} \}; \{BT\(_{i}^{L} \}; \{BT\(_{i}^{H} \}; \{BT\(_{i}^{L} \}, BT\(_{i}^{H} \}; \{TB\(_{i}^{L} \}, TB\(_{i}^{H} \} \). Particularly, either by \( H \) or by \( R \), BSB\(_{i}^{L} \) and TC\(_{i}^{L} \) always belong to different two classes, while their hybrids are divided into different classes from the parents’ ones. In Figure 1, Classes 1 to 6 are described by the colors of yellow, blue, green, purple, gray and red, respectively. It follows from Figure 1 that larger proportion of the green color in BT\(_{i}^{L} \) and BT\(_{i}^{H} \) (that of the yellow color in TB\(_{i}^{L} \) and TB\(_{i}^{H} \)) demonstrate that there exists greater degree of biological similarity between BT\(_{i}^{L} \) and BT\(_{i}^{H} \) (between TB\(_{i}^{L} \) and TB\(_{i}^{H} \)).

To further test robustness of the above trained results, given \( r = 6 \), we choose the 1st-group and the 3rd-group samples (with superscripts \( L \) and \( H \), respectively) as two test sets to see whether the results are the same or not.

In Table 3 and Figure 2, we report the numerical results. The used colors in Figure 2 only be used to show the similarity of fishes within the same figure. In other words, the same color has no any relation in different figures.

From Table 3 and Figure 2, it is clear that 6 out of 8 samples in the 1st-group or the 3rd-group are correctly classified, compared with the trained result from the samples of the 2nd-group. The accuracy rate reaches 75%. In Table A3, we show that the elements in each row of the matrix \( W \) have different orders of magnitude for the 1st-group samples, which can explain inconsistence of the classification results by \( H \) and \( R \) for the 4 samples: BSB\(_{i}^{L} \), BT\(_{i}^{L} \), BT\(_{i}^{H} \) and BT\(_{i}^{H} \).

To further validate the proposed model and algorithms in this paper, we use them to classify more test samples generated by mixing the training set and the test sets.

### Table 1 | Coordinate matrices for the 2nd-group samples.

| Class | BSB\(_{i}^{L} \) | BT\(_{i}^{L} \) | BT\(_{i}^{H} \) | BT\(_{i}^{L} \) | TB\(_{i}^{L} \) | TB\(_{i}^{H} \) | TC\(_{i}^{L} \) |
|-------|----------------|--------------|--------------|--------------|--------------|--------------|--------------|
| \( r = 2 \) | 1525 | 2293 | 1646 | 2843 | 1302 | 1552 | 2060 | 0 |
| 1st | 0 | 0 | 4290 | 0 | 7198 | 6461 | 1057 | 40.79 |
| 2nd | 0 | 0 | 2181 | 743.4 | 3759 | 967.7 | 664.0 | 2353 | 0 |
| 3rd | 0 | 0 | 1.435 | 0 | 2.146 | 1.957 | 2.932 | 11.28 |
| 4th | 0 | 0 | 2342 | 0 | 355.2 | 107.4 | 343.5 | 919.9 | 89.23 | 0 |
| 5th | 0 | 0 | 6080 | 494.8 | 0 | 2787 | 430.6 | 165.2 | 1474 | 0 |
| 6th | 0 | 0 | 4607 | 3888 | 20.74 | 1465 | 1704 | 1128 | 0 |
| \( r = 5 \) | 0.0104 | 0.2348 | 0 | 1.070 | 0 | 0 | 0.1612 | 0.0024 |
| 2nd | 210.5 | 0 | 36.08 | 0.3265 | 21.54 | 75.89 | 0 | 0 |
| 3rd | 0 | 0 | 1412 | 1070 | 0 | 167.9 | 351.8 | 0 | 0 |
| 4th | 0 | 0 | 181.0 | 0 | 290.1 | 841.0 | 0 | 0 |
| 5th | 0 | 0 | 0.0571 | 0 | 0.2183 | 0.9724 | 11.10 | 4.158 |
| 6th | 0 | 0 | 3.622 | 0.6887 | 0.0606 | 0.4686 | 0.0237 | 0 |
| \( r = 7 \) | 0.0001 | 0.1599 | 0.1299 | 0 | 0 | 0.0300 | 0.0020 |
| 2nd | 0 | 0 | 0.0574 | 0 | 0.0076 | 1.769 | 0.0537 | 0.0020 |
| 3rd | 0 | 0 | 0 | 0 | 0.0044 | 0 | 0.1940 | 0 |
| 4th | 0 | 0 | 0 | 0 | 0 | 0 | 290.1 | 0 |
| 5th | 0 | 0 | 0 | 0 | 0 | 0 | 1.435 | 0 |
| 6th | 0 | 0 | 4695 | 0 | 0 | 0 | 0.6887 | 0 |
| 7th | 0 | 0 | 0 | 0 | 0 | 0.2342 | 0 | 0 |
| 8th | 0 | 0 | 0 | 0 | 0 | 0 | 0.0475 | 0 |

Frontiers in Genetics | www.frontiersin.org 6 October 2019 | Volume 10 | Article 911
TABLE 2 | Membership matrix $R$ of the 2nd-group samples.

