Report from Dagstuhl Seminar 17452

Algorithmic Cheminformatics

Edited by

Jakob L. Andersen[1], Christoph Flamm[2], Daniel Merkle[3], and Peter F. Stadler[4]

1 **Tokyo Institute of Technology, JP,** jlandersen@elsi.jp
2 **Universität Wien, AT,** xtof@tbi.univie.ac.at
3 **University of Southern Denmark – Odense, DK,** daniel@imada.sdu.dk
4 **Universität Leipzig, DE,** studla@bioinf.uni-leipzig.de

Abstract

Dagstuhl Seminar 17452 "Algorithmic Cheminformatics" brought together leading researchers from both chemistry and computer science. The seminar was the second in a series of the Dagstuhl seminars and had a focus on concurrency theory as chemical systems are highly concurrent by nature. Within computer science we focused on formal approaches for chemistry and concurrency theory, including process calculi and Petri nets. The participants surveyed areas of overlapping interests and identified possible fields of joint future research.

Seminar November 5–10, 2017 – http://www.dagstuhl.de/17452
1998 ACM Subject Classification F.1.2 Modes of Computation, F.4.2 Grammars and Other Rewriting Systems, F.4.3 Formal Languages, G.2.2 Graph Theory, J.2 Physical Sciences and Engineering
Keywords and phrases Modelling, Simulation, Networks, Semantics / Formal Methods
Digital Object Identifier 10.4230/DagRep.7.11.28

1 Executive Summary

Jakob L. Andersen (Tokyo Institute of Technology, JP)
Christoph Flamm (Universität Wien, AT)
Daniel Merkle (University of Southern Denmark – Odense, DK)
Peter F. Stadler (Universität Leipzig, DE)

Cheminformatics is the application of algorithms, combinatorial approaches, and formal methods from Computer Science to problems in Chemistry. While being formally a very old research field, building the theoretical foundations for Cheminformatics seen from the perspective of state-of-the-art theoretical Computer Science is not at all established research. The second edition of the seminar on "Algorithmic Cheminformatics" brought together researchers working in Chemistry, Cheminformatics, and most importantly the relevant fields in Computer Science related to it. In contrast to the first Dagstuhl meeting in 2014, we specifically focused on the analysis of the behaviour of chemical systems in terms of reaction networks. This includes both networks inferred from experimental data, as well as networks implicitly specified by for example formal grammars. We integrated experts in concurrency theory, in particular using process calculi, Petri nets, and related formal approaches. State-of-the-art results in these fields are hardly used to infer qualitative and/or quantitative properties of chemical reaction systems, which are highly concurrent systems

Algorithmic Cheminformatics, *Dagstuhl Reports*, Vol. 7, Issue 11, pp. 28–45
Editors: Jakob L. Andersen, Christoph Flamm, Daniel Merkle, and Peter F. Stadler

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

by nature. Most current modeling approaches in chemistry are either very abstract and aimed at formal algebraic properties of reaction networks, or use precise modeling on a very fine grained level such as the quantum mechanical one where computational costs prevent handling of more than a few molecules. In this seminar we therefore seeked to advance discrete modeling approaches for Systems Chemistry. In addition to bringing together the experts in the respective fields from Computer Science, we also invited wet-lab chemists in order to cross-fertilize the fields and generate mutually beneficial activities.

2 Table of Contents

3 Overview of Talks

3.1 Chemical Reaction Networks, Pathways, and Realisability

Jakob L. Andersen (Tokyo Institute of Technology, JP)

Joint work of Jakob L. Andersen, Christoph Flamm, Daniel Merkle, Peter F. Stadler

We present a modelling framework for pathways in chemical reaction neworks. The model is based on a generalisation of network flows to directed multihypergraph, and a restriction to integer hyperflows allows for analysis of pathwys on a mechanistic level. We addiionally introduce a set of necessary constraints for a pathway to be considered either catalytic or autocatalytic. Using ideas from Petri nets we then analyse the realisability of pathways from a concurrency perspective.

