Adrian Woolfson

Synthetic life

In 1540, the German cartographer Sebastian Münster published the first accurate map of the African subcontinent. Contrary to the Ptolemaic view – in which Africa, Antarctica, and part of Asia formed a single southern land mass known as *Terra Incognito* – Africa emerged as a discrete entity in Münster's representation. Improvements in shipping and navigational techniques, such as triangulation and calculation of longitude, made this map possible. These methods enabled transoceanic voyagers to locate their positions in the absence of landmarks, thus facilitating the exploration of new areas.

Münster's map was also remarkable for its unusual depiction of Africa's wildlife. In his exposition, the Dark Continent teemed not only with conventional creatures like elephants and parrots, but also with mythical ones such as one-eyed Monoculi. The incorporation of imaginary beasts suggests Münster's anticipation of the synthetic future of life, and indeed his tacit appreciation of the fact that material existence represents only a fraction of natural and artificial possibility.

In the spirit of Münster, it is possible to explore the idea of compiling a 'library of all possible creatures,' a database of DNA sequences of all species, past, present, and future. With this database, we may be able someday to recreate extinct species and even create entirely new ones.

Like Münster's rudimentary yet imaginative map of Africa, the library – also called 'DNA Sequence Space' – has a distinct mathematical reality. But in contrast to conventional terrestrial domains, this space is boundless, and so appears on first inspection to defy cartographical representation or even rational exploration. Fortunately, like the corporeal continent that underwrites Münster's accurate but nevertheless fanciful depiction, at least a small portion of this apparently limitless landscape can be mapped. This is significant, as it is this region that – much like the former coalfields of the industrial North England – may most economically yield rich seams of potential life. What we need, though, is a method of predicting the location of these 'coalfields,' which contain the

Adrian Woolfson is CEO of ProteinLogic and teaches medicine at the University of Cambridge. He is the author of "Life Without Genes" (2000) and "An Intelligent Person's Guide to Genetics" (2004).

mathematical structures that are able, in principle, to encode processes of life, as well as the means of deciphering them.

Navigating the contours of this complex, rugged, and tortuous mathematical terrain is onerous and not without its own dangers. But its exploration and eventual mapping will ultimately be far more important than the discovery of the Americas, Antarctica, or any other continent. For within this vast and mostly uncharted Borgian space, all life's possibilities may be found: the morphological secrets of existing life as well as of every potential living thing. The moment we complete this project, life will become dissociated from the natural evolutionary processes that have shaped it from its inception. New artificial modes of creation will then supplement, perhaps even supplant, such conventional historical mechanisms.

Having sketched the nature of the project, its complexity, and its potential impact, I would like to trace an outline for its execution. Like the peripatetic mariners that navigated the Earth's uncharted oceans without the luxury of maps, we will hardly be able to achieve success in a task of this magnitude without an armory of appropriate technological innovations – the sextants and longitudes of our day. Fortunately, a selection is already close at hand.

But before we can go over the steps in creating this database, we must first acknowledge that this expansive mathematical edifice is not metaphysical. Instead, it is grounded in mathematical Platonism, which recognizes the immutable and eternal mathematical reality of logical propositions. In this case, the logical propositions pertain to the description and construction of organisms, whether living, extinct, or never before realized. A simple thought experiment can help illustrate this idea.

Imagine, for instance, a creature that is half tiger and half dog. Now imagine that this tiger-dog becomes extinct. The fact that the tiger-dog once existed means that, in principle, a genome sequence capable of computing it exists. Thus, the potential to construct this creature predated its existence and indeed persists beyond its extinction. But even if the tiger-dog had never existed, the possibility of its existing at some point in the future would have remained intact.