| Class | $BSB^{12}$ | $BT_1^{32}$ | $BT_2^{32}$ | $BT_3^{32}$ | $TB_1^{32}$ | $TB_2^{32}$ | $TB_3^{32}$ | $TC^{12}$ |
|-------|------------|-------------|-------------|-------------|-------------|-------------|-------------|-----------|
| 1st   | 0          | 0           | 0.2826      | 0           | 0.6049      | 0           | 0.6132      | 0         |
| 2nd   | 0.3713     | 0.3880      | 0           | 0.8713      | 0           | 0           | 0.2588      | 0.1597    |
| 3rd   | 0          | 0.4268      | 0.4047      | 0           | 0.1333      | 0           | 0           | 0         |
| 4th   | 0          | 0.1653      | 0.0080      | 0           | 0           | 0.1029      | 0.8493      | 0.1280    |
| 5th   | 0          | 0           | 0.1329      | 0           | 0           | 0           | 0.1507      | 0         |
| 6th   | 0.6287     | 0           | 0.1718      | 0.1287      | 0.1589      | 0.1507      | 0           | 0         |

FIGURE 1 | Membership degrees of the 2nd-group samples.

We first mix the training set and the 3rd-group test set. The obtained results are listed in Table 4. Table 4 demonstrates that compared with the trained result, 13 out of 16 samples are correctly classified by both of the membership and coordinate matrices, which includes all the samples in the 2nd-group and the 5 samples in the 3rd-group: $BSB^{12}$, $BT_1^{32}$, $BT_2^{32}$, $TB_1^{32}$ and $TB_2^{32}$. The accuracy rate is as high as 81.25%. Additionally, for the 5 species of fish ($BSB$, $BT_1$, $BT_2$, $TB_1$, and $TB_2$), the replicated samples of each fish are correctly classified into the same class in our test experiments, which also validates the proposed model and algorithms in this paper.

Next, we compute the classification result of all 24 samples (8 samples in the training set, 16 samples in the two test sets). The results are given in Table 5. From Table 5, we know that 17 out of 24 samples are correctly classified by the membership matrix or the coordinate matrix, which excludes $BSB^{12}$, $BT_1^{32}$, $BT_2^{32}$, $TB_1^{32}$, $TB_2^{32}$, $TB_1^{32}$, and $TC^{12}$. The accuracy rate achieves 70.83%, compared with the trained results. In this test, for the 4 species of fish ($BSB$, $BT_1$, $TB_1$, and $TB_2$), the replicated samples of each fish are correctly classified into the same class.

In summary, by all of the above test experiments, the average accuracy rate is 75.52% even if there exists larger detection error of the input initial sample data (see our subsequent correlation analysis). These tests further verifies that the proposed model and algorithm in this paper can be used to efficiently classify the distant multi-generation hybrid fishes based on their transcriptomic profile.

Correlation Analysis

To find out the reasons why the replicated samples are incorrectly classified such that the accuracy rate may be reduced, we calculate the correlation matrix of the sample data to reveal possible detection errors of the input initial data. In Figure 3, the correlation coefficient matrix of the 24 samples is concisely plotted.

From Figure 3, it is easy to see that the sample of $BSB^{12}$ is only weakly correlated with the two replicated samples $BSB^{12}$ and $BSB^{12}$. Their correlation degree is even less than that between the samples of different fish $BSB^{12}$ and $TB_1$. It can explain why $BSB^{12}$ can not be clearly classified into the same class of $BSB^{12}$ and $BSB^{12}$ (revisiting the results in Table 5). Conversely, Figure 3 shows that in the 1st-group, the sample $BSB^{12}$ has greater correlation with the other 3 samples: $TB_1^{32}$, $TB_2^{32}$ and $TC^{12}$, which answers why the class of $BSB^{12}$ can not be clearly identified in Table 3.

From Figure 3, we can also find out similar reasons for the unsatisfactory classification of $BT_1^{32}$, $BT_2^{32}$, $BT_3^{32}$ in Tables 4 and 5. Actually, (1) owing to lower correlation among $BT_1^{32}$, $BT_2^{32}$ and $BT_3^{32}$, they can not be classified into the same class even if they are the three replicated samples. (2) In the 3rd group, the class of $BT_3^{32}$ can not be...
clearly identified in Table 3 since its sample is more correlated with the other 5 samples: BS, BB, TT, BT, and TC.

Similarly, because the sample of BT, F, L, is only little correlated with the two replicated samples BT, F, L, and BT, F, L, it cannot be classified into the same class of BT, F, L, in Table 5.

