



## Finding Functional Features in Saccharomyces Genomes by Phylogenetic Footprinting

Paul Cliften, *et al.*  
*Science* **301**, 71 (2003);  
DOI: 10.1126/science.1084337

**The following resources related to this article are available online at [www.sciencemag.org](http://www.sciencemag.org) (this information is current as of January 9, 2007):**

**Updated information and services**, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/301/5629/71>

**Supporting Online Material** can be found at:

<http://www.sciencemag.org/cgi/content/full/1084337/DC1>

This article **cites 34 articles**, 19 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/301/5629/71#otherarticles>

This article has been **cited by** 286 article(s) on the ISI Web of Science.

This article has been **cited by** 93 articles hosted by HighWire Press; see:

<http://www.sciencemag.org/cgi/content/full/301/5629/71#otherarticles>

This article appears in the following **subject collections**:

Genetics

<http://www.sciencemag.org/cgi/collection/genetics>

Information about obtaining **reprints** of this article or about obtaining **permission to reproduce this article** in whole or in part can be found at:

<http://www.sciencemag.org/help/about/permissions.dtl>

# Finding Functional Features in *Saccharomyces* Genomes by Phylogenetic Footprinting

Paul Cliften,<sup>1</sup> Priya Sudarsanam,<sup>1</sup> Ashwin Desikan,<sup>1</sup>  
Lucinda Fulton,<sup>2</sup> Bob Fulton,<sup>2</sup> John Majors,<sup>3</sup> Robert Waterston,<sup>1,2</sup>  
Barak A. Cohen,<sup>1</sup> Mark Johnston<sup>1\*</sup>

The sifting and winnowing of DNA sequence that occur during evolution cause nonfunctional sequences to diverge, leaving phylogenetic footprints of functional sequence elements in comparisons of genome sequences. We searched for such footprints among the genome sequences of six *Saccharomyces* species and identified potentially functional sequences. Comparison of these sequences allowed us to revise the catalog of yeast genes and identify sequence motifs that may be targets of transcriptional regulatory proteins. Some of these conserved sequence motifs reside upstream of genes with similar functional annotations or similar expression patterns or those bound by the same transcription factor and are thus good candidates for functional regulatory sequences.

Functional non-protein-coding DNA sequences, such as gene regulatory elements, are difficult to recognize because they are usually short, often degenerate, and can reside on either strand of DNA at variable distances from the genes they control. Because functional sequences tend to be conserved through evolution, they can appear as “phylogenetic footprints” in alignments of genome sequences of different species (1–3). To investigate the use of phylogenetic footprinting for identifying gene regulatory elements on a genome-wide scale, we compared the genome sequences of six yeast species. On the basis of our initial analysis of the *Saccharomyces* phylogeny (4), we selected for sequencing three *Saccharomyces sensu stricto* (“strict sense”) species that are relatively closely related to *Saccharomyces cerevisiae* (*S. mikatae*, *S. kudriavzevii*, and *S. bayanus*). Their intergenic sequences average 59 to 67% identity to their *S. cerevisiae* orthologs in global alignments (table S1). At this evolutionary distance, nonfunctional sequences have diverged enough to allow many functional sequence signals to stand out from the “noise,” but the sequences retain enough overall similarity to enable their alignment. Because of the relatively high degree of similarity of sequences at these evolutionary distances, genome sequences of several species need to be compared to lend sufficient acuity to the phylogenetic footprints.

We also chose to include in the analysis two *Saccharomyces* species that are more distantly related to *S. cerevisiae* (*S. castellii* and *S. kluyveri*) (5). To estimate the degree to which these species have diverged from *S. cerevisiae*, we compared the sequences of synonymous codons, because few of their intergenic sequences align to their *S. cerevisiae* orthologs. *S. castellii* and *S. kluyveri* average 33.9% identity to *S. cerevisiae*, compared with 54.5% identity for the sensu stricto species (table S1). Several *Saccharomyces* species are approximately this diverged from *S. cerevisiae* (4). We chose these particular species for sequencing because they are thought to have the smallest genomes (6). In addition, *S. kluyveri* is of particular biological interest because of an unusual aspect of its physiology: Unlike most other *Saccharomyces* species, it does not primarily ferment glucose (7). We included these species in the analysis mainly for two reasons. First, we wished to compare the relative utility of sequences at different phylogenetic distances for comparative sequence analysis. Second, based on our preliminary analysis (4), we expected that many intergenic regions would be too highly conserved among the sensu stricto species to provide adequate definition of functional sequences, and more divergent sequences were expected to sharpen the definition of conserved sequence motifs. An additional reason for obtaining genome sequences of the more distantly related species is that, unlike the case with the closely related species, functional domains of their proteins are often apparent in multiple sequence alignments.

The sensu stricto species are so closely related that their genome organization is almost identical; only a few chromosomal rearrangements have occurred in these species, making their chromosomes almost complete-

ly syntenic with their *S. cerevisiae* counterparts (8, 9). In contrast, many chromosomal rearrangements have occurred in the genomes of the two species that are more distantly related to *S. cerevisiae*, resulting in relatively short stretches of chromosomes that are syntenic with the *S. cerevisiae* genome (8).

**Genome sequencing.** When a complete, highly accurate genome sequence is available, as it is for *S. cerevisiae*, a relatively small amount of sequence data of related genomes is sufficient for substantial comparative analysis. We therefore sought only twofold to threefold genome coverage in random (“shotgun”) sequence reads (10) of the five genomes, which is expected to yield 85 to 95% coverage of the sequence of each genome (11). We then used a semiautomated method for closing the gaps between assembled contigs (12). Because we were primarily interested in identifying functional elements in intergenic sequences, we limited our finishing efforts to gaps between contigs that fall in these regions of the genome. These strategies made this a relatively economical project. Details of sequence assembly and finishing are presented in table S2.

**Improving genome annotation.** DNA sequence comparisons can be used to refine genome annotation. Even for a small, well-studied, and extensively annotated genome like that of *S. cerevisiae*, in which genes are relatively easy to recognize, our analysis affected annotation of more than 10% of the genes. We were able to predict 43 previously unannotated genes [all of them small open reading frames (ORFs) fewer than 100 codons in length, including 7 with an intron], based on their degree of sequence conservation (13). We also predict that 515 annotated genes are false, based on the nonsense codons and frameshift mutations present in their orthologs in the other species (14). We noticed several likely oversights in the current annotation of the *S. cerevisiae* genome (15). For example, the intron branch sequence tACTAAC appeared frequently as a sequence motif conserved in several orthologous intergenic regions, signaling the presence of potential introns. This allowed us to recognize 40 likely introns that had not been annotated, based on consensus 5' splice donor and 3' splice acceptor sequences flanking the conserved intron branch sequence. Our comparative genome analysis leads to a more accurate gene count for *S. cerevisiae* of 5773 (16), rather than the 6331 genes currently annotated in the *Saccharomyces* Genome Database (SGD) (17).

