Finding All Sorting Tandem Duplication Random Loss Operations

Matthias Bernt[1], Ming-Chiang Chen[3], Daniel Merkle[2], Hung-Lung Wang[3], Kun-Mao Chao[3], Martin Middendorf[1]

[1]Parallel Computing and Complex Systems Group, Department of Computer Science, University of Leipzig, Germany

[2]Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark

[3]Department of Computer Science and Information Engineering, National Taiwan University, Taiwan

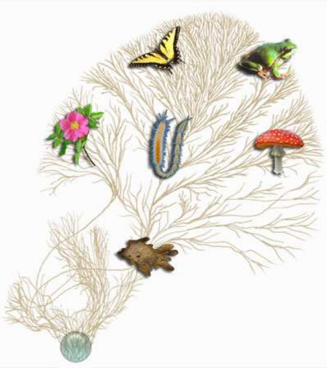
Outline

Genome Rearrangement

>species1
.. G1 G2 G3 G4 G5 G6 G7 ..
>species2
.. G4 G2 G7 G1 -G3 G5 G6 ..

⋮

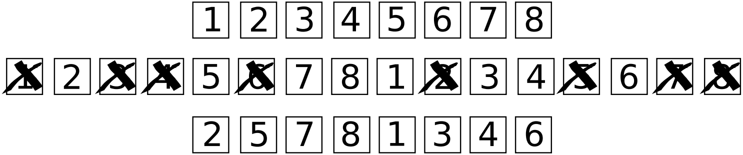
>species*m*
.. G7 G1 G2 G6 G5 G4 G3 ..



www.tolweb.org

Gene arrangements = permutations.

Tandem Duplication Random Loss (TDRL)

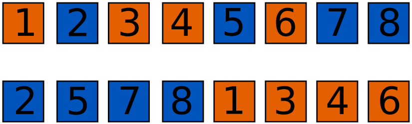


Definition (TDRL)

TDRL $\tau(F,S)$ defined by:

- F the set of elements kept in the first copy and
- S the set of elements kept in the second copy.

Tandem Duplication Random Loss (TDRL)



Definition (TDRL)

TDRL $\tau(F,S)$ defined by:

- F the set of elements kept in the first copy and
- S the set of elements kept in the second copy.

Relevance of TDRLs

S. Purpuratus — *C. miniata*

ND4L CO2 K ATP8 ATP6 CO3 -S2 ND3 ND4 H S1 ND5 -ND6 CYTB F 12S E T P -Q N L1 -A W C -V

E P N L1 W -V ND4L CO2 K ATP8 ATP6 CO3 -S2 ND3 ND4 H S1 ND5 -ND6 CYTB F 12S T -Q -A C

C. Sloani — *E. Plecanoides*

ATP8 ATP6 CO3 G ND3 R ND4L ND4 H S1 L1 ND5 -ND6 -E CYTB T -P F 12S V 16S L2 ND1 I -Q M ND2 W

R ND4L ND4 H S1 -E CYTB F 12S V 16S L2 ND1 -Q W ATP8 ATP6 CO3 G ND3 L1 ND5 -ND6 T -P I M ND2

S. fontinalis — *B. nectabanus*

-Q M ND2 W -A -N -C -Y CO1 -S2 D CO2 K ATP8 ATP6 CO3 G ND3 R ND4L ND4 H S1 L1 ND5 -ND6 -E CYTB T -P F 12S V 16S L2 ND1

ND2 W -Y CO1 -S2 D CO2 K ATP8 ATP6 CO3 G ND3 R ND4L ND4 H L1 ND5 -ND6 -P F 12S V 16S ND1 -Q M -A -N -C S1 -E CYTB T L2

L. polyphemus — *S. coleoptrata*

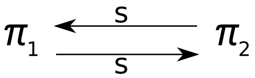
ND3 A R N S2 E -F -ND5 -H -ND4 -ND4L T -P ND6 COB S -ND1 -L2 -L -RNL -V -RNS I -Q M ND2 W -C -Y

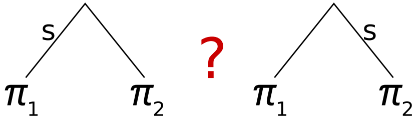
A R S2 E -F -ND5 -ND4L ND6 S -ND1 M -C -Y ND3 N -H -ND4 T -P COB -L2 -L -RNL -V -RNS I -Q ND2 W

