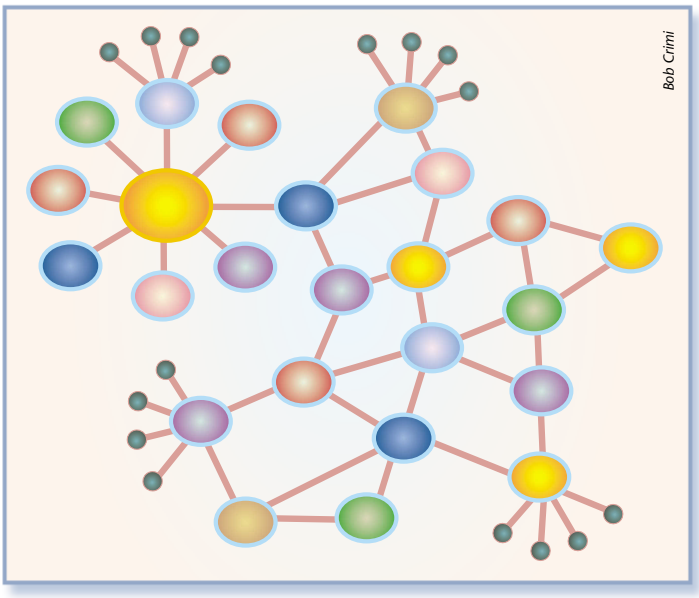
The promoter connection

Thomas Werner

Institute of Experimental Genetics, GSF-National Research Institute for Environment and Health, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany and Genomatix Software GmbH, Landsberger Strasse 6, D-80339 Munich, Germany. e-mail: werner@gsf.de

The availability of the complete genomic sequence of yeast now enables elucidation of molecular mechanisms governing gene expression patterns. New results from the yeast genome and recent advances in predicting and finding human promoters support the use of similar combinatorial approaches to study genome-wide transcriptional regulation in humans.

Since the sequence of the yeast genome was determined in its entirety several years ago, it has served as a catalog of the genes contained in a single genome. On the other hand, large-scale gene expression studies using DNA microarrays now provide clues to the dynamic processes of life, especially gene transcription, that are not immediately evident from the genomic sequence. Ultimately, however, array experiments give no direct clues to the underlying regulatory processes, which are determined to a large extent by the promoters of the genes. On page 153 of this issue, Yitzhak Pilpel and colleagues[1] now show how to link these two sets of data—genomic sequence and expression arrays—by associating the occurrence of transcription-factor binding sites (TF sites) in the yeast genome with expression profiles of groups of genes.



Only connect. The hub-and-spoke-style system of transcriptional regulation in yeast, as represented by an abstract version of the 'combinogram' described by Pilpel and colleagues[1] (see page 153).

The idea is well established. What makes this study remarkable is that Pilpel *et al.*[1] applied the approach to a genome-wide analysis and focused on TF sites common to sets of promoters before looking at the expression profiles of those genes. By concentrating on TF sites, they were able to distinguish between coregulated and coexpressed genes: coregulation is the consequence of a common molecular mechanism, whereas coexpression may occur by mere coincidence.

Other researchers have already observed limited correlation of individual TF sites with gene expression patterns[2,3]. However, such correlations remain generally weak, as most TF sites occur in many promoters, regardless of the particular expression pattern. In mammalian systems in particular, transcriptional specificity often depends on combinations of at least two TF sites, rather than being an intrinsic property of individual sites. Several studies have proved this point quite convincingly for individual genes[4–6]. Pilpel *et al.*[1] have taken advantage of the somewhat simpler yeast system to make a genome-wide analysis based on the concept of synergistic TF-site action.

It is, of course, very satisfying to see that the conclusions drawn from a few individual cases apparently hold in general, at least for the yeast genome. This generalization extends to the interactions between a variety of TFs within yeast regulatory networks, in which different functionalities can be conferred on one factor by its association with different cofactors. This is an important advance over the boolean-style networks established for the description of metabolic regulatory networks, which usually record only the dependence of genes without a clue to the molecular basis[7]. The current study, in contrast, identifies relationships on the molecular basis of synergistically acting TF sites, ensuring functional resolution to the level of nucleotide sequence. Although statistical correlations are of limited value on their own, they are greatly strengthened by combination with experimental observations. In this respect, Pilpel *et al.*[1] give an elegant demonstration of the synergistic power of integrating 'wet-lab' analyses with bioinformatics.

The modular concept of genome-wide transcription control that comes out of this study may be important in resolving a real oxymoron presented by genome analyses: a highly complex organism, such as a human, and the relatively simple worm *Caenorhabditis elegans* have only a relatively moderate difference in mere gene number, in spite of the clearly visible differences between them. The similarity in numbers of genes becomes even more pronounced when we turn to the difference between human and chimpanzee, where not only is the gene count almost identical, but the amino-acid sequences of the encoded proteins most likely differ only at single positions (on the basis of the expected 99% identity at the nucleotide-sequence level).

Leroy Hood brought out this point in his talk at the Bio2001 meeting in San Diego (24–28 June 2001) with his statement that "The difference between man and monkey is gene regulation." I agree with that view, and would like to extend it to the molecular level: the difference between an individual TF site and transcriptional function is the modular context

of a promoter or enhancer within the genomic sequence.

