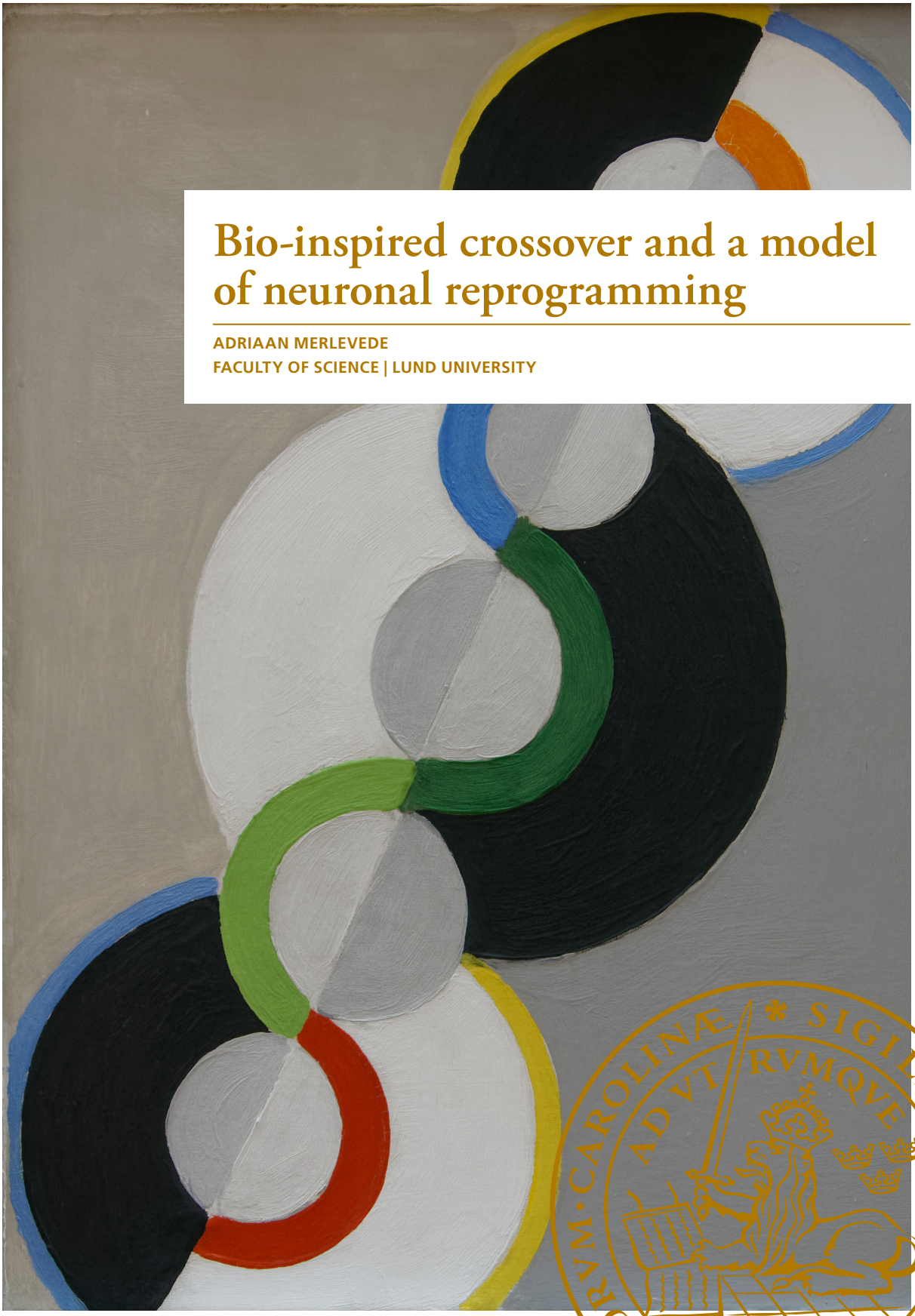
Bio-inspired crossover and a model of neuronal reprogramming

Merlevede, Adriaan

2020



Bio-inspired crossover and a model of neuronal reprogramming

ADRIAAN MERLEVEDE
FACULTY OF SCIENCE | LUND UNIVERSITY

Printed by Media-Tryck, Lund 2020 NORDIC SWAN ECOLABEL 3041 0903

Lund University
Faculty of Science
Department of Astronomy and Theoretical Physics
Computational Biology and Biological Physics
ISBN 978-91-7895-450-6

Bio-inspired crossover and a model of neuronal reprogramming

Bio-inspired crossover and a model of neuronal reprogramming

Adriaan Merlevede

LUND UNIVERSITY

Thesis for the degree of Doctor of Philosophy
Thesis advisors: Carl Troein and Victor Olariu
Faculty opponent: Wolfgang Banzhaf

To be presented, with the permission of the Faculty of Science of Lund University, for public criticism in hall K404 at the Department of Astronomy and Theoretical Physics on Monday, the 20th of April 2020 at 13:15.

Organization
LUND UNIVERSITY

Department of Astronomy and Theoretical Physics
Sölvegatan 14A
SE-223 62 LUND
Sweden

Document name
DOCTORAL DISSERTATION

Date of disputation
2020-04-20

Sponsoring organization

Author(s)
Adriaan Merlevede

Title and subtitle
Bio-inspired crossover and a model of neuronal reprogramming:

Abstract

Bio-inspired crossover Natural genomes are affected by various kinds of mutations, manifesting as variations in the genome. This includes modifications to individual base pairs (single nucleotide polymorphism, SNP), but also short insertions and deletions (indels), and variations in the position (translocations) or multiplicity (copy number variations) of genetic sequences. While SNPs are historically the most studied, the other, structural variations are now known to affect a large fraction of the human genome, and can result in clinically relevant expressions in the phenotype. In biology, structural variations are also known to affect evolutionary dynamics of a population. For example, most genes originate through duplication, followed by divergent evolution of the two copies. However, in computer programs simulating evolution, structural variations are rarely included. Their absence makes it difficult to study the evolutionary consequences of translocation and copy number variation in simulated models. Including them may also have positive consequences for bio-inspired computation such as genetic algorithms. In this dissertation, a theoretical framework with algorithmic implementations is presented for sexual reproduction between linear genomes structures with structural variation, including permutations, and DNA-like strings.

