activity by trichostatin A leads to demethylation in *Neurospora crassa*[11]. This suggests that the HDAC complex may interact with DNA methyltransferase, thereby coupling 'maintenance' methylation with deacetylation of newly assembled nucleosomes after DNA replication.

How to complex and when?

Five methyl CpG-binding proteins have been identified in vertebrates. They share limited sequence homology outside the MBD domain (except for the homologous BMD2 and MBD3 proteins), display distinct patterns of distribution in nuclei and appear to have different binding specificities to methylated DNA (ref. 8). These observations, together with the fact that MBD proteins MeCP2, MBD2 and MBD3 are associated with distinct HDAC complexes, suggest that different MBD/HDAC complexes have different roles in gene silencing.

Distinct MBD/HDAC complexes may operate not only during different steps in gene repression, but also during different stages of development. The Mi-2/HDAC complex interacts with DNA-binding proteins to remodel chromatin during lineage determination[12]; it is plausible that the complex is required to maintain densely methylated heterochromatin during different stages of tissue differentiation. Insight into the roles of different MBDs in gene silencing will come from targeted disruption of the genes encoding these proteins *in vivo*. Mice deficient in MeCP2 die around midgestation, as do methyltransferase-deficient mice, but it is unclear whether gene expression is altered in a similar way in these mutant embryos[13].

A reluctant demethylase?

DNA demethylation reactivates gene expression, occurs at different stages of mammalian development and erases the parental methylation patterns in early embryos and developing gametes. In short, it regulates tissue-specific gene expression during differentiation. DNA demethylation may result from the inhibition of the maintenance methyltransferase, DNMT1, or indirectly, through the inactivation of chromatin-remodelling proteins, such as the SWI2/SNF2-like protein encoded by *DDM1* in plants[14,15]. Alternatively, demethylation may arise from direct removal of the methyl moiety from methylated DNA by an as-yet-unidentified demethylase.

As MBD2 is an active demethylase[16], it may have a dual role in the regulation of gene expression, either repressing gene expression through binding methylated DNA, by recruiting HDAC complexes, or reactivating a silent gene by demethylation. Ng *et al.* and Wade *et al.*, however, were unable to demonstrate demethylase activity of MBD2, despite carrying out similar biochemical analyses. It may be that the demethylase activity of MBD2 is extremely labile and vulnerable to inactivation during protein preparation or incubation. Although active demethylation clearly exists in mammalian cells, the demethylase remains elusive.

Through their interaction with methylated DNA, MBDs may also regulate DNA replication, recombination and repair. A finer understanding of how MBDs establish silencing will depend on elucidating the identity of other proteins of the HDAC complexes, and how they interact with MBDs. And critically, we must determine how distinct MBD complexes target different methylated sequences to regulate different aspects of gene silencing. □

1. Eden, S. & Cedar, H. *Curr. Opin. Genet. Dev.* **4**, 255–259 (1994).
2. Li, E. *Genomic Imprinting—Frontiers in Molecular Biology* (eds Reik, W. & Surani, A.) 1–20 (Oxford Univ. Press, Oxford, 1997).
3. Jones, P.L. *et al. Nature Genet.* **19**, 187–191 (1998).
4. Nan, X. *et al. Nature* **39**, 386–389 (1998).
5. Hampsey, M. *Trends Genet.* **13**, 427–429 (1997).
6. Ng, H.H. *et al. Nature Genet.* **23**, 58–61 (1999).
7. Wade, P.A. *et al. Nature Genet.* **23**, 62–66 (1999).
8. Hendrich, B. & Bird, A. *Mol. Cell. Biol.* **18**, 6538–6547 (1998).
9. Jeppesen, P. & Turner, B.M. *Cell* **74**, 281–289 (1993).
10. Jablonka, E., Goitein, R., Marcus, M. & Cedar, H. *Chromosoma* **93**, 152–156 (1985).
11. Selker, E.U. *Proc. Natl Acad. Sci. USA* **95**, 9430–9435 (1998).
12. Kim, J. *et al. Immunity* **10**, 345–355 (1999).
13. Tate, P., Skarnes, W. & Bird, A. *Nature Genet.* **12**, 205–208 (1996).
14. Li, E., Bestor, T.H. & Jaenisch, R. *Cell* **69**, 915–926 (1992).
15. Jeddeloh, J.A., Stokes, T.L. & Richards, E.J. *Nature Genet.* **22**, 94–97 (1999).
16. Bhattacharya, S., Ramchandani, K.S., Cervoni, N. & Szyf, M. *Nature* **397**, 579–583 (1999).

Do you dig my groove?

