Pattern Discovery in Bioinformatics

Laxmi Parida

IBM T J Watson Research Center, Yorktown Heights & Courant Inst. of Mathematical Sciences, New York University

IMA, May 4, 2007
What?: Context
- Biology as an information science
- Discovery Process
  - String patterns
  - Topological patterns
- Permutation patterns

How?: Permutations → PQ trees
- Gene Proximity Analysis
- Statistics of permutations

More?: PQ Variations
- Phylogeny etc.
- Population genomics
Information in Biology

Organization

- Ecosystem
- Species
- Organism
- Physiology
- Metabolism
- Network
- Function
- Structure
- Sequence
- Complexity

Comparative Genomics
Pharmacogenomics
Physiome
Metabolomics
Proteomics
Functional Genomics
Structural Genomics
Genome

What?: Context
How?: Permutations → PQ trees
More?: PQ Variations

Biology as an information science
Discovery Process
Permutation patterns

Laxmi Parida
Pattern Discovery in Bioinformatics
Discovering Information

- Given a large dataset
  (an indiscriminate body of evidence)
- Can anything be inferred/discovered?
  (questions from the answers)
  - How?
    - Forbidden Phenomenon
    - Unique Phenomenon (pattern-free)
    - Repeating Phenomenon (patterns)
Pattern Discovery

- Define the phenomenon
  - non-unique (occurs at least $k > 1$ times)
  - define occurrence

- Discover the phenomenon
  - Let $D$ be the discovered set on input $s$, then
    - if $p \in D$ then, $p$ is a pattern
    - if $p$ is a pattern, then $p \in D$
What are the common patterns?

\[ s_1 = \ldots g_1 g_2 g_3 g_4 g_5 g_6 g_7 \ldots \]

\[ s_2 = \ldots g_8 g_2 g_3 g_4 g_5 g_9 g_0 \ldots \]
What are the common patterns?

\[ s_1 = \ldots g_1 \underline{g_2 g_3 g_4 g_5} g_6 g_7 \ldots \]

\[ s_2 = \ldots g_8 \underline{g_2 g_3 g_4 g_5} g_9 g_0 \ldots \]
Traditional (string) Pattern Discovery
How bad is the scenario?

String patterns: $O(n^2)$
Is there a combinatorial way of reducing the number?
Is there a combinatorial way of reducing the number?

Duality

Pattern description

Location/occurrence description
Is there a combinatorial way of reducing the number?

Let $P$ be the set of all patterns on a given input string $s$. ($p_1 \in P$) is non-maximal with respect to ($p_2 \in P$) if both of the following hold.

1. Each occurrence of $p_1$ on $s$ is covered by an occurrence of $p_2$ on $s$.
2. Each occurrence of $p_2$ on $s$ covers $l \geq 1$ occurrence(s) of $p_1$ on $s$.

A pattern ($p_2 \in P$) is maximal, if there exists no ($p_1 \in P$) such that $p_2$ is non-maximal w.r.t. $p_1$.
Is there a combinatorial way of reducing the number?

Let $P$ be the set of all patterns on a given input string $s$. $(p_1 \in P)$ is \textit{non-maximal} with respect to $(p_2 \in P)$ if both of the following hold.

(1) Each occurrence of $p_1$ on $s$ is covered by an occurrence of $p_2$ on $s$.

(2) Each occurrence of $p_2$ on $s$ covers $l \geq 1$, occurrence(s) of $p_1$ on $s$.

A pattern $(p_2 \in P)$ is \textbf{maximal}, if there exists no $(p_1 \in P)$ such that $p_2$ is non-maximal w.r.t. $p_1$. 

Laxmi Parida  Pattern Discovery in Bioinformatics
Is the definition “good”?

