Report from Dagstuhl Seminar 14452

Algorithmic Cheminformatics

Edited by

Wolfgang Banzhaf[1], Christoph Flamm[2], Daniel Merkle[3], and Peter F. Stadler[4]

1 **Memorial Univ. of Newfoundland, CA,** banzhaf@cs.mun.ca
2 **Universität Wien, AT,** xtof@tbi.univie.ac.at
3 **University of Southern Denmark – Odense, DK,** daniel@imada.sdu.dk
4 **Universität Leipzig, DE,** peter.stadler@bioinf.uni-leipzig.de

Abstract

Dagstuhl Seminar 14452 "Algorithmic Cheminformatics" brought together leading researchers from both chemistry and computer science. The meeting successfully aimed at bridging in the apparent gap between the two disciplines. The participants surveyed areas of overlapping interests and identified possible fields of joint future research.

Seminar November 2–7, 2014 – http://www.dagstuhl.de/14452
1998 ACM Subject Classification 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 Graph Transformation Systems, Graph and Hypergraph Invariants, Graph Comparison, Network Flows, Hypergraphs, Formal Languages, Algebraic Chemistry
Digital Object Identifier 10.4230/DagRep.4.11.22

1 Executive Summary

Wolfgang Banzhaf
Christoph Flamm
Daniel Merkle
and Peter F. Stadler

Dagstuhl Seminar 14452 "Algorithmic Cheminformatics" was organized to intensify the interactions between chemistry and computer science. While the thriving field of bioinformatics/computational biology is a success story of a lively and extensive inter- and trans-disciplinary collaboration between life sciences and computer science, this is much less so in cheminformatics. After a quick raise of a plethora of computational approaches for chemical problems in the 1960–1970s, the field mainly settled down on machine learning approaches in the late 1990s. Over last two decades, computer science plays a comparably marginal role in chemistry research and education.

This is a puzzling state of affairs as chemistry, and in particular the emerging field of systems chemistry, has to offer a wide range of non-trivial computational problems that are very different from those in the well-established areas of quantum chemistry, molecular dynamics, or physical chemistry, for which physics-style models and numeric mathematics have been established as the methods of choice. In particular, complex chemical networks capable of algorithmic self-assembly under far-from-equilibrium conditions, seem to possess a deep connection to the theory of computation, information recoding and compiler theory.

Algorithmic Cheminformatics, *Dagstuhl Reports*, Vol. 4, Issue 11, pp. 22–39
Editors: Wolfgang Banzhaf, Christoph Flamm, Daniel Merkle, and Peter F. Stadler

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

Dagstuhl Seminar 14452 therefore specifically aimed to establish the connection between theoretical computer science, graph theory and related fields of discrete mathematics, and complexity theory on the one hand and chemistry on the other hand. Several key areas where covered by one or more presentation and extensive discussions among the participants. Topic ranged from formalizing chemical transformations, autocatalytic molecular systems, and the design of chemical experiments, via model checking and key graph algorithm, to chemical information technology and models for the origin of life. Dagstuhl Seminar 14452 successfully brought together wet-lab chemists with theoretical computer scientists and researchers with a focus on bioinformatics and initiated an, as we feel, very fruitful frist step towards cross-boundary research.

2 Table of Contents

3 Overview of Talks

3.1 Composition of Transformation Rules

Jakob Lykke Andersen (University of Southern Denmark – Odense, DK)

We use labelled graphs and labelled graph transformation rules to model molecules and reactions patterns. This enables the generation of large chemical spaces from concise descriptions, while still maintaning an explicit model of each atom and bond. Here we present the general concept of composing transformation rules, which we as example use for increasing the practical performance of the algorithms for generating reaction networks. Using rule composition as a fundamental operation we furthermore illustrate how atom-tracing via isotope labelling can be modelled, exemplified by two chemical systems: the enzyme mechanism for beta-lactamose, and the two glycolysis pathways Embden-Meyerhof-Parnas and Entner-Doudoroff.

3.2 Structure formation of peptide foldamers from first principles

Carsten Baldauf (FHI – MPG Berlin, DE)

Joint work of Schubert, Franziska; Pagel, Kevin; Rossi, Mariana; Warnke, Stefan; Salwiczek, Mario; Koksch, Beate; von Helden, Gert; Blum, Volker; Baldauf, Carsten; Scheffler, Matthias
Main reference F. Schubert, K. Pagel, M. Rossi, S. Warnke, M. Salwiczek, B. Koksch, G. von Helden, V. Blum, C. Baldauf, M. Scheffler, "Native like helices in a specially designed β peptide in the gas phase," Physical Chemistry Chemical Physics, 17:5376–5385, 2015.
URL http://dx.doi.org/10.1039/C4CP05216A

