J Mol Evol (1987) 24:252–259

Journal of
Molecular Evolution

# Molecular Evolution of the *Saccharomyces cerevisiae* Histone Gene Loci

M. Mitchell Smith

Department of Microbiology, Box 441 Jordan Building, School of Medicine, University of Virginia, Charlottesville, Virginia 22908, USA

**Summary.** The core histone genes of *Saccharomyces cerevisiae* are arranged as duplicate nonallelic sets of specifically paired genes. The identity of structural organization between the duplicated gene pairs would have its simplest evolutionary origin in the duplication of a complete locus in a single event. In such a case, the time since the duplication of one of the genes should be identical to that since duplication of the gene adjacent to it on the chromosome. A calculation of the evolutionary distances between the coding DNA sequences of the histone genes leads to a duplication paradox: The extents of sequence divergence in the silent component of third-base positions for adjacent pairs of genes are not identical. Estimates of the evolutionary distance between the two H3–H4 noncoding intergene DNA sequences are large; the divergence between the two separate sequences is indistinguishable from the divergence between either of the regions and a randomly generated permutation of itself. These results suggest that the duplication event may have occurred much earlier than previously estimated. The potential age of the duplication, and the attractive simplicity of the duplication of both the H3–H4 and the H2A–H2B gene pairs having taken place in a single event, leads to the hypothesis that modern haploid *S. cerevisiae* may have evolved by diploidization or fusion of two ancient fungi.

**Key words:** Histone genes — Gene conversion — Diploidization — Yeast

*Offprint requests to:* M. M. Smith

## Introduction

The histones provide a useful and important gene family in which to study the possible principles and mechanisms of eukaryotic molecular evolution. The extreme conservation of the protein sequences over time potentially permits the analysis of early events in evolutionary history. Additionally, the histone genes were among the first eukaryotic genes to be molecularly cloned by recombinant DNA techniques. Thus the structural organizations and DNA sequences of the genes have been determined in a wide variety of organisms (Kedes 1979; Hentschel and Birnstiel 1981; Maxson et al. 1983).

The haploid genome of the yeast *Saccharomyces cerevisiae* contains two copies of each of the genes for the core histones. The genes are arranged as two nonallelic sets of H2A–H2B gene pairs and two nonallelic sets of H3–H4 gene pairs. In each pair, the genes are divergently transcribed and are separated by 600–800 nucleotides of intergene DNA sequence. Each pair of genes is on a separate genetic linkage group. All eight genes are transcribed and translated into histones (Hereford et al. 1979; Smith and Murray 1983; Smith 1984). The duplicated gene pairs are functionally redundant in the laboratory: Any single pair of genes may be deleted without loss of cell viability, although the growth rates of the strains may be affected under certain conditions (Grunstein et al. 1984; M. Smith and V. Stirling, unpublished). This dispersed duplicated arrangement poses questions regarding the origin and evolution of the loci. This report presents the results of an examination

**Table 1.** Description of DNA sequences

| Gene ID[a] | Sequence type | Yeast strain | Accession number | Locus name | Base positions[b] |
|---|---|---|---|---|---|
| HTA1 | Histone H2A | *S. cervisiae* | J01325 | YSCH2A1 | 264–656 |
| HTB1 | Histone H2B | *S. cervisiae* | J01327 | YSCH2B1 | 203–592 |
| HTA2 | Histone H2A | *S. cervisiae* | J01326 | YSCH2A2 | 264–656 |
| HTB2 | Histone H2B | *S. cervisiae* | J01328 | YSCH2B2 | 185–574 |
| HHT1 | Histone H3 | *S. cervisiae* | X00724 | YSCH34CI | 612–208 (c) |
| HHF1 | Histone H4 | *S. cervisiae* | X00724 | YSCH34CI | 1265–1570 |
| HHT2 | Histone H3 | *S. cervisiae* | X00725 | YSCH34CII | 744–340 (c) |
| HHF2 | Histone H4 | *S. cervisiae* | X00725 | YSCH34CII | 1427–1732 |
| IG1 | H3–H4 copy-I intergene DNA | *S. cervisiae* | X00724 | YSCH34CI | 726–1153 |
| IG2 | H3–H4 copy-II intergene DNA | *S. cervisiae* | X00725 | YSCH34CII | 864–1294 |
| GALIG | GAL1–GAL10 intergene DNA | *S. cervisiae* | K02115 | YSCGAL | 242–673 |
| ScCYC | Iso-1-cytochrome c | *S. cervisiae* | J01319 | YSCCYC1 | 252–575 |
| SpCYC | Cytochrome c | *S. pombe* | J01318 | YSPCYC | 448–771 |
| ScADH | Alcohol dehydrogenase | *S. cervisiae* | J01313 | YSCADHI | 754–1794 |
| SpADH | Alcohol dehydrogenase | *S. pombe* | J01341 | YSPADH | 277–1329 |

[a] Identifications of the DNA sequence entries used from release 40.0 of the GenBank genetic sequence data bank

[b] A "(c)" indicates that the sequence was complementary to the GenBank entry

of the time of duplication of the core histone gene loci. Analysis of the gene family organization, the protein sequences of the duplicated pairs, and the DNA sequences of the genes leads to some surprising observations.