**Alignment of intergenic sequences.** Our primary goal was to identify functional non-protein-coding sequences. We therefore compared entire intergenic regions, which are relatively short in *Saccharomyces* [average ~500 base pairs (bp)]. Intergenic sequences of the sensu stricto species are similar enough that most orthologous sequences can be accurately identified and aligned with CLUSTALW (18), but the

<sup>1</sup>Department of Genetics, <sup>2</sup>Genome Sequencing Center, <sup>3</sup>Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, 660 South Euclid Avenue, St. Louis, MO 63110, USA.

\*To whom correspondence should be addressed. E-mail: mj@genetics.wustl.edu

two distantly related species are so diverged from *S. cerevisiae* that their intergenic sequences almost never aligned to the sequence of their *S. cerevisiae* ortholog. The orthologous intergenic sequences of the distantly related species could only be identified by aligning the sequences of their associated predicted proteins [using BLASTX (19)]. Over half of the intergenic regions are available from all four sensu stricto species (20); 40% are available from all six species (21). The genome sequences are available in the public databases (22); the CLUSTALW alignments of the intergenic sequences can be obtained at [www.genetics.wustl.edu/saccharomycesgenomes/](http://www.genetics.wustl.edu/saccharomycesgenomes/).

The four-way CLUSTALW alignments of orthologous intergenic sequences of the closely related sensu stricto species had an average sequence identity of 37.1%. This is significantly more identity than expected for nonselected (neutral) sequences at these phylogenetic distances (~16%, see table S1). The distribution of sequence identity within intergenic regions was not uniform: A peak of conservation spanned approximately 125 to 250 bp upstream of the translational start codon (23), suggesting that this region is enriched in regulatory sequence elements (Fig. 1A) (24). This is consistent with the view that most regulatory sequences in yeast promoters lie relatively close to the genes they regulate (25). It also suggests that a substantial number of the conserved sequence elements are conserved because they are functional, rather than because of their shared ancestry (the latter case would be expected to result in their more uniform distribution in intergenic regions). This is in contrast to the relative uniformity of sequence identity in intergenic regions downstream of genes (i.e., in terminators), which harbor no promoter elements (Fig. 1B). Genes that

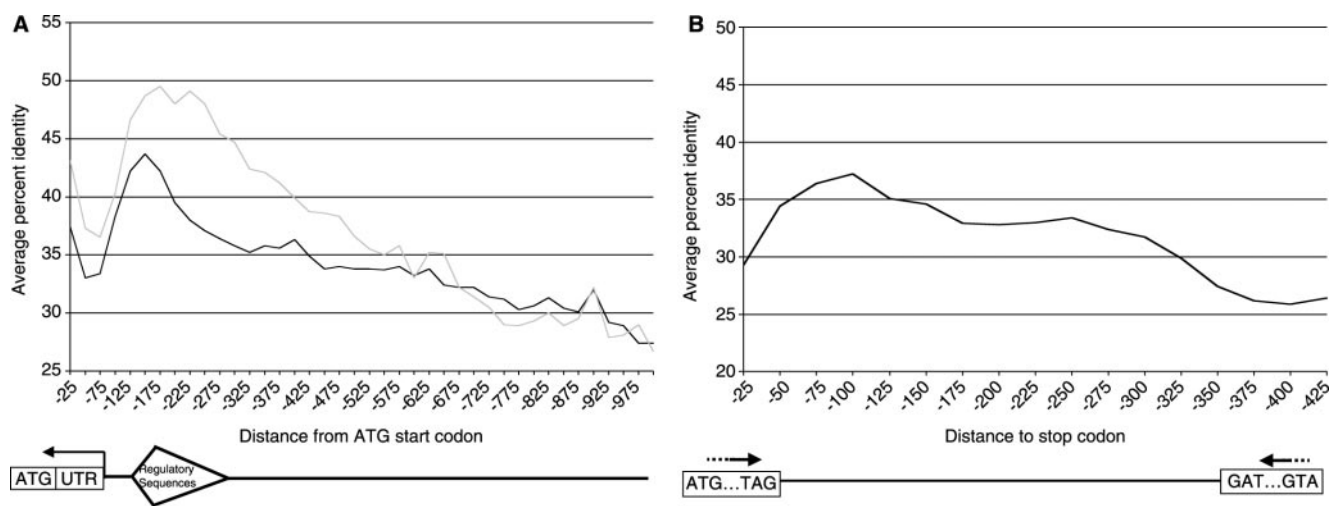
have in their promoter a conserved TATA box (a key promoter sequence element that is the binding site for TATA box-binding protein, the protein around which RNA polymerase II and its many associated proteins assemble) have a broader and higher peak of conservation (Fig. 1A), suggesting that their promoters contain more regulatory sequences than the average promoter, or that they are more slowly evolving than the average promoter. The lower average sequence identity that is 75 to 100 bp upstream of the ATG codon suggests that there may be a spatial restriction on regulatory sequences that prevents them from acting close to the transcriptional start site.

Several of the intergenic regions have highly conserved sequences immediately upstream of the translational start site that are longer and more conserved than the 6- to 10-nucleotide (nt) length expected for transcriptional regulatory elements (26). These sequences are not misannotated coding regions, because they are not encompassed in ORFs. The location of these conserved sequences makes them good candidates for translational regulatory elements. [Indeed, two of them are upstream of genes known to be translationally regulated (27, 28).] Ribosomal protein genes showed the highest degree of sequence identity within 30 bp of the translational start codon (fig. S1) (29).

**Identification of conserved sequence motifs.** We searched for conserved sequence motifs in two overlapping sets of orthologous intergenic sequences: 3523 four-way alignments of intergenic regions of the closely related sensu stricto species, and 3084 six-way sequence comparisons that included sequences of both of the more distantly related species. The advantage of analyzing only the sequences of the closely re-

lated sensu stricto species is that they can be aligned in multiple sequence alignments, which provide a powerful visual tool for identifying evolutionarily conserved sequences. The value of the multiple sequence alignments is proportional to the degree that the regulatory sequence architecture is maintained among the different species, and this is high among the sensu stricto species. In addition, because functional sequences are expected to be in the same position and orientation in the closely related species, unaligned sequence motifs, which could easily occur by chance, are not considered, thereby restricting the amount of sequence to be searched.

