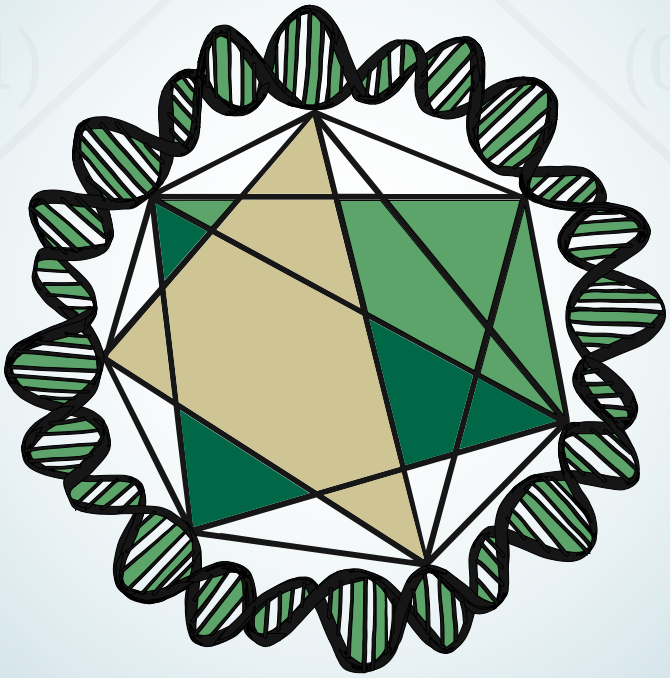
26TH INTERNATIONAL CONFERENCE ON DEVELOPMENTS IN LANGUAGE THEORY

DLT-2022

MAY 9 - 13

TAMPA, FLORIDA, USA



Discrete and Topological Models in Molecular Biology

DTMB-2022

DEPARTMENT OF MATHEMATICS AND STATISTICS

UNIVERSITY of SOUTH FLORIDA

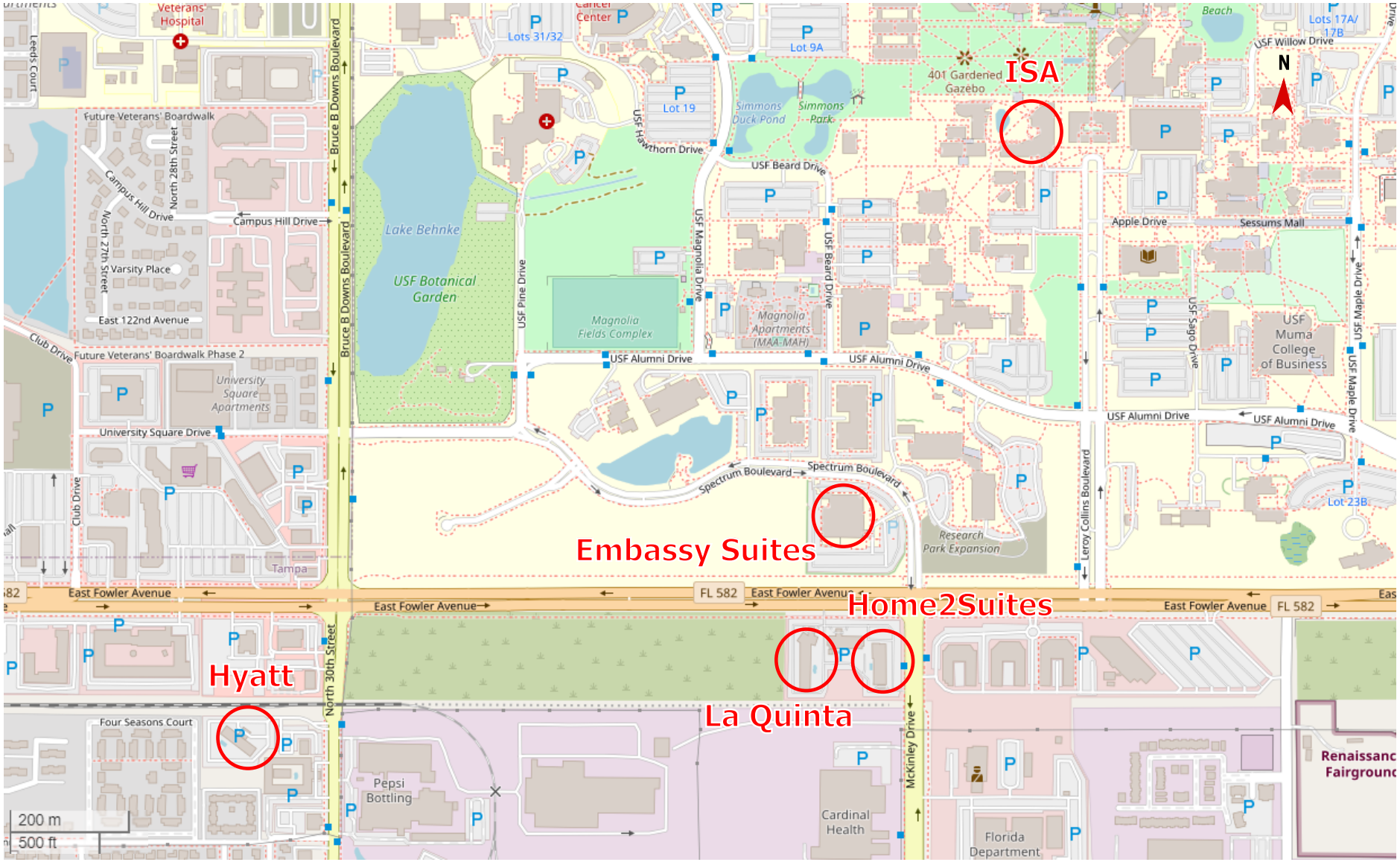
Contents

General Information

Organizers at University of South Florida: Lina Fajardo Gómez, Margherita Maria Ferrari, Nataša Jonoska, Abdulmelik Mohammed, Masahico Saito, Dmytro Savchuk

DLT2022 Program Chairs: Volker Diekert, Mikhail Volkov

Venue and Accommodation



Local Information

Places to Eat (ordered by proximity to ISA building):

- Ichiban Japanese Cuisine & Sushi Bar, 2786 E Fowler Ave, Tampa, FL 33612. 1.2 miles SW.
- Taj Indian Cuisine, 2734 E Fowler Ave, Tampa, FL 33612. 1.3 miles SW.
- World of Beer, 2815 E Fowler Ave, Tampa, FL 33612. 1.3 miles SW.
- Carrabba's Italian Grill, 5515 E Fowler Ave, Temple Terrace, FL 33617. 1.6 miles SE.
- Mr. Dunderbak's Biergarten and Brewery, 14929 Bruce B Downs Blvd, Tampa, FL 33613. 1.7 miles N.
- Acropolis Greek Taverna, 14947 Bruce B Downs Blvd, Tampa, FL 33613. 1.7 miles N.
- Wood Fired (Pizza), 2822 E Bearss Ave, Tampa, FL 33613. 2.1 miles N.
- Petra Restaurant (Middle-Eastern cuisine), 4812 E Busch Blvd, Tampa, FL 33612. 2.3 miles N.
- Thai Ruby, 15319 Amberly Dr, Tampa, FL 33647. 2.5 miles N.
- Koizi Endless Hibachi & Sushi Eatery, 17012 Palm Pointe Dr, Tampa, FL 33647. 5 miles NE.

Places of Interest (ordered by proximity to USF):

Descriptions from Google Maps or attraction's website.