The study by Pilpel *et al.*[1] is certainly a milestone in linking genomic sequences (rather than cDNAs) with functional high-throughput data such as expression arrays. But how well can this approach be transferred from yeast to mammals, especially humans? The modular design of promoters in yeast and humans is apparently very similar from an organizational point of view, but there are also profound differences. The first and most daunting is that human promoters are much more elusive than their yeast counterparts. Whereas yeast promoters can be easily found upstream of open reading frames, human promoters might be located tens of kilobases upstream of their reading frames because of non-coding leader exons and introns. Human promoters are also generally more complex than yeast promoters and use quite different sets of TFs.

Fortunately, most of these obstacles to large-scale analysis of human promoters have been overcome. The first prerequisite, availability of the human genomic sequence, was met last year. Recent significant advances in promoter prediction[8] now enable identification of a large proportion of human promoters, as was demonstrated on chromosome 22 (ref. 9). Compilations of human and other vertebrate TF sites are available[10–12], although still incomplete. Thus the basis has been laid for large-scale analysis of human sequences and expression array results. Somewhat more sophisticated analysis strategies will be required, because of the larger number and more complex organization of human promoters. The principles successfully applied by Pilpel *et al.* should, however, hold true for humans. This means that their study is good news for everyone looking beyond merely descriptive results to the molecular basis of expression array data. □

1. Pilpel, Y., Sudarsanam, P. & Church, G.M. *Nature Genet.* **29**, 153–159 (2001).
2. Brazma, A., Jonassen, I., Vilo, J. & Ukonnen, E. *Genome Res.* **8**, 1202–1215 (1998).
3. van Helden, J., Andre, B. & Collado-Vides, J. *J. Mol. Biol.* **281**, 827–842 (1998).
4. Yuh, C.H., Bolouri, H. & Davidson, E.H. *Science* **279**, 1896–1902 (1998).
5. Klingenhoff, A., Frech, K., Quandt, K. & Werner, T. *Bioinformatics* **15**, 180–186 (1999).
6. Fessele, S., *et al. FASEB J.* **15**, 577–579 (2001).
7. D'haeseleer, P., Liang, S. & Somogyi, R. *Bioinformatics* **16**, 707–726 (2000).
8. Scherf, M., Klingenhoff, A. & Werner, T. *J. Mol. Biol.* **297**, 599–606 (2000).
9. Scherf, M., *et al. Genome Res.* **11**, 333–340 (2001).
10. Quandt, K., Frech, K., Karas, H., Wingender, E. & Werner, T. *Nucleic Acids Res.* **23**, 4878–4884 (1995).
11. Chen, Q. K., Hertz, G. Z., Stormo, G. D. *Comp. Appl. Biosci.* **11**, 56—566 (1995).
12. Wingender, E., *et al. Nucleic Acids Res.* **29**, 281–283 (2001).

The adaptable *lin-39*

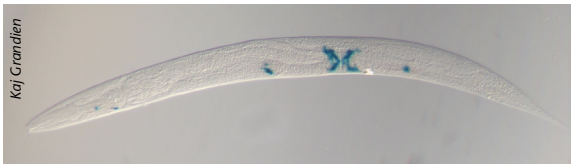
Helen M. Chamberlin

Department of Molecular Genetics, Ohio State University, Columbus, Ohio 43210, USA. e-mail: chamberlin.27@osu.edu

Comparative studies of nematode development provide a powerful framework for investigating the evolution of developmental mechanisms. A recent report also demonstrates how comparative work can inform our understanding of basic developmental signaling pathways. In particular, investigation of the differences in vulva development between *Caenorhabditis elegans* and *Pristionchus pacificus* has clarified the molecular relationship between an epidermal growth factor–Ras–MAP kinase signaling pathway and downstream Hox transcription factor activity.

A molecular geneticist will often submit a gene sequence to a database for comparison and will be pleased to identify regions of sequence similarity. The assumption is that sequence similarity allows one to extrapolate a possible molecular function for the gene based on previous work on other species. In contrast, molecular geneticists with an interest in evolutionary biology may focus on the regions of sequence differences, reasoning that they might point to molecular functions that differ, and thus underlie evolutionary change.

This strategy has been used in a study by Kaj Grandien and Ralf Sommer[1], reported in a recent issue of *Genes & Development*. They have investigated the differences in the nematode gene *lin-39*, a member of the nematode *Hox* cluster that is required for developmental specification in the mid-body region[2]. They show that the differences in *lin-39* function between two nematode species—*C. elegans* and *P. pacificus*—derive not from differences in the protein, but probably from differences in gene regulation. More noteworthy, however, is that their evolutionary perspective provides a conceptual context in which a specific hypothesis about the relationship between LIN-39 and an epidermal growth factor receptor (EGFR)–Ras–MAP kinase (MAPK) signal transduction pathway can be tested. Their work illustrates how comparative developmental genetics provides a unique vantage from which to clarify the molecular mechanisms of developmental pathways.



X marks the spot. *C. elegans lin-39::LacZ* transgene expression restricted to the vulva muscles of an adult nematode. Photo courtesy of Kaj Grandien.

The vulva in evolution

Comparative analyses of *C. elegans* and *P. pacificus* rely on a significant body of work that details the cellular and molecular features of vulva development in both species. In each, the vulva (egg-laying structure) forms in the mid-body region, and is produced by the specialized division and differentiation of ventral epidermal cells. The prospective vulva cells represent only a subset of cells capable of forming vulva tissue, and they divide only in response to a signal(s) from the overlying gonad. The differences between the two species are in the details. For example, in *C. elegans*, the EGF-related signal LIN-3 derives from a single gonadal cell (the anchor cell) and stimulates an EGFR–Ras–MAPK signal transduction pathway in the responding epidermal