Another problem with variants on either side of P vs. NP divide

Leonid Chindelevitch

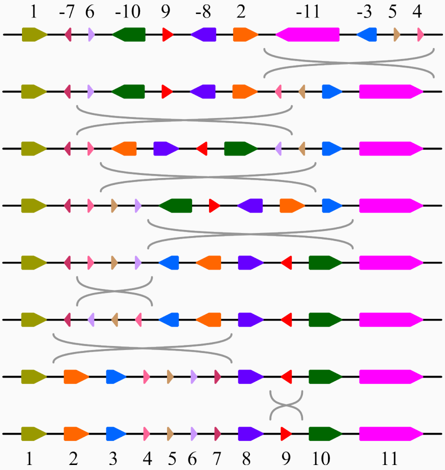
28 March 2019

Theory Seminar, Spring 2019, Simon Fraser University

Background

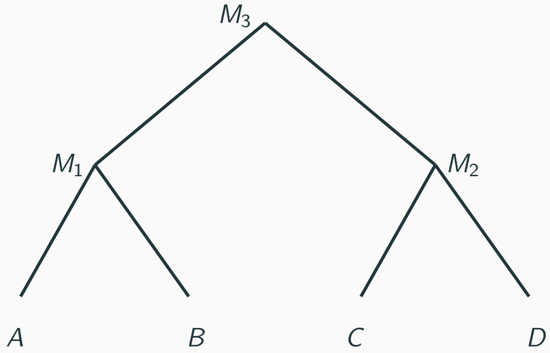
From mice to men through genome rearrangements

Mouse X-Chromosome



Human X-Chromosome

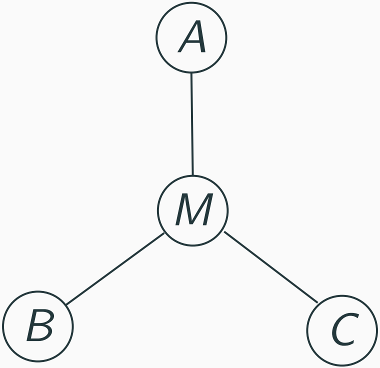
Ancestral Reconstruction



Input: Tree and genomes A, B, C, D

Output: Ancestral genomes M_1, M_2, M_3

The Median of Three

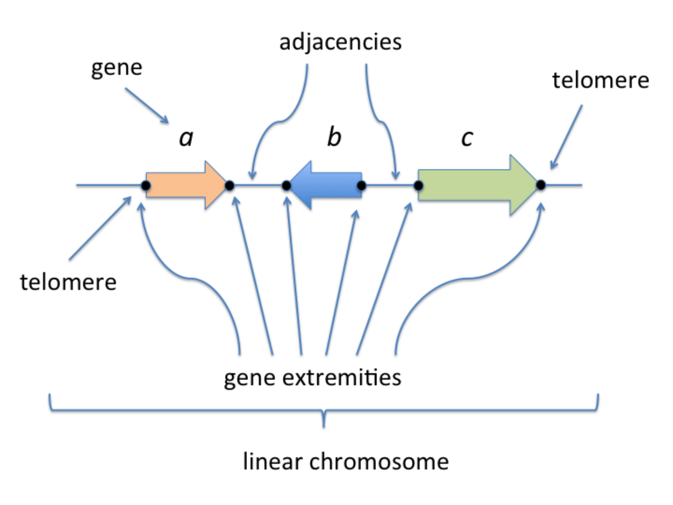


Input: Genomes A, B, C

Output: Genome M (the median, AKA the lowest common ancestor) which minimizes

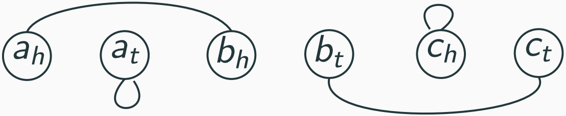
$$d(A, M) + d(B, M) + d(C, M)$$

Genome Elements

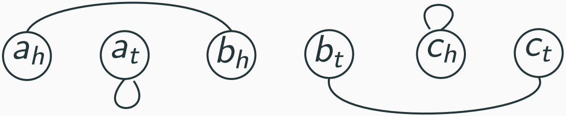


Adjacencies: $\{a_h, b_h\}, \{b_t, c_t\}$; telomeres: a_t, c_h

Genome Representations



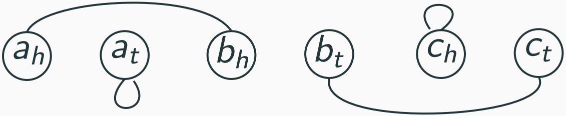
Genome Representations



$$\begin{array}{c|cccccc} & a_t & a_h & b_t & b_h & c_t & c_h \\ \hline a_t & 1 & 0 & 0 & 0 & 0 & 0 \\ a_h & 0 & 0 & 0 & 1 & 0 & 0 \\ b_t & 0 & 0 & 0 & 0 & 1 & 0 \\ b_h & 0 & 1 & 0 & 0 & 0 & 0 \\ c_t & 0 & 0 & 1 & 0 & 0 & 0 \\ c_h & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

This is a *genome matrix*.