Relevance of TDRLs

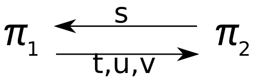
Asymmetry provides additional phylogenetic information

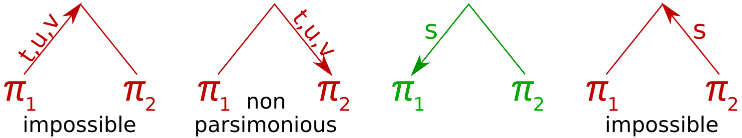
Symmetric operation:
(Inversion, Transposition, ...)





Asymmetric operation:
(TDRL)





Sorting by TDRLs

Definition (Sorting by TDRLs)

Given a permutation π. Find a shortest sequence of TDRLs $\tau_1, \ldots, \tau_{d(\pi)}$ such that $\pi \circ \tau_1 \circ \ldots \circ \tau_{d(\pi)} = \iota$.
The TDRL distance is the length of the sequence, denoted by $d(\pi)$.

Definition (Chain of a permutation π)

A chain of a permutation π is a maximal list $(e_1, \ldots, e_k)$ of elements of π where $e_{i+1} = e_i + 1$ and $\pi^{-1}(e_i) < \pi^{-1}(e_{i+1})$.
Number of chains of π: $\rho(\pi)$.

Sorting $\equiv$ Merge the chains in order to get one chain.
Indexing Scheme: $c < c'$ iff $\forall e \in c, \forall e' \in c' : e < e'$

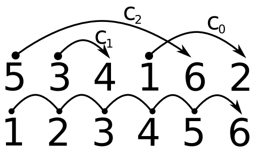
Sorting by TDRLs

Definition (Sorting by TDRLs)

Given a permutation π. Find a shortest sequence of TDRLs $\tau_1, \ldots, \tau_{d(\pi)}$ such that $\pi \circ \tau_1 \circ \ldots \circ \tau_{d(\pi)} = \iota$.
The TDRL distance is the length of the sequence, denoted by $d(\pi)$.

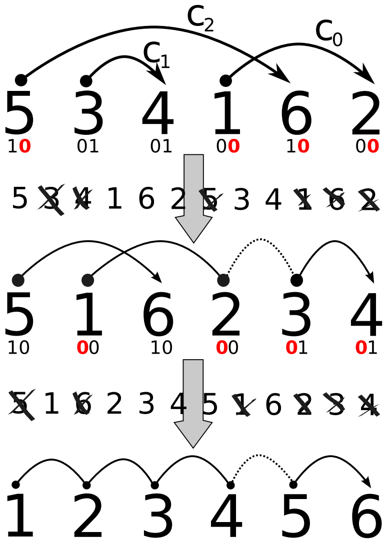
Definition (Chain of a permutation π)

A chain of a permutation π is a maximal list $(e_1, \ldots, e_k)$ of elements of π where $e_{i+1} = e_i + 1$ and $\pi^{-1}(e_i) < \pi^{-1}(e_{i+1})$.
Number of chains of π: $\rho(\pi)$.



Sorting $\equiv$ Merge the chains in order to get one chain.
Indexing Scheme: $c < c'$ iff $\forall e \in c, \forall e' \in c' : e < e'$

Sorting by TDRLs [Chaudhuri et al., SODA, 2006]



TDRL Distance:

$d(\pi) = \lceil \log_2(\rho(\pi)) \rceil$

Radix-Sort inspired algorithm:

- Get the binary representation of the chain index of each element
- In the i-th step: keep the elements of chains with a 0 at the i-th least significant bit in the first copy

All Sorting TDRLs

Problem Definition

Question 1: Are there alternative sorting TDRL scenarios?

Definition (All Sorting TDRLs)

Find the set of TDRLS $\{\tau : d(\pi \circ \tau) < d(\pi)\}$.

Question 2: How many sorting TDRLs are there?

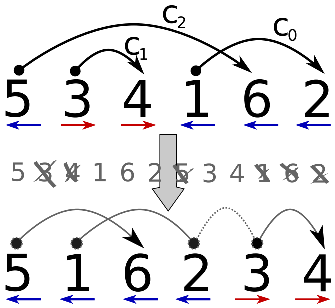
Definition (Number of Sorting TDRLs)

Determine $|\{\tau : d(\pi \circ \tau) < d(\pi)\}|$.