We study the structure formation and dynamics of peptides and peptide foldamers using first-principles methods,[1] specifically we employ density-functional theory (DFT) corrected for van der Waals interactions. Navigating the conformational space of such flexible (bio-)oligomers is a challenge in itself that we currently tackle with force field based pre-sampling (with basin hopping or replica-exchange molecular dynamics) and then complement with extensive DFT calculations. We compare our results to actual gas-phase experiments, i. e., ion mobility mass spectrometry and vibrational spectroscopy (especially to work by K. Pagel at FU Berlin and G. von Helden at FHI Berlin). I would like to cover three topics in my contribution: With peptides that feature central prolyl-peptide bonds and that model β-turns, we studied the effect of monovalent cations on the structure formation. Cations locally disrupt the hydrogen-bonding network and enforce, by favorable electrostatic interactions, otherwise not observed conformations on the peptide's backbone.[2] Helix formation of peptides Ac-Alan-LysH+ in the gas phase has been studied for years now.[3] We added a new direction by studying the effect of increased backbone flexibility on the helix forming properties. For that the β-peptide Ac-(β2hAla)6-LysH+ was designed and investigated. We demonstrated for the first time that β-peptides from acyclic monomers can form native-like helices (similar to 310, α, π). At the same time, the stability order of the three helix types seems to be inverted with respect to their natural α-peptide counterpart. Last I would like to briefly introduce our efforts towards a conformational search and sampling approach that is entirely based on DFT and avoids the use force fields. The performance of the

genetic-algorithm search is assessed by comparison to data for capped amino acids (in house reference data, to be published) and two non-natural α/γ hybrid peptides.[4]

References

1 C. Baldauf, H.-J. Hofmann. Ab initio MO Theory – An Important Tool in Foldamer Research: Prediction of Helices in Oligomers of ω-Amino Acids. *Helv. Chim. Acta* 2012 (95), 2348–2383.
2 C. Baldauf, K. Pagel, S. Warnke, G. von Helden, B. Koksch, V. Blum, M. Scheffler. How Cations Change Peptide Structure. *Chem. Eur. J.* 2013 (19), 11224-1-1234.
3 M.F. Jarrold. Helices and Sheets in vacuo. *Phys. Chem. Chem. Phys.* 2007 (9), 1559–1671.
4 R. Kusaka, D. Zhang, P.S. Walsh, J.R. Gord, B.F. Fisher, S.H. Gellman, T.S. Zwier. Role of Ring-Constrained γ-Amino Acid Residues in α/γ-Peptide Folding: Single-Conformation UV and IR Spectroscopy. *J. Phys. Chem. A* 2013 (117), 10847–10862.

3.3 Explicit-state model checking

David Dill (Stanford University, US)

Main reference D. L. Dill, "A Retrospective on Murφ," in O. Grumberg, H. Veith (eds.). "25 Years of Model Checking - History, Achievements, Perspectives," LNCS, Vol. 5000, pp. 77–88, Springer, 2008; pre-print available from symposium webpage.
URL http://dx.doi.org/10.1007/978-3-540-69850-0_5
URL https://www7.in.tum.de/um/25/pdf/Dill.pdf

I did a tutorial on techniques developed in the model-checking/protocol verification community for solving the reachability problem for large implicitly-defined state graphs, with the thought that similar methods might be useful for discovering new synthesis pathways in organic chemistry. I described the basics of explicit state model checking and the optimizations of hash compaction, symmetry reduction, and partial order reduction.

References

1 C. Norris Ip and David L. Dill. Better verification through symmetry. *Formal Methods in System Design*, 9(1–2):41–75, 1996.
2 Ulrich Stern and David L. Dill. Parallelizing the Murφ verifier. In Orna Grumberg, editor, *9th Int' Conf. on Computer Aided Verification*, volume 1254 of *Lecture Notes in Computer Science*, pages 256–67. Springer-Verlag, June 1997. Haifa, Isreal, June 22–25.
3 Ulrich Stern and David L. Dill. Using magnetic disk instead of main memory in the Murφ verifier. In *10th Int'l Conf. on Computer Aided Verification*, pages 172–183, June 1998.

3.4 Molecular Codes

Peter Dittrich (Universität Jena, DE)

Joint work of Görlich, Dennis; Dittrich, Peter
Main reference D. Görlich, P. Dittrich, "Molecular Codes in Biological and Chemical Reaction Networks," PLoS ONE, 8(1):e54694, 2013.
URL http://dx.doi.org/10.1371/journal.pone.0054694

A molecular code is a mapping that can be realized by a reaction network and that is contingent, that is, the same reaction network can realize a different mapping using the same set of molecules as a domain and codomain of molecules as the original mapping. With

this definition we can distinguish different chemistries with respect to their ability to realize molecular codes. It seems that chemistries acquired by life can realize much more molecular codes than chemistries found in non-living systems, like the atmosphere photo-chemistry or combustion chemistries. Thus measuring contingency of mappings in a strict sense as demonstrated here might be a useful tool to get a better understanding of the emergence and of evolution of meaningful information and semiotic communication in living systems.

3.5 Finding the K Best Synthesis Plans

Rolf Fagerberg (University of Southern Denmark – Odense, DK)

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

Retrosynthetic synthesis planning works by transforming a target molecule into simpler precursor structures. In one version of this, first a set of bonds in the target molecule is chosen, whose removal then defines the precursor structures. In a next step, a sequence for creating these bonds is chosen, leading to a base synthesis plan. Cost measures, based on e. g. yield/loss of the reactions involved, can be associated with plans. Methods for picking the best plan for a given bond set are known, based e. g. on dynamic programming. Here, we propose to model synthesis plans for a given bond set as hyperpaths in a hypergraph. As a consequence, a polynomial time algorithm to find the K shortest hyperpaths [1] can be adapted to computing the K best synthesis plans for the bond set. Since any modeling and cost measure definition necessarily leaves out many real-world details, finding a (small) set of good plans – as opposed to merely one – allows for a chemist to choose a good plan based on additional chemical knowledge and actual wet-lab feasibility.