Genome Representations



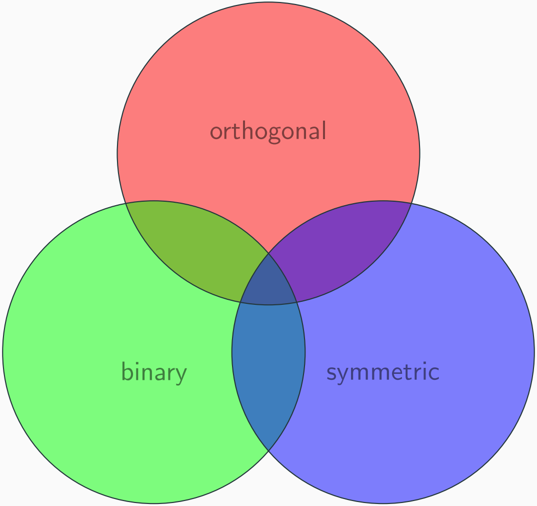
$$\begin{array}{c|cccccc} & a_t & a_h & b_t & b_h & c_t & c_h \\ \hline a_t & 1 & 0 & 0 & 0 & 0 & 0 \\ a_h & 0 & 0 & 0 & 1 & 0 & 0 \\ b_t & 0 & 0 & 0 & 0 & 1 & 0 \\ b_h & 0 & 1 & 0 & 0 & 0 & 0 \\ c_t & 0 & 0 & 1 & 0 & 0 & 0 \\ c_h & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

This is a *genome matrix*.

Genome matrices can be represented by involutions: $(a_h\ b_h)(b_t\ c_t)$.

Properties of Genome Matrices

- binary matrices that satisfy $A = A^T = A^{-1}$
- even dimension n (but we can relax this assumption)



Rank Distance

The **rank distance** between two genome matrices is the rank of their difference

$$d(A, B) = r(A - B)$$

Properties

- $d(A, B) \geq 0$; $d(A, B) = 0$ if and only if $A = B$
- $d(A, B) = d(B, A)$
- $d(A, C) \leq d(A, B) + d(B, C)$

Rank Distance

The **rank distance** between two genome matrices is the rank of their difference

$$d(A, B) = r(A - B)$$

Properties

- $d(A, B) \geq 0$; $d(A, B) = 0$ if and only if $A = B$
- $d(A, B) = d(B, A)$
- $d(A, C) \leq d(A, B) + d(B, C)$

This is a **metric** on the space of genome matrices (and matrices in general).

Equivalence of Rank Distance and the Cayley Distance

Lemma

Consider permutations matrices P,Q, with permutation representations $\pi, \tau \in S_n$, respectively. Then

$$d(P,Q) = ||\tau\pi^{-1}||$$

where $||\cdot||$ is the minimum number of cycles in a 2-cycle decomposition.

Equivalence of Rank Distance and the Cayley Distance

Lemma

Consider permutations matrices P,Q, with permutation representations $\pi, \tau \in S_n$, respectively. Then

$$d(P, Q) = ||\tau\pi^{-1}||$$

where $||\cdot||$ is the minimum number of cycles in a 2-cycle decomposition.

Remark

$||\cdot||$ is a metric on permutations, also referred to as the Cayley distance.

Equivalence of Rank Distance and the Cayley Distance

Lemma

Consider permutations matrices P,Q, with permutation representations $\pi, \tau \in S_n$, respectively. Then

$$d(P,Q) = ||\tau\pi^{-1}||$$

where $||\cdot||$ is the minimum number of cycles in a 2-cycle decomposition.

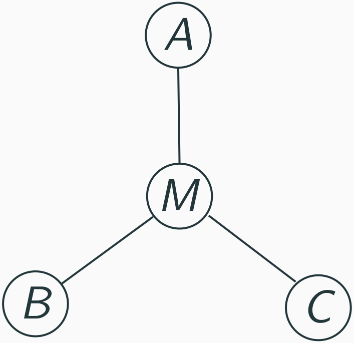
Remark

$||\cdot||$ is a metric on permutations, also referred to as the Cayley distance.

Remark

We may as well work with involutions in S_n instead of genome matrices.

The Rank Median Problem

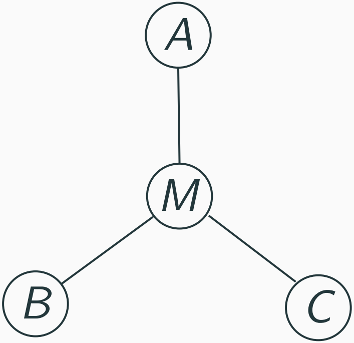


Input: Genome *Matrices* A, B, C

Output: Matrix M (the median) which minimizes

$$s(M; A, B, C) = d(A, M) + d(B, M) + d(C, M)$$

The Rank Median Problem



Input: Genome *Matrices* A, B, C

Output: Matrix M (the median) which minimizes

$$s(M; A, B, C) = d(A, M) + d(B, M) + d(C, M)$$

What kind of matrix should M be?

