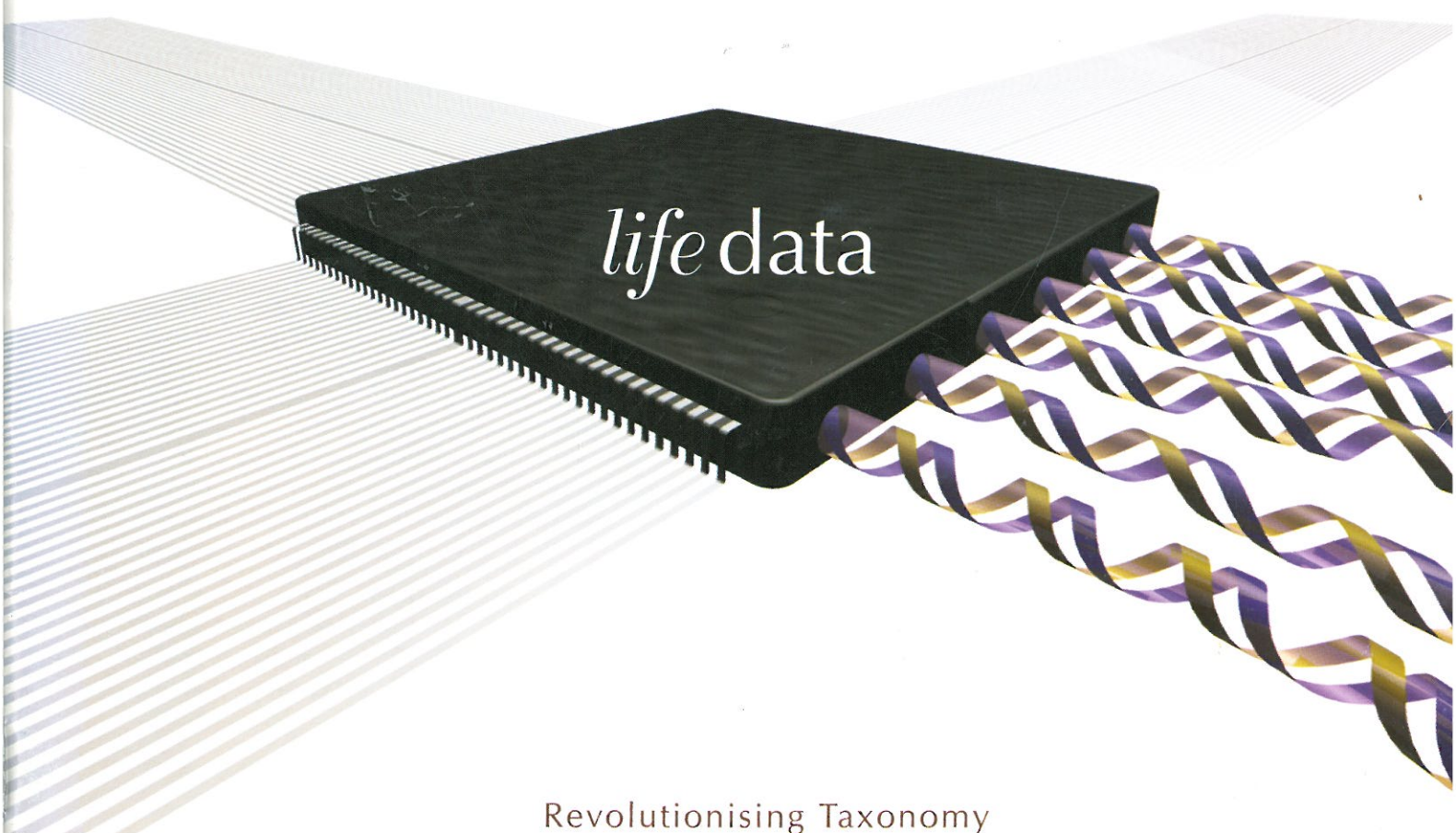
BIOLOGIC

No.12
August 1997
ISSN 1320-6028

Research School of Biological Sciences
Institute of Advanced Studies
Australian National University



Revolutionising Taxonomy

Interacting with Sequences

VIDE – Virus Databases

Unfolding Folding Protiens

Contents . . .

The Research School of Biological Sciences is one of Australia's leading centres for basic biological research and graduate training. Since its inception in 1967, it has focussed on three domains: plant science, genetics, and neuroscience. This work is carried out in 10 research groups, organised as follows:

Visual Sciences ~ Group leader: Dr Ian Morgan
Plant Cell Biology ~ Group leader: Dr Peter John
Molecular Genetics and Evolution ~ Group leader: Dr Des Clark-Walker
Plant Microbe Interactions ~ Group leader: Prof Barry Rolfe
Bioinformatics ~ Group leader: Dr Georg Weiller
Ecosystem Dynamics ~ Group leader: Prof Ian Noble
Molecular Plant Physiology ~ Group leader: Prof John Andrews
Environmental Biology ~ Group leader: Prof Graham Farquhar
Photobioenergetics ~ Group leader: Prof Barry Osmond
Developmental Neurobiology ~ Group leader: Prof Richard Mark

If you would like to know more about any of the research activities, please contact the above Group Leaders at:
Research School of Biological Sciences, Institute of Advanced Studies,
Australian National University, GPO Box 475, Canberra, ACT 2601.
or the Director's Secretary phone: (02) 6249 3841, or fax: (02) 6249 0758

Biologic is published by the Research School of Biological Sciences, Institute of Advanced Studies, Australian National University. Written and edited by Sarah Vandermark and Barry Osmond. Artwork/design by James Whitehead and Tara Goodsell. Photographs by Jeff Wilson. Printed by Union Offset Printing, Fyshwick, A.C.T.

Editorial . . .

– *a few words about Bioinformatics*

Bioinformatics is the theme encompassed by this twelfth edition of *Biologic*. The OECD Megascience Forum Working Group on Biological Informatics said Bioinformatics, 'encompasses research in, and development of, tools and approaches for electronic acquisition, dissemination, storage, querying, retrieval, visualisation, integration, analysis, synthesis and sharing (which includes electronic means of collaboration) of biological data.' Prof Adrian Gibbs of RSBS simply regards Bioinformatics as the transformation of biological data into biological information.

No matter what the perspective taken upon the precise definition of bioinformatics, one thing is for certain – this field of research is HOT! Bioinformatics is a product of the information age – the coupling of computer power and biological data. Growth of public databases which house DNA sequence data has driven the need for programs which can interrogate and analyse databases, as well as for the designers of such programs, programmers with an appropriate understanding of molecular biology. An article in the journal *Science* reported that scientists with a combined degree in molecular genetics and computer science, were being head hunted by some of the biggest drug firms in America and being offered starting salaries of $US 65,000 (*Science*, vol 272, 21 June 1996, p 1730-1732). This demand stems from the belief that databanks will become 'the source of most, if not all, new drug targets,' by early next century (ibid).

The Australian National University has embraced the advent of bioinformatics and the accompanying challenge to keep up with the rapid pace of development in the area, by establishing a Bioformatics Laboratory in the 'Centre for Molecular Structure and Function' or CMSF. This Centre represents an, agreement between scientists in the Research School of Biological Sciences, the Research School of Chemistry (RSC), the John Curtin School of Medical Research and the Faculty of Science, to share equipment and knowledge. This collaborative effort is reflected in the contribution of articles to this edition, with stories not only from RSBS, but also from the RSC.

This edition begins with 'Revolutionising Taxonomy', a story which exemplifies the utility of combining biological data with computers and the extraordinary degree of foresight demonstrated by the two scientists who decided to take taxonomy out of the biological backwaters and into the information age. Dr Georg Weiller (RSBS) is one of a handful of scientists in Australia who are designing computer programs to analyse DNA databases. The second story, 'Interacting with sequences', looks at Dr Weiller's unique approach of using graphics in order to 'see' the relationship between different sequences. The largest biological database on the Internet, is all about viruses and the development of this database is described in 'VIDE (Latin: to see).' One of the most intractable biological problems is predicting a protein's shape from its sequence and thereby, elucidating the relationship between its structure and its function. The fourth article, 'Unfolding Folding Proteins', investigates how a particularly creative chemist and computer programmer, Dr Andrew Torda, has approached this problem.

Our cover symbolises the significance of bioinformatics in modern biology. Australia's first network server in bioinformatics, ***life.anu.edu.au***, which was established in RSBS in the early 1990s by David Green (now at Charles Sturt University) still supports much of the School's database research. The articles in this issue of Biologic, as well as links to related databases, can be found at the RSBS web site: ***biology.anu.edu.au/internal/publications/biologic/home.html.***

I hope that you find bioinfomatics to be as interesting and challenging as I did!

Sarah Vandermark

REVOLUTIONISING taxonomy

Traditional Identification

20* Spikelets <9 mm long, or if longer then the lemma either 2-cleft and awned or mucronate from the sinus or entire and awned from near the apex with an awn twice or more the length of the lemma and the lemma linear.

21 Glumes rigidly ciliate on the keel; lemma entire or 2-cleft with the lobes sometimes aristate, and awned or mucronate from the sinus, 5-11-nerved; panicle spicate, sometimes dense ..... *Echinopogon*

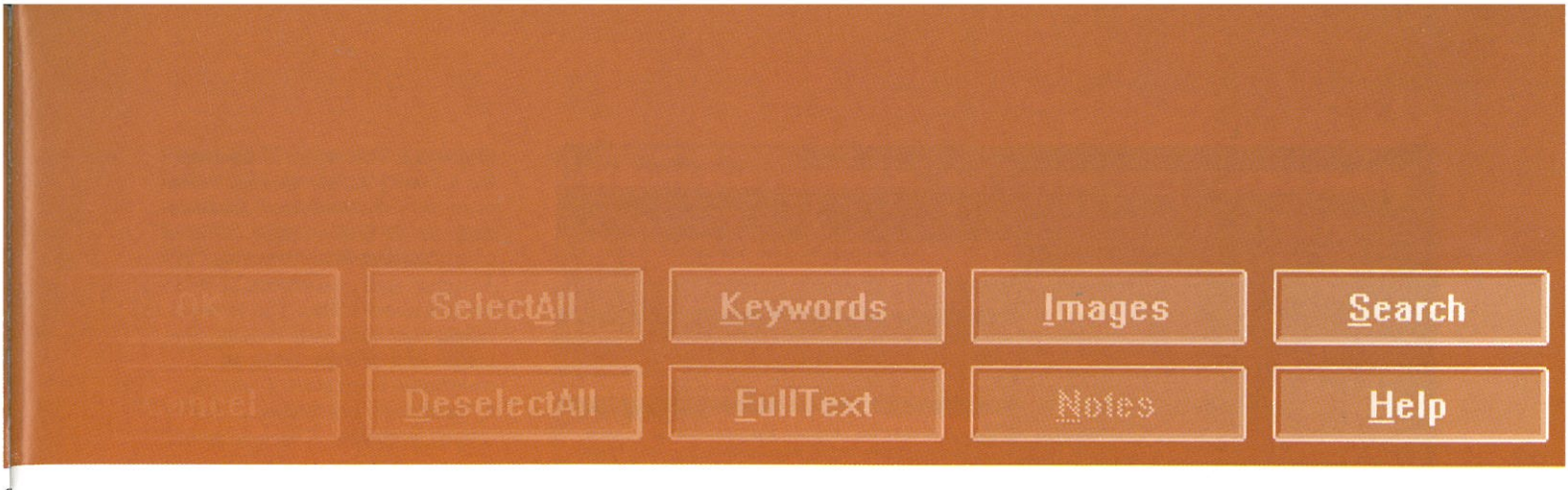
21* Glumes smooth or scabrous on the keel but not ciliate; lemma entire or minutely (often 4-) toothed.

