pseudogenes seem more likely to have arisen by some rare event rather than by the extensive mobility that characterizes mobile SINEs [G. R. Daniels and P. L. Deininger, *Nature* **317**, 819 (1985)].
43. A. F. Smit, *Curr. Opin. Genet. Dev* . **6**, 743 (1996).
44. R. F. Ketting, S. E. J. Fischer, R. H. Plasterk, *Nucleic Acids Res.* **25**, 4041 (1997).
45. G. Bernardi, *Annu. Rev. Genet.* **29**, 445 (1995); B. Dujon *et al.*, *Nature* **369**, 371 (1994).
46. The abundance of *C. elegans* ESTs does not directly reflect expression levels, because they are derived from cDNAs in which more abundantly expressed genes were partially selected against (*6*, *7*).
47. T. M. Barnes, Y. Kohara, A. Coulson, *Genetics* **141**, 159 (1995).
48. This approach is also being used for the human genome (Sanger Centre, Washington University Genome Sequencing Center, *Genome Res.*, in press).
49. For methodological details, see (*20*) or (*21*). For biochemical procedures, see R. K. Wilson and E. R. Mardis, in *Genome Analysis: A Laboratory Manual*, B. Birren, E. D. Green, S. Klapholz, R. M. Myers, J. Roskams, Eds. (Cold Spring Harbor Laboratory Press, Plainview, NY, 1997), vol. 1, pp. 397–454. For software packages, see (*20*) or (*21*) and S. Dear *et al.*, *Genome Res.* **8**, 260 (1998); M. Wendl *et al.*, *ibid.*, p. 975; J. D. Parsons, *Comput. Appl. Biosci.* **11**, 615 (1995); and M. Cooper *et al.*, *Genome Res.* **6**, 1110 (1996).
50. B. Ewing, L. Hillier, M. C. Wendl, *Genome Res.* **8**, 175 (1998); B. Ewing and P. Green, *ibid.*, p. 186.
51. P. Green, personal communication.
52. J. K. Bonfield, K. F. Smith, R. Staden, *Nucleic Acids Res.* **23**, 4992 (1995).
53. D. Gordon, C. Abajian, P. Green, *Genome Res.* **8**, 195 (1998).
54. H. M. Robertson, *Genome Res.* **8**, 449 (1998).
55. This work has been supported by grants from the U.S. National Human Genome Research Institute and the UK MRC. We would also like to thank the many members of the *C. elegans* community who have shared data and provided encouragement in the course of this project.

Zinc Fingers in *Caenorhabditis elegans*: Finding Families and Probing Pathways

Neil D. Clarke and Jeremy M. Berg

REVIEW

More than 3 percent of the protein sequences inferred from the *Caenorhabditis elegans* genome contain sequence motifs characteristic of zinc-binding structural domains, and of these more than half are believed to be sequence-specific DNA-binding proteins. The distribution of these zinc-binding domains among the genomes of various organisms offers insights into the role of zinc-binding proteins in evolution. In addition, the complete genome sequence of *C. elegans* provides an opportunity to analyze, and perhaps predict, pathways of transcriptional regulation.

Less than 15 years ago, it was suggested that repeated sequences found in transcription factor IIIA (TFIIIA) of *Xenopus* might fold into structural domains stabilized by the binding of zinc to conserved cysteine and histidine residues (*1*–*3*). Klug and co-workers further noted that "it would not be surprising if the same 30 residue units were found to occur in varying numbers in other related gene control proteins" (*1*). This proposal proved remarkably prescient: *Caenorhabditis elegans*, for example, turns out to have more than 100 such proteins, and the number of domains per protein varies from one to perhaps as many as fourteen. Unanticipated at the time, though, was the fact that the zinc-binding motif found in TFIIIA is just one of many small zinc-binding domains, a number of which are involved in gene regulation. The properties of a few of these domains have been summarized recently (*4*).

Eukaryotes contain a much greater number of proteins with well-characterized zinc-binding motifs than do bacterial and archaeal organisms (Table 1). The complete genome of *Caenorhabditis elegans* (a metazoan), in conjunction with that of *Saccharomyces cerevisiae* (a yeast), presents a special opportunity to examine the range and diversity of these gene families in eukaryotes. Furthermore, because some of these zinc-binding motifs are sequence-specific DNA-binding proteins, the availability of nearly complete sequence information also permits a preliminary analysis of the distribution of potential binding sites within the entire genome. Such analyses may prove to be of value in deducing development control pathways and in more fully defining the characteristics of eukaryotic promoters.

Department of Biophysics and Biophysical Chemistry, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA.