Basic Properties

Observations

- *Elements kept in the 1st (2nd) copy are moved to the left (right)*
- *The order of the elements kept in the same copy is not changed*



Restricted TDRLs

Definition (Restricted TDRL)

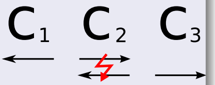
All elements of a chain are kept in the same copy.

Proposition

Two chains c_i and c_j can be connected with a TDRL iff $j = i + 1$.
This can be done by keeping the elements of c_i in the 1st copy and the elements of c_{i+1} in the 2nd copy.

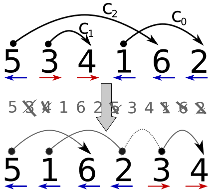
Proposition

Three chains c_i, c_{i+1}, and c_{i+2} can not be connected at once.
$\Rightarrow$ *Restricted TDRL distance = TDRL distance*



Reformulation

- Restricted TDRL $\equiv$
  Binary string s of length $\rho(\pi)$:
  $s_i = 1 \leftrightarrow c_i$ is kept in the 1st copy
  $s_i = 2 \leftrightarrow c_i$ is kept in the 2nd copy
- c_i and c_{i+1} get connected iff $s_i s_{i+1} = 12$



s_0	s_1	s_2
1	2	1

Question:

- How many strings of length n with at least k 12-transitions?

String Count

Number of strings of length n which have exactly k 12-transitions.

$$s_1 = 1 \wedge s_n = 1 \rightarrow \binom{n-1}{2k}$$
$$s_1 = 1 \wedge s_n = 2 \rightarrow \binom{n-1}{2k-1}$$
$$s_1 = 2 \wedge s_n = 1 \rightarrow \binom{n-1}{2k+1}$$
$$s_1 = 2 \wedge s_n = 2 \rightarrow \binom{n-1}{2k}$$

1▲2▲1 1▲2 2▲1 1

1▲2▲1 1▲2 2▲1▲2

2 2▲1 1▲2 2▲1 1

2 2▲1 1▲2 2▲1▲2

$$= \binom{n+1}{2k+1}$$

Number of strings of length n which have at least k 12-transitions.

$$= \sum_{i=k}^{\lfloor \frac{n}{2} \rfloor} \binom{n+1}{2i+1}$$

Result

Theorem

For a permutation π with ρ chains there are

$$\sum_{i=\rho-2^{\lceil \log_2(\rho) \rceil -1}}^{\lfloor \frac{\rho}{2} \rfloor} \binom{\rho+1}{2i+1}$$

sorting restricted TDRLs.

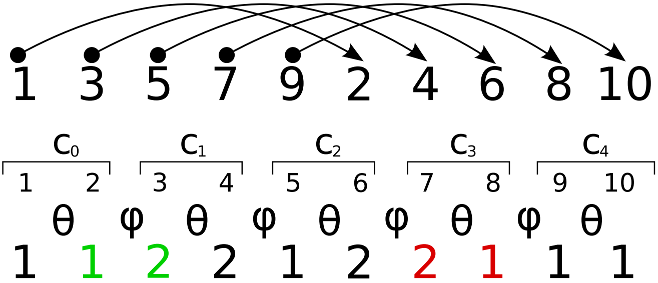
For $\rho = 2^x$:

- Only one sorting TDRL
- Only one sorting TDRL scenario

In general:

- Each sorting scenario is unique after $\left\lceil \log_2(2^{\lceil \log_2(\rho) \rceil} - \rho + 1) \right\rceil$ sorting TDRLs.

All Sorting TDRLs

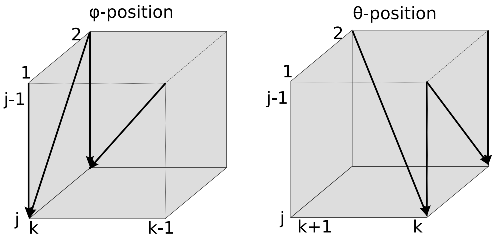


- 21 at θ-positions breaks a chain $\longrightarrow$ ρ increased by 1
- 12 at ϕ-positions connect chains $\longrightarrow$ ρ decreased by 1

Question: How many binary strings of length n with $k \leq \Delta\rho$.