Types of medians

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

Types of medians

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

Generalized median: minimizer of $d(A, M) + d(B, M) + d(C, M)$ over all *real valued matrices*

$$\begin{bmatrix} -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{bmatrix}$$

Types of medians

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

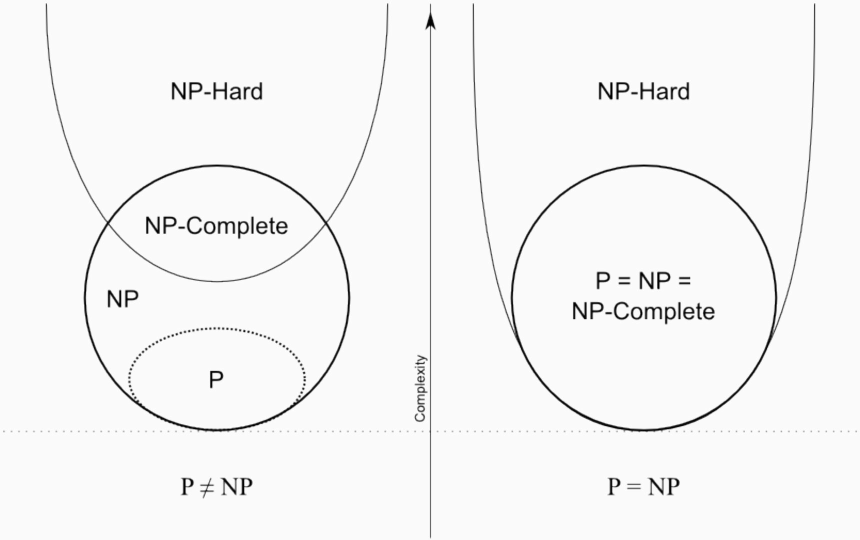
Generalized median: minimizer of $d(A, M) + d(B, M) + d(C, M)$ over all *real valued matrices*

$$\begin{bmatrix} -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{bmatrix}$$

Genome median: minimizer of $d(A, M) + d(B, M) + d(C, M)$ over all *genome matrices*

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

P, NP, and NP-hard



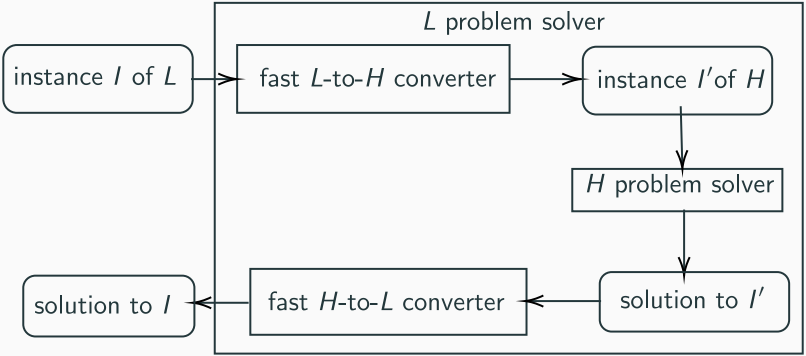
Problems with variants on both sides of the P vs. NP divide

Problem type	P variant	NP-hard variant
Cover	Edge cover	Vertex cover
Satisfiability	2-CNF-SAT	3-CNF-SAT
Graph mapping	*Graph isomorphism*	Subgraph isomorphism
Optimization	Linear programming	Integer programming
Median-of-three	**Generalized median**	**Genome median**

NP-hard

NP-hard is the set of problems which are "at least as hard as hardest problems in NP".

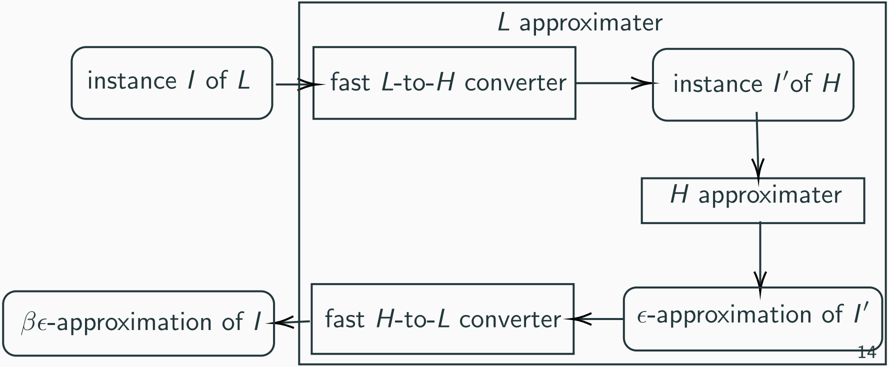
i.e. there is a polynomial time *reduction* from any problem $L \in NP$ to $H \in$ NP-hard.



