Medians of permutations and gene orders

Sylvie Hamel
Université de Montréal

Work in collaboration with:

- Guillaume Blin
- Maxime Crochemore
- Stéphane Vialette

from Université Marne-la-Vallée
Medians of permutations and gene orders

Sylvie Hamel
Université de Montréal

Research trainees/programmers:
- Quentin Dejean
- Anthony Estebe
- Kelsey Lang

from Université Montpellier II and University of Victoria
Biological context:

1: Permutations
Biological context:

2: Signed Permutations
The general problem:

Given \( m \) permutations \( \pi_1, \pi_2, \ldots, \pi_m \) of \( \{1, 2, \ldots, n\} \) and a distance function \( d \), the median problem is to find a permutation \( \pi^* \) that is the “closest”, under the distance \( d \), to the \( m \) given permutations.
Biological context: phylogeny
Blanchette and Sankoff*:

In 1997, they study the problem of finding a median of 3 genomes (permutations) under the breakpoint distance.

The **breakpoint distance** between two permutations $\pi^1$ and $\pi^2$ is the number of pair of elements which are adjacent in $\pi^1$ but not in $\pi^2$.

Blanchette and Sankoff*:

In 1997, they study the problem of finding a median of 3 genomes (permutations) under the breakpoint distance.

\[
\pi^1 = [1, 2, 3, 4, 5, 6, 7, 8] \quad \text{distance} = 4
\]

\[
\pi^2 = [1, 5, 6, 8, 4, 3, 2, 7] \quad \text{distance} = 4
\]

Blanchette and Sankoff*:  
They described efficient heuristics to find a median of three circular genomes that have, or do not have, the same gene content.

Pe’er and Shamir**:  
They show that the median problem of three permutations or signed permutations under the breakpoint distance is NP-complete.

The Kendall-$\mathcal{T}$ distance:

Counts the number of pairwise disagreements between two permutations i.e.

$$d_{KT}(\pi, \sigma) = \sum_{i < j} \left( (\pi[i] < \pi[j] \text{ and } \sigma[i] > \sigma[j]) \right.$$

$$\left. \text{ or } (\pi[i] > \pi[j] \text{ and } \sigma[i] < \sigma[j]) \right)$$

The Kendall-$\mathcal{T}$ distance is equivalent to the “bubble-sort” distance i.e. the number of transpositions needed to transform one permutation into the other one.

We have $$d_{KT}(\pi, \iota) = \text{inv}(\pi)$$
Example:

\[ \pi = [1, 4, 2, 5, 3] \]

\[ \sigma = [3, 4, 1, 2, 5] \]

\[ d_{KT}(\pi, \sigma) = \]
Example:

\[ \pi = [1, 4, 2, 5, 3] \]

\[ \sigma = [3, 4, 1, 2, 5] \]

\[ d_{KT}(\pi, \sigma) = 1 + \]
Example:

$$\pi = [1, 4, 2, 5, 3]$$

$$\sigma = [3, 4, 1, 2, 5]$$

$$d_{KT}(\pi, \sigma) = 1 + 1 +$$
Example:

\[ \pi = [1, 4, 2, 5, 3] \]

\[ \sigma = [3, 4, 1, 2, 5] \]

\[ d_{KT}(\pi, \sigma) = 1 + 1 + \]
Example:

\[ \pi = [1, 4, 2, 5, 3] \]

\[ \sigma = [3, 4, 1, 2, 5] \]

\[ d_{KT}(\pi, \sigma) = 1 + 1 + 1 + \]
Example:

\[ \pi = [1, 4, 2, 5, 3] \]

\[ \sigma = [3, 4, 1, 2, 5] \]

\[ d_{KT}(\pi, \sigma) = 1 + 1 + 1 + \]
Example:

\[ \pi = [1, 4, 2, 5, 3] \]

\[ \sigma = [3, 4, 1, 2, 5] \]

\[ d_{KT}(\pi, \sigma) = 1 + 1 + 1 + \]
Example:

\[ \pi = [1, 4, 2, 5, 3] \]

\[ \sigma = [3, 4, 1, 2, 5] \]

\[ d_{KT}(\pi, \sigma) = 1 + 1 + 1 + 1 + \]
Example:

\[ \pi = [1, 4, 2, 5, 3] \]

\[ \sigma = [3, 4, 1, 2, 5] \]

\[ d_{KT}(\pi, \sigma) = 1 + 1 + 1 + 1 + 1 + 1 + \]