Figure 1. In identification keys, descriptions are presented at each step of the identification. Users have to decide between two alternatives at each step. A specimen of *Echinopogon* (Hedgehog grass) can be identified by a number of features of the grass flower *(left)* such as the long, sharp appendage (awn) and the stiff hairs on the keel of the outer bract (e.g. 21, 21*).

Have you ever felt frustrated when using a traditional 'classification key' to identify an exciting find? For example, with specimen in hand, you eagerly begin the process of working, step by step, through set questions, which must be answered either 'yes' or 'no', in order to move closer to a positive identification. Question 1 may ask, 'Is the sample an animal or plant?' It is a plant and you are directed to; 'Go to question X,' which asks, 'Is it a flowering plant?' Well, this particular specimen does not have a flower and therefore you are unsure of its status (see Figure 1). Immediately, your search is thwarted or at least postponed until you are able to obtain a complete specimen!

Imagine the luxury of an interactive identification tool where you 'tell' a computer about your specimen and in return, it 'tells' you what you have! How about an identification tool which goes beyond simple 'yes' or 'no' answers and can distinguish between answers such as; 'unknown', 'variable' and 'not applicable'. A specimen does not even need to be complete. The powerful computerised 'DELTA system' is a dream come true for taxonomists, it can do all this and more!

For the past thirty years computer tools for aiding taxonomy have been designed that attempt to revitalise the tedious taxonomic process. By far the most successful is DELTA, DEscription Language for TAxonomy, developed by Dr Mike Dallwitz (CSIRO) and Dr Les Watson (Research School of Biological Sciences). DELTA is a set of computer programs designed to store and manipulate a database of taxonomic information. Comparatively, DELTA is the Rolls Royce of databank software and a silver service provider of programs facilitating easy access and utilisation of stored information.



The flow to DELTA

The reason why the DELTA format is so successful is due to the unique collaboration between Drs Dallwitz and Watson. Dr Dallwitz, a computer programmer, specialised and developed the software in conjunction with Dr Watson's taxonomic expertise. In this way, there was always a very close association between program development and the needs of its potential users.

Dr Dallwitz studied at the Australian National University, completing two honours degrees, one in pure mathematics and the other in physics, prior to writing his PhD. After a temporary lectureship in the Physics Department, he moved to the CSIRO's Division of Entomology, where he is now a Senior Principal Research Scientist.

Some say that Dr Watson's interest in biology was kindled by the fact that the boys' school he attended in Staffordshire, England, did not teach biology, forcing him to make two trips per week to the local girls' school! He claims that his collection of local butterflies kindled a permanent fascination for natural selection. Either way, Dr Watson chose to study botany and never lost his enthusiasm for taxonomy. He was also an early explorer of the potential applications for computers in this field.

The Research School of Biological sciences was still embryonic when Dr Watson arrived in 1969, joining the Taxonomy Unit. In 1972, Drs Dallwitz and Watson discovered their mutual enthusiasm for the challenge of automating taxonomy with computer technology and initiated their collaborative work, which over the next quarter of a century, became one of the most important collaborations in modern taxonomy.

'The Watson-Dallwitz collaboration has transformed taxonomy from a dry, name-generating backwater of biology into a modern branch of information science, standing alongside, and complementary to molecular biology, in its ability to provide information and generate novel ideas and questions,' says Prof Adrian Gibbs (RSBS).

DELTA – the product

The beauty of DELTA is the flexibility it provides to users wanting to record taxonomic information as well as to those wishing to utilise a database. The DELTA format caters for all types of taxonomic information such as; counts, measurements and text, with provision for comments and qualifications.

The programs in the DELTA system allow the recorded data to be used in a variety of ways, from producing descriptions and identification tools, to exporting the data in formats required by other programs. Such programs may, for example, analyse the evolutionary relationships amongst the organisms. The DELTA system may also be used as a state of the art desk-top publishing program, producing descriptions of the organisms stored in the database in several different forms including a fully formatted text for hard copy publication, or in electronic form such as HTML for the Internet.

When DELTA is used as an identification tool, it turns the traditional classification process on its head. The program INTKEY, for interactive identification and information retrieval, is a crucial program within the DELTA system, designed to give the investigator a high degree of freedom and flexibility when querying a database. To identify a specimen the investigator interrogates the database, selecting questions which best describe the specimen at hand. These chosen 'characters' can be placed in any order desired, so the user controls the identification process unlike a traditional taxonomic key. It is also possible to extract data, as opposed to identifying an individual. For example, it is easy to retrieve lists of species sharing certain characteristics or inhabiting a certain geographical area. Yet, there is another degree of freedom in the DELTA system, the software and database may be stored on a lap top computer, making it conveniently transportable for use in the field!

It is no wonder that the DELTA system is in widespread use throughout the world and has been applied to all types of organisms including; viruses

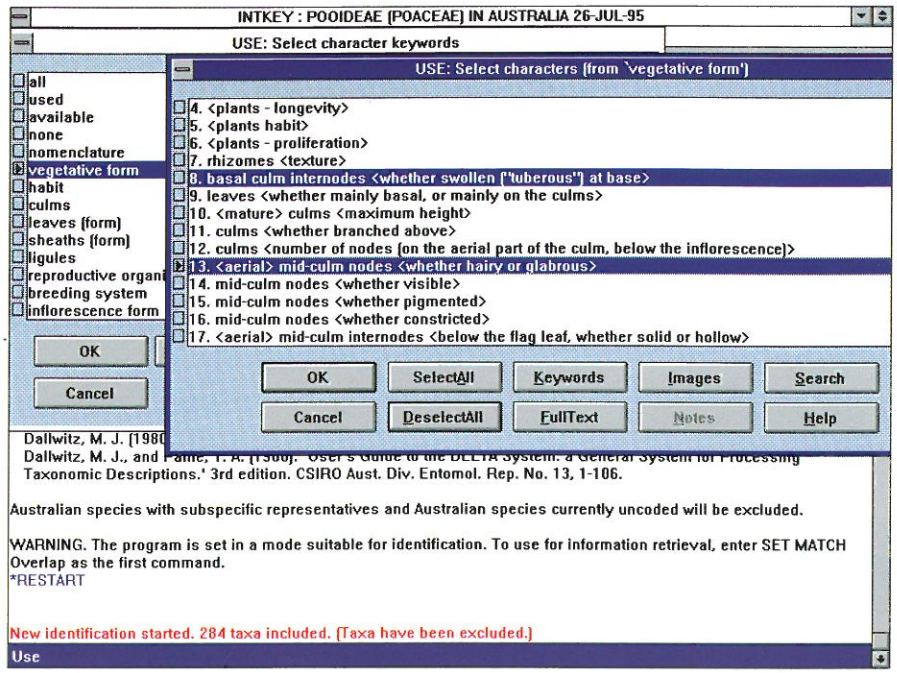


Figure 2a. Four steps *(shown in the following screen-dumps)* in the computer identification of 'Bulbous oatgrass' or 'Onion twitch', a native of Europe and Asia, introduced in Australia, and now a weed in the south-eastern states. The keyword "vegetative form" was chosen. This produced a window listing all associated characters. Characters 8 "basal culm internodes..." and 13 "mid-culm nodes..." were selected.

(see *VIDE*, page 13), algae, fungi, arthropods, vascular plants and vertebrates. In 1994, Drs Dallwitz and Watson released an interactive CD-ROM, constructed using DELTA, called *The Families of Flowering Plants*. The review of this CD in the scientific journal *Nature*, comments upon its interactive format and states, '. . . it is not too grandiose to think that *The Families of Flowering Plants* represents a stage in plant taxonomy as important as the publication of Linnaeus's *Genera Plantarum* in 1737.'

Grasses and DELTA

During the development of DELTA, extensive trials were carried out upon Dr Watson's data on grass genera. Although now retired in Western Australia, Dr Watson retains a Visiting Fellowship with RSBS and is actively working on maintaining his databases and assisting the Western Australian Herbarium construct a DELTA dataset of the genera of Western Australia. At RSBS, Dr Carolyn Weiller is extending Dr Watson's work on grass genera to the species level. The aim of this enormous task is to provide

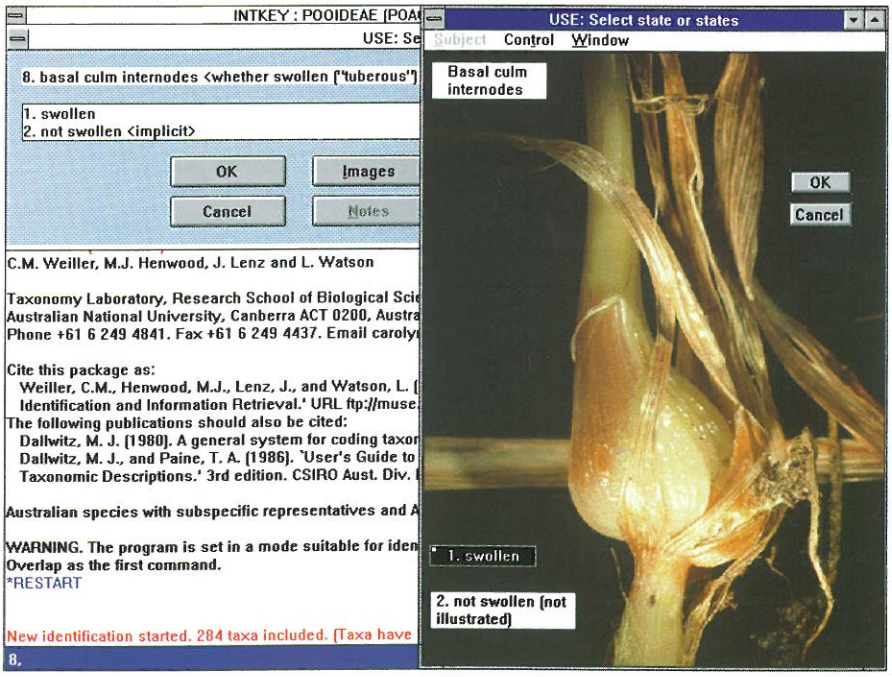


Figure 2b. The two options of character 8 are presented in a window. This window also displays a series of buttons allowing, for example, 'Images' and 'Notes' associated with a character to be displayed. Characters may be directly selected from the image. Here feature "1. swollen" has been selected.

