Experiments

Partitioning into Colorful Components by Minimum Edge Deletions

<u>Sharon Bruckner</u>¹ Falk Hüffner² Christian Komusiewicz² Rolf Niedermeier² Sven Thiel³ Johannes Uhlmann²

¹Institut für Mathematik, Freie Universität Berlin

²Institut für Softwaretechnik und Theoretische Informatik, TU Berlin

³Institut für Informatik, Friedrich-Schiller-Universität Jena

3 July 2012



Experiments

Multiple Sequence Alignment



Experiments

Multiple Sequence Alignment



Experiments

Multiple Sequence Alignment





Experiments

Multiple Sequence Alignment



Idea

Use alignment graph constructed by local alignment to reconstruct global alignment.

search School

Experiments

Multiple Sequence Alignment



Idea

Use alignment graph constructed by local alignment to reconstruct global alignment.

search School

Experiments

Multiple Sequence Alignment



Idea

Use alignment graph constructed by local alignment to reconstruct global alignment.

search School ...onal Biology and Scientific Computing

Experiments

Colorful Components

Part of a Multiple Sequence Alignment pipeline suggested by Corel, Pitschi & Morgenstern (Bioinformatics 2010).



Colorful Components

Part of a Multiple Sequence Alignment pipeline suggested by Corel, Pitschi & Morgenstern (Bioinformatics 2010).

COLORFUL COMPONENTS

Instance: An undirected graph G = (V, E) and a coloring of the vertices $\chi : V \to \{1, ..., c\}$.

Task: Delete a minimum number of edges such that all connected components are *colorful*, that is, they do not contain two vertices of the same color.



Other application: Wikipedia interlanguage links

	Labyrinthulomycetes - Wikipedia, the	free enc	ycloped	ia – Icew	easel	×
<u>F</u> ile <u>E</u> dit <u>V</u> iew Hi <u>s</u> t	ory <u>B</u> ookmarks <u>T</u> ools <u>H</u> elp					
W Labyrinthulomycete	s - Wikipedi 💠					~
W en wikipedia.	org/wiki/Labyrinthulomycetes			☆ ~ @	Google	Q 🏠
-	<i>• • • •</i>				Log in / cre	eate account
Bern O.					• • • • • • •	
1 QC W	Article Talk	Read	Edit Vi	ew history	Search	Q
1 44 7						
a l	Labyrinthulomycetes					-
WIKIPEDIA The Free Encyclopedia	From Wikipedia, the free encyclopedia					
	The Laburinthulem restor (ICDN) or Laburinthu	la a ^[1]				
Main page	(ICZN), or Slime nets are a class of protists that r	roduce			Slime nets	
Contents	a network of filaments or tubes. ^[2] which serve as tr	acks		herd /		
Featured content	for the cells to glide along and absorb nutrients for	them.				
Current events	There are two main groups, the labyrinthulids and					
Random article	thraustochytrids. They are mostly marine, common	ılv			1 De	
Donate to Wikipedia	found as parasites on alga and seagrass or as	<i>.</i>			N.S.	
Interaction	decomposers on dead plant material. They also inc	lude			- Charles - Char	
. The officers	some parasites of marine invertebrates.					
► IOODOX	Although they are outside the cells, the filaments a	re				
 Print/export 	Although they are outside the cells, the filaments a surrounded by a membrane. They are formed and	re				
Print/export Languages	Although they are outside the cells, the filaments a surrounded by a membrane. They are formed and connected with the cytoplasm by a unique organell	re e called		×3000 10	3.0060	Una
Print/export Languages Česky	Although they are outside the cells, the filaments a surrounded by a membrane. They are formed and connected with the cytoplasm by a unique organell a sagenogen or bothrosome. The cells are uninucle	re e called ate	The cel	×3000 Io	etwork of filaments Aplanoo	una chytrium sp.
Front/export Languages Česky Deutsch	Although they are outside the cells, the filaments a surrounded by a membrane. They are formed and connected with the cytoplasm by a unique organell a sagenogen or bothrosome. The cells are uninucle and typically ovoid, and move back and forth along	re e called ate the	The cel	with the n	etwork of filaments Aplanoo entific classification	oss chytrium sp.
Iodiox Print/export Languages Česky Deutsch Español	Although they are outside the cells, the filaments a surrounded by a membrane. They are formed and connected with the cytoplasm by a unique organell a sagenogen or bothrosome. The cells are uninucle and typically lovidi, and move back and forth along amorphous network at speeds varying from 5-150 pringta. Agenage the labeighthulide the cells are area lower to be an another the set of the set of the set of the set of the set of the set of the set of the set of the set of the set of the set of the set of the set of the set of the set of the set of the set of set of s	re e called ate the μm per	The cell	with the n Science: S	etwork of filaments Aplanoo entific classification ota	om chytrium sp.
▶ riolbox ▶ Print/export ► Languages Česky Deutsch Español 日本語	Although they are outside the cells, the filaments a surrounded by a membrane. They are formed and connected with the cytoplasm by a unique organell a sagenogen or bothrosome. The cells are uninucle and typically ovoid, and move back and forth along amorphous network at speeds varying from 5-150 minute. Among the labyrinthulids the cells are encl. within the tube, and means the theoretic threads the cells are encled.	re e called eate the μm per osed	The cell Domair Kingdor	with the n Scin : Eukary m: Chrom	etwork of filaments Aplanoo entific classification ota alveolata	oss chytrium sp.
 Print/export Languages Česky Deutsch Español 日本語 Македонски 	Although they are outside the cells, the filaments a surrounded by a membrane. They are formed and connected with the cytoplasm by a unique organell a sagenogen or bothrosome. The cells are uninucle and typically oxid, and move back and forth along amorphous network at speeds varying from 5-150 minute. Among the labyrinthulids the cells are encl within the tubes, and among the thraustochytrids ti attached to their sides.	re e called the μm per osed hey are	The cell Domair Kingdor Phylum	with the n Scient: Eukary m: Chrom I: Hetero	etwork of filaments <i>Aplanoo</i> entific classification ota alveolata kontophyta	aan