The Cys_2His_2 Family

The zinc-stabilized domains of TFIIIA are known as "zinc fingers" or Cys_2His_2 domains. The consensus sequence for this family is (Phe, Tyr)-X-Cys-X_{2-4}-Cys-X_3-Phe-X_5-Leu-X_2-His-X_{3-5}-His (*5*–*7*). In both *C. elegans* and the yeast *S. cerevisiae*, roughly 0.7% of all proteins contain one or more Cys_2His_2 zinc finger domains (Table 1). However, the distribution of these domains within proteins is rather different in the two organisms. In yeast, the majority of zinc finger proteins contain exactly two domains, and only a few (~10%) have more than two. In contrast, there are more zinc finger proteins in *C. elegans* that have three or more Cys_2His_2 domains than there are proteins that have exactly two (Fig. 1) (*8*). On the basis of the sequences of mammalian and *Drosophila* zinc finger proteins, it appears that the distribution of Cys_2His_2 domains among *C. elegans* proteins is typical of multicellular organisms.

The GATA, LIM, and Hormone Receptor Families: Implications for Metazoan Evolution

The GATA domain, the LIM domain, and the DNA-binding domains from nuclear hormone receptors each include a four-cysteine zinc-binding domain that can be clustered into the same structural superfamily, and it is possible that they share a common evolutionary origin (Fig. 2) (*9*, *10*). In addition to the Cys_4 superfamily domain, LIM domains contain a similar LIM-specific Cys_2HisCys zinc motif, whereas the hormone receptors have a second and distinct Cys_4 domain. GATA proteins frequently contain a pair of Cys_4 superfamily domains.

Normalized to the number of genes in their respective genomes, the number of GATA and LIM domain homologs is similar in *C. elegans* and *S. cerevisiae*. In striking contrast, the hormone receptor family is completely absent in yeast but is the largest single family of zinc-binding domains in *C. elegans*. In fact, with over 200 family members, the hormone receptors make up nearly 1.5% of the entire coding sequence of *C. elegans*. The differences in the distribution of nuclear hormone receptors in *C. elegans* and *S. cerevisiae* may be relevant to the evolution of multicellular animals. As has been noted before, the evolution of hormone receptors may have been a key event in the development of cell-cell communication and the origins of multicellularity in the metazoa (*11*).

The ligand-binding domains of the hormone receptors have diverged considerably more than the DNA-binding domains. Applying the same criterion for significance to both the DNA- and ligand-binding domains of the hormone receptor family, only about 10% of the open reading frames (ORFs) that have a DNA-binding domain

appear to have a ligand-binding domain. However, among genes containing hormone receptor DNA-binding domains, the scores for potential ligand-binding domains are typically higher than those seen in ORFs that do not have the DNA-binding domain. For example, over 40% of the DNA-binding domain ORFs have ligand-binding domain scores that exceed by more than 2 SD the mean score for ORFs that lack the DNA-binding domain. Furthermore, when we used a hidden Markov model (HMM) constructed from some of these top-scoring worm domains, over 90% of the DNA-binding domain ORFs (most of which were not used in constructing the HMM) now had ligand-binding domain scores that exceeded those for unrelated genes by more than 3 SD. We believe, therefore, that most of the hormone receptor homologs in *C. elegans* do have sequences related to the ligand-binding domain.

Identification of Genes That May Be Regulated by the TRA-1A Zinc Finger Protein

Several of the common zinc-binding motifs function as sequence-specific DNA-binding domains, including the Cys_2His_2 zinc fingers. With a complete genome sequence in hand, a comprehensive analysis of potential binding sites becomes possible; this, in turn, raises the possibility that certain aspects of transcriptional regulation might be predictable on the basis of genomic sequence analysis. As a test case, we conducted a preliminary analysis of potential TRA-1A–binding sites in the *C. elegans* genome. TRA-1A, which is a product of the *tra-1* gene, was chosen for this analysis because its binding specificity has been well characterized (*12*) and because it belongs to a subfamily of Cys_2His_2 proteins that is of exceptional biological interest. TRA-1A is a close homolog of *Drosophila cubitus interruptus* (segment polarity gene) and of human GLI (oncogene) and GLI3 (cranio-facial development) (*13*). Furthermore, a crystal structure has been determined for the zinc finger region from GLI bound to a DNA site (*14*).