References

1 Lars Relund Nielsen, Kim Allan Andersen, Daniele Pretolani. Finding the K shortest hyperpaths. *Computers & Operations Research*, 32 (2005) 1477–1497

3.6 Beyond Numbers: Physical Simulation with Complex States

Harold Fellermann (Newcastle University, GB)

Traditional test tube chemistry typically operates with few purified chemical reagents. This convenience is lost when chemical systems approach the complexity of living systems which are prone to involve a multitude of spatially arranged interacting reagents. Particularly when molecules are able to polymerize and form heteropolymers, the space of potential chemical species becomes infinite. While a plethora of computational techniques exists for finite, small-sized reaction systems, computational tools for combinatorial chemistries over infinte dimensional system spaces are still an area of active development. My talk will introduce the use of formal calculi to infer reactions in complex reaction spaces from a set of axiomatic reaction rules. The first part of the talk will demonstrate how stochastic semantics can infer reaction rates among members of a combinatorial chemical library composed of self-replicating polymers. An example chemistry is formed by self-replicating binary heteropolymers and

I will present how symmetries in the reaction network give rise to a previously unreported selection pressure toward sequences with low information content [1]. The second part of the talk showcases how formal calculi can express a more complex reaction space where reagents are encapsulated in possibly nested nano-compartments such as vesicles. DNA tags on these compartments allow for targeted microfluidic manipulation, and DNA computing operations can be used to either control or report on the outcome of reaction cascades. The overall framework allows for programmable chemistry and has been applied to both chemical production as well as molecular computing [2, 3].

References

1 S. Tanaka, H. Fellermann, S. Rasmussen. *Europhys Lett* 107(2), 28004, 2014
2 H. Fellermann and L. Cardelli. *R. Soc. Interface* 11(99), 2014
3 H. Fellermann, M. Hadorn, R. Füchslin and N. Krasnogor. *ACM J. Emerg. Tech. Comp. Sys* 11(3):24, 2014

3.7 Artificial Chemical Life

Martin Hanczyc (University of Trento, IT)

Joint work of Hanczyc, Martin; Ikegami, Takashi
Main reference M. M. Hanczyc, T. Ikegami, "Chemical basis for minimal cognition." Artificial Life, 16(3):233–243., 2010.
URL http://dx.doi.org/10.1162/artl_a_00002

We present liquid-liquid droplet systems as physical embodiments of concepts from complex systems, computer science and origin of life research. By creating droplet systems containing chemistry that are far from equilibrium, dynamics are observed governed by fluid mechanics. Dynamic properties presented are self-motion, droplet division and chemotaxis. Developing such dynamics from physical systems approaching equilibrium to something more interesting such as systems capable of decision making and evolution are discussed. Comparisons to dissipative systems as well as the Game of Life are presented.

3.8 Algorithms for Autocatalytic Sets

Wim Hordijk (SmartAnalytiX.com – Lausanne, CH)

Joint work of Hordijk, Wim; Steel, Mike
Main reference W. Hordijk, "Autocatalytic sets: From the origin of life to the economy," BioScience, 63(11):877–881, 2013.
URL http://dx.doi.org/10.1525/bio.2013.63.11.6

Life is a functionally closed and self-sustaining chemical reaction network. During the 70s, several formal models of life based on this definition were developed independently, including hypercycles, autopoietic systems, the chemoton model, and autocatalytic sets. The notion of autocatalytic sets has been formalized as RAF theory more recently, and studied in detail in a theoretical and computational way. This has led to more insight into the possible emergence, structure, and further evolution of such sets. These studies were mostly done using a simple model of a chemical reaction system known as the binary polymer model, but the formal

framework has also been applied to real chemical networks such as an experimental system of cooperative RNA molecules and the metabolic network of E. coli. In this talk I will give an overview of the formal RAF framework and its main results, with an emphasis on algorithmic aspects of various computational problems related to studying the existence and structure of autocatalytic sets.

3.9 A Maximalism Design principle for Chemical Experiments

Takashi Ikegami (University of Tokyo, JP)

Joint work of Hanczyc, Marti; Takashi Ikegami
Main reference T. Ikegami, M. M. Hanczyc, "The search for a First Cell under the Maximalism Design Principle," Technoetic Arts, 7(2):153–164, 2009.
URL http://dx.doi.org/10.1386/tear.7.2.153/1

