On the Minimum Copy Number Generation Problem in Cancer Genomics

Letu Qingge, Xiaozhou He, Zhihui Liu, and Binhai Zhu

March 12, 2019

Overview

1 Introduction

2 The Minimum Copy Number Generation Problem
- Other formulations
- A greedy solution for the MCNG Problem

3 Conclusions

Motivation

- Some tumor types' hererogeneity is better reflected in the genomic rearrangements and duplications rather than small mutations.
- Easier to produce phylogeny on copy number profiles (CNPs) than genomic data in such cases.
- Need to be able to compute the minimum number of duplications/deletions to transform one profile into another.

General Definitions

Definition 1

A genome is a string $G = \Sigma^n$ where $\Sigma = \{g_1, g_2, ..., g_m\}$ is a set of m genes and $n \geq m$.

Definition 2

$cnp(G) =< c_1, c_2, ..., c_m >$ where $c_i \geq 0$ is the number of times g_i appears in G.

General Definitions

Definition 1

A genome is a string $G = \Sigma^n$ where $\Sigma = \{g_1, g_2, ..., g_m\}$ is a set of m genes and $n \geq m$.

Definition 2

$cnp(G) =< c_1, c_2, ..., c_m >$ where $c_i \geq 0$ is the number of times g_i appears in G.

Definition 3

A genome G is exemplar if it is a permutation of the set of genes $g_1, g_2, ...g_m$.

General Definitions

Definition 4

Two operations on CNPs, duplications and deletions.

- $\texttt{duplicate}(< c_1, c_2, ..., c_m >, i, j)$
  $=< c_1, c_2, ..., c_i + 1, c_{i+1} + 1, ...c_j + 1, ...c_m >$
- $\texttt{delete}(< c_1, c_2, ..., c_m >, i, j)$
  $=< c_1, c_2, ..., max(c_i - 1, 0), max(c_{i+1} - 1), ...max(c_j - 1), ...c_m >$

General Definitions

Definition 4

Two operations on CNPs, duplications and deletions.

- $\texttt{duplicate}(< c_1, c_2, ..., c_m >, i, j)$
  $=< c_1, c_2, ..., c_i + 1, c_{i+1} + 1, ...c_j + 1, ...c_m >$
- $\texttt{delete}(< c_1, c_2, ..., c_m >, i, j)$
  $=< c_1, c_2, ..., max(c_i - 1, 0), max(c_{i+1} - 1), ...max(c_j - 1), ...c_m >$

Remark 1

Functions above don't necessarily represent possible duplication and deletion events in a genome.

General Definitions

Definition 5

Two operations on CNPs, duplications and deletions.

- `duplicate`(G, i, j, k)
  Creates a copy of the genome from index i to j and inserts it at some k s.t. $k < i$ and $k \geq j$.
- `delete`(G, i, j)
  Deletes a segment of the genome from index i to j.

General Definitions

Definition 5

Two operations on CNPs, duplications and deletions.

- `duplicate`(G, i, j, k)
  Creates a copy of the genome from index i to j and inserts it at some k s.t. $k < i$ and $k \geq j$.
- `delete`(G, i, j)
  Deletes a segment of the genome from index i to j.

Definition 6

A duplication is considered *tandem* if $k = j$ or $k = i$.

Genomes vs Profiles

Consider an input genome $< a, b, c, d, e, f >$ and a desired CNP $H =< 1, 3, 2, 3, 3, 1 >$. When only being able to perform operations on the input CNP, we have minimum number of 3 operations seen by

$$< 1, \underline{1, 1, 1, 1}, 1 > \rightarrow < 1, \underline{2, 2, 2, 2}, 1 > \rightarrow < 1, 3, \underline{3}, 3, 3, 1 > \rightarrow H$$

Genomes vs Profiles

Consider an input genome $< a, b, c, d, e, f >$ and a desired CNP $H =< 1, 3, 2, 3, 3, 1 >$. When only being able to perform operations on the input CNP, we have minimum number of 3 operations seen by

$$< 1, \underline{1, 1, 1, 1}, 1 > \rightarrow < 1, \underline{2, 2, 2, 2}, 1 > \rightarrow < 1, 3, \underline{3}, 3, 3, 1 > \rightarrow H$$

Now consider a set of 2 operations on the input genome

$$< a, \underline{b, c, d, e}, f > \rightarrow < a, b, c, \underline{d, e, b}, c, d, e, f > \rightarrow G'$$

$$\mathrm{cnp}(G' =< a, b, c, d, e, b, d, e, b, c, d, e, f >) = H$$

1 Introduction

2 The Minimum Copy Number Generation Problem

- Other formulations
- A greedy solution for the MCNG Problem

3 Conclusions

Problem Statement

Definition 7 (MCNG Problem)

Given a genome G and a copy number profile H, is there a sequence of at most k operations that can convert G into some G' such that $\texttt{cnp}(G') = H$?

We can easily constrain the problem in many ways:

- Tandem duplications only
- Exemplar genomes only
- No deletions

Problem Statement

Theorem 2

The MCNG problem when duplications must be tandem is NP-Hard

We show this by reducing from the exact cover by 3-sets problem (X3C).

Definition 8 (X3C)

Given $S = \{S_1, S_2, ..., S_m\}$ s.t. $|S_i| = 3$ and $S_i \subseteq X = \{x_1, x_2, ..., x_n\}$ and $|X| = 3q$, is there a selection of q 3-sets S_i such that their union is the base set X?