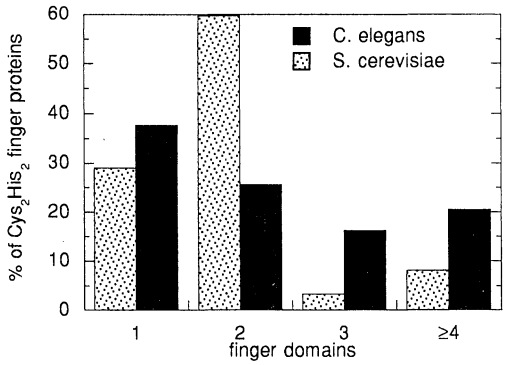

Fig. 1. Distribution of finger domains among Cys_2His_2 zinc finger proteins in *C. elegans* and *S. cerevisiae*.

In *C. elegans*, *tra-1* activity is necessary for animals to develop into females or hermaphrodites (*15–18*). As the last in a line of global regulators of sexual development, *tra-1* controls the specialized pathways that lead to sex-specific development in different tissues. One way *tra-1* activity could lead to female animals is by repressing genes whose expression would otherwise lead to the development of male-specific features. Expression of *mab-3*, for example, leads to the development of male-specific peripheral sense organs in males, but in females and hermaphrodites *tra-1* activity blocks this pathway by reducing *mab-3* mRNA levels (*19*, *20*). This reduction in the steady-state level of *mab-3* transcripts could be due to direct repression by TRA-1A.

Using DNA sequences that were selected in vitro for tight binding to TRA-1A, the binding-competent gene product of *tra-1*, we constructed an HMM for TRA-1A–binding sites (*12*, *21*). HMMs provide a probabilistic definition of binding sites on the basis of the nucleotide frequencies observed experimentally at each position and are presumably a more realistic predictor of in vivo binding sites than are simple consensus sequences. We used the TRA-1A HMM to identify about 1300 potential binding sites in the *C. elegans* genome (*22*). The distribution of these sites within 5′ extragenic regions differs from random distributions in the existence of five genes that have three or more upstream TRA-1A sites (Fig. 3A). Strikingly, *mab-3* is among this very small subset of genes, which supports the idea that *mab-3* transcription is repressed directly by the binding of TRA-1A.

Some of the other genes besides *mab-3* that have a large number of potential TRA-1A sites might also be regulated by TRA-1A. The most interesting of these other genes is *lin-31* (Fig. 3A) (*23*). Like *mab-3*, *lin-31* is required for development of sex-specific tissues and is a putative transcription factor. Unlike *mab-3*, though, *lin-31* is required for development of a lineage that is female- and hermaphrodite-specific rather than male-specific (*23*). Thus, if *tra-1* does regulate *lin-31* (which remains to be shown experimentally), it might well be expected to activate transcription rather than repress it. The remaining three genes identified in the upstream binding site analysis are not related to sexual development in any obvious way. However, one is a TATA-binding protein associated factor (TAF) and another is homologous to a protein with antiproliferative activity (*24*). Whether expression of these genes is affected by *tra-1* is unknown at present.

The *mab-3* gene product is itself a putative transcription factor containing a novel zinc-binding motif (*20*). Because our analysis of the *C. elegans* genome "predicted" the regulation of *mab-3* by TRA-1A, we extended this analysis by attempting to predict genes that might be regulated in turn by *mab-3*. Data from unpublished binding site selection experiments (*25*) were used to construct an HMM that was then used to search the *C. elegans* genome with a cutoff score chosen to yield a number of sites similar to that found in the TRA-1A analysis. As shown in Fig. 3B, the distribution of these binding sites

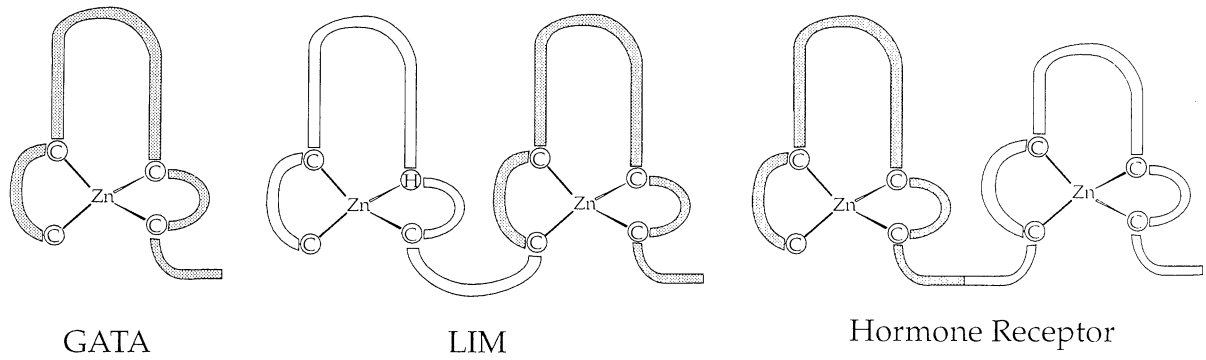

Fig. 2. Schematic views of the zinc-binding regions from the GATA, LIM, and hormone receptor families.

upstream of *C. elegans* genes indicates that fewer genes have significantly large numbers of sites than was the case with TRA-1A. Nevertheless, there is an excess of genes with two sites over the number expected from a random distribution, and there is one gene with three upstream sites. Intriguingly, the gene with three upstream sites is the *C. elegans* ortholog of *hunchback*, a *Drosophila* gene that encodes a Cys_2His_2 zinc finger transcription factor important in development.

