FEED ME DATA

The iPlant programme was designed to give plant scientists a new information infrastructure. But first they had to decide what they wanted, finds **Heidi Ledford**.

In April 2008, Richard Jorgensen found himself in front of a group of expectant researchers gathered in Cold Spring Harbor Laboratory, New York. The pressure was on: Jorgensen had recently been placed in charge of iPlant, a US$50-million, five-year programme funded by the National Science Foundation (NSF). The project was supposed to tackle the biggest computation questions in plant biology — and his job was to unite the community behind the effort. "The plant sciences are being given the opportunity to lead," he told the assembled crowd.

But after two days of the meeting, the researchers were not clear where that leadership would be taking them. Brainstorming sessions had repeatedly slipped into guessing games as participants tried to infer what Jorgensen wanted as the project's 'grand challenges'. "I'm still not sure I understand, in all honesty, what this is and what this is supposed to do," said June Medford, a plant synthetic biologist at Colorado State University in Fort Collins, during one session. Jorgensen, who is based at the University of Arizona in Tucson, was refusing to offer concrete suggestions for what the grand challenges should be for fear of unduly influencing the participants. "I'm officially agnostic," he said, squinting in the midday sun on the last day of the meeting. "My role is more like being a therapist, in a way."

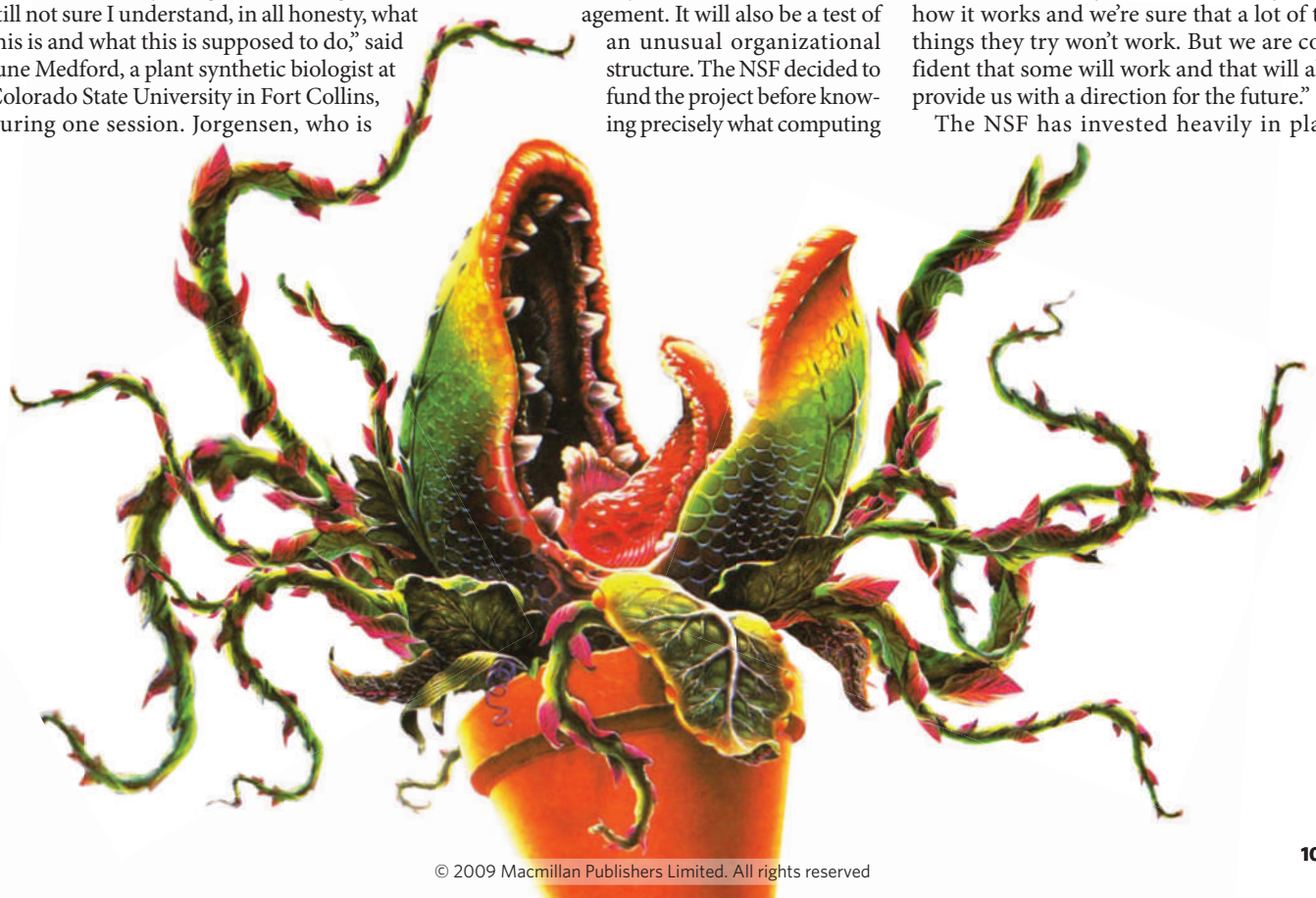
One year and several workshops later, Jorgensen's therapy seems to be paying off. Groups in the plant community now have half a dozen grand-challenge projects that tackle everything from evolutionary genetics to the mathematical modelling of plant development — and earlier this year the iPlant organizers announced which two they will pursue first. If the initial projects work out, the whole effort could be extended for another five years with an additional $50 million.