For the same reason of weaker correlation, in Tables 4 and 5, the three replicated samples of TB, F, L (TC) are also classified into the different classes. It is believed that if the detection errors of samples can be controlled to be small enough, the proposed model and algorithms in this paper can provide a more
TABLE 4 | Results for the mixed samples of the 2nd/3rd group.

| Coordinate matrices | Class | BSB<sup>4<sup>th</sup></sup> | BSB<sup>3<sup>rd</sup></sup> | BT<sup>3<sup>rd</sup></sup> | BT<sup>2<sup>nd</sup></sup> | BT<sup>2<sup>nd</sup></sup> | BT<sup>1<sup>st</sup></sup> | BT<sup>1<sup>st</sup></sup> |
|---------------------|-------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 1<sup>st</sup>      | 0     | 0.119           | 0.0451          | 0.0094          | 0               | 0               | 0.0208          | 0.0486          |
| 2<sup>nd</sup>      | 0     | 169.6           | 2662            | 17.07           | 1801            | 2057            | 0               | 27.20           |
| 3<sup>rd</sup>      | 0.4607| 0               | 0.5200          | 0.0245          | 0               | 0.3763          | 2.5665          | 2.5614          |
| 4<sup>th</sup>      | 0     | 139.7           | 545.9           | 129.5           | 3907            | 0               | 3.533           |
| 5<sup>th</sup>      | 2719  | 2378            | 922.6           | 669.6           | 276.0           | 0               | 0               |
| 6<sup>th</sup>      | 0     | 0.4276          | 0               | 1.414           | 1.342           | 1.940           | 0.1942          | 0.0491          |

| Membership matrices | Class | BSB<sup>4<sup>th</sup></sup> | BSB<sup>3<sup>rd</sup></sup> | BT<sup>3<sup>rd</sup></sup> | BT<sup>2<sup>nd</sup></sup> | BT<sup>2<sup>nd</sup></sup> | BT<sup>1<sup>st</sup></sup> | BT<sup>1<sup>st</sup></sup> |
|---------------------|-------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 1<sup>st</sup>      | 0     | 0.1093          | 0.0705          | 0.0813          | 0               | 0               | 0.0702          | 0.0894          |
| 2<sup>nd</sup>      | 0     | 0.1001          | 0.6086          | 0.0140          | 0.4571          | 0.4481          | 0               | 0.0340          |
| 3<sup>rd</sup>      | 0.4360| 0               | 0.3210          | 0.0256          | 0               | 0.1797          | 0.8340          | 0.8241          |
| 4<sup>th</sup>      | 0     | 0.1616          | 0               | 0.3274          | 0.1147          | 0.0102          | 0               | 0.0204          |
| 5<sup>th</sup>      | 0.5640| 0.4885          | 0               | 0.2465          | 0.1746          | 0.0803          | 0               | 0               |
| 6<sup>th</sup>      | 0     | 0.1404          | 0               | 0.3052          | 0.2536          | 0.2818          | 0.0957          | 0.0322          |

satisfactory result of classification. Actually, for the three species of fish: BT<sub>F</sub>, BT<sub>L</sub> and BSB<sub>F</sub>, their three replicated samples can always classified into the respective same class (see Tables 4 and 5), which may be related with higher correlation between them as shown in Figure 3.

Genes Of High Expression

In the end of this section, based on our classification result from the 2nd-group samples, we answer what are the differently expressed genes in all the six classes. By definition, we know that each column of the base matrix W gives the feature of gene expression for each class of fish. Since the sample of each class consists of 20,093 genes, we only list a part of the highly expressed genes for each class. When r = 6, the highly expressed genes are reported in Table A1 and Figures 1(a), 1(b) and 1(c). From the numerical results in Table A1 and Figures 1(a), 1(b) and 1(c), it follows that there exists stronger genetic similarity between the BSB (parents) and the hybrids. Actually, the BSB (the 6<sup>th</sup> class) has 3 shared highly expressed genes with BT<sub>F</sub> (the 1<sup>st</sup> class), 45 shared highly expressed genes with BT<sub>L</sub> (the 2<sup>nd</sup> class) and 12 shared highly expressed genes with BT<sub>F</sub> (the 4<sup>th</sup> class). In contrast, the TC (the 5<sup>th</sup> class) does not have any shared highly expressed genes with their hybrids, which implies that their hybrids seem to look more like BSB, rather than TC, regardless of reciprocal hybrids.

Apart from one-by-one comparison in Table A1, we also statistically analyze the numbers of shared highly expressed genes for more than three classes of fish. The reported results in Table A2 demonstrate that BSB (6<sup>th</sup> class) has higher hereditary conservatism than TC (5<sup>th</sup> class). Actually, by comparing the numbers of shared highly expressed genes among BSB, TC and the hybrids, it is clear that the gene expression profile of their grandchildren looks more like BSB (6<sup>th</sup> class), rather than TC (5<sup>th</sup> class).

It is also noted that in Table A2, there are no shared expressed genes between BT<sub>F</sub> (3<sup>rd</sup> class) and TB<sub>F</sub> (1<sup>st</sup> class), or between BT<sub>L</sub> (3<sup>rd</sup> class) and TB<sub>F</sub> (4<sup>th</sup> class), and there only exist 3 shared highly expressed genes between the BT<sub>F</sub> (2<sup>nd</sup> class) and TB<sub>F</sub> (1<sup>st</sup> class). It suggests that the trait separation occurs between these hybrids.
In addition, from Table A2 and Figure A1, it follows that the hybrids have larger transcript intersection than that between the hybrids and the parents. Actually, there are 277 shared highly expressed genes among the hybrids (BSB) and the hybrids (BT); there are only less than 45 shared highly expressed genes between the hybrids (offspring) and their parents. Actually, there are 277 shared highly expressed genes among TB, TB (1st class), BT (2nd class) and TB (4th class), in contrast, there are only less than 45 shared highly expressed genes between the parent (BSB) and the hybrids (BT).
DISCUSSION

In our numerical experiments, it is found that the nonnegative factorization of the matrix $A$ is not unique. In particular, if we choose different initial matrices $W^0$ and $H^0$, the base and coordinate matrices $W$ and $H$ may be different. However, our numerical experiments show that for Algorithms 1 and 2, different choices of $W^0$ and $H^0$ do not affect the final result of classification. For example, as $r = 6$, the result of classification always is the same for any $W^0$ and $H^0$, which can show robustness of our classification method.