It turns out that a tiger-like dog did once exist in Tasmania until only very recently. A visitor to the Hobart Zoo in 1933 might have marveled at the last living example of the Tasmanian tiger, also known as the Thylacine or *Thylacinus cynocephalus*. This curious and somewhat unlikely creature resembled a large, short-haired dog with a tail reminiscent of a kangaroo's. Its yellow-brown coat had a distinctive striped pattern, as did its rump and tail. These markings gave the Thylacine its tiger-like appearance. Concerned about the loss of domestic animals to this successful predator, the Van Diemen's Land Company put a bounty on the Thylacine around 1830. Numbers of the Thylacine declined rapidly, and before long, the species was close to extinction. The Hobart Zoo Thylacine was the last-known captive example; and despite occasional reports of sightings in the wild, the Thylacine is now thought to be extinct. All that remains is a disparate collection of photographs, skeletons, and a short grainy black-and-white film in which the Hobart Zoo Thylacine is seen pacing frenetically back and forth in a painfully inadequate and featureless enclosure.

In what sense might we describe an extinct Thylacine as having a timeless

Platonic reality in mathematics that transcends the existence of any individual example? A few samples of preserved Thylacine flesh and bone persist in museum collections around the world. Indeed, a team headed by Don Coglan of the Evolutionary Biology Unit at the Australian Museum managed to recover intact DNA from some specimens. The fact that the attempt met with only limited success reflects the poor state of the starting material, rather than any intrinsic obstacle to the resurrection of this extinct organism.

Had the DNA been better preserved, it would have been possible, at least in theory, to reconstruct a Thylacine genome in its entirety. We could then have placed the reconstructed genome into an appropriate egg, either natural or artificial, which we might have then used to generate some or all of the structural and functional characteristics of a historical Thylacine. Naturally, some features might not have been encoded in its DNA. Its cry, for example, was probably transmitted through cultural rather than genetic means. Such aspects of the historical Thylacine are therefore likely to be irretrievable. This is not to say, though, that we could not infer the broad features of these types of characteristics.

Once we can create a physical facsimile of an extinct Thylacine by amplifying and pasting together fragments of DNA recovered from the flesh of a formerly living example, it becomes clear that natural material is not a prerequisite for such activities. Indeed, if a DNA sequence that corresponded exactly or closely to that of a historical Thylacine were discovered by chance in DNA Sequence Space, this purely artificial sequence might just as readily provide the genomic substrate for the generation of a facsimile Thylacine.

From here, it follows that, like the Thylacine, which once existed and has subsequently become extinct, there must be countless other creatures that have not had the chance to exist and for which we could find perfectly plausible DNA sequences. We might discover such sequences mathematically on a computer or define them artificially from first principles.

The history of life on Earth has simply been too short to realize more than a fraction of the wealth of possibilities contained with DNA Sequence Space. The exploration of this space by natural processes like natural selection has furthermore been subject to constraints and historical contingencies. These have ensured that the history of life on Earth, like a river snaking through an unpredictable and perilous mountain landscape, has taken a well-prescribed, narrowly defined, and ultimately highly constrained pathway. In some cases, we have overlooked potential creatures simply because we had no easily accessible route by which to reach them. In others, chance events may have extinguished rivulets of life, ensuring that we neither explored in the first instance nor revisited certain regions of DNA Sequence Space.

Once we have established the framework of this grand enterprise, we must first address the technological issue of how to compile a sufficient – or better, exhaustive – database of DNA sequences. This task requires sequencing the genomes of all known living things and of any recoverable genetic material from extinct creatures. Sequencing is the process that deciphers a genome's unique combinatorial string of four different DNA building blocks. This string chemically encodes the core information that directs the assembly and operation of

living things. This information basically translates into the proteins that underpin the structure and function of all known life on Earth.

Whereas current sequencing technologies are able to decipher the genomes of small organisms like yeast and bacteria in a matter of months, a complete inventory of the genomes of all known species, many of which are far larger, will require significant improvements in DNA sequencing technology. Without such innovations, a task of this magnitude is both unrealistic and untenable. Such technological advances, however, are forthcoming; and it is easy to imagine a time in the near future when we will be able to obtain complete genome sequences of any complexity within a matter of minutes, or even instantaneously.