3.2 Bio-Model-Kit – A Framework for BioModel-Engineering

Mary Ann Blätke (IPK Gatersleben, DE)

Representing the complexity of biomolecular components in a consistent and straightforward way is one of the challenges in biomodel engineering. We developed a modular modelling framework for biomodel engineering, called BioModelKit, addressing this problem. In the BioModelKit framework, modules are designed for the purpose of model composition. A module is a molecule-centred Petri net model representing the functionality and interactions of a biomolecule. Defined interfaces shared among the modules allow the modular composition of models without manual adjustments. We distinguish between mechanistic and causal module types, to capture the different types of networks and omic fields. Each module includes an annotation file holding descriptive information and cross-links to related references. The algorithmic model mutation of composed models allows mimicking the effect of gene knock-outs and structural mutations of a biomolecule. Modularly composed models can also be transformed into spatial models to represent the movement of biomolecules, the cell geometry, compartments, and the distribution of molecules. Modules can be constructed using direct and reverse engineering approaches. Existing models, e.g SBML or Boolean models, as well as OMIC data can be another source to obtain modules. The modular model composition and its extensions are implemented in a web-based tool that is supported by a MySQL database holding the Petri net graphs and annotations of submitted modules. The BioModelKit framework mentioned extensions overs a versatile and unifying tool for biomodel engineering.

3.3 Introduction to Process Algebras

Cinzia Di Giusto (Laboratoire I3S – Sophia Antipolis, FR)

We introduce the main concepts behind process algebras.

Process Algebras are mathematical formalism that allow the specification, the analysis and the verification of properties of concurrent and distributed systems. Processes are agents that work in parallel, exchange informations and take decisions depending on the acquired informations. The semantics of such processes is given in terms of structural operational semantics (SOS) rules and labelled transition systems are built and composed by using the different operators of the given process algebra.

In order to compare processes, behavioural equivalences are used to abstract from unwanted details and identify those processes that behaves "similarly". We introduce trace and bisimilarity equivalences.

Finally we briefly present behavioural types which specifically describe protocols in distributed systems. The type system ensures well-typed processes to enjoy properties such as communication safety, protocol fidelity, and progress.

3.4 On the way to Synthetic Biology: Enzyme Cascade Optimisation at the University of Greifswald Robotic Platform LARA

Mark Dörr (Universität Greifswald, DE)

Main reference M. Dörr, M. P. C. Fibinger, D. Last, S. Schmidt, J. Santos-Aberturas, D. Böttcher, A. Hummel, C. Vickers, M. Voss, U. T. Bornscheuer: "Fully Automatized High-Throughput Enzyme Library Screening Using a Robotic Platform", Biotechnol. Bioeng. , 113 (7), 1421–1432, 2016.
URL http://dx.doi.org/10.1002/bit.25925

Automated evolutionary experiments with (ultra) high-throughput in synthetic biology and protein design are state-of-the-art approaches in the bio-chemical laboratory. They generate new challenges on the technical realisation, e.g., detection of very tiny samples and their changes over time, and on the computational side, e.g., data evaluation of the large amounts of data points, machine learning and "machine based understanding" of the underlying experiments and finally closed feedback cycles of automated evolution. We demonstrated our different technical approaches, like our robotic protein screening platform LARA (lara.uni-greifswald.de) [2] for macroscopic samples. And microfluidic set-ups utilizing millions of nano- and picoliter sized droplets with individual chemical composition (Prof. Bornscheuer lab, Biochemistry, Uni-Greifswald). Droplet formation and double droplet formation (water-in-oil-in water emulsions) has been explained. The experimental challenges of "Synthetic Biology" and the introduction of new genes that are completely foreign to a species have been demonstrated by the example of the oligo-epsilon-caprolactone production in E.coli [1], [2]). Tuning different protein expression levels can be addressed by splitXFP systems (XFP: colour X Fluorescent Protein) [3]. This is an important tool to adjust a good substance flux through a synthetic metabolic pathway. Furthermore protein long-term stability is an issue that was addressed by automated temperature selection experiments.

References

1 S. Schmidt, C. Scherkus, J. Muschiol, U. Menyes, T. Winkler, W. Hummel, H. Gröger, A. Liese, H.-G. Herz, U. T. Bornscheuer. *An Enzyme Cascade Synthesis of ϵ-Caprolactone and Its Oligomers*. Angew. Chem. Int. Ed. 2015, 54 (9), 2784–2787.

2 M. Dörr, M. P. C. Fibinger, D. Last, S. Schmidt, J. Santos-Aberturas, D. Böttcher, A. Hummel, C. Vickers, M. Voss, U. T. Bornscheuer. *Fully Automatized High-Throughput Enzyme Library Screening Using a Robotic Platform*. Biotechnol. Bioeng. 2016, 113 (7), 1421–1432.