- *Busch Gardens Tampa Bay*, 10165 McKinley Dr, Tampa, FL 33612. Venerable theme park with thrill rides, African animals, live entertainment & attractions. About 2.7 miles S.
- *Lettuce Lake Park*, 6920 E Fletcher Ave, Tampa, FL 33637 ($2 per car). A very nice nature park featuring boardwalks over the lake. A place to see alligators and other wildlife. About 3.4 miles E.
- *ZooTampa at Lowry Park*, 1101 W Sligh Ave, Tampa, FL 33604. Celebrated zoo offering giraffe feeding, educational programs and a water play area. About 8.7 miles SW.
- *Tampa Theatre*, 711 N Franklin St, Tampa, FL 33602. Built in 1926, this ornate movie palace features a Wurlitzer organ plus films, shows & other events. About 12.7 miles S.
- *Riverwalk*, 100 E Madison St, Tampa, FL 33602. A downtown area walking trail by the river with a number of attractions along the line including Amalie Arena, Armature Works, The Florida Aquarium and Sparkman's Wharf. About 12.8 miles S.
- *The Florida Aquarium*, 701 Channelside Dr, Tampa, FL 33602. Aquarium with sharks, penguins, stingray touch tanks, a wild dolphin cruise and a water playground. About 13.3 miles S.

- *Museum of Fine Arts*, 255 Beach Dr NE, St. Petersburg, FL 33701. Exhibits spanning 4,000 years, including African art, European paintings & American photography. About 35 miles SW.
- *The Dali Museum*, 1 Dali Blvd, St. Petersburg, FL 33701. It houses the largest collection of Dalí's works outside Europe. About 35 miles (40 minute drive) SW.
- *Chihuly Collection*, 720 Central Ave, St. Petersburg, FL 33701. The Chihuly Collection is a stunning, permanent collection of world-renowned artist Dale Chihuly's unique artwork. About 35 miles SW.
- *Clearwater Beach*. About 36 miles (50 minute drive) W. Other alternatives: *Sarasota Beach* (more touristy, 73 miles S), and *Sand Key Beach* (less touristy, 38 miles W).
- *John and Mable Ringling Museum of Art*. 5401 Bay Shore Rd, Sarasota, FL 34243. The most celebrated items in the museum are 16th–20th-century European paintings, including a world-renowned collection of Peter Paul Rubens paintings. About 53 miles (1 hour drive) S.
- *River Ventures - The Legends of Adventure*, 498 SE Kings Bay Dr, Crystal River, FL 34429. Swim with mantees, take a tour on luxury pontoons, or on a houseboat mantee tour. About 66 miles (1 hr 15 mins drive) N.

26th International Conference on Developments in Language Theory

DLT2022 Program & Abstracts

DLT Committees

Program Committee

- Volker Diekert (*cochair*), University of Stuttgart, Germany
- Yo-Sub Han, Yonsei University, Republic of Korea
- Artur Jeż, University of Wrocław, Poland
- Jarkko Kari, University of Turku, Finland
- Shankara Narayanan Krishna, Indian Institute of Technology, Bombay, India
- Alexander Okhotin, St. Petersburg State University, Russia
- Joël Ouaknine, Max Planck Institute for Software Systems, Saarbrücken, Germany
- Svetlana Puzynina, St. Petersburg State University, Russia
- Narad Rampersad, University of Winnipeg, Canada
- Helmut Seidl, Technical University of Munich, Germany
- Mikhail Volkov (*cochair*), Ural Federal University, Russia
- Marc Zeitoun, University of Bordeaux, France

Steering Committee

- Marie-Pierre Béal (Marne-la-Vallée, France)
- Mikołaj Bojańczyk (Warsaw, Poland)
- Cristian S. Calude (Auckland, New Zealand)
- Volker Diekert (Stuttgart, Germany)
- Yo-Sub Han (Seoul, Republic of Korea)
- Juraj Hromkovič (Zurich, Switzerland)
- Oscar H. Ibarra (Santa Barbara, USA)
- Nataša Jonoska (Tampa, USA)
- Juhani Karhumäki (Turku, Finland)
- Martin Kutrib (Giessen, Germany)
- Giovanni Pighizzini (Milan, Italy), *chair*
- Michel Rigo (Liège, Belgium)

- Antonio Restivo (Palermo, Italy)
- Wojciech Rytter (Warsaw, Poland)
- Kai Salomaa (Kingston, Canada)
- Shinnosuke Seki (Tokyo, Japan)
- Mikhail Volkov (Ekaterinburg, Russia)
- Takashi Yokomori (Tokyo, Japan)

Honorary Members

- Grzegorz Rozenberg (Leiden, Netherlands)
- Arto Salomaa (Turku, Finland)

Monday, May 9 2022, Room: ISA 1051

1:00pm–1:50pm	Registration
1:50pm–2:00pm	Opening
2:00pm–3:00pm	**Jarkko Kari** *Algebraic methods for periodicity in multidimensional symbolic dynamics*
3:00pm–3:30pm	Coffee break
3:30pm–4:00pm	**Michel Rigo, Manon Stipulanti and Markus Whiteland** *Binomial complexities and Parikh-collinear morphisms*
4:00pm–4:30pm	**Émilie Charlier, Célia Cisternino and Manon Stipulanti** *A full characterization of Bertrand numeration systems*
4:30pm–5:00pm	**Dora Bulgakova, Anna Frid and Jérémy Scanvic** *Prefix palindromic length of the Sierpinski word*

Tuesday, May 10 2022, Room: ISA 1051

9:00am–9:30am	Breakfast
9:30am–10:30am	**Volodymyr Nekrashevych** *Non-deterministic transducers*
10:30am–11:00am	Coffee break
11:00am–11:30am	**Florian Stober and Armin Weiss** *The power word problem in graph products*
11:30am–12:00pm	**Zoran Šunić** *On one-counter positive cones of free groups*
12:00pm–12:30pm	**Stefan Hoffmann** *Automata-theoretical regularity characterizations for the Iterated shuffle on commutative regular languages*
12:30pm–2:00pm	Lunch
2:00pm–3:00pm	**Delaram Kahrobaei** *Group–based cryptography in the quantum era*
3:00pm–3:30pm	Coffee break
3:30pm–4:00pm	**Ondrej Klima and Jonatan Kolegar** *Well quasi-orders arising from finite ordered semigroups*
4:00pm–4:30pm	**Aistis Atminas and Vadim Lozin** *Deciding atomicity of subword-closed languages*
4:30pm–5:00pm	**Szymon Łopaciuk and Daniel Reidenbach** *The Billaud conjecture for $\|\Sigma\| = 4$, and beyond*
5:00pm–7:00pm	Reception / Poster session

Wednesday, May 11 2022, Room: ISA 1051

9:00am–9:30am	Breakfast
9:30am–10:30am	**Paola Bonizzoni** *How can formal languages help pangenomics?*
10:30am–11:00am	**Tomoyuki Yamakami** *Kolmogorov complexity descriptions of the exquisite behaviors of advised deterministic pushdown automata*
11:00am–11:30am	Coffee break
11:30am–12:00pm	**C. Aiswarya, Sahil Mhaskar and M. Praveen** *Checking regular invariance under tightly-controlled string modifications*
12:00pm–12:30pm	**Hyunjoon Cheon, Joonghyuk Hahn and Yo–Sub Han** *On the decidability of infix inclusion problem*
12:30pm–1:00pm	**Francesco Dolce and Pierre-Adrien Tahay** *Column representation of Sturmian words in cellular automata*
1:00pm–2:00pm	Lunch Free afternoon

Thursday, May 12, Room: ISA 1051

10:00am–10:30am	**Volker Diekert** Announcement of awards: Salomaa Prize and the best paper award
10:30am–11:00am	**Jozef Jirasek and Ian McQuillan** *Visit-bounded stack automata*
11:00am–11:30am	Coffee break
11:30am–12:00pm	**Giovanni Pighizzini, Luca Prigioniero and Šimon Sádovský** *Performing regular operations with 1-limited automata*
12:00pm–12:30pm	**Andrea Frosini, Ilaria Mancini, Simone Rinaldi, Giuseppe Romana and Marinella Sciortino** *Logarithmic equal-letter runs for BWT of purely morphic words*
12:30pm–2:00pm	Lunch
2:00pm–3:00pm	**Joel Day** *Word equations in the context of string solving*
3:00pm–3:30pm	Coffee break
3:30pm–4:00pm	**Ryoma Sin'Ya** *Measuring power of locally testable languages*
4:00pm–4:30pm	**Elias Heikkilä, Pyry Herva and Jarkko Kari** *On perfect coverings of two-dimensional grids*
4:30pm–5:00pm	**Oscar Ibarra and Ian McQuillan** *On the complexity of decision problems for counter machines with applications to coding theory*
7:00pm–10:00pm	Conference dinner