A model of neuronal reprogramming In multicellular organisms, cells typically follow a hierarchical progression from embryonal stem cells to increasingly specialised cell types. However, by manipulating the expression of key genes, it is possible to reverse this specialisation, or convert directly between terminal cell types. This technology has been applied to convert human skin cells to neurons, with promising applications in disease modelling and regenerative medicine. Understanding the gene regulation network (GRN) governing this transition is crucial for improving the efficiency of this conversion in the future. In this dissertation, we integrated known interactions described in the literature into a holistic GRN model. We measured the activity of genes that are known to play pivotal roles during a reprogramming process, and compared these data to our modelling system. We found that the reprogramming process could be accurately modelled. The interaction between two genes, PTB and nPTB, played a different role from its usual interpretation in the literature, acting as a negative feedback loop.

Key words
computational evolution, recombination, gene regulatory network, direct reprogramming

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title

ISBN
978-91-7895-450-6 (print)
978-91-7895-451-3 (pdf)

Recipient's notes

Number of pages
127

Price

Security classification

Signature Adriaan

Date 2020-03-10

Bio-inspired crossover and a model of neuronal reprogramming

Adriaan Merlevede

LUND UNIVERSITY

Cover illustration front: *Cercles simultanés* (simultaneous circles), Robert Delaunay (1934). Oil on cardboard. Photographed by J.L. Lacroix for the Musée de Grenoble. Artwork depicts genetic recombination in the eyes of a sleep deprived PhD student.

Faculty of Science, Department of Astronomy and Theoretical Physics

ISBN: 978-91-7895-450-6 (print)
ISBN: 978-91-7895-451-3 (pdf)

Printed in Sweden by Media-Tryck, Lund University, Lund 2020

Media-Tryck is a Nordic Swan Ecolabel certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se

MADE IN SWEDEN

Contents

Acknowledgements

My PhD studies have been a lot of things, but most of all, the past four and a half years have been a deeply humbling experience. I have been fortunate to work with and meet people who are exceptionally intelligent, interesting, and, most importantly, kind. For this I owe a few sincere statements of gratitude.

First and foremost, my supervisors, Carl and Victor. You have both given me great freedom and encouragement to develop the ideas and methods presented in this thesis. Thank you for being patient with me, and for leaving your door open for my foolish questions.

Thank you also to all the other colleagues at the department, who have been welcoming and supportive.

I moved to a foreign country to work on this research, and I feel strongly that all the wonderful people I shared this time with have contributed greatly to whatever modest achievements may be found in the following pages. Najmeh has been an important benefactor, sponsoring a healthy dose of sanity at work by always listening to my troubles and stories. Generous donations of beer and politics were supplied by Dinko and Jasmina, who, almost immediately when I arrived, adopted me into their social circle (consisting somehow exclusively of unusually brilliant and kindhearted gentlefolk). I managed all these years to not get stabbed thanks to helpful reminders from Matthäus, my partner in crime (the crime is beer). He also introduced me to Anna, who is best described as a ray of American sunshine and a much needed break from the usual Swedish weather. Fiona's sage advice is worth the weight of half this thesis in gold. To all these people, and many more who are not mentioned, I say thank you for being you, with me. You should be proud.

My family have been intensely involved as the world's most deluxe travel agency. Even more importantly, though, they have been an inexhaustible well of love and support. There is a sense of absurdity in thanking someone whose care is so self-evident that it is axiomatic to my nature, but here it is, irrational and inadequate: thank you.

Finally, in an equally absurd gesture, I thank the love of my life, Jeroen, whose help and encouragement have meant the world to me. Thank you for sharing the burden with me.

Abstract

Bio-inspired crossover Natural genomes are affected by various kinds of mutations, manifesting as variations in the genome. This includes modifications to individual base pairs (single nucleotide polymorphism, SNP), but also short insertions and deletions (indels), and variations in the position (translocations) or multiplicity (copy number variations) of genetic sequences. While SNPs are historically the most studied, the other, structural variations are now known to affect a large fraction of the human genome, and can result in clinically relevant expressions in the phenotype. In biology, structural variations are also known to affect evolutionary dynamics of a population. For example, most genes originate through duplication, followed by divergent evolution of the two copies. However, in computer programs simulating evolution, structural variations are rarely included. Their absence makes it difficult to study the evolutionary consequences of translocation and copy number variation in simulated models. Including them may also have positive consequences for bio-inspired computation such as genetic algorithms. In this dissertation, a theoretical framework with algorithmic implementations is presented for sexual reproduction between linear genomes structures with structural variation, including permutations, and DNA-like strings.

A model of neuronal reprogramming In multicellular organisms, cells typically follow a hierarchical progression from embryonal stem cells to increasingly specialised cell types. However, by manipulating the expression of key genes, it is possible to reverse this specialisation, or convert directly between terminal cell types. This technology has been applied to convert human skin cells to neurons, with promising applications in disease modelling and regenerative medicine. Understanding the gene regulation network (GRN) governing this transition is crucial for improving the efficiency of this conversion in the future. In this dissertation, we integrated known interactions described in the literature into a holistic GRN model. We measured the activity of genes that are known to play pivotal roles during a reprogramming process, and compared these data to our modelling system. We found that the reprogramming process could be accurately modelled. The interaction between two genes, PTB and nPTB, played a different role from its usual interpretation in the literature, acting as a negative feedback loop.

Popular science summary

Bio-inspired crossover

Evolution is the mechanism through which nature created all varieties of life on our planet. In order to understand this process, it is crucial to not only catalogue the natural history of species, but also to create deeper insights in the properties and limitations of Darwinian evolution. In natural science, such understanding is typically gained by comparing a wide range of possible scenarios, in the form of experiments that are deliberately designed to verify or falsify a scientific hypothesis. However, experimentation is difficult in evolutionary science, in part because many interesting phenomena tend to occur only over millions of years, but also because we are able to observe life on only one planet, and thus only one natural evolutionary process. Because all species on Earth share common ancestry, it is not always possible to tell which common features are necessary consequences of evolution and natural selection, and which features are random properties of our ancestors that were inherited by all living species. For example, we do not know if genetic information could be stored using other mechanisms than the DNA code. It may be that other molecules besides DNA are not as efficient at carrying hereditary information, and that the evolution of DNA-based life was inevitable on our planet; but we cannot rule out that an early life form evolved to use DNA as a chance event out of many possible alternatives, and then multiplied so that other hereditary mechanisms never arose.

