Combinatorics and Statistics of Gene Clusters

Laxmi Parida

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

November 26-30, 2007

An everyday fundamental question...

What is the (statistical) significance associated with a *common* gene cluster q?

A Naive (straightforward) Approach

- Assumption: Let the input be generated by a stationary, *iid* source which emits x_i with probability p_{x_i}.
- $$q = \{x_1, x_2, \ldots, x_l\}$$

$$\mathbb{P}_q = (p_{x_1})(p_{x_2})\ldots(p_{x_l})$$

A more complex model (naive approach)

- Assumption: Let the input be generated by a stationary, *iid* source which emits x_i with probability p_{x_i}.
- $$q = \{x_1(i_1), x_2(i_2), \ldots, x_l(i_l)\}$$

$$\mathbb{P}_q = \left(\frac{(i_1 + i_2 + \ldots + i_l)!}{i_1!\, i_2!\, \ldots\, i_l!} \right) (p_{x_1})^{i_1} (p_{x_2})^{i_2} \ldots (p_{x_l})^{i_l}$$

(multinomial coefficients)

Some troubling issues....

- What is p_{x_i}?
  - A gene x_i is not commonly known to occur too many times
- If $i \neq j$, how do p_{x_i} and p_{x_j} compare?
- $\vdots$

Some troubling issues....

- What is p_{x_i}?
  - A gene x_i is not commonly known to occur too many times
- If $i \neq j$, how do p_{x_i} and p_{x_j} compare?
- ⋮

Can we do without guessing p_{x_i}?

The proposed model (structured clusters)

- Recall: Genes occur on chromosomes (linearly arranged)
  - Or, on a network (*common connected components*)

The proposed model (structured clusters)

- Recall: Genes occur on chromosomes (linearly arranged)
  - Or, on a network (*common connected components*)
- Let the collection of all subclusters within the common cluster q that occurs in the species being compared be $\mathcal{S}$.

The proposed model (structured clusters)

- Recall: Genes occur on chromosomes (linearly arranged)
  - Or, on a network (*common connected components*)
- Let the collection of all subclusters within the common cluster q that occurs in the species being compared be $\mathcal{S}$.

What is the probability of occurrence of subclusters $\mathcal{S}$, given the occurrence of cluster q?

The proposed model (structured clusters)

- Individual p_{x_i} is irrelevant
- The occurrence of each x_i is equally likely (or not)
- Well-posed question

The proposed model (structured clusters)

- Individual p_{x_i} is irrelevant
- The occurrence of each x_i is equally likely (or not)
- Well-posed question, but a combinatorics nightmare!

The proposed model (structured clusters)

- Individual p_{x_i} is irrelevant
- The occurrence of each x_i is equally likely (or not)
- Well-posed question, but a combinatorics nightmare!

The 64K $ question: Is this solvable?

Roadmap

Context

- Permutation patterns

Permutations → PQ trees

- Gene Proximity Analysis
- Statistics of permutations

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 (πpatterns)

$$s_1 = \ldots g_1 \boxed{g_2\ g_3\ g_4\ g_5}\ g_6\ g_7\ \ldots$$
$$s_2 = \ldots g_8 \boxed{g_5\ g_2\ g_4\ g_3}\ g_9\ g_0\ \ldots$$

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)

$$\{g_2, g_3, g_4, g_5\}$$

πPatterns 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 πPatterns

How bad is the scenario?

Permutation patterns: $\mathcal{O}(n^2)$

Maximal πpatterns

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.

Maximal πpatterns

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.

nonmaximal π*patterns* $\Leftrightarrow$ *subclusters*

πpatterns (nested & straddling)

$$s_1 = \ldots g \boxed{\boxed{a\,c\,d\,b}\ \boxed{e\,f\,g}}\ e\,b\ldots$$

$$s_2 = \ldots b \boxed{\boxed{g\,f\,e}\ \boxed{d\,a\,b\,c}}\ f\,b\ldots$$

$$p = \{a, b, c, d, e, f, g\}, \quad (1)$$

$$nonMaximal(p) = \{\{e, f\}, \{f, g\}, \{e, f, g\}, \quad (2)$$

$$\{a, b, c, d, \}\}. \quad (3)$$

Theorems on πPatterns

Theorem:

Let R = {Q'| Q' 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?

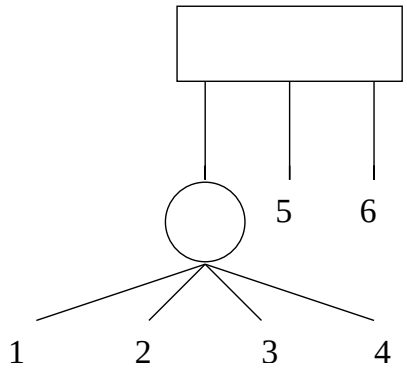
A collection of sets

$\{5, 6\}$

$\{1, 2, 3, 4\}$

$\{1, 2, 3, 4, 5\}$

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

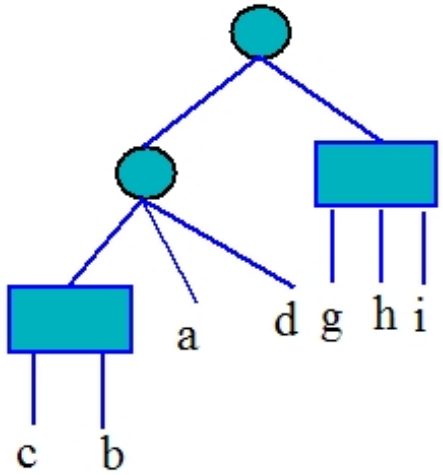


The answer is YES for this set.
All such sequences captured by the PQ tree.