Friday, May 13, Room: ISA 1051

9:00am–9:30am	Breakfast
9:30am–10:30am	**Helmut Seidl** *Origin equivalence for macro tree transducers*
10:30am–11:00am	Coffee break
11:00am–11:30am	**Ekaterina Shemetova, Alexander Okhotin and Semyon Grigorev** *Rational index of languages with bounded dimension of parse trees*
11:30am–12:00pm	**Andreas Maletti and Andreea-Teodora Nász** *Weighted tree automata with constraints*
12:00pm–12:30pm	**Olivier Carton** *Preservation of normality by unambiguous transducers*
12:30pm–2:00pm	Lunch

Algebraic Methods for Periodicity in Multidimensional Symbolic Dynamics

Jarkko Kari, University of Turku

A d-dimensional configuration is a coloring $c : \mathbb{Z}^d \longrightarrow A$ of the infinite grid by elements of a finite set $A \subseteq \mathbb{Z}$. It is natural to express such a configuration as a formal power series with d variables $X = (x_1, \ldots, x_d)$ where the coefficient of the $X^{\boldsymbol{u}}$ term is $c(\boldsymbol{u})$ for all $\boldsymbol{u} \in \mathbb{Z}^d$. Invariance of c under the translation by $\boldsymbol{v} \in \mathbb{Z}^d$ then means that the difference (Laurent) polynomial $X^{\boldsymbol{v}} - 1$ *annihilates* the power series in the sense that its formal product with the series is the null series. More generally, we say that a polynomial p *periodizes* c if the formal product pc is strongly periodic. All periodizing polynomials of c form a polynomial ideal, and we can use algebraic geometry to study the structure of this ideal $\mathrm{Per}(c)$. We call a polynomial a *line polynomial* if it has at least two non-zero terms and the exponents of the terms lie on a single line. If $\mathrm{Per}(c)$ contains a line polynomial then clearly c is periodic in the direction of the line. If $\mathrm{Per}(c)$ contains a non-zero polynomial then it can be proved using a dilation lemma and Hilbert's Nullstellensatz that $\mathrm{Per}(c)$ contains a product of line polynomials [1, 2]. In the two-dimensional case $d = 2$ one can further show that $\mathrm{Per}(c)$ is a principal ideal generated by a product of line polynomials. It follows in the two dimensional case that if $\mathrm{Per}(c)$ contains a polynomial without line polynomial factors then c is strongly periodic. Our methods can be applied, for example, on *low-complexity configurations*, containing at most $|D|$ patterns of a finite shape $D \subseteq \mathbb{Z}^d$. We have shown that a two-dimensional uniformly recurrent configuration that has low complexity with respect to a convex shape D must be periodic [4, 5]. This implies that any 2D subshift containing a low complexity configuration with respect to a convex shape D also contains a periodic configuration. We have also shown that any low complexity configuration (with respect to any shape D) of the well-known *Ledrappier subshift* is periodic, and this result can be extended to many other algebraically defined subshifts [6].

jkari@utu.fi

Bibliography

[1] Kari, J., Szabados, M. An algebraic geometric approach to Nivat's conjecture. *In: Proceedings of ICALP 2015*, part II, vol. 9135 of *LNCS*, pp. 273–285 (2015).

[2] Kari, J., Szabados, M. An algebraic geometric approach to Nivat's conjecture. *Information and Computation* **271**, 104481, (2020).

[3] Kari, J. Low-complexity tilings of the plane. In: *Proceedings of DCFS 2019*, vol. 11612 of *LNCS*, pp. 35–45. Springer (2019).

[4] Kari, J., Moutot, E. Decidability and periodicity of low complexity tilings. *In: Proceedings of STACS 2020*, vol. 154 of *LIPIcs*, pp. 14:1–14:12 (2020).

[5] Kari, J., Moutot, E. Decidability and periodicity of low complexity tilings. *Theory of Computing Systems* (2021). https://doi.org/10.1007/s00224-021-10063-8.

[6] Kari, J., Moutot, E. Nivat's conjecture and pattern complexity in algebraic subshifts. *Theoretical Computer Science* **777**, 379–386 (2019).

Non-Deterministic Transducers

Volodymyr Nekrashevych, Texas A&M University

We will discuss automata transducers that are non-deterministic on finite words, but deterministic on infinite one-sided sequences. They naturally appear as automata defining simple groups with interesting properties (like amenability, torsion, or intermediate growth). We will discuss open questions related to them and their connections with other classes of automata and languages.

nekrash@math.tamu.edu

Group-Based Cryptography in the Quantum Era

Delaram Kahrobaei, City University of New York and University of York (UK)

In this talk, I present an overview of the current state-of-the-art in post-quantum group-based cryptography. I describe several families of groups that have been proposed as platforms, with special emphasis in polycyclic groups and graph groups, dealing in particular with their algorithmic properties and cryptographic applications. I then describe some applications of combinatorial algebra in fully homomorphic encryption. In the end we discussed several open problems in this direction. Reference: https://arxiv.org/abs/2202.05917 accepted by the Notices of the American Mathematical Society.

dkahrobaei@gc.cuny.edu

How Can Formal Languages Help Pangenomics?

Paola Bonizzoni, University of Milano-Bicocca

Graph pangenomics is a new emerging field in computational biology that is changing the traditional view of a reference genome from a linear sequence to a new paradigm: a sequence graph (pangenome graph or simply pangenome) that represents the main similarities and differences in multiple evolutionary related genomes. The speed in producing large amounts of genome data, driven by advances in sequencing technologies, is far from the slow progress in developing new methods for constructing and analyzing a pangenome. Most recent advances in the field are still based on notions rooted in established and quite old literature on combinatorics on words, formal languages and space efficient data structures. In this talk I will answer to the question posed by the title of my talk by presenting some challenging problems in pangenomics.

paola.bonizzoni@unimib.it

Word Equations in the Context of String Solving

Joel Day, Loughborough University

String solvers are tools for automatically reasoning about words over some finite alphabet. They are commonly used during static analysis and verification of string manipulating programs. A fundamental problem which string solvers need to address is solving word equations, usually in combination with additional constraints involving e.g. string lengths or regular languages. In this talk, I will present some recent results on the topic of word equations as well as some open problems and challenges related to word equations which are of particular interest in the context of string solving.

J.Day@lboro.ac.uk

Origin Equivalence for Macro Tree Transducers

Helmut Seidl, Technical University of Munich

We consider a notion of origin for deterministic macro tree transducers with look-ahead which records for each output node, the corresponding input node for which a rule-application generated that output node. With respect to this natural notion, we show that origin equivalence is decidable — whenever the transducers are weakly self-nesting. The latter means that whenever two nested calls on the same input node occur, then there must be at least one other node (a terminal output node or a call on another input node) in between these nested calls. We also indicate that for monadic input alphabets, equivalence of the transducers can be reduced to origin equivalence – whenever unrestricted self-nesting is allowed.