Reduction

Transform an instance of X3C to MCNG by creating a genome G such that

$$G = < T_1, p_1, T_2, p_2, ..., p_{m-1}, T_m >$$

- Each T_i corresponds to the elements in S_i and each p_i is a peg gene.
- Let $f(x_i)$ denote the frequency of x_i in the multiset of $\bigcup_j S_j$.
- For each p_i, we set the target copy number to 1.
- For each x_i, we set the target copy number to $f(x_i) + 1$.

Reduction

Transform an instance of X3C to MCNG by creating a genome G such that

$$G =< T_1, p_1, T_2, p_2, ..., p_{m-1}, T_m >$$

- Each T_i corresponds to the elements in S_i and each p_i is a peg gene.
- Let $f(x_i)$ denote the frequency of x_i in the multiset of $\bigcup_j S_j$.
- For each p_i, we set the target copy number to 1.
- For each x_i, we set the target copy number to $f(x_i) + 1$.

Remark 3

We can obtain a duplication of $T_i, T_{i+1}...T_j$ in $j - i + 1$ operations. Therefore we have an average of 1 operation per T block.

Proof

Lemma 4

Let G' be obtained from G with the allowed operations and $cnp(G') = H$. Let $t = G' - G$ and $|t| = 3q$ and $cnp(t) = I_{3q}$. Then

t is canonical (a concatenation of T_is obtained by canonical operations) $\iff$ t is obtained with q allowable operations from G.

Corollary 5

If the MCNG problem admits a solution of q tandem duplication or deletion operations, then the X3C instance has a solution.

Proof ctd

Forward.

As we showed earlier, we can obtain each T_i with 1 operation on average. □

Backward.

If t is not obtained in a canonical way, then at least one T block is obtained using a non-canonical operation. However, this would mean that T is obtained using more than one operation. Since there are still $q-1$ other T blocks to obtain, then at this point we need at least $q+1$ total operations to obtain them all. □

Exemplar case

Remark 6

When the input genome P is exemplar, whether the corresponding Exemplar Minimum Copy Number Generation (EMCNG) problem is NP-hard is open.

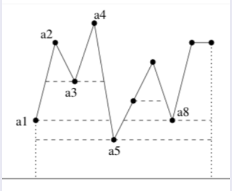
Theorem 7

In the Exemplar Minimum Copy Number Generation problem with tandem duplications, if $c_i = 2^{a_i}$ with $a_i \geq 0$, for all $i = 1..m$, then the optimal solution can be computed in time linear in terms of the length of the input, plus $O(m \log m)$.

Proof

Proof.

Computing trapezoids takes $O(m \log m)$ time, and there are $O(m)$ of them. The solution cost is the sum of all trapezoids' heights.



□

Remark 8

The paper shows an example of where the greedy algorithm is not optimal for the general case.

MCNG$\leq$ and MCNG$\geq$

Definition 9 (MCNG$\leq$ ($\geq$))

Given a sequence G over m genes and a CNP $C = C_1 \cdot C_2$ where $C_1 =< c_1, c_2, ..., c_q >$ and $C_2 =< c_{q+1}, c_{q+2} ..., c_m >$ and an integer k, is there a sequence of at most k deletions (tandem duplications) that result in G' such that $\mathrm{cnp}(G')[1..q] \leq (\geq) C1$ and $\mathrm{cnp}(G')[q+1..m] = C2$

MCNG$\leq$ and MCNG$\geq$

Definition 9 (MCNG$\leq$ ($\geq$))

Given a sequence G over m genes and a CNP $C = C_1 \cdot C_2$ where $C_1 =< c_1, c_2, ..., c_q >$ and $C_2 =< c_{q+1}, c_{q+2}..., c_m >$ and an integer k, is there a sequence of at most k deletions (tandem duplications) that result in G' such that $\mathrm{cnp}(G')[1..q] \leq (\geq) C1$ and $\mathrm{cnp}(G')[q+1..m] = C2$

Theorem 9

Both problems can be reduced to set cover via a similar method as the one shown earlier.

Corollary 10

There is no $O(\log n)$ factor approximation for either problem above unless $P = NP$.

Algorithm

1. For current sequence G' and $\texttt{cnp}(G') = C'$, let $Z = H - C'$
2. Construct a string s as follows:
   For each index i in G', let $g_j = G[i]$. We now assign $s[i] = +$ if $Z[i] > 0$, $s[i] = -$ if $Z[i] < 0$, and $s[i] = x$ otherwise.
3. Find longest substring of $+$ or $-$. If it is $-$, perform a delete operation on that substring. Otherwise, perform a duplication on the substring.
4. Repeat

Results

- Duplication-heavy vs deletion-heavy simulations didn't make a significant difference in algorithm perfomance.
- Performance dropped with increase in k, m, and duplication/deletion size.
- Initial seed is always a permutation of GG i.e. a permutation of a doubling of G.

1 Introduction

2 The Minimum Copy Number Generation Problem
- Other formulations
- A greedy solution for the MCNG Problem

3 Conclusions

Conclusions and Future Directions

- MCNG with tandem duplications is NP-Hard, but are there any approximation or FPT algorithms for it? MCNG($\geq$)($\leq$) we know is only approximable to $O(\log n)$ and has no FPT.
- The Exemplar MCNG with duplications only is tractable when the desired CNP is all powers of 2. It is open whether or not it is polynomially solvable when the desired CNP is arbitrary.
- The MCNG problem with no restrictions is open.