PQ Trees Revisited.....

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

cbadghi
dacbihg
abcdihg
ghidacb
ihgdabc
:
:
:



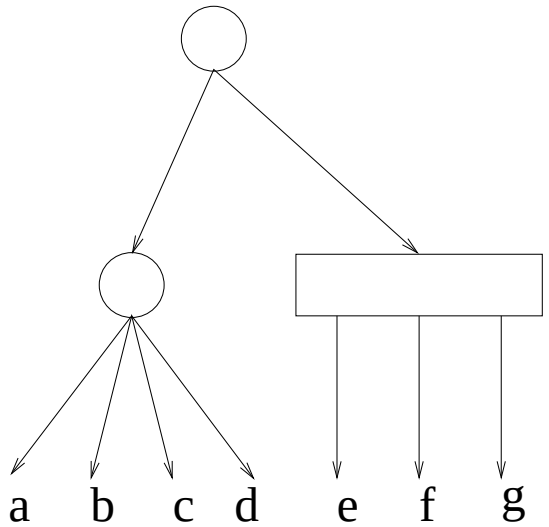
Maximal πPatterns

(Notation, PQ Tree notation)

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

Linear Notation

$$\begin{aligned} p &= \{a, b, c, d, e, f, g\}, \\ nonMaximal(p) &= \{\{e, f\}, \{f, g\}, \{e, f, g\}, \\ &\quad \{a, b, c, d, \}\}. \end{aligned}$$



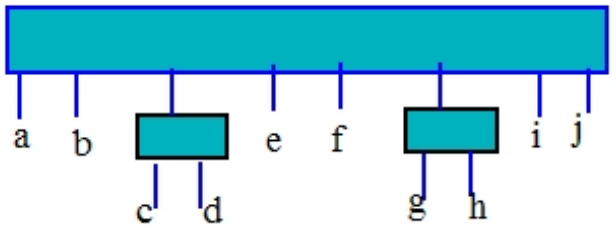
$$p = ((a, b, c, d)\text{-}(e\text{-}f\text{-}g)).$$

πPattern (maximal)

S = **abcdefghijabdcefhgij** (input size 20, no of patts 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,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}

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.,

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

Then M is unique.

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)

(software available: `http://www.mit.edu/oweimann/BIO/`)

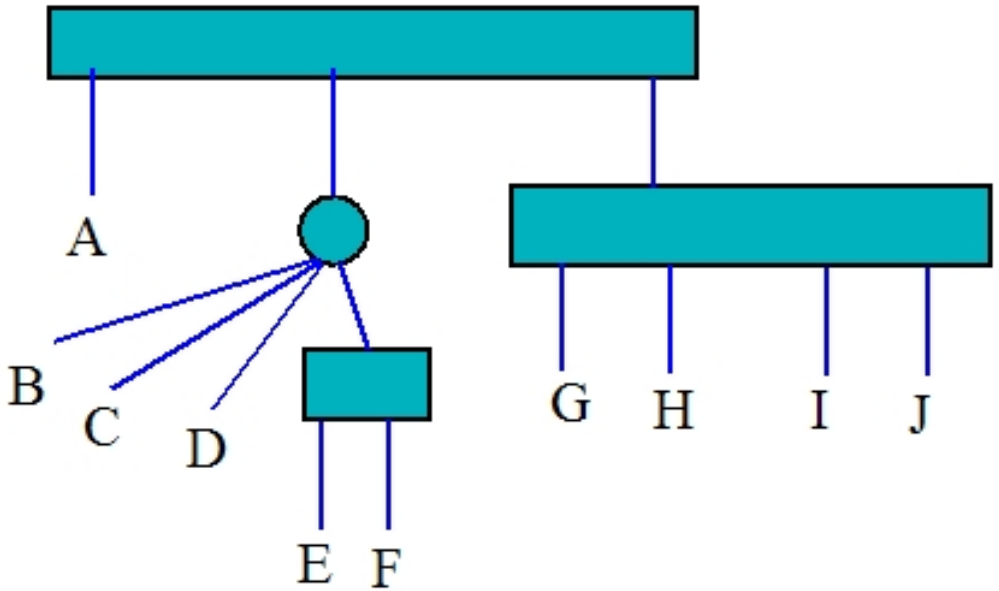
Gene Proximity Analysis on Whole Genomes

(CPM 05, JCB 06)

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

Human & Rat

πPattern (166)