These results have been obtained jointly with Sebastian Maneth.

seidl@in.tum.de

Workshop on Discrete and Topological Models in Molecular Biology

DTMB2022 Program & Abstracts

Monday, May 9, Room: ISA 1061 except as indicated

1:00pm–1:50pm	Registration
1:50pm–2:00pm Room: ISA 1051	Opening
2:00pm–3:00pm Room: ISA 1051	**Jarkko Kari** *Algebraic methods for periodicity in multidimensional symbolic dynamics*
3:00pm–3:30pm	Coffee break
3:30pm–4:30pm	**De Witt Sumners** *Knots in biology and fluid dynamics*
4:30pm–5:00pm	**Chad Giusti** *Iterated integrals for time-varying persistence diagrams*

Tuesday, May 10, Room: ISA 1061

9:00am–9:30am	Breakfast
9:30am–10:30am	**Mariel Vázquez** *Topological considerations in genome biology*
10:30am–11:00am	Coffee break
11:00am–11:30am	**Rob Scharein** *Knot shadows in the field*
11:30am–12:30pm	**Sarah Olson** *Modeling cell motility: from agent based models to continuous approximations*
12:30pm–2:00pm	Lunch
2:00pm–2:30pm	**Radmila Sazdanović** *Data, relations and their shape*
2:30pm–3:00pm	**Paweł Dłotko** *Shapes and their meaning in biology*
3:00pm–3:30pm	Coffee break
3:30pm–4:00pm	**Noureen Khan** *Encoding virtual tangles for topological conformation of DNA bounds*
4:00pm–4:30pm	**Asja Radja** *Patterns on single cell surfaces*
4:30pm–5:00pm	**Giuliana Indelicato** *A jigsaw puzzle for Thermus virus* P23-77
5:00pm–7:00pm	Reception / Poster session

Wednesday, May 11, Room: ISA 1061 except as indicated

9:00am–9:30am	Breakfast
9:30am–10:30am Room: ISA 1051	**Paola Bonizzoni** *How can formal languages help pangenomics?*
10:30am–11:00am	**Emanuela Merelli** *The topology of RNA folding*
11:00am–11:30am	Coffee break
11:30am–12:00pm	**Abdulmelik Mohammed** *Genome rearrangement dynamics in the ciliate Oxytricha trifallax*
12:00pm–12:30pm	**Ian McQuillan** *Inference and machine learning with Lindenmayer systems*
12:30pm–1:00pm	**Stella Hartono** *Understanding the relationship between co-transcriptional R-loops and DNA topology*
1:00pm–2:00pm	Lunch
2:00pm–3:00pm	**Peter Bubenik** *Topological data analysis for biological images and video*
3:00pm–3:30pm	Coffee break
3:30pm–4:00pm	**Tom Needham** *Hypergraph co-optimal transport*
4:00pm–4:30pm	**Daniela Genova** *Equivalence and minimization of reaction systems*

Thursday, May 12, Room: ISA 1061

10:30am–11:00am	**Yuanan Diao** *The ropelengths of alternating knots*
11:00am–11:30am	Coffee break
11:30am–12:30pm	**Vladimir Itzkov** *The topology of Boltzmann machines: learning and dimensionality*
12:30pm–2:00pm	Lunch
2:00pm–2:30pm	**Francesca Storici** *Models for DNA–RNA interactions in double-strand break repair*
2:30pm–3:00pm	**Michela Quadrini** *RNA abstractions for structural comparison and classification*
3:00pm–3:30pm	Coffee break
3:30pm–4:30pm	**Christine Heitsch** *Can geometric combinatorics improve RNA folding predictions?*
4:30pm–5:00pm	**Svetlana Poznanović** *Improving RNA branching predictions: advances and limitations*
7:00pm–10:00pm	Conference dinner at Embassy Suites

Friday, May 14, Room: ISA 1061

9:00am–9:30am	Breakfast
9:30am–10:30am	**Ion Petre** *Network controllability: algorithmics and applications in medicine*
10:30am–11:00am	Coffee break
11:00am–12:00pm	**Lila Kari** *Mathematical representations of DNA sequences and the biodiversity grand challenge*
12:00am–12:30pm	**Matthew Macauley** *Artifacts of synchrony in Boolean models of tryptophan*
12:30pm–2:00pm	Lunch

Algebraic Methods for Periodicity in Multidimensional Symbolic Dynamics

Jarkko Kari, University of Turku

A d-dimensional configuration is a coloring $c : \mathbb{Z}^d \longrightarrow A$ of the infinite grid by elements of a finite set $A \subseteq \mathbb{Z}$. It is natural to express such a configuration as a formal power series with d variables $X = (x_1, \ldots, x_d)$ where the coefficient of the $X^{\boldsymbol{u}}$ term is $c(\boldsymbol{u})$ for all $\boldsymbol{u} \in \mathbb{Z}^d$. Invariance of c under the translation by $\boldsymbol{v} \in \mathbb{Z}^d$ then means that the difference (Laurent) polynomial $X^{\boldsymbol{v}} - 1$ *annihilates* the power series in the sense that its formal product with the series is the null series. More generally, we say that a polynomial p *periodizes* c if the formal product pc is strongly periodic. All periodizing polynomials of c form a polynomial ideal, and we can use algebraic geometry to study the structure of this ideal $\mathrm{Per}(c)$. We call a polynomial a *line polynomial* if it has at least two non-zero terms and the exponents of the terms lie on a single line. If $\mathrm{Per}(c)$ contains a line polynomial then clearly c is periodic in the direction of the line. If $\mathrm{Per}(c)$ contains a non-zero polynomial then it can be proved using a dilation lemma and Hilbert's Nullstellensatz that $\mathrm{Per}(c)$ contains a product of line polynomials [1, 2]. In the two-dimensional case $d = 2$ one can further show that $\mathrm{Per}(c)$ is a principal ideal generated by a product of line polynomials. It follows in the two dimensional case that if $\mathrm{Per}(c)$ contains a polynomial without line polynomial factors then c is strongly periodic. Our methods can be applied, for example, on *low-complexity configurations*, containing at most $|D|$ patterns of a finite shape $D \subseteq \mathbb{Z}^d$. We have shown that a two-dimensional uniformly recurrent configuration that has low complexity with respect to a convex shape D must be periodic [4, 5]. This implies that any 2D subshift containing a low complexity configuration with respect to a convex shape D also contains a periodic configuration. We have also shown that any low complexity configuration (with respect to any shape D) of the well-known *Ledrappier subshift* is periodic, and this result can be extended to many other algebraically defined subshifts [6].

jkari@utu.fi

Bibliography

[1] Kari, J., Szabados, M. An algebraic geometric approach to Nivat's conjecture. *In: Proceedings of ICALP 2015*, part II, vol. 9135 of *LNCS*, pp. 273–285 (2015).

[2] Kari, J., Szabados, M. An algebraic geometric approach to Nivat's conjecture. *Information and Computation* **271**, 104481, (2020).

[3] Kari, J. Low-complexity tilings of the plane. In: *Proceedings of DCFS 2019*, vol. 11612 of *LNCS*, pp. 35–45. Springer (2019).

[4] Kari, J., Moutot, E. Decidability and periodicity of low complexity tilings. *In: Proceedings of STACS 2020*, vol. 154 of *LIPIcs*, pp. 14:1–14:12 (2020).

[5] Kari, J., Moutot, E. Decidability and periodicity of low complexity tilings. *Theory of Computing Systems* (2021). https://doi.org/10.1007/s00224-021-10063-8.

[6] Kari, J., Moutot, E. Nivat's conjecture and pattern complexity in algebraic subshifts. *Theoretical Computer Science* **777**, 379–386 (2019).