We often assume that evolution proceeds from simple to complex in various levels of description, e. g. morphology and behavioral structures. An example can be found in a suture line pattern of ammonites or the increment of the genomic size from prokaryotes to eukaryotes. In the simulation of artificial life, we tend to simulate how simple creatures develop into complex ones. Increasing complexity corresponds to the direction of evolution (or time so to speak) similar to the second law of thermodynamics. It should be noted that high entropy is not equal to high complexity, as we think living systems absorb the entropy (i. e. neg-entropy) by increasing their structural complexity. Recently, there have been many attempts to make a minimal cell in the nano and micro scales. Here the possible minimal cell would be a liposome that contains a minimal metabolic cycle for self-maintenance and self-replication. However, none of these attempts have succeeded so far (see e. g. [6]). We thus have conducted a chemical experiment: add oleic anhydride oil phase to highly alkaline water phase (pH 12) to see how the hydrolysis of the anhydride proceeds in a glass plate. Immediately the oil begins to react with the water causing the oil phase to break up into smaller spherical droplets, several to hundreds of microns in diameter. These droplets are like gliders in the game of life moving freely in the space and interacting with each other. Different from the game of life, the droplets can change direction spontaneously and coming into contact they never fuse together. In other words, they are far more robust than gliders. Also the droplets have finite life spans of less than 30 minutes and are sensitive to factors in the external environment such as pH. We argue that the mechanism of the movement is caused by the coupling of the hydrolysis reaction at the interface with the fluid dynamics of the droplet. Because of this coupling, chemical reaction lasts much longer than without the coupling. This is a "half living" state as it sustains the non-equilibrium state by its own self-regulation. A key point is that the environmental conditions (such as pH, product concentration, Reynolds number„,) are self-organized by the system itself rather than being prepared by the experimenter. If we try to obtain the same behavior by preparing the oil phase in high pH water along with some product of the reaction, the moving droplets never appear. The moving state, i. e. chemical gliders, appears through radical self-construction of the environment. We state that we cannot rely on the power of self-organization in searching for the origin of life. This is because both self-organization and the rich complex initial state are required. The ratio of the two is determined by the "rareness" of the event. Self-organization tends to simplify the final outcome limiting it to a low degree of complexity, while the low complexity assures the robustness of the outcome. The rich and complex initial state prevents the system from falling into a simple state. Therefore, we call our principle the Maximalism design principle.

References

1 Hanczyc, M.M. and Ikegami, T. Chemical Basis for Minimal Cognition. *Artificial Life.* 16(3):233–243, 2010.

2 Hanczyc, M.M., Toyota,T., Ikegami, T., Packard, N. and Sugawara, T. Chemistry at the oil-water interface: Self-propelled oil droplets, *J. Am. Chem. Soc.* (2007); 129(30):9386–9391.

3 Ikegami, T and Suzuki, K. From Homeostatic to Homeodynamic Self. *BioSystems*, 91 (2008) 388-400.

4 Matsuno, H., Hanczyc, M.M. and Ikegami, T. Self-maintained Movements of Droplets with Convection Flow. *LNAI* 4828, 2007, pp. 179–188.

5 Ikegami, T. and Hanczyc, M.M. The search for a First Cell under the Maximalism Design Principle. *Technoetic Arts* 7(2):153–164, 2009.

6 Luisi, L. The Emergence of Life: From Chemical Origins to Synthetic Biology. Cambridge University Press 2006.

3.10 Autocatalysis and evolution in metabolic networks

Ádam Kun (Eötvös Lorand University – Budapest, HU)

Joint work of Kun, Ádam; Papp, Balázs; Szathmáry, Eörs
Main reference Á. Kun, B. Papp, E. Szathmáry, "Computational identification of obligatorily autocatalytic replicators embedded in metabolic networks," Genome Biology, 9(3):R51, 2008.
URL http://dx.doi.org/10.1186/gb-2008-9-3-r51

Autocatalysis is central to biology. Life is autocatalytic: given food, a bacterial cell can grow and divide into two more or less identical bacteria. The same argument can be made at different level of biological organization, like genes, enzymes, cells, multicellular organism, eusocial societies, etc. The minimal living entity has 3 subsystems, each of which is autocatalytic: metabolism, membrane (compartmentalization) and information storage. During the evolution of life, living organism were preceded by infrabiological systems that only had two of the above subsytems. In the widely accepted RNA world hypothesis the infrabiological system consisted of the informational and the metabolic subsystem, and compartmentalization occurred later. The compartmentalization of ribozymes (RNA enzymes) lead to the first living system, and thus this is the first major evolutionary transition. Metabolism has obligatory autocatalytic cycles. An obligatory autocatalytic cycle is one whose constituents cannot be produced from the externally available (food) molecules. It turns out that ATP, the universal energy molecule of living systems, is universally obligatory autocatalytic. We have so far analysed 24 Bacterial, 4 Archaeal, and 2 Eukaryotic genome-based metabolic reconstructions. In each of them ATP was found to be obligatory autocatalytic. In certain organisms, other cofactors, such as NAD, CoA, THF, was also found to be obligatory autocatalytic. [2] Plausible metabolic networks can be generated, but then we need to understand how and in what order did reactions in the network get enzymatic catalysis. On one hand, having more reaction catalysed result in higher yield, but more enzymes also require more biomass component to be produced to replicate the whole system. Moreover, side reactions are problematic. We have shown by numerical simulation of an enzyme-catalysed reaction chain that specialist enzymes can appear, but only after the invention of linked genes (chromosome). [1] There is still much to do. The employed reaction network was very simple, not even approaching the complexity of real metabolic network.

Algorithmic, modelling and computational innovation is required to make the transition from toy-models (which are still computationally hard) to more realistic ones.

References

1 Szilágyi A.; Kun Á.; Szathmáry E. Early evolution of efficient enzymes and genome organization. *Biology Direct* 7:38, 2012.

2 Kun Á.; Papp B.; Szathmáry E. Computational identification of obligatorily autocatalytic replicators embedded in metabolic networks. *Genome Biology* 9:R51, 2008.