## Sequences and Data Analysis

Release 40.0 of the GenBank sequence library (Bolt Beranek and Newman Inc, Cambridge, MA) was used as the source of the DNA sequences in this study, except for the *Schizosaccharomyces pombe* histone gene sequences, which were taken directly from Matsumoto and Yanagida (1985). The sequences studied are summarized in Table 1. Random sequences were generated from the histone intergene DNA by a simple algorithm in which bases were selected at random, using a random-number-generator function call, until the count of each nucleotide was the same as in the original sequence.

Evolutionary distances were calculated using the three-parameter divergence matrix as described by Kimura (1980, 1981). Coding region DNA sequences were aligned by similarities in the amino acid sequence translations; the program of Wilbur and Lipman (1983) was used to help position minor gap insertions in the DNA. Alignment of the histone H3–H4 intergene DNA sequences was more difficult because of the extensive divergence; in general the sequences were aligned at their respective "TATA" regulatory sequences and a variety of methods were used to introduce gap insertions to improve homology. Insertion/deletion gaps were assigned with computer assistance using combinations of dot matrix alignment (Smith and Andresson 1983); the algorithms of Korn et al. (1977), Goad and Kanehisa (1982), and Wilbur and Lipman (1983); and finally manual inspection.

## Results

The basic structural organization of the core histone genes in the *S. cerevisiae* haploid genome is summarized in Fig. 1. The striking feature of this arrangement is the symmetry of the duplication; that is, two distinct unlinked but functionally related gene pairs, and not simply a single gene, have been duplicated. It would be particularly interesting to understand the origins and evolution of this gene arrangement. As a first step toward this end it is reasonable to ask how long ago the duplication of the genes took place within the yeast genome.

### *The Duplication Paradox*

A straightforward calculation of the time of duplication of the histone gene pairs by the alignment of the coding sequences results in the surprising observation that the paired genes appear to have been duplicated at quite different times. This discrepancy has been noted earlier based on a comparison of amino acid sequences (Choe et al. 1982; Smith 1984). The consequences of this point will be further elucidated here and examined in greater detail at the DNA sequence level.

Silent, or synonymous, nucleotide substitutions at the third position of amino acid codons have been shown to occur at a rate that is independent of the rate of amino acid substitution in a variety of genes, and particularly among different genes adjacent to each other on the same DNA sequence (Miyata et al. 1980; Kimura 1981). For example, although amino acid substitutions have occurred at one of the lowest known rates in the H4 histone protein, the rate of base substitution at the third position of H4 codons is high, comparable with those determined for a variety of other genes (Kimura 1977; Graur 1985).

Table 2 presents the calculated values of the evolutionary distance $K_s'$ estimated from the silent component of all third-position substitutions for each of the duplicated histone genes. Surprisingly, the H3 genes appear to have been diverging almost six times longer than the H4 genes. Similarly, the H2B genes are about twice as divergent as the H2A genes and 10 times as divergent as the H4 genes.

The structural similarity of the related gene pairs would seem to have its simplest origin in the duplication of an original complete locus in a single event. That is, it seems likely that the H3 and H4 genes were duplicated together rather than in two separate events. The latter scenario would require first the duplication of one gene of the pair, the H3 gene for example, and then at a later time the duplication of the other (the H4 gene). In addition, the second event would have to have produced exactly the original gene configuration at the duplication locus. Thus, the duplication paradox is that genes that logically should have been duplicated at the same time appear to have been duplicated independently.

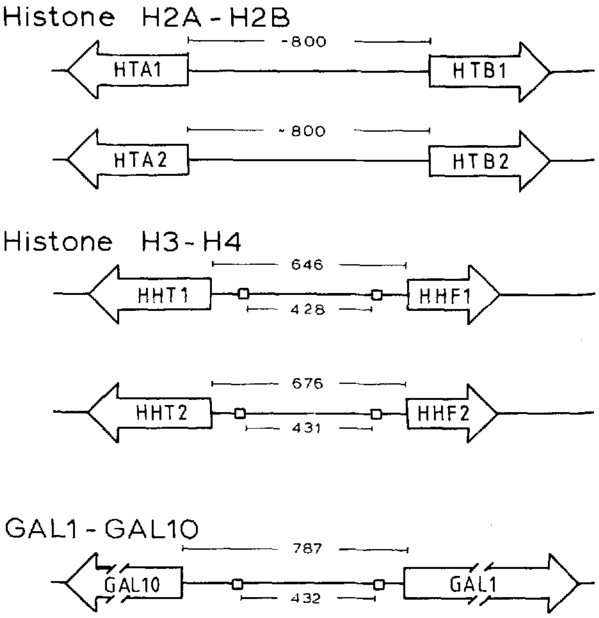

**Fig. 1.** Genomic organizations of the core histone genes and the GAL1 and GAL10 genes in *Saccharomyces cerevisiae*. The arrows encompass the translational start and stop codons for each gene, and their directions show the polarity of transcription. The arrows for GAL1 and GAL10 coding sequences are hatched to show that they have been truncated and are not drawn to scale. The small open boxes in the histone H3–H4 and GAL1–GAL10 intergene regions represent the "TATA" regulatory sites. The distances between the ATG start codons, and between the TATA box signals, are indicated in base pairs. The intergene lengths for the H2A–H2B pairs have been estimated from mapping studies since the complete sequences are not known. Gene designations are as in Table 1

