QUASI-BEST MATCH GRAPHS 38th TBI WINTERSEMINAR IN BLED

Annachiara Korchmaros joint work with David Schaller, Marc Hellmuth, Peter F. Stadler

Bioinformatics Group, University of Leipzig

February 16, 2023





BMG REFERENCES

- species trees". In: Journal of mathematical biology 80.5.
- best match graphs". In: *Theoretical Computer Science* 809.
- 304.
- Mathematical Biology 80.3.
- Mathematical Biology 82.
- binary-explainable best match graphs". In: *Algorithms* 14.4.

- assignments in best match graphs". In: *Journal of mathematical biology* 82.3.
- graphs". In: Theoretical Computer Science 865.
- for Molecular Biology 15.1.

Geiß, Manuela, Edgar Chávez, et al. (2019). "Best match graphs". In: Journal of mathematical biology 78.

Geiß, Manuela, Marcos E González Laffitte, et al. (2020). "Best match graphs and reconciliation of gene trees with

Hellmuth, Marc, Manuela Geiß, and Peter F Stadler (2020). "Complexity of modification problems for reciprocal

Korchmaros, Annachiara (2021). "The structure of 2-colored best match graphs". In: Discrete Applied Mathematics

Manuela, Geiß, Peter F Stadler, and Marc Hellmuth (2020). "Reciprocal best match graphs". In: Journal of

Schaller, David, Manuela Geiß, Edgar Chávez, et al. (2021). "Corrigendum to "Best match graphs"". In: Journal of

Schaller, David, Manuela Geiß, Marc Hellmuth, et al. (2021a). "Arc-completion of 2-colored best match graphs to

— (2021b). "Heuristic algorithms for best match graph editing". In: Algorithms for Molecular Biology 16.1.

— (2021c). "Least resolved trees for two-colored best match graphs". In: *arXiv preprint arXiv:*2101.07000.

Schaller, David, Manuela Geiß, Peter F Stadler, et al. (2021). "Complete characterization of incorrect orthology

Schaller, David, Marc Hellmuth, and Peter F Stadler (2021). "A simpler linear-time algorithm for the common refinement of rooted phylogenetic trees on a common leaf set". In: Algorithms for Molecular Biology 16.1.

Schaller, David, Peter F Stadler, and Marc Hellmuth (2021). "Complexity of modification problems for best match

Stadler, Peter F et al. (2020). "From pairs of most similar sequences to phylogenetic best matches". In: Algorithms

BMG REFERENCES

- **Geiß**, Manuela, Edgar Chávez, et al. (2019). "Best match graphs". In: *Journal of mathematical biology* 78.
- Geiß, Manuela, Marcos E González Laffitte, et al. (2020). "Best match graphs and reconciliation of gene trees with species trees". In: *Journal of mathematical biology* 80.5.
- Hellmuth, Marc, Manuela Geiß, and Peter F Stadler (2020). "Complexity of modification problems for reciprocal best match graphs". In: *Theoretical Computer Science* 809.
- Korchmaros, Annachiara (2021). "The structure of 2-colored best match graphs". In: *Discrete Applied Mathematics* 304.
- Manuela, Geiß, Peter F Stadler, and Marc Hellmuth (2020). "Reciprocal best match graphs". In: *Journal of Mathematical Biology* 80.3.
- Schaller, David, Manuela Geiß, Edgar Chávez, et al. (2021). "Corrigendum to "Best match graphs"". In: Journal of Mathematical Biology 82.
- Schaller, David, Manuela Geiß, Marc Hellmuth, et al. (2021a). "Arc-completion of 2-colored best match graphs to binary-explainable best match graphs". In: *Algorithms* 14.4.
- (2021b). "Heuristic algorithms for best match graph editing". In: *Algorithms for Molecular Biology* 16.1.
- (2021c). "Least resolved trees for two-colored best match graphs". In: *arXiv preprint arXiv*:2101.07000.
- Schaller, David, Manuela Geiß, Peter F Stadler, et al. (2021). "Complete characterization of incorrect orthology assignments in best match graphs". In: *Journal of mathematical biology* 82.3.
- Schaller, David, Marc Hellmuth, and Peter F Stadler (2021). "A simpler linear-time algorithm for the common refinement of rooted phylogenetic trees on a common leaf set". In: *Algorithms for Molecular Biology* 16.1.
- Schaller, David, Peter F Stadler, and Marc Hellmuth (2021). "Complexity of modification problems for best match graphs". In: *Theoretical Computer Science* 865.
- Stadler, Peter F et al. (2020). "From pairs of most similar sequences to phylogenetic best matches". In: *Algorithms for Molecular Biology* 15.1.