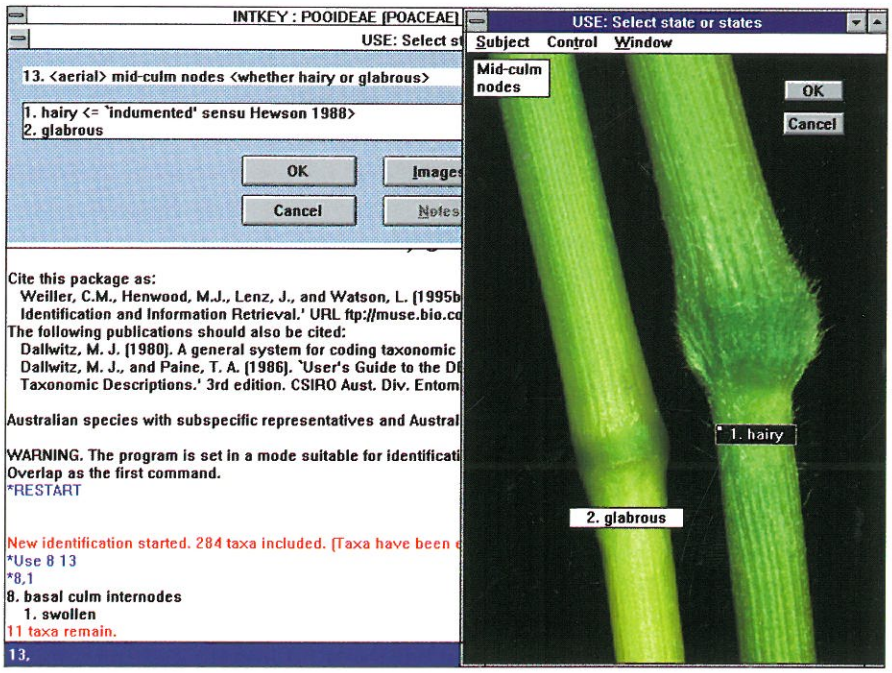


Figure 2c. The previous step reduced the number of taxa remaining to 11 from 284 (see lower left screen). Now character #13 is presented and feature "1. hairy" is selected.

species level databases for the approximately 1375 native and introduced grasses occurring in Australia. For the majority who believe that when it comes to grasses, 'once you've seen one you've seen them all', be assured that there is much variability within the grass family (see Figure 2).

Dr Weiller's research is sponsored by the Australian Biological Resources Study and will contribute to two volumes of the *Flora of Australia*. Apart from providing data on grasses, she coordinates the entry and updating of information for over 1000 of the Australian taxa. This is a delicate balancing act, as Dr Weiller explains, 'the character list is very extensive and provides much more information than is required, or in fact desired, by editors of a hard copy product such as the *Flora*.'

Currently data for three quarters of the Australian grass species are contained in DELTA databases. 'Examining your data with INTKEY can give you valuable insights into your data and can quickly show such things as the degree of similarity between taxa,' she says. The databases also contain

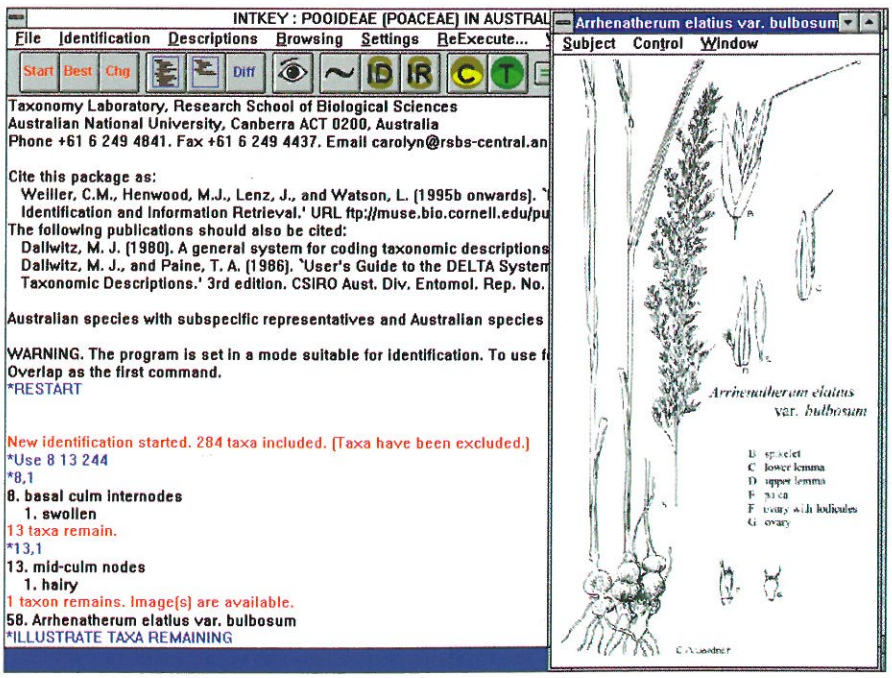


Figure 2d. The number of taxa has now been reduced to one and the specimen has been tentatively identified as "*Arrhenatherum elatius* var. *bulbosum*". An image of the taxon is displayed. The identification may be verified by viewing a complete description of the species.

descriptions of introduced grasses, many of which have been previously described. Dr Weiller has found, however, that these descriptions do not necessarily reflect how these introduced species appear in the Australian environment!

The grass datasets are available as INTKEY packages and contain detailed morphological data as well as details of geographical distribution, habitat, nomenclature and references. Also included are hundreds of illustrations, available at the click of a mouse button, depicting microscopic features of the plants and whole-plant line drawings. An on-line glossary is also available providing explanations and definitions of terms and characters.

Where the river meets the sea

The development of the DELTA system required an intriguing amount of foresight and brilliance on behalf of Drs Dallwitz and Watson. This is especially true when it is considered that they initiated their research long before the majority of people believed that computers would become everyday tools. The story of the development of DELTA also provides a clear example of the need for the support of basic research. For many years, Drs Dallwitz and Watson's research may have seemed obscure and not necessarily 'applied', when in the long term, DELTA catalysed a paradigm shift in taxonomy and is currently being utilised throughout the world.

Visit the DELTA web site and experience revolutionary taxonomy: *http://www.keil.ukans.edu/delta/*

If you want to know more ...

- The Grasses of Australia: Subfamily Pooideae. Second Edition (1995) (A taxonomic database in DELTA format distributed with the INTKEY program.)
  C.M. Weiller, M.J. Henwood, J. Lenz and L. Watson.
  http://www.keil.ukans.edu/delta/www/intkeyp.html

- Pooideae of Australia: Descriptions and illustrations.
  C.M. Weiller, M.J. Henwood, J. Lenz and L. Watson (1995 onwards).
  http://www.keil.ukans.edu/delta/www/descrip.html

- You are welcome to contact Dr Weiller on (02) 6249 4841.

Dr Carolyn Weiller

Profile

Adrian Gibbs

Now a keen balloonist, Professor Adrian Gibbs learnt an important lesson from his initial encounter with a computer. While working for his PhD in plant virology, at Rothamsted Experimental Station, England, he was privy to the use of one of the first computers. The stature of this computer was perhaps more impressive than its capacity. It occupied an entire room and had roughly a kilobase of memory! When Gibbs received the computerised analysis of his data, it did not provide the answer he was anticipating. Upon expressing his doubts to the computer programmer concerning the result, he was told that his expectations must be incorrect, as "it's been through the *electronic* computer". The young Gibbs, insisted that his data be reanalysed. It then became apparent that there had been an error - only one tenth of the data had got into the computer - and the first result was, indeed, incorrect. 'While it is true that computers never lie, this does not mean that one stops keeping an eye on them' says Prof Gibbs.

Since his early interaction with computers Prof Gibbs has developed an interest in analysing data. In particular, he is interested in non-mathematical analysis, as he claims to suffer from innumeracy and prefers graphical methods of presenting data. In this manner, relationships within a data set are visualised as a graph as opposed to a set of indigestible numbers. Prof Gibbs believes that, he, like most biologists, uses statistics, 'as a drunk uses a lamp post - for support rather than illumination!' With this belief at heart, Prof Gibb's career could be summarised as a personal crusade in information science, where biological data is turned into biological information, with the use of computers.

One of Prof Gibbs' early successes was a very simple and more informative method to assess the relatedness between two sequences of nucleotides or amino acids. The so called 'dot-plot', was devised in 1971, in collaboration with the late Dr George McIntyre (CSIRO Division of Mathematical Statistics). A matrix is constructed by placing the two sequences to be aligned along adjacent edges of a rectangle (see Figure 1). With one sequence running along the top of the matrix and the other down the side, nucleotides on the horizontal axis are scored against all the nucleotides on the vertical axis, so that when a row and a column have matching letters a dot is marked. Sequence similarities are immediately obvious as diagonal runs of dots (see Figure 1). In this way, amino acid or nucleotide sequences may be compared in dot-plots.

The ability to assess the relatedness of sequences visually without statistical analysis was quite a radical development. As dot-plots were such a successful tool for DNA analysis, they formed the basis from which other methods were derived (see *Interacting with sequences*, page 8).

Prof Gibbs met Drs Mike Dallwitz and Les Watson in 1971, the time when they were initiating their work on computerising taxonomic information (see *Revolutionising taxonomy*, page 2). He realised that their system, called DELTA, provided an ideal method for storing the information he and other virologists were collecting (see Figure 2). Prof Gibbs' foresight initiated fifteen years of collaboration among 200 plant virologists, who set up the VIDE database on plant viruses, produced numerous publications and, most recently, one of the world's biggest hypertext databases, 'Plant Viruses Online' (see *VIDE*, page 13).

Figure 2. An early photo of the players involved in RSBS-CSIRO collaboration during development and establishment of major DELTA databases. Program developer, Mike Dallwitz *(left)* and plant systematic botanist, Les Watson *(front-centre)* and virologist, Adrian Gibbs *(back)*, with colleagues Cornelia Büchen-Osmond *(front-left)* and Karen Crabtree *(front-right)*.

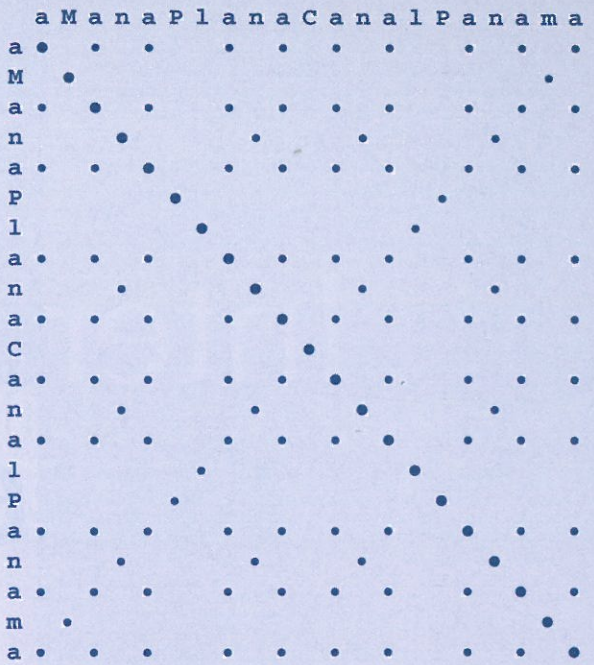


Figure 1. A dot diagram, in its simplest form, compares two sequences by placing them on two sides of a matrix and putting dots at intersections of rows and columns that contain the same letter. The dot diagram above shows how the idea works by comparing the phrase " A man, a plan, a canal, Panama" with itself, after omitting gaps and punctuation. The main diagonal run of dots (top left to bottom right) is intact because the sequence is being compared to itself, however there are short runs on other diagonals resulting from the repetitions of the sequence "ana", and the reverse main diagonal shows that the sequence is palindromic (i.e. reads the same in both directions).

Most recently Prof Gibbs has helped establish the Centre for Molecular Structure and Function (CMSF) at the Australian National University. The Centre is unusual as it is not a Centre built of concrete, it is an agreement to promote the cross fertilisation of ideas and sharing of resources between three research schools, JCSMR, RSBS and RSC and the Science Faculty of the ANU. The advance of both biotechnology and computing ushered in a new era of research, one where scientists require sophisticated equipment for analysing genes and gene products, as well as computer programs capable of analysing vast lengths of sequence data. The CMSF plays a vital role by providing the links to the expensive infrastructure scientists now require to remain active in international research.

