Joint Optimal Classification and Pairing of Human Chromosomes

Joint Optimal Classification and Pairing of Human Chromosomes

By

Pravesh Biyani

B.Tech, EE Dept, IIT Bombay

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Masters of Applied Science

McMaster University

MASTERS OF APPLIED SCIENCE (2004) McMaster University
Electrical and Computer Engineering Hamilton, Ontario

TITLE Joint Optimal Classification and Pairing of Human Chromosomes

AUTHOR: Pravesh Biyani, B. Tech.
Indian Institute of Technology, Bombay, India (2002),
SUPERVISOR: Dr. Xiaolin Wu

NUMBER OF PAGES: x, 50

To my parents:

Amarchand and Shrikanta Biyani

Abstract

In this thesis, we reexamine the problems of computer-aided classification and pairing of human chromosomes. Traditionally researchers have dealt with the problem of classification and pairing separately. In our work, we propose to jointly optimize the solutions of these two very closely related problems. The combined problem is formulated into one of optimal three-dimensional assignment with an objective function of maximum likelihood. This formulation poses two technical challenges: 1. estimation of the posterior probability that two chromosomes form a pair and the pair belongs to a class, and 2. good heuristic algorithms to solve the three-dimensional assignment problem which is NP-hard. In our work, we present various techniques to solve these problems. We also generalize our algorithms to cases where the cell data are incomplete as often encountered in practice.

Acknowledgement

I am indebted to my advisor Professor Xiaolin Wu for his guidance and care. It is an honor as well as pleasure to work with him. Merely watching Dr.Wu's mind in action has been a wonderful experience. I thank him for instilling in me a passion for clarity of thoughts and expression.

I would also like to thank Dr. Abhijit Sinha, who helped me in all my difficulties, academic and otherwise. I stopped counting the frequency of knocking Abhijit's door long back. I will always remember the coffee-table-chats we had about everything under the sun.

I am grateful to Dr. Qiang Wu for his expert help and insights on various issues of the chromosome classification problem. Without his help, the project would not have been possible. I would also like to thank Dr. Kirubarajan, Dr. Shirani and Dr. Dumitrescu for their suggestions during the course of my thesis.

I would like to thank all my department mates, Zhe, Nima, Pouya, Cindy, Shawns, Maggy, Ehab, Sathyan, Bala and Ashish for their friendship, support and for making my stay an extremely pleasant experience.

Many thanks to my dear friend Roli for always lending her shoulders for me to cry.

Finally, I wish to acknowledge that without the unconditional love of my parents, nothing was possible. This thesis is dedicated to them.

Contents

List of Figures

List of Tables

List of Abbreviations

NN	Neural Networks
MLP	Multilayer Perceptron
3D	Three-dimensional

Chapter 1

Introduction

The past decade has seen a tremendous growth in the research and applications of Bioinformatics. Bioinformatics is a field involving the analysis and exploration of biological information using computer technology and statistical methods [1]. It has many applications in life sciences, medicine, and pharmaceutical industry [2]. In particular, the recent explosion in the data generated by the research and experiments in the fields of genetics and molecular biology has created pressing needs for sophisticated computational methods to extract meaningful information. Bioinformatics has thus become an important interdisciplinary field of close interaction between biosciences and information technology.

Chromosomes are genetic information carriers and human chromosome analysis constitutes an important procedure in the clinical and cancer cytogenetics studies, especially in the prenatal screening and genetic syndrome diagnosis [3]. One of the aims of chromosome analysis is the creation of a karyotype. Chromosome karyotyping refers to the classification and subsequently a formatted display of the chromosomes found in a cell. At certain stage of the life cycle of the cell, these chromosomes exist as separate bodies which, when appropriately stained by chemicals (e.g. giemsa, iodo-acetate) may be made visible under high resolution microscope. Figure 1.1[1] shows the image of a metaphase

[1]Courtesy Dr Q. Wu, Advanced Digital Imaging Research, Texas

stage cell spread in which the chromosomes have been stained by giemsa such that each exhibits a banding pattern (Giemsa-bands) along its length. Figure 1.2 shows the image of a karyotype of the chromosomes in Figure 1.1. Karyotype images are used in the clinical diagnostic tests, such as in amniocentesis, to determine if all the chromosomes appear normal and are present in the correct numbers. Abnormal cells can have an excess or deficit of a chromosome and can have structural defects too.

Figure 1.1: A Giemsa-stained metaphase cell spread

1.1 Chromosome and Karyotyping

There are generally 46 chromosomes in a normal human cell. Normal cells contain pairs or the homologues for chromosome classes $1 - 22$, which are called the autosomes, and the gender chromosome pair X and X for a female or X and Y chromosomes for a male. The chromosomes are characterized by what we call as chromosome features. The use of features rather than the image itself makes the karyotyping procedure easier, faster and

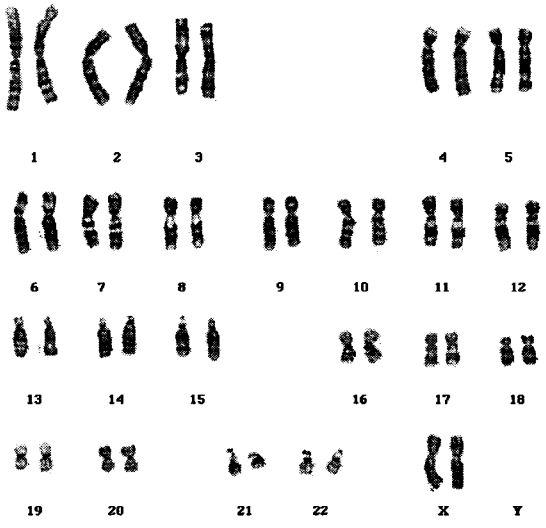


Figure 1.2: A karyotype of the chromosomes in Fig. 1

more accurate. The features are of two types: scalar like length, size, area, centromere index etc, and, vector like, banding patterns. The size can be measured either as the length or area of the chromosome, or the average of both, normalized by all the chromosomes in the cell. The banding patterns, on the other hand, are represented by the banding profiles, which are the projected profiles of chromosome image intensity averaged perpendicular to their medial axes.

Visually, the chromosome pairs or homologues, appear similar, that is the features would be close. This implies that the features like length, etc would be closer in magnitude for the paired chromosomes. Also, the chromosomes belonging to the same class will have similar shaped banding patterns. Similarity in the features can be noticed in Figure 1.2. Hence, in summary, the problem of chromosome karyotyping is to assign an identity (1−22, X,Y) to a chromosome, given a vector containing its grey level or intensity measurements (banding profiles) or other measured features (length, area etc) and some training vectors with known identities. A helpful additional output would include the degree of certainty

with which the chromosomes are identified and paired.

In many cytogenetic laboratories, the chromosome analysis and karyotyping is done manually. This process is long, repetitive, time consuming and hence expensive. This is the main motivation for the development of automatic chromosome classification techniques. Great efforts have been made to develop such automated karyopting techniques in the last 25 years. However, all of them have had a limited success and yield poor results compared to the trained cytotechnician [4]. One main reason for these relatively subgrade performance is the insufficient use of expert knowledge and experience. Another drawback of these methods is that the system always requires operator interaction to separate touching and/or overlapping chromosome and also to verify the classification results.

The next section describes the motivation behind the strategy used in our work and a brief idea of our technique used for the chromosome analysis.

1.2 Motivation and Methodology

Over the past decade a number of attempts have been made to develop efficient methods for automated karyotyping of human chromosomes. These efforts range from simple feature matching [5], Bayesian analysis, Markov Networks [6] and Neural Networks [4, 7, 8], the details of which will be explained in the chapter 2 of the thesis. However, during recent years, most of the work on automated chromosome classification has been based on the Neural Networks (NN) approach [4]. It has been shown to have achieved one of the best results in the chromosome classification analysis. In our work, we start with the conjecture that the neural network approach cannot guarantee the globally optimal solution to the classification problem. It is essentially an iterative process of classifying one chromosome or forming one homologue pair at a time. During this process the neural network method essentially disregards the interdependencies between the current classification decision and the past and future ones. Thus, the obvious weakness of such

approach is that it might only settle for a locally optimal solution.

The technique based on the transportation formulation was reported by Tsao et al [9]. As opposed to the NN, this framework achieves optimal classification. However, this approach was only for the normal somatic cells, in which all the 46 chromosomes are extracted from the image of metaphase and are presented to the classification algorithm. The transportation algorithm can find the globally optimal solution with respect to the given cost of assigning an chromosome to a class.

However, for incomplete cells (those with less than 46 chromosomes), the transportation algorithm cannot be applied directly. In chapter 3, we generalize the transportation algorithm based classification of human chromosomes for incomplete cells. However, we notice that this scheme does not use the fact that the features of the pairs are close, irrespective of the class they belong to. That is, there is a distinct correlation between the features of the paired chromosomes in a cell. Therefore, we look at the problem in a new perspective of pairing rather than classification. In chapter 3, we introduce the joint pairing and the classification framework based on the complete and the bi-partite graph matching. This approach aims at exploiting the similarity or the correlation between the actually paired chromosomes. It is essentially a two step process in which pairing is done in the first step using the maximum weight graph matching and then the pairs are classified according to their joint classification probability using the bi-partite graph matching.

The outline of the remaining part of the thesis is as follows. Chapter 4 presents the main contribution of the work, in which we first identify two inherent drawbacks of chromosome classification by the transportation based approach. The first drawback in this approach is the disregard of correlations between features of chromosomes in the objective function of the transportation algorithm, and the other is the fact that it doest not use the feature similarities within a cell. We then propose, as a natural cure to these weaknesses, the approach of three-dimensional assignment that optimizes chromosome classification

and pairing simultaneously. The optimal solution to the above discrete optimization problem gives the pairing and classification simultaneously. This new approach can also make use of joint statistics which were neglected by previous methods. In chapter 4, we also deal with the problem of estimating posterior probabilities that are required by the 3D assignment algorithm. The 3D assignment problem is NP-hard in nature. Hence, we have to devise a heuristic strategy to approach the optimal solution of the problem. We propose two heuristic solutions to the problem, first based on the lagrangian relaxation algorithm and the other based on the intelligent search in the feasible solution set. The better performance of the proposed methods in practice is verified by experimental results that are presented in chapter 5. The results clearly demonstrate the strength of the new formulation. Chapter 6 of the thesis concludes the work with a discussion on the scope of the future work in classification and pairing of human chromosomes.