Knots in Biology and Fluid Dynamics

De Witt Sumners, Florida State University

Knots in DNA can interfere with vital cellular life processes, but they can also provide experimental insight into the mechanism of enzymes that facilitate these processes. This talk will discuss the topology of topoisomerase (the enzyme that solves DNA entanglement) and site-specific recombinase (an enzyme that changes the genome). Very similar to DNA recombination is vortex reconnection in fluids, and the topology of anti-parallel vortex reconnection will be discussed.

sumners@math.fsu.edu

Iterated Integrals for Time-Varying Persistence Diagrams

Chad Giusti, University of Delaware
Joint work with: Darrick Lee

Persistent homology is a useful tool for characterizing mesoscale structure in complex systems, usually via functional summaries of persistence diagrams. When studying time-varying systems, we obtain sequences of persistence diagrams. Functional summaries of such parameterized families of persistence diagrams are difficult to construct, and those in common use often lack important theoretical guarantees that we rely on when studying static persistence diagrams and their summaries. Here, we describe a describe a general framework for constructing computable feature sets for time-varying persistence diagrams. We describe such a set of features that provides guarantees of stability, universality, and characteristicness, making it well-suited for standard statistical and machine learning tasks. Finally, we demonstrate their use by imputing model parameters for simulated swarms from sparse, time- and size-inhomogeneous samples of the positions of their constituent elements.

cgiusti@udel.edu

Tuesday, May 10

Topological Considerations in Genome Biology

Mariel Vázquez, University of California Davis

The genetic code of viruses and of living organisms is encoded in very long DNA or RNA molecules, which are tightly packaged in confined environments. Understanding the geometry and topology of nucleic acids is key to understanding the mechanisms of viral infection and the inner workings of a cell. We use techniques from knot theory and low-dimensional topology, aided by discrete methods and computational tools, to ask questions about the topological state of a genome. I will illustrate the use of these methods with examples drawn from recent work in my group.

mrlvazquez@ucdavis.edu

Knot Shadows in the Field

Rob Scharein, Hypnagogic Software

A new app for Android and iOS mobile devices will be presented that allows users to capture an image of a knotted object. The app then identifies all the possible knot types corresponding to the shadow of the knot. Application to identifying knot types in electron micrographs of DNA will be demonstrated.

rob@knotplot.com

Modeling Cell Motility: From Agent Based Models to Continuous Approximations

Sarah Olson, Worcester Polytechnic Institute
Joint work with: Michael Yereniuk and Simone Cassani

Movement is ubiquitous to cells and takes on many different modes, depending on the particular cell type and the surrounding environment. In the presence of no cues or sources, a cell will often move in a non-biased and non-persistent direction corresponding to a random walk. To understand cellular motility, we can utilize agent based models (ABM) where cells are modeled as individual agents (points in the domain) with a given state where cells are able to move and/or transition between states based on a set of rules. The agent based modeling framework is often simple to understand and develop, but they can be difficult to understand in terms of long term dynamics and stability. Through several examples, we will show how continuum limits can be derived and analyzed to investigate questions such as the average time spent in a particular state and the probability of a cell's location at a given time point. In the case of cancer cells being treated with different drugs, we will show how we can model absorption of drugs and resulting state changes. In addition, we will illustrate that if the cell status in these experiments is known, we can determine the cellular absorption rate of these drugs.

sdolson@wpi.edu

Data, Relations and Their Shape

Radmila Sazdanović, North Carolina State University

TDA provides tools for discovering relevant features of data by analyzing its shape. In this context we develop tools for visualizing maps between high dimensional spaces with the goal of discovering relations between data sets with expected correlations. Examples include analyzing relations between numerical and polynomial knot and graph invariants, as well as breast cancer subtype characterization. This is joint work with P. Dlotko, D. Gurnari, and D. Scofield.

rsazdan@ncsu.edu

Shapes and Their Meaning in Biology

Paweł Dłotko, Dioscuri Centre in TDA

Shape is an abstract concept that is intuitive for humans, but still lacks good mathematical description. Using topology and geometry, and recently topological data analysis, we are making very first steps in formalizing and quantifying the concept of shapes. In this talk, I will discuss a few projects in which we are adopting existing or providing new tools from topological data analysis to solve biologically relevant problems. Starting from neurons and their abstract shapes, through bones and histopathological images, Covid19 data, ending up with breast cancer datasets we will explain how understanding of the underlying 'shapes hidden data' helps us to get some insides on the underlying biology.

pdlotko@impan.pl

Encoding Virtual Tangles for Topological Conformation of DNA Bounds

Noureen Khan, University of North Texas at Dallas

Despite being discovered some twenty years ago, knotted proteins remain a valuable discovery for theoretical knot theorists. Tangle analysis is being used to determine the topological shape of DNA segments, and simple invariants to study the proteins which bind segments of DNA. A fascinating aspect of our research is discovering how DNA can be colored and how a virtual knot theory invariant called tricolorability can provide insight into enzyme action. By encoding, we search for possible DNA conformations on the basis of virtual coloring, a computational method for finding the topological conformation of DNA bound within a protein complex.

Noureen.Khan@unt.edu

Patterns on Single Cell Surfaces

Asja Radja, Harvard University

Patterns are ubiquitous on biological cell surfaces, and we have only begun to develop an understanding of their formation. In this talk, I draw inspiration from rigid, extracellular surface patterns found on cells: radiolaria (Ernst Haeckel's favorite marine protozoa) and pollen grains, two incredibly morphologically varied systems with strikingly similar surface patterns, and discuss if there is a common mechanism in their pattern formation that can be described by a single, generalized theory. I highlight our experiments done in the more tractable system of the two (pollen) and expand upon how our modified Landau-Ginzburg theory of pollen wall development in this system may be applied to organisms in the entirely disparate kingdom, protozoa, expanding upon an idea first hinted towards in D'Arcy Wentworth Thompson's On Growth and Form.

radja.asja@gmail.com

A Jigsaw Puzzle for Thermus Virus P23-77

Giuliana Indelicato, Politecnico di Torino

Thermus phage P23-77 is a virus infecting bacteria that live in hot springs. It self-assembles from copies of two types of coat proteins and the resulting capsid layout does not conform to the classical geometry of spherical viruses. Two questions therefore arise: which assembly mechanism can lead to this arrangement, and what is the evolutionary advantage of the specific layout of the hexamers found in this virus.
We first classify all possible spherical tilings obeying icosahedral symmetry that can be constructed by these two species of proteins and are consistent with biologically relevant exclusion rules.
We also investigate various sets of candidates for local assembly rules for capsids made of these two types of proteins, and define a set of simple local rules that yield the observed layout.
Finally, we show that there is a correlation between sites of stress concentration in the capsid and the structure of the hexamers at those sites, which suggests that the observed arrangement of the capsid proteins helps reinforce the shell against internal stresses.

giuliana.indelicato@polito.it

Wednesday, May 11

How Can Formal Languages Help Pangenomics?

Paola Bonizzoni, University of Milano-Bicocca

Graph pangenomics is a new emerging field in computational biology that is changing the traditional view of a reference genome from a linear sequence to a new paradigm: a sequence graph (pangenome graph or simply pangenome) that represents the main similarities and differences in multiple evolutionary related genomes. The speed in producing large amounts of genome data, driven by advances in sequencing technologies, is far from the slow progress in developing new methods for constructing and analyzing a pangenome. Most recent advances in the field are still based on notions rooted in established and quite old literature on combinatorics on words, formal languages and space efficient data structures. In this talk I will answer to the question posed by the title of my talk by presenting some challenging problems in pangenomics.

paola.bonizzoni@unimib.it

The Topology of RNA Folding

Emanuela Merelli, University of Camerino

Investigating how the interactions between the elementary units of RNA can determine the 3-D shape of molecules is fundamental to understanding their biological functions. In this talk, we discuss how RNA shapes may reveal some features typical of a model of computation.

emanuela.merelli@unicam.itu

Genome Rearrangement Dynamics in the Ciliate Oxytricha trifallax

Abdulmelik Mohammed, University of South Florida
Joint work with: Richard V. Miller, Jaspreet S. Khurana, Yi Feng, Rafik Neme, Laura Landweber, Masahico Saito and Nataša Jonoska