Interacting with Sequences

'DNA sequences and computers were made for each other.' Dr Georg Weiller, Bioinformatics Group Leader in RSBS, illustrates his point with the example of the human genome project. Four million nucleotides from the human genome have been sequenced and require only 1 megabyte of computer memory to be stored. Furthermore, in its entirety, the human genome sequence is expected to fill a meagre 100 megabytes of memory. That's equivalent to about five scanned colour images, which by comparison hold little information!

Dr Weiller never dreamed that his interests in molecular biology and computer programming would become a practical combination. Having trained as a molecular biologist, it was the misadventure of writing his Masters thesis on a typewriter, that spawned his enthusiasm for computers and computer programming. Scientists today are able to sequence vast lengths of DNA and then store this data in so called 'databanks'. The process of examining databanks has become crucial. Dr Weiller is at the forefront of DNA analysis and is expert in designing programs to 'get the information out' of these databases. He further explains that, 'the field of DNA analysis has taken off and it is primarily due to the simplicity of the DNA alphabet, which is comprised of four discrete nucleotides; adenosine, cytosine, guanosine and thymidine. It is this simplicity which makes sequence data so amenable to computer analysis.'

Seeing is believing

Comprehending visual information, particularly patterns, is a very well developed human skill, exemplified by the abundance of traffic signs and advertising billboards cluttering our modern environment. Most of us can rapidly comprehend a simple graph – identify trends, intuitively fill in missing data points and notice small variations amongst the 'big picture'. It is impossible to accomplish an equivalent degree of interpretation when the same data is presented as long columns of data.

When designing computer programs for DNA analysis, Dr Weiller recognised the considerable advantage of combining the processing power of computers with the superior ability humans have for interpreting graphical information. Accordingly, he is developing interactive computer graphics as an interface for analysing and examining DNA sequences. 'I am specialising in computer programs that compare different sequences and present their relationships in the form of interactive graphs,' explains Dr Weiller. 'Instead of printing a table of statistics, the graph remains on screen, maintaining the association between the graph and the original sequence data. This allows the investigator to modify directly aspects of the graph and indirectly specify the most suitable parameters for their investigation. In this way, they can really explore their data,' he says.

A number of programs that produce different types of 'plots' have been designed by Dr Weiller and his colleagues to enable scientists to more easily 'explore their data'. Two of them are presented below.

Distance-Plots

One of the most common questions asked about sequences concerns their similarity or relatedness. There are complicated algorithms which, when accompanied by a variety of statistical tests for significance, measure this. Most scientists, however, prefer to view the sequences they are comparing directly and utilise their experience to judge their relatedness.

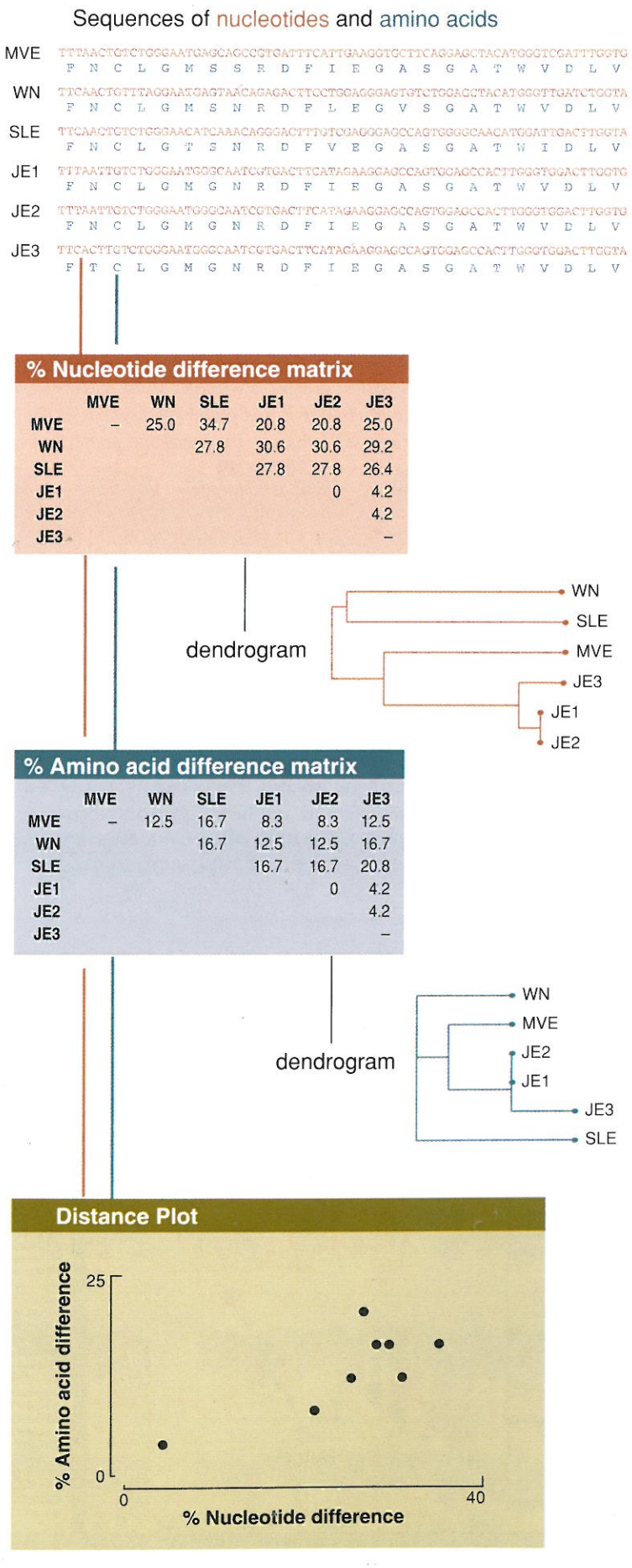
The relationships between organisms have traditionally been represented as dendrograms or taxonomic trees. Dr Weiller points out that, 'a severe limitation of dendrograms is that they can only represent one particular clustering pattern, when often many alternative clusterings are of similar importance.' This is a more pressing problem today, as it is now possible to highlight previously undetected differences between taxa, at the gene sequence level.

Dr Weiller has overcome the limitations of dendrograms by combining two ingredients, 'distance matrices' and a simple scatter plot, to produce the distance-plot. What makes this graphical DNA analysis so versatile is that taxa may be compared, using various distance measures derived from different genes or different parts of the same gene, indeed any of their features that can be expressed as distances. For example, two 'distance matrices' may be calculated, one representing the percentage difference between the respective taxa's nucleotide sequences, and the other their amino acid sequences. The degree of correlation between two matrices is illustrated in a resulting distance-plot and the investigator can 'see' the rates of change of relatedness between nucleotides and amino acids (see Figure 1). Depending upon the data compared, the plot may provide insight into selection pressures and evolutionary processes affecting the respective taxa.

DIPLOMO is the acronym for DIstance PLOt MOnitor, the computer program developed by Dr Weiller, which not only constructs the scatter plots but assists the investigator's exploration of their data. Groups of data points, displayed in a plot, may be labelled using symbols or colours and these symbols are retained when the matrices are changed. 'This makes it easy to follow the properties of taxa when they are compared using a variety of distance measures,' says Dr Weiller.

To give an example of DIPLOMO's potential, Dr Weiller decided to examine the selection pressure upon the envelope protein gene of the lentiviruses, the virus group which includes the two types of human immunodeficiency viruses. It has long been known that the human immunodeficiency viruses (HIV types 1 and 2) have evolved from monkey or simian immunodeficiency viruses (SIV). Two famous scientists, Gould and Eldridge, developed the idea

Figure 1. Steps in the generation of a distance plot.



that evolution proceeds by a succession of 'punctuated equilibria'. This implies that if an organism remains in a particular environment over a long period of evolutionary time, it will adapt to the point where selection for further evolutionary change diminishes. However, if that organism's environment changes (for example, when a virus enters a new host), its genes will experience different selection pressures and may well undergo rapid change.

With this theory in mind Dr Weiller chose to examine how strongly amino acid changes had been selected in the envelope proteins of HIV and its known relatives. He decided to compare the pairwise differences in amino acid sequences with the pairwise differences in nucleotide sequences, between the two viruses. He expected that a distance-plot of these differences would illustrate whether there had been a greater rate of change in the envelope protein of the recently arisen HIV than in the envelope protein of its wild relatives (see Figure 2).

When viewing the resulting distance-plot it is immediately obvious that the ratio of amino acid to nucleotide changes is not the same in all pairwise comparisons. Different populations or isolates of HIV and SIV are included in the plot, providing a surprising degree of insight into the evolution of these viruses. The position of HIV-1, (red 1) on the extreme left, indicates that its envelope protein has acquired more amino acid changes for each nucleotide change than HIV-2 (purple 2) and all of the SIV (purple +) isolates. This suggests that HIV-2 is slightly more stable and hence, probably better adapted to its human host than HIV-1. It is interesting to note that some of the SIV isolates (SIVmac (blue M)) are in the same part of the scatter plot as HIV-1, suggesting that they have also recently invaded a new host. Thus, it is not surprising to learn that these SIV isolates are restricted to primate research center's where macaques are housed together with and crossinfected by sooty mangabey monkeys. Like HIV1, but unlike the other SIV, the isolates are highly pathogenic, killing their new host within weeks of infection.

These results clearly demonstrate that distance-plots are indeed a sensitive, effective and simple tool for detecting changes in evolutionary pressures.

Phylogenetic profiles

Most recently, Dr Weiller has designed another ingenious program called 'PhylPro', short for 'phylogenetic profile'. He explains, 'the analysis of relatedness using, for example, tree diagrams or distance plots, is fundamentally flawed if the representative sequences being compared have inconsistent histories. What if one part of a gene comes from one source and the remainder from another?'

Various mechanisms speed up evolution. Among the best known are the different forms of genetic recombination, which allow the combination of independently evolved genes. This seems to be the main reason why most higher organisms have developed sexual reproduction, as this process allows genetic material from different parents to recombine. The discovery of the many different types of 'jumping genes' or mobile genetic elements has shown that pieces of genes from diverse sources can be combined to improve existing or create new genes

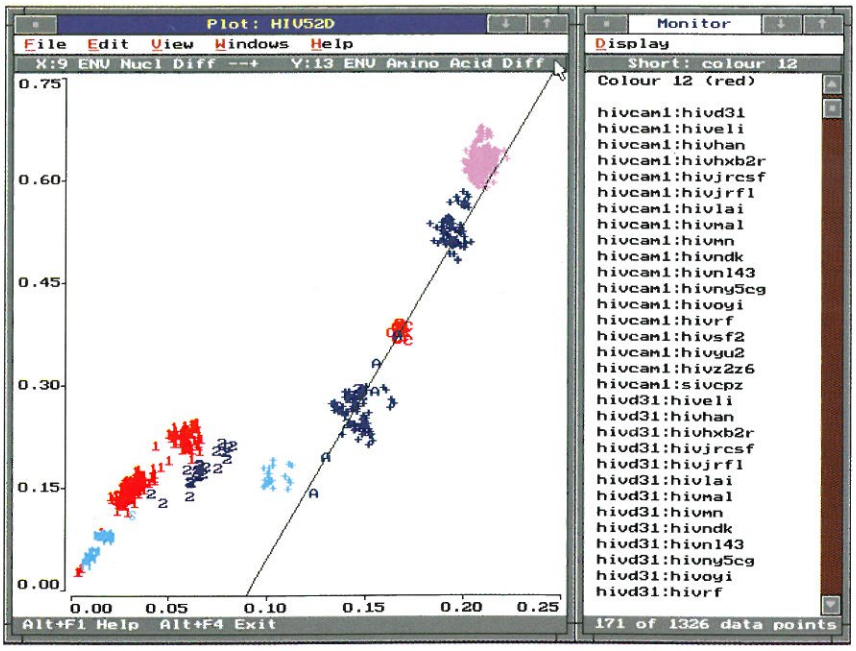


Figure 2. DIPLOMO screen. The plot window displays a distance-plot of the *env* gene of 40 HIV/SIV viruses. The X and Y axes represent the proportion of 3rd codon position and amino acid differences respectively. A linear regression line is fitted to comparisons between SIVagm isolates (purple A). The monitor window displays the taxon names of all comparisons represented in red.