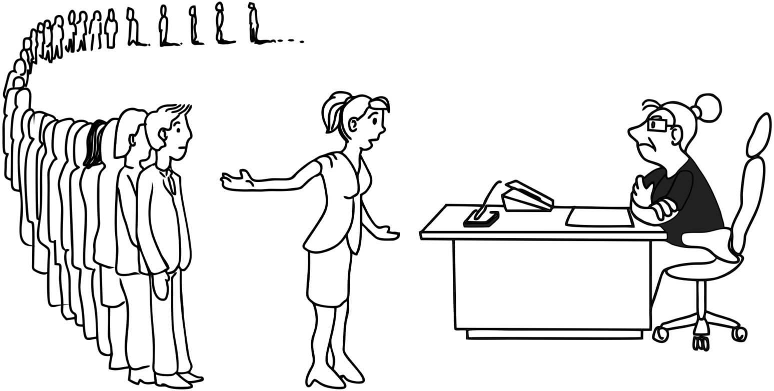
APX-hard

APX is the set of problems which have polynomial time constant-factor approximation algorithms.

APX-hard is the set of problems where there exists a *polynomial time approximation scheme reduction* from any problem $L \in$ APX to any problem $H \in$ APX-hard.


L approximater
instance I of L
fast L-to-H converter
instance $I'$of H
H approximater
ϵ-approximation of I'
fast H-to-L converter
$\beta\epsilon$-approximation of I

Computational Complexity



“I can’t find an efficient algorithm, but neither can all these famous people.”

The Generalized Median problem

Properties of the Median

- Lower Bound

$$d(M, A) + d(M, B) + d(M, C) \geq \frac{d(A, B) + d(B, C) + d(C, A)}{2} := \beta$$

- At least one of the "corners" (input genomes) is a $\frac{4}{3}$ approximation of the median
- The lower bound is achieved if and only if

$$d(M, A) = \frac{d(A, B) + d(C, A) - d(B, C)}{2}$$

  and likewise for $d(M, B)$ and $d(M, C)$.
- Not every A, B, C can achieve the lower bound β, e.g.:

$$A = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, B = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Approximating Matrix Medians

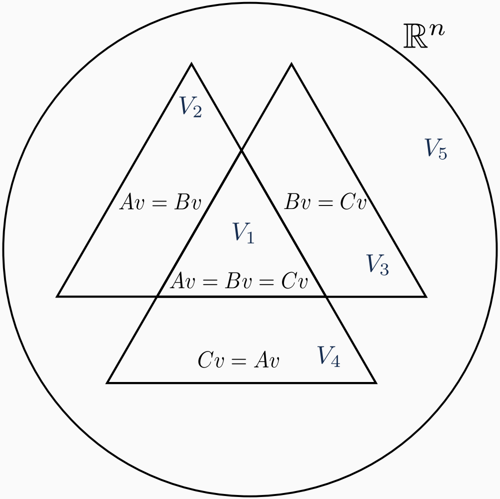
- Interesting Property

Theorem

For any three $n \times n$ matrices A, B, and C there is a median M satisfying: for all vectors $v \in \mathbb{R}^n$ such that $Av = Bv = Cv$, we have $Mv = Av$.

- We define the invariant $\alpha := \dim(\{v | Av = Bv = Cv\})$.
- For permutations, this can be computed in $O(n)$ time via graph union.
- Can we say the same if we have $Av = Bv$? We don't know [yes for orthogonal A, B, C].
- However, we can act on this idea.

Subspace decomposition



Approximation Algorithm

$$\begin{tikzcd}
\text{Subspaces} & V_1 \arrow[d] & V_2 \arrow[d] & V_3 \arrow[d] & V_4 \arrow[d] & V_5 \arrow[d] \\
\text{Orthonormal Bases} & S_1 \arrow[d] & S_2 \arrow[d] & S_3 \arrow[d] & S_4 \arrow[d] & S_5 \arrow[d] \\
\text{Projection Matrices} & P_1 \arrow[d] & P_2 \arrow[d] & P_3 \arrow[d] & P_4 \arrow[d] & P_5 \arrow[d] \\
 & M_A = AP_1 & + AP_2 & + BP_3 & + AP_4 & + AP_5 \\
\text{Median Candidates} & M_B = BP_1 & + BP_2 & + BP_3 & + AP_4 & + BP_5 \\
 & M_C = CP_1 & + BP_2 & + CP_3 & + CP_4 & + CP_5
\end{tikzcd}$$

- $\frac{4}{3}$ approximation factor for genome matrices
- if $V_5 = \{0\}$ then each candidate is a median (its score is β)
- In general, $\dim(V_5) := 2\delta$, where $\delta := \alpha + \beta - n$ is called the "deficiency" of the triplet A, B, C.