Hybridization is considered as the rapidly driving forces that shape epigenetic modifications in plants and parts of lower vertebrate (Liu et al., 2016; Mallet, 2005). The merge of divergent genome always results in a ‘genomic and transcriptome shock’ in newborn hybrid (Ren et al., 2017b; Wu et al., 2016; Ren et al., 2016). Analysis on the expression changes after hybridization, including expression dominance and expression bias related to specific function-regulated genes, always provides us insights into the molecule mechanism of various phenotypes including heterosis (Ren et al., 2016; Zhou et al., 2015). However, the multiple regulatory mechanism and complex protein interaction network restricted our ability to investigate the underlying regulation in hybrid.

It is noted that in this research, we choose the 2nd-group samples as the training set, instead of the 1st-group or 3rd-group, and the latter is regarded as test samples to verify the trained result. One of the reasons for our doing so lies in that correlation analysis of the three-group samples indicates that each sample in the second-group is better correlated with the other replicated samples than those in the other two groups.

The proposed model and algorithms in this paper can be extended to solve more practical engineering problems from other fields. For example, if we can collect sufficient transcriptome data of patients possibly suffering from breast cancer, we can apply the proposed model and algorithms to identify the classes of patients, even development of the relevant smart aided-system of diagnosis for the sufferers.

CONCLUSIONS

In this paper, we have constructed a classification model for the distant multi-generation hybrid fishes based on transcriptome data, and developed an efficient algorithm, called the modified spectral conjugate gradient algorithm, for solving such a model.

In virtue of our model and algorithm, we have obtained a satisfactory classification for a given full-length transcriptome data of fish samples, and the differently expressed genes of each class have been identified. Our results are first obtained by a training set of samples, then are tested by many test samples generated by different ways.

Main results are stated as follows.

(1) Even for input data with larger detection error, the average accuracy rate of classification still achieves 75.52% in all the test experiments. It suggests that our model and algorithms are promising in classifying the distant multi-generation hybrid fishes.

(2) Owing to the weakest intersection of highly expressed genes between BSB and TC, they are deterministically divided into two classes. However, there exists a higher transcript intersection between them and their hybrids. These findings have further deeply mined the biological genetic characteristics of distant hybridization generated by BSB and TC, based on optimization techniques and transcriptome data.
(3) Although the hybrids of TC and BSB have been divided into different classes, the hybrids display higher transcript intersection. Since the transcript intersection of the hybrids and the parents is smaller than that among the hybrids, it can be concluded that the hybrid progeny of TC and BSB has significant hybrid characteristics, which may be useful to carry out trait improvement in practice.

(4) Since $BT_5$ and $TB_5$ are classified to two different classes, where there only exist 3 shared genes of high expression, it is concluded that there exists larger trait separation in the third generation of TC and BSB hybrid progeny ($BT_5$ and $TB_5$). In other words, both $BT_5$ and $TB_5$ are a good variety for the reproduction of fish.

(5) Since there are no shared genes of high expression between $BT_5$ and $TB_5$, they belong to two different classes (1st and 3rd classes). It implies that the reciprocal hybrids in the first generation of TC and BSB ($BT_5$ and $TB_5$) have larger biological distinction.

DATA AVAILABILITY STATEMENT

The genome assembly used in this study was downloaded from NCBI BioProject database (BioProject: http://www.ncbi.nlm.nih.gov/bioproject/) under accession numbers PRJNA269572. All raw mRNA-Seq data were downloaded from the NCBI Sequence Read Archive (http://trace.ncbi.nlm.nih.gov/Traces/sra/) under accession number SRP050891.

AUTHOR CONTRIBUTIONS

ZW conceived and designed the study, wrote and revised the paper. JT designed and implemented the algorithm to analyze the data, and wrote the paper. LR, YX and SL did all the relevant experiments, collected the data and revised the paper.

FUNDING

This research was supported by the National Science Foundation of China (Grant 71671190), National Key Research and Development Program of China (2018YFD0901202), National Science Foundation of China (31772902), and State Key Laboratory of Developmental Biology of Freshwater Fish (2018KF003).

ACKNOWLEDGMENTS

The authors would like to express their thanks to the anonymous referees for their constructive comments on the paper, which have greatly improved its presentation.

