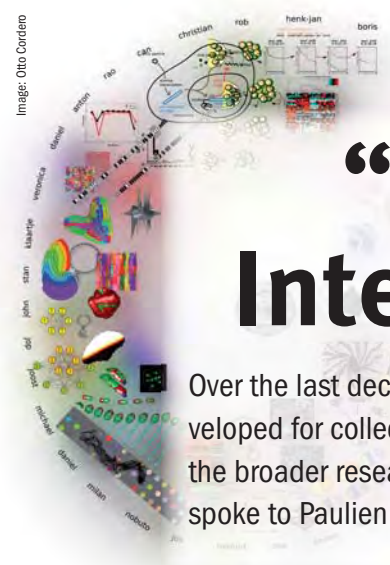

Image: Otto Cordero

A conversation with Paulien Hogeweg, Utrecht, NL

"Thinking in the Most Interesting Simplification"

Over the last decades, bioinformatics has been commonly associated with the computational methods developed for collection and analysis of genomic data. But the concept, as originally defined, rather refers to the broader research area studying life as a dynamic information processing system. Alejandra Manjarrez spoke to Paulien Hogeweg who, together with her colleague Ben Hesper, coined the term in the early 1970s.

Lab Times: Let's start by talking about your early steps into conceiving information processing as a property of life and how this led you to define the concept of bioinformatics.

Hogeweg: First of all, I indeed thought information processing should be in the centre of the analysis. At that time, most theoretical work was in quantitative terms, instead of pattern and information processing. So, coming from the field of theoretical biology, I had the feeling that the focus there should change. Then, Ben Hesper actually thought that we had to give it a name, so he came up with the term 'bioinformatics'.

"It is very remarkable that the first university computer had been bought by the biologists, rather than mathematicians or physicists."

At that time you were already a theoretical biologist. What were you doing?

Hogeweg: Actually, at that time I was just starting as a PhD student. I had come from ecology and, after finishing my Master degree, I decided I wanted to go into theory. Initially, there were no positions available, so I started as a volunteer at a theoretical biology lab in Leiden. There I started to work with Ben Hesper and it was then when we started to think about information processing. That's how it all began.

What projects were you carrying out?

Hogeweg: I didn't have a specific topic to begin with. Having come from the ecology side, I was interested in pattern recognition and pattern analysis of datasets – in fact, what people call now 'bioinformatics' in the strict sense. But we realised that, in order to understand the processes, we needed not only static data analysis but also dynamical modelling.

Was this already done with computers?

Hogeweg: Yes. But at that time, computers were still far away in bigger rooms and you couldn't really reach them. In Leiden, there was a computer available for students and it is very remarkable that there the first university computer had been bought by the biologists, rather than mathematicians or physicists. So I started to work a little bit with the data we had but Ben, coming from the crystallography field, had done a lot of programming so he was able to do that job.

You told us you come from an ecology background, how did you start doing theory in the area of molecular biology?

Hogeweg: Well, from very early on I was interested in how genotypes map into phenotypes – and I'm still very much interested in that. Actually, the very first thing I did for my PhD studies was related to this mapping. My boss at that time was Aristid Lindenmayer and he had devised the so called L-systems, which are grammar rewiring systems with a number of symbols, for example 0 and 1, and rules; for example, 1 transforming into 11 after every recursive step. Lindenmayer thought about them as an abstract mathematical framework for development but they can be translated into geometric structures and we studied them as prototype for genotype-phenotype mapping. We had, for example, a system of 1s and 0s that also included brackets '[]', which represented side branches. Thus, within this framework of symbols and their translation into branching structures, we took all possible rules of transformation and studied how the variation of the branching patterns were related to the grammar that produced them. In this very simple way, you can see already that this is a very nonlinear process: a minor change at the genotypic level, the grammar, could lead to a major change in the pattern. By the way, the programme and the 3D graphics that we did for this were all very fancy for the time and the computer we were using for it was behind this big wall. For entering, you just had to give your card to the operator and, after a while, you got it back. So I'll never forget that, while we were working on these fancy branching patterns, the man in charge came and said with surprise, "What's this?" He had never seen such a thing.

It was actually in this period when you moved to Utrecht, right? Did you have computer facilities when you arrive then?

Hogeweg: Well, I started in Leiden in 1969 and in December 1970 I moved halftime to Utrecht. From then I slowly shifted until I was here full-time. When we arrived, we had this old Tektronix [see photo on p. 26]. It was a present for Lindenmayer because he moved to Utrecht University but he couldn't use it and, in the end, it was mine. It was connected to the central computer of Utrecht University but I also could use it to connect to the Leiden computers.

Paulien Hogeweg

Since the late 1960s, Paulien Hogeweg has pioneered different conceptual and computational approaches to study life. Her interests include eco-evolutionary processes in space, animal behaviour, prebiotic evolution, morphogenesis and genotype-phenotype mapping. She has published more than 150 articles in the field of theoretical biology and has been a member of the editorial board of the *Journal of Theoretical Biology*, *Bulletin Mathematical Biology*, *Biosystems* and *Artificial Life Journal*, among others. Her career has been mostly developed at Utrecht University.