(T, σ) rooted gene tree, leaf coloring σ on leaf set L(T)



v, t, w duplications/speciations



BMG(T, σ)



qBMG((T, a, u))

Best Match Graphs - Definition

- (*T*, σ) rooted gene tree, leaf coloring σ on leaf set *L*(*T*)
- ▶ $y \in L(T)$ is a **best match** of $x \in L(T)$ if
 - 1. $\sigma(x) \neq \sigma(y)$ and
 - 2. $lca(x, y) \leq lca(x, z)$ for all $z \in L(T)$ with $\sigma(z) = \sigma(y)$



v, *t*, *w* duplications/speciations



BMG(T, σ)



qPBMG((T, a, u))

- (T, σ) rooted gene tree, leaf coloring σ on leaf set L(T)
- ▶ $y \in L(T)$ is a **best match** of $x \in L(T)$ if
 - 1. $\sigma(x) \neq \sigma(y)$ and
 - 2. $lca(x, y) \leq lca(x, z)$ for all $z \in L(T)$ with $\sigma(z) = \sigma(y)$



v, *t*, *w* duplications/speciations



BMG(T, σ)



 $q \mathbf{B} \mathbf{M} (\mathcal{T}, \mathcal{T}, \mathcal{O}, \mathcal{U})$

- (T, σ) rooted gene tree, leaf coloring σ on leaf set L(T)
- ▶ $y \in L(T)$ is a **best match** of $x \in L(T)$ if
 - 1. $\sigma(x) \neq \sigma(y)$ and
 - 2. $lca(x, y) \leq lca(x, z)$ for all $z \in L(T)$ with $\sigma(z) = \sigma(y)$



v, *t*, *w* duplications/speciations

(*G*, σ) is **BMG**(*T*, σ) if vertices=leaves colored by σ and $x \to y$ iff *y* is a best match of *x* on (*T*, σ)



- (T, σ) rooted gene tree, leaf coloring σ on leaf set L(T)
- ▶ $y \in L(T)$ is a **best match** of $x \in L(T)$ if
 - 1. $\sigma(x) \neq \sigma(y)$ and
- 2. $lca(x, y) \leq lca(x, z)$ for all $z \in L(T)$ with $\sigma(z) = \sigma(y)$
- (*G*, σ) is **BMG**(*T*, σ) if vertices=leaves colored by σ and $x \to y$ iff *y* is a best match of *x* on (*T*, σ)



v, *t*, *w* duplications/speciations



- (T, σ) rooted gene tree, leaf coloring σ on leaf set L(T)
- ▶ $y \in L(T)$ is a **best match** of $x \in L(T)$ if
 - 1. $\sigma(x) \neq \sigma(y)$ and
- 2. $lca(x, y) \leq lca(x, z)$ for all $z \in L(T)$ with $\sigma(z) = \sigma(y)$
- (*G*, σ) is **BMG**(*T*, σ) if vertices=leaves colored by σ and $x \to y$ iff y is a best match of x on (T, σ)



v, *t*, *w* duplications/speciations



BMGs are

properly colored, no loops, no multiple edges

BMGs are

- properly colored, no loops, no multiple edges
- color-sink-free (every vertex has a best match for every species)

dges natch for every species)

BMGs are

- properly colored, no loops, no multiple edges
- color-sink-free (every vertex has a best match for every species)
- not a hereditary class (induced subgraphs are not necessary BMGs)

dges natch for every species) s are not necessary BMGs)

BMGs are

- properly colored, no loops, no multiple edges
- color-sink-free (every vertex has a best match for every species)
- not a hereditary class (induced subgraphs are not necessary BMGs)