Many shared properties of living species — such as DNA but also details of cell structure, metabolism, and the genetic underpinnings of embryonal development, for example — have great influence on the evolutionary process. They determine which evolutionary adaptations are possible, how populations react to environmental change, how new species can form and which interactions are possible between species. If some of these aspects of life as we know it are the result of random events in our shared evolutionary past, this implies that there are many other ways for evolution to progress, which we cannot observe in nature. Understanding how evolution happens in populations with other hereditary mechanisms, cell structures, and so on, would give us greater insight in the properties and limitations of evolution in general. This knowledge, in turn, would lead to technological innovation in some widely used computational techniques that are inspired by evolutionary theory. Understanding evolution also has consequences for cancer research, since tumours form out of a runaway evolutionary process wherein cells that normally cooperate to form tissues and organs instead compete on an individual basis in natural selection. Finally, if there is life on other planets, most scientists expect that it must have been formed by Darwinian evolution, so that a more general evolutionary theory might give a better idea of what to look for in the search for extraterrestrial life.

When traditional experimentation is difficult, computer simulations are useful as alternative experimental tools. In computational evolution experiments, scientists create virtual environments wherein evolution occurs. This is useful because these evolutionary processes are entirely separate from our own natural history, and are thus not shaped by the same random events. In addition, there are few limitations on the experimental design of an environment that is built with programming code, so that many different kinds of evolutionary processes can be compared.

One of the topics that are studied in evolutionary theory for which this computational approach would be useful is the evolution of how DNA sequence is structured. Over many generations, genes can become ordered differently, or they can be copied or deleted, so that the overall structure of the sequence changes. It is not fully understood what drives these changes and how they in turn affect the evolution of a species. Simulated evolution could be a useful instrument for researchers in this area, because the structural properties of the virtual DNA can be varied in different experiments. Thus, different aspects of the evolution of DNA structure can be explored separately and compared with alternatives, allowing specific hypotheses to be tested. However, there are some structural changes that can occur in natural DNA, which have never been implemented in digital evolution programs. In this dissertation, it is discussed how realistic DNA-like structure can be incorporated in digital genes for use in evolutionary simulations. In particular, we deal with the design of algorithms that imitate sexual reproduction, wherein two genetic architectures are combined into one offspring. Our goal is to enable future research in simulated evolution to uncover the causes and consequences of DNA architecture using virtual experiments.

A model of neuronal reprogramming

The human body is composed of trillions of cells. Despite the fact that all these cells have the same DNA, they come in thousands of different types with different shapes and functions. For example, a white blood cell is specialised in destroying foreign invaders in the blood, while a muscle cell can contract and expand with force. Cells are able to be so varied despite their common DNA blueprint because not all genes are equally active in all cell types, resulting in different kinds of proteins being present in the cell. Crucially, because the activity of genes is itself regulated by proteins, when some groups of genes are activated, they produce proteins that activate themselves and deactivate other genes. Cells maintain their cell type through this mechanism of self-regulating gene activity.

In the past decade, medical scientists have discovered that the specialisation of a cell into a particular type can be undone by destabilising this self-regulation in a lab. Specifically, some viruses can be genetically engineered so that they activate or deactivate a particular gene when they infect cells in a Petri dish, which is a standard laboratory technique in

biomedical science. By applying this technique to manipulate key genes which are crucial for the self-regulation that maintains cell type, cells can be reprogrammed to a different type. When reprogramming human cells, fibroblasts (the cell type that composes most of the skin, excluding hair follicles, blood vessels, etc.) are often used as the source material, because they can be easily gathered from a patient without invasive surgery.

A particularly well-studied and useful kind of reprogramming is to turn skin cells into neurons, which is the type of brain cells and nerves. By giving doctors direct access to a patient's own neurons in a Petri dish, this technology could enable new diagnostic tests and revolutionise the way some brain diseases are studied. Even more spectacularly, it may be possible to inject reprogrammed brain cells into the skull of patients suffering from neurodegenerative diseases such as Alzheimer's, which are characterised by a catastrophic loss of brain matter. Using cells that originate from the patient's own body sidesteps ethical issues associated with sourcing stem cells and organs from other humans, and also minimises the risk of the body rejecting tissue from an incompatible donor.

Fibroblast-to-neuron reprogramming of human cells was first performed in 2012, one year after a critical breakthrough in mice. Since then, much research has been done to improve the efficiency of the method, and some understanding has been gained into which gene activation and deactivation mechanisms are crucial to flip the switch between the two cell types. However, self-regulation of gene activity in the real cell is a complex system that can only be understood as a whole, not as a collection of individual gene regulation mechanisms. In a publication that is included in this dissertation, we created a mathematical model of this gene regulation system, which combines many different activation and deactivation mechanisms, based on the measured activity of certain important genes during a skin-to-neuron reprogramming process. This model is able to predict which experimental procedures can successfully turn fibroblasts into neurons, and gives us more insight in how exactly the cell genes decide to change from one cell type to another. Medical scientists may learn from our model to develop more efficient reprogramming methods in the future, and put neuronal reprogramming one step closer to medical application.

Introduction

The origin of life is a topic that has puzzled humans for millennia. Today, we reap the benefits of our ancestors' curiosity, having successfully solved some of life's greatest mysteries. We find ourselves in solidarity with all other living things on our planet, from humble fields of grass to majestic whales: all species were created gradually over time through an ongoing process of evolution. This creative capacity of nature is one of the most vital elements of life, on our planet and in general, and can be harnessed in fascinating technological applications.

The modern science of evolution began in earnest with the publication of Charles Darwin's *Origin of Species* in the middle of the 19th century; in full: *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* [1]. The significance of this publication was not in the idea that species have common ancestors and change over time; although living things were mainly thought to be organised in fixed categories throughout history, scholars had conjectured the evolutionary origin of species to various degrees of accuracy since the ancient Greeks. As Darwin emphasised in the full title of his book, his monumental contribution was the evidence with which he could support *natural selection* as a causal mechanism for evolution. Some traits are correlated with an increased ability to survive, gather resources, and reproduce; if successful individuals have more offspring that inherit their characteristics, successful traits will increase in the population over time. This mechanism, accumulated over generations, causes populations to adapt to their environment.

Darwin used the term natural selection because his contemporaries were already familiar with (what Darwin called "artificial") selection of crops and livestock. By carefully choosing individuals with favourable traits for breeding, it was known that farmers could influence populations of crops and livestock to enhance desirable qualities such as smaller seeds in plants or larger muscles in animals. Darwin's key insight can be formulated as stating that human interference ("man's use or fancy"), that is the creative intention from a conscious mind, was not essential to this process. Selection could arise spontaneously from the elements of the struggle for life, because of a statistical tendency of traits that aid in survival and reproduction to increase over generations. In other words, creation and invention of

complex forms and behaviours emerged from natural processes that could be studied and understood.