A few genomic sequences to which *Drosophila* Hunchback binds in vitro have been identified (*26*, *27*). On the basis of these data, we could perhaps extend our predictions for this regulatory pathway one more step by assuming that *C. elegans* Hunchback recognizes the same binding sites as *Drosophila* Hunchback. However, our experience with MAB-3 and its close *Drosophila* homolog DSX indicates that such predictions should await experimental determination of binding site specificity for the *C. elegans* protein. The in vivo binding specificity of *Drosophila* DSX^M (the male-specific product of the *doublesex* gene) must be fairly similar to that of MAB-3, because ectopically expressed DSX^M can functionally replace *mab-3* to some extent (*20*). Furthermore, there are sequences to which both proteins will bind in vitro with reasonable affinity (*25*). Nevertheless, the distribution of binding sites obtained by in vitro selection experiments is quite different for the two homologs (*25*, *28*), and use of *Drosophila* DSX binding site data (instead of the MAB-3 data) gives a distribution of predicted binding sites in the *C. elegans* genome that is not substantially different from a random distribution (*29*).

Potential Autoregulation by the GATA Homolog ELT-1

As a final example of how binding site distributions can be used to assess regulatory issues in a complete genome, we considered the *C. elegans* GATA family member *elt-1* (*30*). Spieth *et al.* suggested previously that the *elt-1* gene may be autoregulated because there are multiple matches to a consensus GATA-binding site [(A/T)GATA(G/A)] within a few hundred base pairs upstream of its initiation codon (*30*). However, because there are more than 200,000 matches to the consensus GATA site in the *C. elegans* genome, the question arises as to whether the number of GATA sites upstream of *elt-1* is unusually large. In an analysis similar to that described above for the *tra-1* and *mab-3* gene products, we searched the *C. elegans* genome for matches to the canonical GATA recognition site and determined the number of these sites within 500 base pairs of the first predicted exon for each gene. This distribution was then compared with a set of random distributions. As shown in Fig. 4, the number of sites associated with *elt-1* does, in fact, place it among a set of 25 genes that have an unusually large number of such sequences.

The Challenges of Regulatory Pathway Prediction

Counting the number of upstream sites that exceed some threshold for similarity to a binding sequence is a rather simple-minded approach to predicting transcriptional regulation and one that will undoubtedly lead to some incorrect predictions. Among the complicating factors that are not captured by simple enumeration of binding sites are the spacing and

Table 1. Zinc-binding domains were identified with HMMs and the HMMER program package version 1.8.4 (*21*, *38*). Only motifs involved in DNA binding or in protein-protein interactions are considered here; enzymes that use catalytically active zinc sites are ubiquitous but were not examined. The C3H and DM HMMs were constructed from published sequence alignments, with the addition of *C. elegans* ORF K08B12.2 to the DM alignment (*20*, *39*). All other HMMs were from the Pfam database (*38*). A threshold of 10 bits was used as the criterion for significance for all database hits reported here. The database of *C. elegans* ORFs was current as of 10 July 1998 and is available with other supplementary information at www.sciencemag.org/feature/data/985286.shl. The database of *S. cerevisiae* ORFs (orf_trans-.fasta) was obtained from genome-ftp.stanford.edu/pub/yeast/yeast_ORFs. Databases for *Escherichia coli* (ecoli.faa) and *Methanococcus jannaschii* (mjan.faa) were obtained from ncbi.nlm.nih.gov/genbank/genomes/bacteria in their respective subdirectories. A general overview of the data sets and analysis is available at www.sciencemag.org/feature/data/c-elegans.shl.