The compilation of an exhaustive database will also necessitate an extensive trawl of every available niche and microenvironment, so as to capture as many examples of the different species on Earth as is practically possible. But once we have obtained a database of genome sequences ranging from the most insignificant Amazonian beetle and obscure microorganism to antelopes and zebras, we will be able to begin searching for predictive architectural features that facilitate the interrogation and subsequent interpretation of target unknown sequences.

Following the successful compilation of an extensive, and ideally complete, DNA sequence database, we will need to establish a universal algorithmic machine capable of computing the structure and function of any organism from the abstract mathematical notation of its genomic structure. For example, if we fed the genomic sequence of a giraffe into this hypothetical machine, it should recognize that the inputted sequence represents a giraffe-like creature, or simply, a giraffe. Similarly, when we enter the genomic sequence of a flamingo, the machine should conclude that the sequence belongs to a pink, long-necked bird – or better, it should infer that this is a flamingo.

Preferably, the algorithm should predict not just the morphology of the organism, but its biochemistry and behavior as well. And a truly universal algorithmic engine should be capable of computing with a certain degree of accuracy the morphological structure of any organism, even if the sequence is artificial or the organism has never formally existed.

The construction of this universal algorithmic machine is materially contingent on the successful completion of the DNA sequence database. Only then will the algorithmic machine have exposure to enough sequences for it to be adequately trained. Success in this domain thus depends on the maturation of the science of comparative genomics, namely, the process by which the features of the genome of one species are systematically compared with the features of another. Ideally, the discovery of structural genomic patterns correlating with specific macroscopic features, such as a fin or a wing; microscopic features, such as the architecture of a liver sinusoid as opposed to that of connective tissue; and molecular features, such as a metabolism based upon oxygen as opposed to hydrogen cyanide, will be performed *in silico* by specialized machine code capable of automatically discerning such fundamental relationships.

Once we have a machine able to compute the likely structure – both internal and external – of the organism the sequence represents, we will need to develop a search engine that can navigate DNA Sequence Space efficiently, and sys-

tematically compute every possible sequence housed within the space. In this way, we could, in principle, get descriptions of the sequences of all actual and potential living things. After the exhaustive exploration and testing of a significant portion of DNA Sequence Space, we could then use the information to form a comprehensive 'Life Map.' It would contain all of the inferred biological possibility extracted from a circumscribed region of DNA Sequence Space. Although the majority of the creatures unearthed in this process would not exist or have existed, we would assign each potential organism a coordinate on this map.

The next issue to address is whether it is possible to translate the abstract logic of synthetic genomes into the molecular hardware of living creatures. In the case of simple synthetic organisms that mimic the fundamental features of their natural asexual counterparts, generating artificial cells capable of accommodating the artificial genomes should suffice. In the case of more complex synthetic organisms that mimic the features of sexually reproducing organisms, however, we will have to develop a way to construct artificial eggs so as to enable the information in synthetic genomes to be read.

The development of a predictive algorithm able to compute DNA sequences of uncertain provenance leads logically to the possibility of a methodology capable of both designing and constructing new genomes from first principles. Using such constructional principles, we should be able to generate organisms with entirely new properties, some or all of which may never have been encountered in the natural world, either individually or in combination. The introduction of such novel properties may have many benefits, including, for example, advances in medicine, improvements in food and energy production, and the colonization of hostile environments both on Earth and elsewhere.

