UNIVERSITY OF CALIFORNIA

Los Angeles

The Effect of Various Histone H4 Sequences on Yeast Viability

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Biology

by

Gary Bryce Fogel

1998

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

The dissertation of Gary Bryce Fogel is approved.

David Chapman

David Chapman

Charles Marshall

Clifford Brunk

Clifford Brunk, Committee Chair

University of California, Los Angeles

1998

DEDICATION

To my mother for her dedication and

to Joanne for her tireless support and love.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

I would first like to thank the members of my committee Dr. David Chapman, Dr. David Eisenberg, Dr. Charles Marshall, and Dr. Robert Wayne for their support and guidance. Most importantly I would like to thank my advisor Dr. Clifford Brunk for giving me the opportunity to work in his lab. He helped design and construct much of the equipment and computer programs essential to this dissertation. More importantly, he taught me the power of parallel processing and for that I will always be grateful.

Several members of the Life Sciences Administration were essential over the years at UCLA. Annie Alpers, Annette and Richard Klufas, Margaret Kowalczyk, Rhonda Murotake, and Jocelyn Yamadera all contributed in various ways to the success of my stay at UCLA. Dr. Joseph Cascarano, Dr. Bill Ebersold, Dr. Steve Strand, Dr. Peter Narins and Dr. Kenneth Jones were a joy to communicate with. I would like to thank Dr. Dohn Glitz for the use of his oligonucleotide synthesizer. Most importantly I would like to thank both Dr. Roger Bohman for being a true inspiration in teaching and for giving me the opportunity to serve as a Teaching Assistant in the largest class ever to be taught at UCLA. I would be remiss if I did not also thank Dr. Austin MacInnis for his service as Team Captain of the Biology softball team and for letting me remember the smell of the newly mowed grass in left field every now and then.

On becoming a Fellow of the Center for the Study of the Evolution and Origin of Life, I was introduced to Dr. Bill Schopf, who provided an example of enthusiasm and leadership at the Wednesday Evening Evolution Group meetings and introduced me to Erik Schultes and John Bragin among others. Fellowship support from CSEOL was invaluable in the completion of this dissertation research.

Dan Laserna, Micheal Passaides, Jennifer Palmore, Kenneth Chen, Aslan Turer, Jacob Lacayo, Olivia Mereno, Vahe Sarkissian, Tooraj Bereliani, Sarah Ter-Minasyan, Dmitriy Niyazov, Roxanne Torabian-Bashardoust, Joe Leung, Marianne Makely, and Natalie Zahr trained in the laboratory as undergraduates and each assisted in the completion of my degree. I hope that they learned as much from me as I did from each of them. I also hope that one day they will also have the opportunity to instruct a student under their guidance.

Several graduate students, post-docs, and visiting professors including Margaret Harmon, Marta DeJesus, Patrick and Parvie Navas, Bobby and Mimi Khan, Bo and Helge Anderson, and Torsten Graybill shared my struggle for existence and helped me achieve my goals during my stay at UCLA. Of special recognition are Bruce Olsen, Chris Collins, Jinliang Li, Mark Boian and last (but certainly not least) Erik Avaniss-Aghajani, who not only encouraged me in the lab, but introduced me to the art of trout fishing. Completing my Ph.D. would have been far more difficult if it were not for the efforts and friendship of Erik.

Most of all I wish to thank my mother, father, brother, Joanne and the Lee family for their constant love and support during my seven years of labor. Thanks also go to my grandmother and the many four leaf clovers she collected for me. I'm sure their luck rubbed off on me now and then. I would especially like to thank my father for reiterating the question "tell me something I don't know about" throughout my childhood. After all those years, I can finally respond in the manner I hoped for with the completion of this dissertation.

I would like to thank the following for copyright permission and co-authorship in specific chapters:

Chapter 2. Reprinted from *Expression of Tetrabymena Histone H4 in Yeast*, by Gary B. Fogel and Clifford F. Brunk, Copyright 1997, Biochimica et Biophysica Acta 1354:116-126 (1997), with permission from Elsevier Science.

Chapter 3. Preprinted from *Temperature Gradient Chamber for Relative Growth Rate Analysis of Yeast*, by Gary B. Fogel and Clifford F. Brunk. Analytical Biochemistry. In Press. Copyright 1998 Academic Press. With permission from Academic Press. Patent application is pending with the United States Patent Office.

Appendix 1: Work completed with the assistance of Dmitriy Niyazov and Clifford F. Brunk.

Appendix 2. Work completed with the assistance of Christopher R. Collins, Jinliang Li and Clifford F. Brunk.

VITA

April 7, 1968	Born, La Jolla, California
1991	Bachelor of Arts, Biology University of California, Santa Cruz Santa Cruz, California
1991-1993	Graduate Student Fellow Center for the Study of the Evolution and Origin of Life Institute of Geophysics and Planetary Physics University of California, Los Angeles Los Angeles, California
1991-1996	Teaching Assistant/Associate/Fellow Department of Biology University of California, Los Angeles Los Angeles, California
1993	Candidate in Philosophy, Biology University of California, Los Angeles Los Angeles, California
1994	Schectman Award for Teaching Excellence Department of Biology University of California, Los Angeles Los Angeles, California
1995-1996	U.C.L.A. Graduate Student Award College of Letters and Sciences University of California, Los Angeles Los Angeles, California
1997	Scherbaum Award for Research Excellence Department of Biology University of California, Los Angeles Los Angeles, California
1997-1998	Research Assistant Department of Biology University of California, Los Angeles Los Angeles, California

PUBLICATIONS AND PRESENTATIONS

Fogel, G.B. and Brunk, C.F. (1998) Cassette Mutagenesis of a Conserved Core Region of *Saccharomyces cerevisiae* Histone H4. Poster presented at the 1998 Biology Research Symposium, Department of Biology UCLA.

Collins, C.R., Fogel, G.B., Li, J., and Brunk, C.F. (1998) A Survey of Prokaryotic Genome Size and SSU rRNA Copy Number for Microbial Relative Abundance Estimation. Poster presented at the 1998 Biology Research Symposium, Department of Biology UCLA.

Fogel, G.B. (1998) Sequence Reconstruction from a Simulated DNA Chip Using Evolutionary Programming. In: *Evolutionary Programming VII: Proceedings of the Seventh International Conference on Evolutionary Programming.* (in press).

Fogel, G.B. and Brunk, C.F. (1998) Temperature Gradient Chamber for Relative Growth Rate Analysis of Yeast. *Analytical Biochemistry* (in press).

Fogel, D.B., Fogel, G.B. and Andrews, P.C. (1998) On the Instability of Evolutionary Stable Strategies in Small Populations. *Ecological Modeling* (in press).

Fogel, G.B. (1997) The Application of Evolutionary Computation to Selected Problems in Molecular Biology. In *Evolutionary Programming VI: Proceedings of the Sixth International Conference on Evolutionary Programming.* eds. P.J. Angeline, R.G. Reynolds, J.R. McDonnell, and R. Eberhart. Springer-Verlag, Berlin. pp. 23-33.

Fogel, G.B. and Brunk, C.F. (1997) Mutagenesis of *Saccharomyces cerevisiae* Histone H4 Using Cassettes Derived From Amino Acid Replacement Matrices. Presented at the Fifth Annual Meeting of the Society for Molecular Biology and Evolution. Garmisch-Partenkirchen, Germany.

Fogel, G.B. and Brunk, C.F. (1997) Expression of *Tetrahymena* Histone H4 in Yeast. *Biochimica et Biophysica Acta.* 1354:116-126.

Fogel, D.B., Fogel, G.B. and Andrews, P.C. (1997) On the Instability of Evolutionary Stable Strategies. *BioSystems.* 44:135-152.

Fogel, G.B. and Brunk, C.F. (1996) A Method to Assay Functional and Non-Functional Modifications to a Universally Conserved Region of *Saccharomyces cerevisiae* Histone H4. Poster presented at the Gordon Research Conference in Molecular Evolution, Ventura, California.

Fogel, G.B. (1995) Viable Alterations to *Tetrahymena* Histone H4 Using Random Oligonucleotide Mutagenesis. Presented at the Fifth Annual West Coast Ciliate Conference, U.C. Santa Barbara, Santa Barbara, California.

Fogel, G.B. and Fogel, D.B. (1995) Forecasting Precipitation in Los Angeles Using Evolutionary Programming. In *Mission Earth: Modeling and Simulation for a Sustainable Future*. ed. A. M. Wildberger, Society for Computer Simulation, San Diego, CA. pp. 65-70.

Fogel, G.B. and Fogel, D.B. (1995) Continuous Evolutionary Programming: Analysis and Experiments. *Cybernetics and Systems*. 26(1):79-90.

Fogel, D.B. and Fogel, G.B. (1995) Evolutionary Stable Strategies are Not Always Stable Under Evolutionary Dynamics. In *Evolutionary Programming IV: Proceedings of the Fourth Conference on Evolutionary Programming*. eds. J.R. McDonnell, R.G. Reynolds, and D.B. Fogel. MIT Press, Cambridge, MA. pp. 565-577.

Fogel, G.B. (1994) Histone Variability in the Ciliates. Presented at the Fourth Annual West Coast Ciliate Conference, U.C. Santa Barbara, Santa Barbara, California.

Fogel, G.B. and Brunk, C.F. (1994) An Assay of Potential Variability in Ciliate Histone H4. Poster presented at the Second Annual Meeting of the Society for Molecular Biology and Evolution, University of Georgia, Athens, Georgia.

Fogel, G.B. and Brunk, C.F. (1994) Using Random Oligonucleotides to Determine Variability in *Tetrahymena* Histone H4. Poster presented at the American Association for the Advancement of Science Annual National Conference, San Francisco, California.

Fogel, G.B. (1993) A Paleoenvironmental/Taphonomic Analysis of the San Diego Formation at Southern La Jolla, California. Presented at the 1993 California Paleontology Conference. U.C. Los Angeles, Los Angeles, California.

Fogel, G.B. (1993) The Protein Folding Problem and the Potential Application of Evolutionary Programming. In *Proceedings of the Second Annual Conference on Evolutionary Programming*. eds. D.B. Fogel and W. Atmar. Evolutionary Programming Society, San Diego, CA. pp. 170-177.

Fogel, G.B. (1993) Random Oligonucleotide Mutagenesis of a Highly Conserved Histone Core Sequence in *Tetrahymena*: an Analysis of Histone Variability. Presented at the Third Annual West Coast Ciliate Conference, Loyola Marymount University, Los Angeles, California.

Brunk, C.F., Fogel, G.B., and Navas, P.A. (1993) Unusual Patterns in Nucleotide Substitutions Among Species of *Tetrahymena*. Presented at the First Annual Meeting of the Society for Molecular Biology and Evolution, U.C. Irvine, Irvine, California.

Fogel, D.B., Fogel, L.J., Atmar, W., and Fogel, G.B. (1992). Hierarchic Methods of Evolutionary Programming. In *Proceedings of the First Annual Conference on Evolutionary Programming*. eds. D.B. Fogel and W. Atmar. Evolutionary Programming Society, San Diego, CA. pp. 175-182.

ABSTRACT OF THE DISSERTATION

The Effect of Various Histone H4 Sequences on Yeast Viability

by

Gary Bryce Fogel

Doctor of Philosophy in Biology

University of California, Los Angeles, 1998

Professor Clifford Brunk, Chair

Histone H4 is one of the most conserved proteins known. The low rate of non-synonymous substitutions over long evolutionary periods suggests that histone H4 is under severe selective constraints. Previous mutational analyses of histone H4 suggest that the central core region of the protein is less permissive to modification than the amino or carboxy termini. The focus of this dissertation was to introduce novel mutations to the core of histone H4 and develop an *in vivo* assay for viability for each of these modifications in the yeast *Saccharomyces cerevisiae*.

A background on the evolutionary history of histone H4 is included in Chapter 1. In order to test a large number of mutations simultaneously in histone H4, the wild-type yeast histone H4 was replaced by the highly divergent histone H4 from *Tetrahymena thermophila*. This replacement forces yeast to utilize a histone with 21 amino acid replacements out of 102 amino acids relative to the wild-type protein. Growth characteristics of these yeast were assayed over a range of temperatures suggesting that

histone H4 still functions in yeast even after a large number of alterations were scattered throughout the protein.

To test the relative growth rates of yeast utilizing modified histone H4 proteins, a temperature gradient incubator was developed. In combination with computer imaging technology, this incubator can be used to assay 24 yeast strains simultaneously across a user-specified temperature gradient.

By using oligonucleotide cassette mutagenesis, a series of modifications were introduced to a universally conserved core region in histone H4. Using the first cassette, a sequence space of 256 possible histone H4 proteins was assayed in both the yeast histone H4 and the *Tetrahymena* histone H4 contexts. The analysis suggests that different sets of modifications will work in each of these contexts. Using the second cassette, a sequence space of 13,824 possible histone H4 proteins was assayed in the yeast histone H4 background. 9.2% of these sequences were viable in yeast. This suggests that there is a much larger space of potentially viable histone H4 sequences than has been realized in light of the extreme historical conservation of this protein.