### *The Paradox is Not the Result of Codon Bias*

The *S. cerevisiae* histone genes, in common with many other yeast genes, use a small subset of possible amino acid codons (Bennetzen and Hall 1982a; Smith and Andrésson 1983). A strong bias in preferred codon usage by a gene may introduce errors into the comparative distance calculations (Miyata et al. 1980; Miyata and Hayashida 1981). If this codon bias has a functional significance, a third-base substitution in glycine codons, for example, may not be neutral, since the H3 and H4 genes use GGT exclusively as the codon for glycine (Smith and Andrésson 1983). Thus, the apparent differences in the rates of divergence of the histone genes could be due to different proportions of extremely biased codons.

The data for yeast codon usage suggest that substitutions in the third base position of the codons for alanine, serine, threonine, valine, and isoleucine should be neutral in *S. cerevisiae* histone genes. Thus, a neutral-mutation rate was calculated for the histone genes considering only these five codons. Table 2 shows the distance calculations for this limited biased-codon analysis ($K_b$). These distances are

**Table 2.** Estimates of evolutionary distance

| Comparison | K | $K_s'$ | $K_b$ |
|---|---|---|---|
| HTA1–HTA2 | – | 0.18 ± 0.05 | 0.21 ± 0.07 |
| HTB1–HTB2 | – | 0.41 ± 0.09 | 0.65 ± 0.20 |
| HHT1–HHT2 | – | 0.23 ± 0.05 | 0.45 ± 0.14 |
| HHF1–HHF2 | – | 0.04 ± 0.02 | 0.06 ± 0.05 |
| IG1–IG2 (no gaps) | 2.41 ± 0.54 | – | – |
| IG1–IG1 shuffle (no gaps)[a] | 1.87 ± 0.27 | – | – |
| IG1–IG2 (15% gaps) | 1.01 ± 0.15 | – | – |
| IG1–IG1 shuffle (15% gaps)[a] | 0.86 ± 0.09 | – | – |
| IG1–GALIG (15% gaps) | 1.42 ± 0.29 | – | – |
| ScHistone–SpHistone[b] | – | 0.73 ± 0.05 | – |
| ScCYC–SpCYC | – | 0.90 ± 0.24 | – |
| ScADH–SpADH | – | 0.99 ± 0.17 | – |

[a] IG1 shuffle sequences were generated by randomly shuffling the bases in the IG1 sequence as described in Sequences and Data Analysis section

[b] Comparison of the *S. cerevisiae* and *S. pombe* histone gene protein coding sequenes. The K value presented is the weighted average of all the pairwise comparisons of the homologous gene copies

```
         .       20.         .       40.         .       60.         .       80.         .
TTATATATGTATTTCA--ACAAGGAGCCGTG-GAGGGTAC-CAAAAGAAA---ATCGCCCGGG-TATTTCGTTTATCTTTTACGCTAAAA
|||||||    |      ||||| |||   | ||| |  | ||| |       ||| | |    ||| |  |    |     |
TTATATACAACTGCGGTCACAAGAAGCAACGCGAGAGAGCACAACACGCTGTTATCACGCAAACTATGTTTTGACACCGAGCCATAGCCG

      100.         .      120.         .      140.         .      160.         .      180.
---------GTCAAGGAGA--GATATTACGGCCAAGATCGCAAAGGTGCAGAGCAAGGAAATG-TGAGA------AATTGTGAGA-ACGA
         ||||    |   |||| |     | | ||  | ||    ||     ||||     | ||       ||| ||   |  ||
TGATTGTGCGTCACATTGGGCGATAATGAACGCTAAATGACCAACTCCCATCCGTAGGACGCCCTTAGGCGTGCCAATAGTTTCACGCGC

         .      200.         .      220.         .      240.         .      260.         .
-TAATGT--ATGGGAC--AATGCGAAAATGTGAGAACGAG--AGC--AAAAATCTTTTTTGTATCTCCCCGCCGAATTTGGAAACCG---
 |||||   |  |  |  || |   || ||||       |  | |   ||| |  |  | |           || ||||| |   |
TTAATGCGAAGTGCTCGGAACGGACAACTGTGGTCGTTTGGCACCGGGAAAGTGGTACTAGACCGAGAGTTTCGCATTTGTATGGCAGGA

      280.         .      300.         .      320.         .      340.         .      360.
CGTTCTGAAAACTTCGCATCTTCACATAGTAAAACCGTTCCGAGCGCTTCTCCCCATAATGGTTAGTGGTAAAAACCGAAGTTGTTTACT
|||||||  | |||||| |||  |  |  |     ||   |||           ||| |   |         |||||    |      |
CGTTCTGGGAGCTTCGCGTCTCAAGCTTTTTCGGGCG---CGAAATGCAGA--CCAGACCAGAAC------AAAACAACTGACAAGAAGG

         .      380.         .      400.         .      420.         .      440.         .
TTAGCAAATGCCCGCGAATACGGTGGTTAAATTGCCACCCCCCCTTCCCCATTCATTGGGTAAAGACCAATTTGATGGATAAATTGGTTG
      |||       |||| ||  ||| | | ||       |||   |  ||  ||               | |||  |||   || |||
CGTTTTAATT-----TAATATGTT-GTTCACTCGCG------CCTGGGCTGTTG-TTATTCGG-------CTAGATACATAC-GTGTTTG

      460.         .      480.
TGGAAAAGGTCTAATTCTTTTTCCTATAAATA
||       | |     | | || |||| |
TGCG-----TATGTAGTTATATCATATATAA
```