F is induced subgraph of *G* if $x, y \in V(F)$ and $(x, y) \in E(G) \Rightarrow (x, y) \in E(F)$



 $BMG(T, \sigma)$

dges natch for every species) s are not necessary BMGs)



BMG(T, σ)[{x', y', z'}]

BMGs are

- properly colored, no loops, no multiple edges
- color-sink-free (every vertex has a best match for every species)
- not a hereditary class (induced subgraphs are not necessary BMGs)

F is induced subgraph of *G* if $x, y \in V(F)$ and $(x, y) \in E(G) \Rightarrow (x, y) \in E(F)$

Bio connection: Can we recognize BMGs?



dges natch for every species) s are not necessary BMGs)

► ab|c is a (rooted) triple if lca(a, b) is descendent in T of lca(a, c) = lca(b, c)



▶ ab|c is a (rooted) **triple** if lca(a, b) is descendent in *T* of lca(a, c) = lca(b, c)

•
$$(G, \sigma)$$
 digraph, vertex-colored by σ
• $\mathcal{R}(G, \sigma) \coloneqq \{ab | b' \colon \sigma(a) \neq \sigma(b) = \sigma(b')\}$

 $xy \mid y' \in \mathcal{R}(G, \sigma)$





'), $ab \in E(G)$, and $ab' \notin E(G)$ informative triples



T





•
$$(G, \sigma)$$
 digraph, vertex-colored by σ
• $\mathcal{R}(G, \sigma) \coloneqq \{ab|b' \colon \sigma(a) \neq \sigma(b) = \sigma(b')$
• $\mathcal{F}(G, \sigma) \coloneqq \{ab|b' \colon \sigma(a) \neq \sigma(b) = \sigma(b')\}$

 $xy \mid y' \in \mathcal{R}(G, \sigma)$





'), $ab \in E(G)$, and $ab' \notin E(G)$ informative triples '), $b \neq b'$, and $ab, ab' \in E(G)$ forbidden triples

. . . .

T

lca(a,b) des









• $\Re(G, \sigma) \coloneqq \{ab|b': \sigma(a) \neq \sigma(b) = \sigma(b'), ab \in E(G), \text{ and } ab' \notin E(G)\}$ informative triples • $\Re(G, \sigma) \coloneqq \{ab|b': \sigma(a) \neq \sigma(b) = \sigma(b'), b \neq b', \text{ and } ab, ab' \in E(G)\}$ forbidden triples



(*ii*) there exists (T, σ) displaying

• <u>Problem</u>: different trees associated to $BMG(G, \sigma)$, how to choose the most parsimonious???

- \blacktriangleright $T_{L'}$ is a **restriction** of *T* to a subset *L'* of leaves of *T*



Problem: different trees associated to $BMG(G, \sigma)$, how to choose the most parsimonious???

- \blacktriangleright $T_{L'}$ is a **restriction** of *T* to a subset *L'* of leaves of *T*



Problem: different trees associated to $BMG(G, \sigma)$, how to choose the most parsimonious???

BEST MATCH GRAPHS - UNIQUE LRT

- \blacktriangleright $T_{L'}$ is a **restriction** of *T* to a subset *L'* of leaves of *T*



Problem: different trees associated to $BMG(G, \sigma)$, how to choose the most parsimonious???

(*T*, σ) is **LRT** if there is NO $T' = T_{L'}(+ \text{ inner edge contractions})$ st $BMG(T, \sigma) = BMG(T', \sigma)$

BEST MATCH GRAPHS - UNIQUE LRT

- \blacktriangleright $T_{L'}$ is a **restriction** of *T* to a subset *L'* of leaves of *T*



Theorem

Every BMG has a unique LRT¹

Problem: different trees associated to $BMG(G, \sigma)$, how to choose the most parsimonious???