3 J. Santos-Aberturas, M. Dörr, G. S. Waldo, U. T. Bornscheuer. *In-Depth High-Throughput Screening of Protein Engineering Libraries by Split-GFP Direct Crude Cell Extract Data Normalization*. Chemistry and Biology 2015, 22 (10), 1406–1414.

3.5 Reduction of Models of Intra-Cellular Signaling Pathways

Jérôme Feret (ENS – Paris, FR)

We use the flow of information between the states of sites in bio-molecular compounds in rule-based models of signaling pathways, so as to identify which correlations have a real impact on the dynamics of the models. The others can be safely abstracted away which leads to a factorisation of the underlying differential systems. As a result, we get a fully automatic procedure to generate a reduced ODE systems, without having to consider the ODE equations of the initial network.

3.6 Introduction to recent developments in Computational and Systems Chemistry

Christoph Flamm (Universität Wien, AT)

Over the last decade the focus of experimental approaches in systems chemistry and the field of origin of life research has shifted from the study of single molecules with close structural proximity to biological building blocks to the high-throughput analysis of complex mixtures of molecules which often possess little to no direct biological relevance. This development, i.e. the manipulation of diverse molecular mixtures, is not mirrored by theoretical advances in computational chemistry. Novel computational approaches capable of handling concurrency and causality in complex reactive mixtures are required to deconvolute the intricate chain of events that resulted in the experimental observation. Computational approaches could be even further exploited as unbiased discovery tools for novel chemical mechanisms instead of only arbitrating between competing experimental hypothesis. The current excitement in deep machine learning approaches does unfortunately as-well not carry over to chemistry since most of the chemical knowledge is sealed away in proprietary databases and not accessible on the large scale in the public domain. The problematic sketched above will be discussed in depth in my presentation illustrating certain aspects with experimental and computational examples from recent literature.

3.7 Behavioural Model Checking of Dynamic systems, with a Focus on Reaction Networks in a Multiscale Scenario

David Gilbert (Brunel University – Uxbridge, GB)

I describe how model checking is used over the dynamic behaviour of models of biochemical reaction systems, which can be modelled in Petri nets, but also in other formalisms, and uses descriptions in temporal logic. The approach can also be applied to time-series laboratory data. Our approach can be applied to model checking the behaviours of very large models, for example of whole genome metabolism of E.coli, using libraries of properties of interest of both metabolite concentrations as well as of reaction activity over time.

I also show how this behaviour-based model checking can be applied to multi-level [and multi-scale] models which have been constructed using Coloured Petri nets, with an example of planar cell polarity in Drosophila fly wing.

I also show how temporal logic descriptions can be automatically derived from sets of examples.

Finally I put this methodology in the context of design for synthetic biology, and within a multiscale scenario, and speculate how this could be relevant to cheminformatics.

References

1 D. Gilbert, M. Heiner, Y. Jayaweera, C. Rohr. *Towards dynamic genome-scale models*. Briefings in Bioinformatics, 2017. https://doi.org/10.1093/bib/bbx096
2 O. Parvu, D. Gilbert. *A Novel Method to Verify Multilevel Computational Models of Biological Systems Using Multiscale Spatio-Temporal Meta Model Checking*. PloS one 11 2016:(5). https://doi.org/10.1371/journal.pone.0154847
3 Q. Gao, D. Gilbert, M. Heiner, F. Liu, D. Maccagnola and D. Tree. *Multiscale Modelling and Analysis of Planar Cell Polarity in the Drosophila Wing*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012:10:2. http://doi.org/10.1109/TCBB.2012.101
4 D. Maccagnola, E. Messina, Q. Gao and D. Gilbert. *A Machine Learning Approach for Generating Temporal Logic Classifications of Complex Model Behaviours*. In Proceedings of WSC 2012 (Winter Simulation Conference December 9-12, 2012, Berlin).

3.8 Parameterizing Rule-based Systems

Harold Fellermann

Joint work of Harold Fellermann, Annunziata Lopiccolo, Benjamin Shirt-Ediss

Rule-based modelling is a powerful technique for the simulation and analysis of complex chemical and molecular biological systems. Its strength results from its ability to capture a polynomial number of reactions among a combinatorial set of reactants in few rules that model general interactions. Parameterizing rule-based systems with experimental data is an important yet relatively little explored field.