Human chromosome 1

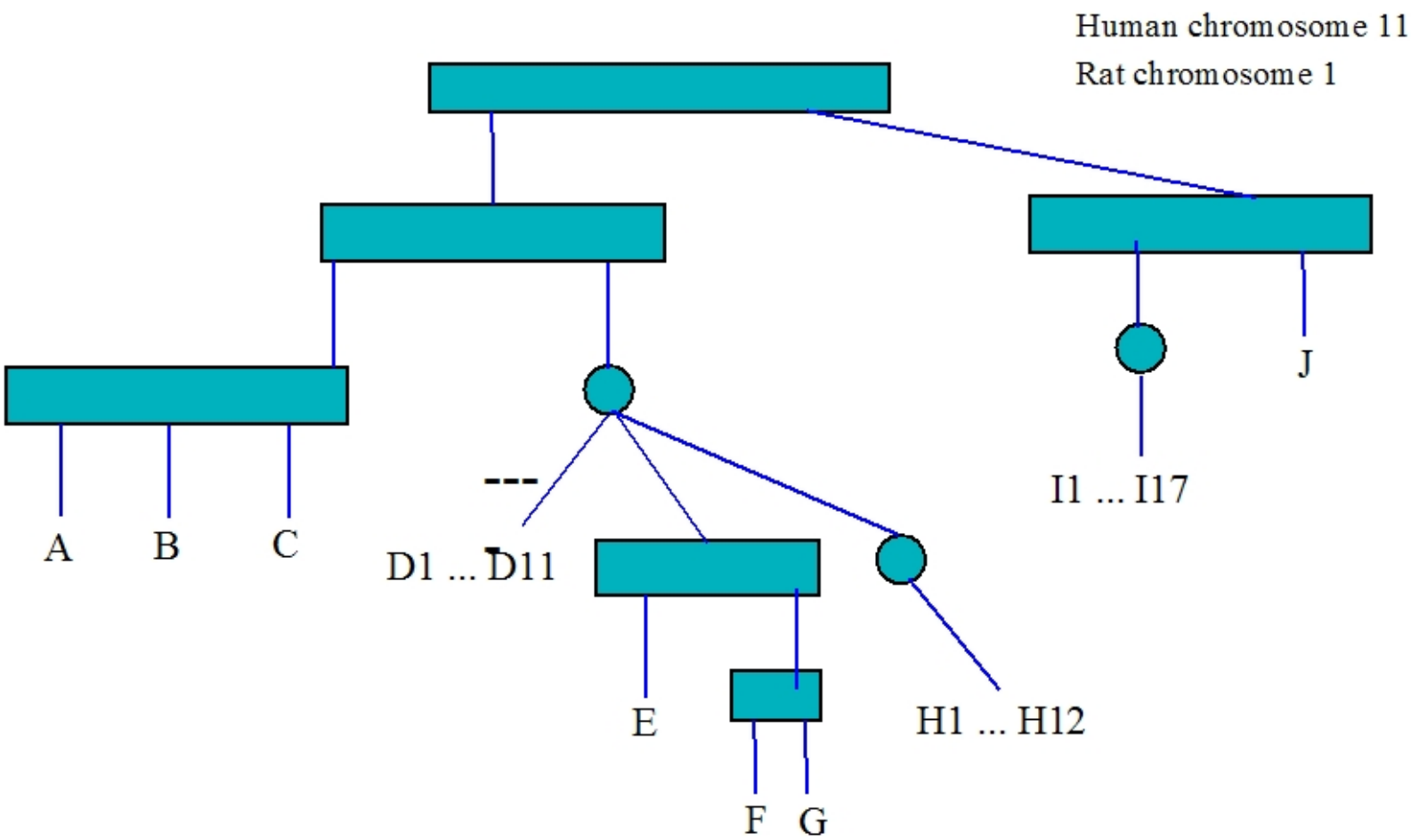
Rat chromosome 13

ABCDEFGHIJ

JIHGDBFECA

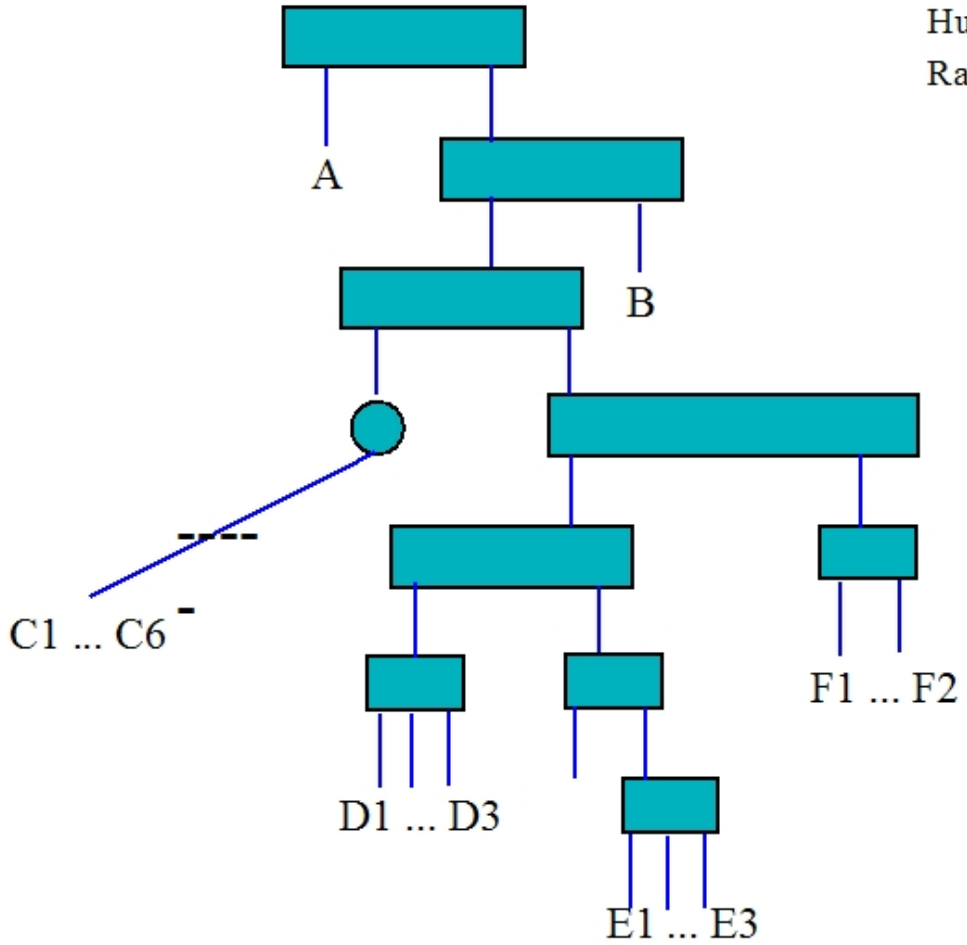
Human & Rat

πPattern (303)



Human & Rat

πPattern (250)



Human chromosomes 10,11
Rat chromosome 1

Proximity- Summary

	Number of all patterns	Number of maximal patterns
E Coli K-12 & B Subtilis	15,000	450
human & rat	1,574,312	504

(Joint work with Revital Eres, Oren Weimann, Gadi Landau)

Statistical Significance

Two related but distinct questions:

1. *Naive model: Given n random permutations of k genes, what is the probability that K of these n contain the cluster q ?*

2. *Structured cluster model: 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 ?*

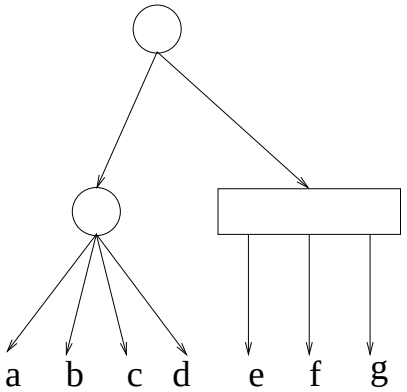
Statistical Significance of Large Gene Clusters, Laxmi Parida, Journal of Computational Biology, 14(9), pp 1145–1159, 2007.

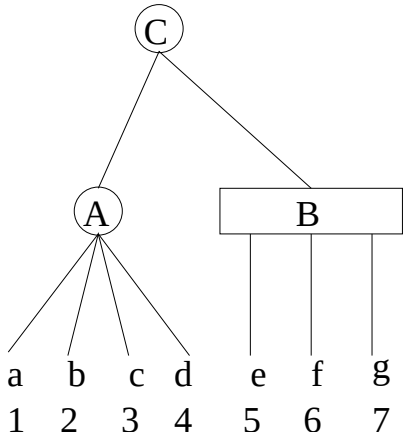
How to answer Question 2 (for $K = 2$)?

- Recall: PQ T in relation with multiple occurrences
  - a single occurrence has no PQ T
  - *Consensus PQ* of $o_1, o_2 \in Fr(T)$ is not necessarily T
- Use one occurrence as reference WLOG

How to answer Question 2 (for $K = 2$)?

- Recall: PQ T in relation with multiple occurrences
  - a single occurrence has no PQ T