Chapter 2

Background and Related Work

Over the last three decades, there has been an extensive research on the automated chromosome analysis and karyotyping. The earliest attempt to develop algorithms for the task was reported in 1973 by Castleman and Wall [10]. In 1980, Ledley et. al. [11] set a requirement of 90% correct classification as a performance criterion for an algorithm to be of clinical value. They used the discriminant analysis technique for the classification of Giemsa banded preparations. The banded density profiles were subjected to fourier analysis, and the first 9 terms of the fourier sine and cosine series were used as features for a discriminant analysis training set. These authors were able to achieve a correct classification rate of 86.1% on the training set. Other authors used the same method but used higher order harmonics, as high as 25 terms [10, 11]. This approach encountered the problems of differential contraction in the chromosomes and the resulting influence on the fundamental frequency and all higher order harmonics. Template matching was used by Neurath et. al. [5] to classify human G-banded chromosomes. This technique used preassigned templates to match the chromosomes and best match was used for the classification. Through this technique the researchers tried to replicate the cytogeneticist's method and performance.

Other old approach to solve the classification problem includes the matching of a

template banding profile with the given chromosome banding profile and estimating the correlations. [13]. It proved to be useful for the study of a small series of human chromosomes [5]. In 1980, Vanderheydt et. al. [14] developed an algorithm based on a split and merge procedure used to describe the chromosome profile in a tree structure. By defining supplementary branches between the tree nodes, i.e. nodes which correspond with the black and bright intervals of the profile, a graphical description of the chromosome was generated. Fuzzy set theory was used to interpret this graph. An operators aggregation structure was applied for the interpretation of chromosomes. This method also achieved limited success.

The use of neural networks for chromosome classification was suggested in the early 90's. The neural networks approach held the promise of overcoming most limitations of the older methods for chromosome classification. This is because NN's easily trainable and are capable of creating arbit partitions in the feature space.

In 1993, Granum and Thomason in [6] used the Markov networks model for the classification of chromosomal banding pattern structure. In their paper, the chromosome banding pattern was treated like a symbol string and features were computed through string-network alignment computation of a similarity measure. Thus, a feature vector is formed using this similarity measure from different classes and is used for the pattern recognition. In the training phase, a Markov network is constructed from a set of learning strings for each pattern class using data driven interference. Subsequent classification of test strings is carried out by aligning the string with its candidate networks and using the resulting similarity measure as the input to the maximum-likelihood and nearest-neighbor classifiers. The above method has two limitations. First, Markov models are complex. There are many parameters which must be incorporated into the model, such as transition probabilities and substring probabilities to specific classes. These parameters detract from the direct comparison, which is made between chromosomes within the same cell. Second, the technique did not use scalar features, hence it was suboptimal.

Sweeney et. al. proposed probabilistic neural network for chromosome classification in 1994 [17]. Lerner et. al. proposed Multilayer Perceptron based Neural Network model for the same task [4]. The authors argue that neural network can only give results as good as transportation algorithm if the cells are complete (that is each class has exactly two chromosomes). They propose a MLP neural network based scheme for the incomplete cells. All the approaches involving neural networks are inherently sub-optimal as the neural network approach cannot guarantee the globally optimal solution to the classification problem. It is essentially an iterative process of classifying one chromosome or forming one homologue pair at a time, disregarding the interplays between the current classification decision and the past and future ones. The weakness of such an approach is that it can only settle for a local optimum.

In summary, all the previous approaches reviewed so far share two main characteristics

1. All of them considered classification as the principle problem and regarded pairing decisions as the implicit by-product of classification.

2. Very few of them used all the available features in one framework. All the neural networks approaches ignored the dependency of the classification and pairing decisions between two chromosomes and hence only attained a local optimal solution.

In this thesis, we take care of both of the above issues.

Chapter 3

Optimal Classification and Pairing of Chromosomes: Graph-Matching and Transportation Approach

In this chapter, we introduce the two approaches to find the globally optimal solution for the chromosome classification and the chromosome pairing problem. The first approach is based on the transportation problem. Tso et. al. [9] used the transportation algorithm to solve the problem. For normal somatic cells (all the 46 chromosomes are extracted from the image of metaphase spread and presented to the classification algorithm), the transportation algorithm can find the globally optimal solution in maximum likelihood sense, with respect to given estimated posterior probabilities of chromosomes being in different classes.

In this chapter, we are interested in classification of chromosomes from either complete or incomplete cells. The problem of incomplete data cell often occurs when two or more chromosomes are overlapped and therefore unsuitable for karyotyping analysis, or chromosomes are missing due to abnormalities or specimen preparation artifacts. In this case the original transportation algorithm does not apply, because the number of

chromosomes in a given class is not a known prior. The focus of this chapter is on algorithmic approach to finding the globally optimal solution of chromosome classification and pairing. We first show that chromosome classification with incomplete cell data can also be formulated as a transportation problem, just like in the case of complete cell data classification where each class has exactly two chromosomes. This allows the globally optimal chromosome classification to be computed in polynomial time. Then, we turn to the problem of homologue pairing. Although homologue pairing can be a by-product of chromosome classification, the pairing problem is worth investigating in its own right. Within a given cell homologue pairing can be performed using the feature measurements of the concerned cell, whereas chromosome classification relies on statistics drawn from a training set consisting of a large number of cells of different individuals. Considering that many features of a chromosome such as the length, area, centromeric ratios, have less variations within a cell than between different cells, it is possible to obtain more robust homologue pairing result from the features of a given cell than the pairing derived from the output of the transportation algorithm. To this end we develop a homologue pairing technique of maximum-weight graph matching. This technique forms all homologue pairs simultaneously under a maximum likelihood criterion, hence offers a globally optimal solution to the problem, as opposed to the locally optimal solutions of greedy and neural network algorithms. Furthermore, upon completing homologue pairing, chromosomes classification can be done by maximum-weight graph matching as well. This new approach may produce better classification results in some cases. The next section describes the classification using the transportation problem formulation. The second and the third section deals with the graph-matching based approach for the pairing and classification respectively.

3.1 Globally Optimal Classification

We abstract a cell as a set of objects X_i, $1 \leq i \leq N$, $N \leq 46$. Each of the objects corresponds to a chromosome in the cell. Note that the subscripts i are just an arbitrary indexing of objects that do not necessarily relate to the class labels of the chromosomes. As mentioned before, an object (chromosome) X_i is characterized by a feature vector $\mathbf{f}_i = (f_1, f_2, \cdots, f_n)$. The features f_t, $1 \leq t \leq n$, are of two types: scalar features such as chromosome size, length, intensity, centromeric ratios, the number of bands in the banding profile, and etc., and the vector feature that is the banding profile of the chromosome. The objects can be classified into K classes. The value of K is 23 for a male cell and 24 for a female cell. Each object belongs to one class, C_k. For normal somatic cells (i.e., $N = 46$), each class C_k, $1 \leq k \leq 22$, has exactly two objects. The gender class C_{23} has one or two objects for female or male respectively. Likewise, the class C_{24} has one object for female, or is empty for male.

For the problem of chromosome classification from incomplete cell data, each cell has less than 46 objects. (i.e., $N < 46$). This calls for the relaxation in the constraint on the number of chromosomes in each class and thus allowing inequalities. Specifically, let us define an $N \times K$ binary matrix M whose elements are either 0 or 1. The rows of the matrix correspond to the objects and the columns to the classes. The matrix has following properties,

$$\sum_{1 \leq k \leq K} M_{ik} = 1, 1 \leq i \leq N; \sum_{1 \leq i \leq N} M_{ik} \leq 2, 1 \leq k \leq 22;$$

$$\begin{aligned} \sum_{1 \leq i \leq N} M_{i,23} &= \begin{cases} \leq 2 & male \\ \leq 1 & female \end{cases} \\ \sum_{1 \leq i \leq N} M_{i,24} &= \begin{cases} = 0 & male \\ \leq 1 & female \end{cases} \end{aligned} \tag{3.1}$$

Each row of the matrix M has exactly one element being 1 and the rest being 0,

because each chromosome belongs to only one class. Columns 1 through 22 of the matrix have at most two elements equal to 1, corresponding to the possibilities that a class has zero, one, or two chromosomes when some chromosomes are missing. Clearly, matrix M with the above properties corresponds to a possible classification of N chromosomes into K classes.

Let $P(X_i \in C_k|\mathbf{f}_i)$ be the posterior probability that $X_i \in C_k$ given the feature vector $\mathbf{f}_i$. Then by the very nature of the chromosome classification, we formulate the problem of optimal classification of chromosomes as one of maximum likelihood estimation:

$$M_{opt} = \arg\max_{M \in S} \prod_{i=1}^{N} \sum_{k=1}^{K} M_{ik} P(X_i \in C_k|\mathbf{f}_i), \tag{3.2}$$

where S is the set of all classification matrices M. By taking logarithm, (3.2) becomes equivalent to

$$M_{opt} = \arg\max_{M \in S} \sum_{i=1}^{N} \log \sum_{k=1}^{K} M_{ik} P(X_i \in C_k|\mathbf{f}_i) \tag{3.3}$$

subject to the constraints (A.1). Since for each given i only one M_{ik} is non-zero, (3.3) is equivalent to

$$M_{opt} = \arg\max_{M \in S} \sum_{i=1}^{N} \sum_{k=1}^{K} M_{ik} \log P(X_i \in C_k|\mathbf{f}_i) \tag{3.4}$$

which is clearly an integer programming problem.