Chapter 1

INTRODUCTION

The nucleosome is the primary repeating unit of DNA in eukaryotic chromatin (24, 28). Nucleosomes are octamers containing two each of histone proteins H2A, H2B, H3, and H4 as well as 146 base pairs of DNA wrapped around the histones in a left-handed spiral (36). Linker histone H1 (or H5 in the case of avian erythrocytes) is positioned between nucleosomes on a variable length segment of linker DNA in the formation of a "chromatosome" and is involved with regulating gene activity (4). The nucleosome octamer has been further characterized as having a tripartite organization of a central $(H3\text{-}H4)_2$ tetramer flanked by two H2A-H2B dimers with an approximate diameter of 70Å and length of 55Å when resolved at 3.1Å (1, 6, 51).

The histone octamer has unique selective constraints. Preservation of pairings between neighboring histones in the formation of dimers, pairings between histone dimers in the formation of tetramers, as well as the maintenance of the proper pitch, outer diameter, and local pattern of electrostatic character to complement the binding of the DNA helix on the outer surface of the nucleosome must all be maintained by the octamer (30). These significant constraints maintain a very high selective pressure on all of the histone proteins. The 102 amino acid histone H4 is widely considered to be one of the most evolutionarily conserved proteins characterized to date. There exists a dramatic lack of diversity in histone H4 sequences between eukaryotes as different as cows and pea plants which have nearly identical histone H4 amino acid sequences (98% similarity) (9). Such extreme selection is thought to curtail an exploration of alternative amino acid sequences that could function in the role of histone H4.

Histone H4 contains both an amino-terminal "tail" and a more highly conserved central globular "core." The amino-terminal "tail" apparently plays a role in gene regulation (16). RNA synthesis *in vitro* can be prevented by the presence of nucleosomes at transcription initiation sites (25) and coding regions of eukaryotes are more sensitive to DNase I activity during transcription suggesting that active DNA is free of nucleosomes *in vivo* (52). Han et al. (18) demonstrated that an arrest of histone H4 mRNA synthesis and corresponding nucleosomal loss from the *PHO5* gene promoter led to strong activation of the *PHO5* gene. Several similar studies documenting the roles for histone proteins in the regulation of gene expression suggest that nucleosomes, and in particular the amino-terminal tail of histone H4, can be held directly responsible for differences in gene activity (11, 17, 22, 40, 43, 54).

Striking results via deletion experiments of both the amino- and carboxy-terminal regions of yeast histone H4 demonstrated not only that the amino-terminus was a participant in the specific derepression of the silent mating loci *HMLα* and *HMR***a**, but also effected chromatin structure and cell cycle duration (23). Similar deletion experiments in the amino-terminus of histones H2A and H2B failed to show this effect on yeast mating suggesting a regulatory effect unique to histone H4. Whitlock and Stein (53) demonstrated that after digestion of nucleosomes with trypsin and corresponding loss of amino-terminal tails, histones still retained the ability to fold DNA *in vitro* suggesting that only the core and carboxy-termini were necessary for nucleosomal formation. This result was later confirmed *in vivo* with the chromatin of *Xenopus* embryos (13).

In contrast, Megee et al. (29) noted that within the amino-terminus, the four lysine residues at positions 5, 8, 12, and 16 in histone H4 were essential for function. Replacement of arginine or asparagine at all four positions was lethal in yeast. However, replacement of any one of the positions with arginine gave viable cells with a much reduced DNA replication rate. Simultaneous replacement of all four lysine positions with glutamine gave viable cells with altered phenotypes including sterility, slower progress through G_2/M phase, temperature-sensitive growth, and a prolonged period of DNA replication. In other words, through the use of directed point mutations and creation of amino acid replacements, convincing evidence was developed for the role of the histone H4 amino-terminal lysine residues in gene expression. Hong et al. (20) noted that reversible acetylation and modulation of ionic conditions *in vivo* also help to regulate the effects of the amino-terminus in gene expression. Although simultaneous replacement of all four positions with either arginine or asparagine leads to non-viable yeast cells (29), deletion of the amino terminus unexpectedly leads to viable yeast cells (23, 29). Therefore, deletion of the histone H4 amino terminus appears to have less severe consequences for viability than the replacement of the first four lysine residues by arginine.

The Histone Fold Motif and the Evolution of Histone Proteins

Amino acid sequences of the histone H4 core display an unusually low degree of sequence divergence across all eukaryotic taxa. All four types of histones have very low sequence similarity and yet share a common tertiary motif known as the "histone fold" in their core (1). This similarity is much greater than would have been predicted on the basis of primary sequence information alone (Figure 1.1). Each histone fold appears to be a tandem duplication of two similar helix-loop-helix (HLH) motifs (HLH1 and HLH2) creating a full motif that is 65 residues in length (30). The fold is thought to be essential for the histone dimerization and chromatin assembly as well as to provide specific residues for DNA binding on the surface of the nucleosome (13). Hydrophobic residues believed to be critical for histone dimerization are highly conserved in evolution. Modification of these amino acids would be expected to be lethal except in the unlikely event of two successful simultaneous amino acid replacements in two adjoining histones. The HLH2 motif in histone H4 is known to be more conserved than its neighbor HLH1; this is consistent with the theory that the helix regions of HLH2 are most important for protein-protein interactions. A lysine-arginine pair in the strand region of HLH2 is highly conserved and appears to be crucial for DNA binding (30). Another region of DNA binding is located in the strand region of HLH1 and is believed to aid in the wrapping of DNA around the octamer, such that the DNA maintains a curvature allowing it to coil tightly around the nucleosome (30).

The common belief that the histone fold is confined to the histone protein family has recently been contested (6, 33, 34, 49). Within the eukarya, CCAAT-binding transcription factor subunits CBF-A and CBF-C from *Gallus gallus, Homo sapiens, Mus musculus* and *Rattus norvegicus* (7,33), the TBP-associated factors dTAF$_{II}$42 and dTAF$_{II}$62 from *Drosophila* (56), hTAF80, hTAF31, and hTAF20/15 subunits of TFIID from *Homo sapiens* (19) contain histone fold motifs or histone octameric-like structures. Even more interesting, however, was the recent discovery that several archaeal DNA binding proteins show similarity to the histone fold such as the *hanlA* gene product from *Thermococcus* (39), HMt from *Methanobacterium thermoautotrophicum* (47), and HMfA and HMfB from *Methanothermus fervidus* (15, 42, 48). No eubacterial proteins identified to date show similarly to the histone fold and curiously no histone proteins have yet been discovered in the crenarchaeota. Whereas eukaryotic histones only form heterodimers (H2A+H2B and H3+H4), archaeal histones form both heterodimers and homodimers and the ratio of dimers changes *in vivo* with respect to growth conditions (35). Together, these data suggest: 1) the histones are homologous to archaeal ancestral DNA binding proteins (35), 2) histones and some eukaryotic transcription factors have diverged from the common histone fold ancestor and that ancestral protein could have acted as a homodimer, and 3) eukarya and archaea are more closely related to each other than either is to the bacteria (Figure 1.2). This provides further agreement with recent phylogenies derived from rRNA (8, 10, 31) or proteins such as TFIIB (32), V-type and F-type ATPases (14) and EF-Tu or EF-G (3, 21). In comparison to the archaeal histone-like proteins (HLPs), the histone H4 family is most similar and is thought to be the progenitor of all of the other core histones in eukaryotes (33). Therefore, histone H4 is a crucial link to the understanding of the evolution and subsequent diversification of the histone proteins in the eukaryotes.

Histone H4 in Lower Eukaryotes

A major distinction between the prokaryotes and eukaryotes is the ability of eukaryotic nuclei to package large lengths of DNA into nuclei on the order of five nanometers in diameter using histone proteins. Within the eukaryotes, the protists offer a wide range of unique DNA packaging problems and solutions. Therefore, analysis of protist histones should offer unique opportunities to study histone evolution. For example of all eukaryotes, only the uninucleate dinoflagellates lack true histones and instead utilize a set of "histone-like" proteins reminiscent of the archaeal proteins to fold DNA (37). Within the ciliated protozoa genus *Tetrahymena,* there exists a large degree of histone H4 primary sequence variation. *Tetrahymena* also have a large difference in histone H4 primary sequence relative to the vertebrates (41), suggesting the possibility of a reduced selection pressure on the histones in *Tetrahymena* allowing for an exploration of a different region of histone H4 protein "sequence space."

Biochemical studies on the histone genes of ciliates have been restricted thus far to a few genera including *Tetrahymena, Paramecium, Stylonychia, Oxytricha,* and *Euplotes.* In general, all of the ciliates display high diversification of histone H4 amino acid sequence, perhaps related to their unusual nuclear dimorphism. Ciliates possess two functional nuclei: a somatic macronucleus and a germline micronucleus. The apparent sequence diversity of histone H4 in *Tetrahymena* and other ciliates makes these proteins interesting with respect to their plasticity and evolution. However, ciliates are difficult to manipulate in the lab especially with regards to the transformation of plasmid DNA. Yeast, on the other hand, share many technical advantages suitable for manipulation in the lab (44). These include rapid growth, ease of replica plating and mutant isolation, well-defined genetics, and DNA transformation techniques. As a result, the yeast *Saccharomyces cerevisiae* is recognized as an ideal eukaryote for molecular biological studies.

Kayne et al. (23) developed a system for yeast known as the "glucose shift viability assay." This system can be used in combination with a yeast shuttle vector to express mutant histone H4 genes and assay fitness *in vivo*. The procedure was used previously to study the effect of deletions in wild-type yeast histone H4 (23). We have utilized the "glucose shift viability assay" to test the function of *Tetrahymena* histone H4 proteins in yeast in which the endogenous yeast histone H4 gene has been replaced by a modified histone H4 gene. The results of these experiments are the subject of Chapter 2. When forced to utilize the *Tetrahymena* histone H4 gene, yeast cells can survive although they have several alterations in their phenotype. This surprising result occurs even with the large number of amino acid replacements scattered throughout the protein. Our research is the first successful trans-species expression of a histone H4 gene *in vivo*. Similar research using yeast to express the histone H5 from *Xenopus laevis* (45), H1 from the sea urchin *Pasmmechinus miliaris* (26) and histone H2A from *Tetrahymena thermophila* (27) have also been successful.

Measurement of Fitness

After the successful incorporation of a novel gene into a yeast cell, the effects of this replacement to the fitness of the organism must be quantitatively measured. Commonly measured phenotypic responses include effects on mating, colony size or color, and the duration of cell cycle phases. An assay of cell growth at different temperatures offers a relatively easy metric of phenotypic difference. However, growth rate measurements are usually quite time-consuming involving clonal growth in liquid culture, serial dilution, and particle counting measurements for each specified temperature to be analyzed. Such measurements were part of the analysis of the ability of *Tetrahymena* histone H4 to function in yeast shown in Chapter 2. In order to generate a more rapid technique for the measurement of growth rates, a new "temperature gradient incubator" was constructed. This instrument gives parallel growth of yeast clones across a specified temperature gradient and can be quantified by subsequent computer analysis. A description of this instrument is the focus of Chapter 3.

Definition of Sequence Space

The evolution of histone H4 has been an exploration of the potential amino acid sequences capable of meeting the combined demands of several different selective pressures. These selective pressures do not act uniformly across the entire amino acid sequence. For instance, the histone fold in the core of the histone proteins is the backbone of all histone structure. Amino acid sequences that fail to fold into the shape of the histone fold have lethal consequences and will be eliminated by natural selection.

It is possible to view the evolution of histone H4 as a path through the immense space of potential amino acid sequences of length 102 amino acids (the length of histone H4). This space containing 20^{102} possible sequences is clearly too large to sample completely even over the billions of years of eukaryotic life on Earth. For instance, the number of hydrogen atoms in the universe has been estimated to be on the order of 10^{108} atoms (12). In a theoretical context, a fitness landscape (55) could be applied to the set of all histone sequences where the height of the landscape corresponds to the fitness of that sequence in a given eukaryotic (for instance: *Saccharomyces cerevisiae*). By incorporating *in vitro* mutation, we have attempted to describe the landscape of a very small region in the sequence space surrounding the wild-type *Saccharomyces cerevisiae* histone H4 using growth rate as a marker of fitness.

In vitro selection and amplification techniques have recently been used to assay and isolate RNA sequences with specified biochemical properties from a random set of RNAs (46). Using this *in vitro* "directed evolution" approach, new classes of catalytic RNAs (5), ribozymes that specifically cleave single-stranded DNA (38), RNAs with a higher affinity for binding T4 DNA polymerase (50) , and ribozymes capable of forming peptide bonds (57) have been isolated. In effect, these methods search a sequence space of random RNAs, assay their functionality *in vitro*, and select for desired functions.

The construction, transformation, expression and subsequent fitness analysis of a suite of modified histone H4 genes in yeast helps to define the character of the fitness landscape surrounding the wild-type yeast histone H4 protein. In turn, these results provide clues towards the evolutionary conservation of this protein in the eukaryotes. A series of cassettes introduced into an absolutely conserved region in the core of histone H4 (in essence, creating new sequences scattered across the fitness landscape surrounding the wild-type sequence) is the subject of Chapter 4. A mutagenesis strategy that uses a PAM matrix as a predictor of viable or non-viable amino acid replacements is described in Chapter 5. An exploration of histone sequence space with new sequences helps to establish a better understanding of the permissible variation in histone H4.

While constructing the expression vector necessary for the expression of histone H4 genes in yeast, field inversion gel electrophoresis conditions which enhanced the separation of DNA fragments less than 50 kb in length were developed (Appendix 1). Our approach utilized a very short pulse of a high electric field in the forward direction and a long pulse of low electric field in the reverse direction where the integrated forward electric field over time was twice that of the reverse field. This field inversion regime can be used to increase the separation of similarly sized fragments relative to their separation under constant field conditions.

