Hindawi
BioMed Research International
Volume 2017, Article ID 5482750, 13 pages
https://doi.org/10.1155/2017/5482750

Research Article

An Improved Binary Differential Evolution Algorithm to Infer Tumor Phylogenetic Trees

Ying Liang, Bo Liao, and Wen Zhu

College of Information Science and Engineering, Hunan University, Changsha, China

Correspondence should be addressed to Bo Liao; dragonbw@163.com

Received 2 September 2017; Accepted 18 October 2017; Published 27 November 2017

Academic Editor: Tao Huang

Tumourigenesis is a mutation accumulation process, which is likely to start with a mutated founder cell. The evolutionary nature of tumor development makes phylogenetic models suitable for inferring tumor evolution through genetic variation data. Copy number variation (CNV) is the major genetic marker of the genome with more genes, disease loci, and functional elements involved. Fluorescence in situ hybridization (FISH) accurately measures multiple gene copy number of hundreds of single cells. We propose an improved binary differential evolution algorithm, BDEP, to infer tumor phylogenetic tree based on FISH platform. The topology analysis of tumor progression tree shows that the pathway of tumor subcell expansion varies greatly during different stages of tumor formation. And the classification experiment shows that tree-based features are better than data-based features in distinguishing tumor. The constructed phylogenetic trees have great performance in characterizing tumor development process, which outperforms other similar algorithms.

1. Introduction

Cancer is the most serious and dangerous disease to human health in the world. Over the past few decades, researchers have been working on the diagnosis and treatment of cancer. Owing to these great efforts, our understanding of cancer has been greatly improved, and early clinical diagnosis and reliable treatment are critical for cancer [1]. Cancer is the result of an imbalance in the cell cycle of the organism. Each cell of the organism contains a complete genome and has great spontaneity [1]. When the genome is no longer regulated by normal tissue and the spontaneity of cells is activated, then cancer develops. Tumor cells succumb to different evolutionary pressures and result in constant replication, growth, invasion, and metastasis [1].

In the early days, Nowell [2] proposed the "clonal evolution" theory that combines evolutionary biology with tumor biology. The model suggests a tumor is most likely to start with a mutated cell. Owing to the expansion of one or more cell subclones, tumor cells show high heterogeneity, which is an important characteristic of tumor development [3]. These tumor cells show significant differences even in the same tissue of the same individual. It has been shown that tumor heterogeneity is evolving along with tumor progression [3]. Tumor heterogeneity has been shown to have a significant impact on the diagnosis and treatment of cancer [3, 4].

Because of the evolutionary nature of tumor development, phylogenetic models were used to infer tumor evolution through genetic variation data [5]. Navin et al. [6] found that a single breast tumor may contain multiple cell subclones, and their chromosome copy numbers vary considerably via single-cell DNA copy number data on CGH platform. The development of next-generation sequencing allows people to infer SNVs and their allele frequencies in heterogeneous tumor cell populations. Because of the huge number of SNVs, inference of a complete tumor progression model to explain the observed data has encountered computational difficulties. Nik-Zainal et al. [7] reconstructs phylogenetic tree from inferred SNV frequencies based on two assumptions: (i) no mutation occurs twice in the course of cancer evolution and (ii) no mutation is ever lost. Strino et al. [8] proposed a linear algebra approach based on the two hypotheses to limit the number of possible trees, which can handle up to 25 SNVs. Detection of clones based on SNV frequency data is necessary for inferring phylogeny. Jiao et al. [9] proposes PhyloSub, a Bayesian nonparametric model, to infer the phylogeny and genotype of the major subclonal lineages represented in the population of cancer cells. Miller et al. [10] proposed a variational Bayesian mixture model to identify the number and genetic composition of subclones by analyzing the variant allele frequencies. Hajirasouliha et al. [11] formulate the problem of constructing the subpopulations of tumor cells from the variant allele frequencies (VAFs) as binary tree partition and present an approximation algorithm to solve the max-BTP problem. El-Kebir et al. [12] formulate the problem of reconstructing the clonal evolution of a tumor using SNV as the VAF factorization problem and derives an integer linear programming solution to the VAF factorization problem. Popic et al. [13] propose LICHeE, a novel method to infer the phylogenetic tree of cancer progression from multiple somatic samples. Because of copy number alterations, loss of heterozygosity (LOH), and normal contamination, the allele frequencies of related SNV need to be corrected [14]. Copy number variation is segment loss or duplication of genome sequence ranging from kilo bases (Kb) to mega bases (Mb) in size, which covers 360 Mb and encompasses hundreds of genes, disease loci, and functional elements [15]. CNVs affect gene expressions in human cell-lines, which also play a major role in cancer [16]. Subramanian et al. [17] develop a novel pipeline for building trees of tumor evolution from the unmixed tumor copy number variations (CNVs) data. Oesper et al. [18] introduce ThetA, an algorithm to infer the most likely collection of genome and its proportions in a sample, and identify subclonal CNVs using high-throughput sequencing data. Ha et al. [19] also present a novel probabilistic model, TITAN, to infer CNA and LOH events while accounting for mixtures of cell populations, thereby estimating the proportion of cells harboring each event. Some tumor progression analysis tools combine VAFs of SNVs and population frequencies of structure variations to reconstruct subclonal composition and tumor evolution. PhyloWGS [20] uses copy number alterations to correct the VAFs of affected SNVs and greatly improves subclonal reconstruction compared to existing methods. As tumor is a heterogeneity system, Jiang et al. [21] propose Canopy to identify cell populations and infer phylogenies using both somatic copy number alterations and single-nucleotide alterations from one or more samples derived from a single patient. Li and Xie [22] propose a software package called PyLOH to deconvolve the mixture of normal and tumor cells using copy number alterations and LOH information. Yu et al. [23] introduce CloneCNA to address normal cell contamination, tumor aneuploidy, and intratumor heterogeneity issues and automatically detect clonal and subclonal somatic copy number alterations from heterogeneous tumor samples. El-Kebir et al. [24] develop SPRUCE to construct phylogenetic trees jointly from SNVs and CNAs, which overcomes complexities in simultaneous analysis of SNVs and CNAs.

The samples of the above studies are mixture of cancer cells and stromal cells; analyzing single cells is the most informative approach to assess the heterogeneity within a tumor [5]. Single-cell analysis is not only one more step towards more-sensitive measurements, but also a decisive jump to a more-fundamental understanding of biology [25]. Navin et al. [26] obtain robust high-resolution copy number profiles by sequencing a single cell and infer about the evolution and spread of cancer by examining multiple cells from the same cancer with the Euclidean metric. Traditionally used Euclidean or correlation distances for tree reconstruction from copy number profiles are ill-suited, owing to the dependent and nonidentical distribution of rearrangement events [5]. Fluorescence in situ hybridization (FISH) is a technique that can be used to count the copy number of DNA probes for specific genes or chromosomal regions in potentially hundreds of individual cells of a tumor. Pennington et al. [27] develop a new method combined with expectation maximization to infer unknown parameters for identifying common tumor progression pathways by taking advantage of information on tumor heterogeneity lost to prior microarray-based approaches on a set of fluorescent in situ hybridization (FISH) data. Chowdhury et al. [28–30] propose a software FISHtrees to build evolutionary trees of single tumors with FISH data. FISHtrees models gain or loss of genetic regions at the scale of single genes, whole chromosomes, or the entire genome, including variable rates for different gain and loss events in tumor evolution [30]. Later, Gertz et al. [31] present FISHtrees 3.0, which implements a ploidy-based tree building method based on mixed integer linear programming. The ploidy-based modeling in FISHtrees 3.0 includes a new formulation of the problem of merging trees for changes of a single gene into trees modeling changes in multiple genes and the ploidy [31]. Here, we propose an improved binary differential evolution algorithm to infer phylogenetic trees (BDEP) using CNV data of cervical cancer and breast cancer. The cervical cancer dataset contains the copy number profiles of four genes, and breast cancer dataset is up to eight genes. Liu et al. [32] show that, on average, each cancer can be explained with around six different marker sets. Tumor phylogenetic tree inference can be treated as minimum Steiner tree problem in directed graph, which is a NP-hard problem. BDEP uses differential individual to search for the best approximate solutions, with the help of individual's difference information and neighborhood optimal information to update. BDEP overcomes the weakness that differential evolution algorithm can only be used in continuous search space with advantages of fast convergence and strong robustness.