In general, integer programming problem is NP-hard. However, by a second reflection, we can convert the problem (3.4) into one of transportation. Now let us introduce $46 - N$ dummy objects X_i and let $P(X_i \in C_k|\mathbf{f}_i) = \frac{1}{K}$, for $N + 1 \leq i \leq 46$, $1 \leq k \leq K$. The classification matrix for the complete set of objects is $\bar{M}$ obtained by augmenting M with $46 - N$ rows. The optimal $\bar{M}_{opt}$ is obtained by solving the transportation problem

$$\bar{M}_{opt} = \arg\max_{\bar{M} \in \bar{S}} \sum_{i=1}^{46} \sum_{k=1}^{K} \bar{M}_{ik} \log P(X_i \in C_k|\mathbf{f}_i) \tag{3.5}$$

subject to

$$\sum_{1\leq k\leq K} \bar{M}_{ik} = 1, 1 \leq i \leq 46; \sum_{1\leq i\leq 46} \bar{M}_{ik} = 2, 1 \leq k \leq 22;$$

$$\sum_{1\leq i\leq 46} \bar{M}_{i,23} = \begin{cases} = 2 & male \\ = 1 & female \end{cases}$$
$$\sum_{1\leq i\leq 46} \bar{M}_{i,24} = \begin{cases} = 0 & male \\ = 1 & female \end{cases} \quad (3.6)$$

Note that if $\sum_{i=1}^{46}\sum_{k=1}^{K} \bar{M}_{ik} \log P(X_i \in C_k|\mathbf{f}_i)$ is maximal, then $\sum_{i=1}^{N}\sum_{k=1}^{K} M_{ik} \log P(X_i \in C_k|\mathbf{f}_i)$ is maximal, too.This is because

$$\sum_{i=1}^{46}\sum_{k=1}^{K} \bar{M}_{ik} \log P(X_i \in C_k|\mathbf{f}_i) = (46 - N) \log K^{-1} + \sum_{i=1}^{N}\sum_{k=1}^{K} M_{ik} \log P(X_i \in C_k|\mathbf{f}_i).$$

Consequently, M_{opt} (i.e. the globally optimal chromosome classification) can be found by solving the transportation problem (3.5), which can be done in $O(N^2Klog(N)+N^2log^2N)$ time. In our work, the transportation problem was solved by using Bertsekas's auction algorithm [24]. Detailed discussion of the above approach is presented in the appendix A of the thesis. The remaining challenge is how to estimate the posterior probability $P(X_i \in C_k|\mathbf{f}_i)$. The details of the estimation of the above classification probabilities will be presented in the chapter 4 of the thesis.

3.2 Homologue Pairing by Maximum-weight Graph Matching

In this section, we investigate the problem of automated formation of homologue pairs of a cell. The chromosome pairing and classification are closely related. Indeed, the transportation algorithm for chromosome classification also implicitly solves the problem of homologue pairing. The accuracy of homologue pairing by the transportation algorithm

depends on the discriminating power of the estimated posterior probability $P(X_i \in C_k|\mathbf{f}_i)$. Note that $P(X_i \in C_k|\mathbf{f}_i)$ needs to be estimated necessarily from a training set of a large number of different cells. The feature vector $\mathbf{f}_i$ for a chromosome can vary significantly among different types and even among the same type of chromosomes from different cells of the same individual. In contrast such variations are of much lesser degree between two chromosomes of a given class within a cell (i.e. the homologues). Inter-cell variations are much greater than within-cell variations because practically different cells are imaged at somewhat different stages of mitosis. These observations lead us to a new and more robust paradigm for chromosome pairing and classification.

From the perspective above more accurate homologue chromosome pairing can be obtained by exploiting the similarities between the two homologue chromosomes of a cell. This problem naturally induces a binary relation $\cong$ such that $X_i \cong X_j$ if and only if two chromosomes X_i and X_j form a homologue pair. This binary relation among chromosomes is exhibited by their feature vectors $\mathbf{f}_i$, $1 \leq i \leq N$. Our task is to statistically infer the binary relation $\cong$ among all chromosomes from all observable feature vectors.

Any chromosome pairing can be identified with a permutation function $\tau(i)$ of integers $i \in \{1, 2, \cdots, N\}$, where N is a positive even integer and $\tau(i) = j$ if and only if the chromosomes X_i and X_j form a pair. Clearly, $\tau(i) = j$ if and only if $\tau(j) = i$, moreover $\tau(i) \neq i$, for all i and j. Let $P(X_i \cong X_j|\mathbf{f}_i\mathbf{f}_j)$ be the posterior probability that $X_i \cong X_j$ given the features of X_i and X_j. Then the problem of optimal pairing of chromosomes can be formulated as one of maximum likelihood estimation:

$$\tau_{opt} = \arg\max_{\tau} \prod_{i=1}^{N} P(X_i \cong X_{\tau(i)}|\mathbf{f}_i\mathbf{f}_{\tau(i)}), \tag{3.7}$$

over all pairings τ. By taking logarithm of (4.1), we have equivalently,

$$\tau_{opt} = \arg\max_{\tau} \sum_{i=1}^{N} \log P(X_i \cong X_{\tau(i)}|\mathbf{f}_i\mathbf{f}_{\tau(i)}). \tag{3.8}$$

In order to solve the optimization problem of (3.8), let us construct a graph G with N

nodes, one node for each chromosome X_i, and with $\binom{N}{2}$ edges, one edge for each pair of X_i and X_j. The weight of such an edge is $\log P(X_i \cong X_j|\mathbf{f}_i\mathbf{f}_j)$. It is easy to see that any pairing τ corresponds to a set of $N/2$ edges which cover all N vertices once and only once. In other words, M represents a match of vertices of the graph G. Now the maximization problem of (3.8) becomes one of maximum-weight graph matching.

This graph problem can be solved in $O(NZ \log N)$ time where N is the number of vertices [19] and Z is the number of edges in the graph. In our case $Z = O(N^2)$. However, we can reduce Z to $O(N)$, and hence reduce the complexity to $O(N^2 \log N)$ without affecting the optimality of the solution by deleting edges whose weights are too small.

We stress that given the posterior probability $P(X_i \cong X_j|\mathbf{f}_i\mathbf{f}_j)$ for pairing, the maximum-weight graph matching can solve the maximum likelihood estimation problem of (3.8) exactly over all possible pairings. The validity of the pairing results depends on the quality of the estimated posterior probability $P(X_i \cong X_j|\mathbf{f}_i\mathbf{f}_j)$.

3.3 Classification via Bipartite Graph Matching

Once the chromosomes of a cell are paired by the maximum-weight graph matching algorithm, we can proceed to classify homologue pairs. Consider two paired chromosomes $\mathbf{a}_t = (a_t, \grave{a}_t)$ of features $\mathbf{F}_t$ (the concatenation of the features of a_t and $\grave{a}_t$). Denote by $P(\mathbf{a}_t \in C_k|\mathbf{F}_t)$ the probability that the pair $\mathbf{a}_t$ belongs to chromosome class C_k given $\mathbf{F}_t$. We take the maximum likelihood approach to classify homologue pairs by solving the following optimization problem: $\max_\pi \prod_{t=1}^{K} P(\mathbf{a}_t \in C_{\pi(t)}|\mathbf{F}_t)$, which is equivalent to

$$\max_{\pi} \sum_{t=1}^{K} \log P(\mathbf{a}_t \in C_{\pi(t)}|\mathbf{F}_t), \tag{3.9}$$

where K is the number of homologue pairs and $\pi(i)$ denotes the permutation function of integers $t \in \{1, 2, \cdots, K\}$. Consider the bipartite graph whose two vertex sets are A (the set of all homologue pairs of a given cell) and B (the set of all chromosome class labels).

The edges of the graph connect any node $\mathbf{a}_t \in A$ to any node $b \in B$. The weight of such an edge is $\log P(\mathbf{a}_t \in C_b|\mathbf{F}_t)$. Then the optimization problem (3.9) is equivalent to finding the maximum-weight match in this graph.

In order to estimate $P(\mathbf{a}_t \in C_{\pi(t)}|\mathbf{F}_t)$, we assume that the random events $a_t \in C_\pi(t)$ and $\grave{a}_t \in C_\pi(t)$ are independent of each other for a given $\mathbf{F}_t$, thus

$$P(\mathbf{a}_t \in C_k|\mathbf{F}_t) = P(a_t \in C_k|\mathbf{F}_t)P(\grave{a}_t \in C_k|\mathbf{F}_t).$$

The complexity of weighted bipartite graph matching is $O(N^3)$. As mentioned in the previous section we can reduce the complexity to $O(N^2 \log N)$ by deleting edges of small weights. In contrast, the transportation algorithm has a complexity of $O(N^3 \log N + N^2 \log^2 N)$, which is higher than the graph matching method.

Chapter 4

Joint Classification-Matching of Chromosomes

In the previous chapter, we described how transportation problem formulation can be used for the classification of chromosomes even for incomplete cells. We also described an application of a graph theoretical technique to solve the same problem. However, we realize that both of the above techniques address only a particular aspect of the problem and therefore do not achieve optimum performance in the true sense. In this chapter, we reformulate the classification problem and propose a new technique of joint classification and matching via optimal three dimensionsional (3D) assignment. This NP hard problem has two major aspects. First is the estimation of joint classification probabilities, that is the measure of the cost matrix and other is the algorithmic approach to the optimal solution. Two strategies to reach the optimal solution is presented in this chapter. We also devise a novel probability estimation framework that simultaneously uses all the dependent and correlated features. This probability measure can also be be directly used for the previously mentioned transportation and graph matching based classification algorithms too.

The outline of this chapter is as follows. The first section describes the motivation for

the above problem formulation. The next section develops the objective function for the joint optimization formulation. In section three we describe the heuristics to approach the optimal solution. Finally, the fourth section describes the strategy for the estimation of the joint classification probabilities.

4.1 Motivation

The chromosome classification problem can be treated as one of transportation problem. This approach was shown to be optimal for the complete cells, given the posteriori probabilities. We generalize the same scheme for the classification of even incomplete cells. In the transportation technique, the chromosomes were classified by assigning them to the classes in a maximum likelihood sense. The above task was modelled as an integer programming problem and the optimal solution was obtained by the transportation algorithm.

The second approach used was based on the maximum-weight graph matching. It is essentially a two step process. In the first step, it forms all the homologue pairs simultaneously under a maximum likelihood criterion of pairing. That is, it uses the cost or the distance between the two chromosomes and obtains the pairs which minimize the total sum of the pair-wise distances. After the optimal homologue pairing, chromosome classification is done using the maximum-weight bi-partite graph matching algorithm. This new approach to chromosome pairing and classification may be more robust than the transportation algorithm, because many attributes of a chromosome have less variations within a cell than between different cells. The strength of this algorithm is that it covers both the problems of pairing and classification. But its performance is limited as both the problems of classification and pairing are not tackled simultaneously. As we shall see further that the classification and pairing problem are interrelated.

Chromosomes in a cell have two main characteristics. The first is that within the cell

the two chromosomes of a pair have similar features. The second is that the features $\mathbf{f}_i$ of the individual chromosome X_i are within the expected range of variation for normal chromosomes of the class that X_i actually belongs to. The transportation algorithm based scheme utilizes the second property and ignores the first property. On the other hand, the graph matching based approach exploits the first property, but misses classification information given by the second property. Thus neither of the scheme is optimal because it does not exploit all the available information and pre-knowledge.

To overcome the drawback of the previous two techniques for chromosome classification and pairing, we propose a new scheme which combines both the transportation and the graph matching algorithms. The motivation is to use both the above mentioned properties and perform classification and pairing simultaneously. In the next section we will formulate the approach as an optimization problem.

