100 YEARS AGO

'A Diagram of Heredity' – Francis Galton. The law of heredity which was formulated by myself … and which, as I am exceedingly gratified to learn, is now strongly corroborated by an independent investigation, has recently been illustrated by a useful diagram. … The two parents between them contribute *on the average* one half of each inherited faculty, each of them contributing one quarter of it. The four grandparents contribute between them one quarter, or each of them one sixteenth; and so on, the sum of the series 1/2 + 1/4 + 1/8 + 1/16 &c. being equal to 1, as it should be. … The area of the square diagram represents the total heritage of any particular form or faculty that is bequeathed to any particular individual. It is divided into subsidiary squares, each bearing distinctive numbers, which severally refer to different ancestors. The size of these subsidiary squares shows the average proportion of the total heritage derived from the corresponding ancestors. … The Subject of the pedigree is numbered 1. Thenceforward whatever be the distinctive number of an ancestor, which we will call *n*, the number of its sire is 2*n*, and that of its dam is 2*n* + 1.

All male numbers in the pedigree are therefore even, and all female numbers are odd.
From *Nature* 27 January 1898.

50 YEARS AGO

During a television outside broadcast from Brands Hatch, in Kent, during the afternoon of August 31, 1947, there appeared on the screens and from the loudspeakers of two television receivers … a sudden and intense increase of fluctuation noise. The visual effect was that of a violent snowstorm of the type well known to televiewers due to a motor-car ignition interference, but at a very much more intense level than seen heretofore by either of the observers. … It would seem beyond reasonable doubt that the interference observed was, in fact, solar noise, particularly when the radio of television band-width (5,000 kc./s.) to recording receiver band-width (about 20 kc./s.) is taken into account.
From *Nature* 31 January 1948.

The learning defect of *Vol* mutants resembles that caused by destruction of the mushroom bodies. Is Vol simply a protein required for the function of this brain region, or is it more central to learning? In one very nice experiment, Grotewiel *et al.* introduced a wild-type copy of the *Vol* gene into *Vol* mutants under the control of a heat-shock promoter. Inducible expression of Vol protein by heat shock three hours before the learning assay was enough to rescue learning in the mutants. As the Vol protein disappeared in the hours following heat shock, the ability to learn new olfactory tasks diminished as well. So *Vol* activity is required right at the time of learning — not during the initial development of the mushroom bodies. Furthermore, although they seem to represent a loss of gene function, *Vol* mutations are dominant, indicating that learning is sensitive to levels of Vol. The gene encodes two alternative transcripts, and mutations that disrupt either one are sufficient to give a dominant learning defect. This unusual genetic dosage-sensitivity has been observed for other important learning genes such as the cAMP mutants[3], and it strengthens the argument that *Vol* participates directly in learning.

Exactly how *Vol* acts in learning is less obvious. The Vol integrin could convey a signal to the mushroom-body neurons that acts in parallel with the cAMP signal or modulates its production. Or Vol could maintain certain interactions between neurons. Unfortunately, it is not possible to monitor the activity of *Drosophila* mushroom-body neurons but, because other learning mutants affect synaptic function at more easily studied synapses[6], the same might be true of *Vol*. One interesting possibility is that Vol-mediated adhesion contributes to the strength of synapses between two cells, and that this adhesion is dynamically regulated by learning (perhaps by cAMP signalling itself).

Integrin function can be regulated by intracellular and extracellular signals. For example, lymphocyte integrins are inactive until they receive inflammatory signals (from tissue) that engage their adhesive properties and allow invasion of the blood vessel[7]. Other cell-adhesion molecules such as *Aplysia* APCAM and *Drosophila* fasciclin II are downregulated by signals that alter the strength of synapses[8,9]. Perhaps such adhesion modulation is a more general property of neuronal plasticity. If so, these results raise the interesting (but still very speculative) possibility that learning and memory involve rapid changes in the physical properties of synaptic connections — a literal tightening or loosening of the association between two neurons. □

Cori Bargmann is in the Department of Anatomy, Howard Hughes Medical Institute, University of California, San Francisco, California 94143, USA.
e-mail: cori@itsa.ucsf.edu