(*T*, σ) is **LRT** if there is NO $T' = T_{L'}(+ \text{ inner edge contractions})$ st $BMG(T, \sigma) = BMG(T', \sigma)$

BEST MATCH GRAPHS - UNIQUE LRT

- \blacktriangleright $T_{L'}$ is a **restriction** of *T* to a subset *L'* of leaves of *T*



 (T, σ) is **LRT** if there is NO $T' = T_{L'}(+$ inner edge contractions) st $BMG(T, \sigma) = BMG(T', \sigma)$

Theorem

Every BMG has a unique LRT¹ **b** Build the LRT in polynomial time with MTT algorithm²

Problem: different trees associated to $BMG(G, \sigma)$, how to choose the most parsimonious???

¹Manuela Geiß, Edgar Chávez, et al. (2019). "Best match graphs". In: *Journal of mathematical biology* 78.

²Annachiara Korchmaros et al. (2023). "Quasi-best match graphs". In: *Discrete Applied Mathematics* 331, pp. 104–125.

- \blacktriangleright $T_{L'}$ is a **restriction** of *T* to a subset *L'* of leaves of *T*



 (T, σ) is **LRT** if there is NO $T' = T_{L'}(+$ inner edge contractions) st $BMG(T, \sigma) = BMG(T', \sigma)$

Theorem

Every BMG has a unique LRT¹

Problem: different trees associated to $BMG(G, \sigma)$, how to choose the most parsimonious???

▶ Build the LRT in polynomial time with MTT algorithm² \Rightarrow recognize a BMG in polynomial time

¹Manuela Geiß, Edgar Chávez, et al. (2019). "Best match graphs". In: *Journal of mathematical biology* 78.

²Annachiara Korchmaros et al. (2023). "Quasi-best match graphs". In: *Discrete Applied Mathematics* 331, pp. 104–125.

How do we relate far-away genes?

► How do we relate far-away genes? Limit detection of best matches in (T, σ)



- ► How do we relate far-away genes? Limit detection of best matches in (T, σ) ▶ $y \in L(T)$ is a **quasi-best match** of $x \in L(T)$ if
 - 1. *y* is a best match of *x*, and
 - 2. $lca(x, y) \preceq u(x, \sigma(y)), u: L(T) \times \sigma(L(T)) \rightarrow V(T)$ vertex-set of *T*





BMG(T, σ)

- \blacktriangleright How do we relate far-away genes? Limit detection of best matches in (T, σ) ▶ $y \in L(T)$ is a **quasi-best match** of $x \in L(T)$ if
 - 1. *y* is a best match of *x*, and
 - 2. $lca(x, y) \preceq u(x, \sigma(y)), u: L(T) \times \sigma(L(T)) \rightarrow V(T)$ vertex-set of *T*



 $u(z, \sigma(x)) = z, u(q, \sigma(q)) = q, u(q, s) = \rho \text{ and } s \neq \sigma(q)$

(*G*, σ) is **qBMG**(*T*, σ , *u*) if vertices=leaves colored by σ and $x \to y$ iff *y* is **quasi-best** match of *x*

BMGs vs. QBMGs

► Is $BMG(T, \sigma)$ a $qBMG(T, \sigma, u)$?

► Is $BMG(T, \sigma)$ a $qBMG(T, \sigma, u^{\rho})$? Yes, $u^{\rho}(x, x)$

$$s) := \begin{cases} x & s = \sigma(x) \\ \rho & \text{otherwise} \end{cases}$$

► Is $BMG(T, \sigma)$ a $qBMG(T, \sigma, u^{\rho})$? Yes, $u^{\rho}(x, x)$

▶ When $qBMG(T, \sigma, u)$ is a $BMG(T, \sigma)$?

$$s) := \begin{cases} x & s = \sigma(x) \\ \rho & \text{otherwise} \end{cases}$$

BMGS VS. QBMGS - LIKENESS

- ► Is $BMG(T, \sigma)$ a $qBMG(T, \sigma, u^{\rho})$? Yes, $u^{\rho}(x, x)$
- ▶ When $qBMG(T, \sigma, u)$ is a $BMG(T, \sigma)$?
- orem
 - (G, σ) is a BMG iff (G, σ) is a color-sink-free qBMG.

$$s) := \begin{cases} x & s = \sigma(x) \\ \rho & \text{otherwise} \end{cases}$$

- ► Is $BMG(T, \sigma)$ a $qBMG(T, \sigma, u^{\rho})$? Yes, $u^{\rho}(x, t)$
- When $qBMG(T, \sigma, u)$ is a $BMG(T, \sigma)$?

orem

• (G, σ) is a BMG iff (G, σ) is a color-sink-free qBMG.