4.2 Problem Formulation

Following the same convention as in the previous chapter, lets abstract a cell as a set of objects X_i, $1 \leq i \leq N$, each of which corresponds to a chromosome in the cell. We also know from the previous discussion that the objects can be classified into K classes. An object (chromosome) X_i is characterized by a feature vector $\mathbf{f}_i = (f_1 f_2 ... f_n)$.

We formulate the problem for the case where each class has exactly two objects. Latter, we will show that this formulation can be generalized for the cases when the cells are incomplete by adding dummy variables. The philosophy of adding dummy variables is the same as we did for the incomplete cell with regards to the transportation algorithm.

Let $P(X_i, X_j \in C_k | \mathbf{f}_i, \mathbf{f}_j)$ be the posteriori probability that $X_i \in C_k$ and $X_j \in C_k$ given the feature vectors $\mathbf{f}_i$ and $\mathbf{f}_j$. As we did in the previous chapter, we again formulate the problem as one of the maximum likelihood estimation. Consider all possible three-way assignments of two individual chromosomes and a class, denoted by the three-tuple

(X_i, X_j, C_k), such that chromosomes X_i and X_j are assigned as a homologue pair and the pair is assigned to the class C_k, and furthermore each of the K classes is assigned exactly two chromosomes. By introducing a binary assignment variable $a_{i,j,k}$ such that $a_{i,j,k} = 1$ if (X_i, X_j, C_k) is selected and $a_{i,j,k} = 0$ otherwise, we can cast joint chromosome classification and pairing as the following optimization problem:

$$\max_{a_{i,j,k}} \prod_{k=1}^{K} \sum_{j=1}^{N} \sum_{i=1}^{N} a_{i,j,k} P(X_i, X_j \in C_k | \mathbf{f}_i, \mathbf{f}_j), \tag{4.1}$$

The objective function aims to maximize the likelihood that chromosomes X_i and X_j are a pair and the pair belongs to the class C_k over all possible assignments $a_{i,j,k}$ satisfying the constraints given below. By taking the negative logarithm of (4.1), we convert the problem into a minimization problem

$$\min_{a_{i,j,k}} \sum_{k=1}^{K} \sum_{j=1}^{N} \sum_{i=1}^{N} -log(P(X_i, X_j \in C_k | \mathbf{f}_i, \mathbf{f}_j) a_{i,j,k} \tag{4.2}$$

subject to

$$\begin{aligned}
&\textstyle\sum_{j=1}^{N} \sum_{i=1}^{N} a_{i,j,k} = 2, \quad k = 1, 2, \ldots K \\
&\textstyle\sum_{k=1}^{K} \sum_{i=1}^{N} a_{i,j,k} = 1, \quad j = 1, 2, \ldots N \\
&\textstyle\sum_{k=1}^{N} \sum_{j=1}^{N} a_{i,j,k} = 1, \quad i = 1, 2, \ldots N
\end{aligned}$$

The above optimization problem is known as optimal 3D assignment in the discrete optimization literature [20]. The above formulation not only facilitates the joint optimization of chromosome classification and homologue pairing, it also overcomes an inherent drawback of the transportation and other existing algorithms for chromosome classification in their way of using sample statistics. Consider the event $(X_i \in C_k | \mathbf{f}_i)$ that chromosome X_i belongs to class C_k given $\mathbf{f}_i$. We notice that the two events $(X_i \in C_k | \mathbf{f}_i)$ and $(X_j \in C_k | \mathbf{f}_j)$ are not independent. This dependency is attributed to the presence

of correlation between the features $\mathbf{f}_i$ and $\mathbf{f}_j$ of X_i and X_j in a given cell. In the transportation algorithm for chromosome classification, however, the independence between $(X_i \in C_k|\mathbf{f}_i)$ and $(X_j \in C_k|\mathbf{f}_j)$ is assumed, because $P(X_i, X_j \in C_k|\mathbf{f}_i, \mathbf{f}_j)$ is written in the product form $P(X_i \in C_k|\mathbf{f}_i) \cdot P(X_j \in C_k|\mathbf{f}_j)$ when formulating the maximum likelihood estimation problem as we recall from (3.2). The invalid assumption of statistical independence between $(X_i \in C_k|\mathbf{f}_i)$ and $(X_j \in C_k|\mathbf{f}_j)$ limits the classification performance of the transportation and other existing algorithms. This drawback is overcome by the approach of 3D assignment which uses the true joint probability $P(X_i, X_j \in C_k|\mathbf{f}_i, \mathbf{f}_j)$. The above formulation poses two challenges: the estimation of the probability $(P(X_i, X_j \in C_k|\mathbf{f}_i, \mathbf{f}_j)$ and the algorithm to minimize the objective function.

The above 3D assignment problem is NP hard [21]. Therefore, an optimal algorithm to solve the problem would take enormous CPU time and memory even for small N and K. Thus, reaching the optimal solution in finite computational time is impractical. The next section illustrates the various heuristics we adopted in our work to solve the optimization problem.

4.3 Solution to the 3D assignment Problem

Since, the three dimensional minimization problem in equation 4.2 is NP hard, it is natural to think of suboptimal and heuristic approaches to the solution. Ideally, such approach should not take a huge CPU computation time, or in other words has a run time which is approximately bounded by a polynomial function of the number of objects N and/or the number of classes K.

There are several heuristic techniques reported in the literature to solve the 3D assignment problem [22]. These methods breaks the 3D assignment problem into a number of series of 2D problems and subsequently finds the optimal solution to those 2D assignment problems. This approach has been used in various other fields ranging from target

tracking, image processing to the operations research [23]. There are a number of good techniques to solve the 2D assignment problem optimally, for e.g, auction algorithm [24] and bi-partite graph matching. In our work, we use the auction algorithm to solve the 2D assignment problems. The detailed discussion of the 2D assignment problem and the auction algorithm is discussed in appendix A.

To solve the 3D assignment problem, the authors in [22] used lagrangian relaxation techniques, by relaxing constraints one set a time and incorporating it into the cost function via lagrangian multipliers and minimized the cost of the relaxed 2D assignment problem. One nice feature of the lagrangian relaxation technique is that in addition to a good solution we can also have the measure of the quality of the solution. In [23], the randomized heuristic approach is developed, in which, in each step one of the possible solution is chosen randomly depending on the weight associated with it and the solution is iteratively improved. In our work, we also develop a new heuristic algorithm to solve the 3D assignment problem. It finds a subset of the feasible solution set and an exhaustive search is performed on the subset. Hence, we call this method as the semi-exhaustive search for the solution. Next subsections illustrates the techniques.

4.3.1 Lagrangian Relaxation Algorithm

One way to solve the 3D assignment problem is by the use of the lagrangian relaxation technique. This method breaks the 3D problems as the series of 2D problems and successively solve them to attain a nearly optimal solution for the 3D problem. This technique was first reported to solve a similar assignment problem with a different set of constraints in [22]. We modify their approach to suit our requirements. The basic idea is to relax one set of constraints in equation 4.1 and incorporate it into the cost function with the lagrangian multiplier vector $\mathbf{u} = [u_0, u_1, \cdots, u_N]$. Thus, the new objective function represents a 2D assignment problem with original constraints relaxed. Since, the constraints are relaxed, the dual solution would give a lower bound on the optimal cost. Hence, the

corresponding primal feasible cost will be an upper bound to the optimal cost. The algorithm proceeds by updating the dual until the above duality gap decreases till a predefined level. Next we present the brief summary of steps in the relaxation algorithm.

We first assign the unconstrained lagrangian multipliers, $u_i, i = 1, 2, \ldots, N$ to the third set of constraints. Therefore, the dual function $L(\mathbf{u})$ is

$$L(\mathbf{u}) = \min_{a_{i,j,k}} \sum_{k=1}^{K} \sum_{j=1}^{N} \sum_{i=1}^{N} -log(P(X_i, X_j \in C_k | \mathbf{f}_i, \mathbf{f}_j) a_{i,j,k} + \sum_{i=1}^{N} u_i (1 - \sum_{k=1}^{N} \sum_{j=1}^{N} a_{i,j,k}) \quad (4.3)$$

The above equation can be rewritten as,

$$L(\mathbf{u}) = \min_{a_{i,j,k}} \sum_{k=1}^{K} \sum_{j=1}^{N} \sum_{i=1}^{N} (-log(P(X_i, X_j \in C_k | \mathbf{f}_i, \mathbf{f}_j) - u_i) a_{i,j,k} + \sum_{i=1}^{N} u_i \quad (4.4)$$

subject to

$$\sum_{j=1}^{N} \sum_{i=1}^{N} a_{i,j,k} = 1, \quad k = 1, 2, \ldots K$$
$$\sum_{k=1}^{K} \sum_{i=1}^{N} a_{i,j,k} = 2, \quad j = 1, 2, \ldots N$$

Let $\sum_{i=1}^{N} a_{i,j,k} = \omega_{jk}$. Then, the constraints in equation(4.4) are equivalent to

$$\sum_{j=1}^{N} \omega_{jk} = 1, \quad k = 1, 2, \ldots K$$
$$\sum_{k=1}^{K} \omega_{jk} = 2, \quad j = 1, 2, \ldots N$$

and the objective function equation(4.4) can be modified as

$$L(\mathbf{u}) = \min_{\omega_{jk}} \sum_{k=1}^{K} \sum_{j=1}^{N} d_{jk} \omega_{jk} + \sum_{i=1}^{N} u_i \quad (4.5)$$

where,

$$d_{jk} = \min_i((-log(P(X_i, X_j \in C_k|\mathbf{f}_i, \mathbf{f}_j) - u_i) \tag{4.6}$$

.

This is a 2D assignment problem whose minimization is subject to the remaining constraints in equation 4.4. Therefore, for a given dual vector $\mathbf{u}$, we can convert the 3D assignment problem into an equivalent 2D problem. The rationale for relaxing the third set of constraints is as follows. Both the third and second sets of constraints enforce pairing and are symmetric, hence the two are equivalent when included in the Lagrangian objective function. The first set of constraint enforces the structure that one class has two chromosomes. It is more important to impose the first set of constraints because $P(X_i, X_j \in C_k|\mathbf{f}_{i,2}, \mathbf{f}_{j,2})$ is more biased than $P(X_i \cong X_j|\mathbf{f}_i, \mathbf{f}_j)$. In other words, the statistics for classification has higher discriminating power than the statistics for pairing. Therefore, we choose to relax the third (or equivalently the second) set of constraints in the Lagrangian optimization.