REFERENCES

Amorim, G., Thas, O., Vermeulen, K., Vansteelandt, S., and De Neve, J. (2018). Small sample inference for probabilistic index models. Comput. Stat. Data Anal. 121, 137–148. doi: 10.1016/j.csda.2017.11.005

Benett, B. L., Sasaki, D. T., Murray, B. W., O’Leary, E. C., Sakata, S. T., Xu, W., et al. (2001). SP600125, an anthrapyrazolone inhibitor of Jun N-terminal kinase. PNAS 98 (24), 13681–13686. doi: 10.1073/pnas.251194298

Cai, D., He, X., Han, J., and Huang, T. S. (2010). Graph regularized nonnegative matrix factorization for data representation. IEEE Trans. Pattern Anal. Mach. Intell. 33 (8), 1548–1560. doi: 10.1109/TPAMI.2010.231

Carne, M. U., Allingham, R. R., Ashley-Koch, A., and Hauser, M. A. (2018). Transcriptome analysis of adult and fetal trabecular meshwork, cornea, and ciliary body tissues by RNA sequencing. Transm. Res. Int. 119, 67–75. doi: 10.1016/j.tfores.2019.01.061

Chu, M., Diele, F., Plemmons, R., and Ragni, S. (2004). Optimality, computation, and interpretation of nonnegative matrix factorizations. SIAM J. Matrix Anal. Appl. doi: 10.1.1.61.5758

Cong, L., Xu, T. X., and Wang, Q. (2016). Muscle competing fault prediction based on degradation data and fault data. J. Beijing Univ. Aeronaut. Astronaut. 42 (3), 522–531. doi: 10.13700/j.1001-5965.2015.0175

Deng, S., Wan, Z., and Chen, X. (2013). An improved spectral conjugate gradient algorithm for nonconvex unconstrained optimization problems. J. Optim. Theor. Appl. 157 (3), 820–842. doi: 10.1007/s10957-012-0239-7

Deng, J. L., Xu, Y. H., Wang, G., and Zhu, Y. S. (2019). Identification of potential crucial genes and key pathways in breast cancer using bioinformatic analysis. Front. Genet. 10, 695. doi: 10.3389/fgene.2019.00695

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jaffreuxjoum, M., Servant, N., et al. (2013). A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 14 (6), 671–683. doi: 10.1093/bib/bbs046

El-Shaqi, M. (2017). Dealing with small sample bias in post-crisis samples. Econ. Model. 65, 1–8. doi: 10.1016/j.econmod.2017.04.004

Floriou-Servou, A., von Ziegler, L., Stalder, L., Sturman, O., Privitera, M., Rassi, A., et al. (2018). Distinct proteomic, transcriptomic, and epigenetic stress responses in dorsal and ventral hippocampus. Biol. Psychiatry 84 (7), 531–541. doi: 10.1016/j.biopsych.2018.02.003

Gong, P., and Zhang, C. (2012). Efficient nonnegative matrix factorization via projected Newton method. Pattern Recogn. 45 (9), 3557–3565. doi: 10.1016/j.patcog.2012.02.037

Huang, S., Wan, Z., and Zhang, J. (2018). An extended nonmonotone line search technique for large-scale unconstrained optimization. J. Comput. Appl. Math. 330, 586–604. doi: 10.1016/j.cam.2017.09.026

Jiang, W., and Kalbfleisch, J. D. (2012). Boosting U-statistics: applications in least squares and robust regression. Sankhya B 74 (1), 56–76. doi: 10.1007/s13171-012-0043-2

Kaletsky, R., Yao, V., Williams, A., Runnels, A. M., Tadych, A., Zhou, S., et al. (2018). Transcriptome analysis of adult Caenorhabditis elegans cells reveals tissue-specific gene and isoform expression. PLoS Genet. 14 (8), e1007559. doi: 10.1371/journal.pgen.1007559

Li, W., Lai, D., Zhang, X., Cui, J., Hu, J., and Huang, T. S. (2010). The sterility of allotriploid fish and fertility of female autotriploid fish. Front. Genet. 10. doi: 10.3389/fgene.2019.00377

Li, X., Li, J., Zhao, Y., Zhou, Y., You, C., He, W., et al. (2012). Characteristics of diploid and allotriploid hybrids derived from female Megalobrama amblycephala ambycephala Yihxvale Xenocypris david Bleeker. Aquaculture 324, 126–134. doi: 10.1016/j.aquaculture.2012.08.025

Wan, Z., Huang, S., and Zhang, J. (2018). An extended nonmonotone line search technique for large-scale unconstrained optimization. J. Comput. Appl. Math. 330, 586–604. doi: 10.1016/j.cam.2017.09.026

Jiang, W., and Kalbfleisch, J. D. (2012). Boosting U-statistics: applications in least squares and robust regression. Sankhya B 74 (1), 56–76. doi: 10.1007/s13171-012-0043-2

Kaletsky, R., Yao, V., Williams, A., Runnels, A. M., Tadych, A., Zhou, S., et al. (2018). Transcriptome analysis of adult Caenorhabditis elegans cells reveals tissue-specific gene and isoform expression. PLoS Genet. 14 (8), e1007559. doi: 10.1371/journal.pgen.1007559

Li, W., Lai, D., Zhang, X., Cui, J., Hu, J., and Huang, T. S. (2010). The sterility of allotriploid fish and fertility of female autotriploid fish. Front. Genet. 10. doi: 10.3389/fgene.2019.00377

Li, W., Lai, D., Zhang, X., Cui, J., Hu, J., and Huang, T. S. (2010). The sterility of allotriploid fish and fertility of female autotriploid fish. Front. Genet. 10. doi: 10.3389/fgene.2019.00377

Lin, C., and Jiang, W. (2012). New adaptive Barzilai-Borwein step size and its applications in solving large scale optimization problems. ANZIAM J. 61 (1), 76–98. doi: 10.1017/S1446181118000263