LIX Bioinformatics Colloquium

November 2010
Example:

\[ \pi = [1, 4, 2, 5, 3] \]

\[ \sigma = [3, 4, 1, 2, 5] \]

\[ d_{KT}(\pi, \sigma) = 1 + 1 + 1 + 1 + 1 + 1 \]
Example:

\[ \pi = [1, 4, 2, 5, 3] \]

\[ \sigma = [3, 4, 1, 2, 5] \]

\[ d_{KT}(\pi, \sigma) = 1 + 1 + 1 + 1 + 1 = 5 \]
Kendall-$\mathcal{T}$ distance between a permutation $\pi$ and a set of permutations $A = \{\pi^1, \ldots, \pi^m\}$:

$$d_{KT}(\pi, A) = \sum_{i=1}^{m} d_{KT}(\pi, \pi^i)$$
The problem of finding the median of a set of \( m \) permutations of \( \{1, 2, \ldots, n\} \) under the Kendall-\( \tau \) distance is best known in the literature as the Kemedy Score Problem.

In this problem, we have \( m \) voters that have to order \( n \) candidates from their best-liked candidate to their least-liked one.

The problem then consists in finding a “Kemedy consensus” i.e. an order of the candidates that agree the most with the orders of the voters.
Example:

\[
\begin{align*}
\pi^1 &= [2, 1, 3, 4] \\
\pi^2 &= [4, 1, 2, 3] \\
\pi^3 &= [4, 2, 3, 1] \\
\end{align*}
\]

Disagreement graph \( G(\pi) \) of a permutation \( \pi \) with respect to a set of permutations \( A \):

\[
G(\pi):
\]

\[
\begin{array}{cccc}
4 & 2 & 1 & 3 \\
\end{array}
\]
Example:

\[ \pi^1 = [2, 1, 3, 4] \]
\[ \pi^2 = [4, 1, 2, 3] \]
\[ \pi^3 = [4, 2, 3, 1] \]
\[ \pi = [4, 2, 1, 3] \]

Disagreement graph \( G(\pi) \) of a permutation \( \pi \) with respect to a set of permutations \( A \):

\( G(\pi) \):

\[ 4 \xrightarrow{1} 2 \xrightarrow{d_{KT}(\pi, \pi^1)} \pi \]
\[ 4 \xrightarrow{d_{KT}(\pi, \pi^2)} 2 \]
\[ 4 \xrightarrow{d_{KT}(\pi, \pi^3)} 2 \]

\[ 1 \]
\[ 3 \]
Example:

\[ \pi^1 = [2, 1, 3, 4] \]
\[ \pi^2 = [4, 1, 2, 3] \]
\[ \pi^3 = [4, 2, 3, 1] \]

Disagreement graph \( G(\pi) \) of a permutation \( \pi \) with respect to a set of permutations \( A \):

\[ G(\pi): \]

1

4 \rightarrow 2 \rightarrow 1 \rightarrow 3
Example:

\[ \pi^1 = [2, 1, 3, 4] \]

\[ \pi^2 = [4, 1, 2, 3] \]

\[ \pi^3 = [4, 2, 3, 1] \]

Disagreement graph \( G(\pi) \) of a permutation \( \pi \) with respect to a set of permutations \( \mathcal{A} \):
Example:

\[ \pi^1 = [2, 1, 3, 4] \]
\[ \pi^2 = [4, 1, 2, 3] \]
\[ \pi^3 = [4, 2, 3, 1] \]

Disagreement graph \( G(\pi) \) of a permutation \( \pi \) with respect to a set of permutations \( A \):

\[ G(\pi): \]

\( 4 \) \( \rightarrow \) \( 2 \) \( \rightarrow \) \( 1 \) \( \rightarrow \) \( 3 \)

\( 1 \) \( \rightarrow \) \( 2 \) \( \rightarrow \) \( 1 \) \( \rightarrow \) \( 3 \)

\( 1 \) \( \rightarrow \) \( 2 \) \( \rightarrow \) \( 1 \) \( \rightarrow \) \( 3 \)

