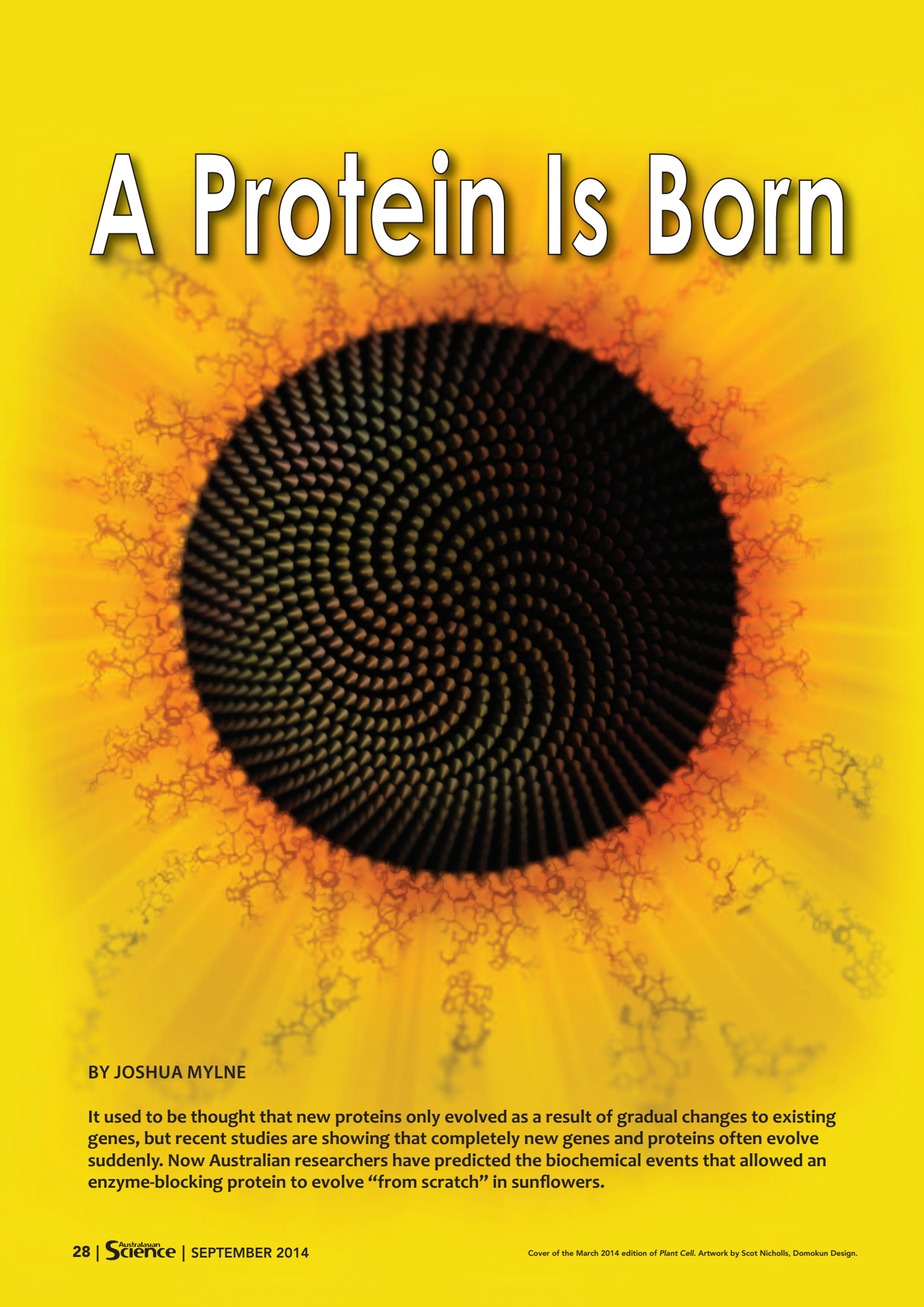
A Protein Is Born

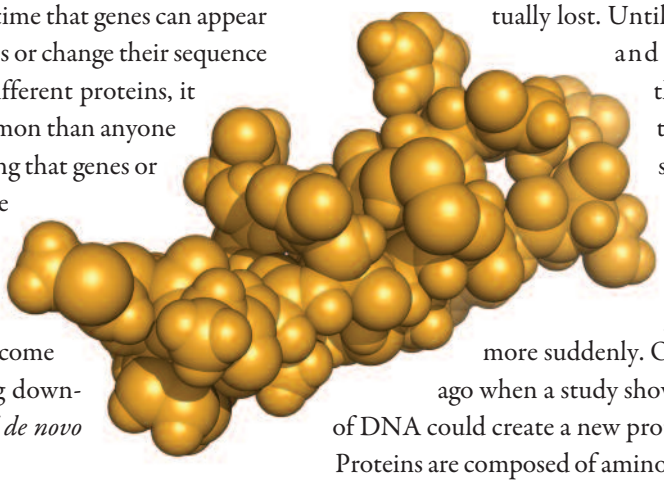


BY JOSHUA MYLNE

It used to be thought that new proteins only evolved as a result of gradual changes to existing genes, but recent studies are showing that completely new genes and proteins often evolve suddenly. Now Australian researchers have predicted the biochemical events that allowed an enzyme-blocking protein to evolve "from scratch" in sunflowers.