1. Grotewiel, M. S., Beck, C. D. O., Wu, K. H., Zhu, X.-R. & Davis, R. L. *Nature* **391,** 455–460 (1998).
2. de Belle, J. S. & Heisenberg, M. *Science* **263,** 692–695 (1994).
3. Davis, R. L. *Physiol. Rev.* **76,** 299–317 (1996).
4. Yin, J. C., Del Vecchio, M., Zhou, H. & Tully, T. *Cell* **81,** 107–115 (1995).
5. Sonnenberg, A. *Curr. Top. Microbiol. Immunol.* **184,** 7–35 (1993).
6. Zhong, Y. & Wu, C. F. *Science* **251,** 198–201 (1991).
7. Tozer, E. C., Hughes, P. E. & Loftus, J. C. *Biochem. Cell Biol.* **74,** 785–798 (1996).
8. Bailey, C. H., Chen, M., Keller, F. & Kandel, E. R. *Science* **256,** 645–659 (1992).
9. Schuster, C. M., Davis, G. W., Fetter, R. D. & Goodman, C. S. *Neuron* **17,** 655–667 (1996).

DNA recognition

Reading the minor groove

Claude Hélène

The design of molecules that recognize specific sequences on the DNA double helix would provide new tools to control gene expression and a rational basis for fresh approaches to drug development. Parts of each base pair are exposed in the two distinct 'grooves' of DNA, the major and the minor grooves, and the sequence information of the DNA duplex is available for read-out from both of them. On page 468 of this issue[1], White *et al.* describe a molecular code for the recognition of the four Watson–Crick base pairs from the minor groove of DNA using hairpin polyamides containing imidazole, pyrrole and 3-hydroxypyrrole rings.

Double-helical DNA can bind different types of ligand which can be classified in two categories: intercalators which insert their aromatic ring between two adjacent base pairs, and groove binders which bind DNA within either groove of the double helix. Intercalators have a limited sequence specificity, because they are interacting only with two base pairs unless they are linked to a groove-recognition element — as exemplified by natural molecules, such as actinomycin; or by synthetic conjugates where an intercalator has been covalently tethered to a major groove ligand, such as an oligonucleotide[2], or to a minor groove ligand such as distamycin[3]. In living organisms, the reading of sequence information on DNA involves binding of proteins to specific control regions of the genes. These proteins exploit all three modes of binding: intercalation or partial insertion of aromatic amino acids, binding of α-helices into the major groove, and binding of loops of polypeptide chains or β-sheets into the minor groove. But no general amino-acid/base-pair code is available yet.

As Fig. 1 shows, it should be possible to distinguish the four base pairs (G•C, C•G, A•T and T•A) from the major-groove side through hydrogen-bonding interactions. Oligonucleotides provide a partial solution to this problem; they recognize oligopurine•oligopyrimidine sequences of double-helical DNA by forming triple-helical complexes[4,5]. But it remains a challenge to design nucleoside analogues that would allow oligonucleotides to recognize all four base pairs (and not only two of them)[6].

Studies in the mid-1980s, on A•T-specific minor-groove ligands, suggested that it should be possible to design ligands that could recognize G•C base pairs by replacing the pyrrole (Py) rings of distamycin or netropsin by imidazole (Im) rings (the so-called lexitropsins)[7,8]. But this strategy had limited success in recognizing extended sequences. The breakthrough came with the discovery that 2:1 complexes could form with two distamycin molecules bound side-by-side to the minor groove of A+T-rich sequences[9] — leading to the idea that covalent linkage of two such monomeric ligands would constitute a strong ligand for the minor groove at A+T-rich sequences[10,11]. Indeed, hairpin polyamides with Py/Py 'pairs' could recognize A•T and T•A base pairs[12]. Introducing imidazole instead of pyrrole rings provided a recognition element for guanine, with Im/Py pairs recognizing G•C and Py/Im recognizing C•G base pairs (the replacement of the C(3)H of pyrrole by N(3) allows for the formation of a hydrogen bond with the NH_2 group of guanine)[13–16]. The degeneracy of the recognition code for A•T and T•A remained, however.

Based on the observation of an asymmetrically placed cleft on the minor groove surface at A•T base pairs, White *et al.*[1] now report that replacing the C(3)H group of pyrrole by a bulkier C(3)OH group (in 3-hydroxypyrrole; Hp) enables hairpin polyamides to recognize all four base pairs with the following molecular code: Py/Im → C•G; Im/Py→ G•C; Hp/Py→T•A; and Py/Hp→A•T. The A•T/T•A discrimination is, however, obtained at the expense of stability, but destabilization is higher when 3-hydroxypyrrole is on the A side rather than the T side. In the example described by White *et al.* (see Fig. 2), replacing the Py/Py pair of *1* by a Py/Hp pair as in *2* decreases binding to the sequence 5′-TGGTCA-3′ 191-fold and that to 5′-TGGACA-3′ by sixfold only. Conversely, when the Py/Py pair is replaced by a Hp/Py pair as in *3*, the destabilization is reversed with 5′-TGGACA-3′ destabilized about 250-fold and 5′-TGGTCA-3′ only sixfold. This result illustrates a common phenomenon in molecular recognition: specificity may be achieved by increased destabilization of interactions at unselected sequences, rather than by enhancing interactions at the specific sites.