There are many optimal algorithms to solve the 2D assignment problem. For our experiments we choose the auction algorithm due to its low complexity and easier implementation as compared to the other algorithms [22]. Since, the constraints are relaxed the solution to the above dual optimization problem is not primal feasible. Nevertheless, it provides us with the lower bound on the optimal cost. Thus, the next step is to obtain the dual vector $\mathbf{u}$ which maximizes $L(\mathbf{u})$. Or,

$$\max_{\mathbf{u}} L(\mathbf{u}) \tag{4.7}$$

To solve the above dual optimization problem, many methods have been suggested in the literature [22]. We chose accelerated subgradient algorithm for our experiments which produces a sequence of updated lagrange multipliers $\mathbf{u}^0 \longrightarrow \mathbf{u}^1 \longrightarrow \cdots .. \longrightarrow \mathbf{u}^\star$. This update is constructed in the following manner. Define $\mathbf{g}$ as the N dimensional subgradient

vector given by the following equation:

$$g_i = 1 - (\sum_{k=1}^{K} \sum_{j=1}^{N} a_{p,j,k}) \quad p = 1, 2, \ldots N \tag{4.8}$$

. **u** is updated as follows:

$$u_i = u_i + (\text{adapt}) * g_i \tag{4.9}$$

Note that, g_i is the measure of the constraint violation. It is zero if the partial feasible 2D subproblem does not violate the constraint. The step-size adapt is used to improve convergence. It is computed using accelerated subgradient method. Other methods are available in literature [25, 26].

Once the dual vector is updated, a corresponding feasible solution is constructed. We define a relative approximate duality gap, d_{gap} as $\frac{|f_p - f_d|}{|f_p|}$, where f_p is the primal feasible cost and f_d is the dual cost. The process stops when d_gap is less than the threshold. For our experiments, this threshold was set to 10^{-5}.

4.3.2 Semi-Exhaustive Search

The feasible solution set of the chromosome classification problem is extremely large. Hence, an exhaustive search for the optimal solution on this set is impractical. In the semi-exhaustive search method, we select a subset of the feasible solution set and search for the optimal solution in the context of the subset. The selection of the subset is done by eliminating those class-chromosome associations, which have a relatively lower probability of participating in the optimal solution. A heuristic association probability measure between the chromosome and the class is defined to facilitate the above elimination approach. Like the previous methods, the semi-exhaustive search for the solution of the 3D assignment problem also works on the principle of breaking the problem into a series of 2D assignment problems. This 2D problem in our case is the chromosome pairing problem. Recall that the pairing problem is a symmetric assignment problem. Thus, the basic

philosophy of the semi-exhaustive search is to first obtain a subset of the feasible solution set, identify all the 2D assignment problems in the subset and solve them to get the best solution in the subset.

Following are the basic steps in the semi-exhaustive search algorithm. It will be preceded by the detailed description.

- Build the three level tree of the objects and the classes.
- Assign heuristic association probabilities to the edges between level 1 (class) and the level 2 (chromosomes).
- Obtain the subset by eliminating those class-object associations which do not satisfy the threshold criteria.
- Perform all possible 2D assignments on this subset.
- Select the assignment with lowest overall cost.

The algorithm starts with building a three level tree. Figure 4.1 shows one example. Nodes in the level 1 denotes the classes and nodes in the second and the third level represents the chromosomes. Each edge between level 1 and level 2 nodes represents a possible assignment of a class C_k to an object $X_i, \forall i$. Each edge between level 2 and level 3 nodes represents the pairing between the objects X_i and $X_j, i \neq j$. Thus, a path from level 1 to level 3 denotes the joint pairing and classification of objects X_i, X_j with class C_k. Figure 4.1 shows a small instance of a tree for two classes and four chromosomes. The weight of the edge between the level 2 and level 3 nodes is the probability $P(X_i, X_j \in C_k|\mathbf{f}_i, \mathbf{f}_j)$. This weight signifies the association of the node k of level 1, which is nothing but class C_k, with nodes X_i and X_j of the levels 2 and 3.

Having build the tree, the next step is to break the 3D assignment problem into a series of 2D assignment problems. It can be seen from the Figure 4.1 that, if we link

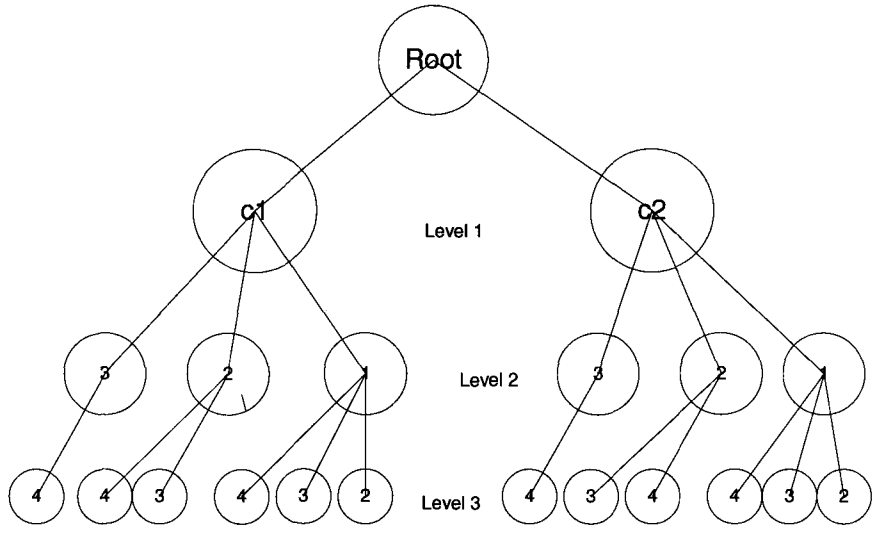


Figure 4.1: Tree for 2 objects and 4 classes with all possible connections

each class (level 1 nodes) to exactly one chromosome (level 2 nodes), then the remaining problem is a 2D assignment problem. For N objects and K classes, there are $\binom{N}{K}$ ways of linking classes to the objects, satisfying the constraints in equation 4.2. Therefore, this is the total number of 2D assignment problems. For $N = 46$ and $K = 23$, this number is of the order 10^{12}. Exhaustive search for all of such problems is therefore impractical. However, the inherent discriminative nature of the chromosome classification and pairing problem motivates us to look for only a subset of the total $\binom{N}{K}$ 2D problems. We observe that there exists a large number of class-chromosome associations which have a very high probability of being absent in the potentially optimal solution. The presence of this discriminative property reduces our total number of 2D problems possible. Thus, we need to find an intelligent way of choosing the subset of those $\binom{N}{K}$ possible permutations, which could be potentially close to the optimal solution. We propose a heuristic to identify the "potential" optimal associations and perform an exhaustive search on those to identify the solution which has the least cost or in other words, is closest to the optimal solution.

We start by defining the association set Υ between the nodes in level 1 and level 2 as follows

$$\Upsilon = \{p(C_k, X_i), \forall k, i\} \tag{4.10}$$

where, $p(C_k, X_i)$ is called the heuristic association probability of the class C_k with the object X_i and is calculated in following manner

$$w(C_k, X_i) = \underset{j}{max} P(X_i, X_j \in C_k | \mathbf{f}_i, \mathbf{f}_j) \tag{4.11}$$

$$p(C_k, X_i) = \frac{w(C_k, X_i).w(C_k, X_i)}{\sum_m w(C_m, X_i) \sum_n w(C_m, X_n)} \tag{4.12}$$

Note that, in the previous equation, the weights are normalized with respect to the classes and the objects.

Having build the association set Υ, we divide the set Υ into K sets I_k for $k = 1, 2, \ldots, K$. This set registers all the possible associations between a class and all the chromosomes (objects). The set I_k is defined as follows:

$$I_k = \{(X_i, p(C_k, X_i)), \forall i = 1, 2 \ldots, N\} \tag{4.13}$$

Next step is to eliminate the unwanted associations and reduce the total number of possible 2D problems. We define a threshold factor λ, and delete all the chromosomes, X_i from the set I_k, for which the $p(C_k, X_i)$ is less than λ times the $\underset{i}{\max}\ p(C_k, X_i)$. This way we delete all the associations that we guess are not the part of the optimal solution. This process is repeated for all the classes. Tree in Figure 4.2 describes the previous step for a simple example of 2 classes and 4 objects. Having build the smaller association set, we now perform an exhaustive search over all the possible cases, and solve all the possible 2D assignment problems. To obtain the best solution, we pick the solution with lowest overall cost according to equation 4.2.

The semi-exhaustive search approach for the optimal solution has interesting properties. The threshold factor λ can be changed according to the total number of problems we

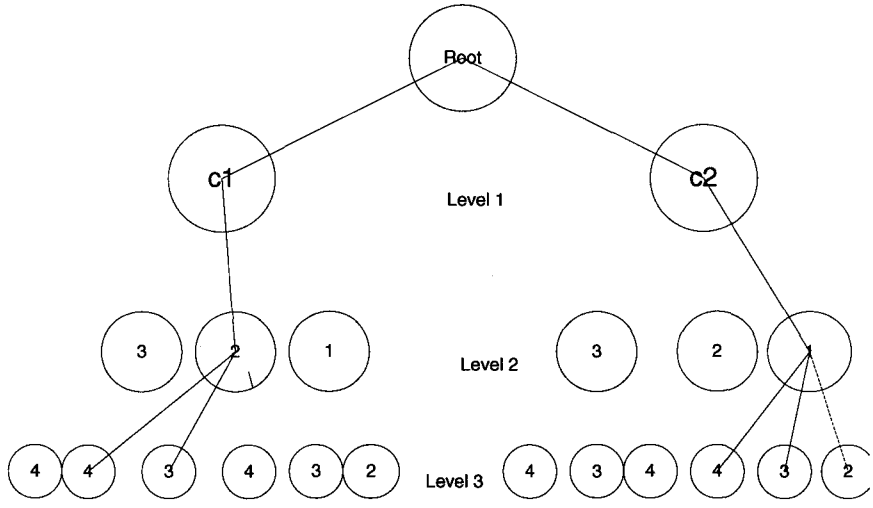


Figure 4.2: Tree depicting the initial assignment of the classes with the object, —- shows the infeasible assignment

want to solve in our search for the lowest cost solution of the assignment problem. The reduction in the value of λ implies inclusion of more associations without excluding the previous associations. In order words, the new solution will always have the an overall cost less than or equal to the previous solution. This proves the convergence properties of our heuristic method.