**Theorem**

Let $M$ be the set of all maximal patterns, i.e.,

$$M = \{ p \in P \mid \text{there is no } (p' \in P) \text{ maximal w.r.t } p \}$$

Then $M$ is unique.
Suffix tree = Trie of suffixes

$s = abcdabcdabcdabcab$

Maximal string patterns: $O(n)$
Again, what are the common patterns?
Topological Motifs (Network motifs)

- $|V_m| = 4$
- $|E_m| = 4$
- $N_{4,4} = 1$
- $|V_m| = 4$
- $|E_m| = 3$
- $N_{4,3} = 2$
- $|V_m| = 3$
- $|E_m| = 3$
- $N_{3,3} = 1$
- $|V_m| = 3$
- $|E_m| = 2$
- $N_{3,2} = 5$
- $|V_m| = 2$
- $|E_m| = 1$
- $N_{2,1} = 4$
- $|V_m| = 1$
- $|E_m| = 0$
Is there a combinatorial way of reducing the number?

Let $P$ be the set of all patterns on a given input data set. ($p_1 \in P$) is non-maximal with respect to ($p_2 \in P$) if both of the following hold.

1. Each occurrence of $p_1$ is covered by an occurrence of $p_2$.
2. Each occurrence of $p_2$ covers an occurrence of $p_1$.

A pattern ($p_2 \in P$) is maximal, if there exists no ($p_1 \in P$) such that $p_2$ is non-maximal w.r.t. $p_1$. 

Laxmi Parida

Pattern Discovery in Bioinformatics
Is there a combinatorial way of reducing the number?

Let $P$ be the set of all patterns on a given input data set. $(p_1 \in P)$ is non-maximal with respect to $(p_2 \in P)$ if both of the following hold.

1. Each occurrence of $p_1$ is covered by an occurrence of $p_2$.
2. Each occurrence of $p_2$ covers an occurrence of $p_1$.

A pattern $(p_2 \in P)$ is maximal, if there exists no $(p_1 \in P)$ such that $p_2$ is non-maximal w.r.t. $p_1$. 
But isomorphism is tricky..
There is a combinatorial way to reduce the output size...

1. Dual notation of motifs in terms of locations

2. **Compact** notation
   - compact vertices, edges and motifs
   - handles explosion due to isomorphisms

Again, what are the common patterns?

\[ s_1 = \ldots g_1 g_2 g_3 g_4 g_5 g_6 g_7 \ldots \]

\[ s_2 = \ldots g_8 g_5 g_2 g_4 g_3 g_9 g_0 \ldots \]
What are the common patterns?

\[ s_1 = \ldots g_1 \boxed{g_2 g_3 g_4 g_5} g_6 g_7 \ldots \]

\[ s_2 = \ldots g_8 \boxed{g_5 g_2 g_4 g_3} g_9 g_0 \ldots \]
Permutation patterns (\( \pi \) patterns)

\[
\begin{align*}
    s_1 &= \ldots g_1 \boxed{g_2 \ g_3 \ g_4 \ g_5} g_6 \ g_7 \ \ldots \\
    s_2 &= \ldots g_8 \boxed{g_5 \ g_2 \ g_4 \ g_3} g_9 \ g_0 \ \ldots
\end{align*}
\]

Genes \( g_i \) in \( s_1 \) and \( g_i \) in \( s_2 \) are orthologous.

Block of genes \( g_2, g_3, g_4, g_5 \)
appear together,
albeit in a different order.
Permutation patterns ($\pi$ patterns)

\begin{align*}
  s_1 &= \ldots g_1 \boxed{g_2 \ g_3 \ g_4 \ g_5} \ g_6 \ g_7 \ \ldots \\
  s_2 &= \ldots g_8 \boxed{g_5 \ g_2 \ g_4 \ g_3} \ g_9 \ g_0 \ \ldots \\
\end{align*}

Genes $g_i$ in $s_1$ and $g_i$ in $s_2$ are **orthologous**

Block of genes $g_2, g_3, g_4, g_5$

appear together,

albeit in a different order

This block is a permutation (pattern)

\[ \{g_2, g_3, g_4, g_5\} \]
Patterns Example
(Pursuit of the Preposterous)

\[ S = abcdefghijabdcefhgij \]