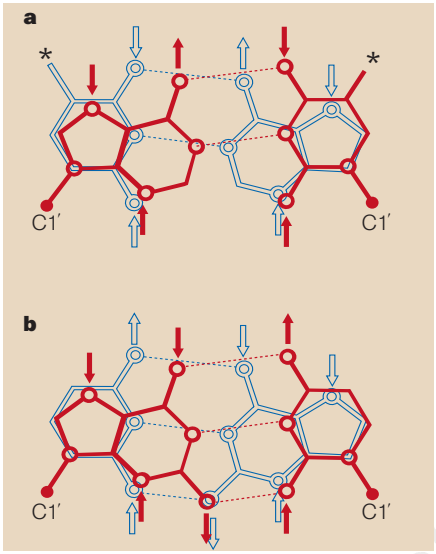

Figure 1 **Differences in DNA base-pair recognition between the major and minor grooves.** a, **A•T superimposed on T•A and,** b, **G•C superimposed on C•G, so that the C1′ atoms of the glycosidic linkages coincide. The arrows indicate the positions where hydrogen bonds may form (but not their direction), and point to the O and N hydrogen-bond acceptors and away from the NH_2 hydrogen-bond donors. The hydrogen-bonding positions in the minor groove (lower edge of the base pairs) nearly coincide when the base pairs are reversed. In contrast, the hydrogen-bond donor and acceptor groups in the major groove (upper edge), as well as the methyl group of thymines (*), occupy different positions following base pair reversal. Major-groove ligands should be much better for discriminating between the four base pairs, but White *et al.*[1] show that minor groove ligands can be designed successfully to achieve this end. (Adapted from ref. 16.)**

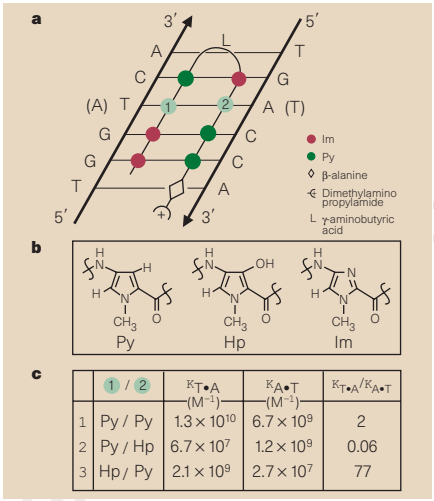

Figure 2 **Hairpin polyamides.** a, **Hairpin polyamide binding to the minor groove of B-DNA, the helix being unrolled so that the base pairs are horizontal and the helix axis vertical.** b, **Chemical structures of the *N*-methyl derivatives of pyrrole (Py), imidazole (Im) and 3-hydroxypyrrole (Hp) used by White *et al.* to build the hairpin polyamides.** c, **Binding constants for the hairpin polyamides with pyrrole and or 3-hydroxypyrrole rings at positions ① and ② and the two sequences where a T•A base pair is replaced by an A•T base pair. The last column shows the ratio of the two binding constants for each of the ①/② combinations.**

White and colleagues' results do not provide evidence for the exact molecular basis of the observed discrimination. It would be surprising if the OH group of 3-hydroxypyrrole plays only a steric role. Hydrogen-bonding interactions might lead to different distortions at A•T and T•A base pairs. Replacement of H by OH at the C(3) position of pyrrole may also change the hydration of the hairpin polyamide, and differential hydrophobic effects might be involved in discrimination of A•T from T•A. The effect might also depend on the neighbouring base pairs. The substrate described by White *et al.* (GAC versus GTC) has G•C and C•G neighbours. The A•T/T•A discrimination is reported to be observed when neighbours are also A•T base pairs, but the extent of discrimination might vary with the environment and the local polymorphism of DNA.