  - *Consensus PQ* of $o_1, o_2 \in Fr(T)$ is not necessarily T
- Use one occurrence as reference WLOG





(1) PQ tree T. (2) Ref PQ Tree T' on 7 consecutive integers.

Working with T' (on consecutive integers)

- Count the size of the *frontier* set of T'

  Interpretation: Each $o \in Fr(T')$ is such that the consensus PQ tree of o and the identity permutation is T'

- What is the bottleneck?

  How many possibilities does a P node (with k children) introduce?
  Each possibility cannot have any internal structure

Working with T' (on consecutive integers)

- Count the size of the *frontier* set of T'

  Interpretation: Each $o \in Fr(T')$ is such that the consensus PQ tree of o and the identity permutation is T'

- What is the bottleneck?

  How many possibilities does a P node (with k children) introduce?
  Each possibility cannot have any internal structure
  We call this P-Arrangement of size k

What is an interval?

interval $[k_1..k_2]$		$\Pi(q_1[k_1..k_2])$	size
non-trivial:			
$[3..4]$	$5\,2\,\boxed{4\,3}\,1$	$\{3,4\}$	2
$[2..4]$	$5\,\boxed{2\,4\,3}\,1$	$\{2,3,4\}$	3
$[1..4]$	$\boxed{5\,2\,4\,3}\,1$	$\{2,3,4,5\}$	4
trivial:			
$[1..5]$	$\boxed{5\,2\,4\,3\,1}$	$\{1,2,3,4,5\}$	5

P-arrangement

An arrangement of size k is a *P-arrangement* if it has no non-trivial intervals.

