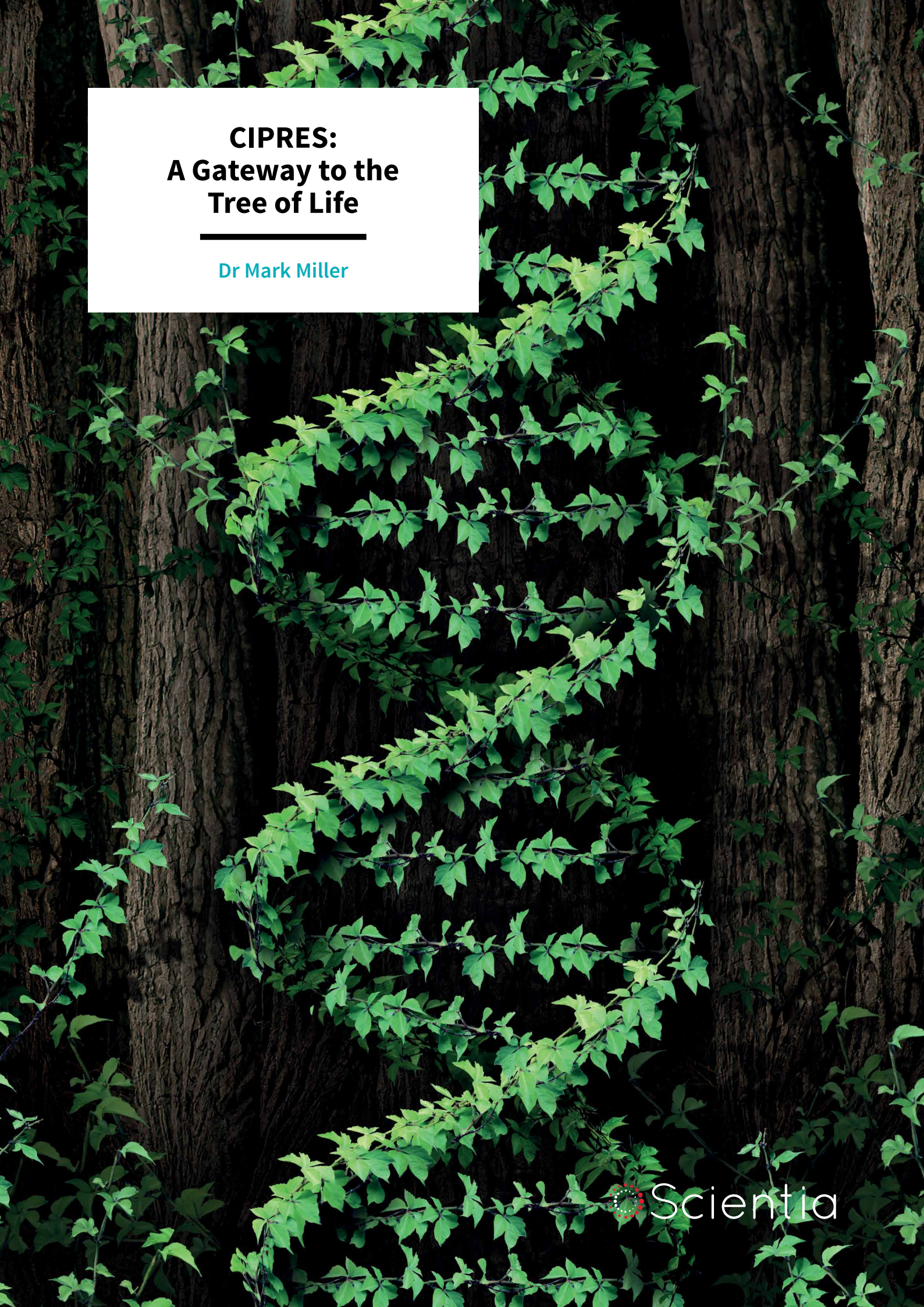


CIPRES: A Gateway to the Tree of Life

Dr Mark Miller

Scientia

CIPRES: A GATEWAY TO THE TREE OF LIFE

Modern genetics provides vast sets of information which can take weeks to assess. Developed by **Dr Mark Miller** and his colleagues at the San Diego Supercomputer Center, the CIPRES Scientific Gateway solves this problem, by bringing supercomputer-powered analysis to researchers across the globe.