2. Methods

2.1. Problem Definition. One copy number variation usually affects the copy number of two or more closely related genes [15]. The genes may change their copy number alone or together with their neighbors located in one copy number variation region, which results in computational difficulties of evolution distance between gene copy number profiles (Figure 1). Shamir et al. propose an algorithm that calculates evolution events in linear time and linear space by backtracking the dynamic programming vector [33]. We adopt the idea proposed by Shamir to calculate the minimum variation events between two copy number profiles. Profiles (u, v) present the evolution distance from the source profile u to the target profile v. As mentioned by Shamir et al. [33],

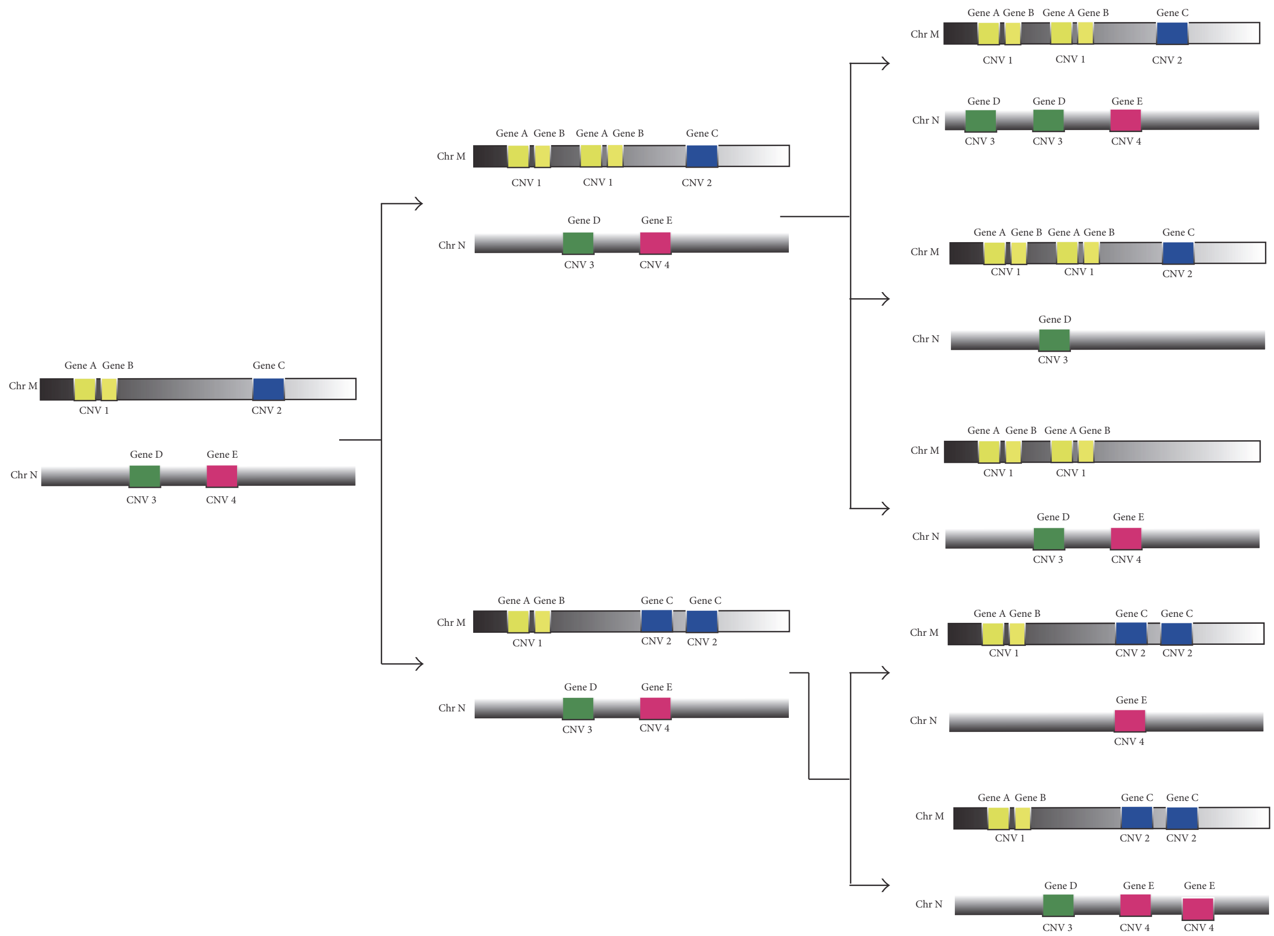


FIGURE 1: The association between CNVs and genes.

if the source profile contains the gene with copy number 0 but the target profile with the gene copy number > 0, the transformation from u to v is unreachable. On the contrary, if the gene has copy number > 0 in the source profile but with the copy number 0 in the target profile, the profiles (u, v) can be inferred. The distance matrix between copy number profiles is asymmetric, which corresponds to directed edges between copy number profiles.

Cells are continuously growing, proliferating, and dying during the tumor progress; the dying cells disappeared but once played an important role in tumourigenesis. Construct a tree to describe evolutionary relationship of observed cells and dying cells can be regarded as Steiner tree problem; the dying cells in Steiner tree are Steiner node. The Steiner tree problem is a classical combinatorial optimization problem, which has important applications in the fields of computer network layout, circuit design, and biological network analysis. In the paper, the tumor phylogenetic tree is a Steiner minimum tree problem in graph, which is proposed by Hakimi [34] and Hwang et al. [35]. The problem can be described as follows: Given a directed connected graph $G = (V, E)$ with observed nodes and all possible Steiner nodes, V, and edges, E, each node presents a copy number profile and each edge presents the evolution direction between nodes. The weight of each edge presents the evolution distance between copy number profiles. There is a subset $P \subseteq V$; each element presents the observed copy number profile of cell. The Steiner tree problem is to find a subtree T of directed connected graph G, which contains all nodes in P with minimal weight sum. The subtree T is the Steiner tree of subset P; the node that exists in T but not in P is the Steiner node. When $P = V$, the Steiner tree problem is minimum arborescence problem, which can be worked out in polynomial time [36]. Otherwise, the Steiner tree problem has no polynomial time solution, which is a NP-hard problem [37]. When the input scale becomes large, it is impossible to find the exact optimal solution in polynomial time. Therefore, a good approximation algorithm will provide a compromise solution for the NP-hard problem.

2.2. The Improved Binary Differential Evolution Model. The differential (DE) evolution algorithm does not depend on the characteristics information of problem, with the help of difference information among individuals to disturb the formation of individual and then to search the entire population space. Greedy competition mechanism is employed to seek the optimal solution of the problem. DE algorithm is a population-based stochastic direct search method, which is based on real number coding [38]. The differential evolution algorithm has the advantages of fast convergence, simple operation, easy programming, and strong robustness, which have been widely used in various fields [39–42]. The DE algorithm contains three basic operations: mutation, crossover, and selection. The initial population is randomly generated and covers the entire search space.

Initial Population. Suppose $X_{i,G} = \{x^1_{i,G}, \ldots, x^n_{i,G}\}$ is the ith individual of generation Gth; n is the dimension of individual; $i = 1, 2, \ldots, M$ is the population scale; $G = 1, 2, \ldots, G_{\max}$ is the maximum evolution generation. The initial population of DE is generated by

$$x^j_{i,0} = \text{rand}_j(0,1)\left(x^j_U - x^j_L\right) + x^j_L, \tag{1}$$

where x^j_U and x^j_L represent the upper and lower bounds of the jth dimension, respectively, and $\text{rand}_j(0,1)$ represents a random number within the range $[0, 1]$.