3.11 Architectures for Self-reproduction: Abstractions, Realisations and a Research Program

Barry McMullin (Dublin City University, IE)

Joint work of McMullin, Barry; Hasegawa, Tomonori; Baugh, Declan
Main reference B. McMullin, "Architectures for Self-reproduction: Abstractions, Realisations and a Research Program," in Proc. of the 13th Int'l Conf. on the Simulation and Synthesis of Living Systems (ArtificialLife'13), pp. 83–90, MIT Press, 2012.
URL http://dx.doi.org/10.7551/978-0-262-31050-5-ch012

It is well recognised that von Neumann's seminal abstraction of machine self- reproduction can be related to the reality of biological self-reproduction – albeit only in very general terms. On the other hand, the most thoroughly studied artificial evolutionary systems, incorporating meaningful self-reproduction, are the coreworld systems such as Tierra, Avida etc.; and these, in general, rely on a purely "self-inspection" mode of reproduction (or, more simply, "replication"). To the extent that the latter has any direct biological analog it would appear to be with molecular level reproduction and evolution in the hypothesised RNA-world. In this presentation I review the details and distinctions between these modes of reproduction. I indicate how the abstract von Neumann architecture can, in fact, be readily realised in coreworld systems; and outline the research program that flows from this. Finally I present some very preliminary results from actually implementing this programme. These validate the overall concept, and the merits (in terms of feasible evolutionary experimentation, with relatively modest computational effort, and in tractable wall clock time); but highlight the rapid onset of a range of difficult challenges in analysing and making sense of the resultant experimental data sets.

References

1 Adami, C. *Artificial Life, An Introduction.* Springer, 1998.

2 Adami, C. and Brown, C. T. Evolutionary learning in the 2D artificial life system "Avida". In Brooks, R. A. and Maes, P., eds., *Proc. Artificial Life IV*, pp. 377–381. MIT Press, 1994.

3 Barricelli, N. Symbiogenetic evolution processes realized by artificial methods. *Methodos*, 9(35-36):143–182 1957.

4 Barricelli, N. Numerical testing of evolution theories. *Acta Biotheoretica*, 16(3):99–126 1963.

5 Burks, A. W., editor *Theory of Self-Reproducing Automata [by] John von Neumann.* University of Illinois Press, Urbana 1966.

6 Dewdney, A. K. Computer recreations: A program called mice nibbles its way to victory at the first core wars tournament. *Scientific American*, 256(1):8–11 1987.

7 Gilbert, W. Origin of life: The RNA world. DOI: 10.1038/319618a0, 319(6055):618–618, 1986.

8 McMullin, B. Architectures for self-reproduction: Abstractions, realisations and a research program. In Adami, C., Bryson, D. M., Ofria, C., and Pennock, T., R., editors, *Artificial Life 13*, pages 83–90. MIT Press 2012.

9 Rasmussen, S., Knudsen, C., Feldberg, R., and Hindsholm, M. The coreworld: Emergence and evolution of cooperative structures in a computational chemistry. *Physica*, 42D:111–134 1990

10 Ray, T. S. An approach to the synthesis of life. In Langton, C. G., Taylor, C., Farmer, J. D., and Rasmussen, S., editors, *Artifical Life II*, volume X of *Series: Sante Fe Institute Studies in the Sciences of Complexity*, pages 371–408. Addison-Wesley Publishing Company, Inc., Redwood City, California 1992.

11 Ray, T. S. An evolutionary approach to synthetic biology: Zen and the art of creating life. *Artificial Life*, 1:179–209 1994.

12 von Neumann, J. Theory and organization of complicated automata. In [5], pages 29–87 (Part One). Based on transcripts of lectures delivered at the University of Illinois, in December 1949. Edited for publication by A.W. Burks 1949.

13 Watson, J. D. and Crick, F. H. C. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171:737–738 1953.

3.12 Generative Chemistries

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

Joint work of Andersen, Jakob L.; Flamm, Christoph; Merkle, Daniel; Stadler, Peter F.
Main reference J. L. Andersen, C. Flamm, D. Merkle, P. F. Stadler, "Generic strategies for chemical space exploration," Int'l J. of Computational Biology and Drug Design, 7(2–3):225–258, 2014.
URL http://dx.doi.org/10.1504/IJCBDD.2014.061649

We use undirected labeled graphs and graph rewriting as a natural model for chemical compounds and chemical reactions, respectively. The combinatorial explosion that is often seen in graph grammar based chemical space exploration makes it infeasible to compute the underlying network with a breadth first expansion approach. We alleviate this problem by introducing a strategy framework that is well suited to explore chemical spaces. Within these spaces, which are formally hypergraphs, we find "Chemical Transformation Motifs" like pathways, autocatalytic patterns, or polymerization, based on various optimization techniques. Solutions provide hypothesis on an atomic level that allow, for example, for direct wet-lab verification. The Formose reaction, the Krebs cycle, the Pentose Phosphate Pathway and many more systems are examples where we applied our approaches. The tool implementing our techniques is called "MØD". The presentation gave a gentle introduction and illustrated the approaches based on several application scenarios.