Fig. 2. Sample alignment of histone H3–H4 intergene sequences. The sequences of IG1 (top) and IG2 (bottom) are compared. The TATA sites of each sequence were aligned and are shaded at each end of the comparison. The internal alignment was keyed by the upstream-activator-site homology between the two sequences. This region, also shaded, is centered around position 280. The dashes represent gaps inserted into each sequence during the analysis to improve nucleotide homology

slightly larger than the $K_s'$ values, particularly for the more divergent genes H2B and H3. Nevertheless, the $K_b$ values also show a significant disparity in the apparent divergence of the different core histone genes—even for genes adjacently paired on the chromosome.

## Intergene Sequence Divergence

A comparison of the intergene sequences can also provide a measure of the distance separating the duplicate gene sets. At present this analysis is restricted to the H3 and H4 gene sets, for which the complete DNA sequences of the intergene regions are available for both loci (Smith and Andrésson 1983). Divergence of the intergene regions is not constrained by the amino acid codons required for protein function. Furthermore, without the overall DNA sequence homology imposed by the coding requirements of the genes, recombination events between the intergene regions are less likely. On the other hand, the alignment of the intergene sequences is more difficult, since they lack a translational reading frame with which to establish a proper registration. Several comparisons have been made to try to compensate for this uncertainty in alignment. The main point to be drawn from the results of the calculations described below is that the H3–H4 intergenic DNA sequences are considerably more divergent than the silent component of third-base changes observed between the coding regions.

The comparisons rests on the following assumptions. First, it is assumed that the intergene regions of the duplicated pairs are derived directly from a common ancestor. The primary support for this assumption is the common gene arrangement of the pairs, as argued in the preceding sections, and the presumptive conservation of upstream activator sequences in the intergene DNA, a point to be illustrated shortly. Second, it is assumed that the alignment of the two intergenic regions may be limited by the locations of the H3 and H4 transcribed sequences. This solves one of the more difficult problems of aligning "spacer" DNA sequences: where to begin and end the comparison. Indeed, the sequence lengths between the H3 and H4 gene TATA box positions are nearly identical between the two copies (Fig. 1). Therefore, in the alignments that follow, the nucleotides between, but not including, the TATA sequences were compared. The functional significance of the conserved length is unknown; however, it is supportive of the evolutionary comparison.

When the intergene DNA sequences located between the TATA box positions were aligned without introducing insertion/deletion gaps, a maximum of 24% nucleotide homology was obtained. The evolutionary distance for this alignment was calculated as K = 2.41 ± 0.54 (Table 2). To obtain a more liberal estimate of the intergene DNA divergence, insertion/deletion gaps were introduced into the two sequences to increase the homology. An example of an extreme alignment is illustrated in Fig. 2 and was constructed from the following considerations. An upstream activation site (UAS), required for expression of the copy-I genes, has been identified in the copy-I intergene DNA by deletion analysis (M.M. Smith and L.R. Karns, unpublished). The copy-II intergene DNA contains a sequence homologous to this UAS, although its function has not yet been tested genetically. For the alignment presented in Fig. 2, the TATA box signals for the genes were aligned, as well as the two presumptive UAS regions (positions 271–291). Insertion gaps were then introduced to produce an alignment with improved homology. The proportion of homologous nucleotides for this alignment is about 42% and the distance between the regions was calculated as K = 0.94 ± 0.09 substitutions per position, ignoring any contribution of the insertion/deletion gaps (Table 2).

The calculated divergence between the H3–H4 intergene sequences will obviously depend on the extent of gap insertion used in the alignment. Sequence gaps, in effect, remove a portion of a sequence from an alignment comparison. As more gaps are introduced into two sequences, the fraction of the total nucleotides compared decreases and the apparent homology between the remaining aligned nucleotides increases. Figure 3 illustrates the effect on the calculated evolutionary distance of removing portions of the sequences from the comparison. When approximately 30% of the matches are removed by introduction of insertion/deletion gaps, the evolutionary distance calculated for the remaining bases is reduced to about 0.50 substitutions per position.