Mutation Operation. Randomly select two different individuals $X_{p_1,G}$, $X_{p_2,G}$ to produce the mutant individual $V_{i,G}$ corresponding to individual $X_{i,G}$ as

$$v^j_{i,G} = x^j_{i,G} + \lambda\left(x^j_{p_1,G} - x^j_{p_2,G}\right), \tag{2}$$

where $x^j_{p_1,G} - x^j_{p_2,G}$ is difference vector and scaling factor λ is a positive control parameter of difference vector.

Crossover Operation. Crossover operation aims at increasing population diversity. The crossover strategy exchanges mutant and old individual's information to generate trial individual $U_{i,G}$. The crossover operation is defined as

$$u^j_{i,G} = \begin{cases} v^j_{i,G} & \text{rand}_j[0,1) \le \text{CR or } j = \text{rand}(i) \\ x^j_{i,G} & \text{otherwise.} \end{cases} \tag{3}$$

The crossover strategy ensures that $U_{i,G}$ has at least one element from $V_{i,G}$. The crossover rate CR can be adjusted by user within the range $[0, 1]$.

Selection Operation. Trial individual $U_{i,G}$ will become a member of the next-generation population, if the fitness function values of $U_{i,G}$ are superior to $X_{i,G}$. Otherwise, the individual $X_{i,G}$ will remain in the next-generation population. The selection operation is defined as

$$X_{i,G+1} = \begin{cases} U_{i,G}, & \text{fitness}\left(U_{i,G}\right) \le \text{fitness}\left(X_{i,G}\right) \\ X_{i,G}, & \text{otherwise.} \end{cases} \tag{4}$$

Perform the above three operations repeatedly until the stopping criterion is satisfied.

2.2.1. Binary Differential Evolution Algorithm. Conventional DE algorithm focuses on the problem of continuous search space, which cannot solve the discrete problem. Also the DE algorithm does not take into account the global or neighborhood optimal individual information. In this paper, we propose a novel binary differential evolution algorithm (BDEP) to solve the Steiner tree problem and further construct tumor phylogenetic tree. In BDEP, trial individual absorbs neighborhood optimal individual information to update at crossover phase. BDEP is different from conventional DE algorithm at initial population operation, mutation operation, and crossover operation. The algorithm flow chart of BDEP is in Algorithm 1.

Candidate Steiner Node Generation. The Steiner tree problem in graph is to find a minimum arborescence which at least contains all nodes in subset P. The set of nodes V in graph G includes the nodes in P and all possible Steiner nodes. Before applying Chu-Liu's algorithm to find the minimum arborescence, it is prerequisite to compute all possible Steiner points. The candidate Steiner node is generated according to the gene copy number profile in subset P. Under maximum parsimony criterion, the evolutionary distance from gene copy number profile to the candidate Steiner node is 1. As a result, the set of nodes V consists of candidate Steiner nodes and subset P, which corresponds to a complete directed graph G.

Individual Encoding. The individual i of binary differential evolution is encoded as a binary string $X_i = (x^1_i, x^2_i, \ldots, x^n_i)$, where x^j_i is a binary variable corresponding to the jth candidate Steiner node and n is the number of candidate Steiner nodes. When $x^j_i = 1$, the ith individual has the jth candidate Steiner node. With the gene copy number profile in set P, each individual represents a phylogenetic tree; the fitness function is the distance sum of the phylogenetic tree. The objective of BDEP is to find a minimum arborescence representing tumor phylogenetic tree.

Initial Population. The population initialization of BDEP is as follows:

$$x^j_{i,0} = \begin{cases} 1 & \text{rand}_j(0,1) < 0.05 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The meaning of i, j, and $\text{rand}_j(0,1)$ is the same as that of conventional DE algorithm.

Mutation Operation. For each individual $X_{i,G}$, randomly select two different individuals $X_{p_1,G}$, $X_{p_2,G}$ to produce the mutant individual $V_{i,G}$ as follows:

$$v^j_{i,G} = \begin{cases} x^j_{p_1,G} \mid x^j_{p_2,G} & x^j_{p_1,G} = x^j_{p_2,G} \\ x^j_{i,G} & \text{otherwise.} \end{cases} \tag{6}$$

For the jth candidate Steiner node, if individuals $X_{p_1,G}$, $X_{p_2,G}$ have the same choice, the mutant individual yields $x^j_{p_1,G}$ or $x^j_{p_2,G}$; otherwise it directly derives from $X_{i,G}$.

Crossover Operation. Social learning is an important way to improve population diversity and self-adaptability. The individual would influence its neighbors: BDEP uses local neighborhood as social learning areas. BDEP adopts the ring

```
Require: The copy number profiles (object nodes set $P$).
    The max generation $G_{\max}$.
    The number of individuals (population scale) $M$.
Ensure: The tumor Steiner tree with the shortest length.
Generate candidate Steiner node according to the copy number profiles, construct a complete directed graph Graph.
Set the generation number $G \leftarrow 0$, initialize a population of $M$ individuals $P_G = \{X_{1,G}, \dots, X_{M,G}\}$ with $X_{i,G} = \{x^1_{i,G}, \dots, x^n_{i,G}\}$
    where $x^n_{i,G} \in \{0,1\}$ is a binary variable.
while stopping criterion is not satisfied do
    Mutation step
    for $i \leftarrow 1$ to $M$ do
        Generate a mutant individual $V_{i,G} = \{v^1_{i,G}, \dots, v^n_{i,G}\}$ from the target individual $X_{i,G}$ and two different individuals
        $X_{p1,G}$, $X_{p2,G}$.
        for $j \leftarrow 1$ to $n$ do
            $v^j_{i,G} = \begin{cases} x^j_{p1,G} \text{ or } x^j_{p2,G} & x^j_{p1,G} = x^j_{p2,G} \\ x^j_{i,G} & \text{otherwise} \end{cases}$
        end for
    end for
    Crossover step
    for $i \leftarrow 1$ to $M$ do
        Search the $r$-neighborhood of individual $V_{i,G}$, the best neighbor of $V_{i,G}$ is $V_{n\text{best},G} = \min_{r\text{-neighborhood}}\text{fitness}$
        Update trial individual $V_{i,G}$ to $U_{i,G}$
        $\text{rand}(i) = \lfloor \text{rand}[0,1) * n \rfloor$
        for $j \leftarrow 1$ to $n$ do
            $u^j_{i,G} = \begin{cases} v^j_{n\text{best},G} & \text{rand}[0.1) \leq \text{CR or } j = \text{rand}(i) \\ v^j_{i,G} & \text{otherwise} \end{cases}$
        end for
    end for
    Selection step
    for $i \leftarrow 1$ to $M$ do
        Evaluate the trial individual $U_{i,G}$
        if $\text{fitness}(U_{i,G}) \leq \text{fitness}(X_{i,G})$ then
            $X_{i,G+1} = U_{i,G}$, $\text{fitness}(X_{i,G+1}) = \text{fitness}(U_{i,G})$
        end if
    end for
    Update the generation count $G \leftarrow G + 1$
end while
return optimal tumor Steiner tree $T$
```

Algorithm 1: An improved binary differential evolution algorithm to infer tumor phylogenetic trees (BDEP).

topology of population with radius r to define local neighborhoods. The r-neighborhood of individual i is represented as $\{R_j \mid |i-j| \leq r,\ j = 0, 1, 2, \dots, M-1\}$. The individual $V_{n\text{best},G}$ represents the best neighbors with minimum fitness value in the r-neighborhood of mutant individual $V_{i,G}$. The cross operation is according to

$$u^j_{i,G} = \begin{cases} v^j_{n\text{best},G} & \text{rand}_j\,[0,1) \leq \text{CR or } j = \text{rand}\,(i) \\ v^j_{i,G} & \text{otherwise.} \end{cases} \tag{7}$$

The crossover strategy exchanges mutant individual and its best neighbor's information to generate trial individual. The crossover rate CR can be adjusted by user within the range $[0,1]$. The crossover strategy ensures that $U_{i,G}$ has at least one element from the best neighbor. The neighborhood radius r depends on population scale and the complexity of problem.

Selection Operation. The selection strategy is similar to conventional DE algorithm; whether the trial individual $U_{i,G}$ could become a member of the next-generation population depends on fitness function values. If the new individual $U_{i,G}$ is superior to old one $X_{i,G}$, $U_{i,G}$ would replace $X_{i,G}$. Otherwise, the individual $X_{i,G}$ will remain in the next-generation population.