Development of a somatic nucleus from a germline nucleus in the ciliate Oxytricha trifallax involves massive genome rearrangement via events of sequence reordering, as well as programmed DNA deletion. We study the timing of removal of deleted DNA segments from the germline by capturing and aligning reads from a set of developmental time points against the reference precursor and product genomes. The presence of a deleted segment at an intermediate time point is captured by a retention score that represents the average of the number of reads that align at the two ends of the deleted region. Vectors of the retention scores of a filtered set of deleted regions were clustered using the k-means algorithm. Clustering reveals that conventional deleted segments, i.e. those that can be deleted without reordering the flanking sequences, tend to be amplified earlier than those removed from scrambled regions, suggesting that conventional deletions may occur earlier. Deleted segments that are removed from the same germline chromosome, as well as those segments whose flanking regions are incorporated into the same somatic chromosome, are significantly more likely to be in the same cluster, suggesting their concurrent processing. Moreover, proximity of deleted segments in the germline chromosome leads to a significantly higher likelihood of their concurrent processing.

abdulmelik@usf.edu

Inference and Machine Learning with Lindenmayer Systems

Ian McQuillan, University of Saskatchewan

Lindenmayer systems (L systems) are a formal grammar system with parallelism in the application of the string rewriting rules. They were created to model multicellular structures present in many biological organisms with inherent self-similarity, especially plants. They have been widely used to create realistic visual simulations of developing plants, and they can also capture mechanisms of development.
We describe our lab's work on applying L systems towards plants. This includes using synthetic images from L system simulations to improve training of artificial neural networks for recognizing features from real plant images. Furthermore, we describe our lab's work on automatic inference of L systems from data, including both deterministic, stochastic, and parameterized L systems.

mcquillan@cs.usask.ca

Understanding the Relationship Between Co-transcriptional R-loops and DNA Topology

Stella Hartono, University of California Davis
Joint work with: Robert Stolz, Shaheen Sultana, Maika Malig, Craig Benham and Frédéric Chédin

R-loops are prevalent and conserved non-B DNA structures consisting of an RNA:DNA hybrid and looped-out single-stranded DNA sequence. R-loops form during transcription where the nascent RNA anneals with the template DNA strand behind the advancing RNA polymerase. Over the last decade, R-loop formation has been associated with both physiological and pathological outcomes from yeast to humans. R-loops have been implicated as key transcriptional and epigenetic regulator as well as site of replication origins. However, aberrant R-loop formations also linked to genomic instability and many human diseases. Therefore, it is important to understand the mechanism(s) by which R-loops form. Previous experimental evidence suggests that DNA sequence and topology affect R-loop formation, but little is known about how exactly these factors interact. We developed a statistical mechanical equilibrium model of R-loop formation in superhelical DNA. This model showed that the significant energy barrier imposed by the formation of junctions can be overcome in two ways: base-pairing energy over favorable sequences and the ability of R-loop structure to partially or fully relax DNA region by absorbing negative superhelicity and returning it to a lower energy state. In vitro transcription assays showed that negative superhelicity is required to form R-loop structure even in favorable region, and its formation relaxed the plasmid. Experimental results from single-molecule R-loop footprinting (SMRF-seq) following in vitro transcriptions also showed a strong agreement with the prediction from the model. Overall, these results elucidate further the interplay between DNA topology and DNA sequence on R-loop formation.

srhartono@ucdavis.edu

Topological Data Analysis for Biological Images and Video

Peter Bubenik, University of Florida

I will present the results of two recent projects applying topological data analysis (TDA) and machine learning (ML) to biological data. In the first, we have developed a new tool, TDAExplore, that combines TDA and ML to both classify biological images and to provide a visualization that is biologically informative. In the second, we use TDA and ML to classify quasi-periodic biological videos and we apply TDA to such a video to produce synthetic periodic videos.

peter.bubenik@ufl.edu

Hypergraph Co-optimal Transport

Tom Needham, Florida State University
Joint work with: Samir Chowdhury, Ethan Semrad, Bei Wang and Youjia Zhou

The concept of a hypergraph naturally generalizes that of a graph by allowing "edges" to contain more than two vertices. Hypergraphs thereby capture multi-way relationships in data, and they have consequently seen a number of applications in higher-order network analysis, machine learning and molecular biology. In this talk, I will describe work on the theoretical foundations of hypergraph theory using ingredients from optimal transport. I will discuss metric and categorical properties of the space of hypergraphs, as well as a computational framework geared towards applications.

tneedham@fsu.edu

Equivalence and Minimization of Reaction Systems

Daniela Genova, University of North Florida

Reaction systems were introduced by A. Ehrenfeucht and G. Rozenberg as a theoretical model of computation capturing the two main features of biochemical reactions: facilitation and inhibition. Each reaction system consists of a finite set of background entities and a finite set of reactions. A result function on the power set of entities determines how computation evolves. Different sets of reactions may induce the same result function, in which case they are called functionally equivalent. We study the equivalence of such sets of reactions and different ways of minimizing them.

d.genova@unf.edu

The Ropelengths of Alternating Knots

Yuanan Diao, University of North Carolina at Charlotte

The ropelength $R(K)$ of a knot K is the minimum length of a unit thickness rope needed to tie the knot. If K is alternating, it is conjectured that $R(K) \geq a\mathrm{Cr}(K)$ for some constant $a > 0$, where $\mathrm{Cr}(K)$ is the minimum crossing number of K. In this talk I will first give a brief introduction to the ropelength problem. I will then show that there exists a constant $a_0 > 0$ such that $R(K) \geq a_0\mathbf{b}(K)$ for any knot K, where $\mathbf{b}(K)$ is the braid index of K. It follows that if $\mathbf{b}(K) \geq a_1\mathrm{Cr}(K)$ for some constant $a_1 > 0$, then $R(K) \geq a_0a_1\mathrm{Cr}(K) = a\mathrm{Cr}(K)$. However if $\mathbf{b}(K)$ is small compared to $\mathrm{Cr}(K)$ (in fact there are alternating knots with arbitrarily large crossing numbers but fixed braid indices), then this result cannot be applied directly. I will show that this result can in fact be applied in an indirect way to prove that the conjecture holds for a large class of alternating knots, regardless what their braid indices are.

ydiao@uncc.edu

The Topology of Boltzmann Machines: Learning and Dimensionality

Vladimir Itskov, Pennsylvania State University
Joint work with: Hannah Rocio Santa Cruz and Biy-Kuang Day

Statistical mechanics models, such as the Boltzmann machines, have been extensively used in physics, machine-learning and computational biology. These models are probability distributions on a Boolean lattice and thus possess natural topological structures. It turns out that the tools of combinatorial algebraic topology are quite useful in approaching problems such as learning and dimensionality reduction in the context of these models. I will describe two use cases for topological tools in "learning by parts" as well as inferring the minimal dimension of a restricted Boltzmann machine.

vladimir.itskov@psu.edu

Models for DNA-RNA Interactions in Double-Strand Break Repair

Francesca Storici, Georgia Institute of Technology

Double-strand breaks (DSBs) in DNA are challenging lesions to repair. Mammalian cells employ at least three DSB repair mechanisms with a preference for non-homologous end joining (NHEJ) over homologous recombination (HR) and microhomology-mediated end joining (MMEJ). Differently from HR, NHEJ and MMEJ do not utilize a DNA template molecule to recover the damaged or lost nucleotides. NHEJ directly ligates the broken DNA ends, while MMEJ exploits the alignment of short microhomologies at the DSB sides, and is associated with deletions of the sequence between the microhomologies. Although in recent years it has become more evident that RNA can interact with DNA in the process of repairing a DSB in DNA, whether a transcript RNA has a direct role in DSB repair mechanisms in mammalian cells is unknown. Moreover, there are almost no mathematical studies to characterize the role of RNA in DSB repair. In this study, we generated next-generation sequencing libraries of DSB repair sites, and we applied both bioinformatics and graph theoretical methods to analyze the sequencing data. We found that transcript RNA facilitates DNA DSB repair in a sequence-dependent manner in human cells. Our results demonstrate an unexpected function of RNA in directing the way DSBs are repaired in human cells. Together, our data provide new avenues to understanding mechanisms of genome integrity and evolution, and to advancing genome editing.

storici@gatech.edu

RNA Abstractions for Structural Comparison and Classification

Michela Quadrini, University of Camerino
Joint work with: Luca Tesei and Emanuela Merelli

RNAs fold into three-dimensional shapes to enable their biological functions. The comparison and classification of RNAs play a fundamental role in understanding their behaviour and grouping similar organisms. The literature has faced these problems with several approaches whose results strongly depend on molecular abstractions and representations. Most of them focus on the secondary structure, a shape abstraction that consists of loops and their compositions.