### *Control Sequence Comparisons*

The initial introduction of gaps into related DNA sequences can potentially improve the chances of detecting a relevant evolutionary alignment. Continued gap insertion, however, will degrade the appropriate alignment at the expense of increased individual matches. At that point, the evolutionary distance between related sequences should approach that calculated for totally unrelated sequences. Two types of comparisons have been made to help place limits on gap insertion. First, the H3–H4 intergene sequences were compared with randomized permutations of their own sequences; second, the H3–H4 intergene sequences were aligned with a totally unrelated *S. cerevisiae* sequence.

Randomized H3–H4 intergene sequences were constructed by randomly shuffling the bases in the sequence. The generated sequences thus had the same base composition as the original intergene sequence but a changed base order. Figure 3 illustrates a comparison of the copy-I H3–H4 intergene sequence with such a random permutation of itself. The pattern of "evolutionary distance" for this control comparison is indistinguishable from that for the comparison of two authentic histone H3–H4 intergene sequences.

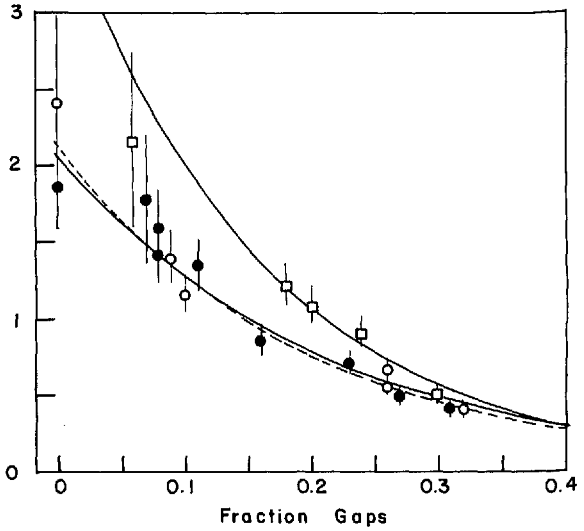

**Fig. 3.** Effect of insertion/deletion gaps on evolutionary distance. The estimate of evolutionary distance, K, is plotted versus the fraction of gaps inserted into the test sequences. The histone IG1–IG2 comparisons are given by the open-circle data points and the lower solid curve. The histone IG1–random–shuffle comparison is given by the closed-circle data points and the dashed curve. The histone IG1–GALIG comparisons are given by the open-box data points and the upper solid curve. The sequences were aligned using the algorithm of Wilbur and Lipman (1983) and varying the gap penalty parameter; for the analysis presented the k-tuple and window parameters were 3 and 20, respectively. The apparent evolutionary distance for each alignment was then determined (Kimura 1980, 1981). The gap fraction was calculated by dividing the total number of gap insertions in both sequences by the length of sequence IG1 (428 nucleotides). The curves are first-order exponentials fit by the nonlinear least-squares program of Johnson and Frasier (1985)

As a second control, the H3–H4 intergene sequences were compared with an unrelated *S. cerevisiae* intergene sequence located between the GAL10 and GAL1 genes. Like the histone intergene regions, the GAL1–GAL10 intergene DNA serves as a bidirectional promoter and control region for a divergently transcribed pair of genes. The length of the region is almost identical to that of the histone H3–H4 regions: There are approximately 430 base pairs between the major TATA signals for the GAL1 and GAL10 genes (Fig. 1). The region's distribution of base composition is similar to that of the H3–H4 intergene regions and it has a central activator sequence controlling GAL1–GAL10 expression (St. John and Davis 1981; St. John et al. 1981; Guarente et al. 1982; Johnston and Davis 1984). Therefore, the structure and function of the GAL1–GAL10 intergene region are similar to those of the histone H3–H4 regions. However, there should be virtually no evolutionary link between the DNA sequences of the H3–H4 and the GAL1–GAL10 intergene regions.

Figure 3 presents the results of calculations comparing the copy-I H3–H4 and the GAL1–GAL10 intergenic sequences. At modest frequencies of gap insertion the two histone gene sequences are more closely related to each other than the copy-I H3–H4 sequence is to the GAL1–GAL10 region. However, as the fraction of gapped nucleotides begins to exceed about 20% the two comparisons become indistinguishable. If the evolutionary relationship between the histone sequences is preserved at 15–20% gap insertion, then the divergence between the two is about 0.7–1.0 substitutions per position.

These must be especially low estimates of the evolutionary distance between the H3–H4 intergene sequences, considering the uncertainties of the alignments, the exemption of gap penalties from the distance calculations, and the loss of resolution obtained as K approaches a value of 1.0 (Kimura 1980, 1981). In addition, the histone intergene sequences are certain to contain regulatory sequences under functional evolutionary constraints. For example, in the alignment of Fig. 2, presumptive regulatory sequences were deliberately aligned and mutations of these sequences would not be expected to be neutral. However, exact values are not necessary for the argument. The striking result of the analysis is that even the smallest (most recent) distance estimates between the intergene sequences are much larger than the most distant calculation for the most divergent histone coding sequences, the H2B genes.