An example of a chemical system that lends itself to rule-based modelling is our current work toward a molecular memory devices (stack data structure) based on DNA hybridization

and strand displacement reactions [1]. Parallel to our experiments, we develop rule-based models that capture device operation with three to four rules. In order to use such models for hypothesis testing, thorough parameterization of reaction rates is paramount.

After briefly summarizing approaches toward parameter optimization, I emphasize the difficulties of systematically performing parameter fits of rule-based models against experimental data: namely (a) proper treatment of stochasticity, (b) potential infinity of the system's state space, (c) generalization of rate constants to context-dependent rate functions, (d) difficulty in quantitatively comparing the model state space against experimental electrophoresis data.

These difficulties are the motivation for a semi-automated qualitative fitting procedure of the models: with an interactive interface, the modeller can explore the impact that parameter choices have on a simulated experiment by visually comparing the experimental images to virtual electrophoresis images generated from model predictions.

We demonstrate that this approach leads to literature consistent parameters and a good qualitative fit. Importantly, the procedure can be used to support hypothesis testing if alternative models are probed against the same set of data.

References

1 H. Fellermann, A. Lopiccolo, J. Kozyra, N. Krasnogor. *In vitro implementation of a stack data structure based on DNA strand displacement* in Martyn Amos *Unconventional Computation and Natural Computation* pp. 87–89. Springer, Heidelberg, 2016.

3.9 How Might Petri Nets Enhance Your Systems Biology Toolkit

Monika Heiner (BTU Cottbus, DE)

In my talk I will give a Petri net perspective on the exciting research field of "Biomodel Engineering".

Petri nets enjoy an intuitive graphical representation and formal semantics alike, thus they are a popular notation for biochemical reaction networks, such as gene regulatory, signal transduction or metabolic networks.

In our scenario, Petri nets serve as umbrella formalism combining different modelling paradigms, where each perspective contributes to a better understanding of the biochemical system under study. In this spirit of BioModel Engineering, we developed over the last two decades our unifying Petri net framework comprising the traditional time-free Petri nets (PN) as well as quantitative, i.e. time-dependent Petri nets such as stochastic Petri nets (SPN – opening the door to the Markovian world), continuous Petri nets (CPN – opening the door to the world of ordinary differential equations), and hybrid Petri nets (HPN – combining the previous two formalisms), as well as their their coloured counterparts. The crucial idea is to have a family of related models, sharing structure, but differing in their kinetic details.

Coloured Petri nets permit, among others, the convenient and flexible encoding of spatial attributes, and thus the modelling of processes evolving in time and space, which are usually treated as stochastic or deterministic partial differential equations (PDE). In our approach, the discretisation of space happens on the modelling level, while traditionally the discretisation is left for the PDE integration methods.

Our framework is supported by a related Petri net toolkit comprising Snoopy, Charlie and Marcie. It has been applied to a couple of case studies.

For more details, please see http://www-dssz.informatik.tu-cottbus.de/BME/BME-tutorial

References

1 M. Heiner and D. Gilbert. *How Might Petri Nets Enhance Your Systems Biology Toolkit*. In Proc. PETRI NETS 2011, Springer, LNCS 6709, pp. 17–37, 2011. https://doi.org/10.1007/978-3-642-21834-7_2

2 D. Gilbert, M. Heiner, Y. Jayaweera and C. Rohr. *Towards dynamic genome scale models*. Briefings in Bioinformatics, 2017. https://doi.org/10.1093/bib/bbx096

3.10 Autocatalytic Sets and Chemical Organizations

Wim Hordijk (Konrad Lorenz Institut – Klosterneuburg, AT)

Joint work of Hordijk, Wim; Steel, Mike; Dittrich, Peter
Main reference W. Hordijk, M. Steel, P. Dittrich, "Autocatalytic sets and chemical organizations: Modeling self-sustaining reaction networks at the origin of life", New Journal of Physics 20:015011, 2018.
URL https://doi.org/10.1088/1367-2630/aa9fcd

Autocatalytic sets are self-sustaining reaction networks in which all molecules mutually catalyze each other's formation from a basic food source. They are believed to have played an important role in the origin of life. Much mathematical and computational work has been done on autocatalytic sets in the form of RAF theory [1], but they have also been constructed experimentally, either with RNA molecules [2] or with peptides [3].