The more distantly related species enhance the analysis in primarily two ways (30). First, because many intergenic regions have a high degree of sequence identity over the length of the CLUSTALW sequence alignments of the four sensu stricto species (31), the more distantly related species provide the sequence divergence necessary for conserved sequence motif identification in these intergenic regions. Second, the length of individual conserved sequence motifs in the more distantly related species correlates better with the length of transcription factor binding sites than it does in the sensu stricto species' sequences, where the conserved sequence motifs are often longer than known binding sites. The average length of sequence motifs conserved in the sensu stricto species' sequence alignments is 10.7 nt, with 43% of them being 10 nt or longer; the average length of conserved sequence motifs identified in the six-way sequence comparison is only 7.3 nt, and only 2% of them are 10 nt or longer. Similarly, the 7.5 nt average length of known conserved protein-binding sites identified in the six-way sequence comparisons correlates better to the length of the binding site than does the 11.7 nt average length



**Fig. 1.** Profiles of the average sequence conservation over the length of intergenic regions for different classes of promoters. The average sequence identity in intergenic regions is 37.1% with a range of less than 10% to almost 95% identity, and a median of 36.8%. The percent identity of CLUSTALW alignments of orthologous intergenic sequences of four sensu stricto species was averaged over the length of the

intergenic region in 25-bp windows. (A) Average sequence identity profile of all 3523 four-way alignments (black), and average sequence identity of intergenic regions containing aligned TATA box sequences (gray). (B) Profile for intergenic regions between 595 convergently transcribed genes (that is, intergenic regions consisting entirely of sequences downstream of genes).

found for the same protein-binding sites in the sensu stricto species' sequence alignments (32).

A number of algorithms have been developed to identify similar sequence motifs within sets of DNA sequences (33–36). Because these algorithms were designed for searching unrelated sequences, they tend to identify a large number of conserved sequence motifs in our related sequences, many of which are likely conserved as a result of their shared ancestry. Therefore, we used more stringent criteria that require the sequence motifs to be precisely conserved in all sequences being compared. A limitation of this approach is that some functional sequences will be missed because they need not be exactly conserved through evolution, but searching among a large number of intergenic regions increases the chance that a functional sequence motif will be found precisely conserved. We focused on ungapped motifs because most of the characterized sequence motifs (62 of 71) are ungapped, and many gapped motifs are likely artifacts as a result of simple sequences associated with ungapped motifs (such as simple

A+T-rich sequences associated with an ungapped sequence motif), which makes it statistically difficult to distinguish the likely real gapped motifs. We expected that this relatively straightforward approach for identification of conserved sequence motifs would be successful because of our judicious choice of the number and phylogenetic distances of sequences to be compared. Indeed, we identified 53 of 62 characterized ungapped transcriptional regulatory motifs that are precisely conserved upstream of at least five genes [we also identified many conserved instances of all nine known gapped sequence motifs (37)].

We identified 8873 conserved 6- to 30-oligomers (mers) in the four-way CLUSTALW alignments of orthologous intergenic sequences of the sensu stricto species. This is significantly more (>150 SDs) than the 1090 that were found in the same sequence alignments when their columns had been randomly shuffled (38). Thus, we are about 88% confident that the 6- to 30-mers do not occur by chance. Our confidence in each *n*-mer increases with its size: The 98%

confidence level is reached with 10-mers, because the number of 10-mers in the shuffled alignments is less than 2% of those in the real alignments (fig. S2). Because the shuffled sequence alignments maintain the same degree of conservation, these results suggest that most of the *n*-mers result from functional selection rather than common ancestry. The six-way sequence comparison that includes the more distantly related sequences yielded 7915 conserved sequence motifs (39), with the 98% confidence level reached with 8-mers [see table S3 (40)]. The most statistically significant *n*-mers seem to be biologically significant because many of them are known binding sites for characterized DNA binding proteins (41).

About one-third (2771) of the 6- to 30-mers identified from multiple sequence alignments of sensu stricto species' intergenic regions and about 20% (1535) of the conserved *n*-mers identified in the six-way sequence comparisons contain at least 1 of 71 known sequence motifs (fig. S3). A few known sequence motifs accounted for the majority of the matching *n*-mers: 13 characterized sequence motifs accounted for about three-fourths of the known *n*-mers identified in the alignments of the closely related sequences; 10 characterized sequence motifs accounted for about three-fourths of the known *n*-mers found in the six-way sequence comparisons. Thus, a few sequence motifs seem to be regulating a large number of genes. The TATA box, which is by far the most frequent of the known conserved sequence motifs (fig. S3), is conserved in surprisingly few intergenic regions (42).

**Predicting novel functional sequence motifs.** In an attempt to identify novel functional sequences among the 6- to 30-mers, we first determined if any tend to reside upstream of genes that are functionally related (43). Considering unknown 6 to 30-mers that occur upstream of several genes, 18 *n*-mers identified from the alignments of the closely related sequences, and 18 *n*-mers identified in the six-way sequence comparisons are upstream of genes significantly enriched for those with similar functional annotations (44) (Table 1), and are thus good candidates for functional sequence motifs.

Another way to predict which of the unknown conserved sequence motifs are functional is to identify those that reside upstream of genes that exhibit a similar pattern of expression. This seems to be a valid way to evaluate the functionality of conserved sequence motifs because the expression profiles of the 736 genes in the *S. cerevisiae* genome whose promoters contain an MCB box [a sequence motif that contributes to regulation of gene expression in the G<sub>1</sub> phase of the cell cycle (12, 45, 46)] were essentially random through the cell cycle (Fig. 2A), but the subset of genes whose MCB box is conserved and aligned in the orthologous sequences of all