 \uparrow

$$(s,s) := \begin{cases} x & s = \sigma(x) \\ \rho & \text{otherwise} \end{cases}$$

• (G, σ) is a qBMG iff there is (T, σ, u) displaying all triples in $\mathcal{R}(G, \sigma)$ but none in $\mathcal{F}(G, \sigma)$

BMGS VS. QBMGS - LIKENESS

- ► Is $BMG(T, \sigma)$ a $qBMG(T, \sigma, u^{\rho})$? Yes, $u^{\rho}(x, \tau)$
- When $qBMG(T, \sigma, u)$ is a $BMG(T, \sigma)$?

orem

• (G, σ) is a BMG iff (G, σ) is a color-sink-free qBMG.

 \uparrow

 $\mathcal{R}(G,\sigma), \mathcal{F}(G,\sigma) \xrightarrow{\text{input to}} \text{MTT Algorithr}$

$$(s,s) := \begin{cases} x & s = \sigma(x) \\ \rho & \text{otherwise} \end{cases}$$

• (G, σ) is a qBMG iff there is (T, σ, u) displaying all triples in $\mathcal{R}(G, \sigma)$ but none in $\mathcal{F}(G, \sigma)$

$$rac{ ext{polynomial time}}{ op} (T,\sigma)$$

BMGS VS. QBMGS - LIKENESS

- ► Is $BMG(T, \sigma)$ a $qBMG(T, \sigma, u^{\rho})$? Yes, $u^{\rho}(x, \tau)$
- When $qBMG(T, \sigma, u)$ is a $BMG(T, \sigma)$?

orem

• (G, σ) is a BMG iff (G, σ) is a color-sink-free qBMG.

 \uparrow

 $\mathcal{R}(G,\sigma), \mathcal{F}(G,\sigma) \xrightarrow{\text{input to}} \text{MTT Algorithm}$

$$(s,s) := \begin{cases} x & s = \sigma(x) \\ \rho & \text{otherwise} \end{cases}$$

• (G, σ) is a qBMG iff there is (T, σ, u) displaying all triples in $\mathcal{R}(G, \sigma)$ but none in $\mathcal{F}(G, \sigma)$

$$\begin{array}{c} m \xrightarrow{\text{polynomial time}} (T, \sigma) \xrightarrow{u^*} (T, \sigma, u^*) \\ \\ u^*(x, s) := \begin{cases} x & x \text{ is sink wrt s, or } s = \sigma(x) \\ \rho & \text{otherwise} \end{cases} \end{array}$$

Property	BMG
hereditary class	no, color-sink free
disjoint union	yes, if partition set have same colours
unique LRT	yes ²
binary explainable iff finite sets of forbidden graphs	hourglass-free ³



¹Annachiara Korchmaros et al. (2023). "Quasi-best match graphs". In: *Discrete Applied Mathematics* 331, pp. 104–125.

²Manuela Geiß, Edgar Chávez, et al. (2019). "Best match graphs". In: *Journal of mathematical biology* 78.

³David Schaller, Manuela Geiß, Peter F Stadler, et al. (2021). "Complete characterization of incorrect orthology assignments in best match graphs". In: *Journal of mathematical biology* 82.3.

Property	BMG	qBMG ¹
hereditary class	no, color-sink free	yes
disjoint union	yes, if partition sets have same colours	yes
unique LRT	2 yes	no
binary explainable iff finite sets of forbidden graphs	3 hourglass-free	no

¹Annachiara Korchmaros et al. (2023). "Quasi-best match graphs". In: *Discrete Applied Mathematics* 331, pp. 104–125.

²Manuela Geiß, Edgar Chávez, et al. (2019). "Best match graphs". In: *Journal of mathematical biology* 78.