(see *Jumping Genes*, in *Biologic* volume 11, for more on transposons). Some of these mobile genetic elements, most notably viruses, carry genetic material across the barriers between species. Thus, genes are not only transferred 'vertically' from parents to offspring but also 'horizontally' and even between different species. It is of no surprise, therefore, that various parts of some genes may have different evolutionary histories.

Dr Weiller has developed the new algorithm to detect recombination. "Given a set of related sequences, the 'phylogenetic correlation' measures for every sequence, whether its relationship to all other sequences in the various segments correlate in different parts of the gene." Dr Weiller explains. When the sequences in the various segments have the same ancestry, then their relationships correlate well. A poor phylogenetic correlation by contrast indicates that segments of different origins have recombined. At any given position of a multiple sequence alignment, the segments immediately to the left and right are compared. By plotting this phylogenetic correlation for every position of a sequence, one obtains the phylogenetic profile of that sequence. A typical phylogenetic profile plot superimposes the profiles of all sequences, and therefore looks like the output of a seismograph (see Figure 3).

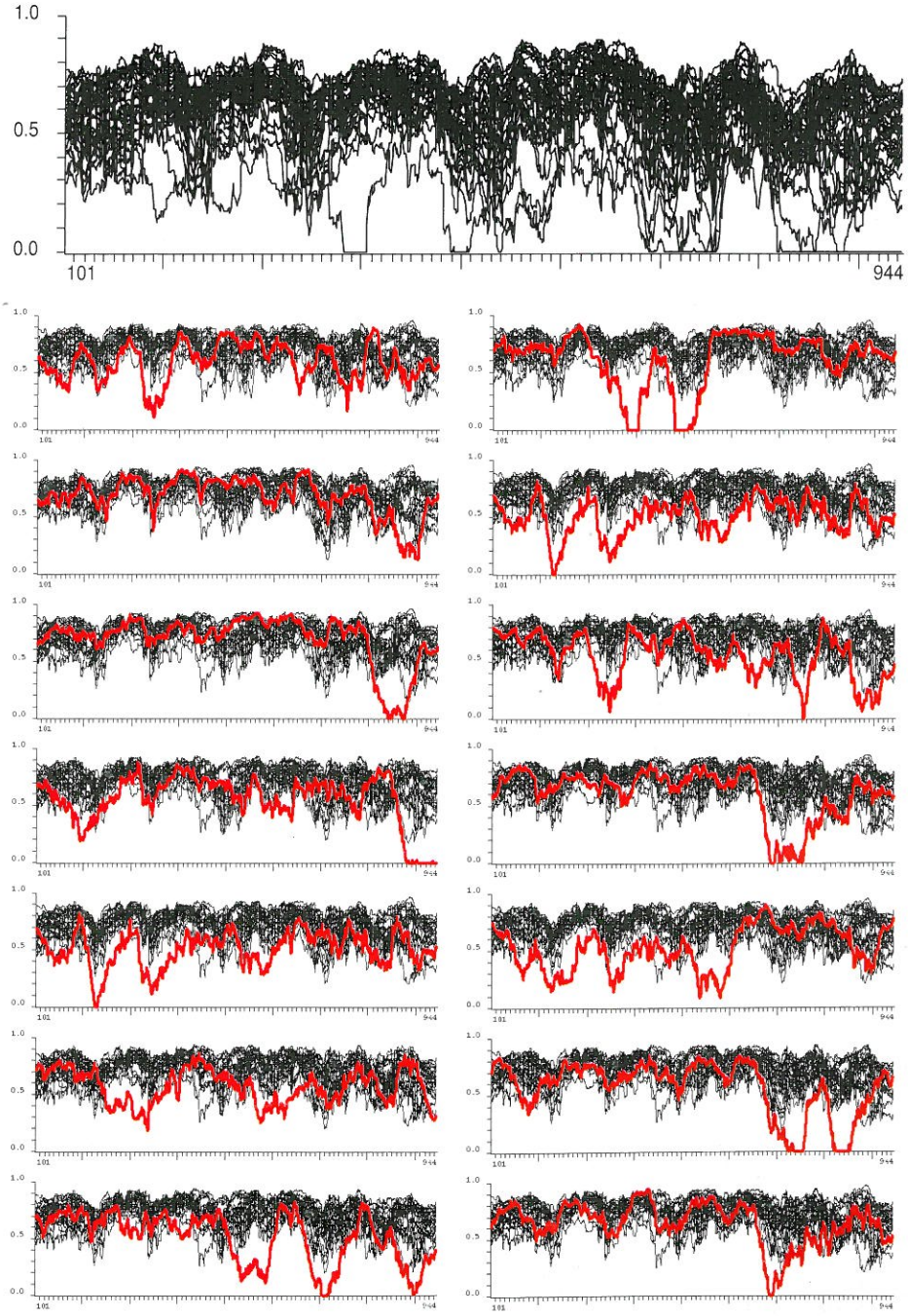


Figure 3. Phylogenetic profile scans of the *gag* locus in 84 HIV1 isolates.
The top graph shows the phylogenetic profiles of 84 different HIV sequences. The graph shows that many of the sequences have sites of little phylogenetic correlation (low values on the graph), which indicate the sites of genetic recombination. The sequences that gave particularly clear recombination signals were subsequently removed from the data-set. The graphs below give the phylogenetic profiles of this reduced data set, containing only the more stable HIV isolates plus one of the clear recombinants shown in red. The analysis shows that HIV viruses recombine frequently. This implies that different types of HIV viruses meet in the same cell, and consequently, that multiple independent infections have occured.
Hence it appears that infection with one HIV virus does not lead to the production of antibodies that protect the individual from further infections.

Each downward pointing spike marks a possible recombination site; the junction between two parts of the sequence that are not directly related.

Dr Weiller tested his new PhylPro program on a number of sequence sets known to contain recombination events. In every case, PhylPro detected the known sites of recombination as well as novel recombination events. 'This clearly demonstrates the enormous sensitivity of the PhylPro method,' says Dr Weiller.

Many vaccine development programs base their strategies upon the belief that the presence of HIV antibodies will protect against further infection from HIV. This implies that only the descendents of one virus strain should be found in any given individual.

Thus HIV is an obvious candidate for testing the value of PhylPro. Dr Weiller and assistant Holger Averdunk developed phylogenetic profiles for 84 different HIV-1 sequences, which encoded the 'group specific antigen' or *gag* gene (see Figure 3). 'We found clear signs of multiple recombination events, with 30 occurring recently,' says Dr Weiller. Various statistical approaches had previously only detected around five recombinant *gag* genes! 'What is most exciting is our finding that many viruses sampled contain genetic material from three or more different HIV-1 strains,' Dr Weiller exclaims.

This result questions the assumption, outlined earlier, that an initial infection with HIV guarantees immunity against other strains of the virus. How can HIV antibodies be seen to provide any 'protection', in the light of evidence showing the presence of three different strains of virus in one host? Dr Weiller cautions, 'this may mean that many vaccination programs are likely to fail.'

Dr Weiller's findings do not, however, prove that antibodies against HIV, i.e. a vaccination, do not protect against the HIV virus. Dr Weiller says, 'developing immunity takes time.' During the first several weeks post infection or post vaccination, antibody levels remain low. This period may be a window of opportunity, when different strains of the virus may infect the same host. "The question is, whether this window is big enough to account for the frequent recombinations of different HIV strains that we have discovered" says Dr Weiller. What does appear certain, is that Phylpro has confirmed that it will be very difficult to design an effective HIV vaccination.

Future for DNA analysis

Recently returned from an international think-tank of bioinformatics experts, Dr Weiller believes that these types of graphical tools for DNA analysis represent only the early stages of a new field of research. In his opinion, 'the current rush to complete sequence databases obscures the excitement and potential of this field. After all, databases are simply a collection of files. Later, when questions are asked it becomes very interesting.' He also believes that, 'sequences are so amenable to computer analysis, it should be possible to deliver answers to many questions.'

If you want to know more ...

• Diplomo: the tool for a new type of evolutionary analysis. Georg F. Weiller and Adrian Gibbs (1995). CABIOS, 11, 535-540

• Phylogenetic Profiles: A graphical method for detecting genetic recombination in homologous sequences. G. Weiller. Mol. Bio. Evol (in press).

Dr Georg Weiller

• You are welcome to contact Dr Weiller on (02) 6249 5916, or via email: weiller@rsbs.anu.edu.au