One result coming out of RAF theory is that often such self-sustaining sets consist of a hierarchical structure of smaller autocatalytic subsets [4]. From a dynamical point of view, the so-called "closed autocatalytic sets" are of primary interest, as they represent the dynamically stable states that are observed in stochastic dynamic simulations of autocatalytic sets.

A related formalism for modeling self-maintaining chemical reaction networks is that of chemical organization theory (COT) [5]. Recently, we have shown several formal relationships between RAF theory and COT, in particular between closed autocatalytic subsets and chemical organizations [6]. Moreover, this close relationship leads to a precise method to enumerate all closed subsets within an autocatalytic set by calculating all its chemical organizations.

In my talk I presented the main ideas behind closed autocatalytic sets, how they play an important role in the stochastic dynamics of autocatalytic sets, and how they can be enumerated using chemical organization theory.

References

1 Hordijk and Steel. BioSystems 152:1–10, 2017.
2 Vaidya et al. Nature 491:72–77, 2012.
3 Ashkenasy et al. PNAS 101:10872–10877, 2004.
4 Hordijk et al. Acta Biotheoretica 60:379–392, 2012.
5 Dittrich and Speroni di Fenizio. Bulletin of Mathematical Biology 69:1199–1231, 2007.
6 Hordijk et al. New Journal of Physics 20:015011, 2018.

3.11 Computing Optimal Synthesis Plans

Rojin Kianian (University of Southern Denmark – Odense, DK)

Joint work of Rolf Fagerberg, Christoph Flamm, Rojin Kianian, Daniel Merkle, Peter F. Stadler

In synthesis planning, the goal is to synthesize a target molecule from available starting materials, possibly optimizing costs such as price or environmental impact of the process.

We demonstrate that synthesis planning can be phrased as a combinatorial optimization problem on hypergraphs by modeling individual synthesis plans as directed hyperpaths embedded in a hypergraph of reactions (HoR) representing the chemistry of interest. As a consequence, a polynomial time algorithm to find the K shortest hyperpaths can be used to compute the K best synthesis plans for a given target molecule. Having K good plans to choose from has many benefits: it makes the synthesis planning process much more robust when in later stages adding further chemical detail, it allows one to combine several notions of cost, and it provides a way to deal with imprecise yield estimates.

Our modeling is not restricted to bond set based approaches, which are otherwise the most widespread—any set of known reactions and starting materials can be used to define a HoR, as long as the reactions are modeled with only one product.This generalization enables computation of synthesis plans using the collected knowledge of chemists, provided by Reaxys. In this way, the expert knowledge of reactions that are well tested in the lab may be combined into plans in ways that have not been thought of before and that are not biased by tradition.

3.12 Graph Transformation, Rule Composition, and Stereochemistry

Daniel Merkle (University of Southern Denmark – Odense, DK)

Joint work of Jakob L. Andersen, Christoph Flamm, Daniel Merkle, Peter F. Stadler
Main reference Jakob L. Andersen, Christoph Flamm, Daniel Merkle, Peter F. Stadler: "Chemical Transformation Motifs – Modelling Pathways as Integer Hyperflows", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. PP(99), IEEE, 2017.
URL http://dx.doi.org/10.1109/TCBB.2017.2781724

Graph transformation form a natural model for chemical reaction systems and provide a sufficient level of detail to track individual atoms. Among alternative graph transformation formalisms the Double Pushout approach, which is firmly grounded in category theory, is particularly well-suited as a model of chemistry. The formal foundations of defining composition of transformation rules using ideas from concurrency theory are presented. In addition of a generic construction several special cases that each have an intuitive chemical interpretation will be considered. The usefulness of these specialised operations by automatically calculating coarse-grained transformation rules for complete chemical pathways, that preserve the traces of atoms through the pathways, will be illustrated. This type of computation has direct practical relevance for the analysis and design of isotope labelling experiments. Examples will be illustrated with MØD, a software package developed for graph-based cheminformatics.

References

1 J. L. Andersen, C. Flamm, D. Merkle, P. F. Stadler. *Chemical Transformation Motifs – Modelling Pathways as Integer Hyperflows*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2018.

2 J. L. Andersen, C. Flamm, D. Merkle, P. F. Stadler. *Rule Composition in Graph Transformation Models of Chemical Reactions*. MATCH Communications in Mathematical and in Computer Chemistry, 2018, to be published.