Other application: Wikipedia interlanguage links

	Labyrinthulomycetes – Wikipedia	, the free en	yclo	pedia – Icew	easel		Ţ
_ile <u>E</u> dit <u>V</u> iew ⊦	li <u>s</u> tory <u>B</u> ookmarks <u>T</u> ools <u>H</u> elp						
₩ Labyrinthulomyce	ites - Wikipedi 💠						
🚱 🛛 W en.wikipe	dia.org/wiki/Labyrinthulomycetes			☆ ∨ @	🛃 V Google	Q	6
Ben a					🕹 Log ir) / create accou	int
Ω W	Article Talk	Read	Edit	View history	Search	C	Q
	Labyrinthulomycetes						
WIKIPEDIA The Free Encyclopedia	From Wikipedia, the free encyclopedia						
Main page Contents Featured content Current events Random article Donate to Wikipedi	The Labyrinthulomycetes (ICBN) or Labyri (ICZN), or Slime nets are a class of protists is a network of filaments or tubes. ¹²¹ which serve for the cells to glide along and absorb nutrien threa are two main groups, the labyrinthulds thraustochrids. They are mostly marine, cor found as parasites on alga and seagrass or ar	nthulea ^[1] that produce a as tracks ts for them. and mmonly s			Slime nets		Ī
Interaction	decomposers on dead plant material. They al some parasites of marine invertebrates.	so include			PR		
Toolbox	Although they are outside the cells, the filame	ents are				ϕ	
Print/export Languages	Netzschleimpilze		.	×3000 10		040 000	
Deutsch	Die Netzschleimpilze oder Schleimnetze		The	Scie	etwork or nlaments Ap entific classificati	on	I
Espandi 日本語	Stramenopilen und sind somit näher mit Braunalger	i, Goldalgen	Kinç	nain: Eukary gdom: Chrom	ota alveolata		
Македонски Norsk (bokmål)	attached to their sides.		Phy Clas	lum: Hetero ss: Labyri	kontophyta inthulomycetes D	CK 2001 OF	

ck Research School Jational Biology tific Computing

onal

Partitioning into Colorful Components by Minimum Edge Deletions

Experiments

Wikipedia interlanguage link graph example



Experiments

Complexity of Colorful Components

• COLORFUL COMPONENTS with two colors can be solved in $O(\sqrt{n}m)$ time by matching techniques.