Russell F. Doolittle

Center for Molecular Genetics, University of California, San Diego, La Jolla, California 92093-0634, USA. e-mail: rdoolittle@ucsd.edu

Finding functions for gene products using genomic sequence as a guide has proved more difficult than anticipated: almost half of the putative open reading frames (ORFs) of recently sequenced microbial genomes fall into the 'hypothetical' category (with respect to function; ref. 1). Many had hoped that the large inventory of proteins with known function would allow most new genes to be identified on the basis of homology. Not surprisingly, the sequence 'miners' are pulling out all stops in their quest to find clues to the function of ORFs. One approach is to identify binding partners with which polypeptide X might interact. After all, many proteins exist as heterooligomers or need to interact with other proteins in one way or other.

Ménage a trois

A recent study reported by Edward Marcotte, David Eisenberg and colleagues in the pages of *Science* describes a novel strategy to identify protein-protein interactions, through a kind of *in silico* 'yeast two-hybrid' approach[2]. It is based on the observation that genes encoding proteins that interact in one organism are sometimes genetically fused (and thus encode a single polypeptide chain) in others (Fig. 1). With this in mind, they used all 4,290 proteins from *Escherichia coli* (as inferred from their ORFs) to search for 'homologues' in other organisms that are fused to other sequences, which, in turn, are homologues of other *E. coli* proteins. They call the fused proteins 'Rosetta Stone' sequences because they provide the clue to a connection between two apparently independent proteins. Using two different search methods and the ProDom database[3] (which contains a listing of all known protein domains), they found 6,809 'triplets', or about 1.5 connections per *E. coli* protein. When they applied the same approach to the 5,800 putative proteins of *Saccharomyces cerevisiae*, they found 45,000 such connections. Ancillary support for the lists was provided by various means, including

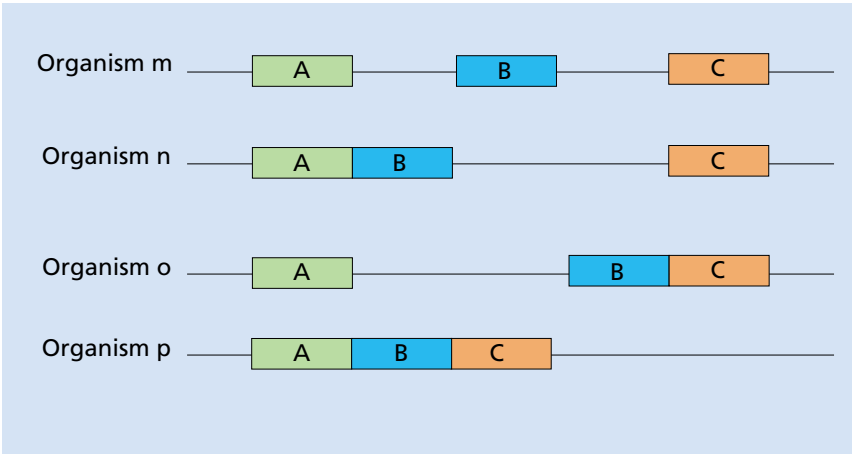

Fig. 1 Basis of the 'Rosetta Stone' approach. Described by Marcotte *et al.*[2], it identifies putative interactions between gene products. The existence of fused genes in some (for example, genes A and B in organism n) but not all organisms implies that proteins encoded by separated genes interact, either physically or functionally.

'phylogenetic profiling'[4], a method developed by the same group.

As the authors wisely acknowledge, not all of these connections are biologically significant; many involve genomically mobile domains that shuffle about from protein to protein. Accordingly, they filtered the retrieved list by removing all connections based on the most commonly exchanged domains, such as immunoglobulin domains and ATP-binding cassettes. In the end, they staked out 749 connections to *E. coli* proteins they felt had promise. They present a sample of five of these triplets, three of which represent genuine interactions, as established by previous experiment. For example, gyrase A and gyrase B are encoded separately in *E. coli* but have homology to different regions of yeast topoisomerase II, the 'Rosetta Stone' that demonstrates a functional connection. The two *E. coli* gyrase proteins do indeed interact. In contrast, two examples were offered of pairs of proteins that are not known to interact, although—perhaps not by chance—the members of each pair are known to represent associated steps in biosynthetic pathways.

The authors also explored more extended relationships of the sort where A and B might be related by one 'Rosetta Stone', and B and C by another (Fig. 1). In this way they found connections between enzymes involved in various metabolic pathways, including those involved in purine biosynthesis and the shikimate pathway (shikimate is a key intermediate in the synthesis of aromatic amino acids). They suggest that the groupings might represent multiprotein complexes, which will come as no surprise to biochemists who reported the enormous catalytic advantage of the shikimate biosynthesis complex almost 30 years ago[5].