3.13 The Symbolic Method & Random Generation

Markus E. Nebel (Universität Bielefeld, DE)

Joint work of Daniel Merkle, Markus Nebel

In this talk we show how methods based on generating functions and complex analysis can be used to count and randomly generate various kinds of molecules taking stereoisomers into account. Based on so-called symbolic equations for the recursive structure of the molecules, precise asymptotic formulae for the number of different conformations of given size (number of carbon atoms) can be derived. Furthermore, the underlying recursive description together with the resulting generating functions can be used to efficiently generate random molecule of given size according to uniform and non-uniform probability distributions. We discuss fixed size as well as approximate size (Boltzmann) sampling approaches.

3.14 Behavioural Equivalences in a Nutshell

Marco Peressotti (University of Southern Denmark – Odense, DK)

Two of the fundamental questions driving research in Concurrency Theory are "what is a process?" and "when is the behaviour of two processes equivalent?". Although a definitive answer is still elusive, remarkable progress has been made in the last three and a half decades. A cornerstone of the theory and perhaps one of the most illustrative example, of behavioural equivalence is Milner and Park's bisimulation. This notion was initially proposed for the study of non-deterministic processes and, as this theory matured, more detailed models of concurrent systems have been developed, some including quantitative data such as cost, probabilities, and time. Today, these techniques are widely used in computer science, but also in other fields, including artificial intelligence, cognitive science, mathematics, modal logics, philosophy, and physics; mainly to explain phenomena involving some kind of circularity or infinite object.

The talk provides an introduction to the theory of behavioural equivalences with special attention to bisimulation and coinduction.

3.15 Exploring the Reaxys network of chemical reactions: Computational history of chemical reactions

Guillermo Restrepo (Universität Leipzig, DE)

The chemical community devotes most of its efforts to synthetic chemistry, therefore knowledge about reactants, catalysts, solvents and several other related aspects of chemical reactions is of important relevance. Part of this knowledge is its history that involves determining the aspects that have shaped chemical reactions to their current state; which are tasks for the history of chemistry. However, analysing the chemical reactions that have been reported in the scientific literature is not any more a subject of the conventional history of chemistry, for the number of substances and reactions grows exponentially. Here we show that a computational approach to the history of chemical reactions sheds light on the patterns behind the development and use of substances and reaction conditions. We explored the more than 45 million reactions gathered in Reaxys database and came across with historical patterns for substances, types of substances, catalysts, solvents, temperatures and pressures of those reactions. It is found that chemists have traditionally used few reactants to produce many different substances. In such synthesis more combinations of about four chemical elements are explored. Despite the exponential growth of substances and reactions, little variation of catalysts, solvents, and reactants is observed throughout time. Regarding reaction conditions, the vast majority of reactions fall into a narrow domain of temperature and pressure, namely normal conditions. We also found and quantified the effect of world wars (WWs) upon chemical novelty during war periods. WW1 took production of new substances and reactions back around 30 years and WW2 around 15. We anticipate this study and especially its methodological approach to be the starting point for the computational history of chemical reactivity, where social and economical contexts are integrated.

3.16 Open Science and Chiminey a Collaborative Platforms for Biochemistry

Heinrich W. Schmidt (RMIT University – Melbourne, AU)

Scientific experiments and engineering modeling increasingly include data and software. A so-called fourth paradigm of scientific discovery has emerged based on these. It enables researchers to accelerate the three other paradigms, empirical, theoretical and computing by learning models from data and by software-defined workflows, sharing of data and collaborating globally at a scale not possible just a few decades ago. We developed Chiminey with quantum physicists and molecular biologists, to ease scaling up stochastic physics and chemistry algorithms from small data sets and laptop use to run straight on massive numbers of nodes in high-performance compute centres or the Australian Open Science Cloud.

This talk draws from two presentations, the first providing some broader insights into the Open Science movement with its FAIR, i.e., findable, accessible, interoperable and reusable resources (data and software) and community support; the second drilling down into specifics of human-centric parallel stochastic modelling with Chiminey.

Some aspects of Chiminey can be found here http://doi.org/10.1016/j.bdr.2017.01.004 and here http://github.com/chiminey/chiminey.

3.17 Brief Comments on Origin of Life Iinterests in Computational Cheminformatics

D. Eric Smith (Santa Fe Institute, US)

The task is to go from the beginning of order in near-surface geochemistry, to the pruned and hierarchical order of biochemistry.