Appendix 2 contains a review of prokaryotic genome sizes and SSU rRNA copy numbers from the literature. A technique to estimate the relative abundance of particular bacterial taxa from a mixed sample is derived. Using the data in Appendix 2, it is possible to determine the percentage of a sample that is composed of a particular prokaryotic taxon knowing an estimate for the weighted average genome size in the sample and the SSU rRNA copy number for the organism of interest. This technique is very useful in applied and environmental microbiology.

REFERENCES

1. Arents, G., R. W. Burlingame, B. -C. Wang, W. E. Love and E. N. Moudrianakis. 1991. The nucleosomal core histone octamer at 3.1Å resolution: A tripartite protein assembly and a left-handed superhelix. Proc. Natl. Acad. Sci. USA. 88:10148-10152.

2. Arents, G. and E. N. Moudrianakis. 1995. The histone fold: A ubiquitous architectural motif utilized in DNA compaction and protein dimerization. Proc. Natl. Acad. Sci. USA. 92:11170-11174.

3. Baldauf, S. L., J. D. Palmer, and W. F. Doolittle. 1996. The root of the universal tree and the original of eukaryotes based on elongation factor phylogeny. Proc. Natl. Acad. Sci. USA. 93:7749-7754.

4. Baldwin, J. P. 1992. Protein-nucleic acid interactions in nucleosomes. Cur. Op. Struct. Biol. 2:78-83.

5. Bartel, D. P. and J. W. Szostak. 1993. Isolation of new ribozymes for a large pool of random sequences. Science. 261:1411-1418.

6. Baxevanis, A. D., J. E. Godfrey and E. N. Moudrianakis. 1991. Associative behavior of the histone $(H3\text{-}H4)_2$ tetramer: dependence on ionic environment. Biochemistry. 30:8817-8823.

7. Baxevanis, A. D., G. Arents, E. N. Moudrianakis and D. Landsman. 1995. A variety of DNA binding and multimeric proteins contain the histone fold motif. Nuc. Acids. Res. 23:2685-2691.

8. Brown, J. R. and W. F. Doolittle. 1997. Archaea and the prokaryote-to-eukaryote transition. Microbiol. and Molec. Rev. 61(4):456-502.

9. DeLange, R. J., D. M. Fambrough, E. L. Smith and J. Bonner. 1969. Calf and pea histone IV. Jour. Biol. Chem. 244(20):5669-5679.

10. Doolittle, W. F. and J. R. Brown. 1994. Tempo, mode, the progenote, and the universal root. Proc. Natl. Acad. Sci. USA. 91:6721-6728.

11. Durrin, L. K., R. K. Mann, P. S. Kayne and M. Grunstein. 1991. Yeast histone H4 N-terminal sequence is required for promoter activation in vivo. Cell. 65:1023-1031.

12. Eigen, M. Steps towards life: A perspective on evolution. Oxford University Press. Oxford. 1992.

13. Freeman, L, H. Kurumizaka and A. Wolffe. 1996. Functional domains for assembly of histone H3 and H4 into the chromatin of *Xenopus* embryos. Proc. Natl. Acad. Sci. USA. 93:12780-12785.

14. Gogarden, J. P., H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, T. Oshima, J. Konishi, K. Denda and M. Yoshida.

1989. Evolution of the vacuolar H$^+$-ATPase: implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. USA. 86:6661-6665.

15. Grayling, R. A., K. Sandman and J. N. Reeve. 1996. Histones and chromatin structure in hyperthermophilic archaea. FEMS Microbiol. Rev. 18:203-213.

16. Grunstein, M. 1990. Nucleosomes: regulators of transcription. Trends Genet. 6:395-400.

17. Grunstein, M. 1992. Histones as regulators of genes. Sci. Am. 267(4):68-74B.

18. Han, M., U. -J. Kim, P. Kayne and M. Grunstein. 1988. Depletion of histone H4 and nucleosomes activates the PHO5 gene in *Saccharomyces cerevisiae*. EMBO J. 7(7):2221-2228.

19. Hoffmann, A., C. -M. Chiang, T. Oelgeschlaeger, X. Xie, S. K. Burley, Y. Nakatani and R.G. Roeder. 1996. A histone octamer-like structure within TFIID. Nature. 380:356-359.

20. Hong, L., G. P. Schroth, H. R. Matthews, P. Yau, and E. M. Bradbury. 1993. Studies of the DNA binding properties of histone H4 amino terminus. J. Biol. Chem. 268(1):305-314.

21. Iwabe, N., K. -I. Kuma, M. Hasegawa, S. Osawa and T. Miyata. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA. 86:9355-9359.

22. Johnson, L. M. 1990. Genetic evidence for an interaction between SIR3 and histone H4 in the repression of the silent mating loci in *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA. 87:6286-6290.

23. Kayne, P. S., U. -J. Kim, M. Han, J. R. Mullen, F. Yoshizaki and M. Grunstein. 1988. Extremely conserved histone H4 N-terminus is dispensable for growth but essential for repressing the silent mating loci in yeast. Cell. 55:27-39.

24. Kornberg, R. D. 1977. Structure of chromatin. Ann. Rev. Biochem. 46:931-954.

25. Lorch, Y., W. LaPointe and R. D. Kornberg. 1987. Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. Cell. 49:203-210.

26. Linder, C. and F. Thoma. 1994. Histone H1 expressed in *Saccharomyces cerevisiae* binds to chromatin and affects survival, growth, temperature, and plasmid stability but does not change nucleosomal spacing. Mol. Cell. Biol. 14(4):2822-2835.

27. Liu, X., J. Bowen and M. A. Gorovsky. 1996. Either of the major H2A genes but not an evolutionary conserved H2A.F/Z variant of *Tetrahymena* can function as the sole H2A gene in the yeast *Saccharomyces cerevisiae*. Mol. Cell. Biol. 16:2878-2887.

28. McGhee, J. D. and G. Felsenfeld. 1980. Nucleosome structure. Ann. Rev. Biochem. 49:1115-1156.

29. Megee, P. C., B. A. Morgan, B. A. Mittman and M. Mitchell Smith. 1990. Genetic analysis of histone H4: Essential role of lysines subject to reversible acetylation. Science. 247:841-845.

30. Moudrianakis, E. N. and G. Arents. 1993. Structure of the histone octamer core of the nucleosome and its potential interactions with DNA. Cold Spring Harbor Symposia on Quantitative Biology, Volume LVIII. pp.273-279.

31. Olsen, G. J. and C. R. Woese. 1997. Archael genomics: an overview. Cell. 89:991-994.

32. Ouzounis, C. and C. Sander. 1992. TFIIB, an evolutionary link between the transcription machineries of archaebacteria and eukaryotes. Cell. 71:189-190.

33. Ouzounis, C. A. and N. C. Kyrpides. 1996a. Parallel origins of the nucleosome core and Eukaryotic transcription for archaea. J. Mol. Evol. 42:234-239.

34. Ouzounis, C. A. and N. C. Kyrpides. 1996b. The core histone fold: Limits to functional versatility. J. Mol. Evol. 43:541-542.

35. Reeve, J. N., K. Sandman and C. J. Daniels. 1997. Archaeal histones, nucleosomes, and transcription initiation. Cell. 89:999-1002.

36. Richmond, T. J., J. T. Finch, B. Rushton, D. Rhodes and A. Klug. 1984. Structure of the nucleosome core particle at 7Å resolution. Nature. 311:532-537.

37. Rizzo, P. J. 1985. Histones in protistan evolution. BioSystems. 18:249-262.

38. Robertson, D. L. and G. F. Joyce. 1990. Selection *in vitro* of an RNA enzyme that specifically cleaves single-stranded DNA. Nature. 344:467-468.

39. Ronimus, R. S. and D. R. Musgrave. 1996. A gene, han1A, encoding an archaeal histone-like protein from the *Thermococcus* species AN1: Homology with eukaryal histone consensus sequences and the implications for delineation of the histone fold. Biochimica et Biophysica Acta. 1307:1-7.

40. Roth, S. Y., M. Shimizu, L. Johnson, M. Grunstein and R. T. Simpson. 1992. Stable nucleosome positioning and complete repression by the yeast a2 repressor are disrupted by amino-terminal mutations in histone H4. Genes and Devel. 6:411-425.

41. Sadler, L. A. and C. F. Brunk. 1992. Phylogenetic relationships and unusual diversity in histone H4 proteins within the *Tetrahymena pyriformis* complex. Mol. Biol. Evol. 9(1):70:84.

42. Sandman, K., J. A. Krzycki, B. Dobrinski, R. Lurz and J. N. Reeve. 1990. HMf, a DNA-binding protein isolated from the hyperthermophilic archaeon *Methanothermus fervidus*, is most closely related to histones. Proc. Natl. Acad. Sci. USA. 87:5788-5791.

43. Schmid, A., K. -D. Fascher and W. Hörz. 1992. Nucleosome disruption at the yeast PHO5 promoter upon PHO5 induction occurs in the absence of DNA replication. Cell. 71:853-864.

44. Sherman, F. 1991. Getting started with yeast. *In* Methods in Enzymology Vol. 194 (Guthrie, C. and Fink, G.R. eds.). Academic Press, Inc. San Diego.

45. Shwed, P. S., J. M. Neelin and V. L. Seligy. 1992. Expression of *Xenopus laevis* histone H5 gene in yeast. Biochimica et Biophysica Acta. 1131:152-160.

46. Szostak, J. W. 1992. In vitro genetics. Trends in Biochem. Sci. 17(3):89-93.

47. Tabassum, R., K. M. Sandman and J. N. Reeve. 1992. HMt, a histone-related protein from *Methanobacterium thermoautotrophicum* ΔH. J. Bacteriol. 174(24):7890-7895.

48. Thomm, M., K. Sandman, G. Frey, G. Koller and J. N. Reeve. 1992. Transcription *in vivo* and *in vitro* of the histone-encoding gene hmfB from the hyperthermophilic archaeon *Methanothermus fervidus*. J. Bacteriol. 174(11):3508-3513.

49. Travers, A. 1996. Building an initiation machine. Curr. Biol. 6:401-403.

50. Tuerk, C. and L. Gold. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science. 249:505-549.

51. Wang, B. -C., J. Rose, G. Arents and E. N. Moudrianakis. 1994. The octameric histone core of the nucleosome. J. Mol. Biol. 236:179-188.

52. Weintraub, H. and M. Groudine. 1976. Chromosomal subunits in active genes have an altered conformation. Science. 193:848-856.

53. Whitlock, J. P. and A. Stein. 1978. Folding of DNA by histones which lack their NH_2-terminal regions. J. Biol. Chem. 253(11):3857-3861.

54. Wolffe, A. 1994. Transcription: in tune with the histones. Cell. 77:13-16.

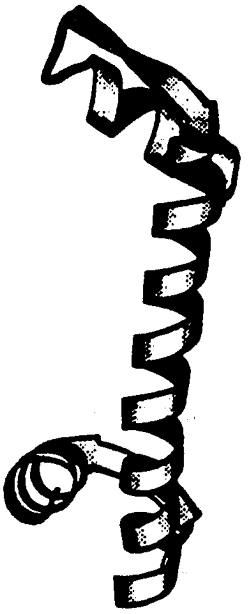
55. Wright, S. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Conference of Genetics*. 1:356-366.

56. Xie, X., T. Kokubo, S. L. Cohen, U. A. Mirza, A. Hoffman, B. T. Chait, R. G. Roeder, Y. Nakatani and S. K. Burley. 1996. Structural similarity between TAFs and the heterotetrameric core of the histone octamer. Nature. 380:316-322.

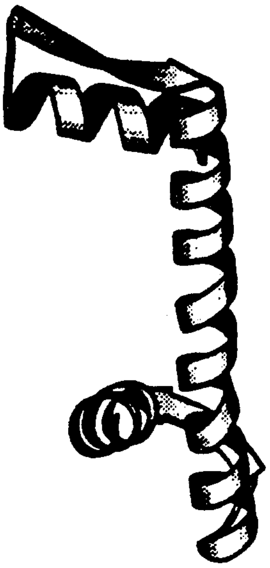
57. Zhang, B. and T. R. Cech. 1997. Peptide bond formation by *in vitro* selected ribozymes. Nature. 390:96-100.

Figure 1.1

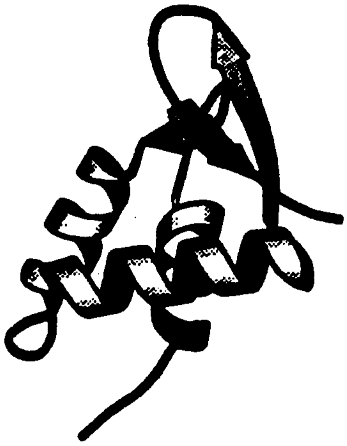
Histone Fold structures of histones H2A, H2B, H4, and H3 from *Gallus gallus* (chicken) represented clockwise from the upper left (adapted from (2)). The structure of the globular part of histone H1 is shown in comparison in the center of the figure. Despite numerous differences in primary amino acid sequence information, all four core histones have a remarkable similarity in terms of their tertiary structure of this central histone fold. The histone folds in H2B, H3, and H4 exhibit the greatest similarity of the four nucleosomal histones, whereas H2A shows an altered orientation of helix I.