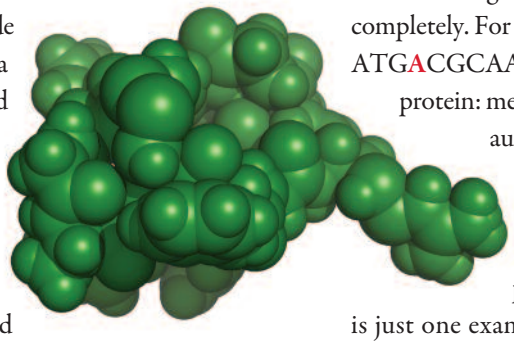
Cover of the March 2014 edition of *Plant Cell*. Artwork by Scot Nicholls, Domokun Design.

Each genome and the proteins it encodes is what makes every organism unique. Although it has been known for a long time that genes can appear suddenly in some species or change their sequence and code completely different proteins, it now seems that this is far more common than anyone expected. Some people are even saying that genes or proteins that appear *de novo* are the major driver of biological innovation.



Most *de novo* gene studies focus on DNA and the new messages that come from them, but we've been looking downstream at a very specific example of *de novo* protein evolution.

The common sunflower has an unusual gene called *PawS1* whose encoded protein is processed into a storage albumin protein. Buried alongside albumin in the PawS1 protein is a second, completely different protein that blocks digestive enzymes.

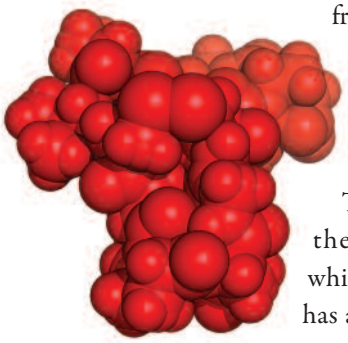


Seed storage albumins are made in great abundance to serve as a degradable source of nitrogen and sulfur for germinating seeds. By contrast, digestion-blocking proteins protect seeds from grain-eating insects.

We were desperate to know how such different proteins ended up being cut from the same precursor protein. By tracing the evolutionary history of *PawS1*, we could predict which steps enabled *PawS1* to make two proteins instead of just the usual one.

Gene Duplication and Divergence

The panapoly of proteins in each organism is encoded by its many genes: most plants and animals have 20,000–30,000. Far from being permanently set, the gene content of organisms can vary due to rare errors that get made as DNA is replicated.



Throughout many genomes there are duplicated regions, which means the organism now has an extra copy of these genes.

The *PawS1* gene has evolved in the daisy family over millions of years to create a panapoly of seed peptides. The four structural models shown on this page were determined using nuclear magnetic resonance spectroscopy.

These "spare" genes are free to evolve by diverging gradually into new and useful genes, or can become disrupted and eventually lost. Until recently, this "duplication and divergence" model was thought to explain much of the observed protein diversity.

Although this view predominated, some studies showed how proteins could evolve much more suddenly. One of the first was 30 years ago when a study showed how a single extra base of DNA could create a new protein.

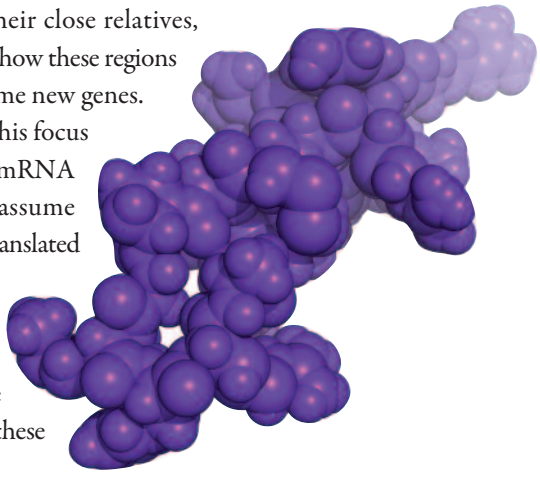
Proteins are composed of amino acids, and three nucleotide bases (or "letters") of DNA determines which of 20 possible amino acids is eventually made. For example, the DNA sequence ATGCGCAAGGTC will encode a small protein comprised of four amino acids: methionine–arginine–lysine–valine.