(size 20)

{a,b}, {a,b,c,d}, {a,b,c,d,e}, {a,b,c,d,e,f}, {a,b,c,d,e,f,g,h}
{a,b,c,d,e,f,g,h,i}, {a,b,c,d,e,f,g,h,i,j},
{b,c,d}, {b,c,d,e,f}, {b,c,d,e,f,g,h}, {b,c,d,e,f,g,h,i,j},
{c,d}, {c,d,e}, {c,d,e,f}, {c,d,e,f,g,h}, {c,d,e,f,g,h,i},
{c,d,e,f,g,h,i,j},
{e,f}, {e,f,g,h}, {e,f,g,h,i,j},
{f,g,h}, {f,g,h,i,j},
{g,h}, {g,h,i,j},
{i,j}

25 \pi \text{Patterns}
How bad is the scenario?

Permutation patterns: $O(n^2)$
Maximal \( \pi \) patterns

Let \( P \) be the set of all patterns on a given input string \( s \). \((p_1 \in P)\) is non-maximal with respect to \((p_2 \in P)\) if both of the following hold.

1. Each occurrence of \( p_1 \) on \( s \) is covered by an occurrence of \( p_2 \) on \( s \).
2. Each occurrence of \( p_2 \) on \( s \) covers \( l \geq 1 \), occurrence(s) of \( p_1 \) on \( s \).

A pattern \((p_2 \in P)\) is \textbf{maximal}, if there exists no \((p_1 \in P)\) such that \( p_2 \) is non-maximal w.r.t. \( p_1 \).
π patterns (nested & straddling)

\[
s_1 = \ldots g \begin{array}{|c|c|c|} \hline a & c & d \hline \end{array} b \begin{array}{|c|c|c|} \hline e & f & g \hline \end{array} e b \ldots
\]

\[
s_2 = \ldots b \begin{array}{|c|c|c|} \hline g & f & e \hline \end{array} d \begin{array}{|c|c|c|} \hline a & b & c \hline \end{array} f b \ldots
\]
\( \pi \) patterns (nested & straddling)

\[ s_1 = \ldots g \begin{array}{|c|c|} \hline a & c \\ b & e \\ d & f \\ g & e \\ \hline \end{array} \ e \begin{array}{|c|c|} \hline b & \ldots \\ \end{array} \]

\[ s_2 = \ldots b \begin{array}{|c|c|} \hline g & f \\ e & \hline \end{array} \ d \begin{array}{|c|c|} \hline a & b \\ c & \hline \end{array} \ f \begin{array}{|c|c|} \hline b & \ldots \\ \end{array} \]

\[ p = \{a, b, c, d, e, f, g\}, \]

\[ \text{nonMaximal}(p) = \{\{e, f\}, \{f, g\}, \{e, f, g\}, \{a, b, c, d\}\}. \]

Biology as an information science
Discovery Process
Permutation patterns

Laxmi Parida
Pattern Discovery in Bioinformatics
Theorems on $\pi$ Patterns

**Theorem:**
Let $R = \{Q' | Q' \text{ is non-maximal w.r.t } Q\}$. Then there exists a permutation $Q''$ of the elements of $Q$, such that for each $Q'$, a permutation of the elements of $Q'$ is a substring of $Q''$

**Corollary 1:**
The ordering is not necessarily complete

**Corollary 2:**
A representation that captures the order of elements of $Q$ along with intervals that captures each of $Q'$ encodes $Q$
**What is a PQ Tree?**

Is there a sequence where the sets are consecutive?