\( 1 \) \( \rightarrow \) \( 2 \) \( \rightarrow \) \( 1 \) \( \rightarrow \) \( 3 \)

\( 1 \) \( \rightarrow \) \( 2 \) \( \rightarrow \) \( 1 \) \( \rightarrow \) \( 3 \)
Example:

\[
\pi^1 = [2, 1, 3, 4] \\
\pi^2 = [4, 1, 2, 3] \\
\pi^3 = [4, 2, 3, 1]
\]

Disagreement graph \( G(\pi) \) of a permutation \( \pi \) with respect to a set of permutations \( A \):
Example:

\[ \pi^1 = [2, 1, 3, 4] \]
\[ \pi^2 = [4, 1, 2, 3] \]
\[ \pi^3 = [4, 2, 3, 1] \]

Disagreement graph \( G(\pi) \) of a permutation \( \pi \) with respect to a set of permutations \( A \):

\( d_{KT}(\pi, \pi^1) \)
\( d_{KT}(\pi, \pi^2) \)
\( d_{KT}(\pi, \pi^3) \)

\( d_{KT} = 5 \)
Our problem:

Given a set of $m$ permutations $A \subseteq S_n$, we want to find a permutation $\pi^*$ such that

$$d_{KT}(\pi^*, A) \leq d_{KT}(\pi, A), \text{ for all } \pi \in S_n.$$

- If $m$ is the cardinality of the set $A \subseteq S_n$, the problem is NP-complete for $m \geq 4$, $m$ even*
- We will then considered the case $m \geq 3$, $m$ odd

Example where there is more than one median:

\[ A = \{ \pi^1 = [1, 2, 3], \pi^2 = [3, 1, 2], \pi^3 = [2, 3, 1] \} \]
Finding a median with a brute force algorithm: 
n! permutations de \{1, 2, 3, \ldots n\}

- n=5, 120 permutations
- n=6, 720 permutations
- n=7, 5040 permutations
- n=8, 40 320 permutations
- n=9, 362 880 permutations
- n=10, 3 628 800 permutations
- n=11, 39 916 800 permutations
- . . .
Lemma 1: If a pair of elements appear in the same order in all permutations of the set $A$, then they also appear in that order in all medians $\pi^*$.

Lemma 1 follows directly from the Extended Condorcet criterion (Truchon, 1998): If there is a partition $(C, C')$ of $\{1, 2, \ldots, n\}$ such that for any $x$ in $C$ and $y$ in $C'$ the majority prefers $x$ to $y$, then $x$ must be ranked above $y$.

The “original” Condorcet criterion, proposed by Marie Jean Antoine Nicolas de Caritat in 1785, marquis de Condorcet, stated that if there is some element of $\{1, 2, \ldots, n\}$ that defeats every other in pairwise simple majority voting, then this element should be ranked first.


M.-J. Condorcet, Essai sur l’application de l’analyse à la probabilité des décisions rendues à la pluralité des voix, 1785.
Lemma 1: If a pair of elements appear in the same order in all permutations of the set $A$, then they also appear in that order in all medians $\pi^*$.

Lemma 1 gives us a set of constraints which $\pi^*$ must satisfy.

Lemma 1 imply that in the disagreement graph of $\pi^*$, there are no arcs of weight $m$. 
Lemma 2: The order of any adjacent pair of elements in a median $\pi^*$ agree with the order of this pair of element in the majority of the permutations in $A$.

So in the disagreements graph of a median, all adjacent nodes are linked by an edge of weight $\leq \left\lfloor \frac{m}{2} \right\rfloor$, where $m$ is the cardinality of $A$. 
Comparaison des temps de l’algorithme force brute et algorithme force brute + contraintes

![Graph showing comparison of algorithm force brute and force brute + constraints]

- 0.001 (8)
- 0.004 (9)
- 0.03 (10)
- 0.26 (11)
- 2.26 (12)

1024 ~17 minutes

LIX Bioinformatics Colloquium

November 2010
Some definitions for our heuristic:

**Definition 1:**

Given \( \pi = \pi_1 \pi_2 \ldots \pi_n \), we call **cyclic movement** of a segment \( \pi[i..j] \), denoted \( c[i..j](\pi) \), the cycling shifting of the elements of the segment to the right or to the left:

\[
c_r[i, j](\pi) = \pi_1 \ldots \pi_{i-1} \pi_j \pi_{i+1} \ldots \pi_{j-1} \pi_{j+1} \ldots \pi_n
\]

\[
c_l[i, j](\pi) = \pi_1 \ldots \pi_{i-1} \pi_{i+1} \ldots \pi_j \pi_i \pi_{j+1} \ldots \pi_n
\]
Some definitions for our heuristic:

**Definition 2:**

Given our set of permutations $A$, we say that a cyclic movement is a $k$-move if

$$d_{KT}(c[i, j](\pi), A) = d_{KT}(\pi, A) + k$$

**Definition 3:**

A good cyclic movement is a $k$-move, where $k < 0$
Example of k-move:

\[ \pi^1 = [1, 3, 8, 4, 6, 2, 7, 5] \]
\[ \pi^2 = [7, 2, 1, 3, 4, 5, 6, 8] \]
\[ \pi^3 = [5, 3, 1, 8, 7, 6, 4, 2] \]
Example of k-move:

\[ \pi^1 = [1, 3, 8, 4, 6, 2, 7, 5] \]
\[ \pi^2 = [7, 2, 1, 3, 4, 5, 6, 8] \]
\[ \pi^3 = [5, 3, 1, 8, 7, 6, 4, 2] \]
Example of k-move:

\[ \pi^1 = [1, 3, 8, 4, 6, 2, 7, 5] \]
\[ \pi^2 = [7, 2, 1, 3, 4, 5, 6, 8] \]
\[ \pi^3 = [5, 3, 1, 8, 7, 6, 4, 2] \]
Example of k-move:

\[ \pi^1 = [1, 3, 8, 4, 6, 2, 7, 5] \]
\[ \pi^2 = [7, 2, 1, 3, 4, 5, 6, 8] \]
\[ \pi^3 = [5, 3, 1, 8, 7, 6, 4, 2] \]
Example of k-move:

\[ \pi^1 = [1, 3, 8, 4, 6, 2, 7, 5] \]
\[ \pi^2 = [7, 2, 1, 3, 4, 5, 6, 8] \]
\[ \pi^3 = [5, 3, 1, 8, 7, 6, 4, 2] \]
Example of k-move:

\[ \pi^1 = [1, 3, 8, 4, 6, 2, 7, 5] \]
\[ \pi^2 = [7, 2, 1, 3, 4, 5, 6, 8] \]
\[ \pi^3 = [5, 3, 1, 8, 7, 6, 4, 2] \]
Example of $k$-move:

$$\pi^1 = [1, 3, 8, 4, 6, 2, 7, 5]$$

$$\pi^2 = [7, 2, 1, 3, 4, 5, 6, 8]$$

$$\pi^3 = [5, 3, 1, 8, 7, 6, 4, 2]$$

$k = -2$

Total = 7

Total = 5
Our to find k-moves:

\[ \pi_1 = [1, 3, 8, 4, 6, 2, 7, 5] \]

\[ \pi_2 = [7, 2, 1, 3, 4, 5, 6, 8] \]

\[ \pi_3 = [5, 3, 1, 8, 7, 6, 4, 2] \]

Theorem: We have that \( c_r[i, j](\pi) \) is a k-move iff

\[
k = 3(j - i) - 2 \sum_{t=i}^{j-1} w_g(\pi)(\pi_t, \pi_j)
\]
Our heuristic:

Given a set of permutations $A = \{\pi^1, \ldots, \pi^m\}$:

1) Take $\pi^1$ as a starting point for the heuristic

2) Find and execute a good move, if any

3) Repeat 2) till there is no more good moves and keep the result as a possible median for $A$

4) Repeat 1) with $\pi^i, 2 \leq i \leq m$

5) Take the best of the $m$ result as one median of $A$
Why do we execute our heuristic on each permutations of A?