To fuse or not to fuse

Why should such gene products be fused, anyway? The authors attribute the existence of 'Rosetta Stone' connections to the thermodynamics of associating protein subunits. The driving force, they feel, is the greater 'affinity' of different parts of a unitary protein product, when compared with that of two polypeptides that are synthesized separately. But do two genes really need to be fused to provide an advantage to the interaction of their gene products? If this were the case, why haven't all organisms availed themselves of this advantage? Why should some creatures have fused proteins whereas others do not? The α and β subunits of acetate Co-A transferase of *E. coli* would appear to stand defiant in the face of the 'thermodynamic' hypothesis. The 'Rosetta Stone' of these is the fused polypeptide succinyl Co-A transferase from humans. But the two genes sit right next to each other on the *E. coli* chromosome, allowing ample opportunity for fusion—which one would have expected to occur by now if fusion really does confer advantage.

The questions of why and how genes encoding different elements that have a common purpose are genomically arranged is longstanding. More than 30 years ago, genes encoding components of the tryptophan biosynthesis pathway, which is made up of five enzyme-catalysed steps, were found to be distributed differently in the genomes of different organisms. In *E. coli*, for example, the two subunits of tryptophan synthetase are encoded by two

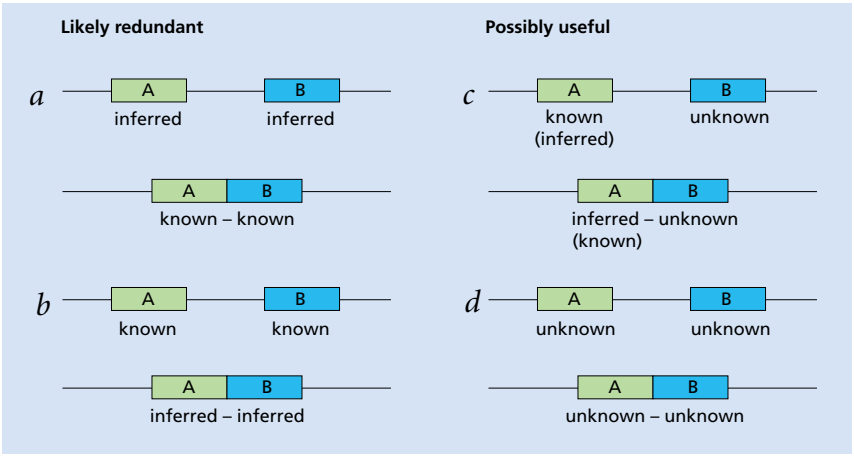

Fig. 2 The amount of new information provided by the 'Rosetta Stone' approach depends on how much is already known about the functions of the linked and/or unlinked genes. For example, if both functions of a fused pair are known (***a***), then it is likely that the functions of the individual gene products would have been inferred by standard sequence searching, and *vice versa* (***b***), as well any likely interaction between them. But if only the function of one of the parts of the fused system is known (***c***), then this approach may yield valuable information, providing a connection between the known and unknown individual genes. Even if no functions are identified for any gene (***d***), knowing that the individual genes have a connection may prove useful.

different (adjacent) genes, in contrast with *Neurospora crassa*, where a single gene encodes a product with both activites[6].

Last year Peer Bork and colleagues published a report[7] showing that interacting proteins could be identified simply on the basis of gene order, at least in Bacteria and Archaea. They searched genomes for evidence of conserved gene order as a clue to interaction, identifying 301 proteins in the process, of which more than three-quarters were found to have been reported to interact by direct experiment. They suggested that the temporal co-ordination of synthesis is advantageous for protein-protein interactions, and critical to interdependent co-translational folding[8]. Ultimately, the new findings of Marcotte *et al.* are quite similar to those of Bork and colleagues, although in the latter case, the genes are adjacent instead of fused. Clearly the thermodynamic advantage of the tether cannot be the only issue here. Perhaps the explanation has to do with differences in bacterial and eukaryotic assembly processes[9], although some 'Rosetta Stones' are provided by bacterial genomes.

One is left wondering why some genes sit next to each other in some bacteria but not in others. Even the gyrase A and B genes, while widely separated in *E. coli,* are adjacent in many other bacteria. Clearly, genomes are in constant foment, and gene order is continually reshuffled[11]. The evolutionary advantages of recombination are widely accepted. So there may be two opposing forces at play here: one that shuffles the genome, and another, that preserves gene associations for functional reasons.