4.4 Estimation of the probabilities using the features

A cytogenetist classifies and pairs chromosomes using scale measurements like lengths, areas and intensities of segmented chromosomes, but he or she relies even more on 2D appearance features such as the 2D contour shape of a chromosome and particularly the banding pattern within the contour, which is scale invariant. For cluster analysis of 2D appearance patterns we factor out the scale measurements such that all chromosomes are mapped into a feature space of fixed rather than variable dimensions using a technique

proposed in [?]. Specifically, each of the segmented chromosomes is transformed into a 10×100 template by cubic spline interpolation, and subsequently normalizing the resulting template image to have zero mean and unit variance. We call this 10×100 template image (see Fig.4.3) for some examples) canonical pattern and use it as a 1000-dimensional feature vector. The canonical pattern characterizes the contour shape and banding pattern of a chromosome irrespective of its size and intensity. In other words the canonical pattern is scale-invariant in terms of both geometry and signal strength.

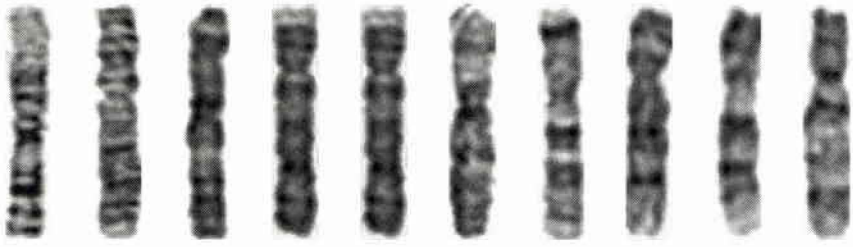


Figure 4.3: Canonical pattern of the chromosomes

Besides the scale-invariant canonical pattern, the length, area, average intensity, and other scale-related features of a chromosome also provide useful information for classification. We denote by $\mathbf{f}_{i,1}$ the canonical pattern of chromosome X_i and by $\mathbf{f}_{i,2}$ the vector of other scale-sensitive features of X_i. The combined feature vector of X_i is $\mathbf{f}_i = (\mathbf{f}_{i,1}, \mathbf{f}_{i,2})$.

Directly estimating $P(X_i, X_j \in C_k | \mathbf{f}_i, \mathbf{f}_j)$ is very difficult, if not impossible, because of the high dimensionality of the problem. In practice one has to resort to some approximation method. To simplify the problem we make the following two assumptions:

1. $P(X_i, X_j \in C_k | \mathbf{f}_i, \mathbf{f}_j) = P(X_i, X_j \in C_k | \mathbf{f}_{i,1}, \mathbf{f}_{j,1}) P(X_i, X_j \in C_k | \mathbf{f}_{i,2}, \mathbf{f}_{j,2})$

2. $P(X_i, X_j \in C_k | \mathbf{f}_{i,1}, \mathbf{f}_{j,1}) = P(X_i \in C_k | \mathbf{f}_{i,1}) P(X_j \in C_k | \mathbf{f}_{j,1})$

The first assumption is reasonable because the scale-invariant canonical chromosome pattern $\mathbf{f}_{i,1}$, which characterizes the generic appearance of chromosome X_i, is independent of

the feature vector $\mathbf{f}_{i,2}$ that consists of absolute measurements of X_i such as area, length, intensity, and etc. The second assumption is made for operational reasons. Based on the above assumptions, we have

$$P(X_i, X_j \in C_k|\mathbf{f}_i, \mathbf{f}_j) = P(X_i \in C_k|\mathbf{f}_{i,1})P(X_j \in C_k|\mathbf{f}_{j,1})P(X_i, X_j \in C_k|\mathbf{f}_{i,2}, \mathbf{f}_{j,2}) \quad (4.14)$$

4.4.1 Estimation of $P(X_i \in C_k|\mathbf{f}_{i,1})$

To estimate $P(X_i \in C_k|\mathbf{f}_{i1})$, we adopt a technique of appearance-based object classification [18], because the feature vector $\mathbf{f}_{i,1}$ represents the normalized appearance of the chromosome X_i. However, the high dimensionality of $\mathbf{f}_{i,1}$ poses a difficulty. To make the problem feasible, we project the 1000-dimensional feature vector $\mathbf{f}_{i,1}$ into a subspace via principal component analysis. Fig. 4.4 shows the images of the first 10 eigen vectors obtained from the PCA of the canonical pattern of the chromosome.

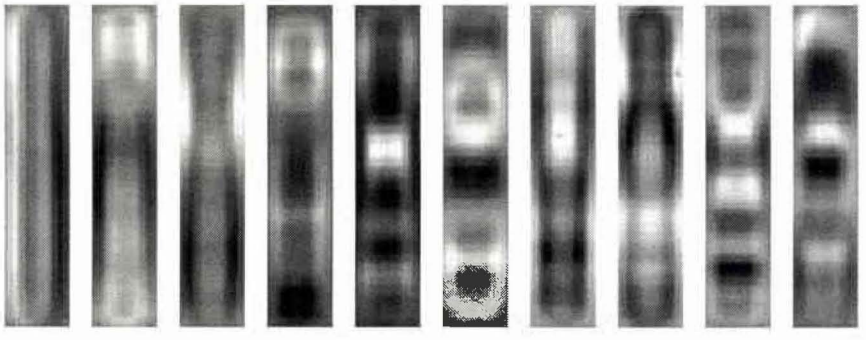


Figure 4.4: Images of the first 10 eigen vectors of the Canonical pattern

In the projected subspace the object features are observed to be approximately multivariate normal distributed. Hence, the posterior probability of the classes given canonical patterns is estimated by

$$P(X_i \in C_k|\mathbf{f}_{i,1}) = \frac{1}{\sqrt{(2\pi)^Q|C_k|}} exp[-\frac{1}{2}(\mathbf{f}_{i,1} - \mu_k)^T\Sigma_k^{-1}(\mathbf{f}_{i,1} - \mu_k)] \quad (4.15)$$

where μ_k is the mean vector and Σ_k is the $Q \times Q$ covariance matrix of class k, $k = 1, 2 \ldots, K$.

4.4.2 Estimation of $P(X_i, X_j \in C_k | \mathbf{f}_{i,2}, \mathbf{f}_{j,2})$

The estimation of the probability of objects X_i and X_j belonging to class k given their correlated features, or $P(X_i, X_j \in C_k | \mathbf{f}_{i,2}, \mathbf{f}_{j,2})$, is the most critical step in the estimation of the overall probability $P(X_i, X_j \in C_k | \mathbf{f}_i, \mathbf{f}_j)$. It is the posterior probability $P(X_i, X_j \in C_k | \mathbf{f}_{i,2}, \mathbf{f}_{j,2})$ that accounts for the joint statistics of correlated features of the homologue pairs, an aspect that was ignored by the previous chromosome classification methods.

To estimate the probability or $P(X_i, X_j \in C_k | \mathbf{f}_{i,2}, \mathbf{f}_{j,2})$, we start with the assumption that the features $\mathbf{f}_{i,2}$ and $\mathbf{f}_{j,2}$ are joint normal distributed, for a given class k. Therefore, we define the probability in the following manner:

$$P(X_i, X_j \in C_k | \mathbf{f}_{i,2}, \mathbf{f}_{j,2}) = \frac{1}{(2\pi)^N |\Sigma_k|^{1/2}} exp[-\frac{1}{2}(\mathbf{F} - \mathbf{U}_k)' \Sigma_k^{-1} (\mathbf{F} - \mathbf{U}_k)] \tag{4.16}$$

where N is the total number of features in consideration, $\mathbf{F} = (\mathbf{f}_{i,2}, \mathbf{f}_{j,2})'$ is the concatenation of the two feature vectors, and $\mathbf{U}_k = (\mu_k, \mu_k)'$ is the concatenation of the mean vectors, where μ_k is the N-dimensional mean vector of the feature vectors $\mathbf{f}_{i,2}$ in class k. Σ_k is the covariance matrix:

$$\Sigma_k = \begin{pmatrix} \Sigma_{1k} & \Sigma_{2k} \\ \Sigma_{2k} & \Sigma_{1k} \end{pmatrix} \tag{4.17}$$

Here, Σ_{1k} is the $N \times N$ covariance matrix of the N features of the chromosomes belonging to class k, whereas Σ_{2k} is the $N \times N$ covariance matrix for the features of homologue pairs in class k. Note that the correlation between the features of the homologue pairs is characterized by Σ_{2k}.

Chapter 5

Experimental Results and Discussions

In this chapter, we present experimental results of various chromosome classification and pairing algorithms, and compare their performances. The outline of the chapter is as follows. The first section describes the main goal of our experiments. Section 2 describes the experimental set up and the various test scenarios. The third section lists detailed experimental results. Finally, the fourth section provides empirical evidence of the fast convergence of the Lagrangian relaxation algorithm when applied to the optimal 3D assignment problem.

5.1 Purpose of Experiments

Automated chromosome classification and analysis are carried out in several steps. They are pre-processing, object segmentation, feature selection and measurement, and lastly the classification stage [3].

The preprocessing stage aims to improve the quality of the cell image by the techniques of noise removal, edge enhancement and contrast improvement. Object segmentation is

to isolate the metaphase chromosomes from the cell image. In the next step the features are extracted from the segmented image of the chromosomes. Finally, classification is done based on optimized statistical inference, such as maximum likelihood estimation. The scope of this thesis is restricted to developing classification and pairing algorithms. Therefore, The main goal of our experiments is to evaluate the accuracy of different algorithms in classifying and pairing chromosomes, given the same data set and the same sample statistics. The next section explains the set up of the experiments.

5.2 Experimental Setup

5.2.1 Data Sets

We used two data sets for our experiments. Both of the data sets are provided to us by Advanced Digital Imaging Research, Houston, Texas. The first data set is a complete set with 16094 chromosomes extracted from 350 cells. We chose this data set for two reasons. First, it is the largest one in size that we can find, providing sufficient sample data for the estimation of high-order joint statistics. Secondly, the data set contains cells captured by different imaging technologies and hence has significant variations in features. Also, some chromosome segmentation errors are present in this data set due to flaws in image processing. The large variances and imperfections of the data set make it a good test case to evaluate the performance and robustness of our algorithms. The second data set is the well-known Copenhagen data set and widely used in the literature as a benchmark [17, 3]. A unique aspect of the Copenhagen set is that more than 90% of the cells have fewer than 46 chromosomes, i.e., incomplete. But on the other hand, the chromosome features of the Copenhagen set are well extracted with a much lower noise level than the first data set.