Lin, C., and Jiang, W. (2012). New adaptive Barzilai-Borwein step size and its applications in solving large scale optimization problems. ANZIAM J. 61 (1), 76–98. doi: 10.1017/S1446181118000263

Lin, C. J. (2007). Projected gradient methods for nonnegative matrix factorization. Neural Comput. 19 (10), 2756–2770. doi: 10.1162/neco.2007.19.10.2756

Liu, S., Liu, Y., Zhou, G., Zhang, X., Luo, C., Feng, H., et al. (2001). The formation of tetraploid stocks of red crucian carp×common carp hybrids as an effect
## APPENDICES

### Results On Highly Expressed Genes

| GeneID | NO   | Elements in matrix $W$ for each class |
|--------|------|--------------------------------------|
| Mam27488 | 4285 | $8.04 \times 10^5$ 0 0 0 0 0 0.2813 |
| Mam12843 | 10739 | $5.91 \times 10^5$ 0 0 0 0 0 0.279 |
| Mam09635 | 15053 | $4.23 \times 10^4$ 0 0 0 0 0 0.1783 |
| Mam27746 | 1069 | 0 0.4815 0 0 0 0 0.2766 |
| Mam05349 | 1066 | 0 0.4789 0 0 0 0 0.1765 |
| Mam04721 | 1278 | 0 0.4775 0 0 0 0 0.1254 |
| Mam18643 | 1610 | 0 0.4789 0 0 0 0 0.1765 |
| Mam28461 | 1720 | 0 0.4815 0 0 0 0 0.2766 |
| Mam18947 | 2485 | 0 0.4815 0 0 0 0 0.2766 |
| Mam03075 | 2654 | 0 0.4815 0 0 0 0 0.2766 |
| Mam04110 | 2974 | 0 0.4815 0 0 0 0 0.2766 |
| Mam21839 | 3102 | 0 0.4815 0 0 0 0 0.2766 |
| Mam04628 | 3760 | 0 0.4815 0 0 0 0 0.2766 |
| Mam08966 | 4306 | 0 0.4815 0 0 0 0 0.2766 |
| Mam29639 | 4534 | 0 0.3065 0 0 0 0 0.2746 |
| Mam11009 | 4467 | 0 0.4815 0 0 0 0 0.2766 |
| Mam30659 | 5940 | 0 0.4815 0 0 0 0 0.2766 |
| Mam01072 | 6294 | 0 0.4815 0 0 0 0 0.2766 |
| Mam02487 | 7396 | 0 0.4815 0 0 0 0 0.2766 |
| Mam07898 | 7412 | 0 0.4815 0 0 0 0 0.2766 |
| Mam20311 | 7557 | 0 0.4815 0 0 0 0 0.2766 |
| Mam05748 | 7783 | 0 0.4815 0 0 0 0 0.2766 |
| Mam22143 | 8170 | 0 0.4815 0 0 0 0 0.2766 |
| Mam16193 | 8446 | 0 0.4815 0 0 0 0 0.2766 |
| Mam28424 | 8827 | 0 0.4815 0 0 0 0 0.2766 |
| Mam25840 | 9858 | 0 0.4771 0 0 0 0 0.1103 |
| Mam13519 | 10285 | 0 0.4815 0 0 0 0 0.2766 |
| Mam25865 | 11901 | 0 0.4815 0 0 0 0 0.2766 |
| Mam19044 | 12352 | 0 0.4815 0 0 0 0 0.2766 |
| Mam18831 | 12585 | 0 0.4815 0 0 0 0 0.2766 |
| Mam05543 | 13326 | 0 0.3065 0 0 0 0 0.2746 |
| Mam13771 | 13506 | 0 0.4815 0 0 0 0 0.2766 |
| Mam28554 | 13652 | 0 0.4815 0 0 0 0 0.2766 |
| Mam05777 | 13715 | 0 0.4792 0 0 0 0 0.1905 |
| Mam07942 | 14000 | 0 0.2847 0 0 0 0 0.2744 |
| Mam07030 | 14089 | 0 0.4789 0 0 0 0 0.1765 |
| Mam17634 | 14312 | 0 0.4815 0 0 0 0 0.2766 |
| Mam19307 | 14829 | 0 0.4815 0 0 0 0 0.2766 |
| Mam00814 | 15556 | 0 0.4779 0 0 0 0 0.1765 |
| Mam10384 | 15720 | 0 0.4815 0 0 0 0 0.2766 |
| Mam00295 | 16707 | 0 0.3065 0 0 0 0 0.2746 |
| Mam11738 | 16870 | 0 0.4815 0 0 0 0 0.2766 |
| Mam20672 | 17245 | 0 0.4815 0 0 0 0 0.2766 |
| Mam27740 | 18056 | 0 0.4815 0 0 0 0 0.2766 |
| Mam18996 | 18725 | 0 0.4815 0 0 0 0 0.2766 |
| Mam22493 | 19575 | 0 0.3065 0 0 0 0 0.2746 |
| Mam17452 | 19798 | 0 0.4815 0 0 0 0 0.2766 |
| Mam00511 | 19852 | 0 0.4815 0 0 0 0 0.2766 |
| Mam24132 | 6027 | 0 0 0 0 0.0342 0 0.2762 |
| Mam14897 | 6151 | 0 0 0 0 0.0342 0 0.2762 |
| Mam04754 | 6751 | 0 0 0 0 0.0342 0 0.2762 |
| Mam09228 | 7106 | 0 0 0 0 0.0342 0 0.2762 |
| Mam17991 | 8808 | 0 0 0 0 0.0342 0 0.2762 |
| Mam05763 | 9053 | 0 0 0 0 0.0342 0 0.2762 |
| Mam09996 | 10053 | 0 0 0 0 0.0342 0 0.2762 |
| Mam09632 | 10428 | 0 0 0 0 0.0342 0 0.2762 |
| Mam10304 | 12794 | 0 0 0 0 0.0342 0 0.2762 |
| Mam19016 | 13794 | 0 0 0 0 0.0342 0 0.2762 |
| Mam03189 | 16523 | 0 0 0 0 0.0342 0 0.2762 |
### TABLE A2 | The number of highly shared genes.