Some recently proven theorems

$$M_I := AP_1 + AP_2 + BP_3 + AP_4 + P_5$$

Theorem: M_I is a median for any genomic inputs A, B, C.

Theorem: $M_I = I + ([AV_1, AV_2, BV_3, AV_4] - V_{14})(V_{14}^T V_{14})^{-1} V_{14}^T$.

Corollary: It is possible to compute M_I in $O(n^\omega)$ time, where ω is Strassen's exponent, in exact or floating-point arithmetic.

Theorem: The matrix M_I is always symmetric and orthogonal for genomic inputs A, B, C.

Special case: If $A = I$, then $\delta = 0$, so $M_A = M_B = M_C = M_I$ and each one has a score of β.

An even faster, $O(n^2)$, algorithm when $\delta = 0$

Theorem: If a matrix M satisfies

$$d(A, M) + d(M, B) = d(A, B),$$

then there exists a projection matrix P such that

$$M = A + P(B - A).$$

- We can ignore the condition that P is a projection matrix.
- This yields the system

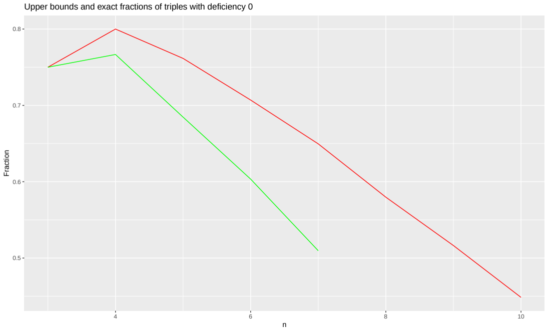
  $$M = A + P(B - A) = B + Q(C - B) = C + R(A - C),$$

  from which we eliminate M and any redundancies.
- It splits into n linear systems with the same left-hand side.
- If A, B, C are permutations, $\delta = 0$, each equation has 2 variables; the Aspvall-Shiloach algorithm solves such systems in $O(n)$ time.

Rarity of the special case $\delta = 0$

Theorem: The fraction of triples with $\delta = 0$ goes to 0 as $n \to \infty$.

Proof: This follows directly from a result in analytic combinatorics.



Challenges with computing V_5

Observation: A basis for the space $\mathrm{im}(A-B) \cap \mathrm{im}(B-C)$ can be computed in $O(n \log n)$.

Proof sketch: Let P, Q be the cycle partitions of $A^{-1}B$,C^{-1}B.

Create a multigraph G with vertices $P \cup Q$ and edges for all $i \in [n]$.

Each parallel edge i, j gives a vector $e_i - e_j \in \mathrm{im}(A-B) \cap \mathrm{im}(B-C)$.

Removing those to get a connected graph G' whose cycle basis $\mathcal{B}$ can be computed from a spanning tree.

Since G' is bipartite, each cycle $C \in \mathcal{B}$ gives rise to the vector $\chi(C^+) - \chi(C^-) \in \mathrm{im}(A-B) \cap \mathrm{im}(B-C)$.

Difficulty: $V_5 = \mathrm{im}(A-B) \cap \mathrm{im}(B-C) \cap \mathrm{im}(C-A)$ may have no nice basis; this construction fails when generalized to hypergraphs.

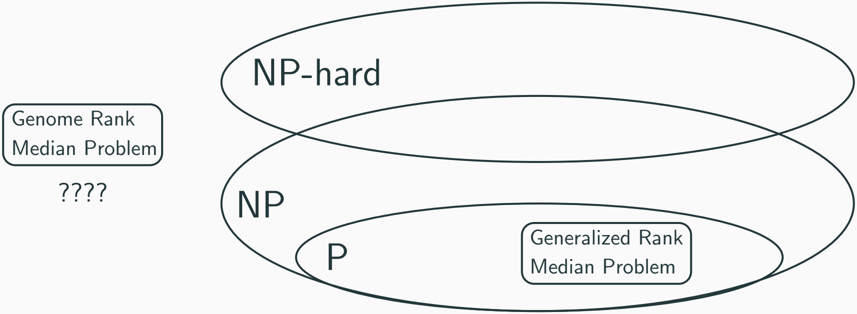
A quartic algorithm for orthogonal matrices

```
while $d(A,B) + d(B,C) > d(A,C)$
    find $u \in \operatorname{im}(A - B) \cap \operatorname{im}(B - C)$;
    $B \leftarrow \left(I - 2\frac{uu^T}{u^T u}\right)B$.
```

Remark

The transformation which multiplies a matrix on the left by $I - 2\frac{uu^T}{u^T u}$ is called a Householder reflection, and is frequently used in numerical analysis.