Experiments

Complexity of Colorful Components

- COLORFUL COMPONENTS with two colors can be solved in $O(\sqrt{n}m)$ time by matching techniques.
- COLORFUL COMPONENTS is NP-hard already with three colors.



Complexity of Colorful Components

- COLORFUL COMPONENTS with two colors can be solved in $O(\sqrt{n}m)$ time by matching techniques.
- COLORFUL COMPONENTS is NP-hard already with three colors.
- COLORFUL COMPONENTS is NP-hard on trees.



Complexity of Colorful Components

- COLORFUL COMPONENTS with two colors can be solved in $O(\sqrt{n}m)$ time by matching techniques.
- COLORFUL COMPONENTS is NP-hard already with three colors.
- COLORFUL COMPONENTS is NP-hard on trees.
- COLORFUL COMPONENTS on trees with *c* colors can be solved in $2^c \cdot n^{O(1)}$ time.



Experiments

Fixed-parameter algorithm

Observation

COLORFUL COMPONENTS can be seen as the problem of destroying by edge deletions all bad paths, that is, simple paths between equally colored vertices.



Experiments

Fixed-parameter algorithm

Observation

COLORFUL COMPONENTS can be seen as the problem of destroying by edge deletions all bad paths, that is, simple paths between equally colored vertices.

Observation

Unless the graph is already colorful, we can always find a bad path with at most *c* edges, where *c* is the number of colors.



Experiments

Fixed-parameter algorithm

Observation

COLORFUL COMPONENTS can be seen as the problem of destroying by edge deletions all bad paths, that is, simple paths between equally colored vertices.

Observation

Unless the graph is already colorful, we can always find a bad path with at most *c* edges, where *c* is the number of colors.

Theorem

COLORFUL COMPONENTS can be solved in $O(c^k \cdot m)$ time, where k is the number of edge deletions.

rearch School rial Biology of Computing

Experiments

Improved fixed-parameter algorithm

Theorem

COLORFUL COMPONENTS can be solved in $O((c-1)^k \cdot m)$ time, where k is the number of edge deletions.



Experiments

Improved fixed-parameter algorithm

Theorem

COLORFUL COMPONENTS can be solved in $O((c-1)^k \cdot m)$ time, where k is the number of edge deletions.

Proof.

If there is a degree-3 or higher vertex v, find a bad path with at most (c - 1) edges by BFS from v. Otherwise, the instance is easy.



Introduction

Complexity and Algorithms

Experiments

Limits of fixed-parameter algorithms

Question

How much further can we improve this algorithm?



Experiments

Limits of fixed-parameter algorithms

Question

How much further can we improve this algorithm?

Exponential Time Hypothesis (ETH)

For all $x \ge 3$, x-SAT, which asks whether a boolean input formula in conjunctive normal form with *n* variables and *m* clauses and at most *x* variables per clause is satisfiable, cannot be solved within a running time of $2^{o(n)}$ or $2^{o(m)}$.



Limits of fixed-parameter algorithms

Question

How much further can we improve this algorithm?

Exponential Time Hypothesis (ETH)

For all $x \ge 3$, x-SAT, which asks whether a boolean input formula in conjunctive normal form with *n* variables and *m* clauses and at most *x* variables per clause is satisfiable, cannot be solved within a running time of $2^{o(n)}$ or $2^{o(m)}$.

Theorem

COLORFUL COMPONENTS with three colors cannot be solved in $2^{o(k)} \cdot n^{O(1)}$ unless the ETH is false.

search School onal Biology and Scientific Computing

Experiments

Weighted version

Problem

If we know that two vertices must belong to the same connected component, we want to be able to simplify the instance by merging them.

Idea

Introduce color sets per vertex and edge weights.





Experiments

Uses of the merge operation

Edge branching

Can branch into two cases: delete an edge, or merge its endpoints.



Uses of the merge operation

Edge branching

Can branch into two cases: delete an edge, or merge its endpoints.

Data reduction

Let $V' \subseteq V$ be a colorful subgraph. If the cut between V' and $V \setminus V'$ is at least as large as the connectivity of V', then merge V' into a single vertex.