Because evolution follows as an inevitable natural consequence from a particular set of preconditions, similar evolutionary processes can be made to occur in synthetic systems. In particular, the efficiency of electronic computing makes simulation an excellent environment to host evolving populations over large numbers of generations. The resulting evolutionary processes are somewhere in between natural and artificial, in Darwin's sense, because creativity emerges naturally from within an artificially constructed system.

Computational evolution is useful to explore particular scenarios of evolution that are otherwise difficult or impossible to observe. In other cases, computational evolution is used to direct the creativity of natural selection to solving complex problems in a wide range of applications. The publications presented in this thesis are situated in different locations on the spectrum between theory and application.

In this introduction, I will begin by motivating the need for computational model systems, presenting evolution as a process that is not fundamentally tethered to biology. Then, I will briefly outline the general structure of evolutionary algorithms, which are the simplest virtual implementations of evolution; followed by a criticism of these algorithms and the need for alternative approaches, as well as a few examples of successful applications of evolutionary algorithms as a model for evolutionary dynamics. Finally, I will give an overview of the papers included in this thesis and sketch their place in the context of my research.

Computational evolution

The need for computational models of evolution does not stem from a shortage of data in volume. Our understanding of natural history today is informed by a comparison of taxonomical diversity that is orders of magnitude larger than the family of finches Darwin observed on Galapagos. We have access to an extensive fossil record, which allows even extinct species to be studied, and puts the diversity of life in a context that spans billions of years. More recently, molecular science and DNA sequencing have made it possible to study species and individuals in exquisite detail, which continues to be invaluable for understanding the dynamics of natural selection.

This wealth of observational data notwithstanding, experimentation is historically rare in evolutionary science. Unfortunately, natural evolution is difficult to manipulate due to its scale, so that evolutionary theory is almost entirely based on the examination of spontaneous occurrences. However, the practice of deliberately constructing particular scenarios that maximise the opportunity for falsifying theories, and suggesting new insights, is a crucial part of the scientific method in most disciplines. Natural observation does not give

such privileged access to precisely those scenarios that most increase our understanding of evolution.

In addition, the most detailed observational methods in evolutionary science, such as genomic and other -omics measurements, are only available as snapshots of a single point in time. Ideally, a process of change should be observed not just retrospectively, but dynamically as it happens. To some extent, ancestral states can be inferred from comparative analyses between extant populations. But many aspects of evolutionary history are permanently lost due to extinction or replacement by other adaptations, and the environmental causes shaping evolutionary events can often no longer be observed.

Experimentation is not entirely absent in evolutionary science. In fact, experimental evolution today is a growing and exciting field. The most famous example from this area of research is undoubtedly Richard Lenski's long-term experiment with *E. coli*. One of its impressive results is the emergence of a subpopulation with the ability to take up and metabolise citrate from its environment. The absence of an effective citrate metabolism is traditionally used in the determination of bacterial species to rule out *E. coli*, so that the experiment is technically the first ever to observe the appearance of a new species in a lab [2]. Evolution experiments such as Lenski's are excellent tools to study evolutionary dynamics [3]. Sequence analysis is now cheap and practical enough to allow investigating population of bacteria in experimental evolution at the level of individual mutations, so that researcher can directly observe the emergence, fixation, competition, and extinction of genotypes. These experiments have given great insights in some fundamental aspects of evolutionary dynamics, such as how adaptation is shaped by random events [2, 4]; how the presence or absence of sexual reproduction affects the fate of mutations [5, 6]; and the effect of having multiple copies of (a part of) the genome on evolutionary dynamics [7, 8, 9].

The limitation of these experiments is in the necessity to use the chemistry of biological organisms as the medium for evolution. This limits the parameters under investigation to the range of possibilities observed in nature, and restricts any inference that can be made to the scope of biological evolution.

It is not obvious that this is indeed a limitation. After all, evolution is typically thought of as a biological process, and the restriction of the scientific method to observable subjects is not specific to evolutionary science. Consider, as a thought experiment, an alien scientist called Portia, living on a planet where all species that once reproduced asexually have become extinct. Experimental evolution can teach Portia and her peers about the evolutionary dynamics of sexual populations, but not asexual ones. As a result, no understanding can be gained about the evolution of asexual life forms through natural observations, leaving Portia blind to the mechanisms that led to the emergence of her own earliest (sexual) ancestor, as well as to the possibilities of life on other planets, and the full range of options for the exploitation of evolution in technology. On Earth, human scientists of course can observe

both sexual and asexual populations, and it is possible to compare how both modes of reproduction affect evolution [5, 6]. However, it is undeniable that we are like Portia with respect to many other aspects of our biology.

A convincing argument for this claim is the simple observation that all life on Earth descends from the same ancestral species, called the last universal common ancestor (LUCA). Its genetics and biochemistry therefore served as a template for all subsequent life forms, including all extant species. However, we know that evolution is strongly bound by its own history. Once a solution to a particular problem is found, its efficiency improves, and it becomes nearly impossible for new alternatives to compete. Common features that are very likely inherited from LUCA include cellular structure, the central dogma (DNA transcribed to RNA according to a particular code, translated to proteins using a particular set of 20 amino acids), and a fairly complex core of biochemistry in cellular metabolism and communication. Indeed, LUCA was already a complex life form, almost certainly imposing a strong bias on the evolutionary options of its descendants. An alternative argument that also exposes such a bias is the fact that all known life forms live on the same planet, with the same chemical composition, solar radiation, gravity, and more generally speaking subject to the same physical laws. Because there is always a possibility that a commonality observed in biology is the result of such a bias, it is impossible to separate the fundamental features of life and evolution from the epiphenomenal, that is, "historical accidents" inherited from a common ancestor, or constraints imposed by the physical reality of our particular environment. These considerations are known succinctly as the $N = 1$ problem, signifying that there is only one known instance of natural evolution. The name borrows notation from statistics to evoke the statistical principle that meaningful inference of general properties is impossible if there are not enough independent observations. In other words, nature does not provide a wide enough basis of evidence to infer fundamental principles of evolution.