Examples: $\boxed{2\ 4\ 1\ 3}$

$\boxed{2\ 4\ 1\ 5\ 3}$

$\boxed{2\ 6\ 4\ 1\ 5\ 3}$

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 (Par07)

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

1. *q' is a P-arrangement, or,*
2. *every interval $[i_1 \dots i_2]$ in q' is such that $i_1 < j < i_2$.*

Theorem illustration (nested arrangements)

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

$9\ 1\ 5\ \boxed{2\ \phi\ 3}\ 6\ 4\ 7\ 10\ 8$ — 2

$9\ 1\ \boxed{5\ 2\ \phi\ 3\ 6\ 4}\ 7\ 10\ 8$ — 5

$9\ \boxed{1\ 5\ 2\ \phi\ 3\ 6\ 4}\ 7\ 10\ 8$ — 6(2)

$9\ 1\ \boxed{5\ 2\ \phi\ 3\ 6\ 4\ 7}\ 10\ 8$ — 6(2)

$9\ \boxed{1\ 5\ 2\ \phi\ 3\ 6\ 4\ 7}\ 10\ 8$ — 7

$\boxed{9\ 1\ 5\ 2\ \phi\ 3\ 6\ 4\ 7\ 10\ 8}$ — 10

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

Signature Lemma

Let q be an arrangement of size k with symbol ϕ in position j with all the intervals $[i_{r1} \ldots i_{r2}]$ satisfying

$$i_{r1} < j < i_{r2},$$

for all $1 \leq r \leq K$. 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)**

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

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

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

Polynomial time dynamic programming solution.

$Nst'(k)$

Number of nested arrangements with viable positions to get $k+1$-sized P-arrangements:

$$\begin{aligned} Nst'(k) &= S(k,2) - S_{cnt}(k,2) \\ &+ \sum_{l=4}^{k} (l-1)S(k,l) - 2S_{cnt}(k,l). \end{aligned}$$

$S(l,u)$, $S_{cnt}(l,u)$

Number of nested arrangements with smallest l and largest u interval sizes:

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

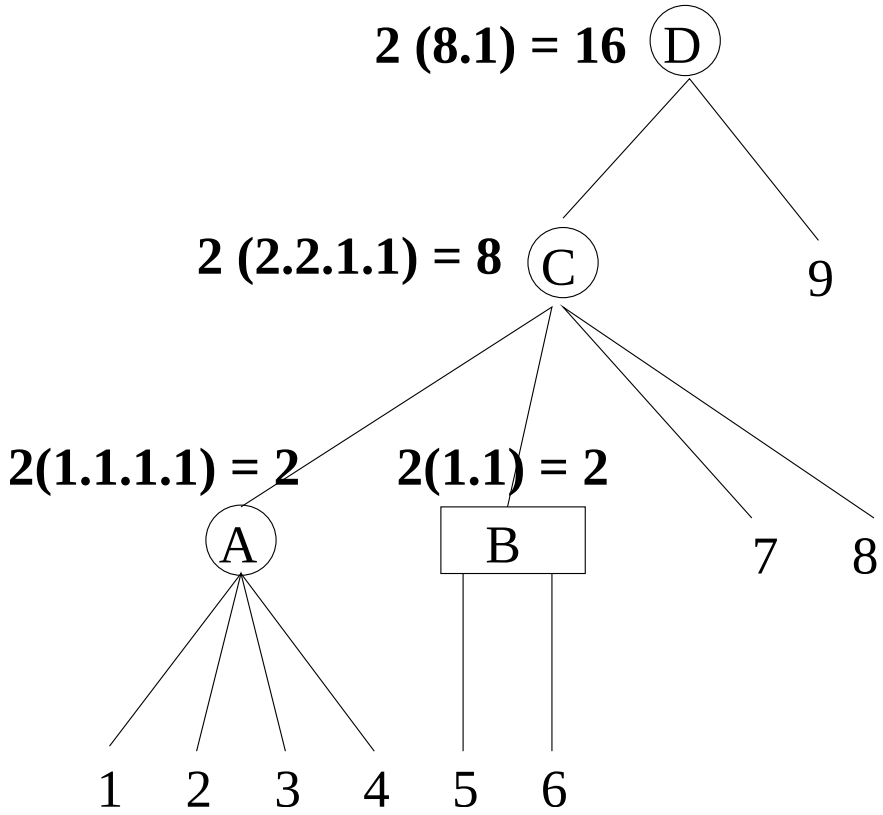
Number of nested arrangements with the extreme element in the smallest interval:

$$S_{cnt}(u,l) = 2S_{cnt}(u\text{-}1,l) + \sum_{y=3}^{u-l} Pa(\Delta_{u-y}) S_{cnt}(u-y,l)$$