	Genome			
	C. elegans	*S. cerevisiae*	*E. coli*	*M. jannaschii*
ORFs searched (*N*)	16,626	6,215	4,289	1,715
Zinc finger	117	42	0	0
Hormone receptor	233	0	0	0
GATA	9	7	0	0
LIM	33	3	0	0
DM	8	0	0	0
Zinc cluster	1	54	0	0
C3H	20	3	0	0
RING finger	97	30	0	0
Nucleocapsid	17	8	0	0
Total	535	147	0	0

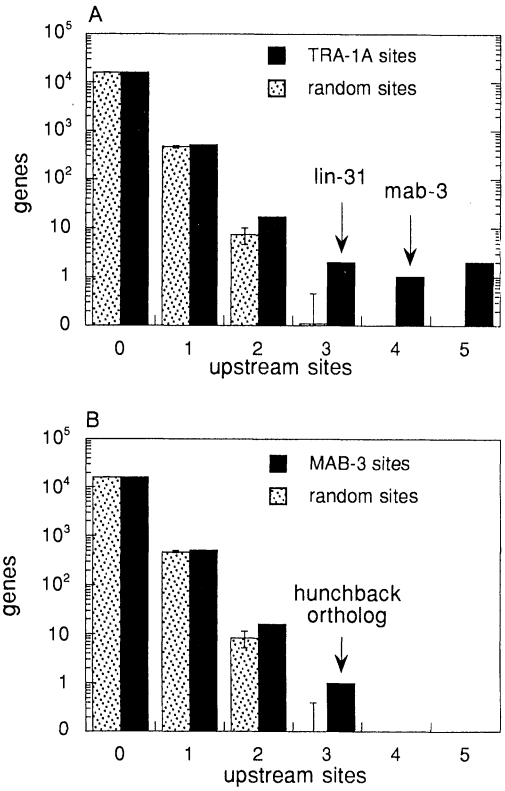

Fig. 3. (**A**) Distribution of potential TRA-1A–binding sites. Of the 1299 TRA-1A sites in *C. elegans* (*22*), 561 are in intergenic regions and no more than 4 kb from the first predicted exon in the gene. The number of genes that have 0, 1, . . . , 5 upstream TRA-1A sites is indicated by the black bars. As a control, 1299 random sites were picked within the genome, and their distribution with respect to the ORFs was determined by the same criteria. This random distribution was generated 100 times, and the mean and standard deviation for the number of genes having a given number of sites were calculated. The stippled bars show the mean random value, and the error bars indicate the standard deviation. Five genes had three or more upstream TRA-1A sites, which is highly significant according to the randomized distributions. The five genes are C03C11.2, Y95B8A_75.a, Y53C12B.5a (*mab-3*), K10G6.1 (*lin-31*), and F08F3.9. Gene names and predicted exons are from genome feature files received 23 June 1998 from J. Spieth of the *C. elegans* Genome Center, Washington University, St. Louis, Missouri. (**B**) Distribution of potential MAB-3–binding sites based on site selection experiments (*25*). The random distributions were calculated as described for TRA-1A but with the number of potential MAB-3–binding sites found, 1346. The gene with three upstream sites is F13D11.2. BLAST searches with this ORF indicated high sequence similarity to Hunchback homologs, and a reciprocal search of the *C. elegans* genome with the *Drosophila* Hunchback sequence showed that this is the only ORF that is strikingly similar (*29*).

orientation of binding sites, cooperative interactions among different proteins, and competition for binding by proteins with similar or overlapping DNA-binding sites. Despite these caveats, much of what we have inferred about the regulation of most genes is based precisely on this kind of simple identification of upstream sequence elements. At the very least, the existence of complete genomic information provides, for the first time, the means to evaluate the statistical significance of such sites without having to make assumptions about the composition and gene distribution of the genome. Furthermore, the example of *mab-3* regulation by *tra-1* offers some encouragement that even this simple approach to the problem could prove fruitful in some cases.

As important as the prospects for prediction is the use of the genome sequence in understanding the complexities of transcription initiation control and in interpreting genome-wide transcription studies (*31*). If we are to really understand how the transcriptional regulation of nearly 20,000 genes is coordinated in *C. elegans*, as opposed to simply cataloging genes and the proteins that affect their expression, then computational analysis of the genome will be an indispensable adjunct to experimental studies.

The modular nature of Cys_2His_2 zinc finger proteins and the relatively simple way in which some members of the family bind DNA had previously led to the idea that simple rules might be found for predicting the sequence specificity of zinc finger proteins (*32*). Indeed, a few rules have been developed and have proven useful in designing proteins to recognize particular DNA sequences (*32*, *33*). However, natural zinc finger proteins are too diverse in terms of both their presumptive DNA-contact residues and the length and sequence of the linkers that connect the fingers for these rules to be usefully applied in a general way in the prediction of specificity.