Repeatedly perform the above three operations until one of the two criteria is satisfied: (i) evolutional iterations reach the maximal generation; (ii) the optimal fitness value is less than the distance sum of minimum arborescence of subset P and stays unchanged in ten consecutive iterations.

3. Results and Discussion

In this section, we apply BDEP to the gene copy number profiles of real tumor and infer the tumor phylogeny of

Table 1: The P value of χ tests between DCIS and IDC.

Sample ID	P value of branches	P value of levels	P value of edges
Patient 1	$4.89E-56$	$8.40E-03$	$5.85E-01$
Patient 2	$4.49E-34$	$5.61E-20$	$9.25E-01$
Patient 3	$1.82E-03$	$1.38E-02$	$8.91E-01$
Patient 4	$5.53E-41$	$1.86E-06$	$2.24E-02$
Patient 5	$2.24E-18$	$4.28E-03$	$5.81E-01$
Patient 6	$4.87E-20$	$5.22E-02$	$3.14E-03$
Patient 7	$6.11E-02$	$1.06E-05$	$1.40E-01$
Patient 8	$2.79E-61$	$1.45E-20$	$2.88E-01$
Patient 9	$1.09E-36$	$1.50E-18$	$7.94E-01$
Patient 10	$6.05E-58$	$1.38E-11$	$9.61E-01$
Patient 11	$1.30E-04$	$5.96E-16$	$8.29E-02$
Patient 12	$7.85E-02$	$7.40E-06$	$4.59E-01$
Patient 13	$2.43E-14$	$4.01E-05$	$9.32E-01$

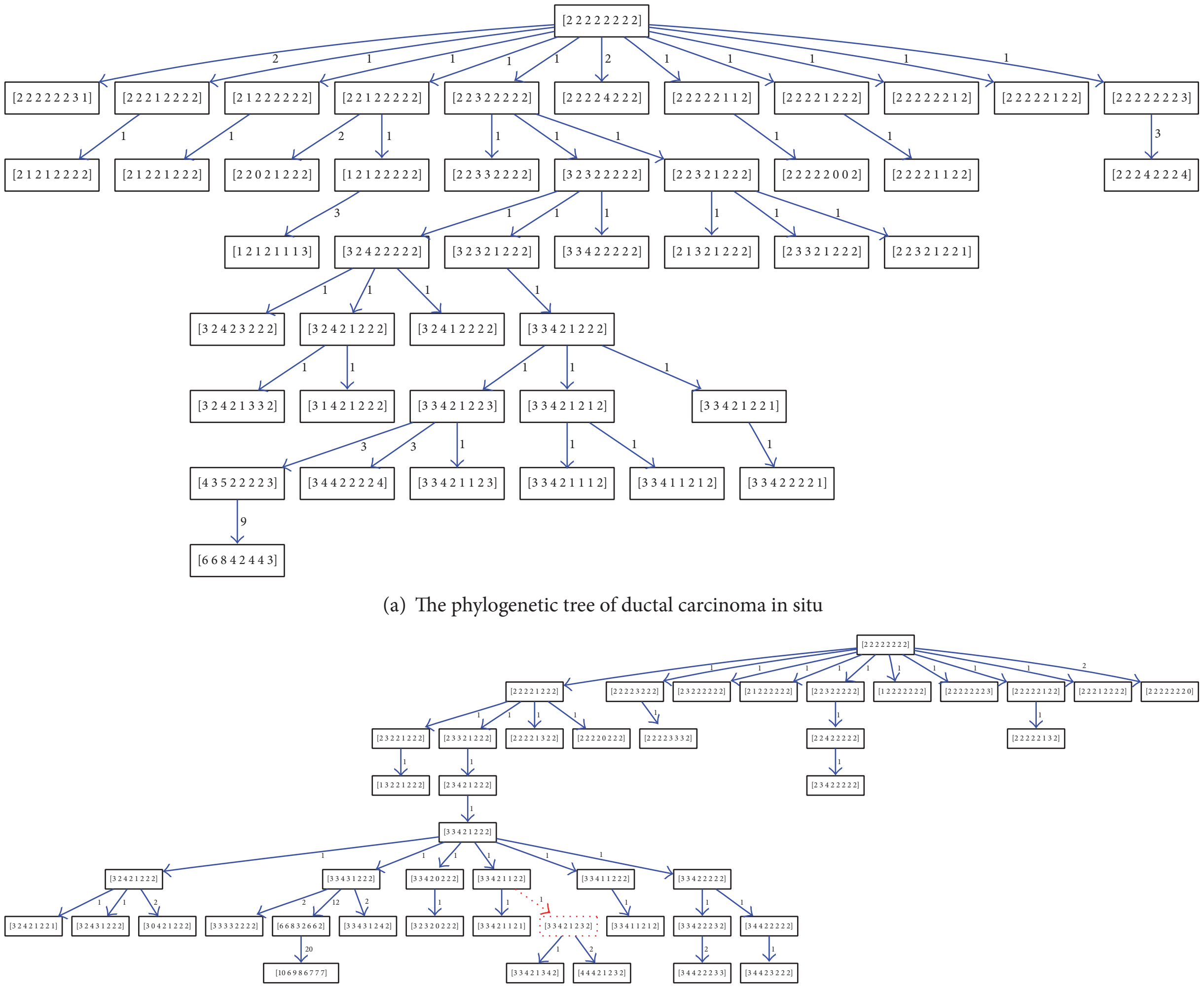
all samples. We study the differences between tumors by statistically analyzing topological features of phylogenetic tree in the following three aspects: branch, level, and edge. And classification experiments are performed to evaluate the merits of these features. The algorithm parameters are set as follows: the max generation $G_{\max}$ is 100; crossover rate (CR) is 0.7 by default; and population size depends on the complexity of the problem ranging from 300 to 500.

3.1. Datasets. Two FISH datasets, cervical cancer and breast cancer, respectively, from Wangsa et al. [43] and Heselmeyer-Haddad et al. [44], are published to visualize copy number changes in tumors based on single-cell analyses. The cervical cancer dataset comprises four probes targeting the genes LAMP3, PROX1, PRKAA1, and CCND1, in pretreatment cervical biopsies from 16 lymph node positive samples and 15 lymph node negative controls from women with stage IB and IIA cervical cancer [43]. The lymph node positive samples contain primary tumors and associated lymph node metastases. The four target genes come from different chromosomes: LAMP3 is a gene located on chromosome 3q26, PROX1 is located on chromosome 1q41, PRKAA1 is located on chromosome 5p19, and CCND1 is located on chromosome 11q13; and altered expression of this gene has been observed in many cancers [43]. The cell number of cervical cancer among 47 cases ranges from 212 to 250 (average cell number is 243), which is not significantly different among primary cancer with positive lymph node, lymph node metastases cases, and lymph node negative controls. But the number of cell gene profiles among them is strikingly different; each gene copy number profile is a tree node in phylogenetic model. The gene profile number of primary cases with positive lymph node ranges from 63 to 187, average being 111. The profile number of lymph node metastases cases ranges from 34 to 115, average being 70. The profile number of lymph node negative controls ranges from 58 to 157, average being 97.

The breast cancer dataset comprises 13 cases of synchronous ductal carcinoma in situ (DCIS) and invasive ductal carcinoma (IDC), which contains eight probes targeting five oncogenes, COX2, MYC, HER2, CCND1, and ZNF217, and three tumor suppressor genes, DBC2, CDH1, and TP53 [44]. The DCIS is considered a precursor lesion for invasive breast cancer, which has a lower degree of chromosomal instability than the IDC [44]. COX2 is located on 1q31.1 and is upregulated in human breast cancer; DBC2 and MYC both are located on chromosome 8; MYC is also upregulated gene in many types of cancers; CDH1 is located on 16q22.1, HER2 and TP53 both are located on chromosome 17, and ZNF217 is located on 20q13.2, which is a strong candidate oncogene for breast and other cancers [44]. The cell number of breast cancer among 26 cases ranges from 76 to 220, average cell number being 142. The cell number and profile number between DCIS and IDC cases are not significantly different. The profile number of DCIS cases ranges from 28 to 143, average being 73. The profile number of IDC cases ranges from 44 to 119, average being 85.