In this talk, motivated by our previous results, we define an RNA secondary structure abstraction called core to more appropriately tackle the problem of comparing and classifying structures with arbitrary pseudoknots. Cores determine secondary structure equivalent classes that can be compared using already introduced techniques such as ASPRA distance and Progressive stem matching. We also formalise core decomposition as loops and their interactions by using a formal language approach over oriented surfaces with boundaries associated with the decomposition itself.

michela.quadrini@unicam.it

Can Geometric Combinatorics Improve RNA Folding Predictions?

Christine Heitsch, Georgia Institute of Technology
Joint work with: Svetlana Poznanović et al.

Accurate prediction of RNA base pairing remains an open problem in computational molecular biology. In the case of RNA viral genomes, the consequences of this have profound ramifications. The branching of an RNA configuration is one of the least accurately predicted characteristics, and possibly the one most amenable to mathematical analysis. We show that it is possible to significantly improve the prediction accuracy for well-defined families, and why the general problem is so difficult.

heitsch@math.gatech.edu

Improving RNA Branching Predictions: Advances and Limitations

Svetlana Poznanović, Clemson University
Joint work with: Carson Wood, Michael Cloer and Christine Heitsch

Minimum free energy prediction of RNA secondary structures is based on the Nearest Neighbor Thermodynamics Model. While such predictions are typically good, the accuracy can vary widely even for short sequences, and the branching thermodynamics are an important factor in this variance. Recently, the simplest model for multiloop energetics—a linear function of the number of branches and unpaired nucleotides—was found to be the best. Subsequently, a parametric analysis coupled with an ad hoc method for parameter search demonstrated that per family accuracy can be improved by changing the weightings in this linear function. We develop a branch-and-bound algorithm that finds the set of optimal parameters with the highest average accuracy for a given set of sequences. Our analysis shows that the previous ad hoc parameters are nearly optimal for tRNA and 5S rRNA sequences on both training and testing sets. Moreover, cross-family improvement is possible but more difficult because competing parameter regions favor different families.

spoznan@clemson.edu

Network Controllability: Algorithmics and Applications in Medicine

Ion Petre, University of Turku

The problem of controlling a dynamic network has a long history in control theory, with roots in a diversity of mathematical methods from complex analysis to topology, graph theory and computational complexity. The basic problem setup is that of a dynamical system represented as a directed graph, with nodes influencing each other's dynamics. Control is sought over a given set of targets, in the sense of being able to change their configuration through external interventions on some well-chosen input nodes in the network, taking advantage of the network topology. We are interested in finding a minimal set of input nodes in the network such that the behavior of the target nodes may be changed arbitrarily through a well-chosen sequence of signals to the input nodes, cascaded throughout the network through its wiring. We focus on formalizations of this network controllability problem that maximize its applicability in biomedicine, including a specific set of targets to choose from (e.g., disease-specific essential genes), a specific set of inputs to choose from (e.g., drug targets), as well as non-linear network topologies. We discuss some of our recent results on the computational complexity of the structural targeted network controllability problem, some fast heuristics for it, and some applications in cancer medicine.

ion.petre@utu.fi

Mathematical Representations of DNA Sequences and the Biodiversity Grand Challenge

Lila Kari, University of Waterloo

Of the estimated 20+ million species of multicellular organisms which share our planet, 95% have still not been catalogued or classified, and do not have a scientific name. To address the Grand Challenge of identifying and classifying all species on Earth, a multitude of techniques have been proposed for genomic sequence analysis and comparison. In this talk we discuss several mathematical representations of DNA sequences, and their use in conjunction with supervised machine learning and unsupervised machine learning techniques for ultrafast, accurate, and scalable genome classification at all taxonomic levels.
This effort is part of BIOSCAN, an international project involving over 1,000 researchers representing more than 40 countries, which uses DNA-based technologies to map and analyze Earth's biodiversity.

lila@uwaterloo.ca

Artifacts of Synchrony in Boolean Models of Tryptophan

Matthew Macauley, Clemson University
Joint work with: Isadora Deal and Robin Davies

The tryptophan (trp) operon in E. coli codes for the proteins responsible for the synthesis of the amino acid tryptophan from chorismic acid. It has been one of the most well-studied gene networks since its discovery in the 1960s. The tryptophanase (tna) operon codes for proteins needed to transport and metabolize it. Both of these have been modeled individually with delay differential equations under the assumption of mass-action kinetics. Recent work has provided strong evidence for bistable behavior of the tna operon, by identifying a medium range of tryptophan in which the system has two stable steady-states, which was reproduced experimentally. In this talk, we will show how a Boolean model can capture this bistability. We will also develop and analyze a Boolean model of the trp operon, and then combine these two to create a single Boolean model of the transport, synthesis, and metabolism of tryptophan. In this amalgamated model, the bistability disappears, presumably reflecting the ability of the trp operon to produce tryptophan and drive the system toward homeostasis. All of these models have longer attractors that we call "artifacts of synchrony", which disappear in the asynchronous automata. This curiously matches the behavior of a recent Boolean model of the arabinose operon in E. coli, and we discuss some open-ended questions that arise along these lines.

macaule@clemson.edu

26th International Conference on Developments in Language Theory & Workshop on Discrete and Topological Models in Molecular Biology

Poster Abstracts

String Attractors of Finite Words: Combinatorial and Algorithmic Perspectives

Giuseppe Romana, University of Palermo

Given a finite word w, a *string attractor* Γ is a subset of positions in w such that every distinct factor of w has at least an occurrence that crosses a position in Γ. The size γ^* of a smallest string attractor for a word has been introduced in the field of data compression, and it represents a natural lower-bound for many well known compression schemes [3].
This poster is intended as an introductionary survey on the different interesting aspects of string attractors for finite words. Some combinatorial properties of the measure γ^* have been explored in [6]. In particular, the behaviour of the measure γ^* when some operations are applied to words and its monotonicity have been investigated.
Combinatorial properties of words have been used to determine the value of γ^* for Standard Sturmian words and Thue-Morse words [6, 5]. Moreover, upper bounds on γ^* for some infinite families of finite words can be deduced by using results on other complexity measures related to LZ-based and BWT-based compression schemes [1, 2].
From an algorithmic point of view, the computation of a smallest string attractor for a word is an NP-complete problem, as well as the computation of other variants of the measure γ^*. Efficient algorithms to check the validity and minimality of a string attractor has been proposed [4].
In this poster, an overview of some combinatorial problems related to string attractors is given. Moreover, an alternative and easy to build optimal time algorithm to check whether or not a set Γ is a string attractor for a word w is presented.

giuseppe.romana01@community.unipa.it

Bibliography

[1] Constantinescu, S., Ilie, L. The Lempel–Ziv complexity of fixed points of morphisms. *SIAM Journal of Discrete Mathematics*, **21**, 2, 466–481, (2007).

[2] Frosini A., Mancini I., Rinaldi S., Romana G., Sciortino, M. Logarithmic equal-letter runs for BWT of purely morphic words. In *DLT*, vol. 13257 of *LNCS*. Springer (2022).

[3] Kempa, D., Prezza, N. At the roots of dictionary compression: String attractors. In *STOC 2018*, pp. 827–840. ACM (2018).

[4] Kempa D., Policriti A., Prezza N., and Rotenberg, E. String attractors: Verification and optimization. In *ESA*, vol. 112 of *LIPIcs*, pp. 52:1–52:13, Dagstuhl, Germany (2018). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[5] Kutsukake, K., Matsumoto, T., Nakashima, Y., Inenaga, S., Bannai H., Takeda, M. On repetitiveness measures of Thue-Morse words. In *SPIRE*, vol. 12303 of *LNCS*, pp. 213–220. Springer (2020).