The probing of regulatory pathways will clearly require careful experimental determination of binding site preferences for all classes of DNA-binding proteins. However, with the acquisition of more (and better) binding data and with the availability of high-throughput technologies to measure transcript levels of essentially all the genes in an organism, the computational analysis of transcriptional regulation is sure to progress rapidly.

Conclusions

Zinc-binding units such as the Cys_2His_2 zinc finger domains are present in a large number of gene products, representing some of the largest protein families in the *C. elegans* genome. Although bacteria and archaea do contain some proteins that bind zinc, they appear to lack the large families of zinc-binding domains like those found in yeast, worms, and other eukaryotes. This suggests that these zinc-binding domains may not be truly ancient units but instead evolved later as genome size and cell sophistication increased. Of particular importance may have been the evolution of efficient mechanisms for zinc homeostasis. Yeast and other eukaryotes have recently been shown to contain proteins for importing and exporting zinc as well as other potential components of such a system (*34*, *35*). If bacteria and archaea did not evolve systems for zinc homeostasis, then the use of zinc-dependent proteins for gene regulation in these organisms may have been disadvantageous.

Comparison of the two available eukaryotic genomes reveals some striking differences. Although several families, such as the Cys_2His_2 zinc finger, RING finger, and nucleocapsid domains, are of comparable size, particularly when normalized for genome size, other families show extremely skewed distributions. As noted above, the hormone receptor superfamily is the largest single family of zinc-binding domains found in *C. elegans*, yet these proteins are not found in yeast. Another family, the zinc cluster proteins, typified by GAL4, is the largest family in yeast, yet only one putative family member (not authenticated) is encoded by the *C. elegans* genome.

Because some of the zinc-binding domains function by sequence-specific interactions with DNA, the completed genome has facilitated preliminary attempts to identify potential gene regulatory pathways in silico. Similar methods could be applied to other DNA-binding proteins of known binding specificity. Further development of such analysis procedures may provide important insights into the myriad gene regulatory pathways that are necessary for the development and growth of multicellular organisms.

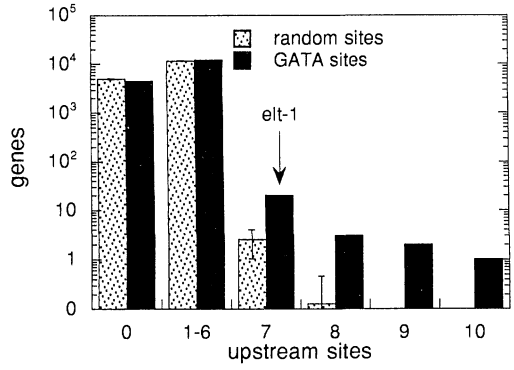

Fig. 4. Distribution of ELT-1 (GATA)–binding sites (black bars). These data were obtained in a manner analogous to that described for TRA-1A (*22*) (Fig. 3) except that only sites within 500 bp 5′ of the first exon were considered. Genes with between one and six GATA sites have been grouped together because the number of genes with *N* GATA sites does not become significantly larger than the number of genes with *N* randomly distributed sites until $N \geq 7$. The mean values from 15 random distributions are indicated by the stippled bars; error bars indicate the standard deviations.