In FISH datasets, gene copy number profiles of each cell are expressed in matrix form, where each row represents a cell case and each column represents a gene probe. The corresponding gene copy number of each cell is a nonnegative integer. The profile with gene copy number of 2 is considered as the root node of tumor evolutionary tree. The datasets can be downloaded at ftp://ftp.ncbi.nlm.nih.gov/pub/FISHtrees/.

3.2. Results on Breast Cancer Datasets. We apply BDEP algorithms to the gene copy number profiles of breast cancer and comparatively analyze the tree topology between paired DCIS and IDC samples. We first analyze the branch features of phylogenetic tree at different stages. The branch is defined as subtree derived from the ith child of the root node. The DBC2 and MYC gene are on chromosome 8, and TP53 and HER2 gene are on chromosome 17. The copy number of genes lying on the same chromosome is easily affected by CNV simultaneously, phylogenetic trees have at most twenty branches, and we use Chi-square test to compare the distribution characteristics of cell numbers of each branch. The P values of Chi-square test from 13 paired samples are listed in Table 1. The P value of Chi-square test less than 0.01 is considered significant. For patients 7 and 12, the branch structures of phylogenetic tree are similar. But the branch



(a) The phylogenetic tree of ductal carcinoma in situ

(b) The phylogenetic tree of invasive ductal carcinoma

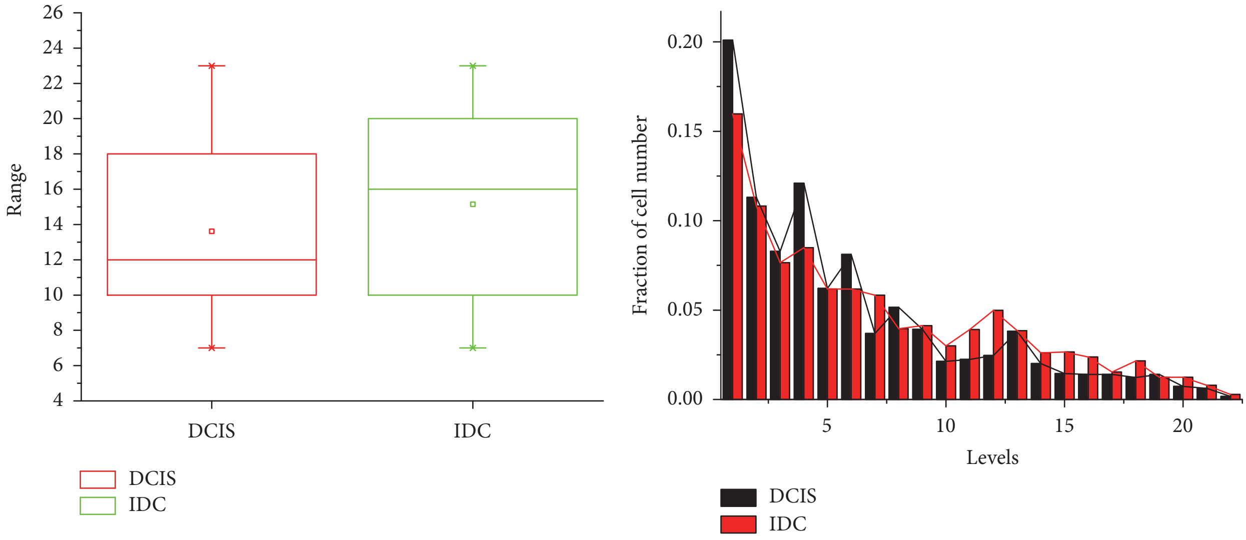
Figure 2: The comparison of BC phylogenetic trees.

structures of the remaining 11 paired samples are significantly different, which means that, under different selection pressures, the pathways of tumor subcellular amplification also change. As shown in Figure 2, which is an example of tumor phylogenetic tree from patient 5, Figures 2(a) and 2(b) are, respectively, from DCIS and IDC samples. The node in red is Steiner node and the weight is evolution distance between two nodes. The DCIS phylogenetic tree is more balanced, with more cells concentrated in the first four levels.

The cells number of phylogenetic tree across levels between DCIS and IDC tumor shows a noticeable difference. The *P* value of Chi-square test across the first twenty-two levels is listed in Table 1; the root node is on level zero. For the 13 paired samples, there are 11 cases with statistical significance. The hierarchical topology of primary and metastasis trees is similar in patients 3 and 6. We also analyze the depth characteristics of trees and corresponding fraction of cell number at each level. From Figure 3(a), the depth of DCIS tree is not distinctly different from IDC. The cell number distribution across different levels is illustrated in Figure 3(b). For the first six levels, the cell distribution of DCIS is more concentrated with a greater proportion compared with IDC. The cells gather in the first six levels up to 66% in DCIS and 55% in IDC. The number of cells decreases with the increment of tree levels, especially for DCIS. We also compare the edge features of phylogenetic trees; each edge is the corresponding gene gain or loss in the tree topological structure. The *P* value of edge statistics is not significantly different between DCIS and IDC except for patient 6, which is listed in Table 1.

3.3. Results on Cervical Cancer Datasets

3.3.1. Statistical Analysis of Tree Feature. BDEP is applied to comparatively analyze the tree topology between paired primary tumor and metastasis samples. The four genes of cervical cancer are on different chromosomes, phylogenetic trees have at most eight branches, and we use Chi-square test to compare the distribution characteristics of cell numbers of each branch. The Chi-square test of branch structure from 16 paired samples shows significant differences, which is listed in Table 2. The tree topology structure of primary and metastasis tumor is quite different. As shown in Figure 4, which is an



(a) The level count comparison of DCIS and IDC phylogenetic tree

(b) The cell number comparison of DCIS and IDC phylogenetic tree

Figure 3: The level characteristics of BC phylogenetic tree.

example of tumor phylogenetic tree from patient 3, Figures 4(a) and 4(b) are, respectively, from primary and metastasis samples. The node in red is Steiner node and the weight is evolution distance between two nodes. The metastasis sample has less copy number profiles, and the corresponding tree has fewer levels but with more balanced and broader topological structure compared with primary one.

In order to find the most decisive gene to distinguish primary and metastasis samples, we analyze the significance of individual gene. For each gene, we compare the cell numbers of branches with gene loss and gain. From Table 2, it is obvious that gene LAMP3 is the most informative gene; there are seven cases showing significant difference (patients 5, 6, 7, 12, 13, 14, and 16), which is consistent with the findings of Kanao et al. [45] and Mine et al. [46]. The overexpression of LAMP3 is associated with an enhanced metastatic potential and may be a prognostic factor for cervical cancer [45]. The gene PRKAA1 is the least with only two significant cases (patients 3 and 11).

For the hierarchical structure of trees, the P value of Chi-square test across the first twelve levels is listed in Table 3. Among the 16 paired samples, there are 14 cases with statistical significance. The hierarchical topology of primary and metastasis trees is distinguishable except for patients 1 and 9. The depth characteristics of trees and corresponding fraction of cell number at each level are illustrated in Figure 5. Whether or not lymph node later metastasized, the level structure of primary tumor is not distinctly different, but much deeper than the metastasized one. The cell distribution of metastasis sample is more concentrated and most of them gather in the first six levels compared with primary stage tumor. The number of cells decreases with the increment of tree levels, especially for metastasis tumor. The cells gather in the first six levels up to 85% in metastasis tumor and 70% in primary tumor. The cells in primary tumor are more evenly distributed and extending to more levels. For the edge feature of phylogenetic tree, all the 16 paired samples show no significant difference, which is similar to breast cancer samples.

For the edge feature of phylogenetic tree, all the 16 paired samples show no significant difference, which is similar to breast cancer samples.