**Table 1.** Conserved sequence motifs upstream of genes with similar function.

Motif	Function*	<i>P</i> value†
<i>From sensu stricto alignments (18)</i>		
CTAAACGA‡	Lipid, fatty acid, and isoprenoid biosynthesis	<1.0 × 10 <sup>-6</sup>
TTGGAG	Lipid, fatty acid, and isoprenoid utilization	<1.0 × 10 <sup>-6</sup>
ACTCTTT‡	Amino acid metabolism	1 × 10 <sup>-5</sup>
TGGCGC	Amino acid biosynthesis	<1.0 × 10 <sup>-6</sup>
GAAAAAG‡	Amino acid biosynthesis	7 × 10 <sup>-5</sup>
AAAGAAA‡	Amino acid transport	1 × 10 <sup>-5</sup>
TGTGGCG‡	Peptide transporters	<1.0 × 10 <sup>-6</sup>
GTACGGAT‡	Ribosome biogenesis	<1.0 × 10 <sup>-6</sup>
TCTAGA‡	Metabolism of cyclic and unusual nucleotides	<1.0 × 10 <sup>-6</sup>
AAGCCACA	Nitrogen and sulfur utilization	<1.0 × 10 <sup>-6</sup>
ATAGAAA	Fermentation	1 × 10 <sup>-5</sup>
AGATCT‡	Phosphate transport	<1.0 × 10 <sup>-6</sup>
AACGCCG‡	Peroxisome	1 × 10 <sup>-5</sup>
TGTTTTAT	Cytoskeleton-dependent transport	<1.0 × 10 <sup>-6</sup>
CGCCGG	Vacuolar degradation	<1.0 × 10 <sup>-6</sup>
GTGCAC	Homeostasis of metal ions (Na, K, Ca, etc.)	<1.0 × 10 <sup>-6</sup>
TTTTTCCT‡	Chromosome function	8 × 10 <sup>-5</sup>
ACGCCAAA	Centrosome	<1.0 × 10 <sup>-6</sup>
<i>From six-way alignments (18)</i>		
GGAAAA‡	C-compound and carbohydrate utilization	<1.0 × 10 <sup>-6</sup>
TCGTTTA‡	Lipid, fatty acid, and isoprenoid metabolism	<1.0 × 10 <sup>-6</sup>
GGAATT	Amino acid metabolism	<1.0 × 10 <sup>-6</sup>
AGAAAT‡	Ribosome biogenesis	<1.0 × 10 <sup>-6</sup>
CATACA‡	Ribosome biogenesis	<1.0 × 10 <sup>-6</sup>
TACGTA	Translation	<1.0 × 10 <sup>-6</sup>
TTCAAG	Pre-mRNA splicing	<1.0 × 10 <sup>-6</sup>
ATATGT	Ion transporters (Na, K, Ca, NH <sub>4</sub> , etc.)	<1.0 × 10 <sup>-6</sup>
TATTGT	Chromosome function	<1.0 × 10 <sup>-6</sup>
AACAAC	Chromosome function	<1.0 × 10 <sup>-6</sup>
AAGGAA	Chromosome function	<1.0 × 10 <sup>-6</sup>
GAAACA	Pheromone response, mating-type determination	<1.0 × 10 <sup>-6</sup>
AAAACG	Directional cell growth (morphogenesis)	<1.0 × 10 <sup>-6</sup>
AAGGGAA‡	Cell polarity and filament	<1.0 × 10 <sup>-6</sup>
TTGCAA	Peroxisome	<1.0 × 10 <sup>-6</sup>
TGTGGC	Nitrogen and sulfur utilization	<1.0 × 10 <sup>-6</sup>
AGAGAG‡	Nitrogen and sulfur utilization	<1.0 × 10 <sup>-6</sup>
TTTCTT‡	Plasma membrane	<1.0 × 10 <sup>-6</sup>

\*Functional categories from MIPS (<http://mips.gsf.de/proj/yeast/CYGD/db/index.html>). †The probability that this set of genes is enriched in this functional classification by chance (52). ‡In these cases, several overlapping sequence motifs were found as enriched in the same functional category.

RESEARCH ARTICLES

**Table 2.** Unknown sequence motifs upstream of genes with coherent expression.

Sequence motif	Condition(s)*	EC†‡
<i>From sensu stricto alignments (39)</i>		
TCCCTT	MMS	0.393
TCCAGAA	DNA damage	0.476
CAACTTT	DNA damage	0.333
ACGGAT	DNA damage	0.489
GATTGA	DNA damage	0.333
CAGAAC	DNA damage	0.333
TAATAG	DNA damage	0.393
	Stress	0.357
TTTCAGA§	DNA damage	0.352
	Stress	0.619
TGTACGG§	DNA damage	0.6
	Stress	0.4
GCGATGC§	DNA damage	0.303
	Stress	0.53
CAAGGG§	DNA damage	0.333
	Stress	0.321
ACTGAA§	DNA damage	0.393
	Stress	0.321
CGATGCC§	DNA damage	0.81
	Mitochondria	0.429
TGTTCT§	Mitochondria	0.306
	Stress	0.429
CAAACAA	Stress	0.333
ACTCTTT§	Stress	0.321
AACTTTT	Stress	0.333
ATGCGATG	Stress	0.41
AAACAAGA§	Stress	0.381
TAATCT	Stress	0.467
AAAGTA	Stress	0.4
TCCGTA	Stress	0.429
TTTCTAGA	Stress	0.333
ACATTC	Stress	0.381
ATACCT	Stress	0.333
CCCTTAAA	Cell cycle	0.357
AACGCCAA§	Cell cycle	0.533
TTGCCACT§	Cell cycle	0.4
TAAACAAT	Cell cycle	0.333