In this short presentation I review the properties of biochemistry that suggest that certain chemistry is easier than other, at either the mechanistic or the network level, and that this has entrained the large-scale structure of metabolism and evolution.

A second topic is the role of metals in biochemistry and in our inference about the earliest relevant geochemistry, and the progress and goals for incorporating them into graph-grammar systems.

In order to try to find a relevant sense of causation for the chemistry and topology of core metabolism, drawing on the above evidence, however, we would need to know how readily the same properties that suggest an easiest or best solution (reaction redundancy, network completion, autocatalysis, etc.), which exist in actual biochemistry, would be found at arbitrary other points in the space of possible planetary chemical precursor systems. This is needed to disambiguate the confound between a pattern that is laid down in geochemistry as s kind of chemical path of least resistance, versus a frozen accident of evolutionary dynamics in a later era.

The combination of being able to model metals as gateways, at the elementary reaction level, and to simulate very large networks and to efficiently query for network properties that putatively aid evolution or stability, is the advantage that graph grammars seem most likely to promise.

A particular goal in the use of metals is to systematically search for catalysis-related phenomena – once they are demonstrated to exist in some context – for other contexts in which the same phenomena may be found. We might want to follow what has successfully been done in small-molecular organocatalysis, to abstract mechanism so that we can see when the same fundamental mechanism can be achieved at all levels from free ions, to soluble metal-ligand complexes, to mineral unit cells, and ultimately to metal-center cofactors or enzymes that have been tuned by evolution. When common mechanisms exist, this provides a path for continuity in which catalysts can be replaced and evolved, while leaving the overall chemical synthetic network in place from prebiotic contexts all the way to extant biochemistry.

3.18 CRNs: A Contact Point Between Cheminformatics and Recent Topics of Interest in the NEQ Statistical Physics Community

D. Eric Smith (Santa Fe Institute, US)

This presentation had three parts. The first was a brief review of the problem framework known as Chemical Reaction Network theory, and a few of its major results (mostly in convex analysis) from the deterministic methods begun in the 1970s. The second part was an introduction to some areas of interest in modern non-equilibrium statistical mechanics, particularly the question of how much about dynamics can be understood by reference to the local-equilibrium entropy, and what entropy concept takes its place when even local conditions are distant from equilibrium. The third part was a small worked example, showing how the methods of Integer Linear Programming that are common in the topological analysis of CRNs and Petri Nets, can be used to characterize the transport of probability in what are known as "complex-balanced" steady states, the largest class of non-equilibrium systems for which the major elements of Equilibrium Thermodynamic structure are retained.

The worked example illustrated three results in the relation between topology and the probabilistic dynamics of CRNs:

The first is that complex-balanced flows preserve a combinatorial density of states which is a unifying feature of conventional Gibbsian thermal equilibria. The second is that complex-balanced flows preserve a factorization between the Large-Deviations behavior of the environment and that of the system, which is expressed in the property that all steady states are exponential tilts of one another, or of an invariant combinatorial density of states. The third result is that under the same conditions, probability flow at steady states mirrors the flow of the Balanced Integer Hyperflows. This redundant situation, in which all transport quantities are governed by the same small collection of integer coefficients, may be understood as another expression of a well-known fact (proved by Anderson, Craciun, and Kurtz) that the steady states can be written as product-Poisson densities (or slices through them), which means that they are "minimal-information" distributions, with all higher-order moment behaviors dictated in an invariant manner by the values of the mean particle numbers.

3.19 Bisimulations for Differential Equations

Mirco Tribastone (IMT – Lucca, IT)

Large-scale dynamical models hinder our capability of effectively analyzing them and interpreting their behavior. We present an algorithm for the simplification of polynomial ordinary differential equations by aggregating their variables. The reduction can preserve observables of interest and yields a physically intelligible reduced model, since each aggregate corresponds to the exact sum of original variables.

Tool implementation: http://sysma.imtlucca.it/tools/erode/

3.20 Open Instrumentation

Klaus-Peter Zauner (University of Southampton, GB)

Joint work of Martin Fischlechner, Jonathan West, Klaus-Peter Zauner
URL https://dropletkitchen.github.io/

A brief overview of recent progress in artificial chemical networks based on lipid coated Belousov Zhabotinsky (BZ) droplets in oil showed the use of laser-cut acrylic templates to position hundreds of BZ droplets in desired network configurations [1]. From this simple application of digital fabrication for custom laboratory tools the presentation focused on the potential of the confluence of low-cost embedded systems and modern fabrication techniques such as 3D printing and laser-cutting for the fast evolution of scientific instrumentation.