### *Relative Time Scale of the Duplication*

The calculations for the intergene sequences suggest that the evolutionary distance between the duplicated genes is immeasurably large. It is therefore useful to consider a case where protein coding sequences can be aligned unambiguously and a large divergence is expected. The histone genes from the fission yeast *S. pombe* have recently been cloned and sequenced (Choe et al. 1985; Matsumoto and Yanagida 1985). Thus, the evolutionary distance between the histone coding sequences of *S. cerevisiae* and *S. pombe* can be compared with the distance between the two H3–H4 intergene sequences within *S. cerevisiae* itself. Haploid *S. pombe* contains three sets of H3–H4 gene pairs, one set of H2A–H2B pairs, and a second, single H2A gene. The silent component of third-base substitutions was calculated for all pairwise comparisons of homologous histone genes from the two yeasts. Table 2 shows that the average distance for the histone coding sequence divergence is approximately 0.69 bases per position.

The sequences of other genes are known for both *S. cerevisiae* and *S. pombe*. The cytochrome c genes (Smith et al. 1979; Russell and Hall 1982) and the alcohol dehydrogenase genes (Bennetzen and Hall 1982b; Russell and Hall 1983) are two examples. The $K_s'$ values for comparisons of these genes are also shown in Table 2.

## Discussion

The distance calculations for the *S. cerevisiae* histone H3–H4 intergene sequences suggest that they have been separated a long time. While uncertainties remain in the absolute evolutionary alignment of the two regions, the variety of matches and control comparisons used in the current work place limits on the reasonable values of the divergence. These considerations argue that the distance between the two loci is about 1 substitution per base position—perhaps more. Interestingly, this divergence is greater than that for the histone gene coding sequences of *S. cerevisiae* and *S. pombe*, calculated as the silent component of third-base changes. Two other genes from the two yeasts, the cytochrome c and alcohol dehydrogenase genes, also have diverged by nearly 1 base per position. These results lead to the interesting hypothesis that the *S. cerevisiae* H3 and H4 genes were duplicated at a time in evolution that is on the same order as the time of divergence of *S. cerevisiae* from the line leading to *S. pombe* and onward towards *Drosophila* (Mao et al. 1982).

When the intergene results are combined with the measurements on the coding sequences, the results provide evidence for the convergence of genetic information, over evolutionary time, between the homologous nonallelic histone loci. This support is derived from two lines of evidence. First, the large evolutionary distance between the H3–H4 intergenic DNA segments argues that the loci were duplicated much earlier than the times derived from the coding DNA would suggest. Second, the duplication paradox associated with the paired histone genes is most easily resolved by a single duplication event and occasional corrective recombination between individual coding sequences. The most practical implication of this analysis is that coding-DNA comparisons between duplicated genes within the same genome may provide unreliable measurements of the time of duplication of the genes. Some important questions remain unanswered by the present analysis, however.

### *Mechanism of Coding-DNA Convergence*

The mechanism by which divergence between the nonallelic coding sequences has been corrected is unknown. The problem is similar to that of the maintenance of sequence homology among other dispersed repeated gene families. It is only possible to speculate on the basis of the known genetic properties of present-day yeast. In nature *S. cerevisiae* exists primarily as a homothallic diploid. As such, it likely undergoes frequent cycles of meiosis, germination, mating, and vegetative growth. Therefore one possible mechanism for the convergence of the duplicated histone coding sequences is nonallelic gene conversion. The short lengths of DNA involved in the coding-region correction event and the limitation of the extent of the recombination by the divergent flanking DNA are in line with the properties of meiotic gene conversion (Fogel et al. 1981). There are now several experimental precedents for such a mechanism's maintaining sequence similarity in dispersed repeated genes (Scherer and Davis 1980; Ernst et al. 1981; Jackson and Fink 1981). Recently Jinks-Robertson and Petes (1985) demonstrated that the rate of meiotic gene conversion of nonalleles can be high—nearly as high as that for allelic genes.

### *Duplication by Diploidization*

Finally, there is the question of the duplication event itself. As long as the duplication was placed in recent history, the possible mechanisms were limited predominantly to intragenomic rearrangements, such as unequal sister chromatid exchange, chromosome translocation, and transposition (Stiles et al. 1981; McKnight et al. 1981). However, a difficulty is imposed by the unlinked arrangement of the core histone gene pairs: Under these mechanisms it would be necessary to invoke two or more duplication events, one for the H2A–H2B pair and one for the H3–H4 pair.

On the other hand, the hypothesis that the histone gene pairs were duplicated in a more ancient event, perhaps even predating *Saccharomyces,* makes possible a mechanism for copying both gene pairs at once. Both duplications could be accounted for in a single event if modern haploid *S. cerevisiae* evolved by a diploidization or the fusion of two fungal genomes, each carrying one set of core histone gene pairs. In the latter case, the ancient parents are presumed to have evolved from a common progenitor containing one set of H3–H4 genes and one set of H2A–H2B genes. The histone gene duplication then occurred as a fusion of the two primitive yeasts to give a defective diploid genome that went on to evolve into the haploid yeast *S. cerevisiae.* Because of the coding sequence homology imposed by the conserved histone amino acid sequences, rare nonallelic gene conversion events occasionally became fixed in the population during this evolution.