Dynamic Programming Approach

$a^{x}_{j,k}$: Number of possible binary strings of length j ending with $x \in \{1, 2\}$ that change the number of chains by k.



Dynamic Programming Approach

$a^x_{j,k}$: Number of possible binary strings of length j ending with $x \in \{1,2\}$ that change the number of chains by k.

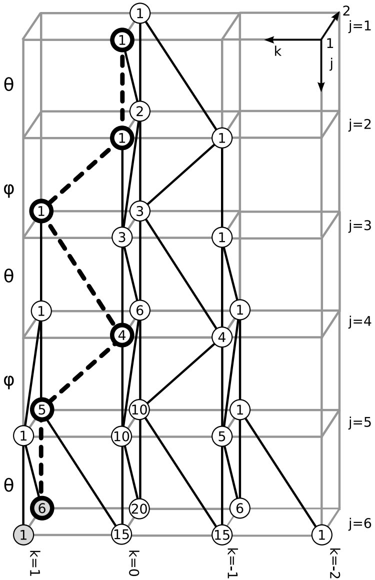
$$a^2_{j,k} = \begin{cases} a^1_{j-1,k+1} + a^2_{j-1,k} & \text{if } p_{j-1} = \phi \\ a^1_{j-1,k} + a^2_{j-1,k} & \text{else} \end{cases}$$

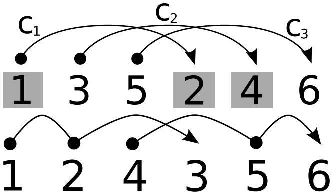
$$a^1_{j,k} = \begin{cases} a^1_{j-1,k} + a^2_{j-1,k-1} & \text{if } p_{j-1} = \theta \\ a^1_{j-1,k} + a^2_{j-1,k} & \text{else} \end{cases}$$

$a^1_{1,0} = 1$, $a^2_{1,0} = 1$ other values initialised to 0

Number of sorting TDRL: $$\sum_{i=\rho-2^{\lceil \log_2(\rho) \rceil -1}}^{\lfloor \frac{\rho}{2} \rfloor} a^1_{n,i} + a^2_{n,i}$$

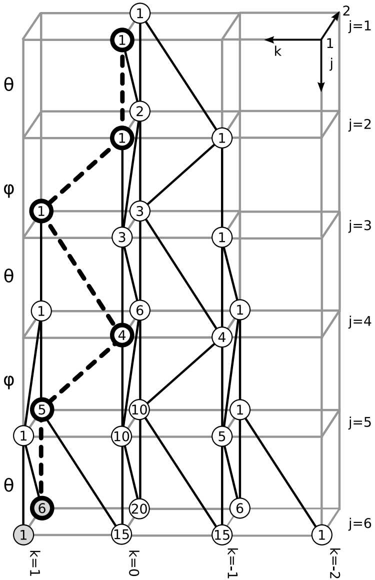
Dynamic Programming Approach





Sorting TDRL events can be enumerated by backtracking.

Equalities



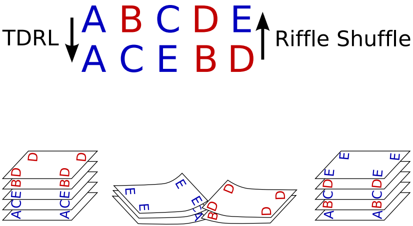
Number of sorting TDRLs:

$$\sum_{i=0}^{2^{\lceil \log_2(\rho(\pi))\rceil}-\rho(\pi)} \binom{n}{i}$$

Number of sorting restricted TDRLs:

$$\sum_{i=0}^{2^{\lceil \log_2(\rho(\pi))\rceil}-\rho(\pi)} \binom{\rho(\pi)}{i}$$

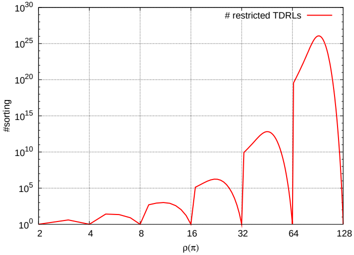
Riffle Shuffle

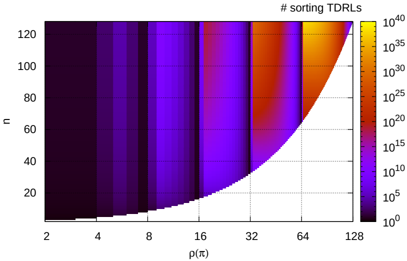


Riffle Shuffle Distance:

- Schwenk *Elementary Problem: E3143*, Am. Math Mon., 1986
- Schwenk *E3143*, Am. Math Mon., 1988

Number of Sorting TDRLs



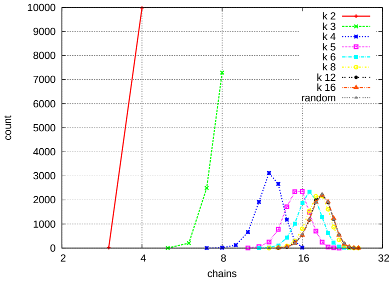


- Only 1 sorting TDRL for $\rho = 2^x$

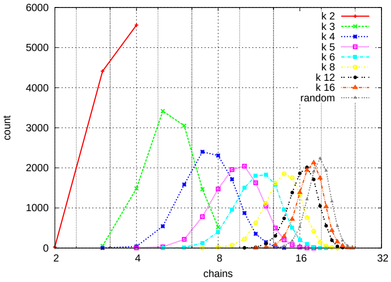
Number of Chains in Simulated Permutations

- Apply k “random” TDRLs on ι

Random TDRLs



Random TDRLs in random interval



- Unique sorting scenarios for $\rho \in \{2, 4\}$ (and $\rho = 8$) are very likely

Mitochondrial Data

Previously unpublished scenario of two TDRLs:

S. fontinalis

CO1 -S2 D CO2 K ATP8 ATP6 CO3 G ND3 R ND4L ND4 H S1 L1 ND5 -ND6 -E CYTB T -P F 12S V 16S L2 ND1 I -Q M ND2 W -A -N -C -Y

CO1 -S2 K R ND4L ND4 H -ND6 -E CYTB T -P F 12S V 16S L2 ND1 I -Q M ND2 -A -N -C -Y D CO2 ATP8 ATP6 CO3 G ND3 S1 L1 ND5 W

CO1 -S2 K R ND4L ND4 H -ND6 -E CYTB T -P F 12S V 16S L2 ND1 I -Q M ND2 -A -N -C -Y D CO2 ATP8 ATP6 CO3 G ND3 S1 L1 ND5 W

CO1 H -ND6 -E CYTB 12S V 16S L2 ND1 -A -N -Y D CO2 ATP8 ATP6 CO3 G ND3 S1 L1 ND5 -S2 K R ND4L ND4 T -P F I -Q M ND2 -C W

P. myriaster

Support:

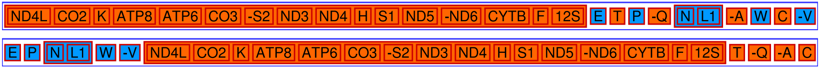
- 4 chains $\longrightarrow$ unique scenario
- *P. myriaster* $\longrightarrow$ *S. fontinalis* needs 3 TDRLs
- Reversal distance is 15
- Transposition distance is 7
- There exists no TDRL median with score 2 or less
- Fragments of duplicated sequences support TDRLs [Miya05]

Conclusion

- Method for enumerating all sorting TDRLs
- Closed formulas for the number of sorting TDRLs
- Identification of unique TDRL sorting scenarios possible
- Identification of likely unique TDRLs scenario in mitochondria

Thank You!


NC_005929 17538nt
Cucumaria miniata
0K
1K
2K
3K
4K
5K
6K
7K
8K
9K
10K
11K
12K
13K
14K
15K
16K
17K
COX1
R
E
CR
P
N
L1
W
V
ND4L
COX2
K
ATP8
ATP6
COX3
S2
ND3
ND4
H
S1
ND5
ND6
CYTB
F
12S
T
CR
Q
A
C
M
Y
G
L2
ND1
ND2
16S


ND4L CO2 K ATP8 ATP6 CO3 -S2 ND3 ND4 H S1 ND5 -ND6 CYTB F 12S E T P -Q N L1 -A W C -V
E P N L1 W -V ND4L CO2 K ATP8 ATP6 CO3 -S2 ND3 ND4 H S1 ND5 -ND6 CYTB F 12S T -Q -A C