But it is also necessary to recognize the inevitable occurrence of 'impossible' creatures within mathematical space. Although morphologically plausible, such creatures would be incapable of initiating or sustaining a life process. For example, imagine a DNA sequence that encodes an oak tree the height of the Empire State building. Such an organism might be morphologically plausible, but it is unlikely to function at the physiological level. Other creatures, though on first inspection appearing sustainable, might on closer reflection be shown to be incapable of existence within the parameters of physics and chemistry. A butterfly with wings the size of tennis courts might survive in a weightless environment, but it is unlikely to fly when exposed to the gravitational pull of the Earth. This is not to say, though, that such creatures might not flourish in some as yet unidentified alternative world.

Synthetic life might, in fact, employ software and hardware technologies different from the DNA 'software' and protein 'hardware' that have formed the informational and structural substrates for life probably from its inception. Although evolution by natural selection has chosen these technologies as the core technologies of all life, it is possible to imagine alternative technologies that might form the basis of the essential informational and structural chemistry of living things someday. But to discover such creatures we may need to hunt in places other than DNA Sequence Space.

Besides these physical constraints, we must acknowledge limitations to the computability of DNA sequences. The

entirety of this enterprise is predicated on the assumption that the class of logical propositions representing the structure and function of living things is computable, at least in principle. Although this is likely to be the case many times, we should not be greatly surprised when things occasionally turn out differently. The intrinsic noncomputability of certain sequences has two principal components. The first is an exclusively mathematical consequence issuing from Gödel's theorem: there are likely to be 'undecidable' logical propositions that cannot be shown definitively to be either true or false. And just as there are logical propositions whose solutions defy computation, there are likely to be genomes and aspects of living things that are similarly not computable.

Second is the fact that the DNA sequence alone does not contain all the important components of the information of living things. The phenomenon of genomic imprinting is responsible for silencing some genes by selective 'epigenetic' chemical modification, in a process known as methylation. The loss of such essential information by the representation of a sequence isolated from its methylation imprint might in some or many instances render interpretation impossible.

This insurmountable constraint prescribes a finite limit to the mathematical space we can navigate. It is possible to imagine, though, that we might one day overcome this inability to incorporate epigenetic factors by the mathematical construction and subsequent exploration of an 'Epigenetic Space.' This is an even more complex computational task, however, and may consequently, at least for the time being, be unattainable.

Despite these constraints, it should be possible to commence what is likely to be the greatest enterprise of the twenty-first century. With the basic universal algorithmic machine and synthetic tool kit in place, humanity will at that point enter a new age of mathematical cartography: the constructional, and principally computational, science of synthetic life will enable the delineation of qualitatively different types of maps than those created by conventional cartographers. These new virtual maps will allow us to catalog the creatures that, like Ebenezer Scrooge's Christmas ghosts, inhabit both the past, present, and future, and which populate the knotted and twisted mathematical landscapes of the 'library of all possible creatures' – a single definitive and exhaustive inventory of all living possibility.

It is impossible to predict the consequences of innovations within the field of synthetic life, and of the paradigmatic shift from natural life that is generated by historical processes of natural selection, to synthetic life that is designed *in silico* and subsequently constructed from first principles. The end of natural selection as the principal agent of speciation will be an unprecedented milestone in human existence. Needless to say, the consequences will be far-reaching, as the distinction between natural and artificial will become nothing more than a historical curiosity. Indeed, the question as to whether something is natural or artificial might itself become quite absurd (in a manner reminiscent of the question mooted by Alan Turing of whether it might be possible to demonstrate that a machine able to convince us that its behavior is the product of conscious awareness is actually capable of conscious awareness).

This technological transition depends, at least for the time being, on the preservation of the natural world and on the systematic documentation of all its con-

tents. This absolute dependence on the extraction of genetic information from multiple and diverse genomic sequences highlights the importance of every single species on Earth, however obscure and apparently irrelevant. It consequently demonstrates *par excellence* the importance of preserving our environment and each of the niches within it, and of maintaining an archival example of every known species on Earth in a genomic Noah's Ark. This biological zoo of genomic material holds the key to the exploration of the mathematical zoo, upon which the continuation of the human species, and indeed all life, may one day depend.