This argument highlights the necessity to generalise our understanding of evolution as independent from our own biology. While this need may appear abstract, similar generalisations are often useful in science. For example, Newton's realisation—immortalised by the myth of the falling apple—that the celestial movement of the planet is governed by the same general principle as the more mundane gravity of objects on Earth not only led to a more accurate model of astronomy, but revolutionised physics in general and ultimately proved to be a prerequisite for many concrete applications. Similarly, a more general understanding of evolutionary theory is likely to positively affect research in other fields. For example, evolutionary theory has already been critical to the development of many techniques in bio-inspired computing, such as evolutionary optimisation which is discussed in the next section. A deeper understanding of evolution can be expected to motivate more innovations in this area. Furthermore, evolution has relevance in medical science. Tumour growth is essentially a process of natural selection between individual cells, wherein the cooperation that is required for the functioning of a multicellular organism is no longer

prioritised. There is already a substantial overlap between the findings of experimental evolution and cancer research [10] (e.g. the recurrence of mutator phenotypes [3, 7, 11] and transient aneuploidy [12], or the consequences of clonal interference). Finally, in the context of astrobiology, NASA defines life as "a self-sustaining chemical system capable of Darwinian evolution" [13, 14]. Clearly, an understanding of evolutionary theory that is as independent as possible from terrestrial particularities is necessary to shape realistic expectations for future discoveries in this field. From a practical point of view, certain concrete questions, such as whether evolution is preconditioned by the presence of chemicals that spontaneously form compartimentalising membranes, can help accentuate or rule out planets and moons in the search for life. These same insights could narrow down the search for explanations about the ultimate origin of life on our own planet.

In order to get around the $N = 1$ problem and achieve a broader perspective on evolution, researchers can investigate a variety of evolutionary processes taking place in artificial environments, built for the purpose of experimentation. The goal in such experiments is not to imitate biology. Rather, similarity and dissimilarity compared to biological systems must be deliberately controlled in order to maximise the opportunity to falsify theories, and inspire new ones [15]. Computational environments are especially viable, because programming offers great freedom of experimental design, as there are no practical constraints related to physical realisation of the experiment. Although there are still limits due to finite memory and computational resources, computational methods can simulate relatively large populations over many generations, and gives researchers access to error-free measurements.

Due to this freedom of design, there is more than one way to implement evolution in a virtual environment. However, one particular template is the oldest, simplest, most popular, and most relevant for the publications in this thesis. I will sketch an outline for these methods in the next section.

Evolutionary optimisation

Evolution emerges from the interplay between reproduction, selection, and mutation. Scientists first implemented these mechanisms as computer algorithms in the 1950s and 60s, in an attempt to model evolutionary dynamics. Simultaneously, similar computer programs were published that were designed to evolve optimal or near-optimal solutions for complex problems of different kinds. These algorithms were given names such as *the genetic algorithm* [16], *evolution strategies* [17], *evolutionary programming* [18], and later also *genetic programming* [19, 20] and *evolution programs* [21]. The conceptual space separating these different algorithms has since been filled up by various extensions and generalisations, so that the distinction is no longer salient. Therefore, another name, *evolutionary*

algorithms [22], was introduced in the 1990s to refer to the generic approach, though *genetic algorithms* is also in popular use. *Evolutionary computation* is an umbrella term that includes more bio-inspired population-based optimisation methods. In this introduction I will keep to the basic principle of reproduction, selection, and mutation applied to virtual data. I will refer to this algorithmic template using the abbreviation GA (evolutionary algorithms), to distinguish it from more specific implementations or more general conceptualisations of computer-based evolution.

In GA, a group (population) of data structures (individuals) is evolved by the repeated application of fixed procedures that represent reproduction, selection, and mutation. Data structures in this case can be bit strings, permutations, real vectors, etc., but also strings or syntax trees representing computer code. The only required property is that each individual can be interpreted as a solution to a particular problem, and associated with a measure of solution quality called the "fitness" of the individual. The population is then gradually modified by adding and removing individuals, which involves a selection procedure that favours solutions with higher fitness over others, so that the population evolves towards progressively better solutions. In this way, GA leverages the inherent creativity of evolution by natural selection to evolve approximate solutions to complex problems.

The initial population is typically made up of randomly generated individuals. Then, modification of the population happens in discrete steps called generations. Each generation, a subset of individuals is 'killed', that is, removed from the population. By skewing the choice of victims to lower-fitness individuals, this culling is a mechanism of selection. Also each generation, some individuals are 'reproduced', meaning that a new individual is added to the population, using another as a template. Reproduction can also induce selection by preferring to use higher-fitness individuals as templates.

During reproduction, two 'genetic operators' are applied to create variation in the newly introduced individual. The crossover operator mixes in genetic information from a second template individual, similar to sexual reproduction in nature. The mutation operator modifies the newly produced individual with random changes, and is therefore likely to introduce new genetic variants into the population.

The GA approach is capable of producing complex solutions to difficult problems by applying the principles of natural selection to an initially random population. Unlike traditional algorithms, there is no guarantee that GA will find the best possible solution to a problem, even if allowed to run for an arbitrary number of generations. In many situations, however, this is not necessarily a weakness. In particular, for many problems it is most likely impossible to find the optimal solution to a problem in a feasible time frame, but even in such cases GA algorithms are often able to find approximate solutions quickly. The other main benefit of GA as an optimisation method is their extreme flexibility. Because they rely on evolutionary principles, and not specific domain knowledge, GA is widely applicable in

many different areas of computation. This is also the reason that the same fundamental algorithm has so many different names. The fact that the same principle can solve numerical problems, write computer programs, design neural networks, and evolve biological life reflects the remarkable generality of evolution.

Open-ended evolution

Algorithms in the GA family are creative in the sense that they invent high-fitness solutions within the constraints set by the programmer. This creativity emerges from the evolutionary process, and is able to design solutions without explicit human supervision. In fact, evolved solutions are often quite different from human ones. A common problem with GA is that the evolutionary process finds 'loopholes' in the fitness function, abusing some limitation of the simulation to cheat the programmer's intentions in ways that are sometimes surprisingly parallel to nature [23].

Yet, the true creativity of natural evolution is not in solving a pre-defined problem or set of problems. On the contrary, biological evolution produces endless variation with no apparent goal. Evolutionary processes that occur in GA, by contrast, move towards a certain solution, defined by the fitness function, and converge when the mutation and crossover operators can no longer improve the best individual in the population. The tendency of GA to find loopholes in the fitness function is a form of creativity, but also exemplifies the limitation of progress towards the simplest possible, good-enough solution.