Complexity of Rank Median Problems



The Genome Median Problem

Genome Rank Median Problem is NP-hard and APX-hard

Theorem

The genome rank median problem of three genomes (GMP) is NP-hard and APX-hard.

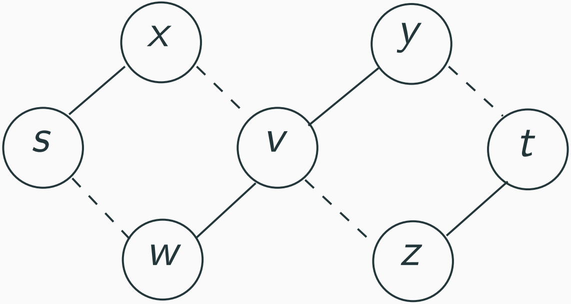
Proof.

By reduction from the *breakpoint graph decomposition problem* (BGD). □

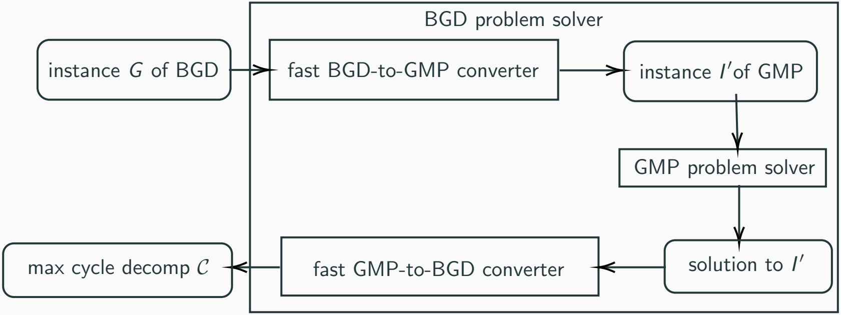
Breakpoint Graph Decomposition Problem

Objective (NP-hard): find a maximum alternating cycle decomposition $\mathcal{C}$ of a balanced bicolored graph G.

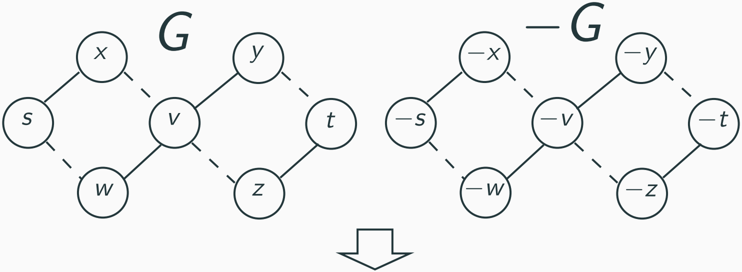
Objective (APX-hard): find an alternating cycle decomposition $\mathcal{C}$ of a balanced bicolored graph G which minimizes $|\mathcal{B}| - |\mathcal{C}|$.



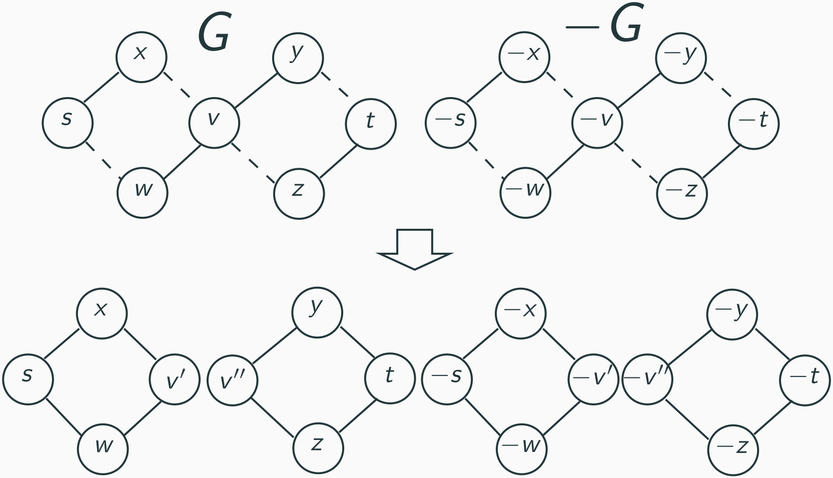
BGD Reduction Plan



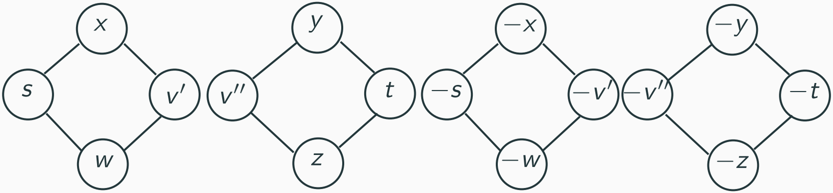
Transforming BGD into GMP



Transforming BGD into GMP



Transforming BGD into GMP



$$\pi_1 = id$$

$$\pi_2 = (v' \ v'')(-v' \ -v'')$$

$$\pi_3 = (v' \ x \ s \ w)(t \ y \ v'' \ z)(-w \ -s \ -x \ -v')(-z \ -v'' \ -y \ -t)$$