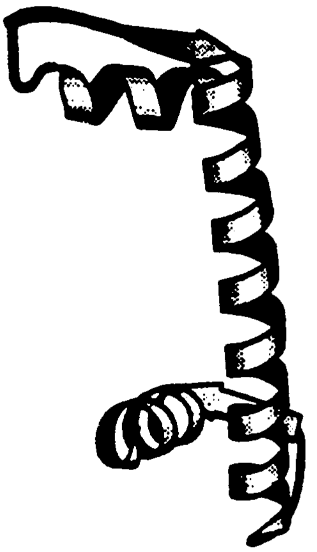
H2A



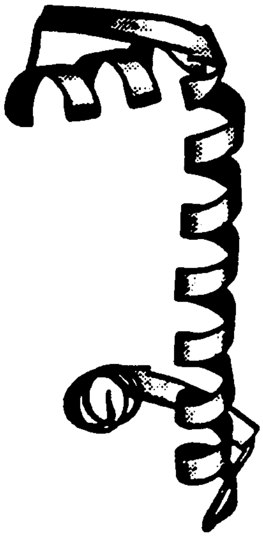
H2B



H1



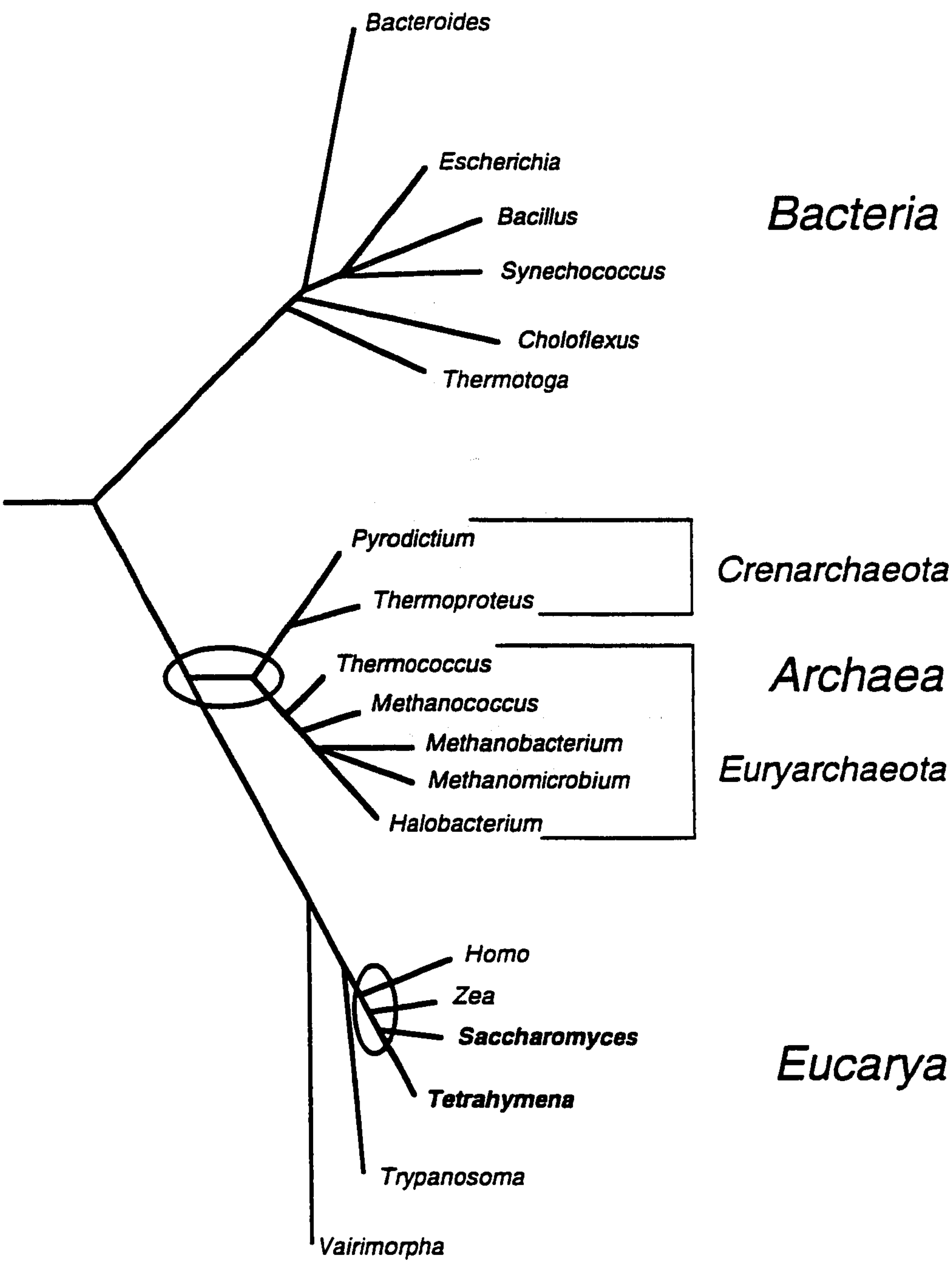
H3



H4

Figure 1.2

Phylogenetic tree of life derived from small-subunit rRNA sequence information (adapted from (31)). The tree was rooted following analysis of duplications in protein sequences (21). Areas of controversy or likely inaccuracies in the tree are noted by an oval. Only organisms within the Euryarchaeota and Eucarya are thought to contain homologous proteins utilizing a histone fold motif. The positions of the yeast *Saccharomyces* and ciliate *Tetrahymena* within the Eucarya are highlighted in bold font. The histone H4 proteins from these organisms are central to the focus of this study.


Bacteroides
Escherichia
Bacillus
Synechococcus
Choloflexus
Thermotoga
Bacteria
Pyrodictium
Thermoproteus
Crenarchaeota
Thermococcus
Methanococcus
Methanobacterium
Methanomicrobium
Halobacterium
Archaea
Euryarchaeota
Homo
Zea
Saccharomyces
Tetrahymena
Trypanosoma
Vairimorpha
Eucarya

Chapter 2

ELSEVIER

Biochimica et Biophysica Acta 1354 (1997) 116–126

BIOCHIMICA ET BIOPHYSICA ACTA

BBA

Expression of *Tetrahymena* histone H4 in yeast

Gary B. Fogel, Clifford F. Brunk *

Department of Biology, University of California, Los Angeles, 405 Hilgard Av., Los Angeles, CA 90095-1606, USA

Received 10 March 1997; accepted 16 April 1997

Abstract

Histone H4 is one of the most conserved proteins known. The very low rate of nonsynonymous substitution in H4 suggests that it fulfills an essential function in virtually all eukaryotes. While the majority of histone H4 sequences differ only slightly from the general consensus H4 sequence, yeast and *Tetrahymena* sequences diverge substantially from both the consensus and from each other. This study demonstrates that despite this divergence, when *Saccharomyces cerevisiae* cells are forced to use the *Tetrahymena thermophila* histone H4 protein, they are viable although they have a reduced growth rate, are temperature-sensitive relative to wild-type, have a lengthened G2 phase, and show a dramatic repression of mating. An amino acid replacement at position 33 in the protein improves the growth rate of these cells growing at temperatures above 28°C. This replacement changes a proline to a serine and is a further divergence from both the *Tetrahymena thermophila* and *Saccharomyces cerevisiae* histone H4 sequences. Thus, the replacement and expression of a non wild-type histone H4 in yeast offers measurable effects on cell growth, identifying amino acids required for optimal yeast functioning. © 1997 Elsevier Science B.V.

Keywords: Yeast; Histone; H4; Evolution; (*Tetrahymena*)

1. Introduction

In eukaryotic nuclei, DNA-nucleosome interactions form the basis of DNA compaction into chromatin. The nucleosome is composed of about 140 base pairs of DNA wrapped around a histone octamer containing two molecules each of histone H2A, H2B, H3, and H4, of which the H3–H4 dimer plays a central role [1,2]. Histone H4, a protein of 102 amino acids in length, is one of the most conserved proteins known. A very low rate of nonsynonymous substitution suggests that its function is essential and very similar in virtually all eukaryotes [3–5].

The functional interactions of histones with DNA and with other histones in the nucleosome are responsible for the extreme evolutionary conservation of the histones. It has been historically assumed that virtually any amino acid substitution in the core histones would be detrimental to the function of the nucleosome, and therefore not favored by natural selection. A difference of only two amino acids between calf and pea histone H4, prompted DeLange et al. [6] to suggest that the entire amino acid sequence was functionally necessary 'within narrow limits of tolerable alteration'. Of the histone H4 amino acid sequences currently determined (Genbank 96.0) [7], the majority are very similar, with yeast (*Saccharomyces* and *Schizosaccharomyces*) and ciliates (*Tetrahymena*, *Oxytrica* and *Glaucoma*) sequences representing substantial differences from the consensus H4

* Corresponding author. Fax: +1 (310) 206-3987; E-mail: cbrunk@ucla.edu

0167-4781/97/$17.00 © 1997 Elsevier Science B.V. All rights reserved.
PII S0167-4781(97)00078-X

sequence [8.9]. Another unusual H4 sequence from *Entamoeba* has recently been determined [10].

While the majority of histone H4 sequences differ from the consensus H4 sequence by less than 4%. *Saccharomyces cerevisiae* differs by 8% and *Tetrahymena thermophila* by 19%. Both *S. cerevisiae* and *T. thermophila* have two histone H4 genes (*HHF1* and *HHF2*). which each produce identical proteins [11.12]. When comparing *S. cerevisiae* and *T. thermophila* histone H4 proteins. there are 21 amino acid differences. one deletion and one insertion for a total of 23 differences. This represents a difference of nearly one-quarter of the protein. These changes occur throughout the protein. including the hydrophobic core. although deletions in the hydrophobic core are thought to be lethal [13].

Despite the generally high level of evolutionary conservation among histone H4 proteins. *S. cerevisiae* cells are viable with a large portion of the amino terminus of the protein deleted [14]. Cells with a deletion of residues 4–28 (or 2–26) are viable: however. they are not capable of mating and display temperature sensitivity relative to cells with the normal histone H4 gene [13.15]. Cells are also viable with a small deletion of residues 100–102. at the carboxy terminus of histone H4 [13]. Whitlock and Stein [16] determined that histones which lack their amino terminal regions retain the ability to fold DNA into a nucleoprotein complex. In *S. cerevisiae* and in many other eukaryotes. the first four lysine residues in histone H4 are subject to reversible acetylation. which is believed to play a critical role in histone-protein interactions necessary for chromatin function and gene regulation [17–20]. Cells continue to grow following amino terminal deletions which remove these lysine residues. suggesting these residues are not essential for cell viability [14]. However. if these lysine residues are replaced by arginines. which cannot be acetylated. the viability of the colonies is dramatically reduced [15.21]. Thus. the deletion of the histone H4 amino terminus appears to have less severe consequences for viability than the substitution of the first four lysine residues by arginine.

Tetrahymena histone H4 is unique for several reasons. It incorporates an unblocked alanine at its amino terminus instead of acetylserine [22]. and in the case of *Tetrahymena pyriformis*. H4 is known to have two primary sequence variants in the same cell [22]. *Tetrahymena* contain both macro- and micronuclei and some histones are known to be macronuclear specific (hv1 and hv2). However. neither of these macronuclear specific histones is thought to be closely related to histone H4 [22]. Macro- and micronuclei have similar relative amounts and types of histones. except that micronuclei contain three peptides that associate with linker regions of chromatin in the role of histone H1 [23]. However. no differences between the histone H4 of macro- and micronuclei have been detected in *T. thermophila*.

Previous experiments have demonstrated that *T. thermophila* histone H4 cannot replace calf histone H4 during nucleosome reassociation experiments in vitro. suggesting that the *T. thermophila* histone H4 protein may not be able to replace the endogenous H4 protein of another eukaryote [24.25]. Recent investigations have utilized yeast as a model eukaryote for the expression of distantly related histone genes. These studies have either expressed a non-endogenous histone in yeast [26–28]. replaced an endogenous yeast histone with a modified histone [13.14.30]. or replaced the yeast histone with a histone from another species [29]. Shwed et al. [26] studied the expression of *Xenopus laevis* linker histone H5 in yeast by placing the histone H5 gene into a yeast expression vector. Similarly. Linder and Thoma [27] and Miloshev et al. [28] tested the role of histone H1 by expression in yeast. Yeasts are believed not to contain either an endogenous H5 [26] or H1 [28]. making them appropriate systems for such experiments. Most relevant to our report is the work by Liu et al.. [29] who successfully replaced the H2A gene in yeast with the major H2A genes of *T. thermophila*. Although the yeast and *Tetrahymena* H2A genes have significant differences in their amino acid sequences. this study demonstrated the first trans-species complementation of histone function and provided clues to the evolution of the H2A.F/Z histone variant found in *Tetrahymena* by creating viable yeast cells.

The data reported here suggest that the replacement of the endogenous *S. cerevisiae* histone H4 gene with the *T. thermophila* histone H4 gene generates viable yeast cells with several phenotypic modifications. This replacement introduces a large number of amino acid substitutions over the entire protein. Several residues are not critical for viability suggesting that the potential variability in histone H4 may be much greater than the set of sequences currently represented in Genbank (release 96.0) [7]. Thus, the replacement and expression of a non wild-type histone H4 in yeast offers measurable effects on cell growth, identifying amino acids required for optimal yeast functioning.

