Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

# Modular verification of chemical reaction network encodings via serializability analysis

Matthew R. Lakin [a,*], Darko Stefanovic [a,b], Andrew Phillips [c,*]

[a] Department of Computer Science, University of New Mexico, Albuquerque, NM, USA
[b] Center for Biomedical Engineering, University of New Mexico, Albuquerque, NM, USA
[c] Microsoft Research, Cambridge, UK

ARTICLE INFO

*Article history:*
Received 30 October 2014
Received in revised form 31 March 2015
Accepted 16 June 2015
Available online xxxx

*Keywords:*
Chemical reaction networks
Modular verification
Serializability
DNA strand displacement

ABSTRACT

Chemical reaction networks are a powerful means of specifying the intended behavior of synthetic biochemical systems. A high-level formal specification, expressed as a chemical reaction network, may be compiled into a lower-level encoding, which can be directly implemented in wet chemistry and may itself be expressed as a chemical reaction network. Here we present conditions under which a lower-level encoding correctly emulates the sequential dynamics of a high-level chemical reaction network. We require that encodings are transactional, such that their execution is divided by a "commit reaction" that irreversibly separates the reactant-consuming phase of the encoding from the product-generating phase. We also impose restrictions on the sharing of species between reaction encodings, based on a notion of "extra tolerance", which defines species that may be shared between encodings without enabling unwanted reactions. Our notion of correctness is serializability of interleaved reaction encodings, and if all reaction encodings satisfy our correctness properties then we can infer that the global dynamics of the system are correct. This allows us to infer correctness of *any* system constructed using verified encodings. As an example, we show how this approach may be used to verify two- and four-domain DNA strand displacement encodings of chemical reaction networks, and we generalize our result to the limit where the populations of helper species are unlimited.



## 1. Introduction

Recent successes in DNA nanotechnology have demonstrated the ability of scientists and engineers to exert unprecedented control over matter at the nanoscale. This control has been used in dynamic DNA logic circuits using a range of molecular architectures [1–4] as well as in the assembly of static structures based on tile assembly [5] or on folding of scaffold strands [6]. Recent work has combined the dynamic and static domains by using a dynamic DNA circuit to trigger self-assembly reactions [7] or reconfigure existing structures [8,9]. Theoretical work has explored the computational power of DNA tile assembly processes [10] and of dynamic DNA reactions [11].

In the theory underlying these efforts, chemical reaction networks (CRNs) play a critical role. Previous work has shown that stochastic CRNs are Turing-universal, provided that an arbitrarily small probability of error is allowed [12], and the fundamental limits of deterministic CRN computation have been explored [13]. Therefore, CRNs are a convenient and powerful

* Corresponding authors.
*E-mail addresses:* mlakin@cs.unm.edu (M.R. Lakin), darko@cs.unm.edu (D. Stefanovic), aphillip@microsoft.com (A. Phillips).

http://dx.doi.org/10.1016/j.tcs.2015.06.033
0304-3975/

programming language for synthetic biochemical systems, which enable researchers to define the desired interactions as a CRN and explore them through simulation using established tools and methods. When such a high-level formal CRN has been designed, in order to test it in the laboratory the abstract CRN species must be mapped to actual chemical species whose interactions correspond to those of the formal CRN. From a computer science perspective, this can be thought of as *compiling* the CRN into another CRN that encodes a lower-level behavioral description of the system. Thus, a common feature of these encodings is that a single-step reaction in a formal CRN is implemented as a multi-step process in its encoding, which comprises a number of individual chemical reactions. When an encoding of a multi-reaction system is executed, the executions of the individual reaction encodings can, and almost certainly will, be interleaved with each other. Thus, there are many possibilities for bugs in the design of CRN encodings due to unwanted interleavings of reactions that may, for example, prematurely generate or consume certain species. It is therefore desirable to develop proofs of correctness for CRN encodings. However, verifying such massive, concurrent systems is non-trivial because a large number of interactions may be possible from any given state.

Since the inputs and outputs from this process are both CRNs, this enables the definition of a hierarchy of CRNs, each of which encodes the behavior of the CRN above it in the hierarchy. In this paper we will focus on the verification of a single encoding step in which a formal (higher-level) CRN is translated into an encoding (lower-level) CRN. However, the same techniques could be used to verify each step of a multi-step encoding process separately.