3.3.2. The Classification Evaluation on Tree Features. The performance to predict the state of the tumor according to topological features of trees is crucial, which provides diagnostic guidance for accurate medical treatment. We evaluate the tree features through classification experiments and compare them with the features directly from data. We use the support vector machines (SVM) as classifier, which is implemented in an open source machine learning Scikit-learn module for Python [47]. We perform three classification experiments on CC dataset and the average accuracy of 100 tests is considered as experimental result. The three classification experiments are as follows:

(1) Distinguishing primary from its corresponding metastatic samples, which is a 16 versus 16 samples' classification

(2) Distinguishing nonmetastasis primary from primary samples, which is a 15 versus 16 samples' classification

(3) Distinguishing primary and nonmetastasis primary samples from metastatic samples, which is a 16 versus 15 versus 16 samples' classification.

The dataset is divided into four parts: three of them are training sets and the remaining one is test set. The extracted features from tree topology are branch, level, and edge. There are two features derived from data: (i) maximum copy number of each gene; (ii) average copy number of each gene. BDEP also compares with the published FISHtrees algorithm [30], which is a state-of-the-art algorithm for

Table 2: The P value of branches χ tests between primary and metastasis samples of cervical cancer.

Sample ID	P value	P value of LAMP3	P value of PROX1	P value of PRKAA1	P value of CCND1
Patient 1	$2.56E-15$	$7.86E-01$	$3.01E-06$	$2.16E-01$	$4.97E-01$
Patient 2	$6.87E-18$	$4.05E-02$	$7.49E-03$	$9.56E-01$	$6.32E-01$
Patient 3	$1.23E-48$	$8.71E-01$	$2.90E-01$	$2.22E-03$	$3.80E-48$
Patient 4	$1.00E-48$	$3.74E-01$	$1.55E-10$	$5.24E-02$	$1.41E-01$
Patient 5	$1.39E-17$	$4.65E-05$	$6.50E-02$	$5.00E-01$	$8.74E-01$
Patient 6	$1.20E-18$	$3.20E-09$	$6.01E-02$	$3.51E-02$	$4.48E-03$
Patient 7	$3.64E-28$	$1.96E-06$	$5.76E-01$	$9.55E-01$	$5.09E-02$
Patient 8	$8.17E-72$	$5.47E-01$	$1.99E-20$	$1.11E-01$	$3.45E-03$
Patient 9	$1.52E-30$	$6.03E-02$	$7.52E-02$	$8.10E-01$	$9.01E-01$
Patient 10	$8.15E-10$	$4.22E-02$	$6.22E-01$	$1.44E-01$	$9.26E-06$
Patient 11	$1.21E-31$	$5.65E-01$	$6.07E-01$	$1.84E-12$	$5.63E-05$
Patient 12	$6.98E-55$	$1.15E-26$	$1.41E-06$	$5.67E-01$	$7.89E-01$
Patient 13	$4.71E-73$	$6.11E-35$	$9.56E-01$	$1.89E-02$	$1.39E-03$
Patient 14	$2.70E-18$	$2.29E-06$	$1.48E-02$	$5.17E-02$	$1.20E-02$
Patient 15	$7.77E-22$	$6.39E-01$	$1.72E-03$	$2.36E-02$	$3.81E-01$
Patient 16	$1.19E-27$	$3.06E-03$	$8.23E-01$	$7.50E-01$	$3.53E-01$

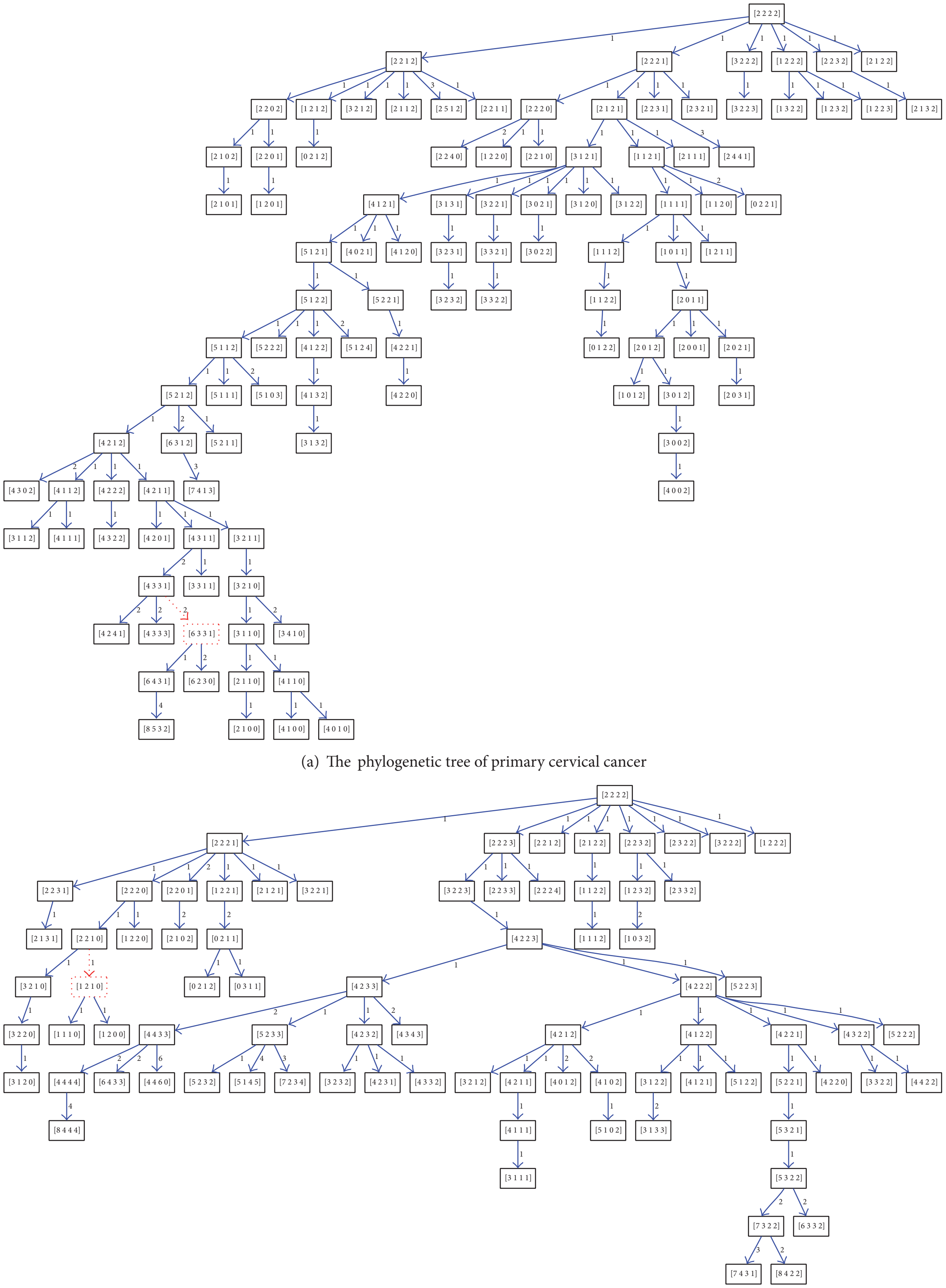
Table 3: The P value of levels and edges χ tests between primary and metastasis samples of cervical cancer.

Sample ID	P value of levels	P value of edges
Patient 1	$2.16E-02$	$9.35E-01$
Patient 2	$9.81E-09$	$6.48E-01$
Patient 3	$3.66E-17$	$8.04E-01$
Patient 4	$1.43E-05$	$9.06E-01$
Patient 5	$2.79E-07$	$3.34E-01$
Patient 6	$6.19E-09$	$6.82E-01$
Patient 7	$3.46E-04$	$9.64E-01$
Patient 8	$1.22E-07$	$7.97E-01$
Patient 9	$1.30E-02$	$9.25E-01$
Patient 10	$2.17E-09$	$8.28E-01$
Patient 11	$3.84E-10$	$4.98E-01$
Patient 12	$1.92E-15$	$2.49E-01$
Patient 13	$6.76E-17$	$2.87E-01$
Patient 14	$2.34E-06$	$6.75E-01$
Patient 15	$7.85E-03$	$6.48E-01$
Patient 16	$1.02E-16$	$9.90E-01$

phylogenetic tree based on FISH platform; the result is shown in Figure 6. The experiment distinguishing primary from its corresponding metastatic samples works best, followed by the classification between primary samples. The effect of distinguishing primary, nonmetastasis primary, and metastatic samples is poor for all features. Among all the features, the level feature achieves the highest accuracy, which shows that the degree of cell differentiation varies widely for tumors of different states. The data-based average feature shows in general the worst performance. Also interestingly, the Chi-square tests of branch structure are significant for all 16 paired samples, but classification effect is not as good as expected, even worse than edge feature. FISHtrees works better than BDEP for branch structure feature, but not for edge and level features. Overall, the classification accuracy of tree-based feature is better than data-based feature.