[6] Mantaci S., Restivo A., Romana G., Rosone G., Sciortino M. A combinatorial view on string attractors. *Theorertical Computer Science* **850**, 236–248 (2021).

Graph-Theoretic Analysis of High-Throughput DNA Sequencing Data

Tejasvi Channagiri, University of South Florida
Joint work with: Margherita Maria Ferrari, Youngkyu Jeon, Penghao Xu, Francesca Storici and Nataša Jonoska

DNA double-strand breaks are dangerous lesions in cells that can lead to mutations, cell death, and cancer. In this project we study the effect of transcript RNA on DNA double-strand break repair. We design two different DNA sequences that produce two types of transcript RNA (A & B) to study how they affect double-strand breaks in the same gene location. Using mathematical methods based on graph theory, we analyze DNA sequences and visualize the sequence variations in the repaired DNA, thus gaining insight into RNA's role in DNA repair.

tchannagri@usf.edu

Topological Measures of DNA Scrambling

Lina Fajardo Gómez, University of South Florida
Joint work with: Margherita Maria Ferrari, Nataša Jonoska and Masahico Saito

We describe the complexity of DNA recombination processes according to the different rearrangement pathways taken by the molecules. Certain species of ciliates are model organisms to study these processes because they undergo massive rearrangements during reproduction. A transcriptionally active macronucleus is destroyed and regenerated from a germline micronucleus where the majority of genes are fragmented into several segments that may be out of order or reversed, separated by so-called "junk" DNA. An alignment of short repeat sequences, called pointers, at the endpoints of gene fragments guides the rearrangement. Double-occurrence words (DOWs), where every symbol corresponding to a pointer appears exactly twice, can be associated to genetic sequences where recombination happens. In this case, the deletion of certain subwords in DOWs indicates that the recombination process has taken place at that location. A word graph is a graph which has DOWs as vertices, while edges indicate when a word can be obtained from another through the deletion of specific subwords. We construct a geometric object based on the graph by attaching to it triangles, squares, and other related higher dimensional polytopes. On this structure we study geometric features like the number of loops, holes and cavities, which represent different recombination pathways. This model relates the complexity of the rearrangement to topological features of the word graph describing it, which may be used to compare DNA recombinant processes in different genes and across species.

fajardogomez@usf.edu

Topological Data Analysis of Pattern Formation of Human Induced Pluripotent Stem Cell Colonies

Iryna Hartsock, University of Florida
Joint work with: Daniel Cruz, Eunbi Park, Jack Toppen, Peter Bubenik and Melissa Kemp

Pluripotent stem cells have the ability to differentiate into different types of cells. We use various concentrations of a drug called doxycycline to increase the rate of cell differentiation which results in different pattern formations of stem cell colonies. We apply topological data analysis to images of stem cell colonies with various levels of doxycycline to study and detect changes in cell pattern formations.

irynakuz@ufl.edu

RNA-Mediated DNA Double-Strand Break Repair in Human Cells

Youngkyu Jeon, Georgia Institute of Technology

Our group has demonstrated that RNA is an alternative template for the repair of double-strand break (DSB) in DNA. In addition to observing RNA-templated DNA repair by synthetic RNA oligonucleotides in yeast and human cells, we showed that endogenous transcript RNA can directly template DSB repair in its own DNA in cis in yeast cells. Here, we examined whether RNA can also be a template to mediate DNA DSB repair in mammalian cells. We developed a genetic assay to study RNA-mediated DNA DSB repair in mammalian cells by transferring the yeast 'cis system' into a DNA plasmid (cis plasmid) to perform a transient assay. We exploit the clustered regularly interspaced short palindromic repeats (CRISPR)-associated protein 9 (Cas9) system with the Cas9 endonuclease and guide RNA (gRNA) to direct Cas9 to the chosen target site on the cis plasmid to generate a DSB. Products of DNA DSB repair on the cis plasmid were captured by Next-Generation Sequencing (NGS) by our customized NGS library prep method, which enables sequencing near the DSB site of the cis plasmid. We induced a DSB (1-DSB) or DNA gap (2-DSBs) to study how a transcript RNA impacts DNA DSB repair in a sequence-dependent manner through specific DSB repair mechanisms such as non-homologous end joining (NHEJ), microhomology-mediated end joining (MMEJ), and RNA-templated DNA repair. Frequencies of DNA DSB repair were determined by the number of sequencing reads, and specific sequences produced by each DSB repair mechanisms were used for calculating frequencies of each DSB repair mechanisms. Through the NGS data analyses, we found that endogenous RNA transcripts can mediate DNA DSB repair in a sequence-dependent manner through NHEJ, MMEJ and RNA-templated repair.

yjeon39@gatech.edu

The Number of Rational Links with a Given Deficiency

Dawn Ray, University of North Carolina at Charlotte
Joint work with: Yuanan Diao and Michael Finney

Let $\mathcal{U}_n$ be the set of un-oriented and rational links with crossing number n, a precise formula for $|\mathcal{U}_n|$ was obtained by Ernst and Sumners in 1987. In this paper, we study the enumeration problem of oriented rational links. Let Λ_n be the set of oriented rational links with crossing number n and let $\Lambda_n(d)$ be the set of oriented rational links with crossing number n ($n \geq 2$) and deficiency d. In this paper, we derive precise formulas for $|\Lambda_n|$ and $|\Lambda_n(d)|$ for any given n and d and show that

$$\Lambda_n(d) = F^{(d)}_{n-d-1} + \frac{1+(-1)^{nd}}{2} F^{(\lfloor \frac{d}{2} \rfloor)}_{\lfloor \frac{n}{2} \rfloor - \lfloor \frac{d+1}{2} \rfloor}$$

where $F^{(d)}_n$ is the convolved Fibonacci sequence.

dawray@fastmail.fm

Machine Learning and Topological Data Analysis in Stem Cell Pattern Formation

Alexander Ruys de Perez, Georgia Institute of Technology

Pluripotent stem cells play a critical role in embryonic development, and studying their differentiation into the different germ layers is an important task. We ask whether the relative positioning of the cells provides enough information to classify those cells as differentiating into a specific tissue. To this end we train a neural network on a selection of stem cell colonies, each treated with a morphogen that induces differentiation into the cells of a particular germ layer. We examine whether the persistent homology features taken from the colony's cell locations can be used to make an accurate prediction about what morphogen was applied.

amrp3@gatech.edu

Modeling Functional Redundancy in Microbial Community

Sandra Annie Tsiorintsoa, Clemson University
Joint work with: Sharon Bewick and Matthew Macauley

In recent years, many microbiome habitats, such as human guts, soils and oceans, have been simplified as a result of human activity. By choosing less complex and varied diets, for example, we decrease the number of different chemicals available to our gut microbes, decreasing gut microbiome diversity and causing a poor digestive health. Likewise, practicing monoculture farming instead of polyculture diminishes soil nutrients availability to microbes resulting in loss of soil fertility. Many studies show that simplified habitat complexity leads to less diversity in microbial communities. What is less clear is if this simplicity also affects functional redundancy, which is the number of species that perform a given function, of these communities. High levels of functional redundancy are important, because they contribute to ecosystem stability. To answer this question, we use metacommunity models to explore the connection between functional redundancy and habitat complexity. Specifically, we consider various paradigms for local community assembly within a larger metacommunity, including environmental filtering and niche partitioning. Our model for environmental filtering indicates that functional redundancy is constant with respect to the local habitat complexity. As for niche partitioning, we observe that functional redundancy rises with the local habitat complexity. These models suggest that different modes of community assembly yield different relationships between habitat complexity and functional redundancy. We explore these findings as they pertain consequences for maintaining stable microbial ecosystem services in anthropogenically simplified landscapes.

stsiori@clemson.edu