iPlant will not just be a test of data-management. It will also be a test of an unusual organizational structure. The NSF decided to fund the project before knowing precisely what computing tools it would be paying for, leaving the scientists to decide. "People would ask, 'Why do we have to go to all of this trouble? You know what we need — just go build it'," Jorgensen says. "But the NSF decided that you have to have a buy in from the users first, or you're going to build something they don't really want. And I think they're right."

Model infrastructure

The outcome of iPlant could have repercussions for the broader biological research community, as it is also struggling to integrate and process a torrent of computational data. The iPlant model is one that the NSF may want to use for constructing 'cyberinfrastructure' in other fields, says Peter McCartney, the NSF's programme officer for iPlant. "It's a grand experiment," he says. "We don't really know how it works and we're sure that a lot of the things they try won't work. But we are confident that some will work and that will also provide us with a direction for the future."

The NSF has invested heavily in plant



ILLUSTRATIONS: GEFFEN COMPANY/RONALD GRANT ARCHIVE

biology over the past decade, particularly in high-throughput 'omics' projects. The agency has funnelled about $200 million into a project to determine the function of all 25,000 or so genes in the model plant *Arabidopsis thaliana*, and has also contributed to a large, interagency programme for genomics projects in other plants. These and other efforts have generated rich databases and computational tools that are open for the community at large to use, but programme managers at the NSF realized that a problem loomed ahead. "Here was a community in which there had been a substantial investment in tools, but there was some concern about how these tools were going to work together, and how they were going to persist long-term," says McCartney.

This problem is not unique to the plant sciences. Researchers typically build databases the quickest way they know how, without necessarily considering whether they will work with other databases. And once a database is in place, it is very difficult to alter it, says Graham McLaren, programme leader in bioinformatics data management for the Generation Challenge Program of the Consultatative Group on International Agricultural Research in Texcoco, Mexico: "I would say it's easier to get someone to change their spouse than their database."

Some research communities in the physical sciences, such as astronomy and particle physics, tackled these issues long ago by agreeing on a unified cyberinfrastructure. But the problem is relatively new in the biological sciences. When the NSF looked to help by building a cyberinfrastructure project in the biological fields it funds, it decided that plants were an ideal place to start, McCartney says. Plant biology covers a very broad and disparate community that studies many model organisms, making data incompatibility a particularly acute problem. Integrating these data could have societal benefits in terms of agriculture and conservation — and, says Jorgensen, the field already has a long history of collaborative projects.