11/19

Experiments

Merge-based heuristic

Idea

Repeatedly merge the two vertices "most likely" to be in the same component, while immediately deleting edges connecting vertices with intersecting color sets.



Merge-based heuristic

Idea

Repeatedly merge the two vertices "most likely" to be in the same component, while immediately deleting edges connecting vertices with intersecting color sets.

We always merge the endpoints of the edge that maximizes cut cost minus merge cost.

- *merge cost*: weight of the edges that would need to be deleted when merging
- cut cost.

$$3w(\{u, v\}) + \sum_{w \in V \mid \{\{u, w\}, \{v, w\}\} \subseteq E} \min\{w(\{u, w\}), w(\{v, w\})\}$$





• We generated one COLORFUL COMPONENTS instance for each multiple alignment instance from the BAliBASE 3.0 benchmark. We restricted the experiments to the 135 of instances that have at most 10 colors. CDFTODDOLILIKLGFFEVWLTSDTEIGLFSAMVLLDRAGLSEPKVIGRARELVAEALRVOILRSRAO RIPGFRDLSOHDOVNLLKAGTFEVLMVSDEEMSLFTAVVLVDRSGIENVNSVEALOETLIRALRTLIMKN GFOLLTODDKFTLLKAGLFDALFVTDAEIGLFCAIVLIDRPGLRNLELIEKMYSRLK PGFDKLCOEDKVLLLKTASLEILLVCETRLALFSSLVLLDRPNLRDPAAIEEIRDR FAKRLPFFTGKVSTDDOVAMLKGCCMEVIVLTETEIAMLKAIIVFDRPRIOHIDEIRNIODSLLOSLRYYVMDKRO/ DFAKOLPGFLOLSREDOIALLKTSAIEVMLLNDAEFALLIAISIFDRPNVODOLOVERLOHTYV Max Planck Resear *OLIVEFAKGLPAFTKIPQEDQITLLKACSSEV*MMLDNVEYALLTAIVIFURPGLEKAQLVEAIQ

Data reduction: Largest connected component

- (1) originally
- (2) after data reduction in the COLORFUL COMPONENTS formulation
- (3) after data reduction in the weighted formulation

		(1)			(2)			(3)	
	n	т	С	n	т	с	п	т	С
average	504	921	6.2	407	697	4.7	354	607	5.3
median	149	232	6	46	90	5	42	58	5



Experiments

Branching algorithms: running time

	< 1 s	1 s to 10 min	> 10 min
bad-path branching	61	6	68
merging branching	70	9	56



Experiments

Branching algorithms: running time

	< 1 s	1 s to 10 min	> 10 min
bad-path branching	61	6	68
merging branching	70	9	56

Note

In ongoing research, we are able to solve several more instances to optimality with integer linear programming (ILP) based approaches.



Experiments

Heuristics: relative error

	min.	max.	avg.	med.
min-cut heuristic [1]	0 % (1)	70.0%	29.2 %	27.8%
merging heuristic	0 % (76)	12.7 %	0.6 %	0 %

[1] Corel, Pitschi & Morgenstern (Bioinformatics 2010)



Experiments

Sequence alignment quality

DIALIGN with several methods for solving the COLORFUL COMPONENTS subproblem:

	TC score
min-cut heuristic	53.6%
merge heuristic	55.1 %
exact algorithm	56.6%



S. Bruckner et al. (FU Berlin)

Experiments

Sequence alignment quality

DIALIGN with several methods for solving the COLORFUL COMPONENTS subproblem:

	TC score
min-cut heuristic	53.6 %
merge heuristic	55.1 %
exact algorithm	56.6%

DIALIGN with the min-cut heuristic is about 10 percentage points worse than current state-of-the-art multiple alignment methods. Hence, an improvement of 3 percentage points is a sizable step towards closing the gap between DIALIGN and these methods.



- ILP-based solutions
- Application to network alignment
- Relaxation of the colorfulness constraint



Introduction

Complexity and Algorithms

Experiments

Acknowledgements



Falk Hüffner, Christian Komusiewicz, Rolf Niedermeier, Johannes Uhlmann