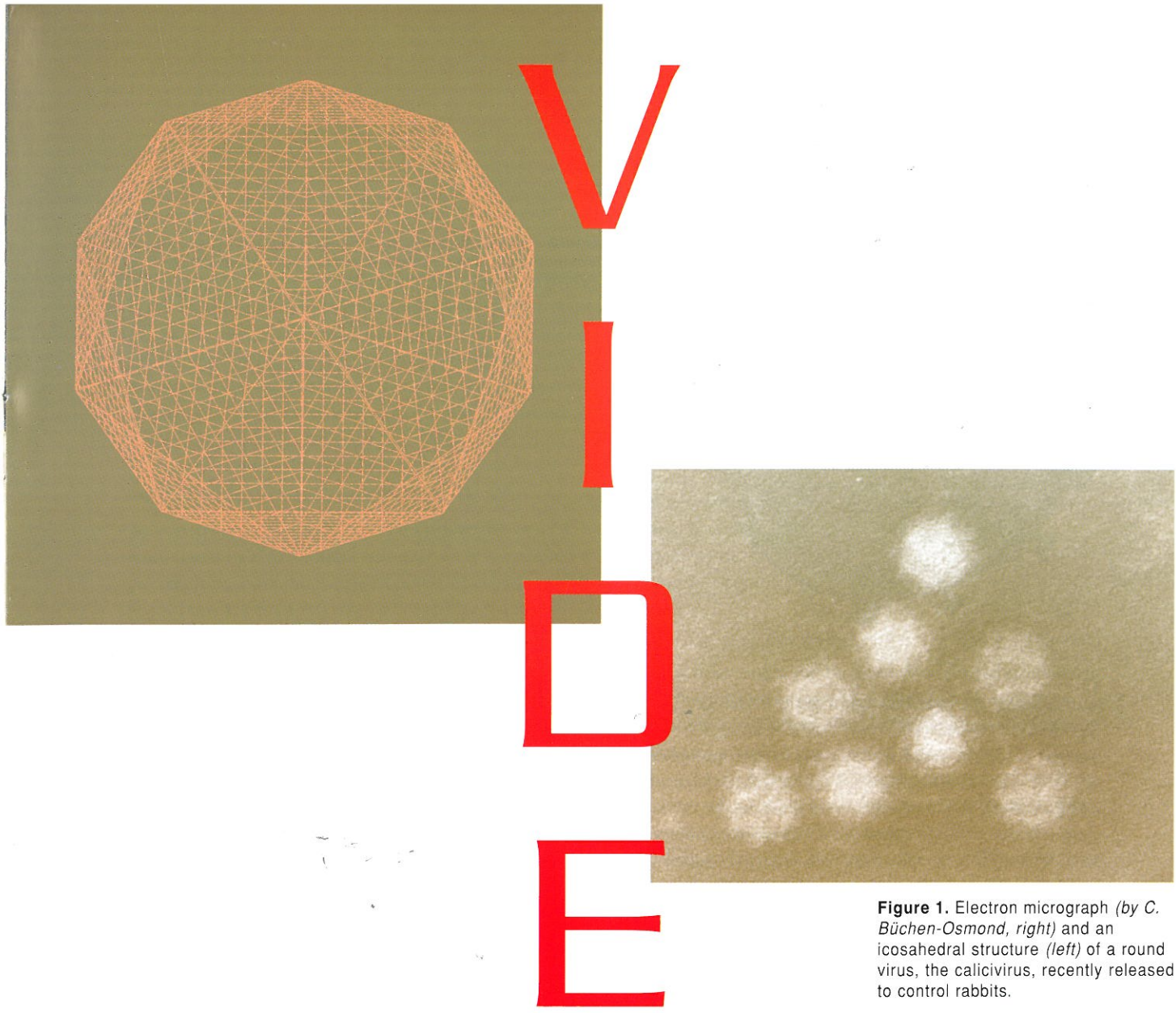


Figure 1. Electron micrograph *(by C. Büchen-Osmond, right)* and an icosahedral structure *(left)* of a round virus, the calicivirus, recently released to control rabbits.

[Latin: to see]

Years ago Professor Adrian Gibbs saw the potential for computer-based virus description, identification, and evolutionary analysis. His word play in Latin led to the Virus Identification Data Exchange project or VIDE.

Begining in the 1970s the viruses of legumes were the first to be tackled by VIDE. Legumes were selected for a number of reasons; they are host to a variety of virus species, are of considerable economic importance and the International Working Group on Legume Viruses were a group of experts willing to collate their information. This early work culminated in publication of the 'Viruses of Legumes' in 1983.

Work on the VIDE database continued, when Dr Cornelia Büchen-Osmond joined the team resulting in the publication 'Viruses of Plants in Australia' in 1988. With support from the Australian Centre for International Agricultural Research, VIDE continued to grow with; 'Viruses of Tropical Plants' appearing in 1990 and 'Viruses of Plants' in 1995.

Dr Cornelia Büchen-Osmond has now extended the vision. She is coordinating an international attempt to develop a virus database describing all viruses of animals, plants, bacteria and fungi, by the year 2000. This is the universal virus database (ICTVdB) available on the World Wide Web. An easily accessible, standardised universal virus database is of obvious benefit to scientists, medical practitioners, veterinarians and many others. The need for such a database was recognised by the International Committee on Taxonomy of Viruses (ICTV) who decided to sponsor its development in 1991.

A visionary virologist

Dr Büchen-Osmond became intensely interested in virology while employed at a health laboratory in Germany. 'I was the officer in charge of chasing up outbreaks of viruses. Samples would arrive from hospitals all over the country and I would have to identify the offending virus using electron

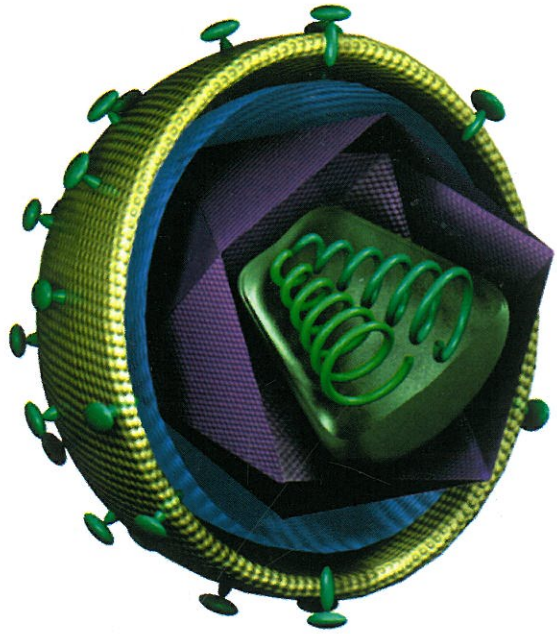


Figure 2. A model of a more complex envoloped virus, the HIV-lentivirus, with an internal icosahedral capsid containing nucleic acid genome and enzymes.

microscopy,' she explains (see Figure 1 and 2). Within thirty minutes of receiving a sample, Dr Büchen-Osmond could identify the family and genus to which the virus belonged. 'I could tell for instance, whether there was a smallpox versus a chickenpox virus. As the initial symptoms of these infections are very similar, an identification was always required urgently,' she says. 'I quickly realised that the better my own knowledge the better I was able to help,' she observes.

At the Research School of Biological Sciences, Dr Büchen-Osmond was responsible for collecting information on Australian plant viruses for the VIDE database. She was attracted to this project as she had a clear understanding of the need for a comprehensive virus database. 'When you have to deal with an unknown species, having a database on hand definitely helps the job,' she says.

The virus databases are constructed using the DEscription Language for TAxonomy software (DELTA), developed by Dr Mike Dallwitz (CSIRO) and Dr Les Watson (RSBS) (see *Revolutionising Taxonomy* page 2). DELTA has attractive attributes, apart from providing a simple and powerful method for storing taxonomic descriptions. The descriptions encoded in DELTA may be easily formatted for publication, like 'Viruses of Plants in Australia', or converted into HyperText Markup Language (HTML) for display on the Internet.

It is not surprising that Dr Büchen-Osmond was selected to coordinate the ICTV's universal database, under sponsorship from the American Type Culture Collection and the US based National Science Foundation, since she has extensive experience with virus classification as well as with the DELTA system. Designing a system which can uniquely identify all viruses presented her with an enormous challenge. As Dr Büchen-Osmond points out, 'There are five thousand species of virus, and if that doesn't sound like many, you have to remember that viruses have strains. Influenza virus, for example, only has four species but hundreds of different strains!'

Code	Morphology	Classification
48.	Family	*Paramyxoviridae*
48.1.	Subfamily	*Paramyxovirinae*
48.1.3.	Genus	*Rubulavirus*
48.1.3.4.	Subgenus	human parainfluenza serogroup 4
48.1.3.0.001.	Type Species	mumps virus
48.1.3.0.014.	Species	Newcastle disease virus

Figure 3. An unequivocal decimal system identifiying a chicken virus, Newcastle disease virus, as a close relative to mumps virus, from family to species in the universal virus database.

Naming a virus

Virus nomenclature is rather complicated. Not only is the traditional scientific nomenclature ill equipped as family, genus and species names can be very similar, but virologists of different nationalities have developed vernacular names for viruses in their own languages. 'The same virus may have many different names, and these names keep changing,' exclaims Dr Büchen-Osmond. Thus, a decimal code has been introduced for the purposes of the database.

To construct the code, all of the families of viruses were listed in alphabetical order and each assigned a two digit number. Subsequent levels of classification, ie subfamily, genus etc, are indicated by a decimal point (see Figure 3). Thus, each virus, whether it be a species or a strain, has a unique and unchangeable 'name' or reference number.

Describing a virus

The backbone to any DELTA database, whether dealing with viruses or any other organism, is its 'character list'. A character list contains many statements, each succinctly describing a unique feature or 'character' such as shape and size. The aim is for different combinations of these characters to specifically identify each and every organism within a particular database. Thus, a unique combination of characters makes up an individual organism's description. Eventually, it is hoped that all species will be classified in a single database, containing one central character list!

The virus character list is comprised of single property statements which comprehensively describe all viruses. Characters may include shape, genome composition, size, host and host range and distribution. The shape or morphology of a virus is the most revealing characteristic for identification (see Figure 4). It is possible to classify some viruses to the level of their genus from their shape alone. For the purpose of the universal virus database character list, the ICTV approved standardised characters for virus morphology and genome properties are used. There was sufficient information in this initial character list to code the families and genera of all viruses.

Conquering the Web

It took four years for Dr Büchen-Osmond to compile the universal virus database, prior to its launch on the Internet, in May 1995 *(http://life.anu.edu.au/viruses/welcome.html)*. It was certainly worth the effort, because by making this massive amount of information readily available, it is now possible for anyone to interrogate the database and thereby identify a virus (see Figure 5). Being on-line also simplifies the maintenance and expansion of the database. When launched, it comprised of 800 description files accounting for all virus families and genera. Some descriptions are accompanied by pictures of the virus, electron-micrographs showing their morphology. In addition, links have been made to both genome and protein sequence databanks, helping to identify particular strains of virus.

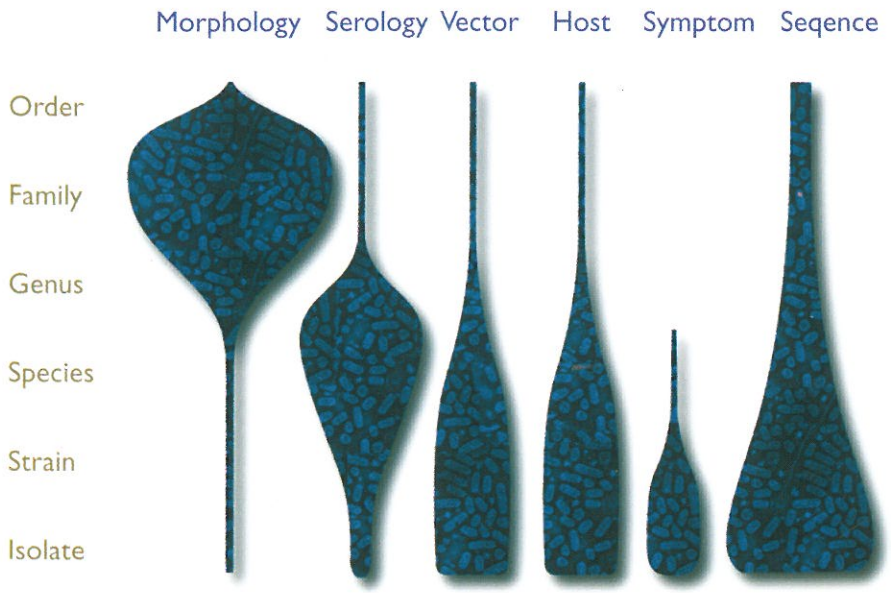


Figure 4. Relative usefulness of different types of information characterise viruses at different taxonomic levels.