Early pioneers advancing evolutionary simulation had not envisioned their proposals in this way. For example, Alan Turing was the first to publish about evolution as a computational concept, in the same paper that proposed the famous Turing test as a means to measure the ability of a machine to exhibit intelligent behaviour [24]. Originally, digital evolution was intended to produce a form of artificial intelligence, which would be achieved by mimicking the evolutionary origins of intelligence in nature. Similarly, early descriptions of artificial evolution by Von Neumann did not predict that their implementation would lead to a stagnant or declining level of complexity unless explicitly motivated towards a certain goal [25]. To date, despite many advances and more realistic approaches than GA, no artificially created evolutionary process has demonstrated the spontaneous tendency to produce a rich ecosystem of species, which is one of the most conspicuous and interesting aspects of evolution in nature [26].

Interestingly, the failure of existing evolutionary models to achieve this goal can be considered as the first triumph of computational evolution in circumventing the $N = 1$ problem. Early pioneers assumed that artificial implementations of evolution would spontaneously generate diverse and complex species because natural observations could not differentiate

between the evolutionary principles causing simple adaptation and problem solving on one hand, and spontaneous generation of variation and complexity on the other hand. Through simulations, we have been able to show that these two concepts are distinct, and therefore that evolution is not as well understood as it is commonly stated to be.

Indeed, it can be said that the origin of species is an unresolved issue at its core, because simulations have clearly demonstrated that mutation and selection are not the only necessary requirements for an evolutionary process to produce such a seemingly endless variety of species. What is it, then, that makes natural evolution fundamentally different to our simulated models, and why have we not been able to capture it? A concrete answer to this question is perhaps the holy grail of simulated evolution.

In the rest of this section, I will briefly discuss two general criticisms of the GA approach to simulated evolution that plausibly explain this shortcoming, and its lack of realism as a model for evolution. Both criticism are based on the observation that GA includes certain aspects of evolution as fixed laws of the virtual environment, whereas the natural analogues of these concepts are only present as abstractions emerging from more fundamental principles. In particular, GA implementations apply a fixed selection procedure with a predefined goal (fixed teleology); in a more abstract sense, GA presupposes a particular notion of what an organism is and how it is distinct from its environment (fixed ontology).

Fixed teleology

The evolutionary process of GA is induced by applying an algorithmic selection procedure to a population. This selection motivates the population to evolve, but in doing so imposes an end goal to evolution in the form of a fitness function. In other words, the GA approach includes a fixed teleology (goal, purpose, function) as part of the natural laws within the virtual environment. Clearly, an evolutionary process with a well-defined end goal cannot produce the kind of open-ended creativity of nature. In nature, the "fitness" of an individual emerges from the interaction between it and its environment (including other organisms) through physical processes. Which traits are fit or unfit depends on the environment, the individual's behaviour or role in it, both of which change over time as an organic result of the evolution of the population as well as other species in the same ecosystem. This gives rise to complex interactions such as predators and parasites employing other individuals as environmental resources, or plants evolving different shapes and sizes to avoid competing with their neighbours for sunlight.

Thomas Ray published these concerns, along with a resolution in the form of an alternative implementation of simulated evolution called Tierra [27]. His system uses computer programs as evolving individuals. However, it differs from GA that evolves computer code in that Tierran individuals are written in a custom assembly language, whose execution

directly affects the computer memory in which the code is embedded. Thus, the virtual memory serves as an ecological environment, filled with computer programs describing interacting behaviours which are executed according to a fixed set of laws, a computational analogue to the laws of nature. Selection emerges organically from the system through the organism's efficiency at reproduction, competition for computational resources, and ability to defy being overwritten by other programs. Finally, mutation is introduced by occasional failure to execute the correct instruction.

A fascinating result from this type of evolutionary experiment is that it is possible to spontaneously generate ecological interactions, and thus multiple species. In the case of Tierra, new species would arise that were more efficient at reproduction due to their smaller code size requiring fewer instructions to copy, which they could achieve by utilising instructions from nearby organisms. However, despite the fact that motivation to evolve arises internally from the system, that motivation is still finite: a fixed set of ecological interactions evolves, but converges towards a steady state, or evolutionary Nash equilibrium, in which no sub-population can further improve its situation.

Fixed ontology

The other type of fixity in GA is in the interpretation of individuals. An individual consists of a particular genetic code, which is interpreted in a fixed way. The shape of the individual is a fixed data structure, and there is a fixed boundary between one individual and its environment. This is also true in systems similar to Tierra: although some programs read and write code of others, each individual is read using a fixed code, implying a limited set of ecological interactions. The rules which allot computational time define where an individual begins and ends.

In nature, what an organism *is* has changed several times. For example, the evolution of multi-cellular organisms involved the complete cooperation of single cells to the point that they became one entity for the purposes of evolution. More often in natural history, it is not so much the delineation of the organism that changes, but its relation to the environment. Such changes are possible because the laws of physics and chemistry to not define a finite set of configurations as 'individuals', which can interact with another category of 'environment' in a finite set of ways. For example, the evolution of the eye did not result from a random mutation discovering that a predefined pattern in its DNA code would produce a sensor device with light as its input and cognitive awareness of an organism's surroundings as its output. Rather, evolution invented the eye, a whole new type of organism-environment interaction, by making creative use of the fact that the physical world contains electromagnetic radiation which interacts with molecules according to certain rules. These same physical principles enabled the evolution of photosynthesis. It is

thought that the richness of physics and chemistry, to let organisms and interaction mechanisms emerge naturally from the composite behaviour of smaller elements, is at the core of natural evolution's forms [28].

Researchers have attempted to produce systems with this same capacity. In artificial chemistry, reproducing entities, the individuals of evolution, are emergent patterns made up of smaller elements, which is thought to have the potential of cracking the problem of open-ended creativity and complexity in evolution. However, the design of a simplified chemistry that is rich enough to hold non-trivial evolving systems, but also simple enough to be computationally feasible, is challenging [29].

Evolutionary algorithms as models

The previous section summarises some of the known shortcomings of GA when it comes to simulating a bio-realistic evolutionary process, in the sense of generating what Darwin called "endless forms most beautiful and most wonderful". Nonetheless, this focus of GA to move towards a fixed evolutionary goal is not a detriment for all applications. On the contrary, in many cases it is a crucial asset. Indeed, it is precisely what allows GA to be used as optimisation heuristics, which rely on an appropriate tuning of the selection procedure to direct the evolution toward meaningful solutions of a predefined problem.

But also for modelling evolutionary processes, the rigid nature of GA implies great control over experimental conditions. Because the direction of selection and organism properties are defined as fixed terms by the programmer, they can be easily modified and measured according to the need to validate or falisify a particular scientific hypothesis. In GA, the programmer has full control over such things as the character of mutation, the frequency of sexual reproduction, and the details of genetic encoding, selection pressure and mechanics. In other words, many of the most fundamental parameters of evolution can be fixed or varied for the purpose of experimental design. For studying evolutionary phenomena that can be simulated by GA, they are easier to design, faster to compute, and more straightforward to interpret than alternative evolutionary simulations.