3.13 Systems Chemistry: Self replicators from dynamic molecular networks

Sijbren Otto (University of Groningen, NL)

Main reference M. Malakoutikhah, J. J-P. Peyralans, M. Colomb-Delsuc, H. Fanlo-Virgós, M. C. A. Stuart, S. Otto, "Uncovering the selection criteria for the emergence of multi-building-block replicators from dynamic combinatorial libraries," J. of the American Chemical Society, 135(49):18406–18417, 2013.
URL http://dx.doi.org/10.1021/ja4067805

How did life start? Can we make life? These are among the most fundamental questions in contemporary science. In this talk I have shown how, starting from a mixture of relatively simple interconverting organic molecules, a set of self-replicating molecules can emerge

spontaneously. Importantly, the process of self-replication as found to be exponential, which is an important characteristic in the context of Darwinian evolution. Mutation of the replicators were enabled by providing the system with different building blocks. When mixed, these systems gave rise to the emergence of two different replicator quasi-species, of which one is the ancestor of the other. Molecular-level insight into the quasi-speciation process showed that outliers in the mutant distribution of the first quasi-species induced the formation of the second quasi-species. The next step is now to achieve replication far from equilibrium, by allowing concurrent replication and destruction processes to take place. The need and challenges for computational approaches that simulate the complex network of reactions were highlighted. Future progress requires a joined experimental and computations approach.

3.14 Systems chemistry and the frontiers of statistical mechanics: fruitful study systems from the Origin of Life

Eric Smith (Santa Fe Institute, US)

The origin of life provides an application domain and a source of questions for cheminformatics. Some questions involve collaboration between computation and laboratory chemistry. Others involve collaboration between computation and statistical physics concepts. All of these have in common a need to extend from a theory of reactions to a theory of reaction systems. Five areas in which chemical experiments and models already exist, which invite a computational analysis, are:

- Radiation chemistry in planetary atmospheres, particularly with respect to oxidation states of carbon, which is important for early organosynthesis.
- Stable isotope chemistry in atmospheres, oceans, and the subsurface, which provides a connection between the rock record and complex atmosphere models.
- Mineral alteration chemistry in hydrothermal systems, which are likely environments for biologically important organosynthesis.
- Organic geochemistry of carbon addition at metal-sulfide surfaces, a class of reactions that may link atmospherically plausible carbon species to prebiotically relevant organic species.
- Organic reactions in solution in the presence of soluble transition metals which form metal-ligand complexes; these are known to create different reaction pathways depending on the metals present.

Another set of conceptual questions is also of interest:

- Can wecreate a theory of non-equilibrium phase transitions in chemistry, combining the understanding of cooperative and collective dynamics from statistical physics with the rich and heterogeneous state-space structure of chemistry.
- - What are the relations of Ross Ashby's "Law of Requisite Variety" in hierarchical control systems, and data rate theorems for error generation and error correction in chemical and catalytic systems?
- Is there a concept of a standard for chemical problems in the origin of life, in which the problem is well-defined independently of the mechanisms used to approximate a solution? Currently models of geochemical settings tend to be one-off case studies, in which the problem formulated and the particular simulation are closely intertwined, and benchmarking of the quality of solutions is difficult or is performed ad hoc.

3.15 Practical Graph Isomorphism

Adolfo Piperno (University of Rome "La Sapienza", IT)

Joint work of McKay, Brendan D.; Piperno, Adolfo
Main reference B. D. McKay, A. Piperno, "Practical Graph Isomorphism, II," Journal of Symbolic Computation, 60(Jan. 2014):94–112, 2014.
URL http://dx.doi.org/10.1016/j.jsc.2013.09.003

I have presented a tutorial reporting the current state of the graph isomorphism problem from the practical point of view. After describing the general principles of the refinement-individualization paradigm and proving its validity, it has been explained how this technique is implemented in several of the key tools in the literature. In particular, the the best known program nauty and an innovative approach called Traces have been described. The presented tools are available at http://pallini.di.uniroma1.it

3.16 Exercises in Molecular Computing and Robotics

Darko Stefanovic (University of New Mexico - Albuquerque, US)

Joint work of Stefanovic, Darko; Stojanovic, Milan; Macdonald, Joanne; Graves, Steven; Lakin, Matthew; Brown, Carl; Olah, Mark

When we say that molecules compute, we mean that they respond to certain inputs, for example, the presence or absence of other molecules, in a precisely defined but potentially complex fashion. The simplest way for a chemist to think about computing molecules is as sensors that can integrate the presence or absence of multiple analytes into a change in a single reporting property. In the talk, I reviewed several forms of molecular computing using concentrations of single-stranded DNA to carry signals. When we began our work, combinatorial approaches to using DNA for computing were used to search for solutions to constraint satisfaction problems. We chose to work instead on logic circuits, building bottom-up from units based on catalytic nucleic acids, focusing on DNA secondary structures in the design of individual circuit elements, and reserving the combinatorial opportunities of DNA for the representation of multiple signals propagating in a large circuit. Such circuit design directly corresponds to the intuition about sensors transforming the detection of analytes into reporting properties. While this approach was unusual at the time, it has been adopted since by other groups working on biomolecular computing with different nucleic acid chemistries. We created logic gates by modularly combining deoxyribozymes as reporting elements with stem-loops as input detection elements. For instance, a deoxyribozyme that normally exhibits an oligonucleotide substrate recognition region is modified such that a stem-loop closes onto the substrate recognition region, making it unavailable for the substrate and thus rendering the deoxyribozyme inactive. But a conformational change can then be induced by an input oligonucleotide, complementary to the loop, to open the stem, allow the substrate to bind, and allow its cleavage to proceed, which is eventually reported via fluorescence. Using multiple modifications, we constructed logic gates, and we constructed large circuits consisting of such logic gates operating in parallel, and orchestrated to carry out arithmetic operations or play simple games of strategy. To enable more complex circuits, with serial connections of gates, we recently developed buffered cascades that use the DNA strand