5.2.2 Performance Evaluation Method

In the performance evaluation the common cross-validation method is adopted. The data set is split into two nearly equal subsets, A and B. The classification accuracy is measured by averaging the test results using subset A for training and subset B for testing, and vice versa. The chromosome features used in our experiments include length, area, density, centromere index, p- arm density, q-arm density, upper and lower density ratio which constitute the individual components of the feature vector $\mathbf{f}_{i,2}$, and as well as the scale-invariant canonical pattern $\mathbf{f}_{i,1}$. The joint statistics of the features and class memberships of the chromosomes are estimated from the samples of the training set. The estimated posterior probabilities $P(X_i \in C_k|\mathbf{f}_{i,1})$ and $P(X_i, X_j \in C_k|\mathbf{f}_{i,2}, \mathbf{f}_{j,2})$ from the training set are then used by the 3D assignment algorithm and the transportation algorithm to classify the chromosomes of the test sets.

5.3 Empirical Results

In this section, we present the comparisons of the correct classification and pairing rates of the three techniques based on transportation, graph-matching and the 3D assignment based algorithms. In our experiments, the cell data is divided into male and female categories, each of which is evaluated separately. This is because the female and the male chromosomes form 23 and 24 classes respectively.

Tables 5.1 and 5.2 list the classification and pairing accuracy of the above mentioned techniques for the complete data set(each cell has 46 chromosomes) for male and female cases respectively. The performance of the three algorithms is compared for two feature sets. The first feature set is $\mathbf{f}_{i,1}$[1] and the second set includes $\mathbf{f}_{i,2}$ and $\mathbf{f}_{i,1}$. It is evident from Tables 5.1 and 5.2 that, on using more features, the accuracies of classification and pairing increase for both transportation and graph matching methods. When scalar features are

[1] $\mathbf{f}_{i,2}$ and $\mathbf{f}_{i,1}$ already defined in Chapter 4 Section 4.4

Algorithms	Correct Classification Rate	Correct Pairing Rate
3-D Assignment(using $\mathbf{f}_{i,2}, \mathbf{f}_{i,1}$)	94.25%	89.56%
Transportation(using $\mathbf{f}_{i,2}, \mathbf{f}_{i,1}$)	94.05%	87.89%
Graph-Matching(using $\mathbf{f}_{i,2}, \mathbf{f}_{i,1}$)	92.37%	86.52%
Transportation(using $\mathbf{f}_{i,1}$)	91.08%	83.82%
Graph-Matching(using $\mathbf{f}_{i,1}$)	89.34%	82.58%

Table 5.1: Comparison of the classification and pairing algorithms for the male test set.

Algorithms	Correct Classification Rate	Correct Pairing Rate
3-D Assignment	94.10%	90.10%
Transportation (using $\mathbf{f}_{i,2}, \mathbf{f}_{i,1}$)	93.9%	89.12%
Graph-Matching (using $\mathbf{f}_{i,2}, \mathbf{f}_{i,1}$)	92.37%	86.52%
Transportation(using $\mathbf{f}_{i,1}$)	92.3%	84.8%
Graph-Matching (using $\mathbf{f}_{i,1}$)	89.34%	82.58%

Table 5.2: Comparison of the classification and pairing algorithms for the female test set.

added, the average error rate of classification decreases by 33% for the transportation method and by nearly 30% for the graph-matching method. Similar trends follow for the pairing accuracy also. Overall, the transportation method performs slightly better than the graph-matching technique with respect to the classification and pairing accuracy of chromosomes. On combining these two methods in the form of joint classification and pairing via the 3D assignment, the classification and pairing error rates decrease further. The classification error rate decreases by only 3% while the pairing rate decreases substantially by around 10% when 3D assignment algorithm is compared with the transportation algorithm. Here again, similar trends follow for both male and female cases.

We also measure the performance of the three algorithms in terms of 100% classifica-

tion (correct classification of all the chromosomes in the cell). The 3D matching algorithm outperforms both transportation and graph-matching based methodologies in the above performance criteria. More than 40% of the cells are perfectly classified in the case of 3D matching algorithm. Table 5.3 shows the accuracy results.

In our experiments, we applied two heuristics to solve the 3D assignment problem. Both the heuristics gave comparable classification and pairing performance. The lagrange relaxation method was executed with the stopping criterion of 100 iterations and for the semi-exhaustive search, the stopping criteria was the execution of a minimum of 1000 2D problems. Of the two methods, the average runtime for the lagrangian relaxation method was nearly 1.5 times of the semi-exhaustive search. Also, as expected, the computational complexity of the auction algorithm based transportation and graph-matching algorithm is much lower as compared to the above mentioned 3D assignment algorithms.

Algorithms	100% Classification Rate
3-D Assignment	40.83%
Transportation(feature set 2)	39.64%
Graph-Matching(feature set 2)	34.32%

Table 5.3: Comparison of the perfect classification rate for the three algorithms

5.4 Incomplete Data set

We also tested the technique of adding dummy chromosomes to incomplete cells which makes the discrete optimization approach for chromosome classification computationally feasible. In our experiment with incomplete cells, we chose the well-known Copenhagen data set in which more than 90% cells have fewer than 46 chromosomes. Table 5.4 compares the performance of different algorithms in terms of the correct classification

Algorithms	Correct Classification Rate
Neural Network	95.6%
Greedy	97.4%
Transportation using dummy chromosomes	98.1%

Table 5.4: Comparison of the classification algorithms for Copenhagen set that contains incomplete cells.

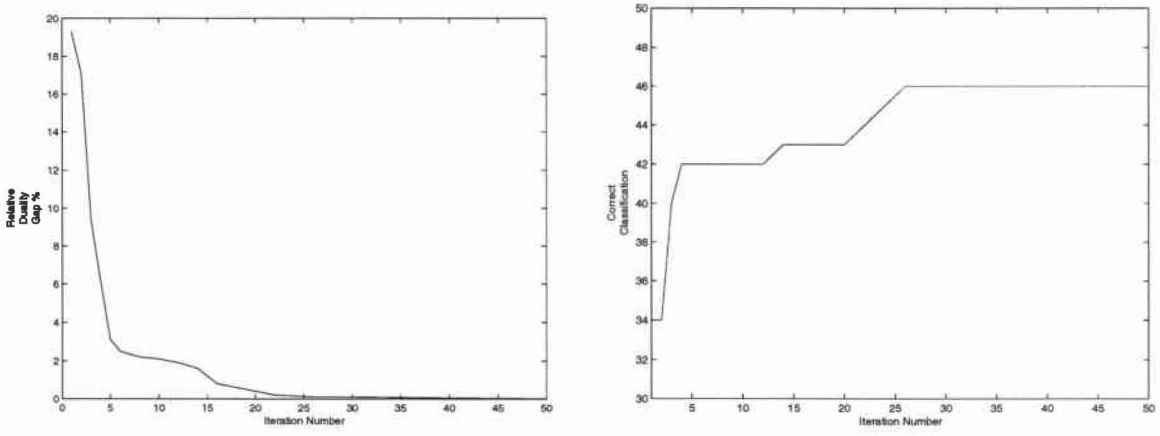
rate. The proposed transportation algorithm for the classification of the incomplete cells outperforms neural network method [16, 17] and the greedy maximum likelihood method for this data set.

5.5 Convergence Properties of Lagrange Relaxation Algorithm

The 3D assignment problem is NP hard. Therefore, an optimal algorithm may take an exponential amount of CPU time in the input size. In our case the input size is the number of chromosomes in the concerned cell. Note that the problem is still intractable even the input size $N = 46$ appears to be modest.

In the previous chapter, we discussed a near optimal 3D assignment algorithm that trades off computation time with accuracy. The algorithm iterates by updating the dual vector which is obtained after the constraints are incorporated into the cost function and the resulting 2D assignment problem is solved. We define a relative approximate duality gap d_{gap}, as $\frac{|f_p - f_d|}{|f_p|}$, where f_p is the primal feasible cost and f_d is the dual cost. The process stops when d_{gap} is less than the threshold. For our experiments, this threshold was set to 10^{-5}. Plot(a) in Figure 5.1 shows the change in the relative gap with the dual vector update iterations. We see that the duality gap decreases as the algorithm proceeds.

Plot(b) in Figure 5.1 shows the increase in accuracy of classification as the duality gap decreases. In other words, the relaxation algorithm converges fast and achieves a near optimal solution.



(a) (b)

Figure 5.1: (a) Plot of relative duality gap % showing the convergence of the relaxation algorithm for a typical case (b) Plot of correct classification rate vs. the number of iterations as the algorithm proceeds.

Chapter 6

Conclusions

Human chromosome classification is a classical and yet challenging pattern recognition problem. Most of the prior work on this problem took either greedy or local optimization approach, notably the Neural Networks (NN) approach. Also, the closely related problems of chromosome classification and homologue pairing were treated separately in the past. These shortcomings have been rectified by this thesis. We proposed an algorithmic approach of optimal three-dimensional assignment to solving both the problems of chromosome classification and pairing simultaneously in a unified framework. The 3D assignment approach allows the utilization of two types of statistical correlations: between the features of chromosomes that belong to a given class but are drawn from different cells, and between the features of chromosomes of different classes but within a given cell.

After the problem was cast into an optimal 3D assignment framework, the next challenge was the statistical estimation of the joint posterior probabilities used in the cost function of the optimization problem. In our work, we use all the scalar features and chromosome banding profile to estimate the above required joint classification and pairing probabilities. The same estimates can also be used in the transportation algorithm for classification and in weighted graph matching algorithm for pairing.

From the discrete optimization point of view, the 3D assignment approach combines

the transportation algorithm for classification and weighted graph matching for pairing. However, the combined problem becomes NP-hard. Algorithmically, we necessarily resort to heuristic approaches to finding good practical solutions. Two techniques were developed: the Lagrangian-type relaxation method and the semi-exhaustive search method.

Experimental results demonstrated an appreciable increase in pairing and classification accuracy when additional statistics are used to estimate the required posterior probabilities in the objective function of the 3D assignment problem formulation. For the same set of features, the 3D assignment approach marginally outperforms the transportation and the weighted graph matching methods in both pairing and classification accuracies.

We also generalized the global optimization approaches of transportation for chromosome classification and minimum-weight graph matching for homologue pairing from complete cells to incomplete cells.