4. Conclusion

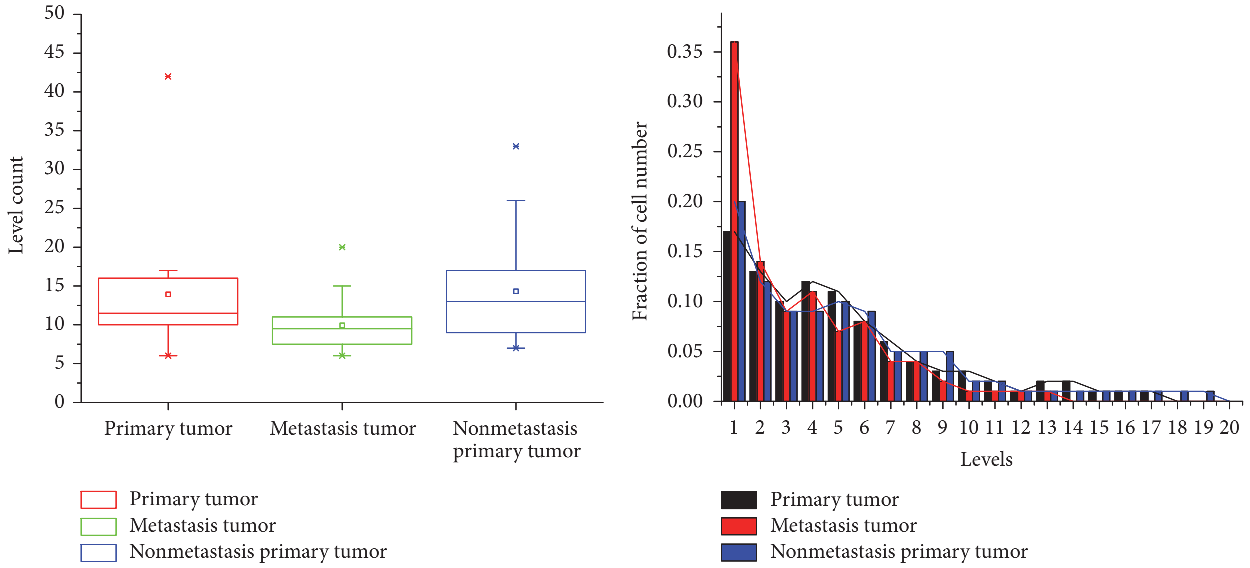
In this paper, we propose a binary differential evolution algorithm (BDEP) to construct tumor phylogenetic tree via CNV data on FISH platform. Tumor phylogenetic tree inference can be treated as minimum Steiner tree problem in directed graph, which cannot be solved in polynomial time unless no Steiner node exists. The binary differential evolution is a heuristic algorithm with advantages of fast convergence and strong robustness, which provides good approximate solutions with reduced running time. Experimental results on real datasets show that the branch and hierarchical structures



(a) The phylogenetic tree of primary cervical cancer

(b) The phylogenetic tree of lymph node metastasis cervical cancer

Figure 4: The comparison of CC phylogenetic trees.



(a) The level count comparison of primary, metastasis, and nonmetastasis primary tumor

(b) The cell number comparison of primary, metastasis, and nonmetastasis primary tumor

Figure 5: The level characteristics of CC phylogenetic tree.

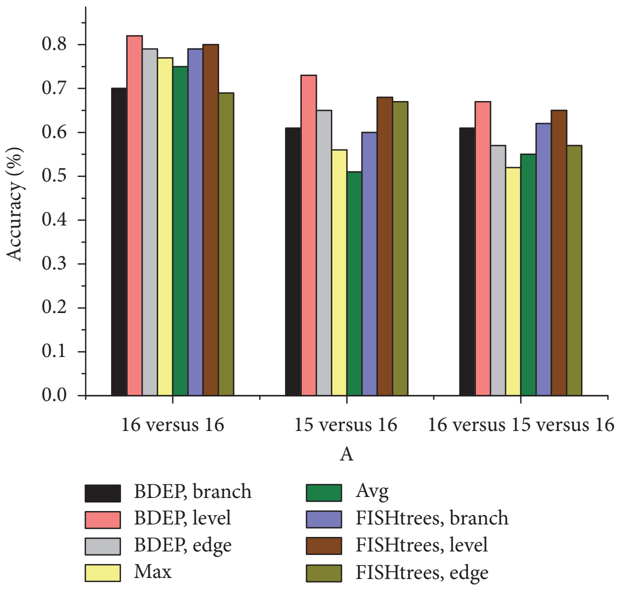


Figure 6: The SVM classification results of different features.

have significant differences for tumors of different states. And the gene under different selection pressures would lead to the different pathways of tumor subcellular expansion. The results on classification experiments show that our tree-based features are in general better than data-based features in distinguishing tumor, which provides more accurate and more comprehensive pathological guidance for clinical diagnosis and treatment. The association between genes is the key point to build and understand tumor progression; combining CNV data with other omics data (RNA and DNA methylation) would be a better strategy for tumor phylogenetic tree inference.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study is supported by the Program for New Century Excellent Talents in University (Grant no. NCET-10-0365), National Natural Science Foundation of China (Grant nos. 11171369, 61272395, 61370171, 61300128, 61472127, 61572178, and 61672214), National Natural Science Foundation of Hunan Province (Grant no. 12JJ2041), and the Planned Science and Technology Project of Hunan Province (Grant nos. 2009FJ3195 and 2012FJ2012).

References

[1] R. Weinberg, The Biology of Cancer. Garland science, 2013.

[2] P. C. Nowell, “The clonal evolution of tumor cell populations,” *Science*, vol. 194, no. 4260, pp. 23–28, 1976.

[3] C. Swanton, “Intratumor heterogeneity: Evolution through space and time,” *Cancer Research*, vol. 72, no. 19, pp. 4875–4882, 2012.

[4] M. Greaves and C. C. Maley, “Clonal evolution in cancer,” *Nature*, vol. 481, no. 7381, pp. 306–313, 2012.

[5] N. Beerenwinkel, R. F. Schwarz, M. Gerstung, and F. Markowetz, “Cancer evolution: Mathematical models and computational inference,” *Systematic Biology*, vol. 64, no. 1, pp. e1–e25, 2015.

[6] N. Navin, A. Krasnitz, L. Rodgers et al., “Inferring tumor progression from genomic heterogeneity,” *Genome Research*, vol. 20, no. 1, pp. 68–80, 2010.

[7] S. Nik-Zainal, P. Van Loo, D. C. Wedge et al. et al., “The life history of 21 breast cancers,” *Cell*, vol. 149, no. 5, pp. 994–1007, 2012.

[8] F. Strino, F. Parisi, M. Micsinai, and Y. Kluger, “TrAp: a tree approach for fingerprinting subclonal tumor composition,” *Nucleic Acids Research*, vol. 41, no. 17, p. e165, 2013.

[9] W. Jiao, S. Vembu, A. G. Deshwar, L. Stein, and Q. Morris, “Inferring clonal evolution of tumors from single nucleotide somatic mutations,” *BMC Bioinformatics*, vol. 15, no. 1, article no. 35, 2014.

[10] C. A. Miller, B. S. White, N. D. Dees et al., “SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution,” *PLoS Computational Biology*, vol. 10, no. 8, Article ID e1003665, 2014.