Back to ... Estimating the frontier size

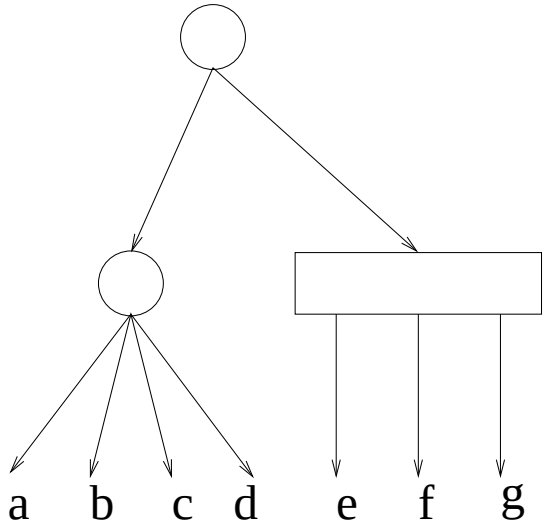
$$\#(A) = \begin{cases} 1 & \text{if } A \text{ is a leaf node,} \\ 2\prod_{j=1}^{c} \#(A_j) & \text{if } A \text{ is a } Q \text{ node,} \\ Pa(c)\prod_{j=1}^{c} \#(A_j) & \text{if } A \text{ is a } P \text{ node.} \end{cases}$$

Estimating the frontier size

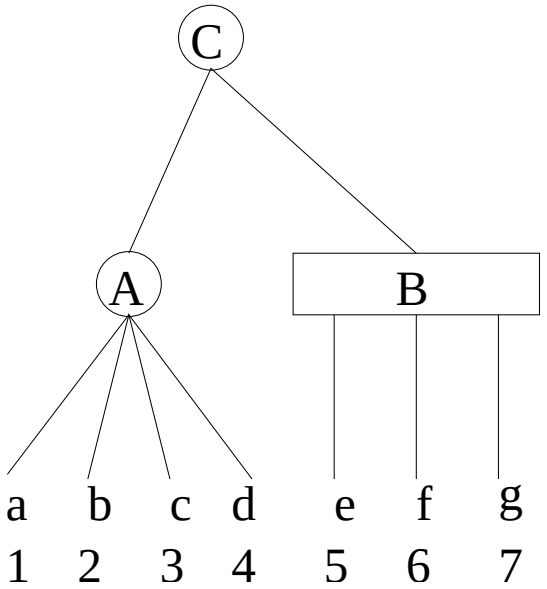


Computation of $\#(X)$ for each node X in the PQ tree.
Note that $\#(A) = \#(B) = 2$, $\#(C) = 8$, and $\#(D) = 16$.

Estimating $Fr(T')$: A complete example

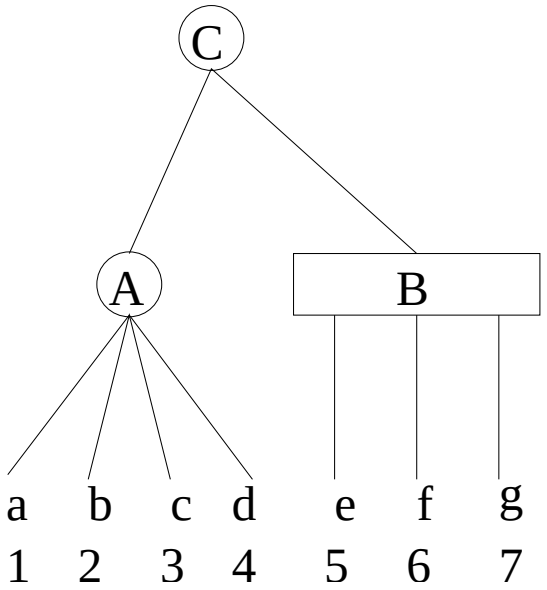


(1) The input PQ tree T.



(2) Numbering the leaf nodes.

Example...



(2) Numbering the leaf nodes & labeling the internal nodes.

Node A

3142	2413

Node B

567	765

Node C

AB	BA

(3) The possible arrangmeents.

Estimating the frontier size

1234567	
3142567	7652413
3142765	5672413
2413567	7653142
2413765	5673142
7654321	

(4) The 10 possible arrangements.

abcdefg	
cadbefg	gfebdac
cadbgfe	efgbdac
bdacefg	gfecadb
bdacgfe	efgcadb
gfedcba	

(5) Arrangements in input alphabet.

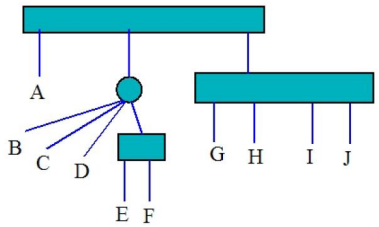
Problems in Combinatorics

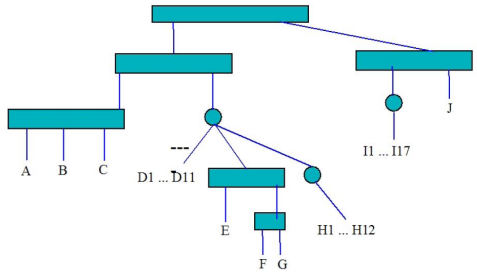
- What is the number of **simple permutations** of degree k?
  What is the number of P-arrangements of size k?
  We give the first explicit formula for this number.
- $\mathcal{O}(n^2)$ time to recognize a simple permutation.
  We give $\mathcal{O}(n)$ time to recognize a P-arrangement.