To achieve selective recognition of a single gene in human cells, a sequence of about 17 base pairs must be recognized[6]. This might be difficult with hairpin polyamides — the repeat distance between two consecutive units does not exactly match that between two consecutive base pairs, and flexible spacers have to be used to restore the register of the recognition elements[17]. But might recognition of fewer base pairs allow for gene-specific effects? Within the cell nucleus, DNA sequences are partly buried because of the nucleosomal organization of chromatin and the large number of proteins bound to the genetic material. If so, shorter recognition sequences might suffice, as most identical sequences might not be accessible. If the target sequences on the DNA double-helix are within gene regulatory regions then, given that they are accessible to regulatory proteins, the same should apply to ligands such as polyamides or oligonucleotides However, other sequences within the transcribed regions are also accessible, as shown for 15-mer oligonucleotides recognizing the major groove of DNA[18]. Further studies are required to tackle these issues.

Finally, we know little about the behaviour of hairpin polyamides in a cellular environment. Hairpin polyamides can inhibit gene-specific transcription in cell cultures at micromolar concentrations even though they bind naked DNA at subnanomolar concentrations[19]. We have no data yet on their uptake and compartmentalization in cells, and we don't know whether they bind to cellular components other than nucleic acids. All of these questions must be investigated for hairpin polyamides and related molecules before their promise as gene-

specific control agents *in vivo* is clearly established. □

Claude Hélène is in the Laboratoire de Biophysique, Muséum National d'Histoire Naturelle, INSERM U.201 - CNRS URA 481, 43 rue Cuvier, 75231 Paris Cedex 05, France.
e-mail: biophys@mnhn.fr

1. White, S., Szewczyk, J. W., Turner, J. M., Baird, E. E. & Dervan, P. B. *Nature* **391,** 468–471 (1998).
2. Sun, J. S. *et al. Proc. Natl Acad. Sci. USA* **86,** 9198–9202 (1989).
3. Bailly, C. *et al. Biochemistry* **33,** 15348–15364 (1994).
4. Le Doan, T. *et al. Nucleic Acids Res.* **15,** 7749–7760 (1987).
5. Moser, H. E. & Dervan, P. B. *Science* **238,** 645–650 (1987).
6. Thuong, N. T. & Hélène, C. *Angew. Chem. Int. Edn Engl.* **32,** 666–690 (1993).
7. Kopka, M. L., Yoon, C., Goodsell, D. S., Pjura, P. & Dickerson, R. E. *Proc. Natl Acad. Sci. USA* **82,** 1376–1380 (1985).
8. Lown, J. W. *et al. Biochemistry* **25,** 7408–7416 (1986).
9. Pelton, J. G. & Wemmer, D. E. *Proc. Natl Acad. Sci. USA* **86,** 5723–5727 (1989).
10. Mrksich, M. & Dervan, P. B. *J. Am. Chem. Soc.* **115,** 9892–9899 (1993).
11. Chen, Y. H. & Lown, J. W. *J. Am. Chem. Soc.* **116,** 6995–7005 (1994).
12. Mrksich, M., Parks, M. E. & Dervan, P. B. *J. Am. Chem. Soc.* **115,** 7983–7988 (1994).
13. Geierstanger, B. H., Mrksich, M., Dervan, P. B. & Wemmer, D. E. *Science* **266,** 646–649 (1994).
14. Trauger, J. W., Baird, E. E. & Dervan, P. B. *Nature* **382,** 559–561 (1996).
15. Singh, M. P., Wylie, W. A. & Lown, J. W. *Magn. Res. Chem.* **34,** F55–F66 (1996).
16. Kopka, M. L. *et al. Structure* **5,** 1033–1046 (1997).
17. Kelly, J. J., Baird, E. E. & Dervan, P. B. *Proc. Natl Acad. Sci. USA* **93,** 6981–6985 (1996).
18. Giovannangeli, C. *et al. Proc. Natl Acad. Sci. USA* **94,** 79–84 (1997).
19. Gottesfeld, J. M., Neely, L., Trauger, J. W., Baird, E. E. & Dervan, P. B. *Nature* **387,** 202–205 (1997).

Genome sequencing

Genes blossom from a weed

Joseph R. Ecker

The tiny weed *Arabidopsis thaliana* (Fig. 1) has received much attention lately, not only from plant scientists (who are already familiar with its attributes), but from genome sequencers, federal granting agencies and even the US Congress[1]. Why is this so? With its small genome[2] of around 120 million base pairs (Mb), its compact growth and the ease with which it can be genetically manipulated, *Arabidopsis* serves as a model for physiological, biochemical, cell biological and developmental studies of over 250,000 species of plant.