A prediction of the genome-fusion model is that many genes, not just the histone genes, would have been duplicated in the defective diploidization event. Consistent with this prediction is the observation that *S. cerevisiae* has a large number of duplicated genes, including cytochrome c (Smith et al. 1979; Montgomery et al. 1980), alcohol dehydrogenase (Bennetzen and Hall 1982b), enolase (Holland and Holland 1980), glyceraldehyde-3-phosphate dehydrogenase (Holland et al. 1981), the yeast ras genes (Defeo-Jones et al. 1983; Dhar et al. 1984; Powers et al. 1984), the mating-type loci (Nasmyth and Tatchell 1980; Strathern et al. 1980), and many of the ribosomal protein genes (Fried et al. 1981; Abovich et al. 1985). It is expected that additional duplicate genes will continue to be found frequently as new sequences are cloned and examined genetically. If these duplicate genes were also copied by the proposed diploidization, then estimates of the evolutionary distance between them should be identical for each pair. As algorithms for gap treatment and spacer-sequence alignment improve, it may become possible to obtain reliable estimates for these cases.

*Acknowledgments.* I would like to thank Michael L. Johnson for helpful discussions and help with the data analysis methods, William R. Pearson for discussions and computer programs, and the Clinical Research Center of the University of Virginia for access to the VAX-11/750 used in some of the calculations. This investigation was supported by Public Health Service grant GM28920. I would also like to acknowledge the support of the American Cancer Society through Faculty Research Award FRA-257.

## References

Abovich N, Gritz L, Tung L, Rosbash M (1985) Effect of RP51 gene dosage alterations on ribosome synthesis in *Saccharomyces cerevisiae.* Mol Cell Biol 5:3429–3435

Bennetzen JL, Hall BD (1982a) Codon selection in yeast. J Biol Chem 257:3026–3031

Bennetzen JL, Hall BD (1982b) The primary structure of the *Saccharomyces cerevisiae* gene for alcohol dehydrogenase I. J Biol Chem 257:3018–3025

Choe J, Kolodrubetz D, Grunstein M (1982) The two yeast histone H2A genes encode similar protein subtypes. Proc Natl Acad Sci USA 79:1484–1487

Choe J, Schuster T, Grunstein M (1985) Organization, primary structure, and evolution of the histone H2A and H2B genes of the fission yeast *Schizosaccharomyces pombe.* Mol Cell Biol 5:3261–3269

Defeo-Jones D, Scolnick E, Koller R, Dhar R (1983) ras-Related gene sequences identified and isolated from *Saccharomyces cerevisiae.* Nature 306:707–709

Dhar R, Nieto A, Koller R, Defeo-Jones D, Scolnick E (1984) Nucleotide sequence of two ras$^{H}$-related genes isolated from the yeast *Saccharomyces cerevisiae.* Nucleic Acids Res 12: 3611–3618

Ernst JF, Stewart JW, Sherman F (1981) The cyc1-11 mutation in yeast reverts by recombination with a nonallelic gene: composite genes determining the iso-cyctochromes c. Proc Natl Acad Sci USA 78:6334–6338

Fogel S, Mortimer RK, Lusnak K (1981) Mechanisms of meiotic gene conversion, or "Wanderings on a Foreign Strand." In: Strathern JN, Jones EW, Broach JR (eds) The molecular biology of the yeast *Saccharomyces.* Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, pp 289–339

Fried HM, Pearson NJ, Kim CH, Warner JR (1981) The genes for fifteen ribosomal proteins of *Saccharomyces cerevisiae.* J Biol Chem 256:10176–10183

Goad WB, Kanehisa MI (1982) Pattern recognition in nucleic acid sequences. I. A general method for finding local homologies and symmetries. Nucleic Acids Res 10:247–263

Graur D (1985) Amino acid composition and the evolutionary rates of protein-coding genes. J Mol Evol 22:53–62

Grunstein M, Rykowski M, Kolodrubetz D, Choe J, Wallis JW (1984) A genetic analysis of histone protein subtypes in yeast. In: Stein GS, Stein JL, Marzluff WF (eds) Histone genes. John Wiley & Sons, New York, pp 35–63

Guarente L, Yocum RR, Gifford P (1982) A GAL10–CYC1 hybrid yeast promoter identifies the GAL4 regulatory region as an upstream site. Proc Natl Acad Sci USA 79:7410–7414

Hentschel CC, Birnstiel ML (1981) The organization and expression of histone gene families. Cell 25:301–313

Hereford L, Fahrner K, Woolford J Jr, Rosbash M, Kaback DB (1979) Isolation of yeast histone genes H2A and H2B. Cell 18:1261–1271