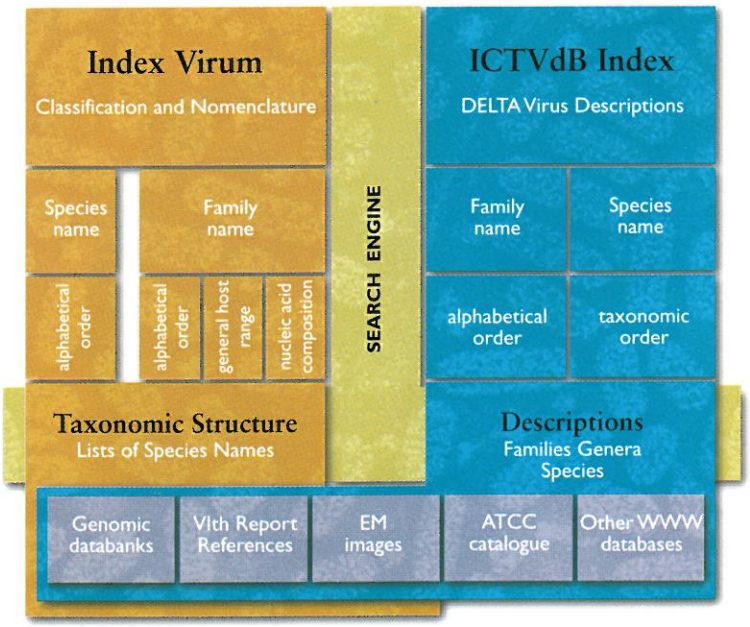


Figure 5. The menu of entry points to the ICTVdB, the universal virus database on the web, recently described as "one-stop-shopping for information on viruses".
http://life.anu.edu.au/viruses/IcTVdB/ictvdb.html

Dr Büchen-Osmond has been very excited by the global response to the on-line database, as there has been an average of two and a half thousand visitors to the database per month. Most visitors are scientists. However, Dr Büchen-Osmond has received some surprising enquires from novelists, medical practitioners and even a Warner Bothers film producer in America, who was 'looking for something that functions like a time bomb of sorts, something with a short fuse that is highly virulent, deadly and scientifically viable,' and not called a virus! She suggested prions in imported hamburger, and as usual, truth turned out to be stranger than fiction, when the role of prions in mad-cow disease burst onto front pages a few months later.

The database, which is already mirrored in Europe and the U.S, will be of greater use to medical and agricultural researchers once it contains data on the thousands of strains of viruses now recognised. The overall objective is to have a filled and fully functional ICTVdB available on the web, and on CD-Rom, by the time of the next International Virology Congress, in Sydney 1999.

Dr Büchen-Osmond is currently designing a computer tool to enable anyone in the world to submit data to the database directly via the Internet. New information is entered into the database by answering questions. New information is reviewed by ICTV study groups, to ensure the quality and accuracy of the data. 'Adding information to the database is very tedious but has to be accurate, so everything has to be done to encourage people to contribute their expertise,' explains Dr Büchen-Osmond.

Seeing the future

In many respects the current contents of the universal virus database is only the tip of the iceberg. Dr Büchen-Osmond estimates that it will take many years to 'complete' the database by entering all the information that is known about viruses. 'Then there are all those viruses of plants and animals which have not been noticed because they are not detrimental to their hosts,' sighs Dr Büchen-Osmond. 'I don't believe I need to worry about being kept busy for the rest of my life!

Ultimately, it will be up to the expertise and goodwill of the virology community as a whole to 'fill in' the database and ensure its success.

If you want to know more ...

• Virus Taxonomy: Classification and Nomclenature of Viruses. FA Murphy, CM Fauquet, DHL Bishop, SA Ghabrial, AW Jarvis, GP Martelli, MA Mayo, MD Summers (eds.)(1995) Sixth Report of the International Committee on Taxonomy of Viruses. Springer Verlag, Wein, New York.

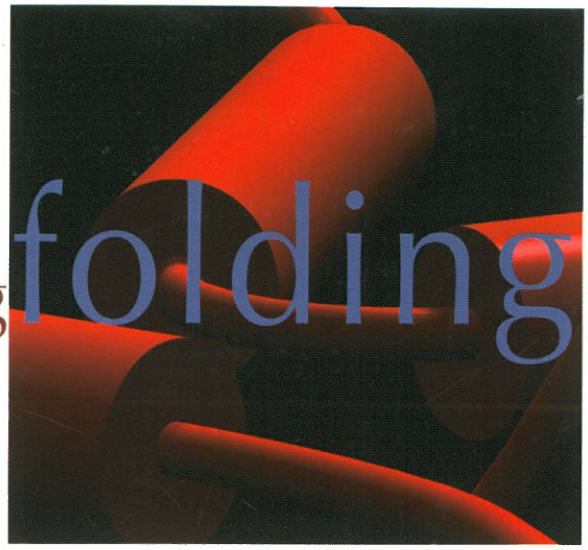
• Towards a universal virus database - progress in the ICTVdB. C Büchen-Osmond and MJ Dallwitz. (1996). Arch Virol 141:392-399.


Dr Cornelia Büchen-Osmond

• A list of viruses can be obtained at: *http://life.anu.edu.au/viruses/Ictv/index.html*

• You are welcome to contact Dr Büchen-Osmond on (02) 6249 4842, or via e-mail: Buchen@rsbs.anu.edu.au.

Unfolding folding proteins



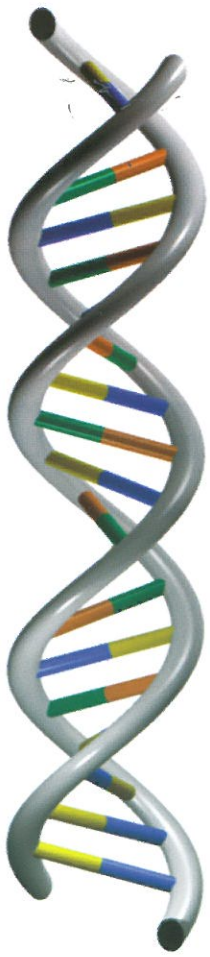


Figure 1. The double helical structure of DNA, one of the best known, Nobel Prize winning examples of insights into biomolecular structure and function.

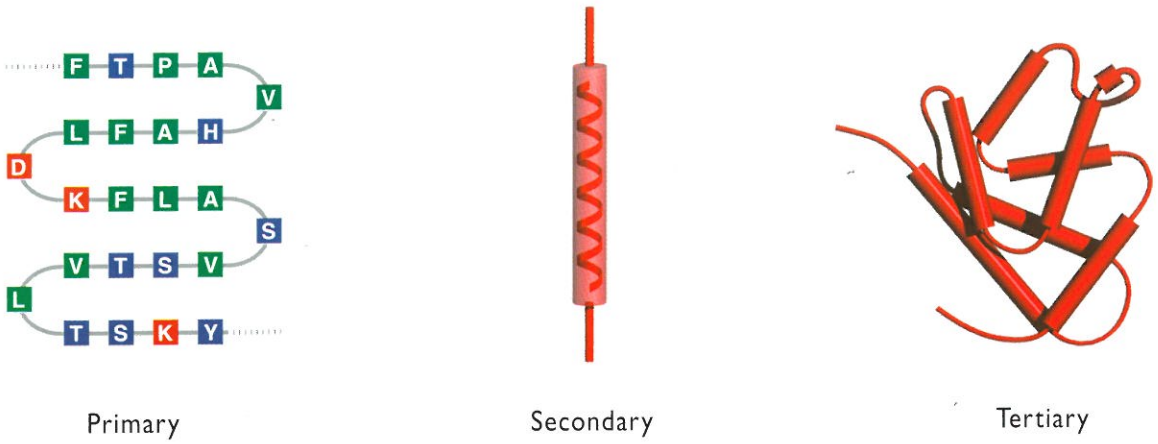
According to Dr Andrew Torda, of the Research School of Chemistry at ANU, the person who can predict the structure of a protein, without experimentation, will be awarded a Nobel Prize. Given the structural diversity of proteins, however, this may well be an unobtainable goal. So, Dr Torda has set somewhat more reasonable objectives and, along with his colleagues, is designing a computer program to recognise the 'most likely' structure of proteins.

The quest to find out how biological systems work is equivalent to the quest for the holy grail. It is sometimes said that the biggest difference between living and non-living processes is the structural complexity of biological molecules. The biologist's quest is a difficult one, as the structure of biological molecules must be determined in order to understand how they function. Proteins are one group of biomolecules with extremely complex structures.

The global flood of DNA sequence information is creating a market for novel ways of predicting the structure of the encoded proteins. Biotech companies, fighting for a competitive advantage, need a 'quick fix' to the protein problem, even if the fix is not one hundred percent correct! This demand for new methods is driving some enterprising approaches to predicting protein structure, as Dr Torda's research shows.

Building Proteins

The relationship between structure and function is, perhaps, best illustrated with the example of DNA. When Watson and Crick, discovered the double helix structure of DNA, in 1953, the elucidation of its function quickly ensued. The structure of DNA is



comparatively simple, being comprised of four kinds of building blocks, the bases, suspended from a sugar-phosphate backbone (see Figure 1). This simple structure serves the purpose of encoding proteins and replicating genetic information. A protein sequence, on the other hand, is built from 20 different kinds of amino acid and this range of building material alone generates a phenomenal order of complexity!

Several decades ago, the three-dimensional structure of the proteins myoglobin and haemoglobin had been determined. Today approximately five thousand different protein structures have been described. Despite the accumulation of all this knowledge, a simple generic set of rules to explain and predict protein structure from an amino acid sequence, let alone predict its function, has not been discovered.

This is not to say that there is little known about the chemistry and physics of protein structure, far from it! This field of research is a sophisticated and intricate science. Descriptions of the different 'levels' of protein structure exist and the terms used to describe them include; primary, secondary, tertiary and quaternary structures (see Figure 2).

Proteins are built from amino acids, joined in a sequence to form a polypeptide chain. The simplest or primary level of protein structure is the sequence of these amino acids. The polypeptide chain twists and turns in a variety of shapes, largely determined by the sequence of amino acids. Some of these secondary shapes or structures have been characterised and are called, for example, beta-turns and alpha-helices. By virtue of being bundled-up in a secondary structure, amino acids which are far apart in the linear sequence, may be brought into proximity and interact with each other. One such interaction is the formation of disulfide bonds when one of these and other interactions are established the tertiary structure of the protein is defined. Once bundled and bonded, proteins may form a conglomerate, where two or more become associated to give the final protein product.

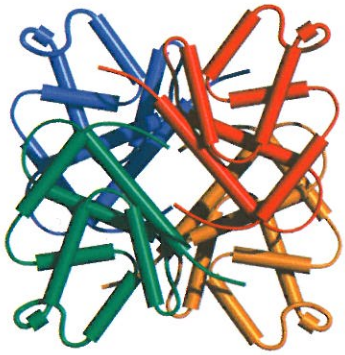
It is interesting to note that proteins with different primary structures, may have very similar tertiary structures.

Looking at proteins with old glasses

The tertiary or three dimensional structure of a protein is of most interest and the most difficult to determine. X-ray crystallography is one method used to 'see' a protein in three dimensions. Basically, the diffraction of x-rays by a protein crystal can be used to determine structure. This method has some significant limitations as it requires crystals of the protein and some proteins simply will not crystallise. In the past, this limitation often led to the analysis of proteins which could be easily crystallised, as opposed to those which were the most interesting or significant! It is fortunate that new techniques in molecular biology can circumvent some of the earlier problems. For example, to crystalise a protein requires a relatively large amount of it. Some protiens are present in normal tissues in very small amounts, however new methods can induce the artificial, 'over expression' or production of chosen proteins. Otherwise X-ray crystallographic studies would still be limited to only those proteins which could be isolated from their natural source in sufficient quantity to facilitate crystallisation. Apart from this limitation, X-ray crystallography is a terribly laborious and time consuming process, and it is said that it requires 'green thumbs'.

Another common technique for determining protein structure is called Nuclear Magnetic Resonance spectroscopy or NMR. Results may be gathered much more rapidly using this technique, however, the big drawback is that NMR only works on small, soluble proteins.