Appendix A

The Transportation Problem and the Auction Algorithm

A.1 Transportation Problem

The transportation problem is a special linear programming problem which has been encountered in many applications [27]. The mathematical formulation for the problem is following:

$$\sum_{j=1}^{n} x_{ij} \leq a_i \quad i = 1, \cdots, m$$

$$\sum_{i=1}^{m} x_{ij} \leq b_j \quad j = 1, \cdots, n$$

$$x_{ij} \geq 0 \quad i = 1, \cdots, m, \quad j = 1, \cdots, n$$

$$minimize \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij}$$

under the hypothesis $\sum a_i \geq \sum b_j \quad a_i \geq 0, \quad b_j \geq 0, \quad c_{ij} \geq 0.$

The first two hypothesis are the necessary conditions for the existence of the program.

The economic interpretation of the transportation problem is as follows. Suppose we have at our disposal at each of the m origins i a quantity a_i of a commodity which we wish to transport to n destinations j, in order to satisfy the demand b_j there. The cost of transportation from i to j is c_{ij}. We wish to minimize the overall transportation cost.

We can always put the problem into a standard form in which the inequalities are replaced by equalities by adding slack variables. The chromosome classification problem can be modelled as the transportation problem too. The classes can be treated as the destination and each chromosome can be considered as an origin. Thus, $b_j = 2$ and $a_j = 1$ in our case. The special structure of the constraints of the above transportation problem enables us to solve problem easily. In our work, we used the auction algorithm given by Bertsekas [24]. To understand how chromosome classification problem can be solved using the auction algorithm, it is necessary to understand the auction algorithm for a symmetric assignment problem.

A.2 Assignment Problem and the Auction Algorithm

The classical symmetric assignment problem is defined as follows. Suppose there are n persons and n objects, with a_{ij} as the cost of assigning the person i to the object j, the objective is to find a set S of person-object pair (i, j), such that each person i and each object j are assigned in at most one pair. That is the number of pairs in S should be exactly n. Auction algorithm was suggested by Bertsekas to solve the above assignment problem.

Auction algorithm can be intuitively understood by considering an economic equilibrium problem. Consider the problem of matching n persons with n objects. Suppose, each object j has a price p_j and that the person has to pay p_j in order to receive the object j. Therefore, the net value of object j to the person i is $a_{ij} - p_j$ and person i would

logically want to be assigned to j_i such that,

$$a_{ij} - p_{j_i} = \max_{j=1,\cdots,n} (a_{ij} - p_{j_i}) \tag{A.1}$$

If the above condition holds for all the objects, the set of prices are at equilibrium. The auction algorithm is based on the above economic equilibrium problem. Starting with any set of prices, the process iterates till all the persons are satisfied according to (A.1). Otherwise, the persons who are not satisfied with the assignment are picked to change the assignment. Let this person i, finds an object j_i which offers maximal value, that is

$$j_i = arg \max_{j=1,\cdots,n} (a_{ij} - p_j) \tag{A.2}$$

then:

1. Exchanges objects with the person assigned to j_i at the beginning of the iteration
2. Sets the price of the best object j_i to the level such that he is indifferent between j_i and the second best object. In order words:

$$p_{j_i} = p_{j_i} + \alpha_i \tag{A.3}$$

where,

$$\alpha_i = v_i - w_i + \epsilon \tag{A.4}$$

v_i is the best object value,

$$v_i = \max_{j=1,\cdots,n} (a_{ij} - p_j) \tag{A.5}$$

and w_i is the second best object value

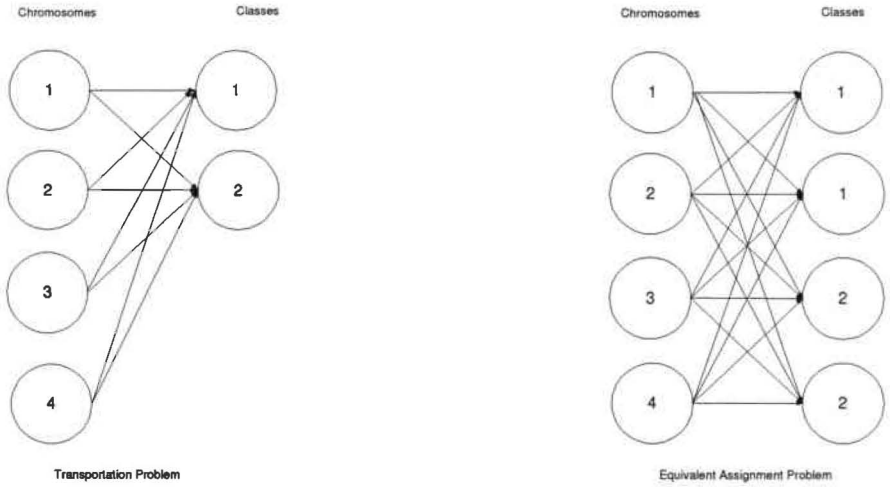
$$v_i = \max_{j \neq j_i} (a_{ij} - p_j) \tag{A.6}$$

and ϵ is the complimentary slackness variable > 0.

This process is repeated until are persons are satisfied.

Note that the above process is the same as the auction process, where at each person raises the price of his desired object by bidding the increment α_i. This makes the bidders own preferred object less attractive to the other potential bidders. Here, ϵ is called complementary slackness variable, which insures that in case when more than one object offers maximum value for the bidder i, the system does not block. Please refer [24] for specific example of above case.

The transportation problem described in the section 1 can be easily converted into a symmetric assignment problem by replacing each source(sink) into a collection of "duplicate" persons. In particular, a source node i with supply a_i is replaced by a_i persons, and a sink node j with demand a_j is replaced by a_j objects. Furthermore, for each arc(i, j) we must create an arc of benefit a_{ij} connecting each person corresponding to i with each object corresponding to j. An example is given in Figure A.1.



(a) (b)

Figure A.1: Illustration of a conversion of chromosome classification based on a transportation problem(a) into an equivalent symmetric assignment problem(b), note that the classes are duplicated in (b)

Bibliography

[1] C.A. Orengo, D.T. Jones, and J.M. Thornton, Bioinformatics, BIOS Scientific Publishers Limited, 2003.

[2] H. Rashidi, and L.K. Buehler, Bioinformatics basics: applications in biological science and medicine, 2nd Edition, Boca Raton, Fla. London:CRC, 2003.

[3] B. Lerner, Towards a completely automatic neural network based human chromosome analysis, IEEE Transactions on Systems, Man and Cybernatics Special Issue on Artificial Neural Networks, vol.28, pt.B, pp. 544-552, 1998.

[4] B. Lerner, H. Guterman, I. Dinstein, and Y. Romem, Human chromosome classification using multilayer perceptron neural network, Int. J. of Neural Syst., vol. 6, pp. 359-370, 1995.

[5] S. Zimmerman, D. Johnston, F. Arright, and M.E Rupp, Automated homologue matching of human G-Banded chromosomes, Computational Biol. Med. vol.16, No.3, pp, 223-233, 1986.

[6] E. Granum, and M.G. Thomason, Automatically inferred Markov network models for classification of chromosomal band pattern structures, Cytometry, vol. 11, pp. 26-39, 1990.

[7] P.A. Errington, and J. Graham, Application of artificial neural networks to chromosome classification, Cytometry, vol. 14, pp. 627639, 1993.

[8] A.M. Jennings, and J. Graham, A neural network approach to automatic chromosome classification, Phys. Med. Biol.38, pp. 959-970, 1993

[9] M. Tso, P. Kleinschmidt, I. Mitterreiter, and J. Graham, An efficient transportation algorithm for automatic chromosome karyotyping, Pattern Recognition Letters, vol. 12, pp. 117-126, 1991.

[10] K.R. Castleman, and R.J. Wall, Automatic Systems for Chromosome Identification, Nobel 23:77-84, 1973

[11] R. Ledley, P.S. Ing, and H.A. Lubs, Human chromosome classification using discriminant analyses and bayesian probability, Computational Biol. Med, vol.10, pp. 209-219, 1980.

[12] F.C.A. Groen, T.K. Ten Kate, A.W.M. Smeulders, and I.T. Young, Human chromosome classification based on local band descriptors, Patt. Rec. Letters, vol. 9, 211-222, 1989

[13] G.H. Granlund, Identification of human chromosomes by using integrated density profiles, IEEE Trans in Biomedical Engineering, BME-23, 182-192, 1976.

[14] L. Vanderheydt, A. Oosterlinck, J. Van Daele, and H. Van den Berghe, Design of a graph-representation and a fuzzy-classifier for human chromosomes, PR, vol. 12, Number 3, 1980.

[15] A. Carothers, and J. Piper, Computer-aided classificatin of human chromosomes: a review, Statistics and Computing, vol.4, pp 161-171, 1994.

[16] R.J. Stanley, J.M. Keller, P. Gader, and C.W. Caldwell, Data-driven Homologue Matching for Chromosome Identification, IEEE Trans. Medical Imaging, vol.17, No.3, pp. 452-462, 1998.

[17] W.P. Sweeney Jr, M.T. Musavi, and J.N. Guidi, Classification of chromosomes using a probabilistic neural network, Cytometry, vol.16, pp. 17-24, May 1994.

[18] Q. Wu, Z. Liu, Z. Xiong, Y. Wang, T. Chen, and K.R. Castleman, On optimal subspaces for appearance-based object recognition, Proc. IEEE ICIP'02, Rochester, NY, USA, 2002.

[19] Z. Galil, S. Micali, and H. Gabow, An O(EV log V) algorithm for finding a maximal weighted matching in general graphs, SIAM J. Comput, vol. 15, 1, 120-130, 1986.

[20] G. Parker and R.L. Rardin, Discrete Optimization, Academic Press, 1988

[21] M.R. Garey and D.S. Johnson, Computers and Intractability: A guide to the theory of NP-Completeness, San Francisco, CA:Freeman, 1979.

[22] K.R. Pattipati, S. Deb, Y. Bar-Shalom and B.R Washburn, A new relaxation algorithm and passive sensor data association problem, IEEE Transcations on Automatic Control, 37, 2(Feb-1992), 198-213

[23] A. Sinha and T. Kirubarajan, A randomized heuristic approach for multidimensional association in target tracking, Proc. SPIE, Signal and Data Processing of Small Targets. Orlando, FL, April, 2004.

[24] D.P. Bertsekas, The auction algorithm: A distributed relaxation method for the assignment problem, Annals of Operations Research, vol. 14, pp. 105-123, 1988.

[25] D.P. Bertsekas, Constrained Optimization and Lagrange Multiplier Methods, Athena Scientific, 1996,

[26] B.T. Poljak: Subgradient method: A survey of Soviet Research, London:pergamon, 1978.

[27] M. Simonhard, Linear Programming, Englewood Cliffs, N.J: Prentice-Hall Inc, 1966.