"It's easier to get someone to change their spouse than their database."
— Graham McLaren

Humble beginnings

Jorgensen was drawn into the field in 2006 as he was preparing for the end of a five-year tenure as editor-in-chief of the *Plant Cell* journal, and making plans for a sabbatical in Mexico. He started having second thoughts when he saw a call from the NSF for proposals in plant-science cyberinfrastructure. "It was a new way to contribute," he says. "It just seemed like one of the most challenging things that I'd encountered and a unique idea." He decided to find out whether his colleagues at the University of Arizona would be interested in taking on the challenge. When his team was awarded the grant, Jorgensen put the sabbatical on hold to coordinate the project. At his suggestion, it became known as iPlant.

Richard Jorgensen went from building genetically modified petunias to plant cyberinfrastructure.

UNIV. ARIZONA

Jorgensen sees iPlant as an opportunity to unify a plant-biology community that has long been split along disciplinary lines — and he knew from the start that recruitment would be key to the project's success. He had to convince ecologists and evolutionary biologists that iPlant was not just about molecular biology and 'omics. He also had to sign up molecular biologists, who quickly assumed the collaboration was just another bioinformatics project. "'Sure, I'll send my bioinformatician to the meeting', was their response," Jorgensen says, "but it's not just the bioinformaticists that we need." He then enticed in dedicated computational biologists and software engineers. "What the NSF has done is forced a kind of shotgun marriage between biologists and computer scientists," he says.

If anyone can unite the community, many say that Jorgensen is the researcher to do it. Well-known but unassuming, he already has the respect of many scientists for his academic achievements and diplomacy. In the late 1980s, for example, Jorgensen and his colleagues at the biotechnology firm DNA Plant Technology in Oakland, California, decided to develop petunias with richer colours by boosting expression of a pigment gene called chalcone synthase. To their surprise, many of the resultant flowers were white: rather than enhancing expression of chalcone synthase, they seemed to have shut it down entirely. Jorgensen left the company to continue investigating the phenomenon, which he named 'cosuppression', and even conducted experiments at his own home for a while before he was given a lab at the University of California, Davis.

Nobel thoughts

Some years later, researchers would realize that some cases of cosuppression, which had turned up time and again when researchers tried to make transgenic plants, were due to a process called RNA interference (RNAi). When Andrew Fire and Craig Mello were awarded the 2006 Nobel Prize in Physiology or Medicine for their work on RNAi in the nematode *Caenorhabditis elegans*, some researchers complained that early contributions made by plant biologists, including Jorgensen, had been overlooked. Jorgensen demurred, pointing to the contributions that Fire and Mello had made to working out the mechanism behind RNAi. "The Nobel prize is not really about making scientists famous — it is about making science interesting and accessible to the public," he wrote in a letter to the journal *Science* at the time (R. Jorgensen *Science* **314,** 1242; 2006). Richard Jefferson, a plant molecular biologist and founder of CAMBIA, a non-profit research institute based in Canberra, Australia, said of Jorgensen: "I think he's the smartest man in plant science — and the most intellectually generous."

Jorgensen needed all those qualities to negotiate his way through the first year of iPlant and to overcome researchers' initial uncertainty. After the Cold Spring Harbor meeting, the NSF solicited proposals for grand-challenge workshops, and selected five that were held over the course of the next year. From those workshops, and a sixth held by the National Center for Ecological Analysis and Synthesis in Santa Barbara, California, emerged six grand-challenge teams, some of which united dozens of researchers. In April this year, the iPlant board of directors — comprised, at Jorgensen's request, of plant biologists and computer scientists rather than iPlant leaders — recommended two projects to focus on for the next two years.

The board gave highest priority to a project already familiar to many plant biologists: developing a plant 'tree of life' to determine the evolutionary relationship between taxa. The NSF has long supported such efforts, including the 'Deep Green' plant phylogeny project of the late 1990s, and the broader Tree of Life project, which included all taxa and will reach the end of its funding in the next year. For Rob Last, a plant biologist at Michigan State University

in East Lansing and associate chairman of the iPlant board of directors, prioritizing the tree-of-life project was a practical decision. "This is a community that has worked together a lot. It has strong leadership," he says. "And the tree of life is a really nice coordinate system that ultimately we should be hanging our data on."