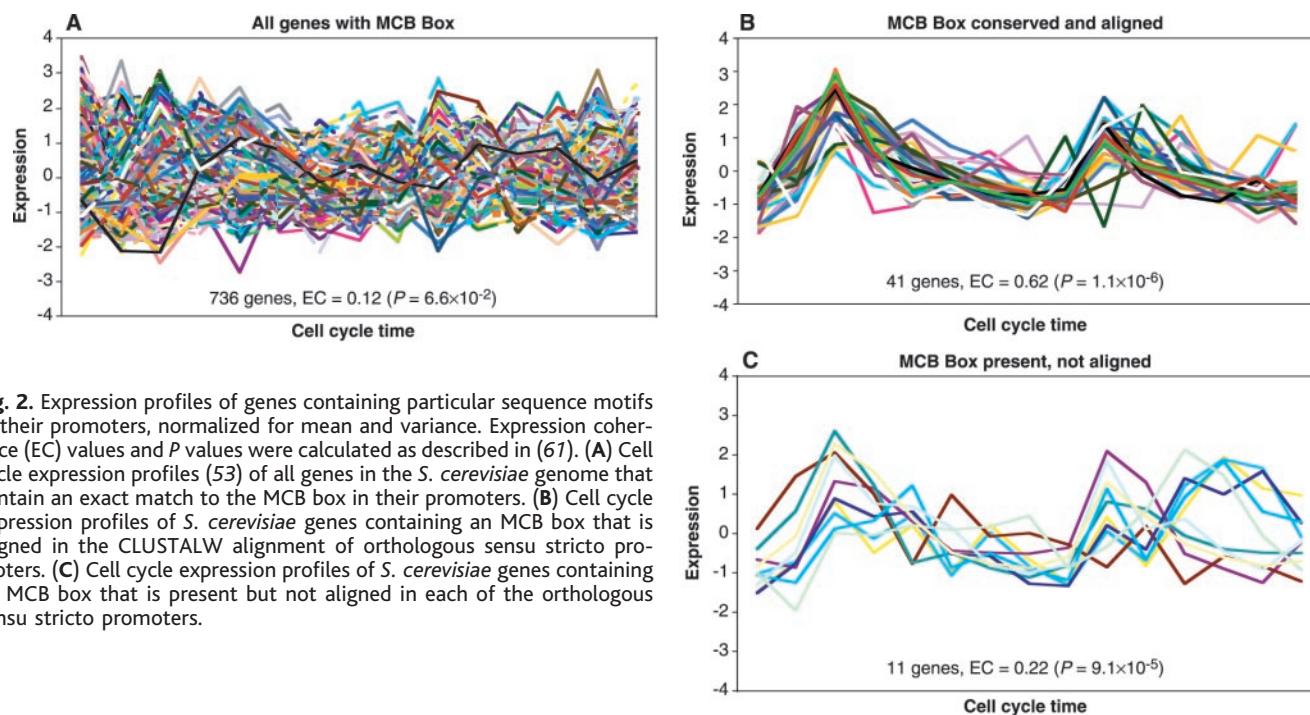
**Table 2.** (Continued).

Sequence motif	Condition(s)*	EC†‡
AAAGAT	Cell cycle	0.321
TATTAG	Cell cycle	0.4
CACCAC	Cell cycle	0.333
TTTTTTGT§	Meiosis	0.451
ACAAAAAC§	Meiosis	0.371
GTTGTTTT§	Meiosis	0.333
ATCAAA	Meiosis	0.357
CGACAC	Meiosis	0.4
TATGTATA§	Pheromone	0.321
GCTACC	Pheromone	0.333
<i>From six-way alignments (13)</i>		
AATGTA§	DNA damage	0.393
AAAAGTA§	DNA damage	0.333
	Stress	0.333
ACATAC	DNA damage	0.357
	Stress	0.6
ATACAT	DNA damage	0.364
	Mitochondria	0.361
	Stress	0.327
TTTTCAT§	Stress	0.35
AGTGAA§	Stress	0.381
CTGAAA§	Stress	0.867
TCAAAT§	Stress	0.333
TTCAAG	Stress	0.372
TAGAAA	Cell cycle	0.4
TTCTTTC§	Cell cycle	0.361
ACAAAA§	Meiosis	0.307
CCCTTT§	Meiosis	0.333

\*The following gene expression profiling data sets were used: cell cycle (53), meiosis (54), MMS damage (55), sporulation (56), stress response (57), DNA damage (58), MAPK (59), mitochondrial dysfunction (60). †Gene expression coherence (0 = no similar expression of the genes, 1 = maximal similarity of expression of the set of genes) was calculated as previously described (54). ‡The probability of each set of genes having the EC score by chance is less than  $10^{-6}$  (61). §In these cases several overlapping sequence motifs were found as enriched in the same functional category.

four sensu stricto species exhibits coherent expression through the cell cycle (Fig. 2B). Even genes in which the MCB box is present but not aligned in the orthologous intergenic sequences of all four species exhibited coherent expression (Fig. 2C), although it is not as obvious as it is in genes that have these motifs aligned. Thirty-nine of the unknown conserved sequence motifs that we identified in sensu stricto species sequence alignments, and 13 of the unknown conserved sequence motifs that we identified in the six-way sequence comparisons that occur in multiple intergenic regions, reside upstream of a set of genes that are significantly enriched for similar gene expression patterns (Table 2) (47).

Potentially functional conserved sequence motifs can also be predicted by identifying those that tend to reside in the intergenic regions to which a particular transcription factor binds. The intergenic regions of the *S. cerevisiae* genome to which 106 known or predicted DNA binding proteins bind have been identified by genome-wide chromatin immunoprecipitation (ChIP) experiments (48). We determined the significance of the overlap between 23 test sets of intergenic regions that contain a conserved occurrence of a known transcription factor binding site with each of the 106 sets of intergenic regions bound by a transcription factor (49). Twenty-one of these 23 sets of intergenic regions overlapped significantly ( $P < 10^{-5}$ ) with the intergenic regions bound by the transcription factor known to bind the site, suggesting that this is a valid approach for predicting functional sequence motifs. Considering unknown conserved sequence motifs that are present in multiple intergenic regions, nine that we identified in sensu stricto species sequence alignments and four that we identified



**Fig. 2.** Expression profiles of genes containing particular sequence motifs in their promoters, normalized for mean and variance. Expression coherence (EC) values and  $P$  values were calculated as described in (61). (A) Cell cycle expression profiles (53) of all genes in the *S. cerevisiae* genome that contain an exact match to the MCB box in their promoters. (B) Cell cycle expression profiles of *S. cerevisiae* genes containing an MCB box that is aligned in the CLUSTALW alignment of orthologous sensu stricto promoters. (C) Cell cycle expression profiles of *S. cerevisiae* genes containing an MCB box that is present but not aligned in each of the orthologous sensu stricto promoters.

in the six-way sequence comparisons significantly overlapped with one of the 106 sets of genes bound by a transcription factor (Table 3). These 13 conserved sequence motifs are candidates for sequences that either bind one of the 106 transcription factors, or are bound by an unknown transcription factor that interacts with one of the 106 known or predicted transcription factors.

In summary, we identified 59 conserved sequence motifs in the four-way sequence alignments and 32 in the six-way sequence comparisons for which there is some evidence of functionality (Tables 1 to 3). Twelve of these were identified in both the four-way and six-way sequence comparisons, leaving 79 unique conserved sequence motifs that are good candidates for functional regulatory sequences (50).