2. Materials and methods

2.1. Plasmid construction

In order to replace the *S. cerevisiae HHF2* histone H4 gene with a *T. thermophila HHF1* histone H4 gene, a specially constructed *S. cerevisiae* strain (UKY403) capable of a 'glucose shift viability test' was used [13]. Both of the endogenous histone H4 genes of strain UKY403 have been replaced with selectable markers [13] and these cells contain a plasmid (pUK421), with a *S. cerevisiae* histone H4-2 gene under the regulation of the *GAL1* promoter as well as a tryptophan marker (*TRP1*) [30]. A plasmid derived from plasmid pUK499 containing the *T. thermophila HHF1* gene under the regulatory control of the *S. cerevisiae HHF2* promoter was constructed.

The plasmid pUK499 containing the yeast *HHF2* gene, (*URA3, CEN3, ARS1, Amp^r* and *Ori*) was used as the primary vector for this construction and has been used in previous experiments [13]. The genotype of UKY403 is: *MATa ade2-101(och) arg4-1 his3-Δ200 leu2-3 leu2-112 lys2-801 (amb) trp1-Δ901 ura3-52 thr- tyr- Δhhf1[HIS3] Δhhf2-[LEU2]/pUK421 (CEN3 ARS1 TRP1 GAL1/HHF2)* [13]. The *T. thermophila HHF1* gene was obtained from the genomic clone p508.8 [31]. The *S. cerevisiae HHF2* and *T. thermophila HHF1* genes have been completely sequenced [11,31] and are aligned in Fig. 1. Both genes have an *Msp* I site at the 5th nucleotide within the coding region. There is an *Sph* I site 44 nucleotides downstream from the termination codon in the *T. thermophila* gene and a compatible *Nla* III site 43 nucleotides downstream from the termination codon in the *S. cerevisiae* gene. The *S. cerevisiae HHF2* gene, from the *Msp* I site to the *Nla* III site, was replaced by the *T. thermophila HHF1* gene from the *Msp* I site to the *Sph* I site, creating a new plasmid pScTt (Fig. 2). The *Msp* I transition at the 5th nucleotide leaves the first amino acid in the protein as a serine rather than an alanine. Therefore, 20 amino acids differ between the *S. cerevisiae HHF2* gene and *T. thermophila HHF1* gene. Expression of the *T. thermophila* histone *HHF1* gene in this plasmid was under the control of the *S. cerevisiae HHF2* promoter.

2.2. Yeast transformation

The pScTt plasmid was used to transform *S. cerevisiae* strain UKY403 employing a lithium acetate yeast transformation protocol [32]. Transformants were selected on galactose media lacking uracil. The galactose induced the expression of the yeast *HHF2* gene on the pUK421 plasmid and the *URA3* marker on the plasmid pScTt provided the selection for a uracil requirement. A similar transformation and selection procedure was performed with pUK499 as a control to generate a yeast strain with the endogenous yeast *HHF2*, under the control of the *S. cerevisiae HHF2* promotor.

The resulting strains — ScTt cells, containing the pScTt plasmid and UKY499 cells, containing the pUK499 plasmid — were checked for growth on glucose media. Using the 'glucose shift viability test' [13], a shift from galactose media to glucose media repressed the histone *HHF1* on pUK421, which was under control of the *GAL1* promoter. Thus, the ScTt cells expressed only the *T. thermophila HHF1* gene while the UKY499 cells expressed only the *S. cerevisiae HHF2* gene.

To ensure that the ScTt cells expressed only the *T. thermophila HHF1* gene, cells that had lost the pUK421 plasmid were isolated. When ScTt cells were grown on glucose minimal media containing tryptophan for an extended period, the pUK421 plasmid was occasionally lost in spite of the fact that it contained a centromere [14]. Enrichment for cells that had lost the pUK421 plasmid was performed by extended growth in synthetic minimal media supplemented with 20 μg/ml tryptophan [33]. Cells lacking the pUK421 plasmid were identified by their requirement for tryptophan. The loss of the pUK421 plasmid was confirmed by Southern analysis of whole cell DNA using the yeast histone H4 sequence as a probe, the pScTt plasmid has an *Sph* I site at the end of the histone gene, while the pUK421 plasmid does

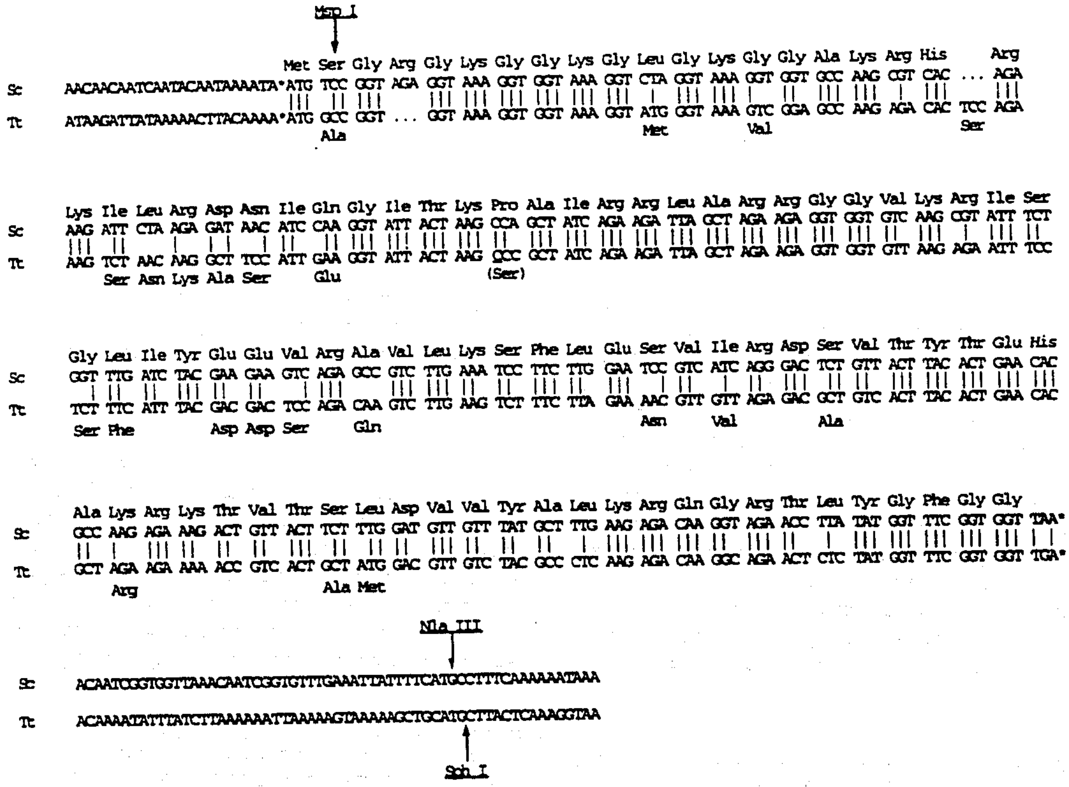


Fig. 1. The aligned DNA sequences for *S. cerevisiae* and *T. thermophila* histone H4 genes. The amino acid sequence of the *S. cerevisiae HHF2* gene is shown above the DNA sequences and differences in the amino acid sequence of the *T. thermophila HHF1* are shown below the DNA sequences. In the mutant gene, MT-1, the 97th nucleotide is a T (shown as *C*) and the corresponding amino acid is a serine. The positions of the *Msp* I, *Nla* III and *Sph* I restriction sites are indicated.

not (Fig. 1). This ensured that the *Tetrahymena* H4 was the only H4 present in the ScTt cells, only one copy of the endogenous yeast H4 was present in UKY499 cells, and that the requirement for tryptophan was derived by a loss of the plasmid rather than a mutation in the *TRP1* marker. A similar elimination of the pUK421 plasmid was performed with the UKY499 cells.

2.3. Growth of yeast

Unless otherwise noted, *S. cerevisiae* were grown in YPD media (1% bacto-yeast extract, 2% bacto-peptone, 2% dextrose, 2% bacto-agar) [33]. Transformed yeast were initially selected on galactose plates lacking uracil (0.67% bacto-yeast nitrogen base [without amino acids], 2% galactose, 2% bacto-agar, 20 μg/ml adenine sulfate, 20 μg/ml arginine, 30 μg/ml lysine, 200 μg/ml threonine, 30 μg/ml tyrosine, 20 μg/ml tryptophan) so that all transformed cells would form colonies. In subsequent assays, cells were grown on dextrose (YPD) plates. Using this 'glucose shift viability test', only the histone H4 gene on the introduced plasmid was expressed. Mutations in the *Tetrahymena* H4 gene that permitted growth at higher temperatures were selected from plates containing ScTt cells grown at 34°C.

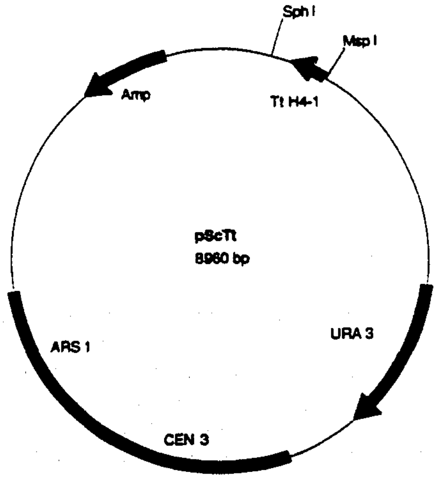


Fig. 2. The plasmid vector construct with the *T. thermophila* histone H4 gene replacing the *S. cerevisiae* histone H4 gene. This vector is a modification of pUK499 [13]. *Amp*, beta-lactamase: *ARS1*, autonomous replicating sequence: *CEN3*, centromere: *URA3*, orotidine-5'-phosphate decarboxylase.

2.4. *Relative growth rates*

To determine relative growth rates at various temperatures, equal numbers of ScTt and UKY499 were inoculated into a single shaking culture of YPD medium. Temperature during growth was controlled to ± 1°C. Growth was measured at 3, 6, 9, 12, and 15 generations. The relative abundance of ScTt and UKY499 strains was determined by Southern analysis [34]. Total cellular DNA was prepared and digested with *Sph* I and the *T. thermophila* histone H4-I sequence was used as a probe [35]. Plasmid pUK499 has a single *Sph* I site yielding a single fragment 9 kb in length, whereas pScTt has two *Sph* I sites and yields two fragments, one of which contains the *HHF2* gene (a fragment 1.3 kb in length). As *CEN3* regulated the plasmid copy number to essentially one copy per cell, the relative amount of each of these fragments was proportional to the number of ScTt and UKY499 cells. Lanes with known amounts of the pUK499 or pScTt on the Southern blot were used to normalize probe hybridization.

In a second assay for the relative growth of ScTt and UKY499 cells, the ratio of ScTt cells to UKY499 cells was determined by plating an aliquot of a mixed culture and incubating the plates at 28°C until small colonies could be counted. These plates were then transferred to 37°C and allowed to grow further. The ScTt colonies did not grow at 37°C, thus only the UKY499 colonies increased in size during the growth at 37°C. The ratio of large colonies to small colonies was proportional to the ratio of UKY499 to ScTt cells.

S. cerevisiae mating efficiencies were estimated by standard mating tests [32]. DNA was isolated [36] and sequences were determined using the double-strand sequencing protocol and the Sequenase kit (U.S. Biochemical).

3. Results

3.1. *Relative growth*

The yeast strains ScTt and UKY499 are isogenic with their sole difference being the histone H4 coding region on their respective plasmids. The *T. thermophila HHF1* probe used in Southern analysis hybridizes well with both the *S. cerevisiae* and *T. thermophila* histone H4 sequence, but should not hybridize with any of the regions flanking the *S. cerevisiae* histone gene. In Southern analyses, a single band was observed with a size corresponding to either the pScTt or pUK499 plasmids (data not shown). This indicated that the hybridization in these cells was exclusively with the plasmid introduced during transformation, as recombination of the histone H4 gene into the yeast genome would have given a different sized band after restriction with *Sph* I.

As the growth temperature was raised, it became apparent that the ScTt cells were temperature-sensitive relative to UKY499 cells (Fig. 3). ScTt had a maximum growth rate at about 28°C, while UKY499 had a maximum growth rate at about 32°C. The maximum growth rate of UKY499 was about 1.5 times that of ScTt. As the temperature decreased, the growth rate of ScTt approached that of UKY499 and the growth of both strains of yeast decreased dramatically. The growth rates at 10°C were so similar that it was very difficult to accurately measure a difference in separate cultures. Measurements were instead made

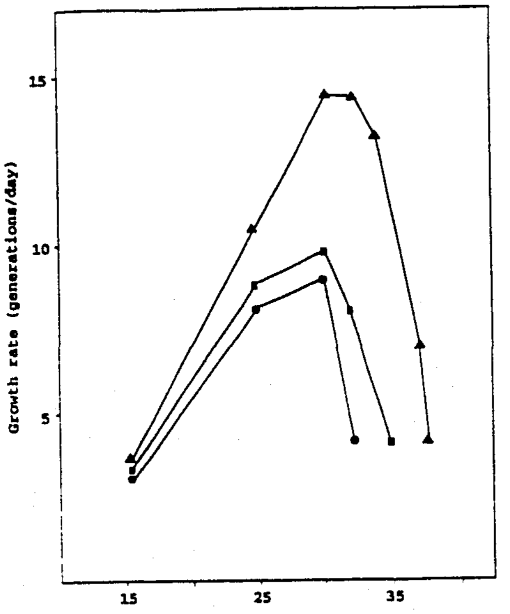


Fig. 3. The relative growth of *S. cerevisiae*, strains UKY499, MT-1, and ScTt as a function of temperature. ▲, UKY499; ■, MT-1; ●, ScTt cells.

in a single culture using Southern analysis. The relative intensities of the pScTt and pUK499 bands were determined by using densitometry analysis against known amounts of pScTt and pUK499 on the same Southern.