References and Notes

1. J. Miller, A. D. McLachlan, A. Klug, *EMBO J.* **4**, 1609 (1985).
2. R. S. Brown, C. Sander, P. Argos, *FEBS Lett.* **186**, 271 (1985).
3. S. Böhm and B. Drescher, *Stud. Biophys.* **107**, 237 (1985).
4. J. Berg and Y. Shi, *Science* **271**, 1081 (1996).
5. J. M. Berg, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 99 (1988).
6. G. Párraga *et al.*, *Science* **241**, 1489 (1988).
7. M. S. Lee, G. P. Gippert, K. V. Soman, D. A. Case, P. E. Wright, *ibid.* **245**, 635 (1989).
8. S. Böhm, D. Frishman, H. W. Mewes, *Nucleic Acids Res.* **25**, 2464 (1997).
9. G. C. Perez-Alvarado *et al.*, *Nature Struct. Biol.* **1**, 388 (1994).
10. J. W. Schwabe and A. Klug, *ibid.*, p. 345.
11. V. Laudet, C. Hanni, J. Coll, F. Catzeflis, D. Stehelin, *EMBO J.* **11**, 1003 (1992).
12. D. Zarkower and J. Hodgkin, *Nucleic Acids Res.* **21**, 3691 (1993).
13. ———, *Cell* **70**, 237 (1992).
14. N. P. Pavletich and C. O. Pabo, *Science* **261**, 1701 (1993).
15. J. A. Hodgkin and S. Brenner, *Genetics* **86**, 275 (1977).
16. J. Hodgkin, *Genes Dev.* **1**, 731 (1987).
17. ———, *Nature* **344**, 721 (1990).
18. T. Schedl, P. L. Graham, M. K. Barton, J. Kimble, *Genetics* **123**, 755 (1989).
19. M. M. Shen and J. Hodgkin, *Cell* **54**, 1019 (1988).
20. C. S. Raymond *et al.*, *Nature* **391**, 691 (1998).
21. S. R. Eddy, *Curr. Opin. Struct. Biol.* **6**, 361 (1996).
22. Data from (*12*) were used to generate in the computer a large number of sequences having the same overall distribution of nucleotide preferences as those determined experimentally for selected *tra-1*–binding sites. The sequences were then used to construct an HMM with the "hmmbuild-f" program in HMMER v2.0 (hmmer.wustl.edu). In the same way, an HMM for the complementary site was constructed with sequences complementary to those used to construct the first HMM. This allows searches to be conducted for both orientations of the site. Because the amount of information in short DNA sequences is low relative to the size of the genome, the significance of even a good match to the HMM is considered suspect by the "hmmsearch" program, and only the first hit in the entire genome is output. To allow all matches in the genome to be detected, we artificially increased the significance of a match by doubling the match state scores in the HMMs. With these modified HMMs, a score threshold of 15 correlated subjectively well with similarity to the consensus TRA-1A site and was used for all further analyses of TRA-1A sites. With this criterion, 1299 sites were found in the *C. elegans* genome. These and other search results can be found at www.sciencemag.org/feature/data/985286.shl. The same procedure was used to identify potential MAB-3 sites with unpublished binding site selection experiments (*25*). At a score threshold of 20, 1346 sites were identified in the *C. elegans* genome. A similar strategy was also used to identify ELT-1–binding sites, except in this case, a score cutoff was chosen that allowed only perfect matches to the consensus sequence (A/T)GATA(A/G). Over 200,000 sites were found.
23. L. M. Miller, M. E. Gallegos, B. A. Morisseau, S. K. Kim, *Genes Dev.* **7**, 933 (1993).
24. C03C11.2 has five upstream *tra-1* sites and is homologous to a protein called Tob

(transducer of erbB-2) (*36*). F08F3.9 has three upstream sites and is homologous to SNAP45, a human TAF (*37*).
25. D. Zarkower, personal communication.
26. J. Treisman and C. Desplan, *Nature* **341**, 335 (1989).
27. D. Stanojevic, T. Hoey, M. Levine, *ibid.*, p. 331.
28. S. Erdman, H. Chen, K. Burtis, *Genetics* **144**, 1639 (1996).
29. N. D. Clarke and J. M. Berg, data not shown.
30. J. Spieth, Y. H. Shim, K. Lea, R. Conrad, T. Blumenthal, *Mol. Cell. Biol.* **11**, 4651 (1991).
31. S. Chu *et al.*, *Science* **282**, 699 (1998).
32. J. R. Desjarlais and J. M. Berg, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 7345 (1992).
33. M. Isalan, A. Klug, Y. Choo, *Biochemistry* **37**, 12026 (1998).
34. D. J. Eide, *Annu. Rev. Nutr.* **18**, 441 (1998).
35. B. Vallee and K. Falchuk, *Physiol. Rev.* **73**, 79 (1993).
36. S. Matsuda *et al.*, *Oncogene* **12**, 705 (1996).
37. R. W. Henry, C. L. Sadowski, R. Kobayashi, N. Hernandez, *Nature* **374**, 653 (1995).
38. E. L. Sonnhammer, S. R. Eddy, R. Durbin, *Proteins* **28**, 405 (1997).
39. M. T. Worthington, B. T. Amann, D. Nathans, J. M. Berg, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 13754 (1996).
40. We thank all involved in the sequencing of the *C. elegans* genome for providing the raw material for this analysis. We especially thank L. Hillier and J. Spieth at the Washington University Genome Sequencing Center for their help in running preliminary searches, providing data files, and promptly answering all of our technical questions. We are also grateful to D. Zarkower for helpful comments on the manuscript, to W. Yi and D. Zarkower for sharing MAB-3 binding data before publication, and to J. Hodgkin for alerting us to the existence of these data. Research on zinc-binding proteins in the laboratory of J.M.B. has been supported by NIH and Sangamo Biosciences. J.M.B. is a member of the Scientific Advisory Board of Sangamo Biosciences. N.D.C. received salary support from the National Institute of Standards and Technology while on a part-time sabbatical.