**Relationship among BSB\textsuperscript{L}, TC\textsuperscript{2L} and hybrids**

| Relationship     | Number | Relationship     | Number |
|------------------|--------|------------------|--------|
| 5th - 6th - 1st  | 107    | 5th - 6th - 2st  | 366    |
| 5th - 6th - 3st  | 58     | 5th - 6th - 4st  | 108    |

**Relationship between TC\textsuperscript{2L} and hybrids**

| Relationship     | Number | Relationship     | Number |
|------------------|--------|------------------|--------|
| 5th - 1st        | 0      | 6th - 1st        | 3      |
| 5th - 2nd        | 0      | 6th - 2nd        | 45     |
| 5th - 3rd        | 0      | 6th - 3rd        | 0      |
| 5th - 4th        | 0      | 6th - 4th        | 12     |
| 5th - 1st - 2nd  | 5      | 6st - 1st - 2nd  | 40     |
| 5th - 1st - 3rd  | 0      | 6st - 1st - 3rd  | 0      |
| 5th - 1st - 4th  | 0      | 6st - 1st - 4th  | 1      |
| 5th - 1st - 3rd  | 2      | 6st - 2nd - 3rd  | 13     |
| 5th - 2nd - 4th  | 587    | 6st - 2nd - 4th  | 125    |
| 5th - 3rd - 4th  | 0      | 6st - 3rd - 4th  | 2      |
| 5th - 1st - 2nd - 3rd | 16 | 6th - 1st - 2nd - 3rd | 27 |
| 5th - 1st - 2nd - 4th | 483 | 6th - 1st - 2nd - 4th | 168 |
| 5th - 1st - 3rd - 4th | 1 | 6th - 1st - 3rd - 4th | 6 |
| 5th - 2nd - 3rd - 4th | 229 | 6th - 2nd - 3rd - 4th | 88 |
| 5th - 1st - 2nd - 3rd - 4th | 2340 | 6th - 1st - 2nd - 3rd - 4th | 499 |

**Relationship among hybrids**

| Relationship     | Number | Relationship     | Number |
|------------------|--------|------------------|--------|
| 1st - 2nd        | 3      | 1st - 3rd        | 0      |
| 1st - 4th        | 0      | 2nd - 3rd        | 7      |
| 2nd - 4th        | 0      | 3rd - 4th        | 0      |
| 1st - 2nd - 3rd  | 4      | 1st - 2nd - 4th  | 277    |
| 1st - 3rd - 4th  | 0      | 2nd - 3rd - 4th  | 171    |