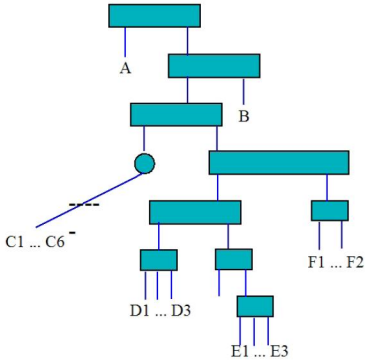
Combinatorics to probabilities

$$pr(T) = \frac{|Fr(T')|}{n!}$$

Back to the clusters







166	0.236×10^{-3}
303	0.376×10^{-4}
250	0.404×10^{-4}

Open Problems-1

Structured clusters with multiplicities

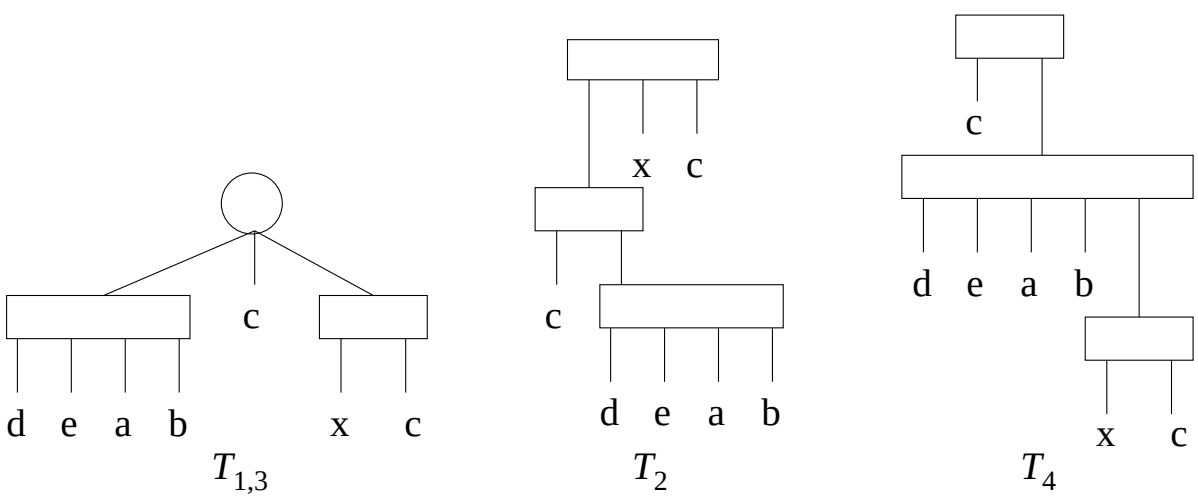
Open Problems-1

Structured clusters with multiplicities

$$p = \{a, b, c(2), d, e, x\},$$

with exactly three occurrences given as

$$\begin{aligned} o_1 &= d\ e\ a\ b\ c\ x\ c, \\ o_2 &= c\ d\ e\ a\ b\ x\ c, \\ o_3 &= c\ x\ c\ b\ a\ e\ d. \end{aligned}$$

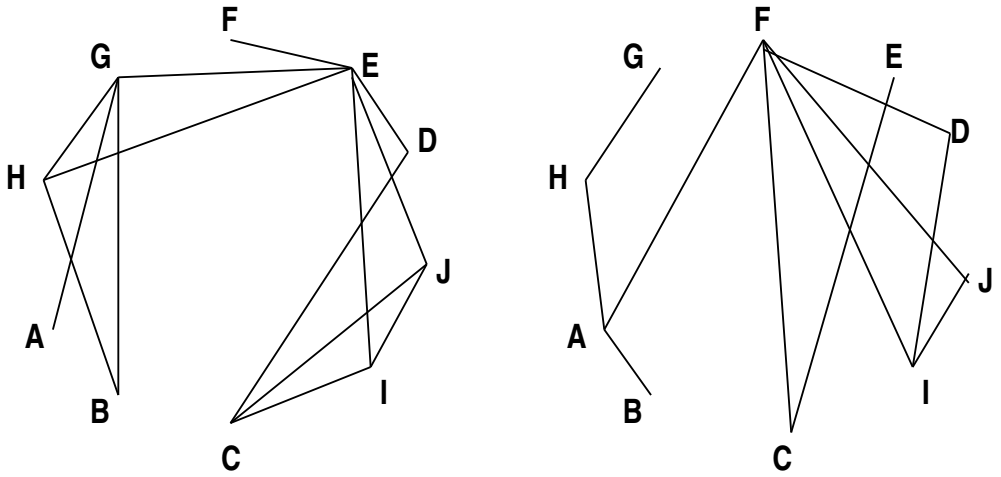


Open Problems-2

Clusters on networks (common connected components)