A collection of sets

{5, 6}
{1, 2, 3, 4}
{1, 2, 3, 4, 5}
{1, 2, 3, 4, 5, 6}

The answer is YES for this set. All such sequences captured by the PQ tree.
PQ Trees Revisited

\{a, b, c, d\}, \{b, c\}, \{g, h\}, \{h, i\}

cbadghi
dacbihg
abcdihg
ghidacb
ihgdabc
  
  a
  |
  
  d g h i
  |
  |
c b
  :
Maximal $\pi$Patterns
(Notation, PQ Tree notation)

- immediate neighbors "-" (Q)
- otherwise "," (P)
- groups "( )" levels in the PQ tree
Linear Notation

\[ p = \{a, b, c, d, e, f, g\}, \]

\[ \text{nonMaximal}(p) = \}\{\{e, f\}, \{f, g\}, \{e, f, g\}, \{a, b, c, d\}\}\].

Linear notation: \(( (a, b, c, d), (e-f-g) )\)
Pattern (maximal)

\[ S = \text{abcdefghijabdcefhgij} \]  (input size 20, no of patt 25)

\{a,b\}, \{a,b,c,d\}, \{a,b,c,d,e\}, \{a,b,c,d,e,f\}, \{a,b,c,d,e,f,g,h\}
\{a,b,c,d,e,f,g,h,i\}, \{a,b,c,d,e,f,g,h,i,j\},
\{b,c,d\}, \{b,c,d,e,f\}, \{b,c,d,e,f,g,h\}, \{b,c,d,e,f,g,h,i\},
\{c,d\}, \{c,d,e\}, \{c,d,e,f\}, \{c,d,e,f,g,h\}, \{c,d,e,f,g,h,i\},
\{c,d,e,f,g,h,i,j\},
\{e,f\}, \{e,f,g,h\}, \{e,f,g,h,i\},
\{f,g,h\}, \{f,g,h,i\},
\{g,h\}, \{g,h,i\},
\{i,j\}

**Distribution:**

\[ \text{a-b-(c-d)-e-f-(g-h)-i-j} \]
Is the definition any good?

**Theorem**

Let $M$ be the set of all maximal patterns, i.e.,

$$M = \{ p \in P \mid \text{there is no } (p' \in P) \text{ maximal w.r.t } p \}$$

Then $M$ is unique.
Gene Proximity Analysis on Whole Genomes
(CPM 05, JCB 06)

- Human, rat genomes (http://bio.math.berkeley.edu/slam)
- 25,422 putative orthologous genes
- 23 human, 21 rat chromosomes
Algorithms

- Find the patterns \((WABI 03, JCB 04)\)
  For a fixed pattern size, the time taken is

  \[
  \mathcal{O}(|\Sigma| + n(\log t)^2 \log |\Sigma|),
  \]

  where

  \[
  t = \mathcal{O}(|\Sigma| + n \log |\Sigma|).
  \]

- Extract maximal form \((CPM 05, JCB 06)\)
  *The Minimal Consensus PQ Tree Algorithm* (linear time)
Human & Rat

\( \pi \text{Pattern (166)} \)

Human chromosome 1
Rat chromosome 13

ABCDEFGHIJ
JIHGDBFECA
Human & Rat

**Pattern (303)**

- What?: Context
- How?: Permutations $\rightarrow$ PQ trees
- More?: PQ Variations

Gene Proximity Analysis
Statistics of permutations

Human chromosome 11
Rat chromosome 1

A B C
D1 ... D11

I1 ... I17

E F G
H1 ... H12

Laxmi Parida
Pattern Discovery in Bioinformatics
**Pattern (250)**

- A
- B
- Human chromosomes 10,11
- Rat chromosome 1
- C1 \ldots C6
- D1 \ldots D3
- E1 \ldots E3
- F1 \ldots F2

**Gene Proximity Analysis**

**Statistics of permutations**

**Human & Rat**

**Pattern Discovery in Bioinformatics**
### Proximity- Summary

<table>
<thead>
<tr>
<th></th>
<th>Number of all patterns</th>
<th>Number of maximal patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>E Coli K-12 &amp; B Subtilis</td>
<td>15,000</td>
<td>450</td>
</tr>
<tr>
<td>human &amp; rat</td>
<td>1,574,312</td>
<td>504</td>
</tr>
</tbody>
</table>

(Joint work with Revital Eres, Oren Weimann, Gadi Landau)
Two related but distinct questions:

1. Given \( n \) random permutations of \( k \) genes, what is the probability that \( K \) of these \( n \) contain the cluster \( q \)?