You continued working on topics related to molecular biology but how did the sequencing era affect your research?

Hogeweg: At the beginning of the 80s, the first public data sets became available. So we got indeed this very first set from the EMBL and one of the very first things we did was to create a multiple alignment programme. It was actually the first programme of this kind [*J Mol Evol,* 20:175-86]. It basically worked by generating a provisional phylogenetic tree, by clustering sequences based on pairwise alignments. Then, the next step was aligning sequences progressively along the tree, to obtain a multiple alignment. This strategy later became well known through the Clustal programmes. With this tool we started to do analysis of specific RNA sequences and, later, focused on evolutionary studies with RNA.

Hogeweg anno 1980

I guess that this was also the starting point of one of your main research lines nowadays: prebiotic evolution.

Hogeweg: Well, we started to work on this topic around the 90s. We had two main approaches. One was focused on the RNA genotype-phenotype mapping and how that evolves, and the other was studying the possibility that all sequences together could be able to contain the information that a single sequence could not. That was our hypercycle work with pattern formation in space.

I am aware that space is important in your research, particularly talking about local interactions but many people prefer working with differential equations, where space is not really taken into account. Why is it important to include space when we establish models of biotic systems?

Hogeweg: Information processing only takes place when things interact and this happens mainly when things are close by – except at the moment: we are interacting while being far away.

...yes, the power of the internet!

Hogeweg: Well, so while working with pattern formation in ecosystems, we were very early on interested in cellular automata [CA] as a modelling tool. At the end of the 80s, I wrote a paper about the use of CAs for ecological research. On the other hand, also in the 80s, we started to develop individual-based [IBM] or agent-based models, where the model entities are information processing individuals rather than populations as in ordinary differential equations [ODE] or patches as in CA. We explored this, first applied to behaviour but later also in an ecological context. However, computer power was still limited and we could only do it in relatively small models. Nonetheless, I

"It is of course interesting to talk to experimentalists but I like to keep my own focus."

would say that we could do enough, especially related to behavioural models.

Can you give us an example of this approach applied to behavioural studies?

Hogeweg: In the early 80s, we had an interesting collaboration with Cor van Honk, who was then looking and working with bumblebees. He had observed a hierarchical organisation in these insect societies but the data didn't make sense to him. We therefore started to analyse them and we could see that there was indeed a certain social structure in the bumblebee colony. It was possible to subdivide into elite and common workers and we were expecting to find some distinction between both, size for example. But we were not able to do any kind of correlation by just studying physical characteristics. We then thought that it might be something behavioural. And to understand how this was generated, we developed our first big individual-based model. In this model we started with one 'bumble', the queen, who later laid eggs, eggs became larvae, larvae became adult and then they became workers, taking care of the next generation of eggs laid by the queen, etc. Thus, in the model we included the population dynamics of a colony plus a behavioural thing named the "DODOM interaction", which was a kind of ritualised dominance behaviour observed in real nests. In this event, bumblebees meet, they touch each other's antennae and, then, one goes away and the other goes straight. We postulated that, when they meet, there is a chance that one wins and the other loses and, in our model, that chance depended on an internal 'D' [dominance] parameter. This parameter started for everyone on the same value. For instance, if two bumblebees that have just born meet, they have the same D value, so they have the same chance of winning. But the one who does win gets a little bit of a boost and its D value becomes a little bit higher while the other's becomes a little bit lower. The more unlikely the outcome – the one with lower D value winning – the higher the boost. That was the main idea of the model. We analysed the output of it in a similar way as real nests are studied, in order to make a valid comparison. We found that adding this behaviour was sufficient to obtain a social structure as observed in the real ones [*Behav Ecol Sociobiol* 12:271-83].

Now a museum piece, the Tektronix computer from the early 70s.

"We need to simplify our models, otherwise we won't understand them."

Listening to this, I understand you had collaborations with experimental groups, how often was this and is it something you still do?

Hogeweg: At that time it was about fifty-fifty. We had collaborations but we also did things on our own. Now, most research is not in direct collaboration with experimentalists. When I want certain data, I can find them either in the literature or in databases or wherever. It is, of course, interesting to talk to experimentalists but I like to keep my own focus. Moreover, there are things that are difficult in the experimental context that can be trivial on a theoretical basis and the other way round, so in direct collaborations it is hard to keep step. However, recently I had a very nice experience collaborating with experimentalists. Veronica Grieneisen, a PhD student jointly working in my group and the experimental molecular genetics group of Ben Scheres, also at Utrecht University. We looked at how the maximum auxin gradient in plants influences root growth [*Nature*, 449:1008-13].

" There are things that are difficult in the experimental context that can be trivial on a theoretical basis and the other way round."

How much has your research changed from what you were doing before? What are your interests now?