**Conclusions.** We have shown that phylogenetic footprinting on a genome-wide scale identifies many statistically significant conserved sequence motifs. Because we compared multiple genome sequences that are as optimally diverged as possible, we were able to predict functional sequence motifs by relatively straightforward methods using fairly stringent criteria for sequence motif definition (i.e., searching for *n*-mers). The fact that most known regulatory sequences turn up as conserved *n*-mers in our analysis validates this approach for identifying functional sequences and bolsters our confidence that many of the novel sequence motifs we identified are likely to be functional. The novel sequence motifs of which we are most confident are the 79 that lie upstream of sets of genes that tend to have similar functional annotations or similar expression or are bound by the same

transcription factors. Of course, these are only predictions of functional sequences; experimental results will be necessary to validate them. The large number of conserved sequence motif predictions provided by comparative DNA sequence analysis should catalyze development and application of the high-throughput experimental methods necessary for testing their function.

#### References and Notes

- R. C. Hardison, J. Oeltjen, W. Miller, *Genome Res.* **7**, 959 (1997).
- D. A. Tagle *et al.*, *J. Mol. Biol.* **203**, 439 (1988).
- R. Hardison *et al.*, *Gene* **205**, 73 (1997).
- P. F. Cliften *et al.*, *Genome Res.* **11**, 1175 (2001).
- The strains are *S. bayanus* (623-6c), *S. mikatae* (IFO 1815), *S. castellii* (NRRL Y-12630), *S. kluyveri* (NRRL Y-12651), and *S. kudriavzevii* (IFO 1802).
- R. F. Petersen, T. Nilsson-Tillgren, J. Piskur, *Int. J. Syst. Bacteriol.* **49 Pt 4**, 1925 (1999).
- K. Moller *et al.*, *Biotechnol. Bioeng.* **77**, 186 (2002).
- B. Lorente *et al.*, *FEBS Lett.* **487**, 101 (2000).
- G. Fischer, S. A. James, I. N. Roberts, S. G. Oliver, E. J. Louis, *Nature* **405**, 451 (2000).
- Libraries of genomic DNA fragments of the species were constructed in plasmid pOT4. Plating and sequencing of plasmid library subclones as well as sample loading, data collection, and processing were done as described at <http://genome.wustl.edu/tools/protocols/>. The sequences were assembled using the PHRAP (phragment assembly program; [www.phrap.org](http://www.phrap.org/)) (51) using the following parameters: forcelevel 1, minmatch 17, minscore 40, new\_ace, view.
- E. S. Lander, M. S. Waterman, *Genomics* **2**, 231 (1988).
- D. Gordon, C. Desmarais, P. Green, *Genome Res.* **11**, 614 (2001).
- Sequences from the five different species were filtered to remove known protein coding sequences. ORFs 25 to 300 amino acids in length (an initiation codon was not required) were extracted from the sequences, and the encoded peptide sequences were compared with sequences from the different species to identify ORFs that were conserved between species. ORFs whose predicted proteins have significant similarity ( $P < 1 \times 10^{-20}$ ) to translated intergenic regions ("NotFeatures DNA," obtained from the SGD) of *S. cerevisiae* were inspected manually. Finally, multiple DNA sequence alignments of the small ORFs were used to make accurate predictions of start and stop codons.
- Genomic sequences of the five species were compared with annotated genes in *S. cerevisiae* using TBLASTN to identify those that are likely to be false. ORFs were termed false if most ( $\geq 50\%$ ) of the top scoring alignments contain frameshift or stop codons, (for overlapping ORFs, only one species had to have a frameshift or stop codon to be called false). Most of the false ORFs contain frame shifts or multiple stop codons in two or more species.
- S. cerevisiae* genomic sequences and annotations were obtained from the SGD ([www.yeastgenome.org](http://www.yeastgenome.org/); July 24, 2002 version).
- An additional 82 annotated *S. cerevisiae* ORFs do not have significant similarity to sequences in any of the *Saccharomyces* species, and are therefore likely not to be genes, but because we do not have complete genome sequences of the other species we cannot be certain of this.
- All changes to the *S. cerevisiae* genome annotation were submitted to the SGD.
- J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res.* **22**, 4673 (1994).
- S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
- The genome sequences were compared to *S. cerevisiae* proteins (obtained from SGD, [orf.trans.fasta](http://blast.wustl.edu/), July 24, 2002 release) using BLASTX ([http://blast.wustl.edu](http://blast.wustl.edu/)), and the gene boundaries of the top scoring proteins (maximum *P* value of  $1.0 \times 10^{-10}$ ) were recorded. Promoter sequences from the -1 position, relative to the ATG start codon, up to 1000 bp (if available) upstream, or to the next gene, were extracted. [*S. cerevisiae* genes overlap

ping or immediately adjacent to other genomic features (692 of the 6359 total genes listed in SGD) were not included in the analysis.] The orthologous promoter sequences of the sensu stricto species were aligned using CLUSTALW (15). Low-scoring alignments (less than 20% identity) and alignments lacking at least one run of six exact nucleotides were manually inspected. When appropriate, sequences were removed from the alignments to improve the quality (and accuracy) of the alignment. For example, we removed sequences if there were multiple sequences from one species (perhaps duplication of an orthologous gene in the other species) or if there was one sequence not similar enough to align with the rest of the sequences (perhaps a misaligned ortholog). After manually editing the sensu stricto species alignments, sequences from the more distantly related species were added to the files. Because these sequences are too diverged from the other sequences to align accurately we could not use alignments to inform ambiguous orthology calls. Multiple orthologs were retained for a given species unless one of them clearly had a more significant BLASTX hit. In many cases these decisions had to be made manually because of incomplete sequence coverage (that is, partial gene coverage and potential frame shifts or sequencing errors that affect the BLAST score).

Sixty-two percent of intergenic regions (3523) are available from all three sensu stricto species; 86% (4908) are available from two of the three sensu stricto species; 2292 intergenic regions (40% of the total) are available from all six species.

The sequences are available at GenBank (project accession numbers: *S. kluyveri*, AAC000000000; *S. castellii*, AACF000000000; *S. bayanus*, AACG000000000; *S. mikatae*, AAC000000000; *S. kudriavzevii*, AACI000000000). The sequences are also available on the SGD Web site ([www.yeastgenome.org](http://www.yeastgenome.org/)) and at [www.genetics.wustl.edu/saccharomycesgenomes/](http://www.genetics.wustl.edu/saccharomycesgenomes/).

The number of identical residues in multiple alignments of orthologous promoters of the sensu stricto species (after removing terminal gaps in the alignments caused by differences in sequence length) was tabulated for each consecutive 25-bp window from the start of translation backward to the 5' end of the promoter sequence. Essentially identical results were obtained with shorter window lengths.

Information is available at [www.genetics.wustl.edu/saccharomycesgenomes/promoter\\_significance.html](http://www.genetics.wustl.edu/saccharomycesgenomes/promoter_significance.html).