By choosing an appropriate implementation of the general template referred to here as GA, it is often possible to implement fascinating aspects of evolutionary dynamics in a minimal model. For example, natural organisms have evolved their own structure at various levels of organisation with respect to their own evolutionary potential. Ideally, small changes caused by mutations should cause the emergence of new variants so that the organism's descendants can more quickly adapt to changing environments (evolvability). At the same time, the negative effects of mutations should be minimised in order to maintain stable, successful offspring (robustness). It can be shown in GA that these phenomena can be

replicated simply by using a fitness function that changes over time, and suitably defining individuals so that the same phenotype can be encoded by multiple different genetic representations [30,31,32]. Evolution of evolvability involves a 'second-order' type of selection, wherein the long-term evolutionary success of an organism is not only decided by its reproductive efficiency but also by the fitness of its far-off descendants who inhabit a different environment, so it is fascinating to study these mechanisms in a minimal model.

References

[1] Darwin, C. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life* (1859).

[2] Blount, Z. D., Borland, C. Z. & Lenski, R. E. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proceedings of the National Academy of Sciences* **105**, 7899–7906 (2008).

[3] Barrick, J. E. & Lenski, R. E. Genome dynamics during experimental evolution. *Nature Reviews Genetics* **14**, 827–839 (2013).

[4] Lenski, R. E., Rose, M. R., Simpson, S. C. & Tadler, S. C. Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. *The American Naturalist* **138**, 1315–1341 (1991).

[5] Cooper, T. F. Recombination speeds adaptation by reducing competition between beneficial mutations in populations of *Escherichia coli*. *PLoS Biology* **5**, e225 (2007).

[6] McDonald, M. J., Rice, D. P. & Desai, M. M. Sex speeds adaptation by altering the dynamics of molecular evolution. *Nature* **531**, 233–236 (2016).

[7] Voordeckers, K. *et al.* Adaptation to high ethanol reveals complex evolutionary pathways. *PLOS Genetics* **11** (2015).

[8] Rancati, G. *et al.* Aneuploidy underlies rapid adaptive evolution of yeast cells deprived of a conserved cytokinesis motor. *Cell* **135**, 879–893 (2008).

[9] Yona, A. H. *et al.* Chromosomal duplication is a transient evolutionary solution to stress. *Proceedings of the National Academy of Sciences* **109**, 21010–21015 (2012).

[10] Sprouffske, K., Merlo, L. M., Gerrish, P. J., Maley, C. C. & Sniegowski, P. D. Cancer in light of experimental evolution. *Current Biology* **22**, R762–R771 (2012).

[11] Frank, S. A. Genetic predisposition to cancer – insights from population genetics. *Nature Reviews Genetics* **5**, 764–772 (2004).

[12] Dunham, M. J. *et al.* Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences* **99**, 16144–16149 (2002).

[13] Joyce, G. F., Deamer, D. & Fleischaker, G. *Origins of Life: The Central Concepts* (Jones & Bartlett Pub., 1994).

[14] Benner, S. A. Defining life. *Astrobiology* **10**, 1021–1030 (2010).

[15] Lehman, J. & Stanley, K. O. Investigating biological assumptions through radical reimplementation. *Artificial Life* **21**, 21–46 (2015).

[16] Holland, J. Adaptation in natural and artificial systems: An introductory analysis with application to biology. *Control and Artificial Intelligence* (1975).

[17] Beyer, H.-G. & Schwefel, H.-P. Evolution strategies – A comprehensive introduction. *Natural Computing* **1**, 3–52 (2002).

[18] Fogel, L. J., Owens, A. J. & Walsh, M. J. *Artificial Intelligence through Simulated Evolution* (Wiley, 1966).

[19] Brameier, M. F. & Banzhaf, W. *Linear Genetic Programming* (Springer Science & Business Media, 2007).

[20] Koza, J. R. Hierarchical genetic algorithms operating on populations of computer programs. In Sridharan, N. S. (ed.) *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, vol. 1, 768–774 (Morgan Kaufmann, 1989).

[21] Michalewicz, Z. *Genetic Algorithms + Data Structures = Evolution Programs* (Springer Science & Business Media, 1996).

[22] Yu, X. & Gen, M. *Introduction to Evolutionary Algorithms*. Decision Engineering (Springer, 2010).

[23] Lehman, J., Clune, J. & Misevic, D. The surprising creativity of digital evolution. In *The 2018 Conference on Artificial Life*, 55–56 (MIT Press, Tokyo, Japan, 2018).

[24] Turing, A. M. Computing machinery and intelligence. In *Parsing the Turing Test*, 23–65 (Springer, 2009).

[25] Neumann, J. V. *Theory of Self-Reproducing Automata* (University of Illinois Press, 1966).

[26] Banzhaf, W. *et al.* Defining and simulating open-ended novelty: requirements, guidelines, and challenges. *Theory in Biosciences* **135**, 131–161 (2016).

[27] Ray, T. S. Evolution, ecology and optimization of digital organisms. *Santa Fe Institute working paper 92-08-042* (1992).

[28] Taylor, T. Evolution in virtual worlds. In Grimshaw, M. (ed.) *The Oxford Handbook of Virtuality* (Oxford University Press, 2014).

[29] Hutton, T. Evolvable self-reproducing cells in a two-dimensional artificial chemistry. *Artificial Life* **13**, 11–30 (2007).

[30] Cuypers, T. D., Rutten, J. P. & Hogeweg, P. Evolution of evolvability and phenotypic plasticity in virtual cells. *BMC Evolutionary Biology* **17**, 60 (2017).

[31] Crombach, A. & Hogeweg, P. Evolution of evolvability in gene regulatory networks. *PLoS Computational Biology* **4**, e1000112 (2008).

[32] Hu, T., Payne, J. L., Banzhaf, W. & Moore, J. H. Robustness, evolvability, and accessibility in linear genetic programming. In Silva, S., Foster, J. A., Nicolau, M., Machado, P. & Giacobini, M. (eds.) *Genetic Programming*, Lecture Notes in Computer Science, 13–24 (Springer, 2011).

[33] Reeck, G. R. *et al.* "Homology" in proteins and nucleic acids: A terminology muddle and a way out of it. *Cell* **50**, 667 (1987).