**Other relationship**

| Relationship | Number | Relationship | Number |
|--------------|--------|--------------|--------|
| 1st          | 0      | 2nd          | 0      |
| 3rd          | 0      | 4th          | 0      |
| 5th          | 0      | 6th          | 0      |
| 6th          | 0      | 1st - 2nd - 3rd - 4th | 194 |
| GeneID   | NO | Elements in base matrix \( W \) for each class |
|---------|----|----------------------------------|
| Mam01912 1 | \( 6.868 \times 10^{-4} \) | 0.0504 1.706 0.0201 0.3837 9.3883 \times 10^{-4} |
| Mam21118 5 | \( 6.835 \times 10^{-4} \) | 5.7525 \times 10^{-4} 0.2237 7.2198 \times 10^{-4} 0 0 |
| Mam11102 6 | 0.0455 | 0.2249 0.0129 0.0745 0 |
| Mam17081 7 | 0.0671 | 0.1641 0.0389 0.0488 1.3303 \times 10^{-4} |
| Mam07456 8 | 0.0398 | 0.0406 0.0509 0.3191 1.4597 \times 10^{-4} |
| Mam20030 9 | 0.0411 | 0.1378 0.0614 0.3605 1.7073 \times 10^{-4} |
| Mam09854 10 | 0.0300 | 0.2225 0.0552 0.1944 4.3024 \times 10^{-4} |
| Mam29205 13 | 0.0072 | 0.1055 0.0185 0.4066 2.0497 \times 10^{-5} |
| Mam06883 14 | 0.0530 | 0.0647 0.0470 0.3659 0 |
| Mam19604 15 | 0.0415 | 0.2180 0.0617 0.1979 0 |
| Mam09824 16 | 0.0105 \times 10^{-5} | 7.416 \times 10^{-4} 0 6.2839 \times 10^{-4} 0.3519 0 |
| Mam05355 18 | 2.1428 \times 10^{-4} | 0.0662 0.2270 0.0496 0.3109 1.1562 \times 10^{-4} |
| Mam19903 19 | 0.0264 | 0.0106 0.3837 6.1219 \times 10^{-4} |
| Mam23784 20 | 0.0626 | 0.1883 0.0376 0.3666 1.1244 \times 10^{-4} |
| Mam19985 21 | 6.0934 \times 10^{-4} | 6.7898 \times 10^{-4} 0.2239 8.5996 \times 10^{-4} 0 0 |
| Mam02753 22 | 0.0257 | 0.0451 0.151 0.3931 1.2711 \times 10^{-4} |
| Mam23187 23 | 1.2105 \times 10^{-5} | 7.416 \times 10^{-4} 0 6.2839 \times 10^{-4} 0.3519 0 |
| Mam05281 24 | 0.0628 | 0.1627 0.0500 0.1532 0 |
| Mam28834 25 | 0.0263 | 0.2251 0.0131 0.1040 0 |
| Mam23819 26 | 0.0668 | 0.1627 0.0500 0.1532 0 |
| Mam02726 29 | 1.2105 \times 10^{-5} | 7.416 \times 10^{-4} 0 6.2839 \times 10^{-4} 0.3519 0 |
| Mam11598 31 | 0.0149 | 0.0745 0.0141 0.3357 0 |
| Mam01497 32 | 0.0327 | 0.1177 0.0195 0.4134 5.8852 \times 10^{-4} |
| Mam03448 33 | 0.0313 | 0.0204 0.0512 0.1571 2.7127 \times 10^{-4} |
| Mam22869 35 | 0.0217 | 0.1089 0.0178 0.3928 1.1210 \times 10^{-4} |
| Mam20307 36 | 0.0200 | 0.0626 0.0269 0.3948 6.2672 \times 10^{-4} |
| Mam03780 37 | 0 | 0.0039 0 0.4115 2.4943 \times 10^{-4} |
| Mam23080 38 | 0.0583 | 0.0536 0.0456 0.0646 9.1700 \times 10^{-4} |
| Mam23255 42 | 0.0622 | 0.2187 0.0488 1.4495 \times 10^{-4} |
| Mam16330 44 | 0.0250 | 0.0786 0.0235 0.3616 6.7008 \times 10^{-5} |
| Mam27424 45 | 0 | 0.2247 0.0373 0.1112 0 |
| Mam2074 46 | 0 | 0.2247 0.0373 0.1112 0 |
| Mam09837 47 | 0.0404 | 0.1379 0 1.3783 \times 10^{-4} |
| Mam09179 49 | 0.0433 | 0.0725 0 3.250 2.1539 \times 10^{-4} |
| Mam11463 50 | 2.0732 \times 10^{-4} | 0.0538 0.0675 5.8679 \times 10^{-4} 0.0547 2.5670 \times 10^{-4} |
| Mam28606 51 | 0 | 0.0110 0 0.4514 7.0319 \times 10^{-4} |
| Mam05693 52 | 0 | 0.0126 0.0025 0 0.4030 1.4919 \times 10^{-4} |
| Mam20805 53 | 1.2105 \times 10^{-5} | 7.416 \times 10^{-4} 0 6.2839 \times 10^{-4} 0.3519 0 |
| Mam08145 54 | 3.1010 \times 10^{-4} | 0.0512 0.0897 0.0244 0.3974 9.3436 \times 10^{-4} |
| Mam26031 55 | 1.6025 \times 10^{-4} | 0.0060 0.0181 0.0211 0.4431 6.8717 \times 10^{-4} |
| Mam14647 56 | 5.9877 \times 10^{-5} | 2.2448 \times 10^{-4} 0.1418 0.0139 0.3587 8.0921 \times 10^{-4} |
| Mam28671 57 | 2.9578 \times 10^{-5} | 0.0222 0.1423 0.0630 0.2386 1.3530 \times 10^{-4} |
| Mam13555 58 | 5.4066 \times 10^{-5} | 0.0172 0.0934 0.0274 0.3661 3.3498 \times 10^{-4} |
| Mam28404 61 | 0.0254 \times 10^{-4} | 0.0110 0.0558 0 0.2177 1.4219 \times 10^{-4} |
| Mam28665 63 | 1.1933 \times 10^{-3} | 0.0170 0.1906 0.0160 0.4052 1.0805 \times 10^{-4} |
| Mam14143 64 | 3.3864 \times 10^{-4} | 0.0180 0.2079 0.0383 0.3853 0 |
| Mam18854 65 | 0 | 0.0034 0 0.4082 2.1172 \times 10^{-4} |
| Mam22835 66 | 4.7533 \times 10^{-4} | 1.3585 \times 10^{-4} 0.1844 0.0499 0.3343 1.7809 \times 10^{-4} |
| Mam05740 68 | 5.7847 \times 10^{-4} | 0.0658 0.1137 0.0435 0.3608 0 |
| Mam11399 69 | 4.5144 \times 10^{-4} | 0.0660 0.2113 0.0175 0.0761 0 |
FIGURE A1 | Shared highly expressed genes for the second-group samples.