displacement mechanism to mediate between successive stages in the cascade, and structured substrate molecules to sequester input sequences until needed. We also used deoxyribozymes as the active components of molecular assemblies, called "molecular spiders", that can walk over surfaces and tracks prepared with substrates, cleaving the substrates in the process. I reviewed the coarse-grained models we developed to explain the experimentally observed behaviors of these molecular walkers in the bulk, contrasting their simple, uncoordinated gait with the complex, coordinated gait found in nature's protein molecular motors. Despite this simplicity, our models predict that molecular spiders exhibit a long superdiffusive transient while walking on a track, even in opposition to a force. We hope to be able to use them for cargo transport in engineered nanosystems.

3.17 Modeling phototrophic growth: Constraints and optimality in cyanobacterial metabolism

Ralf Steuer (HU Berlin, DE)

Joint work of Steuer, Ralf; Knoop, Henning
Main reference H. Knoop, M. Gründel, Y. Zilliges, R. Lehmann, S. Hoffmann, W. Lockau, R. Steuer, "Flux balance analysis of cyanobacterial metabolism: the metabolic network of Synechocystis sp. PCC 6803," PLoS Computational Biology, 9(6):e1003081, 2013.
URL http://dx.doi.org/10.1371/journal.pcbi.1003081

Cyanobacteria are phototrophic microorganisms of global importance and have recently attracted renewed attention due to their capability to convert atmospheric CO2 into organic compounds, including carbon-based transportation fuels and bulk chemicals. These ongoing efforts to domesticate cyanobacteria as a human resource would greatly benefit from an in-depths understanding of metabolism. Current approaches to harness the biotechnological potential of cyanobacteria are increasingly supported by integrated experimental and computational approaches to understand the systematic properties of phototrophic growth. From a modelling perspective, cyanobacteria are highly attractive organisms whose computational description spans a large number of challenging questions. In particular, a computational description of phototrophic growth must incorporate a hierarchy of processes, ranging from the biophysics of photosynthesis, the biochemistry of carbon fixation, molecular mechanisms of cellular growth, to global oxygen and carbon cycles. The contribution did outline the construction of computational models of cyanobacteria and the analysis of such models using kinetic and constraint-based methods. Computational modelling is supplemented with highly controlled growth experiments in laboratory-scale photobioreactors that allows us to determine relevant exchange fluxes. Modeling strategies include the high-quality reconstruction of stoichiometric models of cyanobacterial metabolism, their experimental validation, as well as a path towards the incorporation of kinetic properties and their temporal coordination into the description of phototrophic growth.

References

1 Henning Knoop, Marianne Gründel, Yvonne Zilliges, Robert Lehmann, Sabrina Hoffmann, Wolfgang Lockau, and Ralf Steuer. Flux balance analysis of cyanobacterial metabolism: the metabolic network of Synechocystis sp. PCC 6803. *PLoS Comput Biol*, 9(6):e1003081, 2013.

2 Henning Knoop, Yvonne Zilliges, Wolfgang Lockau, and Ralf Steuer. The metabolic network of Synechocystis sp. PCC 6803: systemic properties of autotrophic growth. *Plant Physiol*, 154(1):410–422, Sep 2010.
3 Ralf Steuer, Henning Knoop, and Rainer Machné. Modelling cyanobacteria: from metabolism to integrative models of phototrophic growth. *J Exp Bot*, 63(6):2259–2274, Mar 2012.

3.18 Molecular Information Technology

Klaus-Peter Zauner (University of Southampton, GB)

Matter is animate if and only if it uses information processing to persist. My definition of life places primary importance on the computation any organism is required to continuously perform just to maintain its own complex material organization in a living state. In this real-time struggle against entropy evolution yielded information processing capabilities of formidable efficiency. The success of this molecular level computation is apparent in the marvellous architectures seen throughout biology. With laboratory experiments ranging from self-assembly and semi-biotic robotics to wet artificial neuronal networks based on Belousov Zhabotinsky medium compartmentalized in lipid coated droplets I illustrate how aspects of the molecular information processing inherent to all living systems may eventually be harvested for man-made technology. It is my view that adding molecular information technology to our engineering toolbox will allow us to overcome the complexity limit faced by synthetic chemistry and will lead to technology revolution that rivals the advent of organic chemistry.

References

1 Philip H. King, Gareth Jones, Hywel Morgan, Maurits R.R. de Planque, and Klaus-Peter Zauner. Interdroplet bilayer arrays in millifluidic droplet traps from 3D-printed moulds. *Lab on a Chip*, 14:722–729, January 2014.
2 Chris Lovell, Gareth Jones, Steve Gunn, and Klaus-Peter Zauner. Autonomous experimentation: Active learning for enzyme response characterisation. *JMLR: Workshop and Conference Proceedings*, 16:141–155, April 2011.
3 Soichiro Tsuda, Klaus-Peter Zauner, and Yukio-Pegio Gunji. Robot control with biological cells. *BioSystems*, 87:215–223, 2007.
4 Klaus-Peter Zauner. Molecular information technology. *Critical Reviews in Solid State and Material Sciences*, 30(1):33–69, March 2005.