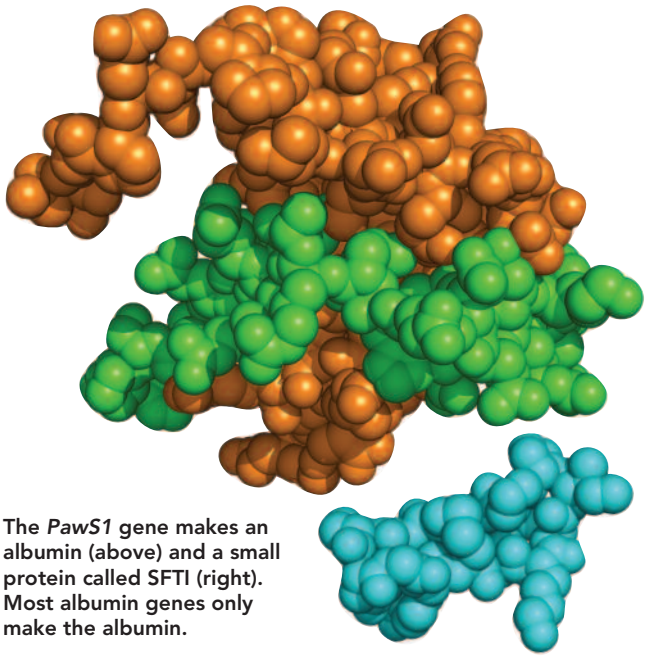
But a single extra DNA base can change the protein completely. For example, with an extra "A", the DNA sequence ATGACGCAAGGTC will now encode a completely different protein: methionine–tryptophan–glutamine–glycine. The authors of this study in 1984 showed how a new enzyme was created by such a change.

A similar process of *de novo* protein evolution is widespread in viruses and is called overprinting. This slippage of the coding sequence is just one example, but there are many ways to evolve new proteins.

Fresh Thoughts on the Origins of Genes

What really changed opinion on the origins of genes was a series of recent studies by different groups that used massively parallel DNA sequencing to identify genetic messages (mRNA) that were unique to specific races or to a single species. Sometimes these unique mRNA messages mapped to DNA regions that were inactive in their close relatives, implying that somehow these regions of DNA had become new genes. Most studies like this focus on DNA and the mRNA they encode, and assume that the genes are translated into proteins. A few studies have gone on to prove that proteins were indeed made from these *de novo* genes.





The *PawS1* gene makes an albumin (above) and a small protein called SFTI (right). Most albumin genes only make the albumin.

An Extra Protein Buried in a Protein

Sunflowers have a gene call *PawS1* that encodes a protein that, at first glance, looks like any other protein that gets matured into a seed storage albumin. On closer inspection, though, one section of its sequence is bigger than usual. It transpires that this overlarge section is in fact an extra protein called SFTI that is clipped out of PawS1 when it is processed into storage albumin.

SFTI is a small protein ring that can block the digestive enzyme trypsin. Nothing like SFTI has been found outside sunflowers, and SFTI has no similarity to its adjacent albumin, being one-tenth its size, a different shape and function.

So how did this completely different and extra protein appear within the sunflower albumin precursor PawS1? This case presented an opportunity to understand the steps that might lead to the birth of a new protein.

Tracing the History of a Gene

To work out where SFTI came from, we cloned many *PawS1* genes from plant species related to sunflowers, specifically the plant family Asteraceae. This family expanded very quickly in the past 2–18 million years and boasts more than 23,000 species. By teaming up with US taxonomists who knew the evolutionary history of the Asteraceae, we were able to connect the genes we cloned with the ancestry of the plants to piece together a model for how SFTI evolved.

We found the situation in sunflowers was much more ancient than we expected, and that the dual-product *PawS1* gene was at least 18 million years old. Many of the small buried proteins being made by *PawS1* genes in Asteraceae looked generally like SFTI, but only the ones from plants most closely related to sunflowers were similar enough to inhibit trypsin too.

We also found a gene we called *PawL1* (*PawS1-Like 1*) that encodes a protein that makes a storage albumin and has an oversized section like PawS1 with many of the protein sequence hallmarks of SFTI. However, in PawL1 this region is smaller and lacks a crucial pair of cysteine residues, and does not get matured into a stable protein.

Taken together, this suggested that SFTI evolved stepwise thus: a normal seed storage albumin gene was subject to an insertion (*PawL1*) that had the potential to make a clipped-out protein that was not stable. The evolution of an internal bond between two cysteine residues then stabilised the buried peptide (PawS1). In the species that eventually led to today's common sunflower, the buried peptide changed in sequence to enable trypsin to bind it. This is why the PawS1-derived buried proteins may be found in several thousand *Asteraceae* species, but the trypsin-blocking SFTI types are found only in a few dozen species closely related to the common sunflower.

We believe this is one of few studies that has retrodicted a biochemical sequence of events for the *de novo* evolution of a protein. To confirm this hypothesis will require a much more in-depth analysis of the *PawS1* and *PawL1* genes of Asteraceae seeds.

As systems go, this dual-destiny protein may be a good one for understanding how a stable protein can evolve *de novo* and it will be exciting to compare it with other examples as they emerge from this rapidly advancing field.

Joshua Mylne is an Australian Research Council Future Fellow appointed jointly to the School of Chemistry and Biochemistry & The ARC Centre of Excellence in Plant Energy Biology at the University of Western Australia.