Novel directions in experimentation often require highly specialized custom made equipment with a concomitant high cost. As a consequence both access to the new technique and the number of allowed design iterations for the instrument are typically severely limited. This situation is about to change and the impact for experimental work is likely to resemble the impact the arrival of the personal computer in the laboratory had with moving data analysis from punch cards on a mainframe to a spread sheet on the researchers desk. A large number of software tools subsequently originated because scientists could iteratively develop themselves the tool that answered their idiosyncratic requirements. By sharing these tools as open source software they were refined by the community to a level that would be unaffordable for any particular research programme and at the same time the tools are widely available.

At present we are at the verge of a similar development with regard to laboratory instruments. Very capable embedded systems that can be programmed with ease[1] complement the flexibility of digital fabrication. The convenience of interfacing with actuators and sensors makes it possible to forsake precision in the components while maintaining overall systems performance through closed-loop control. This in turn makes it possible to use affordable 3D printers and place them directly in the laboratory where the products are used, thus facilitating a cycle where deficiencies in a tool or instrument can result in a modification of the instrument within days, if not hours. By making the files required for the fabrication of the mechanical and electronic hardware, as well as the software available this feedback cycle can transcend the laboratory; and arguably will lead to the refinement of idiosyncratic laboratory tools as was the case with scientific open source software.

Some examples of open instruments–in the sense that the software and hardware plans required to replicate the instruments are available and can be modified and redistributed–for use with microfluidic chips are available[2]; for a collection of links to various instruments see [3].

Aside from the faster development of the instruments, the low cost of open instrumentation offers additional advantages. Equipment is documented in detail which supports both the reproduction of experiments and the evaluation of results. Users can have their personal instrument and whole class sets can be produced instead of sharing a single piece of equipment. Universities and schools with limited funding can access state-of-the art methods. Experiments can be replicated in parallel at a scale that might otherwise not be fundable. In our own research we hope to exploit the adaptability of open instrumentation to develop autonomous experimentation with microfludics and flow chemistry experiments.

[1] joint work with Kai Ming Chang and Maurits R.R. de Planque

References

1 D. P. George, 2014 *MicroPython*. https://micropython.org/

2 M. Fischlechner, J. West, K.-P. Zauner, 2016 *DropletKitchen*. https://dropletkitchen.github.io/

3 Pearce, Joshua, 2013 *Building research equipment with free, open-source hardware*. http://www.appropedia.org/Building_research_equipment_with_free,_open-source_hardware

Participants

- Jakob Lykke Andersen
Tokyo Institute of Technology, JP
- Sarah Berkemer
MPI für Mathematik in den Naturwissenschaften, DE
- Mary Ann Blätke
IPK Gatersleben, DE
- Erick Chastain
Rutgers University – Piscataway, US
- Cinzia Di Giusto
Laboratoire I3S – Sophia Antipolis, FR
- Peter Dittrich
Universität Jena, DE
- Mark Dörr
Universität Greifswald, DE
- Rolf Fagerberg
University of Southern Denmark – Odense, DK
- Harold Fellermann
University of Newcastle, GB
- Jérôme Feret
ENS – Paris, FR
- Christoph Flamm
Universität Wien, AT
- Walter Fontana
Harvard Medical School – Boston, US
- David Gilbert
Brunel University – Uxbridge, GB
- Monika Heiner
BTU Cottbus, DE
- Marc Hellmuth
Universität Greifswald, DE
- Wim Hordijk
Konrad Lorenz Institut – Klosterneuburg, AT
- Rojin Kianian
University of Southern Denmark – Odense, DK
- Daniel Merkle
University of Southern Denmark – Odense, DK
- Markus E. Nebel
Universität Bielefeld, DE
- Nikolai Nøjgaard
Universität Greifswald, DE
- Marco Peressotti
University of Southern Denmark – Odense, DK
- Guillermo Restrepo
Universität Leipzig, DE
- Heinrich W. Schmidt
RMIT University – Melbourne, AU
- D. Eric Smith
Santa Fe Institute, US
- Peter F. Stadler
Universität Leipzig, DE
- Mirco Tribastone
IMT – Lucca, IT
- Klaus-Peter Zauner
University of Southampton, GB