4 Open Problems

The following list of open problems on the interface between chemistry and computer science have been identified and discussed during the seminar:

1. **General simulation methods for self-assembly systems:** Dynamic combinatorial libraries (DCL) are heavily investigated in the emerging field of Systems Chemistry. Upon physical or chemical perturbation these systems show interesting emergent phenomena such as

fiber-formation via self-assembly processes. A formal description of these inherently combinatorial systems and efficient simulation techniques to study their dynamic behavior are not developed, as well as algorithmic methods for their rational design.

2. **Reflexive synthesis plans:** Total synthesis, i. e. the construction of a target compound from smaller, readily available building blocks, is at the core of Organic Synthesis. Synthesis plans, which are tree-like data structures, are used to summarize the necessary information for the synthesis of complex molecular structures. To give synthesis plans the property of being fail-safe they would require the presence of alternative reaction sequences to circumvent reaction steps which turn out to fail in the laboratory. A formal framework is needed to describe and computationally design synthesis plans with this "reflexivity property".
3. **Covering graphs with tiles (particular sub-graphs):** Terpenes and polyketides are complex, pharmaceutically important, natural products, which are formed in plants via "polymerization" of a small number of distinct building blocks of different metabolic origin. Covering these combinatorial molecular objects with their building blocks (viewed as tiles) would allow for the elucidation from which part of the plants secondary metabolism the various substructures of the natural product originated. This knowledge strongly improves the elucidation of unknown metabolic pathways to these important class of molecules.
4. **Random graph models of chemical reaction networks**: The detection of statistically significant motives is based on a proper statistical null model. While the theory of random graph models for simple graphs is well developed, almost nothing can be found for hypergraphs, the usual formalization of chemical reaction networks. Chemistry imposes additional semantic structure onto the network, which needs to be preserved during randomization. A proper null model for chemical reaction networks would for instance allow to detect "chemical transformation motifs" in metabolic networks or chemical reaction networks in general.
5. **Constraint constitutional isomer generator:** An important problem in Mass-Spectroscopy (MS) is the generation of constitutional isomeres from the knowledge of the molecular formula. In many cases additional chemical and physical constrains are known. Efficient algorithms which exploit additional constraints to narrow down the combinatorial space of constitutional isomeres and which are specifically designed for the chemical context are an urgent need.
6. **MS peak perception and interpretation of small molecules:** High resolution mass-spectroscopy, a recently developed experimental technique, allows to collect information on the structure of small (organic) molecules indirectly via the mass peak of the entire molecule and the pattern of mass peaks after fragmentation of the molecule into smaller pieces. On the algorithmic side the structure elucidation from this type of data is far from complete and current methods do not scale up to the huge amounts of data collected during the measurement of the metabolomes of entire organisms.
7. **Inverse mechanism problem (from mechanisms to real molecules):** Dynamical systems theory offers an almost complete set of tools to study the dynamics of a chemical reaction system on the level of the differential equations, which are usually derived from the underlying chemical equations. With the advent of Systems Chemistry, however, the inverse problem of specifying a set of molecules, which implement a particular dynamical system gained increasing attention. By assigning molecules to the abstract variables in the reaction equations constraints in the form of sub-graphs propagate through the reaction mechanism mapping the problem into the realm of constraint-optimization problems.

Participants

- Jakob Lykke Andersen
University of Southern Denmark – Odense, DK
- Carsten Baldauf
FHI – MPG Berlin, DE
- Wolfgang Banzhaf
Memorial University of Newfoundland, CA
- Nikolaj Bjørner
Microsoft Corp. – Redmond, US
- Sebastian Böcker
Universität Jena, DE
- David Dill
Stanford University, US
- Peter Dittrich
Universität Jena, DE
- Andreas Dress
Bielefeld, DE
- Rolf Fagerberg
University of Southern Denmark – Odense, DK
- Harold Fellermann
Newcastle University, GB
- Christoph Flamm
Universität Wien, AT
- Martin Hanczyc
University of Trento, IT
- Marc Hellmuth
Universität des Saarlandes, DE
- Wim Hordijk
SmartAnalytiX.com – Lausanne, CH
- Takashi Ikegami
University of Tokyo, JP
- Ádám Kun
Eötvös Lorand University – Budapest, HU
- Barry McMullin
Dublin City University, IE
- Daniel Merkle
University of Southern Denmark – Odense, DK
- Markus E. Nebel
TU Kaiserslautern, DE
- Sijbren Otto
University of Groningen, NL
- Adolfo Piperno
University of Rome "La Sapienza", IT
- D. Eric Smith
Santa Fe Institute, US
- Pietro Speroni di Fenizio
Dublin City University, IE
- Peter F. Stadler
Universität Leipzig – IZBI, DE
- Darko Stefanovic
University of New Mexico – Albuquerque, US
- Ralf Steuer
HU Berlin, DE
- Klaus-Peter Zauner
University of Southampton, GB