2. Given that a permutation pattern \( q \) occurs \( K \) times in the input, what is the probability of its maximal form given as a PQ tree \( T \)?

Statistical Significance of Large Gene Clusters,
Laxmi Parida, under submission, 2007.
Generalized Question 1:

Let the input be generated by a stationary, \textit{iid} source which emits $x_i$ with probability $p_{x_i}$.

$$q = \{x_1(i_1), x_2(i_2), \ldots, x_l(i_l)\}$$
Generalized Question 1:

Let the input be generated by a stationary, iid source which emits $x_i$ with probability $p_{x_i}$.

$$q = \{x_1(i_1), x_2(i_2), \ldots, x_l(i_l)\}$$

$$\mathbb{P}_q = \left( \frac{(i_1 + i_2 + \ldots + i_l)!}{i_1! i_2! \ldots i_l!} \right) (p_{x_1})^{i_1} (p_{x_2})^{i_2} \ldots (p_{x_l})^{i_l}$$
How to answer Question 2?

- Count the size of the frontier set of $T$
- What is the bottleneck?
  - How many possibilities does a P node introduce?
How to answer Question 2?

- Count the size of the frontier set of $T$
- What is the bottleneck?
  How many possibilities does a P node introduce?
### P-arrangement

<table>
<thead>
<tr>
<th>Interval</th>
<th>P(q1[k1..k2])</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>[k1..k2]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-trivial:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[3..4]</td>
<td>5 2 4 3 1</td>
<td>{3, 4}</td>
</tr>
<tr>
<td>[2..4]</td>
<td>5 2 4 3 1</td>
<td>{2, 3, 4}</td>
</tr>
<tr>
<td>[1..4]</td>
<td>5 2 4 3 1</td>
<td>{2, 3, 4, 5}</td>
</tr>
</tbody>
</table>

trivial:

| [1..5]       | 5 2 4 3 1     | {1, 2, 3, 4, 5} | 5 |
An arrangement of size $k$ is a $P$-arrangement if it has no non-trivial intervals.

Examples:

- 2 4 1 3
- 2 4 1 5 3
- 2 6 4 1 5 3
The central question:

What is the number of $P$-arrangements of size $k$?
The central question:

What is the number of $P$-arrangements of size $k$?

**Theorem**

Let $q$ be a $P$-arrangement of size $k + 1$. Let $q'$ be obtained by replacing an extreme element (either $k + 1$ or $1$) from its position $j$ in $q$, with the empty symbol. Then $q'$ is a $P$-arrangement or every interval $[i_1 \ldots i_2]$ in $q'$ is such that $i_1 < j < i_2$. 
Nested Arrangements

\[ q = 9 \ 1 \ 5 \ 2 \ \phi \ 3 \ 6 \ 4 \ 7 \ 10 \ 8 \]
Nested Arrangements

\[ q = 9 \ 1 \ 5 \ 2 \ \phi \ 3 \ 6 \ 4 \ 7 \ 10 \ 8 \]

\[
\begin{align*}
9 & 1 & 5 & 2 & \phi & 3 & 6 & 4 & 7 & 10 & 8 \\
9 & 1 & 5 & 2 & \phi & 3 & 6 & 4 & 7 & 10 & 8 & 2 \\
9 & 1 & 5 & 2 & \phi & 3 & 6 & 4 & 7 & 10 & 8 & 5 \\
9 & 1 & 5 & 2 & \phi & 3 & 6 & 4 & 7 & 10 & 8 & 6(2) \\
9 & 1 & 5 & 2 & \phi & 3 & 6 & 4 & 7 & 10 & 8 & 10
\end{align*}
\]
Nested Arrangements

\[ q = 9 \ 1 \ 5 \ 2 \ \phi \ 3 \ 6 \ 4 \ 7 \ 10 \ 8 \]