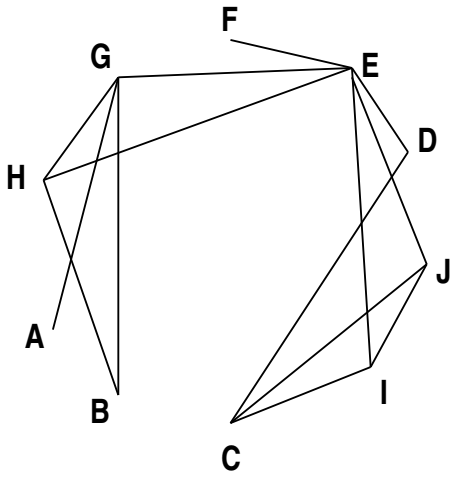
Open Problems-2

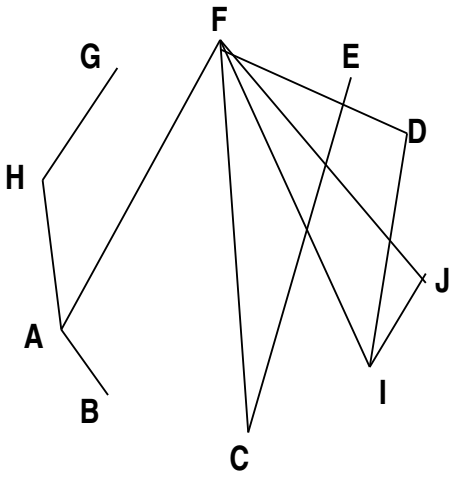
Clusters on networks (common connected components)

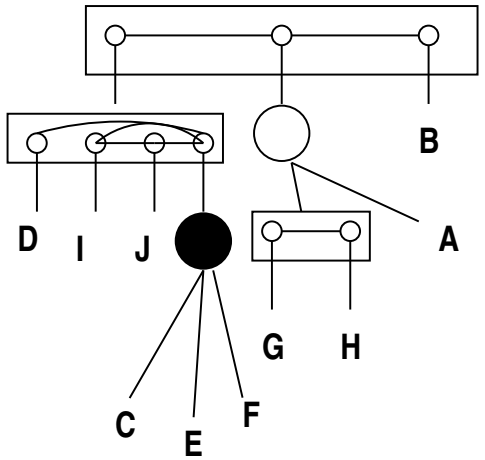


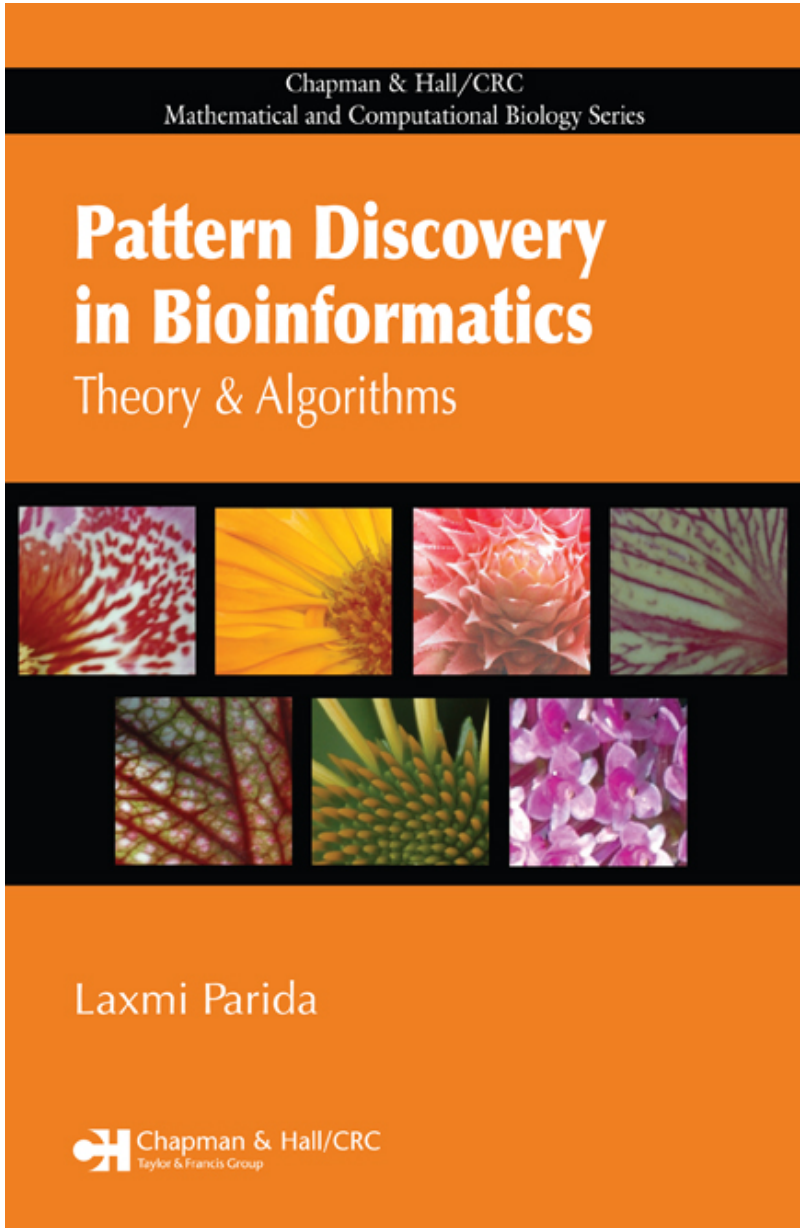
Open Problems-2

Clusters on networks (common connected components)








Chapman & Hall/CRC
Mathematical and Computational Biology Series
Pattern Discovery in Bioinformatics
Theory & Algorithms
Laxmi Parida
Chapman & Hall/CRC
Taylor & Francis Group