The iPlant proposal differs from previous tree-of-life projects in that it does not focus on data collection — iPlant is not allowed to distribute funds for this. Instead, it will concentrate on infrastructure and technology development, says project leader Mike Sanderson, a plant systematicist at the University of Arizona. These computing tools should allow researchers to extract gene sequences and morphological traits from a wide variety of databases, and compile that information into comprehensive evolutionary family trees. The aim is to build trees with the data available for about 50,000 plant taxa, even though the long-term goal of plant phylogeny projects is to generate a tree of the more than 500,000 taxa that are known.

"The NSF has forced a kind of shotgun marriage between biologists and computer scientists."
— Richard Jorgensen

Lack of support for data collection was a common complaint in the early days of iPlant. At the initial Cold Spring Harbor meeting, the question came up repeatedly: why develop tools to unite incomplete databases of varying quality, when what the community needs is more complete data of high quality? With time, and with the knowledge that funding for iPlant does not eat into the NSF's plant-research budget, the community has come to accept the idea. "What we consider high-quality data today may not be considered high quality five to ten years from now, so where do you start?" says Steve Goff, iPlant's director of community interactions. "You have to work with what you have at the time."

The second prioritized project — called 'genotype-to-phenotype' — will explore how variations in genetic sequence relate to the appearance and behaviour of plants and was built by cherry-picking parts of various proposals. One part will make a stab at using new computational tools to study genetic and environmental influences on when a plant commits to flowering — a topic that has long interested farmers. Another aspect will build models of photosynthesis with the ultimate aim of learning how to convert 'C_3 photosynthesis', the kind present in many crop species, into the more efficient form called 'C_4 photosynthesis' that is present in corn and some other plants. A third focus will be on the effects of genotype on responses to climate change. "This is really a unifying grand challenge in biology," says Last.

The selection process has inevitably left some researchers disappointed. One grand challenge proposal aimed to tap into about 500 million digital records from herbaria and ecological study plots around the world, showing the occurrence of plants in different climates and environmental conditions. Linking these data could allow ecologists to monitor how species distributions and habitats have changed over time, with the long-term goal of understanding the impact of climate change. "The data explosion in ecology is enormous," says University of Arizona plant ecologist Brian Enquist. Although iPlant directors have said they hope to tackle some aspects of the proposal, Enquist worries that a piecemeal approach will not suffice. He agrees with the decision to prioritize the tree of life project, but points out that it already has a long history of steady funding. "We have just a kazillion ecological data points, but we have nowhere to go to combine them."

Spreading out

Enquist and his colleagues may have another place to go if iPlant is successful. "[The ecological community] is another that we'd probably be very interested in seeing if this kind of approach would help," McCartney says. But he acknowledges that it's still too early to judge whether iPlant will be a success. That will start to become possible when the first few computing tools are built, probably late this year, and researchers are testing them out. At the moment, those involved are determining where to start, breaking down the broad-sweeping challenge proposals into tasks that can be completed in the next two years.

As for Jorgensen, he's finally taking that sabbatical. With the stress of the project launch behind him, he is hoping to have a little more time for his own research, before a new round of grand-challenge solicitations starts. He is also involved in planning an iPlant meeting for next year, "to give the community a chance to look at what we've started", he says, and to talk about what else the project should do as it matures.

Clearly growth lies ahead. The question for researchers is whether they can grow iPlant into the framework they need: one that is big, strong and fast enough to support the data that they are also busy cultivating. ■

Heidi Ledford is a reporter for *Nature* in Cambridge, Massachusetts.


FANCY/VEER/CORBIS


P. BRYE/ALAMY


P. DUMAS/EURELIOS/SPL


FANCY/VEER/CORBIS


FANCY/VEER/CORBIS

Genomics projects on *Arabidopsis thaliana* (above) and other model species are churning out data.