Aside: Canonical medians

A *canonical median* m_c is a median of π_1, π_2, and π_3 which contains cycles only from π_2.

$$\begin{aligned}
\pi_1 &= id \\
\pi_2 &= (v' \ v'')(-v' \ -v'') \\
\pi_3 &= (v' \ x \ s \ w)(t \ y \ v'' \ z)(-w \ -s \ -x \ -v')(-z \ -v'' \ -y \ -t) \\
m_c &= (v' \ v'')
\end{aligned}$$

Lemma

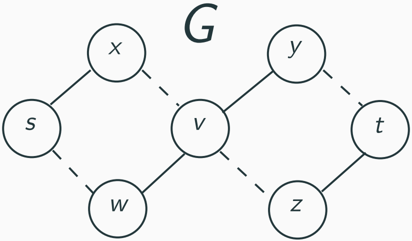
Medians of π_1, π_2, π_3 can be transformed into canonical medians in polynomial time.

Lemma

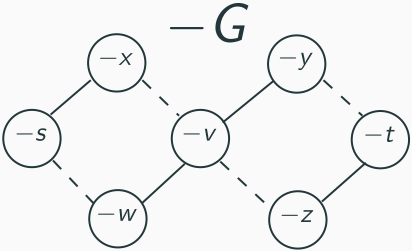
Canonical medians are in bijection to maximum cycle decompositions of G.

Aside: Canonical medians

$$m_c = m(-m)$$



m ⇔ max cycle decomp $\mathcal{C}$



$-m$ ⇔ max cycle decomp -$\mathcal{C}$

Transforming BGD into GMP

$$\Gamma = (v' \ \ - v')(v'' \ \ - v'')(s \ \ - s)(t \ \ - t)(w \ \ - w)(x \ \ - x)(z \ \ - z)(y \ \ - y).$$

Transforming BGD into GMP

$$\Gamma = (v' \ -v')(v'' \ -v'')(s \ -s)(t \ -t)(w \ -w)(x \ -x)(z \ -z)(y \ -y).$$

$$\pi_1 = id$$

$$\pi_2 = (v' \ v'')(-v' \ -v'')$$

$$\pi_3 = (v' \ x \ s \ w)(t \ y \ v'' \ z)(-w \ -s \ -x \ -v')(-z \ -v'' \ -y \ -t)$$

Transforming BGD into GMP

$$\Gamma = (v' \ -v')(v'' \ -v'')(s \ -s)(t \ -t)(w \ -w)(x \ -x)(z \ -z)(y \ -y).$$

$$\pi_1 = id$$

$$\pi_2 = (v' \ v'')(-v' \ -v'')$$

$$\pi_3 = (v' \ x \ s \ w)(t \ y \ v'' \ z)(-w \ -s \ -x \ -v')(-z \ -v'' \ -y \ -t)$$

$$\Big\Downarrow \Gamma$$

$$\pi_1\Gamma = \Gamma$$

$$\pi_2\Gamma = (v' \ -v'')(-v' \ v'')(s \ -s)(t \ -t)(w \ -w)(x \ -x)(z \ -z)(y \ -y)$$

$$\pi_3\Gamma = (v' \ -x)(x - s)\dots$$

Transforming BGD into GMP

$$\Gamma = (v' \ -v')(v'' \ -v'')(s \ -s)(t \ -t)(w \ -w)(x \ -x)(z \ -z)(y \ -y).$$

$$\pi_1 = id$$

$$\pi_2 = (v' \ v'')(-v' \ -v'')$$

$$\pi_3 = (v' \ x \ s \ w)(t \ y \ v'' \ z)(-w \ -s \ -x \ -v')(-z \ -v'' \ -y \ -t)$$

$$\Big\Downarrow \Gamma$$

$$\pi_1\Gamma = \Gamma$$

$$\pi_2\Gamma = (v' \ -v'')(-v' \ v'')(s \ -s)(t \ -t)(w \ -w)(x \ -x)(z \ -z)(y \ -y)$$

$$\pi_3\Gamma = (v' \ -x)(x - s)\ldots$$

$\pi_1\Gamma, \pi_2\Gamma, \pi_3\Gamma$ are involutions, i.e. they are an instance of **GMP**, with genome rank median $m'\Gamma$.