\[
\begin{align*}
9 \ 1 \ 5 & \ 2 \ \phi \ 3 \ 6 \ 4 \ 7 \ 10 \ 8 & 2 \\
9 \ 1 & \ 5 \ 2 \ \phi \ 3 \ 6 \ 4 \ 7 \ 10 \ 8 & 5 \\
9 & \ 1 \ 5 \ 2 \ \phi \ 3 \ 6 \ 4 \ 7 \ 10 \ 8 & 6(2) \\
9 \ 1 \ 5 \ 2 \ \phi \ 3 \ 6 \ 4 \ 7 \ 10 \ 8 & 10
\end{align*}
\]

Signature: \( \text{sig}(q) = 2(1) < 5(1) < 6(2) < 10(1) \)
Signature Lemma

Let \( q \) be a nested arrangement of size \( k \) with symbol \( \phi \) in position \( j \). Let the size of the interval be \( i_r = i_{r2} - i_{r1} + 1 \).

1. *(straddling intervals)* If two such intervals, where one is not nested in the other, are of size \( i \) and \( i' \), then \( i = i' \) and they must overlap in \( i - 1 \) positions.
Signature Lemma

Let \( q \) be a nested arrangement of size \( k \) with symbol \( \phi \) in position \( j \). Let the size of the interval be \( i_r = i_{r2} - i_{r1} + 1 \).

1. (straddling intervals) If two such intervals, where one is not nested in the other, are of size \( i \) and \( i' \), then \( i = i' \) and they must overlap in \( i - 1 \) positions.

2. (uniqueness and form)

\[
\text{sig}(q) = i_1(k_{i_1}) < i_2(k_{i_2}) < \ldots < i_r(k_{i_r}) < \ldots < i_K(k_{i_K}),
\]

is unique with \( k_{i_1} = 1, k_{i_K} = 1, i_K = k \), and each \( k_{i_r}, 1 \leq r < K \), is either 1 or 2.
Formula for number of $P$-arrangements

\[
\begin{align*}
Pa(2) &= 2, \\
Pa(3) &= 0, \\
Pa(4) &= 2, \\
Pa(k) &= Nst'(k - 1), \quad \text{for } k > 4.
\end{align*}
\]

Polynomial time dynamic programming solution.
Number of nested arrangements with viable positions to get $k + 1$-sized $P$-arrangements:

$$Nst'(k) = S(k, 2) - S_{cnt}(k, 2) + \sum_{l=4}^{k} (l - 1)S(k, l) - 2S_{cnt}(k, l).$$
Number of nested arrangements with smallest \( l \) and largest \( u \) interval sizes:

\[
S(u, l) = 4S(u-1, l) + 2S(u-2, l) + \sum_{y=3}^{u-l} \Delta_{u-y} Pa(\Delta_{u-y}) S(u - y, l)
\]

Number of nested arrangements with the extreme element in the smallest interval:

\[
S_{cnt}(u, l) = 2S_{cnt}(u-1, l) + \sum_{y=3}^{u-l} Pa(\Delta_{u-y}) S_{cnt}(u - y, l)
\]
Chloroplast gene order in *Campanulaceae*

Data: 105 genes & 13 extant species
(PRIB 06: Joint work with Enam Karim, Arun Lakhotia)
What more?

- Gapped $\pi$ patterns (WABI 06, JCB 06)
- Bounded $\pi$ patterns (work in progress)
- Can we enrich the PQ structure?
Recall: PQ Tree (proximity information)

\[ s_1 = \ldots g \begin{array}{|c|c|c|} \hline a & c & d \hline b & e & f & g \hline \end{array} e b \ldots \]

\[ s_2 = \ldots b \begin{array}{|c|c|c|} \hline g & f \hline e & d & a & b & c \hline f & b \ldots \end{array} \]
Fortifying the PQ?

\[ s_1 = \ldots g a c d b e f g e b \ldots \]

\[ s_2 = \ldots b g f e d a b c f b \ldots \]
Food for thought...(fortified PQ)