[11] I. Hajirasouliha, A. Mahmoody, and B. J. Raphael, “A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data,” *Bioinformatics*, vol. 30, no. 12, pp. I78–I86, 2014.

[12] M. El-Kebir, L. Oesper, H. Acheson-Field, and B. J. Raphael, “Reconstruction of clonal trees and tumor composition from multi-sample sequencing data,” *Bioinformatics*, vol. 31, no. 12, pp. i62–i70, 2015.

[13] V. Popic, R. Salari, I. Hajirasouliha, D. Kashef-Haghighi, R. B. West, and S. Batzoglou, “Fast and scalable inference of multi-sample cancer lineages,” *Genome Biology*, vol. 16, no. 1, article no. 91, 2015.

[14] A. Roth, J. Khattra, D. Yap et al., “PyClone: statistical inference of clonal population structure in cancer,” *Nature Methods*, vol. 11, no. 4, pp. 396–398, 2014.

[15] R. Redon, S. Ishikawa, K. R. Fitch et al., “Global variation in copy number in the human genome,” *Nature*, vol. 444, no. 7118, pp. 444–454, 2006.

[16] B. E. Stranger, M. S. Forrest, M. Dunning et al., “Relative impact of nucleotide and copy number variation on gene phenotypes,” *Science*, vol. 315, no. 5813, pp. 848–853, 2007.

[17] A. Subramanian, S. Shackney, and R. Schwartz, “Inference of tumor phylogenies from genomic assays on heterogeneous samples,” *Journal of Biomedicine and Biotechnology*, vol. 2012, Article ID 797812, 16 pages, 2012.

[18] L. Oesper, A. Mahmoody, and B. J. Raphael, “THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data,” *Genome Biology*, vol. 14, no. 7, article no. R80, 2013.

[19] G. Ha, A. Roth, J. Khattra et al., “TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data,” *Genome Research*, vol. 24, no. 11, pp. 1881–1893, 2014.

[20] A. G. Deshwar, S. Vembu, C. K. Yung, G. H. Jang, L. Stein, and Q. Morris, “PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors,” *Genome Biology*, vol. 16, no. 1, article no. 35, 2015.

[21] Y. Jiang, Y. Qiu, A. J. Minn, and N. R. Zhang, “Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing,” *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 113, no. 37, pp. E5528–E5537, 2016.

[22] Y. Li and X. Xie, “Deconvolving tumor purity and ploidy by integrating copy number alterations and loss of heterozygosity,” *Bioinformatics*, vol. 30, no. 15, pp. 2121–2129, 2014.

[23] Z. Yu, A. Li, and M. Wang, “CloneCNA: detecting subclonal somatic copy number alterations in heterogeneous tumor samples from whole-exome sequencing data,” *BMC Bioinformatics*, vol. 17, no. 1, article no. 310, 2016.

[24] M. El-Kebir, G. Satas, L. Oesper, and B. J. Raphael, “Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures,” *Cell Systems*, vol. 3, no. 1, pp. 43–53, 2016.

[25] E. Shapiro, T. Biezuner, and S. Linnarsson, “Single-cell sequencing-based technologies will revolutionize whole-organism science,” *Nature Reviews Genetics*, vol. 14, no. 9, pp. 618–630, 2013.

[26] N. Navin, J. Kendall, J. Troge et al., “Tumour evolution inferred by single-cell sequencing,” *Nature*, vol. 472, no. 7341, pp. 90–95, 2011.

[27] G. Pennington, C. A. Smith, S. Shackney, and R. Schwartz, “Reconstructing tumor phylogenies from heterogeneous single-cell data,” *Journal of Bioinformatics and Computational Biology*, vol. 5, no. 2 A, pp. 407–427, 2007.

[28] S. A. Chowdhury, S. E. Shackney, K. Heselmeyer-Haddad, T. Ried, A. A. Schäffer, and R. Schwartz, “Phylogenetic analysis of multiprobe fluorescence in situ hybridization data from tumor cell populations,” *Bioinformatics*, vol. 29, no. 13, pp. i189–i198, 2013.

[29] S. A. Chowdhury, S. E. Shackney, K. Heselmeyer-Haddad, T. Ried, A. A. Schäffer, and R. Schwartz, “Algorithms to model single gene, single chromosome, and whole genome copy number changes jointly in tumor phylogenetics,” *PLoS Computational Biology*, vol. 10, no. 7, Article ID e1003740, 2014.

[30] S. A. Chowdhury, E. M. Gertz, D. Wangsa et al., “Inferring models of multiscale copy number evolution for single-tumor phylogenetics,” *Bioinformatics*, vol. 31, no. 12, pp. i258–i267, 2015.

[31] E. M. Gertz, S. A. Chowdhury, W.-J. Lee et al., “FISHtrees 3.0: Tumor phylogenetics using a ploidy probe,” *PLoS ONE*, vol. 11, no. 6, Article ID e0158569., 2016.

[32] J. Liu, S. Ranka, and T. Kahveci, “Markers improve clustering of CGH data,” *Bioinformatics*, vol. 23, no. 4, pp. 450–457, 2007.

[33] R. Shamir, M. Zehavi, and R. Zeira, “A linear-time algorithm for the copy number transformation problem,” in *27th Annual Symposium on Combinatorial Pattern Matching (CPM 2016)*, vol. 54 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[34] S. L. Hakimi, “Steiner’s problem in graphs and its implications,” *Networks*, vol. 1, no. 2, pp. 113–133, 1971.

[35] F. K. Hwang, D. S. Richards, and P. Winter, *The Steiner Tree Problem*, vol. 53, Elsevier, 1992.

[36] Y.-J. Chu and T.-H. Liu, “On shortest arborescence of a directed graph,” *Scientia Sinica*, vol. 14, no. 10, p. 1396, 1965.

[37] R. M. Karp, “Reducibility among combinatorial problems,” in *Complexity of Computer Computations*, pp. 85–103, Springer, New York, NY, USA, 1972.

[38] R. Storn and K. Price, “Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces,” *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[39] J. Ilonen, J.-K. Kamarainen, and J. Lampinen, “Differential evolution training algorithm for feed-forward neural networks,” *Neural Processing Letters*, vol. 17, no. 1, pp. 93–105, 2003.

[40] R. Joshi and A. C. Sanderson, “Minimal representation multisensor fusion using differential evolution,” *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 29, no. 1, pp. 63–76, 1999.

[41] T. Rogalsky, S. Kocabiyik, and R. W. Derksen, “Differential evolution in aerodynamic optimization,” *Canadian Aeronautics and Space Journal*, vol. 46, no. 4, pp. 183–190, 2000.

[42] R. Storn, “On the usage of differential evolution for function optimization,” in *Proceedings of the Biennial Conference of the North American Fuzzy Information Processing Society (NAFIPS ’96)*, pp. 519–523, June 1996.

[43] D. Wangsa, K. Heselmeyer-Haddad, P. Ried et al., “Fluorescence in situ hybridization markers for prediction of cervical lymph node metastases,” *The American Journal of Pathology*, vol. 175, no. 6, pp. 2637–2645, 2009.

[44] K. Heselmeyer-Haddad, L. Y. Berroa Garcia, A. Bradley et al., “Single-cell genetic analysis of ductal carcinoma in situ and invasive breast cancer reveals enormous tumor heterogeneity yet conserved genomic imbalances and gain of MYC during progression,” *The American Journal of Pathology*, vol. 181, no. 5, pp. 1807–1822, 2012.

[45] H. Kanao, T. Enomoto, T. Kimura et al., “Overexpression of LAMP3/TSC403/DC-LAMP promotes metastasis in uterine cervical cancer,” *Cancer Research*, vol. 65, no. 19, pp. 8640–8645, 2005.

[46] K. L. Mine, N. Shulzhenko, A. Yambartsev et al., “Gene network reconstruction reveals cell cycle and antiviral genes as major drivers of cervical cancer,” *Nature Communications*, vol. 4, article 1806, 2013.

[47] F. Pedregosa, G. Varoquaux, A. Gramfort et al., “Scikit-learn: machine learning in python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.