Genome Rearrangement

Carlos Cotta

Departamento de Lenguajes y Ciencias de la Computación
Universidad de Málaga

http://www.lcc.uma.es/∼ccottap

Comput Eng, Softw Eng, Comput Sci & Math – 2020-2021

Index

1. Lab Session Unit V: Greedy Algorithms
    - Problem Statement
    - Greedy Algorithms Approach

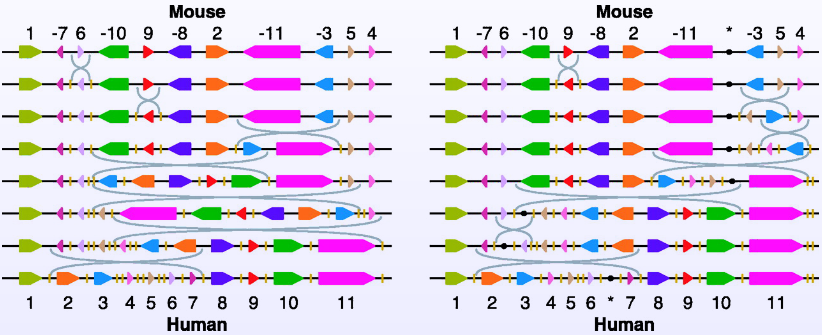
Biological Context

Genes are sequences of nucleotides that encode the synthesis of biochemical material (be it RNA or proteins).

Biological processes can result in mutations in the genome. Some mutations are local and affect a single nucleotide (substitutions, insertions, deletions) while other can be non-local and affect long stretches of the sequence (inversions, transpositions, translocations).

Non-local mutations are rare, but accumulate over the evolutionary timescale. They thus provide a good proxy for evolutionary distance between species.

Biological Context



Different species may have the same genes but in a different order (and orientation).

Problem Statement

Genome as a Permutation

Let the order of genes in an organism be represented as a (signed) permutation $\pi = [\pi_1 \pi_2 \dots \pi_n]$:

- If gene orientation is not important, each $\pi_i \in \{1, \dots, n\}$ is a non-repeated positive integer.
- If gene orientation is important, each π_i can be positive or negative, and $|\pi_i| \in \{1, \dots, n\}$ is a non-repeated positive integer.

Problem Statement

Reversal

Reversal

A reversal $\rho(i,j)$ is an operation that affects the portion of the genome between positions i and j (both inclusive), reversing this segment.

Let $\pi = [\pi_1\pi_2\ldots\pi_n]$. Then, if π is unsigned:

$$\pi\cdot\rho(i,j) = [\pi_1\ldots\pi_{i-1}\,\underbrace{\pi_j\pi_{j-1}\ldots\pi_{i+1}\pi_i}_{\text{reversed segment}}\,\pi_{j+1}\ldots\pi_n]$$

If π is signed, the sign of each π_k, $i \leqslant k \leqslant j$, is flipped as well.

Problem Statement

Minimal Reversal Distance

Let π and π' be two organisms. The reversal distance between them is the minimum number of reversals required to transform one of the permutations into the other.

We can assume w.l.o.g. that one of the permutations, say π', is the positive identity permutation, i.e., $\pi'_i = i$, $1 \leqslant i \leqslant n$.

Sorting Permutations by Reversals

Given a (signed) permutation π, find the shortest sequence of reversals required to transform it into the positive identity permutation.

Unsigned Permutations

The problem of sorting an unsigned permutation by reversals is NP-hard.

Let us consider a naive approach to the problem, namely PrefixSort:

```
for each $i \in \{1, \ldots, n\}$ do
    if $\pi_i \neq i$ then
        $j \leftarrow$ index for which $\pi_j = i$
        $\pi \leftarrow \pi \cdot \rho(i, j)$
    endif
endfor
```

Note that PrefixSort requires at most n reversals to produce the identity permutation.

Unsigned Permutations

PrefixSort in Action

Let $\pi = [67812345]$. PrefixSort requires 6 steps:

[18762345]

[12678345]

[12387645]

[12346785]

[12345876]

[12345678]

But it can be sorted in fewer steps:

[54321876]

[12345876]

[12345678]

Unsigned Permutations

A Greedy Template for Genome Rearrangement

Greedy Genome Rearrangement

```
func GreedyRearrangement ($\downarrow\pi$: Permutation$\langle n \rangle$): List$\langle Op \rangle$
variables
    $sol, candidates$: List$\langle Op \rangle$
    $o, best$: Op
begin
    $\pi \leftarrow [0, \pi, n+1]$                                        // framing the permutation
    $sol \leftarrow \langle\rangle$
    $candidates \leftarrow$ getCandidates($\pi$)                          // get applicable operations
    while $candidates \neq \langle\rangle$ do
        $best \leftarrow \arg\max\{\text{getQuality}(o, \pi) \mid o \in \text{candidates}\}$ // pick best operation
        $sol$.add($best$)
        $\pi \leftarrow best$.apply($\pi$)                                // apply operation
        $candidates \leftarrow$ getCandidates($\pi$)                      // get applicable operations
    endwhile
    return sol
end
```

Unsigned Permutations

Sorting by Breakpoints

Let $0 \leqslant i \leqslant n$. A pair of elements π_i, π_{i+1} are an adjacency if $|\pi_i - \pi_{i+1}| = 1$. Otherwise they are a breakpoint.

$$[0|678|12345|9]$$

The identity permutation is the only one that –when framed (i.e., extended with $\pi_0 = 0$ and $\pi_{n+1} = n + 1$)– has no breakpoints.

It only makes sense to apply a reversal between breakpoints.

A reversal between breakpoints may remove 0, 1 or 2 of them.

Greedy idea: pick the reversal that maximizes the number of breakpoints removed.

Unsigned Permutations

Sorting by Breakpoints

$$\text{getCandidates}(\pi) = \{\rho(i+1, j) \mid i < j, \{i, j\} \subseteq \text{breakpoints}(\pi)\}$$

$$\text{getQuality}(\rho(i, j), \pi) = [|\pi_{i-1} - \pi_j| = 1] + [|\pi_{j+1} - \pi_i| = 1]$$

π	candidates	best	quality
[0|6 7 8|1 2 3 4 5|9]	$\rho(1,3)$, $\rho(1,8)$, $\rho(4,8)$	$\rho(1,3)$	0 (tie)
[0|8 7 6|1 2 3 4 5|9]	$\rho(1,3)$, $\rho(1,8)$, $\rho(4,8)$	$\rho(1,8)$	1
[0|5 4 3 2 1|6 7 8 9]	$\rho(1,5)$	$\rho(1,5)$	2
[0 1 2 3 4 5 6 7 8 9]	–	–	–

This algorithm is not optimal but –if an adequate tie-breaking procedure is picked– it provides a 2-approximation to the optimal.

Complementary Bibliography

N.C. Jones, P.A. Pevzner,
An Introduction to Bioinformatics Algorithms,
MIT Press, Cambridge MA, 2004

Image Credits

- Mouse and human genome rearrangment: Pavel Pevzner and Glenn Tesler, *PNAS*, doi: 10.1073/pnas.1330369100