On page 485 of this issue, Bevan *et al.*[3] describe their analysis of just under 1.9 Mb of contiguous DNA sequence produced by the European Union *Arabidopsis* Genome Project. Unlike the larger genomes of its distant cousins — soybean, corn, wheat and most other agriculturally important crop plants — the *Arabidopsis* genome is chock-full of genes. On average, the authors found one gene every 4,800 bases and, by extrapolation from this gene density, they predict that the maximum number of genes needed to 'grow' a plant is about 21,000. This number is in line with estimates based on complementary DNA sequencing programmes[4,5].

TONI HAYDEN/JOHN INNES CENTRE

Figure 1 ***Arabidopsis thaliana* — a little weed, but a powerful model. A tiny-seeded member of the mustard family, *Arabidopsis* is the model for over 250,000 species of plant.**

Although this sequence represents only about 1.5% of the *Arabidopsis* genome, it provides a 'nuts and bolts' view of the largest contiguous segment of plant DNA sequenced to date. To put things in perspective, it is longer than most of the completely sequenced prokaryote genomes. Like previous reports of functional cataloguing[6] and whole-genome annotation, Bevan and colleagues' snapshot of predicted *Arabidopsis* genes (and the stuff in between) will probably be out of date soon after publication. However, their careful analysis of this 1.87-Mb sequence provides a tantalizing preview of the bricks and mortar needed to build a plant.

Searches of sequence databases with the 389 predicted *Arabidopsis* genes revealed that bits and pieces of just over half (some 209 genes) can be recognized in the genomes

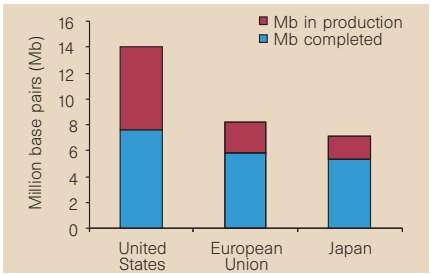

Figure 2 **Progress in the *Arabidopsis* Genome Initiative. As of December 1997, 18.90 million base pairs (Mb) of the *Arabidopsis* genome had been sequenced by the *Arabidopsis* Genome Initiative, 1.9 Mb of which are described by Bevan *et al.*[3]. Another 10.69 Mb of sequence is in production and should be completed by the middle of this year.**

of organisms ranging from the bacterium *Escherichia coli* to humans. Because of the unique nature of plants, the authors had to add several additional categories/subcategories to the functional catalogue[6] of genes. These included genes involved in the production of secondary metabolites, the source material for numerous pharmaceutical products. Moreover, because plants can't up and run from their predators, a category was established for genes involved in disease resistance, defence and responses to a variety of stresses.

Plant species diverged relatively recently, so, when complete, this catalogue of gene sequences will allow *Arabidopsis* to serve as a reference genome or 'gene bank' for all flowering plants. But the flip side of the coin reveals that it may not be easy to assign a function to 46% of the genes. Extrapolating from the data, nearly 10,000 genes in the *Arabidopsis* genome will fall into the category of 'function unknown', providing plant biologists with fertile ground for new investigation, and with a big challenge — to assign functions to these genes.

These estimates are supported by the analysis of a more broadly based collection of largely non-contiguous sequences produced by the *Arabidopsis* Genome Initiative[7,8] (Fig. 2). This international consortium of genome sequencers was established in 1996 as part of the Multinational Coordinated *Arabidopsis* Genome Research Project — a model for international cooperation in science. Bevan *et al.*[3] are involved in this project, as well as genome researchers in France, Japan and the United States. By mid-1998, the US, Japanese and EU groups are expected to deliver around 30 Mb of sequence (about one-quarter of the genome), distributed across the five *Arabidopsis* chromosomes (Fig. 2). These groups will soon be joined by researchers at a genome centre in France, which is just coming on-line.

The current rate of sequencing by the *Arabidopsis* Genome Initiative is on target with the agreed date of 2004 for completion of the genome sequence. However, this year Congress has provided additional funding to the National Science Foundation (NSF), for the establishment of a Plant Genome Research Program and a scaling up of the US effort. These extra funds will allow the target date to be brought forward. In a recent announcement by the NSF, the planned scale-up of the NSF/Department of Energy/US Department of Agriculture Interagency *Arabidopsis* Genome Sequencing Program calls for the genome sequence to be finished by the end of the year 2000. Given this boost in funding — and pending continued support for the other members of the *Arabidopsis* Genome Initiative — the international community of plant scientists should be prepared for a bountiful harvest of genes at the dawn of the new millennium. □