Here we present a powerful, modular framework for proving correctness of CRN encodings. Since CRN encodings are typically defined pointwise by encoding each reaction separately, it is natural to exploit this modularity in our proof technique. We will show that if all individual reaction encodings in the system satisfy certain properties then the whole system may be deduced to be correct in a well-defined sense. Our modular approach to proving correctness will allow us to verify the components of large-scale systems individually, without being limited by the sizes of the corresponding state spaces.

Our approach is inspired by the concept of *serializability* from database theory [14], which requires that interleaved concurrent updates to a database must be equivalent to some serial schedule of those updates. We consider a composition of reaction encodings to be correct if all possible interleavings of their reactions can be rewritten using a small number of simple rules to produce a *serial schedule*. A serial schedule is one in which there is no interleaving between the various reaction encodings, that is, the first reaction encoding runs to completion before the second begins, and so on. We propose serializability as a reasonable notion of correctness for CRN encodings because serialized executions of encodings can be directly related to executions of the underlying reactions. We will use simple rules to rewrite reaction traces in order to serialize them. CRN encodings that are not serializable may display erroneous behaviors that do not correspond to possible behaviors of the formal CRN, because of unwanted crosstalk between individual reaction encodings. Our correctness criteria will allow us to prove that our encodings do not have such problems. We also generalize our results to the case where fuel species populations may be assumed to be unlimited, either because they are relatively very large or are continually fed from an external source. This is of interest in the case of long-running reaction networks such as oscillators [15].

As an example, we will apply our technique to the verification of several encodings of the approximate majority algorithm of Angluin et al. [16] using DNA strand displacement reactions. DNA strand displacement is a simple yet powerful framework for molecular computation, in which an invading strand displaces an incumbent strand that is bound to a template [17,18]. The applications of DNA strand displacement reactions are numerous and varied, including the construction of logic circuits [19,20] and neural networks [21], control of self-assembled nanoscale systems [22,7] and molecular motors [17,23–25]. Here we are particularly interested in applications of DNA strand displacement to implement dynamic behavior expressed as chemical reaction networks: Soloveichik et al. showed that DNA strand displacement *reaction gates* [26] provide a general framework for the implementation of arbitrary CRNs in wet chemistry. A number of encoding schemes for implementing chemical reaction networks using DNA strand displacement have been proposed, such as the four-domain [26] and two-domain schemes [27]—these names refer to the domain-level structure of the strands which denote encoded species in each scheme. In previous work we have explored the use of probabilistic model checking for the verification of two-domain DNA strand displacement systems [28]. However, such approaches are limited by the explosion in the size of the state space as the various species populations increase.

The remainder of this paper is structured as follows. In Section 2 we introduce mathematical notation and preliminary definitions, and we formalize modular CRN encodings in Section 3. In Section 4 we present a (straightforward) completeness proof for reaction encodings, and in Section 5 we present a (more involved) soundness proof. We discuss extensions of these results to handle unlimited fuel populations in Section 6. We present examples of encoding verification in Section 7, and conclude with a discussion and a survey of related work in Section 8. This paper is a revised and extended version of a conference paper (M.R. Lakin, A. Phillips, D. Stefanovic, Modular verification of DNA strand displacement networks via serializability analysis, in: D. Soloveichik, B. Yurke (Eds.), Proceedings of DNA19, Vol. 8141 of Lecture Notes in Computer Science, Springer-Verlag, 2013, pp. 133–146).

## 2. Preliminaries

We now introduce some preliminary mathematical definitions that will be used throughout the paper. Let $\mathbb{N}$ denote the set of natural numbers, including zero. Given a set $X$, we write $\mathbb{N}^X$ for the set of multisets over $X$, defined as the set of all functions $f : X \to \mathbb{N}$, as is standard. By convention we use upper-case boldface symbols for multisets and upper-case italics for sets. We may write multisets explicitly using the notation $\{x_1 = n_1, \ldots, x_k = n_k\}$, where $n_i > 0$ is the count associated

Please cite this article in press as: M.R. Lakin et al., Modular verification of chemical reaction network encodings via serializability analysis, Theoret. Comput. Sci. (2015), http://dx.doi.org/10.1016/j.tcs.2015.06.033