Holland JP, Holland MJ (1980) Structural comparison of two non-tandemly repeated yeast glyceraldehyde-3-phosphate dehydrogenase genes. J Biol Chem 255:2596–2605

Holland MJ, Holland JP, Thill GP, Jackson KA (1981) The primary structures of two yeast enolase genes. J Biol Chem 256:1385–1395

Jackson JA, Fink GR (1981) Gene conversion between duplicated genetic elements in yeast. Nature 292:306–311

Jinks-Robertson S, Petes TD (1985) High-frequency meiotic gene conversion between repeated genes on nonhomologous chromosomes in yeast. Proc Natl Acad Sci USA 82:3350–3354

Johnson ML, Frasier SG (1985) Non-linear least squares analysis. Methods Enzymol 117:301–342

Johnston M, Davis RW (1984) Sequences that regulate the divergent GAL1–GAL10 promoter in *Saccharomyces cerevisiae.* Mol Cell Biol 4:1440–1448

Kedes LH (1979) Histone genes and histone messengers. Annu Rev Biochem 48:837–870

Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267:275–276

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci USA 78:454–458

Korn LJ, Queen CL, Wegman MN (1977) Computer analysis of nucleic acid regulatory sequences. Proc Natl Acad Sci USA 74:4401–4405

Mao J, Appel B, Schaack J, Sharp S, Yamada H, Söll D (1982) The 5S RNA genes of *Schizosaccharomyces pombe.* Nucleic Acids Res 10:487–500

Matsumoto S, Yanagida M (1985) Histone gene organization of a fission yeast: a common upstream sequence. EMBO J 4: 3531–3538

Maxson R, Cohn R, Kedes L (1983) Expression and organization of histone genes. Annu Rev Genet 17:239–277

McKnight GL, Cardillo TS, Sherman F (1981) An extensive deletion causing overproduction of yeast iso-2-cytochrome c. Cell 25:409–419

Miyata T, Hayashida H (1981) Extraordinarily high evolution rate of pseudogenes: evidence for the presence of selective pressure against changes between synonymous codons. Proc Natl Acad Sci USA 78:5739–5743

Miyata T, Yasunaga T, Nishida T (1980) Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc Natl Acad Sci USA 77:7328–7332

Montgomery DL, Leung DW, Smith M, Shalit P, Faye G, Hall BD (1980) Isolation and sequence of the gene coding for iso-2 cytochrome c in *Saccharomyces cerevisiae.* Proc Natl Acad Sci USA 77:541–545

Nasmyth KA, Tatchell K (1980) The structure of transposable yeast mating type loci. Cell 19:753–764

Powers S, Kataoka T, Fasano O, Goldfarb M, Strathern J, Broach J, Wigler M (1984) Genes in *S. cerevisiae* encoding proteins with domains homologous to the mammalian ras proteins. Cell 36:607–612

Russell PR, Hall BD (1982) Structure of the *Schizosaccharomyces pombe* cytochrome c gene. Mol Cell Biol 2:106–116

Russell PR, Hall BD (1983) The primary structure of the alcohol dehydrogenase gene from the fission yeast *Schizosaccharomyces pombe.* J Biol Chem 258:143–149

Scherer S, Davis RW (1980) Recombination of dispersed repeated DNA sequences in yeast. Science 209:1380–1384

Smith MM (1984) The organization of the yeast histone genes. In: Stein GS, Stein JL, Marzluff WF (eds) Histone genes. John Wiley & Sons, New York, pp 3–33

Smith MM, Andrésson ÓS (1983) DNA sequences of yeast H3 and H4 histone genes from two non-allelic gene sets encode identical H3 and H4 proteins. J Mol Biol 169:663–690

Smith MM, Murray K (1983) Yeast H3 and H4 histone messenger RNAs are transcribed from two non-allelic gene sets. J Mol Biol 169:641–661

Smith M, Leung DW, Gillam S, Astell CR, Montgomery DL, Hall BD (1979) Sequence of the gene for iso-1-cytochrome c in *Saccharomyces cerevisiae.* Cell 16:753–761

Stiles JI, Friedman LR, Sherman F (1981) Studies on transposable elements in yeast. II. Deletions, duplications, and transpositions of the COR segment that encompasses the structural gene of yeast iso-1-cytochrome c. Cold Spring Harbor Symp Quant Biol 45:602–607

St John TP, Davis RW (1981) The organization and transcription of the galactose gene cluster of *Saccharomyces cerevisiae.* J Mol Biol 152:285–315

St John TP, Scherer S, McDonell W, Davis RW (1981) Deletion analysis of the *Saccharomyces* GAL gene cluster. J Mol Biol 152:317–334

Strathern JN, Spatola E, McGill C, Hicks JB (1980) Structure and organization of transposable mating type cassettes in *Saccharomyces cerevisiae.* Proc Natl Acad Sci USA 77:2839–2843

Wilbur WJ, Lipman DJ (1983) Rapid similarity searches of nucleic acid and protein data banks. Proc Natl Acad Sci 80: 726–730

Received November 5, 1985/Revised May 29, 1986