K. Struhl, *Annu. Rev. Genet.* **29**, 651 (1995).

The CLUSTALW alignments of 59 genes were at least 75% identical in the 25 nt upstream of the ATG codon (fig. S2).

J. Vilardell, J. R. Warner, *Mol. Cell Biol.* **17**, 1959 (1997).

A. G. Hinnebusch, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 6442 (1984).

Sixty percent of alignments of orthologous sequences upstream of genes encoding ribosomal proteins, but only 3% of alignments of sequence upstream of other genes have 70% or greater identity over the 30 nt adjacent to the ATG codon.

N. Rajewsky, N. D. Socci, M. Zapotocky, E. D. Siggia, *Genome Res.* **12**, 298 (2002).

Of 3523 intergenic regions, 390 are greater than 50% identical; an additional 379 are greater than 45% identical.

Known transcription factor binding sites included Gcn4, Hap2, Mbf1, Ndt80, Pho4, Reb1, Rpn4, SCB, Ste12, Upc2, Mac1, and Gln3.

J. D. Hughes, P. W. Estep, S. Tavazoie, G. M. Church, *J. Mol. Biol.* **296**, 1205 (2000).

G. Z. Hertz, G. D. Stormo, *Bioinformatics* **15**, 563 (1999).

C. E. Lawrence *et al.*, *Science* **262**, 208 (1993).

T. L. Bailey, C. Elkan, *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3**, 21 (1995).

The nine known gapped sequence motifs are the binding sites for Gal4, Abf1, Lys14, Leu3, Cha4, Put3, Uga3, Hap1, and Ppr1. The characterized motifs that we miss are long and strictly conserved (six of the seven are 8 nt or greater in length), probably because the motifs have not been sufficiently characterized to define their essential positions.

Four-way CLUSTALW promoter alignments with 40% or less identity were used in the analysis (2377 of 3523 alignments). Forty percent was arbitrarily chosen as a cutoff to remove alignments with too much similarity. The CLUSTALW alignments were first modified to re-

**Table 3.** Unknown sequence motifs correlated with ChIP experiments.

Sequence motif	Associated transcription factor	<i>P</i> value*
<i>From sensu stricto alignments</i> (9)		
TGTACGG†	Fhl1	$2.50 \times 10^{-10}$
TGTATGG†	Fhl1	$1.97 \times 10^{-7}$
GT TCTG†	Fhl1	$6.66 \times 10^{-7}$
TAATCT	Fhl1	$4.21 \times 10^{-7}$
TAGCCA	Fhl1	$1.97 \times 10^{-7}$
TTCTAGA	Hsf1	$1.92 \times 10^{-7}$
GCCAAAG	Smp1	$4.49 \times 10^{-7}$
GGACCC	Smp1	$1.10 \times 10^{-9}$
ATTATCA	Smp1	$2.99 \times 10^{-7}$
<i>From six-way alignments</i> (4)		
TTGAAA†	Fhl1	$1.15 \times 10^{-10}$
ACATAC†	Fhl1	$1.70 \times 10^{-7}$
GTTTAT	Hir1	$3.18 \times 10^{-6}$
	Hir2	$1.74 \times 10^{-7}$
TCTTTC	Sfp1	$3.47 \times 10^{-6}$

\*The probability that the set of intergenic regions with a conserved sequence motif overlaps one of the 106 sets of intergenic regions bound by a transcription factor by chance was calculated by the hypergeometric probability distribution, as described in (54). †In these cases, several overlapping sequence motifs were found as enriched in the same functional category.