It is clear that even at this low temperature, the growth rate of UKY499 was higher than that of ScTt. The growth rate of UKY499 cells exceeded that of ScTt cells by 0.066 generations per day at 10°C. In 15 days, the UKY499 cells had undergone about one more generation than the ScTt cells.

3.2. Mating efficiency

The deletion of the amino terminus of the histone H4 gene of *S. cerevisiae* results in a dramatic decrease in the ability to mate [13,15]. The replacement of the histidine residue at position 18 by either tyrosine or glycine also depresses mating by 20- to 10000-fold [15,21]. We tested the mating ability of ScTt and UKY499 cells. On synthetic media containing dextrose, approximately equal amounts of UKY499, α*lys1*, and ScTt cells were independently mixed with a*his1* cells and grown at 28°C for 48 h. The mating of ScTt and a*his1* cells was not detectable, apparently being depressed greater than 1000-fold, whereas the mating of both UKY499 and α*lys1* cells with a*his1* appeared normal, suggesting that ScTt is deficient in mating ability. The amino acid substitutions in the *T. thermophila* histone H4 gene relative to the *S. cerevisiae* gene apparently interfered with yeast mating.

3.3. Cell cycle analysis

The deletion of the amino terminus portion of the wild-type *S. cerevisiae* histone H4 gene appears to lengthen the cell cycle by increasing the G2 phase [13,15]. We compared ScTt cells that were growing at 28°C and ScTt cells arrested at 37°C with UKY499 cells growing at 37°C. Table 1 shows the proportion

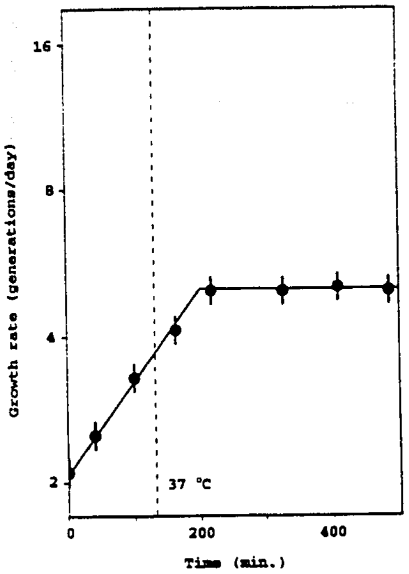


Fig. 4. The growth of ScTt cells following a shift to a non-permissive temperature (37°C). The vertical dotted line indicates the time of temperature shift. The error bars indicate the standard deviation of three measurements of growth. The culture was growing at 28°C prior to the temperature shift.

Table 1
Number of yeast cells in G2 and non-G2 phases and percentage of yeast cells in G2 phase at restrictive and non-restrictive temperatures

Cells	Temp. (°C)	G2	Non-G2	Total observed	% G2
UKY499	37	74	41	115	64
		74	53	127	58
		80	59	108	74
					65 ± 8
ScTt	28	98	19	117	84
		90	31	121	74
		80	39	119	67
					75 ± 8
ScTt	37	93	30	123	76
		84	22	106	79
		82	30	112	73
					76 ± 3

Cells were stained with DAPI and viewed by epifluorescence. Large budded cells with a single nucleus were counted as G2 cells [54]. Percentage of G2 cells in ScTt and UKY499 is statistically significant ($P > 0.05$) using the large sample test for proportions [55].

of cells in G2 and non-G2 phase for these cultures. Relative to UKY499, ScTt have a higher percentage of cells in G2 phase for both growing and arrested cells, indicating that an increase in the G2 phase contributes to the lengthened cell cycle in ScTt cells.

When ScTt cells were shifted from 28°C to a non-permissive temperature (37°C), growth continued at the normal rate for about 0.6 generations and then was abruptly arrested as shown in Fig. 4. When UKY499 cells were deprived of newly synthesized histone H4, they arrested after less than one round of cell division with greater than 90% of the cells in the G2 phase of the cell cycle [30]. However, the arrest of ScTt cells at the non-permissive temperature left the cells distributed throughout the cell cycle. Table 1 indicates that ScTt cells arrested at 37°C are not found in G2 phase at significantly higher proportions than cells growing at 28°C.

3.4. Analysis of a mutant with reduced temperature sensitivity

The temperature sensitivity of yeast cells with the *T. thermophila* histone H4 gene relative to cells with the *S. cerevisiae* histone H4 gene suggested that mutations in the *T. thermophila* histone H4 might possibly reduce the temperature sensitivity of this protein. Several plates with approximately 10^7 ScTt cells were grown at 34°C, which is a non-permissive temperature for ScTt cells. After several days (7–10) a few colonies formed. Fig. 3 shows the growth characteristics of one of these colonies, MT-1, compared to ScTt and UKY499 cells. MT-1 grew at higher temperatures than its parent strain, ScTt, and had a greater maximal growth rate. However, MT-1 was clearly temperature-sensitive relative to UKY499 cells. Apparently, MT-1 carried a mutation(s) relative to the ScTt cells that relieved some of the temperature sensitivity. This mutation probably resided in the histone H4 gene, but could have been localized to another gene (or genes).

In order to localize the mutation in MT-1 to the histone H4 gene, the plasmid containing the histone gene was isolated from MT-1 cells and introduced to yeast strain UKY403. When these transformed cells were grown on glucose, with the plasmid from MT-1 as the sole source of histone H4, they displayed a temperature sensitivity identical to MT-1. This indicated that the mutation(s) responsible for the temperature sensitivity resided on the MT-1 plasmid and was probably localized to the histone H4 gene.

The MT-1 plasmid was isolated and the sequence of the histone H4 gene was determined. This sequence was identical to the *T. thermophila HHF1* gene sequence (Fig. 1) except that the 97th nucleotide in MT-1 was a T rather than a C. This changed the 33rd amino acid in histone H4 from a proline to a serine. At this amino acid position, native histone H4 from *T. thermophila* and *S. cerevisiae* have a proline. Therefore, the amino acid replacement found in MT-1 represents a further divergence from both the *T. thermophila* and *S. cerevisiae* sequences. Apparently, this single amino acid replacement significantly improved the function of the *T. thermophila* histone H4 protein in *S. cerevisiae* cells growing at temperatures above 28°C.

4. Discussion

In spite of the high degree of evolutionary conservation of histone H4 proteins, we find that the *T.*

thermophila histone H4 will function in *S. cerevisiae*. This alteration represents a change of 20% of the amino acids. A further change in the sequence of the *T. thermophila* histone H4 (MT-1) improved its function in *S. cerevisiae*. This suggests that a wide spectrum of histone H4 sequences could function sufficiently well in *S. cerevisiae* to produce viable cells in the laboratory.

The replacement of as few as four amino acids in the amino terminus of the *S. cerevisiae* histone H4 gene has previously been shown to reduce cell viability dramatically [15]. Deletions in the hydrophobic core of this protein are also lethal [13]. However, the substitution of the entire *T. thermophila* histone H4 gene for the *S. cerevisiae* H4 gene yields viable cells. The differences between the *T. thermophila* and *S. cerevisiae* histone H4 protein include changes in the hydrophobic core as well as the hydrophilic amino terminus. It is remarkable how well the *T. thermophila* histone H4 protein functions in yeast, particularly in light of the inability of *T. thermophila* histone H4 to substitute for calf histone H4 in the formation of nucleosomes in vitro [24,25].

When forced to use the *T. thermophila* histone H4 gene with its associated amino acid substitutions, *S. cerevisiae* do have several measurable phenotypic differences from wild-type *S. cerevisiae*. Most notably their growth is sensitive to high temperature. The growth of these cells is arrested at 34°C; about 7°C lower than the arresting temperature for wild type *S. cerevisiae*. At lower temperatures (15–30°C), the growth rate of ScTt approaches that of UKY499, but even at 10°C, UKY499 cells have a higher growth rate.

The cell cycle of *S. cerevisiae* utilizing the *T. thermophila* histone H4 gene appears to be lengthened by an increase in the G2 phase. This represents a similar phenotypic change to *S. cerevisiae* after deletions in the amino terminus of histone H4 [13]. When *S. cerevisiae* cells are deprived of histone H4 synthesis by a 'glucose shift', > 90% of the cells are arrested in the G2 phase [30]. When ScTt cells are shifted to high temperature (37°C), they stop growth in less than one generation; however, the ScTt cells appear to stop throughout the cell cycle. The distribution of cells within the cell cycle is very similar for cultures at permissive temperatures (28°C) and cells arrested by high temperature (37°C). Apparently, they are not arresting at a specific phase of the cell cycle when subjected to high temperature.

There is a significant difference between depriving cells of new histone H4 synthesis and shifting to a non-permissive temperature. In ScTt cells, all of the chromatin contains temperature sensitive histone H4, so that a shift to higher temperature will affect the entire genome. This might be expected to affect all portions of the cell cycle. In *S. cerevisiae*, where new histone H4 synthesis is stopped by a 'glucose shift', all of the previously assembled chromatin should be functional. The requirement for additional histone H4 may be most acute in the G2 phase delaying the cells in this phase of the cell cycle.

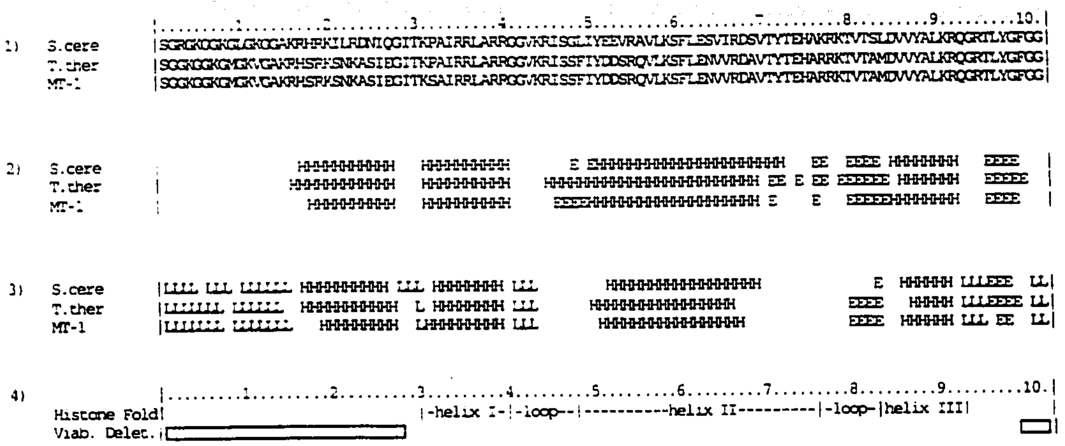
The phenotypic change observed in *S. cerevisiae* utilizing the *T. thermophila* histone H4 gene is most probably associated with impaired function of this foreign histone H4 in nucleosomes containing the histones H2A, H2B and H3 from *S. cerevisiae*. However, the foreign histone H4 may compromise autoregulation of histone H4 gene expression [37]. The translation of the *T. thermophila* histone H4 mRNA is unlikely to be disruptive as the codon utilization of *T. thermophila* and *S. cerevisiae* are very similar [38].

The mating of cells with the *T. thermophila* histone H4 gene is greatly depressed relative to wild-type *S. cerevisiae*. Deletions in the amino terminus, or even a substitution of the histidine at the 18th position in histone H4, substantially decreases the mating efficiency of wild-type *S. cerevisiae* cells [13,15,21]. There are a number of amino acid replacements in the amino terminus of the *T. thermophila* histone H4 gene relative to the *S. cerevisiae* histone H4 (Fig. 1). However, the insertion of a serine residue in the *T. thermophila* histone H4 gene immediately following the histidine at position 18 is the most probable candidate for the significant depression in mating efficiency. The loss of mating function in wild-type *S. cerevisiae* has been previously correlated to the de-repression of the silent mating loci *HMLα* and *HMR*a [13]. This portion of the histone H4 protein is thought to interact with a repressor of the silent loci, the *SIR* gene products, and the insertion of a serine residue in this region apparently disrupts this repression leading to the loss of mating function [13].

Recent X-ray crystallographic analysis of the nucleosomal core histone octamer has revealed a common structural element known as the histone fold [2,39–41]. Although only a low degree of primary structure similarity exists between the various histone types, the histone fold appears to be common to all core histones. Arents and Moudrianakis [41] suggested that the histone fold was common not only to eukaryotes but also occurs in the histone-related archaebacterial proteins HMf_B and HMt_B [42,43] and in other DNA binding proteins. Freeman et al. [44] suggested that the histone fold in H3 and H4 is responsible for nucleosomal integrity and functions essential for cell viability.