- How do a set of transcription factors hone in on a target set of genes?
- use mini-motifs in binding site layouts (say in yeast species *S. cerevisiae, S. paradoxus, S. mikatae*, and *S. bayanus*)
What else is PQ structure good for?
WHAT?: addresses a fundamental curiosity: how did humans populate the earth as we see today? (genetic migratory history)  
WHO?: National Geographic & IBM on a 5-year study  
HOW?: collect data - DNA sample from people all over the world (associate geography)
WHAT? addresses a fundamental curiosity: how did humans populate the earth as we see today? (genetic migratory history)
WHAT? addresses a fundamental curiosity: how did humans populate the earth as we see today? (genetic migratory history)

WHO? National Geographic & IBM on a 5-year study
WHAT? addresses a fundamental curiosity: how did humans populate the earth as we see today? (genetic migratory history)

WHO? National Geographic & IBM on a 5-year study

HOW? collect data-
DNA sample from people all over the world indigenous people (associate geography)
The Genographic Project
Y-SNP Phylogeny

90 (50 - 130) KYA, Hammer and Zegura
59 (40 - 140) KYA, Thomson et al.

69 (56 - 81) KYA, Hammer and Zegura
40 (35 - 89) KYA, Thomson et al.
Y-Haplogroup Distribution

What?: Context
How?: Permutations → PQ trees
More?: PQ Variations

Phylogeny etc.
Population genomics

Pattern Discovery in Bioinformatics
What?: Context
How?: Permutations → PQ trees
More?: PQ Variations
Phylogeny etc.
Population genomics

mtDNA TREE
What?: Context
How?: Permutations → PQ trees
More?: PQ Variations

(Possible) Migration History

Human mtDNA Migrations
Copyright 2002 © Mitomap.org

Laxmi Parida
Pattern Discovery in Bioinformatics
Some Unanswered Questions...
Block of Short Tandem Repeat Polymorphisms (STRP) inverted in a significant fraction of human population:

STRP’s on the short arm of chromosome 8
Block of Short Tandem Repeat Polymorphisms (STRP) inverted in a significant fraction of human population:

STRP’s on the long arm of chromosome 17
The Phylogeny Reconstruction Problem

\[
\Pi \\
A=0234516789 \\
B=0154326987 \\
C=0678321549 \\
D=3210678549 \\
E=0123678945 \\
F=0187632549 \\
G=0123459876 \\
H=0154326789
\]
Assumptions

1. unsigned permutations
2. inversion and translocation
   - $D(\pi_1, \pi_2)$ denote the smallest number of operations, inversion and/or transposition, that takes $\pi_1$ to $\pi_2$. 
Terminology

**Definition**

$F(T)$: The *frontier* of a tree $T$, $F(T)$, is the permutation obtained by reading the labels of the leaves from left to right.

**Definition**

$T \equiv T'$: $T$ is *equivalent* to $T'$, if one can be obtained from the other by applying a sequence of the following transformation rules:

1. Arbitrarily permute the children of a $P$-node, and
2. Reverse the children of a $Q$-node.

**Definition**

$\mathcal{F}(T)$: $\mathcal{F}(T) = \{ F(T') | T' \equiv T \}$.
Example

\[ X = \{ \text{a, b, c, d, e} \} \]

\[ T \]

\[ \text{FRONTIER}(T) = \text{abcde} \]

\[ T' \]

\[ \text{FRONTIER}(T') = \text{edabc} \]

\[ T' \equiv T. \]

\[ \mathcal{F}(T) = \mathcal{F}(T') \]

\[ = \{ \text{abcde, abced, cbade, cbaed, deabc, decba, edabc, edcba} \}. \]
A Tree $T$ such that $\Pi \subseteq F(T(\Pi))$

(\Pi is a set of permutations)

$\pi_1 = 0123456789$
$\pi_2 = 0145238967$
$\pi_3 = 6723894510$
$\pi_4 = 4598326701$
$\pi_5 = 8923674501$
$\pi_6 = 8932674501$

$T(\Pi)$ is the minimal consensus PQ tree of $\Pi$. 