³David Schaller, Manuela Geiß, Peter F Stadler, et al. (2021). "Complete characterization of incorrect orthology assignments in best match graphs". In: *Journal of mathematical biology* 82.3.

Property	BMG
hereditary class	no, color-sink free
disjoint union	yes, if partition set have same colours
unique LRT	yes ²
binary explainable iff finite sets of forbidden graphs	3 hourglass-free



¹Annachiara Korchmaros et al. (2023). "Quasi-best match graphs". In: *Discrete Applied Mathematics* 331, pp. 104–125.

²Manuela Geiß, Edgar Chávez, et al. (2019). "Best match graphs". In: *Journal of mathematical biology* 78.

³David Schaller, Manuela Geiß, Peter F Stadler, et al. (2021). "Complete characterization of incorrect orthology assignments in best match graphs". In: *Journal of mathematical biology* 82.3.

Property	BMG
hereditary class	no, color-sink free
disjoint union	yes, if partition sets have same colours
unique LRT	yes ²
binary explainable iff finite sets of forbidden graphs	3 hourglass-free

subclass of binary trees



¹Annachiara Korchmaros et al. (2023). "Quasi-best match graphs". In: *Discrete Applied Mathematics* 331, pp. 104–125.

²Manuela Geiß, Edgar Chávez, et al. (2019). "Best match graphs". In: *Journal of mathematical biology* 78.

³David Schaller, Manuela Geiß, Peter F Stadler, et al. (2021). "Complete characterization of incorrect orthology assignments in best match graphs". In: *Journal of mathematical biology* 82.3.

Outlook

1. qBMGs do NOT have a unique LRT \rightarrow studying different LRTs of the same qBMG(*G*, σ)

1. qBMGs do NOT have a unique LRT \rightarrow studying different LRTs of the same qBMG(*G*, σ) • Do they have a common core? (they all display $\mathcal{R}(G, \sigma)$).

1. qBMGs do NOT have a unique LRT \rightarrow studying different LRTs of the same qBMG(*G*, σ) • Do they have a common core? (they all display $\mathcal{R}(G, \sigma)$). • Are LRTs of maximal induced BMGs displayed by all LRT of $qBMG(G, \sigma)$?

- 1. qBMGs do NOT have a unique LRT \rightarrow studying different LRTs of the same qBMG(G, σ) • Do they have a common core? (they all display $\mathcal{R}(G, \sigma)$). • Are LRTs of maximal induced BMGs displayed by all LRT of $qBMG(G, \sigma)$? • Is there an efficient algorithm to find all maximal induced BMGs in qBMG(G, σ)?

- Do they have a common core? (they all display $\mathcal{R}(G, \sigma)$). • Are LRTs of maximal induced BMGs displayed by all LRT of $qBMG(G, \sigma)$? • Is there an efficient algorithm to find all maximal induced BMGs in $qBMG(G, \sigma)$?

- 1. qBMGs do NOT have a unique LRT \rightarrow studying different LRTs of the same qBMG(*G*, σ) 2. Study qBMGs where u(x,s) = u(x), i.e. detection limit is color-independent.

• Do they have a common core? (they all display $\mathcal{R}(G, \sigma)$). • Are LRTs of maximal induced BMGs displayed by all LRT of $qBMG(G, \sigma)$? • Is there an efficient algorithm to find all maximal induced BMGs in $qBMG(G, \sigma)$?

1. qBMGs do NOT have a unique LRT \rightarrow studying different LRTs of the same qBMG(*G*, σ) 2. Study qBMGs where u(x, s) = u(x), i.e. detection limit is color-independent.



more details on qBMGs

1. qBMGs do NOT have a unique LRT \rightarrow studying different LRTs of the same qBMG(G, σ) • Do they have a common core? (they all display $\mathcal{R}(G, \sigma)$). • Are LRTs of maximal induced BMGs displayed by all LRT of $qBMG(G, \sigma)$? • Is there an efficient algorithm to find all maximal induced BMGs in $qBMG(G, \sigma)$?

2. Study qBMGs where u(x, s) = u(x), i.e. detection limit is color-independent.



more details on qBMGs

THANK YOU