Comparison of the Complete Protein Sets of Worm and Yeast: Orthology and Divergence

Stephen A. Chervitz, L. Aravind, Gavin Sherlock, Catherine A. Ball, Eugene V. Koonin, Selina S. Dwight, Midori A. Harris, Kara Dolinski, Scott Mohr, Temple Smith, Shuai Weng, J. Michael Cherry, David Botstein

REVIEW

Comparative analysis of predicted protein sequences encoded by the genomes of *Caenorhabditis elegans* and *Saccharomyces cerevisiae* suggests that most of the core biological functions are carried out by orthologous proteins (proteins of different species that can be traced back to a common ancestor) that occur in comparable numbers. The specialized processes of signal transduction and regulatory control that are unique to the multicellular worm appear to use novel proteins, many of which re-use conserved domains. Major expansion of the number of some of these domains seen in the worm may have contributed to the advent of multicellularity. The proteins conserved in yeast and worm are likely to have orthologs throughout eukaryotes; in contrast, the proteins unique to the worm may well define metazoans.

The nematode worm *Caenorhabditis elegans* is only the second eukaryote to have its genome completely sequenced (*1*). The first complete eukaryotic genome sequence, that of the budding yeast *Saccharomyces cerevisiae*, has been reported previously (*2*). Thus, for the first time, it is possible to compare the entire complements of encoded proteins of two highly diverged eukaryotic species, one of which is a unicellular microorganism and the other a multicellular animal.

The first result is quite surprising: Simple sequence comparisons allow one to predict, more often than not, orthologous pairs. In many cases, orthologous pairs can be confidently delineated even within families of highly similar proteins having many members from each organism. In fact, at the most stringent comparison value, ~57% of protein pairs contain just one worm and just one yeast protein. The set of highly conserved proteins is encoded by a minority of the open reading frames (ORFs) in each organism (~40% of yeast and 20% of worm; see Table 1). They carry out the core biological processes shared by these two eukaryotes, such as intermediary metabolism, DNA and RNA metabolism, protein folding, trafficking, and degradation.

The second result is more in line with expectation. Unlike yeast, the worm has a number of specialized, committed cell types with distinct and coordinated programs of gene expression. The differentiation of cell types in the animal is achieved through an elaborate developmental program that has been explored in detail in *C. elegans* (*3*). In contrast, yeast adapts dynamically to its environment by switching on different gene batteries in response to nutrient status, oxygen tension, mating pheromones, and other factors (*4*). It is widely believed that the physical basis of the developmental complexity of a multicellular eukaryote is a system of protein regulators and signal transducers that is significantly more complex than that in unicellular organisms (*5*). Interspecies comparison of the protein domains used in regulation and signal transduction shows that although there is considerable sharing of domains, most of the proteins in which they appear are generally not orthologous. Increasing numbers of multidomain proteins during eukaryotic evolution are thought to have originated largely by domain shuffling (*6*). Indeed, we can predict evolutionary trends including (i) the evolution of new regulatory or signaling domains; (ii) evolution of new domain architectures from shared (presumably preexisting) domains; and (iii) expansion of particular domain families by a series of duplications.

The comparison of 6217 yeast ORFs with 19,099 worm ORFs produces much more information than can possibly be printed here. All of the underlying data, however, can be found in searchable form on our Web site within the Saccharomyces Genome Database (SGD) (genome-www.stanford.edu/Saccharomyces/worm/).

Shared Core Biology of Worm and Yeast: The Orthologs

We set out to compare and contrast the encoded protein complements by identifying both orthologous proteins (*7*), and shared and novel protein domains in yeast and worm. Distinguishing orthologs, which have evolved by vertical descent from a common ancestor and are presumed to carry out the same function (*8*), from paralogs, which arise by duplication and domain shuffling within a genome and hence may have divergent functions, is paramount when carrying out whole genome comparisons (*9*). Failure to do so can result in functional misclassification (*10*) and inaccurate molecular evolutionary reconstructions (*11*). In this part of our analysis, we did not attempt to detect distant homologs, which may be found by using less stringent criteria and more sensitive techniques (*12*).

We compared the predicted proteins of yeast and worm by first carrying out reciprocal WU-BLASTP (*13*) comparisons (that is, each predicted yeast protein against all the predicted proteins of the worm and vice versa). In every case in which a high-scoring pair (HSP) was

S. A. Chervitz, G. Sherlock, C. A. Ball, S. S. Dwight, M. A. Harris, K. Dolinski, S. Weng, J. M. Cherry, and D. Botstein are in the Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305–5120, USA. L. Aravind and E. V. Koonin are at the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. S. Mohr and T. Smith are in the Department of Biomedical Engineering, Boston University, Boston, MA 02115, USA.