$Laxmi Parida
Pattern Discovery in Bioinformatics$
Oriented PQ Trees (oPQ Trees)

1. Minimal consensus of two permutations.
2. All Q-nodes are oriented.
3. The children of P-nodes are labeled with integers.
oPQ Tree Example

\[ \vec{T} = \vec{T}_{\pi_1}(\pi_2) \]

The oPQ tree of \( \pi_1 \) and \( \pi_2 \). Note that \( \text{Size}(\vec{T}_1) = 7 \).
Algorithmic Implication-1

Theorem

Given permutations $\pi_1, \pi_2$ of length $n$ each, $\vec{T}_{\pi_1}(\pi_2)$

1. is unique,

2. $\text{Size}(\vec{T}_{\pi_1}(\pi_2)) = \mathcal{O}(n)$, and,

3. can be constructed in $\mathcal{O}(n)$ time.
Algorithmic Implication-2

**Theorem**

*Given a fixed constant* $c$ *and* $\pi$ *of size* $n$ *and*, let $S$ *a set of non-equivalent oPQ trees* $S$ *be defined as follows:*

$$S = \left\{ \vec{T}_{\pi}(\pi') \mid D(\pi, \pi') = c \text{ and } \pi' \text{ is a permutation of size } n \right\}$$

*Then*

1. $|S| = \mathcal{O}(1)$, and
2. for each $\vec{T} \in S$, $\text{Size}(\vec{T}) = \mathcal{O}(1)$. 

Laxmi Parida

Pattern Discovery in Bioinformatics
An oPQ tree where:

1. Each internal node has at most one leaf node.

2. Each leaf node is labeled with a signed number \( i \) where it represents the \( i \)th block in \( \pi_1 \).

\( \text{(plus sign - identical in } \pi_1 \text{ and } \pi_2, \) 

\( \text{minus sign - block flipped in } \pi_2. \) \)
Designing the oPQ Templates (Nested block inversions)
Two Inversions (Disjoint blocks)

What?: Context

How?: Permutations → PQ trees

More?: PQ Variations

Phylogeny etc.

Population genomics

Two Inversions (Disjoint blocks)

+1 +2 +3 +4 +5

+1 +3 +4 +5

D1

+1 +2 +3 -4 +5

D2

D1

D2

Laxmi Parida

Pattern Discovery in Bioinformatics
Two Inversions (Straddling blocks)

What?: Context
How?: Permutations → PQ trees
More?: PQ Variations
Phylogeny etc.
Population genomics

Laxmi Parida
Pattern Discovery in Bioinformatics
The Permutation Tree Construction Problem

Problem
Given $\Pi$, the PTC problem is to construct the permutation tree $T(V, E)$ of minimum length, where

$$\text{Len}(T) = \sum_{(v_1, v_2) \in E} D(\pi(v_1), \pi(v_2)).$$

Theorem
Given a random collection $\Pi$, the expected number of permutation trees (with small edge length $c$) on $\Pi$ is $o(1)$. 
Experiments

1. sequence of length 200 each
2. transposition or inversion (size 5 to 10 bases)
3. 16 leafnodes and 4-7 internal nodes makeup Π
Results on simulations
(RECOMBCG 06, JCB 06)

(a) inserting and/or deleting random positions,
(b) inserting random segment of size from 2 to 7,
(c) deleting random segments, and,
(d) a combination of (b) and (c).

Number of correct trees out of the 50 experiments:

<table>
<thead>
<tr>
<th></th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bifurcating</td>
<td>39</td>
<td>35</td>
<td>31</td>
<td>28</td>
</tr>
<tr>
<td>Multifurcating</td>
<td>36</td>
<td>28</td>
<td>30</td>
<td>26</td>
</tr>
</tbody>
</table>
Pattern Discovery in Bioinformatics
Theory & Algorithms
Laxmi Parida