The three histone H4 sequences (from strains UKY499, ScTt, and MT-1) were compared using several measures of protein secondary structure, including Chou and Fasman alpha-helix and beta-sheet propensity measures [45], Eisenberg's hydrophobic moment [46], and Kyte and Doolittle's hydropathy [47]. These sequences were also compared using several protein secondary structure prediction algorithms, including NNPredict [48,49], PredictProtein [50,51], and SSP [52]. Not surprisingly, the secondary structure of all three sequences was very similar (Fig. 5). *T. thermophila* histone H4 has slightly more alpha-helix (52.4%) and beta-sheet (15.5%) relative to yeast H4 (49.5% alpha-helix, 11.7% beta-sheet). The mutation in MT-1 occurs in the helix I region of the histone fold and does not appear to change this helix. However, the predicted percentages of alpha-helix and beta-sheet in histone H4 from MT-1 (46.6% and 14.6%, respectively) are closer to yeast. Similarly, the percentage of looped domain for MT-1 was predicted to be the same as yeast H4 (38.8%), whereas *T. thermophila* histone H4 was predicted to be lower (32.0%). This suggests that although the primary sequence of MT-1 diverged from both *S. cerevisiae* and *T. thermophila*, the secondary sequence may have become more typical of *S. cerevisiae* histone H4. The amino acid substitution found in MT-1 appears to have had a 'global' effect on the protein secondary structure.

Replacement and expression of the *Tetrahymena* histone H4 in yeast offers measurable functional effects on cell growth, identifying amino acids required for optimal yeast viability as well as other positions that are not as constrained. A growth disadvantage occurs when yeast are forced to utilize *Tetrahymena* H4, a result that offers insight into protein function through the use of mutations in protein structure. The viability of yeast cells with histone H4 genes which



1)
|.........1.........2.........3.........4.........5.........6.........7.........8.........9.........10.|
S.cere |SGRGKGGKGLGKGGAKRHRKILRDNIQGITKPAIRRLARRGGVKRISGLIYEEVRAVLKSFLESVIRDSVTYTEHAKRKTVTSLDVVYALKRQGRTLYGFGG|
T.ther |SGGKGGKGMGKVGAKRHSRKSNKASIEGITKPAIRRLARRGGVKRISSFIYDDSRQVLKSFLENVVRDAVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG|
MT-1 |SGGKGGKGMGKVGAKRHSRKSNKASIEGITKSAIRRLARRGGVKRISSFIYDDSRQVLKSFLENVVRDAVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG|
2)
S.cere HHHHHHHHHHH HHHHHHHHH E EHHHHHHHHHHHHHHHHHHHHHH EE EEEE HHHHHHH EEEE
T.ther HHHHHHHHHHHH HHHHHHHHH HHHHHHHHHHHHHHHHHHHHHHHHH EE E EE EEEEEE HHHHHH EEEEE
MT-1 HHHHHHHHHH HHHHHHHHH EEEEHHHHHHHHHHHHHHHHHHHHHH E E EEEEEHHHHHHH EEEE
3)
S.cere |LLLL LLL LLLLLL HHHHHHHHHH LLL HHHHHHHH LLL HHHHHHHHHHHHHHHHHH E HHHHHH LLLEEE LL|
T.ther |LLLLLLL LLLLLL HHHHHHHHHHH L HHHHHHHH LLL HHHHHHHHHHHHHHHHH EEEE HHHHH LLLEEEE LL|
MT-1 |LLLLLLL LLLLLLL HHHHHHHHH LHHHHHHHHH LLL HHHHHHHHHHHHHHHHH EEEE HHHHHH LLL EE LL|
4)
|.........1.........2.........3.........4.........5.........6.........7.........8.........9.........10.|
Histone Fold| |-helix I-|-loop--|----------helix II---------|-loop-|helix III| |
Viab. Delet.|

Fig. 5. Secondary structure predictions for histone proteins derived from PredictProtein [50]. (1) Alignment of the three histone H4 sequences used in this study (*S. cerevisiae*, *T. thermophila*, MT-1). (2) Predicted positions of helix (H) and strand (E) secondary structure relative to aligned sequences based on the PHDsec program. (3) Predicted positions of looped domains (L), helix (H) and strand (E) secondary structure based on the SUBsec program. (4) Proposed histone fold [42] relative to aligned sequences and viable deletions to yeast histone H4 [13].

have substantial deletions in the amino terminus or with foreign histone genes (as is the case presented here) suggests that considerable greater latitude in histone H4 sequence can be tolerated than is commonly appreciated. The growth of yeast utilizing altered histone H4 genes in a laboratory culture is. however. not the same as the selection of histone H4 sequences over thousands of generations. Nonetheless, we have found that there is more variation in the sequences of histone H4 genes from different species of *Tetrahymena* than is found among all of the metazoan histone H4 sequences published [53]. It is possible that single-celled eukaryotes have a greater latitude in these sequences. However. it is probable that histone H4 sequences will accept a wide spectrum of substitutions that leave their common structural element. the histone fold. intact. Future mutagenesis experiments will help to clarify this issue.

Acknowledgements

The yeast strains and vectors were provided by Dr. M. Grunstein. A plasmid containing the *Tetrahymena* histone H4 gene was provided by Dr. M. Gorovsky. We gratefully acknowledge helpful advice and discussions with Drs. M. Grunstein. R. Bohman. and D. Fogel. The authors would like to thank three anonymous reviewers for their helpful suggestions. This project was supported by a grant from the Academic Senate of the University of California. Los Angeles. G. Fogel is a Fellow of the Center for the Study of the Evolution and Origin of Life at UCLA.

References

[1] T.J. Richmond. J.T. Finch. B. Rushton. D. Rhodes. A. Klug. Nature 311 (1984) 532–537.

[2] G. Arents. R.W. Burlingame. B.-C. Wang. W.E. Love. E.N. Moudrianakis. Proc. Natl. Acad. Sci. USA 88 (1991) 19148–19152.

[3] C.V.C. Glover. M.A. Gorovsky. Proc. Natl. Acad. Sci. USA 76 (1979) 585–589.

[4] W.-H. Li. D. Graur. Fundamentals of Molecular Evolution. Sinauer Associates. Sunderland. MA. 1991.

[5] T.H. Thatcher. M.A. Gorovsky. Nucl. Acids Res. 22 (1994) 174–179.

[6] R.J. DeLange. D.M. Fambrough. E.L. Smith. J. Bonner. J. Biol. Chem. 244 (1969) 5669–5679.

[7] D. Benson. M. Boguski. D.J. Lipman. J. Ostell. Nucl. Acids Res.. [GenBank Release 96.0] 22 (1994) 3441–3444.

[8] D. Wells. C. McBride. Nucl. Acids Res. 17 (1989) r311–347.

[9] D.S. Harper. C.L. Jahn. Gene 75 (1989) 93–107.

[10] M. Binder. S. Ortner. B. Plaimauer. M. Födinger. G. Wiedermann. O. Scheiner. M. Duchêne. Mol. Biochem. Parasit. 71 (1995) 243–247.

[11] M.M. Smith. O.S. Andrésson. J. Mol. Biol. 169 (1983) 663–669.

[12] S. Horowitz. J.K. Bowen. G.A. Bannon. M.A. Gorovsky. Nucl. Acids Res. 15 (1987) 141–160.

[13] P.S. Kayne. U.-J. Kim. M. Han. J.R. Mullen. F. Yoshizaki. M. Grunstein. Cell 55 (1988) 27–39.

[14] L.K. Durrin. R.K. Mann. P.S. Kayne. M. Grunstein. Cell 65 (1991) 1023–1031.

[15] P.C. Megee. B.A. Morgan. B.A. Mittman. M.M. Smith. Science 247 (1990) 841–845.

[16] J.P. Whitlock. A. Stein. J. Biol. Chem. 253 (1978) 3857–3861.

[17] D. Doenecke. D. Gallwitz. Mol. Cell. Biochem. 44 (1982) 113–128.

[18] V.G. Norton. K.W. Marvin. P. Yau. E.M. Bradbury. J. Biol. Chem. 265 (1990) 19848–19852.

[19] B.M. Turner. A.J. Birley. J. Lavender. Cell 69 (1992) 375–384.

[20] L. Hong. G.P. Schroth. H.R. Matthews. P. Yau. E.M. Bradbury. J. Biol. Chem. 268 (1993) 305–314.

[21] L.M. Johnson. P.S. Kayne. E.S. Kahn. M. Grunstein. Proc. Natl. Acad. Sci. USA 87 (1990) 6286–6290.

[22] M.A. Gorovsky. in: J.G. Gall (ed.). The Molecular Biology of Ciliated Protozoa. Academic Press. San Diego. 1986.

[23] C.D. Allis. C.V.C. Glover. M.A. Gorovsky. Proc. Natl. Acad. Sci. USA 76 (1979) 4857–4861.

[24] C.V.C. Glover. M.A. Gorovsky. Biochemistry 17 (1978) 5705–5713.

[25] M.A.C. Gorovsky. C. Glover. C.A. Johmann. J.B. Keevert. D.J. Mathis. M. Samuelson. Quant. Biol. 42 (1978) 493–503.

[26] P.S. Shwed. J.M. Nellin. V.L. Seligy. Biochimica et Biophysica Acta 1131 (1992) 152–160.

[27] C. Linder. F. Thoma. Mol. Cell. Biol. 14 (1994) 2822–2835.

[28] X. Liu. J. Bowen. M.A. Gorovsky. Mol. Cell. Biol. 16 (1996) 2878–2887.

[29] G. Miloshev. P. Venkov. K. van Holde. J. Zlatanova. Proc. Natl. Acad. Sci. USA 91 (1994) 11567–11570.

[30] U.-J. Kim. M. Han. P. Kayne. M. Grunstein. EMBO J. 7 (1988) 221–229.

[31] G.A. Bannon. J.K. Bowen. M.-C. Yao. M.A. Gorovsky. Nucl. Acids Res. 12 (1984) 1961–1975.

[32] F. Sherman. G.R. Fink. J.B. Hicks. Laboratory Course Manual for Methods in Yeast Genetics. Cold Spring Harbor Laboratory Press. Cold Spring Harbor. 1986.

[33] F. Sherman. in: C. Guthrie. G.R. Fink (eds.). Guide to Yeast Genetics and Molecular Biology. Academic Press. San Diego. 1991.

[34] E.M. Southern. J. Mol. Biol. 98 (1975) 503–517.

[35] J. Sambrook. E.F. Fritsch. T. Maniatis. Molecular Cloning:

a Laboratory Manual. 2nd ed. Cold Spring Harbor Laboratory Press. Cold Spring Harbor. 1989.

[36] P. Philippsen. A. Stotz. C. Scherf. in: C. Guthrie. G.R. Fink (eds.). Guide to Yeast Genetics and Molecular Biology. Academic Press. San Diego. 1991.

[37] R.S. McLaren. J. Ross. J. Biol. Chem. 20 (15) (1993) 14637–14644.

[38] D.W. Martindale. J. Protozool. 36 (1) (1989) 29–34.

[39] G. Arents. E.N. Moudrianakis. Proc. Natl. Acad. Sci. USA 90 (1993) 10489–10493.

[40] B.-C. Wang. J. Rose. G. Arents. E.N. Moudrianakis. J. Mol. Biol. 236 (1994) 179–188.

[41] G. Arents. E.N. Moudrianakis. Proc. Natl. Acad. Sci. USA 92 (1995) 11170–11174.

[42] R. Tabassum. K.M. Sandman. J.N. Reeve. J. Bacteriol. 174 (1992) 7890–7895.

[43] R.A. Grayling. K. Sandman. J.N. Reeve. FEMS Microbiol. Rev. 18 (1996) 203–213.

[44] L. Freeman. H. Kurumizaka. A.P. Wolfe. Proc. Natl. Acad. Sci. USA 93 (1996) 12780–12785.

[45] P.Y. Chou. G.D. Fasman. Adv. Enz. 47 (1978) 45–147.

[46] D. Eisenberg. R.M. Weiss. T.C. Terwilliger. Proc. Natl. Acad. Sci. USA 81 (1984) 140–144.

[47] J. Kyte. R.F. Doolittle. J. Mol. Biol. 157 (1982) 105–132.

[48] J.L. McClelland. D.E. Rumelhart. Explorations in Parallel Distributed Processing. Vol. 3. MIT Press. Cambridge. MA. 1988. pp. 318–362.

[49] D.G. Kneller. F.E. Cohen. R. Langridge. J. Mol. Biol. 214 (1990) 171–182.

[50] B. Rost. Meth. in Enzymol. 266 (1996) 525–539.

[51] B. Rost. R. Casadio. P. Fariselli. in: D. States et al. (eds.). ISMB '96. AAAI Press. Menlo Park. 1996.

[52] V.V. Solovyev. A.A. Salamov. CABIOS 10 (1994) 661–669.

[53] C.F. Brunk. R.W. Kahn. L.A. Sadler. J. Mol. Evol. 30 (1990) 290–297.

[54] J.R. Pringle. L.H. Hartwell. in: J.N. Strathern. E.W. Jones. J.R. Broach (eds.). The Molecular Biology of the Yeast *Saccharomyces cerevisiae*. Life Cycle and Inheritance. Cold Spring Harbor Laboratory Press. Cold Spring Harbor. 1981. pp. 97–142.

[55] J.L. Devore. Probability and Statistics for Engineering and the Sciences. 4th ed. Duxbury Press. Belmont. 1995. 376 pp.

Chapter 3

TEMPERATURE GRADIENT FOR RELATIVE GROWTH RATE ANALYSIS OF YEAST

ABSTRACT

Relative growth is often used as a phenotypic measure to distinguish mutant and wild-type yeast or bacterial strains. Differential growth as a function of temperature is a convenient and accurate means of analyzing differences between strains. Slight differences in the genotypes of two strains frequently results in differential growth as a function of temperature. We have developed a chamber for the simultaneous growth of multiple strains in microtiter plates along a temperature gradient. Image analysis was used to determine colony area and number at various times as a function of temperature. This chamber reduces the time required to measure growth as a function of temperature. This occurs by allowing relative growth to be measured along a temperature gradient where all other conditions are constant. Two strains of yeast (*Saccharomyces cerevisiae*) with a known difference in temperature dependence of growth were used to demonstrate the performance of this chamber.