Transforming BGD into GMP

Proposition

The rank distance is right-multiplication invariant; that is, for $\sigma, \pi, \tau \in S_n$,

$$d(\sigma, \pi) = d(\sigma\tau, \pi\tau)$$

Transforming BGD into GMP

Proposition

The rank distance is right-multiplication invariant; that is, for $\sigma, \pi, \tau \in S_n$*,*

$$d(\sigma, \pi) = d(\sigma\tau, \pi\tau)$$

Corollary

$$s(m'\Gamma; \pi_1\Gamma, \pi_2\Gamma, \pi_3\Gamma) = s(m'; \pi_1, \pi_2, \pi_3)$$

Transforming BGD into GMP

Proposition

The rank distance is right-multiplication invariant; that is, for $\sigma, \pi, \tau \in S_n$*,*

$$d(\sigma, \pi) = d(\sigma\tau, \pi\tau)$$

Corollary

$$s(m'\Gamma; \pi_1\Gamma, \pi_2\Gamma, \pi_3\Gamma) = s(m'; \pi_1, \pi_2, \pi_3)$$

Corollary

$m'\Gamma$ *is a genome median of* $\pi_1\Gamma, \pi_2\Gamma, \pi_3\Gamma$ *if and only if* m' *is a permutation median of* π_1, π_2, π_3*.*

NP-hardness proof sketch

Theorem

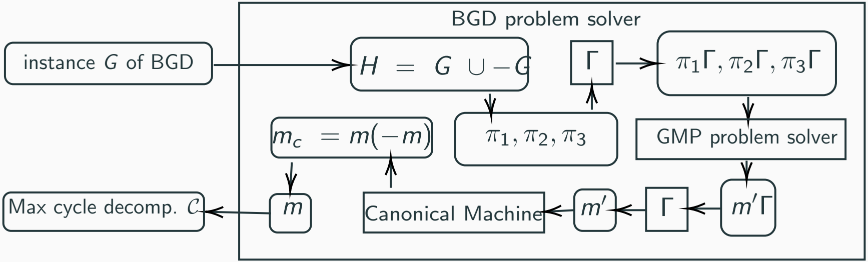
The genome rank median problem of three genomes is NP-hard.

NP-hardness proof sketch

Theorem

The genome rank median problem of three genomes is NP-hard.

Proof.



□

APX-hardness proof sketch

Theorem

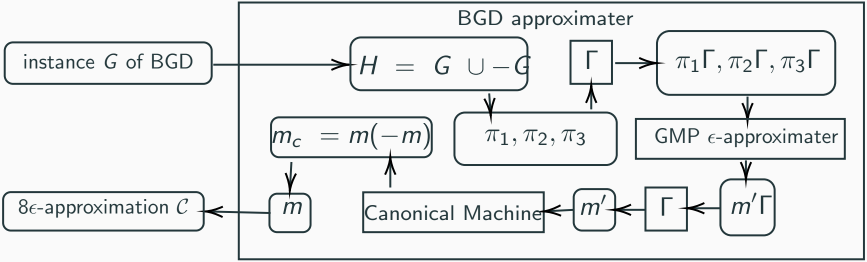
The genome rank median problem of three genomes is APX-hard.

APX-hardness proof sketch

Theorem

The genome rank median problem of three genomes is APX-hard.

Proof.



□

Conclusion and open problems

- We have a general $O(n^{\omega+1})$ algorithm for orthogonal matrices.
- We have a specialized $O(n^{\omega})$ algorithm for symmetric orthogonal matrices.
- We have a $O(n^2)$ algorithm for permutations with $\delta = 0$.

- What properties of input matrices are inherited by medians?
- Partial answer: we know that not all generalized medians are symmetric or orthogonal!
- Can we use convex optimization to find better approximations? What is the best possible ratio for approximating the genome median problem?
- Is there a fast exponential or sub-exponential algorithm for solving this problem?

Thank you for your attention!

Please contact me at *leonid@sfu.ca*.

Acknowledgments

João Meidanis

Cedric Chauve

Pedro Feijão

Sean La

NSERC
CRSNG

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

GenomeCanada

ALFRED P. SLOAN FOUNDATION
1934

References

- J. P. Pereira Zanetti, P. Biller, J. Meidanis. Median Approximations for Genomes Modeled as Matrices. Bulletin of Math Biology, 78(4), 2016.
- L. Chindelevitch and J. Meidanis. On the Rank-Distance Median of 3 Permutations. Proc. 15th RECOMB Comparative Genomics Satellite Workshop. LNCS, vol. 10562, pp. 256-276. Springer, Heidelberg (2017). Journal version in BMC Bioinformatics.
- L. Chindelevitch, S. La and J. Meidanis. A cubic algorithm for the generalized rank median of three genomes. Proc. 16th RECOMB Comparative Genomics Satellite Workshop. LNCS, vol. 11183, pp. 3-27 (2018). Journal version in BMC Algorithms for Molecular Biology.
- R. Sarkis, S. La, P. Feijao, L. Chindelevitch, H. Hatami. Computing the Cayley median for permutations and the rank median for genomes is NP-hard. [Submitted to WABI 2019]