This is not to say that these new genomic approaches aren't both useful and exciting. It seems certain that they will generate valuable data. The fact that all the data presented by Marcotte *et al.* are freely available on the internet will probably set off a rush of activity (their web site at www.doe-mbi.ucla.edu has already hosted more than 1,500 visits). Where to look first? I think that most of the significant connections involving known genes will turn out to be of the sort where the functional relationships were already apparent from ordinary homology searches—with occasional help from gene-order considerations. The most useful 'connections' will be those where the function of only one of the separated components (equivalent to half of the fused 'pair') is known; in these cases a genuine new finding—a clue to the function of a previously 'anonymous' gene—is possible (Fig. 2). Even if no functions have been ascribed to either component, the very existence of the fused gene product implies some kind of connection between two separated genes in the organism under scrutiny, which should get at least some biologists grooving. □

1. Doolittle, R.F. *Nature* **392**, 339–342 (1998).
2. Marcotte, E.M. *et al. Science* **285**, 751–753 (1999).
3. Corpet, F., Gouzy, J. & Kahn, D. *Nucleic Acids Res.* **26**, 323–326 (1998).
4. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. *Proc. Natl Acad. Sci. USA* **96**, 4285–4288 (1999).
5. Gaertner, F.H., Ericson, M.C. & DeMoss, J.A. *J. Biol. Chem.* **245**, 595–600 (1970).
6. DeMoss, J. *Biochem. Biophys. Res. Comm.* **18**, 850–857 (1965).
7. Dandekar, T., Snel, B., Huynen, M. & Bork, P. *Trends Biochem. Sci.* **23**, 324–328 (1998).
8. Thanaraj, T.A. & Argos, P. *Protein. Sci.* **5**, 1594–1612 (1996).
9. Netzer, W.J. & Hartl, F.U. *Nature* **388**, 343–349 (1997).
10. Lawrence, J.G. & Roth, J.R. *Genetics* **143**, 1843–1860 (1996).
11. Siefert, J.L. *et al. J. Mol. Evol.* **45**, 467–472 (1997).

Baby, don't stop!

Alexander S. Mankin[1] & Susan W. Liebman[2]

[1]Center for Pharmaceutical Biotechnology and [2]Department of Biological Sciences, University of Illinois at Chicago, 900 South Ashland Avenue, Chicago, Illinois 60607, USA. e-mail: shura@uic.edu and suel@uic.edu

A large number of human genetic diseases result from mutations that cause the premature termination of the synthesis of the protein encoded by the mutant gene[1]. A study by Elisabeth Barton-Davis and colleagues[2], exploring antibiotic effect on a mouse model of Duchenne muscular dystrophy (DMD) and published in a recent issue of the *Journal of Clinical Investigation*, suggests that gentamicin may provide a readily accessible treatment for disease caused by mutant stop codons.

The stop (nonsense) codons UAA, UAG and UGA signal the termination of protein synthesis. While the anticodons of aminoacyl transfer RNAs (tRNAs) recognize sense codons, leading to the incorporation of a specific amino acid, there are normally no tRNAs with anticodons that precisely match any of the three nonsense codons. Rather, these codons are recognized by proteins that promote the release of the completed polypeptide chain (Fig. 1*a*). When a nonsense codon in a structural gene results from mutation, the protein product is incomplete (or truncated; Fig. 1*b*) and the messenger RNA (mRNA) often rapidly degraded[3].

Start making sense

The synthesis of complete protein from such a mutant gene can be partially restored by unlinked mutations that affect the translational apparatus. For example, tRNAs that have been mutated so that their anticodon can recognize a stop codon will bind to the stop codon in competition with release factors, thereby preventing premature chain termination some fraction of the time. Alternatively, aminoglycoside antibiotics allow normal tRNAs to recognize 'incorrect' codons, including stop codons[4] (Fig. 1*c*). Aminoglycosides interact with the highly conserved decoding centre of the ribosomal RNA (rRNA). This centre normally facilitates accurate codon-anticodon pairing, but when bound with a drug, RNA conformation is altered and accuracy reduced. Depending upon the dose, these drugs may inhibit protein synthesis.

An ideal treatment of genetic disease would be to replace or supplement the mutant gene with a wild-type copy. An alternative approach to treating human genetic diseases caused by premature stop codons would be to engineer mutant human tRNA genes capable of decoding stop codons[5]. A less ideal, but accessible treatment is the use of aminoglycosides[6], as suggested in 1985 by Burke and Mogg. Using cultured mammalian cells, they showed that the aminoglycoside antibiotics paromomycin and G-418 could partially restore the synthesis of a full-size protein from a mutant gene with a premature UAG mutation. Later, G-418 and gentamicin (another aminoglycoside) were shown to restore the expression of the cys-