\[ \pi^1 = [3, 6, 4, 2, 1, 7, 5] \]
\[ \pi^2 = [4, 6, 5, 1, 2, 7, 3] \rightarrow \pi^* = [6, 4, 1, 7, 5, 3, 2] \]
\[ \pi^3 = [1, 7, 5, 3, 6, 2, 4] \]

\[ \pi^1 \rightarrow [3, 6, 4, 1, 2, 7, 5] \]
\[ d_{KT} = 23 \]

\[ \pi^2 \rightarrow [6, 4, 1, 5, 2, 7, 3] \]
\[ d_{KT} = 23 \]

\[ \pi^3 \rightarrow \pi^* \]
\[ d_{KT} = 22 \]
Some results of our heuristic for set of permutations $A$ of cardinality 3:

<table>
<thead>
<tr>
<th>n</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>20</td>
<td>2024</td>
<td>280840</td>
<td>2000</td>
<td>2000</td>
<td>2000</td>
</tr>
<tr>
<td>% of errors</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0.05%</td>
<td>0.25%</td>
</tr>
<tr>
<td>distance difference</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Some results of our heuristic for set of permutations A of cardinality 3:

<table>
<thead>
<tr>
<th></th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>0%</td>
<td>0.05%</td>
<td>0.25%</td>
<td>0.35%</td>
<td>0.6%</td>
<td>1.1%</td>
<td>1.6%</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Adding 0-moves:

$\pi^1 = [3, 6, 4, 2, 1, 7, 5]$

$\pi^2 = [4, 6, 5, 1, 2, 7, 3] \quad \sim \quad \pi^* = [6, 4, 1, 7, 5, 3, 2]$

$\pi^3 = [1, 7, 5, 3, 6, 2, 4]$

$\pi^2 \sim [6, 4, 1, 5, 2, 7, 3]$
Good 0-moves:

A good 0-move is a 0-move that can be immediately follow by any good -k move

\[ \pi^1 = [3, 6, 4, 2, 1, 7, 5] \]
\[ \pi^2 = [4, 6, 5, 1, 2, 7, 3] \]
\[ \pi^3 = [1, 7, 5, 3, 6, 2, 4] \]

\[ \pi^* = [6, 4, 1, 7, 5, 3, 2] \]
Good 0-moves:

A good 0-move is a 0-move that can be immediately follow by any good -k move

\[ \pi^1 = [3, 6, 4, 2, 1, 7, 5] \]
\[ \pi^2 = [4, 6, 5, 1, 2, 7, 3] \]
\[ \pi^3 = [1, 7, 5, 3, 6, 2, 4] \]

\[ \pi^* = [6, 4, 1, 7, 5, 3, 2] \]
Good 0-moves:

\[ c_r[i, j](\pi) \text{ is good 0-move if it is a 0-move and } \]
\[ w_G(\pi)(\pi_{j-1}, \pi_{j+1}) = 2 \text{ or } w_G(\pi)(\pi_{i-1}, \pi_{j}) = 2 \]

\[ \pi^1 = [3, 6, 4, 2, 1, 7, 5] \]
\[ \pi^2 = [4, 6, 5, 1, 2, 7, 3] \]
\[ \pi^3 = [1, 7, 5, 3, 6, 2, 4] \]

\[ \pi^* = [6, 4, 1, 7, 5, 3, 2] \]
Results of our heuristic with 0-moves:

We applied our heuristic on 20,000 triplets of permutations of \{1, 2, ..., n\} for n between 6 and 10 and for each of these triplet, we computed the number of required 0-moves to get to the median.

The maximal number of 0-move permitted was 3.

<table>
<thead>
<tr>
<th>n</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>error %</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% of cases with 0 0-move</td>
<td>100</td>
<td>99.8</td>
<td>99.7</td>
<td>99.5</td>
<td>99.2</td>
</tr>
<tr>
<td>% of cases with 1 0-move</td>
<td>0</td>
<td>0.15</td>
<td>0.2</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>% of cases with 2 0-moves</td>
<td>0</td>
<td>0.05</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>% of cases with 3 0-moves</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
</tr>
</tbody>
</table>
What's left to do:

- Understand why for some given set of permutations there is only one median and for others more than one (we even found 33 medians for a triplet of permutations of \{1, 2, ..., 10\})

- Study the problem under important biological distances on permutations and signed permutations

- Is the median problem of a set of m permutations, m odd, under the Kendall-$\tau$ distance polynomial?

- If we know the medians of a set of permutations $A$ and the medians of set of permutations $B$, does it tell us something on the set of medians of $A \cup B, A \cap B, A \setminus B, \bar{A}$, ...