Hogeweg: Now, considering the huge changes in available data, computer power and conceptual developments, we are still very much interested in this genotype-phenotype mapping and how it evolves. In early times, we were not able to really do large scale evolutionary studies because we lacked computer power. But now, my main interest is how the fact that the genotype maps into the phenotype is an evolved property. In RNA, we can say that it is still a physical property related to energy minimisation. There, many different sequences can result in the same structure and it has been proven that, if you evolve towards a certain structure long enough, the coding will be such that you will have many neutral neighbours – similar sequences giving rise to the same 2D structure. However, in present-day organisms there is, of course, much more that you can do. For instance, we have gene regulatory networks that evolve, deriving in questions such as how mutations go through these regulatory mechanisms to become an overexpressed protein that then becomes phosphorylated, etc. Physicochemical processes are certainly involved but how has that network structure evolved? Most people, when they think about evolution, either think about adaptation to an environment or in neutral events but the important process of evolving this mapping is mostly kept out of consideration. I am very much interested in how it evolves and how this influences the breath of evolution, for example, the adaptation to a certain environment.

You have worked in ecology, behaviour, evolution, molecular biology – a wide range of fields in life sciences. How challenging, as a theoretical biologist, is it to keep up with the knowledge in these different disciplines?

Hogeweg: I think we are really lucky now with the internet. Comparing to what had to be done a number of years ago – breaking into libraries, for example – it's so easy now to find your way into literature! I remember very well that, not so long ago, I used to go once a month to the physics library. For this, as a biologist, you had to go and ask the secretary for the key. I used to browse the latest journals there and I will never forget that this was how I accidently found the extended cellular Potts model of Glazier and Graner. This was a very interesting discovery and later became key in many of our works related to modelling morphogenesis. Now, however, when I want to know something, I browse a lot through the internet. It is so easy to move from one article to another and you can just update yourself so fast with the state-of-the-art in a certain area. This doesn't mean I can do

the experimental work of that field but I know what data are out there and where they come from.

Computer power, easy access to literature... new technologies seem like paradise for theoretical biologists or bioinformaticians, nowadays. What are the new challenges?

Hogeweg: The big challenge is still to find a sensible way to model the complexity of biological organisms. That is certainly not a solved problem yet. We need to simplify our models, otherwise we won't understand them. In that sense, the challenge is to think in the most interesting simplification. For a long time, it was thought that the best approach was to have a single level of description, while I believe that it might be more accurate to have multiple-level models, where higher levels feedback to the lower ones. In other words: looking into vertical models instead of horizontal ones. So, the individual oriented models, the cellular automata, the cellular Potts model, have been, from my point of view, important steps towards this kind of simplification. Now, as I mentioned, this genotype-phenotype mapping as an evolved property can be certainly a new way of tackling the complexity of biological systems. Getting to a level of complexity comparable to what we find in biotic systems is still not happening. Many systems biologists now say that they are doing it, but I honestly believe that nobody knows how to do it. One way people can approach this is through systems of differential equations but then we have the enormous challenge of parameters that we don't know how to solve. Nowadays, we have the high throughput technologies, where we can possibly get closer to parameters and I do expect some progress there, but not in the short-term. I really don't think the solution is around the corner.

"Many systems biologists now say that they are getting to a level of complexity comparable to what we find in biotic systems but I honestly believe that nobody knows how to do that."

What could be a solution?

Hogeweg: One thing we did in that area is to ask whether evolution can help us. The system example we studied was the lac operon in bacteria. People had assumed that this regulatory system behaves as a bistable switch, both in experiments and models. So we took one of these big ugly models with 25 or 30 differential equations. We also used the parameter values people had been measuring in this particular system for 50 years but in our model the parameters of the operon were able to evolve. Given the background of the rest of the metabolism, we wanted to know if the system could evolve towards a bistable switch or not. The results showed very clearly that the system goes away from a bistable switch. It was interesting because, at the same time, there was a paper announcing (again) that the lac operon was a bistable switch. However, reading the supplementary material very carefully, we found a little sentence where the authors clarify that they had done the experiments with artificial inducers but when they tried with lactose, they couldn't find the bistability and they concluded it was probably an experimental flaw. Our conclusion was that it just behaves different because of the artificial inducer and that, in the natural system, this is not a bistable switch. Actually, when we supplied our evolved promoter function with the artificial inducer rather than the lactose, we did find bistability [*Biophys J,* 91:2833-43]. That is just one experiment, and even for this well-studied system, the problem of parameters was certainly reduced. I think that including the evolutionary perspective in the modelling can be partially a solution for these systems of many parameters but I cannot say it is a general solution.

So, there's a lot more work to do but, luckily, as you argue in your recent paper "The Roots of Bioinformatics in Theoretical Biology" [PLoS Comput Biol, *7(3) e1002021], *the concept of bioinformatics is coming back to the original idea of information processing, isn't it?*

Hogeweg: I think that's true and I'm looking forward to interesting research around this. I actually quoted Paul Nurse there, a British cellular biologist who got the Nobel Prize. He recently gave interviews and suggested that we really have to look into informational processing. He argued that, once we start doing this, we will see many counterintuitive results. He said this might be an important 'quantum leap' in Biology, considering this approach as the next step...

...but you've certainly been doing it for a long time and ***Lab Times*** *thanks you for sharing with us your journey along studying life from this interesting – and certainly not so common – perspective.*

Interview: Alejandra Manjarrez