Phylogenetics

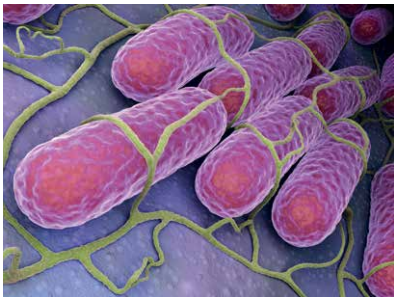
Cross the world and you will run into a dazzling array of life, from desert-dwelling mice to arctic birds, from sun-loving flowers to deep-sea microbes that will never see the light of day. Yet for all of these differences, every organism we find is related to every other one, somehow, even if we need to look back several billion years. But how do we know this? How do we decide whether two creatures are closely or distantly related? How is it even possible to know that *Tyrannosaurus Rex* is more closely related to a chicken than a Great White Shark is to a tuna fish? This is the science of phylogenetics – the study of evolution and how different species give rise to others.

The idea of a 'tree of life' is not a new one. Observant people have been placing animals and plants into groups and building relationships between them for hundreds of years. Yet until Darwin and his ground-breaking work in establishing the theory of evolution, no-one had thought that these groups may have developed from a common ancestor. Phylogenetics was built from this realisation – a scientific approach to plotting out the branches and leaves of the tree.

Early researchers built these plots from common features and appearance – a family of birds that looked like ducks and quacked like ducks were, in all probability, related to ducks. This works well in some situations but is often confused by convergent evolution – widely-separated organisms in a similar environment will often evolve towards a similar 'optimal' form. The classical example here is the eye, which has been independently developed by both vertebrates (mammals, birds, etc) and cephalopods (squid, octopus). Many more exist, and this makes things complicated for the scientist trying to determine how different species fit together.

This is where DNA sequencing enters the picture, which involves reading the 'language' encoded in our DNA. This technique can be used to record the entire sequence of an organism's genetic makeup, or its 'genome'. This genome can then be compared to that of other species to determine how similar they are to one another – a pair of sequences that are highly similar are likely to come from closely-related species, while those that are very different are likely from widely-separated ones.

Naturally, this is more complicated in the real world. Some DNA sequences are very important for all organisms and so remain similar across many



different species – genes involved in the production of RNA and proteins, for example, are almost universally conserved. Thus, scientists need to carefully choose the genes to be compared, and indeed a reliable phylogenetic study requires the analysis of many different genes. And therein lies the current problem.

Drowning in a Sea of DNA

Thanks to the many advances in DNA sequencing, it is now possible to rapidly and cheaply collect vast amounts of genetic data. It currently costs a patient between $1000 and $2000 USD to have their genome completely sequenced. Scientists focusing on smaller sections of DNA can have sequences determined for less than the cost of a coffee. 'DNA sequencing techniques have become very cheap and very fast,' notes Dr Mark Miller of the San Diego Supercomputer Center. 'The phylogenetic field has been catapulted into a world where discovery is no longer limited by the amount of available data.' But all of this data comes with a cost, as this flood of information needs to be *understood*.

Collection of sequence data is fast, simple, and cheap, yet actually analysing this data is a serious bottleneck for the scientific process. 'As more DNA sequence data has become available, routine analyses that used to run in a day on laptop now require a week or a month,' notes Dr Miller. 'The alignment of this data, and analysing the differences are computationally hard. The amount of time required increases exponentially as the length of the DNA sequence and the number of species being examined increases.'

This is a problem for many researchers, who mostly work from standard desktop computers or laptops that lack the processing power to do multiple analyses in parallel. 'Performance is impacted by this kind of analysis,' says Dr Miller, 'and usually only one such analysis can be run per computer.' Pushing the calculations onto dedicated computer clusters is possible, but requires that the researcher actually has access to such a high-performance machine and that the right software is installed, updated and maintained.

CIPRES: The Power Plant of Phylogenetics

This is where CIPRES enters the picture. The CIPRES Scientific Gateway (derived from '**C**yber **I**nfrastructure for **P**hylogenetic **RES**earch') was initially thought up as a response to these problems faced by the scientific community. Developed and improved by Dr Miller and his colleagues over the past decade, it has turned into a central hub for much research going on today. CIPRES provides a web-based access point to a number of up-to-date and efficient phylogenetic programs, all running on local supercomputers. Researchers simply need to register, upload their data, choose from a number of approaches, and then go and grab a cup of coffee. Preferably a coffee to go, as many of the analyses will be finished within minutes rather than the hours or weeks that a personal computer would require.