- move terminal unaligned sequence from the output. The alignments were shuffled 10,000 independent times using the shuffle utility program in SQUID ([www.genetics.wustl.edu/eddy/software/#squid](http://www.genetics.wustl.edu/eddy/software/#squid)). Runs of identical sequence aligned in all four species in the real alignments and the shuffled alignments were extracted and counted.
39. Six-mers are not statistically significant in the six-way comparisons and were therefore not tabulated; *n*-mers longer than 10 nt are quite rare in these comparisons.
  40. Sequences of 100 shuffled multiple intergenic sequence alignments of sensu stricto species were extracted and combined with intergenic sequences from the two distantly related species. *n*-oligomers present in all species were identified in the real promoter sets and compared with those present in the shuffled data sets.
  41. For example, essentially all of the 10-mers conserved in the sensu stricto species' sequences (considered because there is a high degree of confidence that they are not chance occurrences) that occur frequently in the genome (considered because those are likely to be functional) are known: 28 of 32 (present in alignments of sequences upstream of at least seven different genes) are accounted for by the following previously identified sequence motifs: eight variations of the PAC motif, seven variations of the RRPE motif, seven variations of a Ume6 binding site, two variations of the PACE motif (Rpn4 binding site), two variations of an Mbp1 binding site, and one each of the binding sites for Ndt80 and Reb1; the remaining four most frequent *n*-mers were simple A+T-rich sequences. Similarly, 94 of the 160 conserved 10-mers identified in the six-way sequence comparisons correspond to known sequence motifs (table S4), all of which occur upstream of genes known or likely to be regulated by the factor that binds to them. Of the remaining 66 conserved sequence motifs, 46 are A+T-rich, and 11 are immediately upstream of the translation initiation codon of genes encoding ribosomal proteins and thus may be translational regulatory sequences. This leaves only nine that are reasonable candidates for new regulatory sequences, four of which are conserved in the sequences upstream of several genes of similar function (the motifs marked with an asterisk in table S4).
  42. We were surprised to find that only 15.7% of the 3523 multiple alignments of sensu stricto species promoters contain one of the seven sequences that have TATA element function (TATAA, TATATA, TATTTA, CATTTA, TTTAAT, TAATAA, TATAA) (52) conserved and aligned within 250 bp of the translational start codon. Even if the stringency of the search criteria is relaxed to allow for unaligned TATA boxes, promoters containing this sequence element are still in the minority: Only 42.8% of the sensu stricto species' promoters contain any one of the seven TATA sequences (52) anywhere within 250 bp of the translational start codon in all four orthologs. Furthermore, 142 promoters (4%) do not contain a TATA element in any of the four sensu stricto species. Thus, it appears that TATA-containing promoters are the minority in *S. cerevisiae*.
  43. Because many of conserved *n*-mers are longer than the typical 6 to 8 bp that are required for a transcription factor to bind to DNA, we extracted all unique 6- to 8-mers from the longer conserved *n*-mers to test these for functional enrichment and coherent expression. Each unique *n*-mer had to be present in at least five different intergenic regions to test for the functional enrichment or coherent expression.
  44. Functional enrichment was based on the Munich Information Center for Protein Sequences (MIPS) classification of *S. cerevisiae* genes. Functional enrichment and the associated *P* values were calculated as in (53).
  45. N. F. Lowndes, A. L. Johnson, L. H. Johnston, *Nature* **350**, 247 (1991).
  46. E. M. McIntosh, T. Atkinson, R. K. Storms, M. Smith, *Mol. Cell Biol.* **11**, 329 (1991).
  47. Expression coherence was calculated as described previously (54). Expression coherence was calculated for cell cycle (55), meiosis (56), methyl methanesulfonate (MMS) damage (57), sporulation (58), stress response (59), DNA damage (60), mitogen-activated protein kinase (MAPK) (61), and mitochondrial dysfunction (62) data sets.
  48. T. I. Lee *et al.*, *Science* **298**, 799 (2002).
  49. Using high-quality weight matrices for the binding sites of 23 transcription factors whose consensus binding sites are known, we identified: (i) all the occurrences of a particular binding site in the intergenic regions of *S. cerevisiae* using Patser (34), and then (ii) those occurrences that are conserved in the orthologous promoters in the other *Saccharomyces* species and/or are aligned in the CLUSTALW alignments of intergenic sequences of sensu stricto species. Sets of intergenic regions that bind to a particular DNA binding protein come from the data of Lee *et al.* at *P* value < 0.001 (48). The motif alignments for known transcription factor binding sites were generated by applying AlignACE (33) on the appropriate MIPS functional category.
  50. Thirty-six *n*-mers are upstream of genes that are functionally enriched [18 from the sensu stricto sequence alignments and 18 from the six-way sequence comparisons (Table 1)], 52 *n*-mers are identified by coherent expression [39 from the sensu stricto sequence alignments and 13 from the six-way sequence comparisons (Table 2)], and 13 are from upstream of genes bound by a particular transcription factor [nine from the sensu stricto sequence alignments and four from the six-way sequence comparisons (Table 3)]. Twenty-two *n*-mers are found in more than one data set, leaving 79 conserved sequence motifs linked to a function.
  51. P. Green, PHRAP, Department of Genome Sciences, University of Washington, Seattle, WA.
  52. V. L. Singer, C. R. Wobbe, K. Struhl, *Genes Dev.* **4**, 636 (1990).
  53. S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, G. M. Church, *Nature Genet.* **22**, 281 (1999).
  54. Y. Pilpel, P. Sudarsanam, G. M. Church, *Nature Genet.* **29**, 153 (2001).
  55. R. J. Cho *et al.*, *Mol. Cell* **2**, 65 (1998).
  56. M. Primig *et al.*, *Nature Genet.* **26**, 415 (2000).
  57. S. A. Jelinsky, P. Estep, G. M. Church, L. D. Samson, *Mol. Cell Biol.* **20**, 8157 (2000).
  58. S. Chu *et al.*, *Science* **282**, 699 (1998).
  59. A. P. Gasch *et al.*, *Mol. Biol. Cell* **11**, 4241 (2000).
  60. A. P. Gasch *et al.*, *Mol. Biol. Cell* **12**, 2987 (2001).
  61. C. J. Roberts *et al.*, *Science* **287**, 873 (2000).
  62. C. B. Epstein *et al.*, *Mol. Biol. Cell* **12**, 297 (2001).
  63. We thank E. Louis (University of Leicester) for invaluable advice on the *Saccharomyces* phylogeny, and for providing yeast strains; our Washington University colleagues M. Brent, J. Buhler, S. Eddy and members of his lab, S.-W. Ho, and G. Stormo, as well as E. Siggia (Rockefeller University) and R. Young (MIT) for advice and insightful comments on the manuscript. This project was funded by a grant from NIH (RO1 GM63803).

## Supporting Online Material

[www.sciencemag.org/cgi/content/full/1084337/DC1](http://www.sciencemag.org/cgi/content/full/1084337/DC1)

Figs. S1 to S3

Tables S1 to S4

10 March 2003; accepted 21 May 2003

Published online 29 May 2003;

10.1126/science.1084337

Include this information when citing this paper.

# REPORTS

## Catalytic Reduction of Dinitrogen to Ammonia at a Single Molybdenum Center

Dmitry V. Yandulov and Richard R. Schrock\*

Dinitrogen ( $N_2$ ) was reduced to ammonia at room temperature and 1 atmosphere with molybdenum catalysts that contain tetradentate  $[HIPTN_3N]^{3-}$  triamidoamine ligands {such as  $[HIPTN_3N]Mo(N_2)$ , where  $[HIPTN_3N]^{3-}$  is  $[[3,5-(2,4,6-i-Pr_3C_6H_2)_2C_6H_3NCH_2CH_2]_3N]^{3-}$ } in heptane. Slow addition of the proton source  $[[2,6-lutidinium]\{BAR'_4\}]$ , where  $Ar'$  is  $3,5-(CF_3)_2C_6H_3$  and reductant (decamethyl chromocene) was critical for achieving high efficiency (~66% in four turnovers). Numerous x-ray studies, along with isolation and characterization of six proposed intermediates in the catalytic reaction under noncatalytic conditions, suggest that  $N_2$  was reduced at a sterically protected, single molybdenum center that cycled from Mo(III) through Mo(VI) states.

The reduction of dinitrogen ( $N_2$ ) to ammonia ( $NH_3$ ) by various nitrogenase enzymes is one of the most fascinating transition metal-

catalyzed reactions in biology (1-12). Six electrons and six protons produce two equivalents of  $NH_3$  per  $N_2$  in discrete steps at 1 atm

of ambient pressure and mild temperatures, with the aid of one or more transition metal centers (Fe, Mo, or V) within those nitrogenases. Although nitrogenases have been studied for decades (primarily the Fe/Mo nitrogenase), it is still not known today how they accomplish this feat.

With the discovery of the first  $N_2$  complex of a transition metal in 1965 (13) came the hope that many  $N_2$  complexes could be prepared and that an abiological catalytic reduction of  $N_2$  at ambient pressure and temperature with protons and electrons at a well-defined transition-metal site would be forthcoming (14-22). Hundreds of  $N_2$  complexes are now known, but only a few reports of the catalytic reduction of  $N_2$  to  $NH_3$  have appeared (18, 23-27). No reduction of  $N_2$  has been accomplished with a relatively mild reducing agent, and no system has revealed many details of the  $N_2$  reduction steps. The