According to Dr Torda, 'Machines are not going to be powerful enough to tackle the protein structure puzzle in the near future.' There is, therefore, a need to develop clever methods to predict the structure of proteins and Dr Torda is responding to this challenge!



Quaternary

Figure 2. Modelling conventions for representation of protein structures at different levels.

A new set of lenses

Five years ago, Dr Torda was working in Zurich, examining the interaction between atoms in force fields and using these for protein simulations. At this time, an English group of researchers published a paper on predicting protein structure. The paper presented a conceptual leap. Energy was not being described in the traditional 'physical' sense where mathematical functions attempt to reproduce physical energy. Instead they had moved away from old fashioned physical concepts and applied 'pseudo-energy' and liquid simulation methods to protein simulation. 'They had launched themselves head first into blind empiricism,' says Dr Torda. In other words, this was a method for predicting the structure of proteins without worrying about their physical energy.

The old methods of predicting protein structure were unreliable and had often been discredited. As the English researchers had such success while ignoring physical energy, Dr Torda decided to utilise this novel methodology. 'Our ideas ended up converging on theirs by accident,' says Dr Torda. 'I saw their paper and thought we can do this and do it better.' He reasoned that by modifying and improving the new 'pseudo-energy' methods, he might be able to leap-frog the leaders in this field of protein structure.

The English researchers' approach to predicting protein structure used pseudo-energy to reproduce what was occurring in the physical world. Although Dr Torda was working on the same problem as his colleagues, his approach differed. He wanted to design computer programs which could 'discriminate' between a correct and incorrect protein structures (see Figure 3). There are two major differences inherent in his discriminatory approach. Firstly, an ensemble of incorrect structures, ones which do not occur naturally, are a core component of his program! Secondly, he chose to forget about the known reproducible protein structures upon which his colleagues were basing their approach and instead he built mathematical equations which discriminate between the manner in which a protein will and will not naturally fold. 'These methods are completely different and hopefully far cleverer than others in the world,' says Dr Torda.

A fuzzy view

Any amino acid sequence, greater than 100 amino acids, may be analysed using Dr Torda's program. During an analysis, each amino acid in the sequence is examined in every possible position. There will be tens of millions of incorrect positions, however the program's optimising function distinguishes the

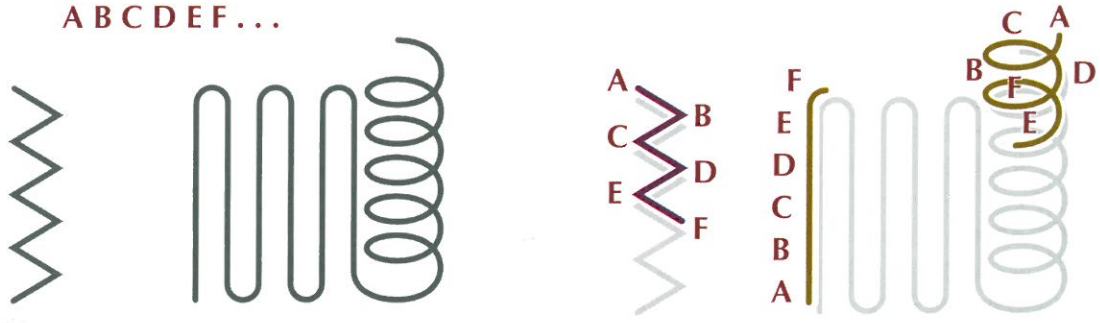


Figure 3. Generation of 'wrong' protein structures from a library of protein folds *(left)*. The test sequence (ABCDEF...) is aligned at every position of the fold library *(right)*.

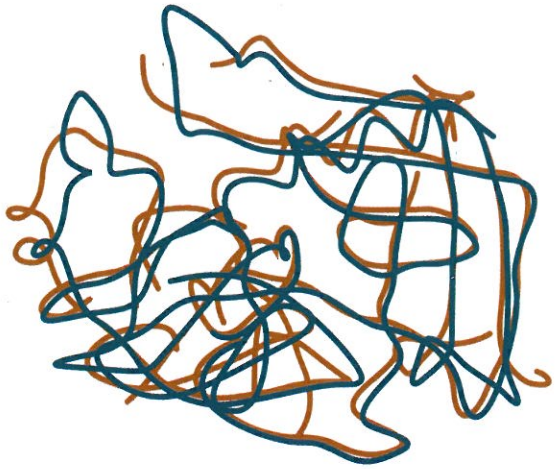


Figure 4. Mock prediction of the structure of the alpha-thrombin protein from its sequence. The blue line shows the correct folding, and the red line the structure guessed using Dr Torda's program.

correct from the incorrect positions. The result of this type of predictive analysis may not necessarily provide the absolutely correct configuration, however, the resulting structure is usually very close to its native configuration (see figure 4).

Dr Torda's research has generated a great deal of interest particularly from drug companies who are constantly finding sequences and wish to know the corresponding protein structure. 'There is currently a wealth of data waiting to be guessed at,' exclaims Dr Torda. It may seem surprising that companies would be willing to have a fuzzy view of a protein's structure. As outlined above, however, if a company were to rely on conventional methods of protein analysis, their answer would be some years and vast amounts of money away!

The main reason why companies are interested in protein structure, is that with this knowledge they may be able to manipulate its function. One enzyme in the washing powder 'Omo', for example, has been genetically engineered to chew up a stain faster, and function over a wider temperature and pH range. Such manipulations can only be done once the structure of a protein is known. Attempting to alter the function of a protein without knowledge of its structure is simply tinkering.

The major benefit from a program such as Dr Torda's is the quick insight it gives into a protein's structure. For some researchers a 'guess' at a protein's structure may be acceptable because it may be too hard or impossible to obtain a structure by alternative methods.

Looking ahead

This is a very new and extremely exciting field of science. According to Dr Torda for the moment this field is highly competitive but in five years time, it may be *passé* with attention focused elsewhere. 'This is science in action, it's all happening right now,' explains Dr Torda.

Looking ahead Dr Torda can see other important applications for his program. The binding of small molecules to proteins or 'drug design', is another hot area of research. With some new algorithms and a new approach, Dr Torda hopes he can recycle his program and predict interactions of this nature.

There is still a lot of work to be done on the protein simulation program, in order to better define its limits and increase positive predictions. Such fine tuning will escalate overall confidence in this type of analysis. 'Programs are never finished,' says Dr Torda, 'Improvement is always around and the methodology will be applicable to other structural problems.'

If you want to know more ...

- Perspectives in protein-fold recognition. A.E.Torda (1997). Curr. Opin. Struct. Biol. 7, 200-205.
- Protein structure prediction force fields: parametrization with quasi-Newtonian dynamics, P. Ulrich, W. Scott, W.F. Van Gunsteren and A.E.Torda (1997). Proteins. 27, 367-384.
- Protein fold recognition without Boltzmann statistics or explicit physical basis, T. Huber and A.E. Torda (1997). Protein Sci. (in press).

Dr Andrew Torda

You are welcome to contact Dr Torda on (02) 6279 8325, or via e-mail: Andrew.Torda@anu.edu.au

BIOLOGIC on the web

Reader Survey

Please xerox and respond, or simply write us a letter.

If future issues of Biologic were only available on the web, I could:

- ☐ access them for personal use
- ☐ access them for class use
- ☐ would prefer hard copy

My situation is that of a:

- ☐ librarian
- ☐ science teacher
- ☐ student
- ☐ other (please specify)

.....

.....

I would like to be alerted to new issues of Biologic:

- ☐ on the web
- ☐ by a poster style circular
  - ☐ mailed to the address below, or
  - ☐ multiple copies (number: .....) to the address below:

Please return to:

Pam Shearmur
Executive Assistant
Research School of Biological Sciences,
GPO Box 475, ACT, 2601

phone (02) 6249 3841
Fax (02) 6249 0758
Email shearmur@rsbs.anu.edu.au

Name and address:

.....

.....

.....

.....

. . . BUT, WHAT ABOUT THE NEXT FIFTY YEARS?

On 1 August 1996, the Australian National University celebrated its first half-century. One of many farsighted steps in the nation's postwar reconstruction, the founding of a National University, a research-only University, was seen by Dr. H.G. Coombs to be the foundation of "*a kind of intellectual powerhouse for the rebuilding of society*".

The present Director of the Institute of Advanced Studies (IAS) at ANU, Professor Sue Serjeantson believes the University was a response to "*two harsh realities*" of Australian postwar reconstruction; that we had been left behind in the nuclear age and that we had fought a war in the Asia-Pacific region with little understanding of its people, politics, cultures and languages. The founding Research Schools were thus devoted to physics, medical research, social sciences and Pacific studies.

Other research schools in chemistry, biology, earth sciences and now information sciences and engineering have been added to the IAS over the years. With centres for resource and environmental studies, for astronomy and astrophysics and for mathematical sciences, the Institute has remained the nation's pre-eminent university research organisation for half a century. In 1960, the ANU amalgamated with the Canberra University College, adding teaching and research in a broad range of disciplines, complementary to those in the Institute. A particularly welcome amalgamation with the Canberra Institute of the Arts in 1992 has built a multi-faceted university of modest size that seeks to be "*one of the world's great research institutions, distinguished by outstanding teaching, guiding students to the frontiers of knowledge and the best standards of scholarship*".

After its 50th Anniversary, we can ask how well it has achieved these goals. In 1995, the Institute was subjected to the most intensive review ever undertaken of university research in Australia. Independent, international reviewers found that "*no other Australian institution and few institutions in the world, can match the high standards of performance that we judge to have been attained by the schools and centres of the Institute*". This conclusion was reinforced by the award of the 1996 Nobel Prize in Physiology or Medicine to Doherty Zinkernagel for research in the John Curtin School of Medical Research of the IAS in the 1970's.

Clearly, the fifty-year experiment of the IAS has been a great success. On every front, whether it be as the catalyst for APEC, or basic research in visual science in RSBS that has led to a new test for glaucoma, likely to be available in every clinic within a decade, the nation continues to reap benefits of a practical kind from IAS research. The freedom the Institute accords to creative individuals to pursue research and scholarship has been instrumental in the growth of a healthier, wealthier and wiser Australia.

Readers of *Biologic*, and especially teachers preparing students for further training in the natural and social sciences, can be proud that, in the IAS and the ANU, the nation indeed possesses one of the world's great universities.

Rather than asking how it is that IAS achieves so much more than the best of Australia's other university research efforts, the review exposed the IAS to continued bureaucratic assaults on its independence, and its resources. An international outcry helped persuade Senator Vanstone to reject these attempts at bureaucratic piracy, but not before she applied the tall-poppy-topping policy to tertiary education generally. The nation-wide budget crisis in universities means that the Research School of Biological Sciences will shed 17% of its staff in 1996-97. Unfortunately, the most mobile are amongst our best.

Although cutting down tall poppies is a national passtime, most of us do recognise achievement and are willing to identify with its strength. Recent attacks on all sectors of education mean that the opportunities for achievement to the highest level, in places like the IAS, are in real danger of evaporating in the life-time of a government that seems to regard high achievement with distain. This may be a politically expedient attitude, but it is a guaranteed path to nowhere. These concerns need to be expressed at all levels of the education system, and to your local member!

These cuts also mean that this may be the last Biologic in hard copy. Please xerox and respond to our survey on the inside back cover, so that we can judge if electronic publication will be both timely and less costly. In the past, we have been heartened by so many enthusiastic responses to *Biologic*, and to hear from you now could greatly boost morale in the School. Given a chance, your best students today can be the high achievers in RSBS over the next 50 years!