The CIPRES Scientific Gateway achieves this speed in two ways. First, it has access to an incredible amount of raw

computing power through a connection to the supercomputers of the National Science Foundation. Secondly, the CIPRES team installs applications (created by others in the field) that run typical phylogenetic analyses in parallel. In other words, rather than taking the data and performing the overall analysis one step at a time, the process is broken up into many different steps, which can be performed alongside each other. This is tricky from a programming point of view but is extremely effective at speeding up the overall process.

'CIPRES allows users to upload their data, and can run as many as 50 analyses concurrently,' says Dr Miller. 'Most analyses also run 10–200 times faster on CIPRES, because we have up-to-date, optimised parallel codes installed on very large computer clusters funded by the US National Science Foundation.'

This essentially makes CIPRES a vital piece of supporting equipment for phylogenetic scientists. By providing access to an assortment of computing resources, the gateway accelerates the pace of research beyond what would otherwise be possible. 'We provide a service that makes it possible to research, much like the waterworks supplies water that makes it possible to live in cities,' explains Dr Miller. 'Most of our users say things would be slower or be halted without the access we provide – what we do is to make research go faster.'

CIPRES has had a noticeable impact on the field of phylogenetics. Access to the gateway is free, allowing researchers from across the world to take advantage of the computing power despite any lack of funding. Indeed, the list of users ranges from NASA through to the Korean National Arboretum. Tens of thousands of researchers have used the gateway to date, with more than 5,500 publications arising through the support of CIPRES in the first nine years of operations. Many ground-breaking studies published in prestigious journals such as *Science* and *Nature* have only been possible due to the rapid analysis of genomic data available through the gateway.

Importantly for scientists, the nuts and bolts of maintaining, tuning, and updating the software are performed by experts at the CIPRES group. This means that scientists can focus on the 'biological thinking' parts of their research, while freeing them from the daily worries of patching software and maintaining computers. 'CIPRES makes it easy for a biologist who is not computer-literate to access and use a highly sophisticated machine to ask important biological questions without months of training,' says Dr Miller.

The Gateway of the Future

So where does Dr Miller and the CIPRES team intend to go from here? They have a number of plans for the future, many of which involve improvements to the user experience. They plan to develop simpler processes for restarting analyses in the middle of a run, as well as complete automation of the upload, input verification and analysis process. They have also just received a grant from the National Science Foundation to implement the rather niftily-named 'cloudburst' into their system. This allows the CIPRES gateway to use cloud computing power when the number of user requests overwhelms even the supercomputer power at their site.

Many other improvements are on the horizon, but Dr Miller is careful not to bring too many ideas onto the field at once. 'We are very judicious with innovation,' he comments, 'we are constantly trying to improve our services – but we have to do so knowing that 10,000 users will be impacted by any changes we make.'

The CIPRES gateway has already shown its value, helping thousands of scientists achieve their work faster than they ever could alone. What will this new wave of improvements bring? Only time will tell – but it is likely to involve further ground-breaking scientific discoveries and ever-more satisfied researchers across the globe.

Meet the researcher

Dr Mark Miller
San Diego Supercomputer Center
La Jolla, CA
USA

Dr Mark Miller received a BS in Biology at Eckerd College in 1972 and PhD in Biochemistry from Purdue University in 1984. He joined the Department of Chemistry at the University of California as a researcher in 1986, where he studied enzyme structure-function relationships. In 2000, he moved across campus to the San Diego Supercomputer Center, where he is currently a Principal Investigator. He is responsible for developing the CIPRES Scientific Gateway, which allows researchers across the world to enjoy high-speed analysis of genomic data.

CONTACT

E: mmiller@sdsc.edu
W: https://www.sdsc.edu/~mmiller/

KEY COLLABORATORS

Wayne Pfeiffer, Distinguished Scientist, San Diego Supercomputer Center

FUNDING

US National Science Foundation
National Institutes of Health

FURTHER READING

MA Miller, T Schwartz, BE Pickett, S He, EB Klem, RH Scheuermann, M Passarotti, S Kaufman, and MA O'Leary, A RESTful API for Access to Phylogenetic Tools via the CIPRES Science Gateway, Evolutionary Bioinformatics, 2015, 11, 43–48.

MA Miller, W Pfeiffer, and T Schwartz, Creating the CIPRES Science Gateway for inference of large phylogenetic trees, in Proceedings of the Gateway Computing Environments Workshop (GCE), 14 Nov. 2010, New Orleans, pp 1–8.

SDSC SAN DIEGO SUPERCOMPUTER CENTER