Research outline

Paper I
Homology and linkage in crossover for linear genomes of variable length

Content We developed a new framework for scoring the performance of crossovers in GA with variable-length linear representations. The framework consists of two dual goals of recombination, which we call the homology score and linkage score. The homology score measures the ability of a crossover to preserve homologous features on the two parents; the linkage score is the ability to mix unique features on the two parents. Our perspective is a return to the root concepts of genetic recombination. In the literature on evolutionary computation, homology is often conflated with sequence similarity, but in biology and bioinformatics it is used in a more fundamental sense, referring to the shared ancestry of sequences (an important distinction [33]). We made the traditionally biological concept of homology measurable in a computational setting and used it to compare several different crossover strategies.

Personal note The idea to work on crossover for flexible linear genomes developed already in the first months of my PhD project, but we did not know at the time how to present our results and fit them into the literature. In the same vein, I tried to study the effect on different genome representations on evolutionary dynamics, but did not converge on a concrete result. After a long break, our image of flexible-length crossover became more focused when we homed in on homology and linkage as the central elements for a productive mental model of crossover with flexible genomes.

Publication status This paper was submitted for publication at PLoS One late in 2018. It was published there in the beginning of 2019 after minor revisions.

Contributions Code for this project was written by Carl Troein, based on existing code from an earlier project. I later (after publication) repeated the implementation (see paper IV). Henrik Åhl and Carl Troein developed the global alignment heuristic included in the paper before I arrived. Under my and CT's supervision, HÅ also worked on a Bachelor project that examined different binary representations of numbers, which seemed related at the time. I have been responsible for data analysis and writing of the manuscript. All aspects of the paper developed through discussion with CT.

Paper II
A quantitative model for cellular decision making in direct neuronal reprogramming

Content This publication is a medical application of systems biology. We measured mRNA concentrations of critical genes during a fibroblast-to-neuron direct reprogramming experiment. Then, we integrated known gene interactions governing the reprogramming process into a gene regulatory network, converted it to a differential equation system, and used an evolutionary algorithm to fit the parameters of the system to our data. We found that the prevailing understanding of the gene interactions responsible for cellular decision making in neuronal reprogramming were insufficient to explain our data. Exploring various alternative models, we found that a simple reinterpretation of the interaction between two genes could produce a greatly improved match between our model and data. In addition, this model could correctly predict the outcome of other experiments.

Personal note I joined this project with the intention of combining theory and application of genetic algorithms in my research, when learning that Victor was looking for someone to implement a non-linear model fitting strategy. I developed a programming framework with the ability to systematically turn custom descriptions of gene regulatory networks into executable code of models with a built-in parameter optimisation. Ultimately, however, the interesting parts of this research turned out to be in applied systems biology. The main challenge for this publication was to extract predictions that would have meaningful consequences for medical science, but also remain conservative in our estimations, which was difficult because experimental constraints permitted only a small data set. I believe we managed to partially surmount this issue by incorporating a lot of existing domain knowledge into our analysis to augment our data, even though the known facts about the gene regulatory network we modelled were heterogeneous and sometimes ambiguous.

Publication status This paper is currently in the review process with Scientific Reports. In response to a request from one of our reviewers, we are waiting for our collaborators to perform an *in vitro* experiment that might validate our model.

Contributions I have been responsible for the formulation of models, developing computational methods, implementation, parameter optimisation, interpretation of results, and writing most of the manuscript, under supervision of Victor Olariu. Our experimental partner, Janelle Drouin-Ouellet, did all lab work, including gathering data, and wrote the corresponding parts of the manuscript. VO and JD originated the project. In parallel with my GA approach, Viktor Drugge developed a parameter optimisation method based on simulated annealing as his Master project, which is not included in the publication.

Paper III
Perfect edge-transmitting recombination of permutations

Content In many crossover strategies for permutations, the goal is to create offspring that consist mostly of edges (pairs of adjacent elements) that are inherited from the parents. Most algorithms do this imperfectly, thus occasionally introducing new edges. A few crossovers implement correct inheritance of edges, but are not able to produce a large variety of possible offspring. We developed a novel crossover algorithm that achieves perfect preservation and produces all possible offspring under that constraint. There is a recognised trade-off in the literature between the correctness of the offspring on one hand, and the variety that a crossover can generate on the other hand. Due to the large variety of offspring our crossover could produce, our findings suggest this trade-off is less stringent than previously thought.

In a sense, paper I and paper III reify the same concepts in different settings. Correct inheritance of either homologous features (in variable-length linear genomes) or parental edges (in permutations) is contrasted with the extent of mixing of genetic information in the offspring. We find, in both cases, that even though there is a logically clear trade-off between these dual goals, in practice a well designed recombination algorithm can recombine genetic information quite well without compromising on correctness.

Personal note This idea originated as a side project of paper IV, when it became clear that a problem I was working on could be reformulated as a permutation crossover. This project also uses the same code base as paper IV. Because I came into this topic from an unusual angle, the manuscript went through various iterations when I realised that concepts I had invented were already described in unfamiliar literature. However, due to this *ab initio* approach, I missed out on some discouraging beliefs that were established in literature, and was able to describe a new relationship between several existing ideas.

Publication status This publication has been submitted to IEEE Transactions on Evolutionary Computation.

Contributions All parts of this research are my contribution, with valuable insights gained from discussion with Carl Troein.

Paper IV
Genetic recombination of linear genomes with flexible structure

Content We generalised the framework presented in paper I to linear genomes which are not only variable in length due to insertions and deletions, but can also be modified with translocations and duplications of genetic sequences. This modification introduces all biologically relevant types of genetic variation to the genetic algorithm setting, excluding those related to the double-strandedness of DNA and its organisation into chromosomes. In the presence of variations in copy number of genetic sequences, homology is a many-to-many mapping between the two genomes, instead of a one-to-one mapping, demanding a more general formulation of the homology and linkage scores developed in paper I. In addition, the cut-point method can not be applied as-is to such genomes by crossover strategies that intend to recombine structural variations on the two parents. We apply concepts from paper III to design a more nuanced generalisation of the cut-point method fit for this purpose.

Personal note This paper was conceived as a successor to paper I. Because one of the concepts introduced here turned out to be a difficult problem with more general application, it became a separate publication (paper III). The work on this and paper III was preceded by a complete reimplementation of the code that Carl Troein wrote for paper I, and a replication of its results to ensure the new framework was comparable.

Publication status This work has not been submitted for publication.

Contributions All parts of this research are my contribution, with valuable insights gained from discussion with Carl Troein.