INTRODUCTION

Temperature sensitivity is a common phenotypic trait used to identify and characterize novel mutant strains (1-3). Typical growth techniques utilize either liquid cultures or petri dishes containing cells placed in incubators each at a fixed temperature. An accurate growth profile requires a large number of incubators and a series of time-consuming measurements. A single device for the parallel growth of multiple strains, where all conditions are identical except for the temperature range, has distinct advantages.

We have developed an apparatus capable of generating a temperature gradient for the simultaneous growth of multiple strains over a wide range of temperatures using common microtiter plates. This chamber can be used with either yeast or bacteria; however, the material presented here focuses on yeast growth. During the incubation period, a picture of each well was captured by a video camera and stored in a computer. Image analysis has been previously used to examine yeast cell volumetric response to osmotic shifts (4), on-line measurement of cell size (5), and for morphological characterization (6). In our system, image analysis was used to measure the colony number and relative area in each well at various times. Using these techniques, a growth profile over a wide temperature range can easily and accurately be determined for each strain.

As a test of the temperature gradient chamber, we analyzed two yeast strains (UKY499 and ScTt) that displayed slightly different growth responses to temperature due to a modification in the yeast histone H4 gene (3). These strains are isogenic except for a difference in the histone H4 genes. Each strain has only a single histone H4 gene, present on a plasmid. The endogenous yeast histone H4 genes have been replaced with selectable markers. The histone H4 proteins in UKY499 and ScTt differ at 23 positions, resulting in a shift of the optimal growth temperature from 33°C in UKY499 to 30°C in ScTt (3). The temperature gradient chamber was used to accurately reproduce this phenotypic response much more quickly than conventional techniques.

MATERIALS AND METHODS

Temperature Gradient Chamber

The cool end of the temperature gradient chamber was connected to a refrigerated water bath (Lauda K-2/R; Brinkmann Instruments) set at a relatively low temperature (10°C). A peristaltic pump (Masterflex pump with a Solid State speed control unit; Cole-Parmer Instrument Co.) was used to circulate cold water from the bath through a copper block at the cool end of the chamber, maintaining the cool end of the temperature gradient at about 20°C (Figure 3.1). The warm end of the temperature gradient chamber was heated electrically to about 50°C. Insulated nichrome wire was wound within a copper block at the warm end of the chamber and the wire was heated with about 75 watts of electrical power, supplied by a variable transformer (Variac; Superior Electric Co.).

The entire growth chamber was surrounded by thick blocks of polyurethane foam (3 cm) for insulation. A linear temperature gradient ($y = 13.218x - 334.77$, $r^2 = 0.999$) of 26°C to 44°C was produced by heat flow from the warm copper block to the cool copper block through the aluminum walls, floor, and lid of the chamber (each 1 cm thick). Temperature was measured with a digital thermometer accurate to 0.1°C (GTH 1160; Greisinger Electronic). The temperature gradient chamber holds six 96-well microtiter plates, arranged with three plates placed on the floor of the chamber with the 12 well sides adjacent to one another. A second level of plates can be placed in an upper layer. Each layer has 24 wells along the temperature gradient and 12 wells across the temperature gradient.

Therefore, 24 strains can be analyzed simultaneously (12 in each of two layers) with 24 well positions along the temperature gradient. The temperature gradient is about 0.7°C per well. Ports in the lid provide gas exchange with the chamber.

Preparation of Yeast Strains

Each strain of yeast (*S. cerevisiae*) was grown from individual colonies to log phase in separate liquid cultures and briefly sonicated to avoid clumping before determination of cell density (Celloscope; Particle Data Inc.). The cells were grown embedded in a low-percentage agar (0.2%) so that colonies were well separated and would not settle. Equal parts of YPD media (7) containing 2000 cells/ml at 28°C and YPD media containing 0.4% bacto-agar at 55°C (Difco) were thoroughly mixed and 200µl aliquots were delivered into the wells of sterile 96-well microtiter plates (Corning) using an octapipetman. The microtiter plates were cooled for 30 minutes at 4°C to solidify the agar before the cells could settle. A cap of mineral oil (50µl) was added to each well to avoid desiccation. Yeast cells grew well under these conditions, suggesting that the agar environment was aerobic. *S. cerevisiae* is capable of anaerobic growth only in the presence of added sterol and unsaturated fatty acids (8).

Analysis of Yeast Relative Growth

After 18 hours of growth, small colonies were observed when viewed through the bottom of the wells at 20x magnification. The addition of mineral oil did not interfere with optics when viewing the colonies from below. Quantitative analysis was performed by collecting images of the colonies using a video camera (Javelin MOS Solid State). The images were stored and analyzed using a computer (Macintosh Quadra 840AV) and image software (NIH Image; version 1.61, developed at the U.S. National Institutes of Health and available on the internet at http://rsb.info.nih.gov/nih-image/). Colony area (projected area) was determined as pixel area using the Density Slice option from NIH Image 1.61 software and converted to mm^2.

The number of colonies in each well and average colony area for each well were determined after various incubation times (23, 28, 32, 48, and 73 hours). The average colony area was determined only for colonies that were clearly non-overlapping with any other colony in the image, with five or more colony areas being used to determine an average. Colony volume, which is proportional to the number of cells in the colony (particularly at small colony area), is equal to 0.752 x $[\text{colony area}]^{1.5}$. For our analysis, colony area was used as a measured metric for growth rate. The average area of the ScTt colonies (in mm^2) was subtracted from the average area of the UKY499 colonies and this difference was multiplied by the difference between the fraction of UKY499 cells that formed colonies and the fraction of ScTt cells that formed colonies. The resulting value was multiplied by 1,000 to produce the metric "delta" (Δ).

RESULTS AND DISCUSSION

After 18 hours, small colonies were observed for strain UKY499 in the range between 22°C and 36°C. However, colonies were not detected for strain ScTt. Differential growth of these strains as a function of temperature between 18 and 48 hours was dramatic (Figure 3.2). After two days of growth, both strains had reached essentially the same maximal colony area in each of the wells that had colonies. Figure 3.2 shows the colony area as a function of temperature at one day intervals over a three day period. These profiles are very similar to the growth rate profiles, previously determined, using liquid cultures in a shaking water bath (3). In figure 3.3, the fraction of cells which form colonies is shown as a function of temperature. The product of the difference in colony area and difference in fraction of cells forming colonies (Δ) for strains UKY499 and ScTt as a function of temperature is displayed in figure 3.4. This metric is a sensitive measure of differential growth as a function of temperature, particularly if the measurement is performed at a time when the colony area is still increasing. Figure 3.4 clearly indicates that the greatest differential growth of strains UKY499 and ScTt occurs between 33°C and 35°C. This approach provides a powerful tool for characterizing different strains.

The growth chamber allows simultaneous analysis of multiple yeast strains along a temperature gradient and can be used to accurately measure even slight differences in the relative growth rates at different temperatures. The temperature gradient chamber produces a detailed growth profile after only two days of incubation. The temperature gradient remains constant on a daily basis assuming constant flow rates at both ends of the device. Similar measurements performed on plates or in liquid culture would require many incubators or water baths and take much more time. In addition, the chamber can easily be used to produce a wide variety of temperature gradients, making it useful for the characterization of a diverse spectrum of cells displaying an anchorage-independent growth phenotype (9). The temperature chamber could also be used for the study of colony morphology of bacterial strains (10) and to monitor the effect of temperature on yeast autolysis (11).

REFERENCES

1. Archambault, J., Jansma, D. B. and Friesen, J. (1996) *Genetics.* **142**, 737-747.

2. Uemura, H., Pandit, S., Jigami, Y., and Sternglanz, R. (1996) *Genetics.* **142**, 1095-1103.

3. Fogel, G. B. and Brunk, C. F. (1997) *Biochem. et Biophys. Acta.* **1354**, 116-126.

4. Berner, J. -L. and Gervais, P. (1994) *Biotech. and Bioengineer.* **43**, 165-170.

5. Yamashita, Y., Kuwashima, M., Nonaka, T., and Suzuki, M. (1993). *J. of Chem. Eng. of Japan.* **26**, 615-619.

6. Pons, M. N., Vivier, H., Rémy, J. F., and Dodds, J. A. (1993) *Biotech. and Bioengineer.* **42**, 1352-1359.

7. Sherman, F. (1991) *in* Methods in Enzymology, Vol. 194 (Guthrie, C. and Fink, G. R. Eds.), pp. 3-21, Academic Press, New York.

8. Verduyn, C., Postma, E., Scheffers, W. A. and van Duken, J. P. (1990) *J. Gen Microbiol.* **136**, 395-403.

9. Lewis, J. C. and Milo, G. E. (1990) *Teratogenesis, Carcinogenesis, and Mutagenesis.* **10**, 351-357.

10. Mitchell, A. J. and Wimpenny, J. W. T. (1997) *J. Appl. Microbiol.* **83**, 76-84.

11. Orban, E., Quaglia, G. B. and Casini, I. (1994) *J. Food Eng.* **21**, 245-261.

Figure 3.1

Schematic view of the temperature gradient chamber. Cool water from a refrigerated circulating water bath maintains the left copper block at approximately 20°C, while the right copper block is electrically heated to approximately 50°C. Heat flow through the aluminum sides (not shown), base, and lid (each 1 cm thick) establishes a temperature gradient. The chamber will accommodate six microtiter plates. The aluminum lid has ports for gas exchange and is held down by four wing nuts which facilitate easy removal after incubation.

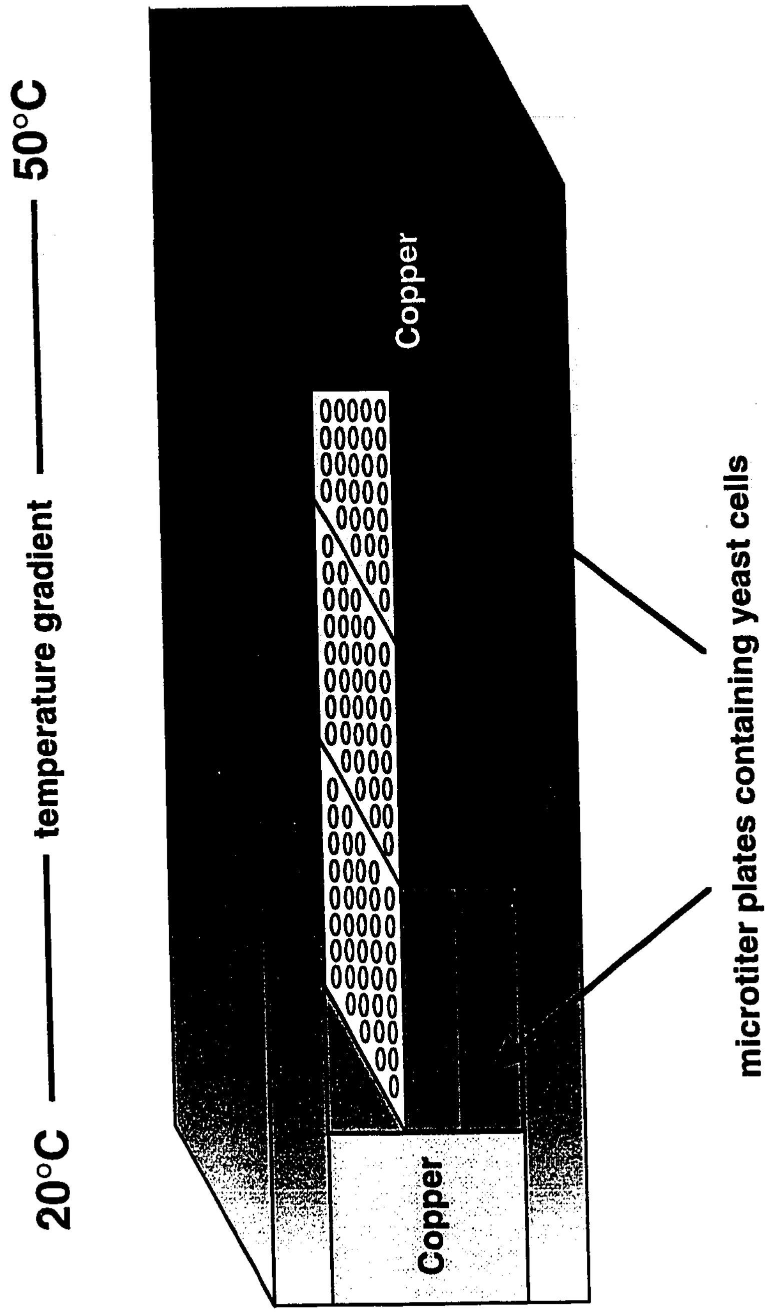

20°C
temperature gradient
50°C
Copper
Copper
microtiter plates containing yeast cells

Figure 3.2

Relative growth rates of yeast strains UKY499 (solid squares) and ScTt (open squares) in the temperature gradient chamber. The average colony area in each well as a function of temperature is shown (error bars show one standard deviation). Strain ScTt has a lower optimal growth temperature than strain UKY499 (3).

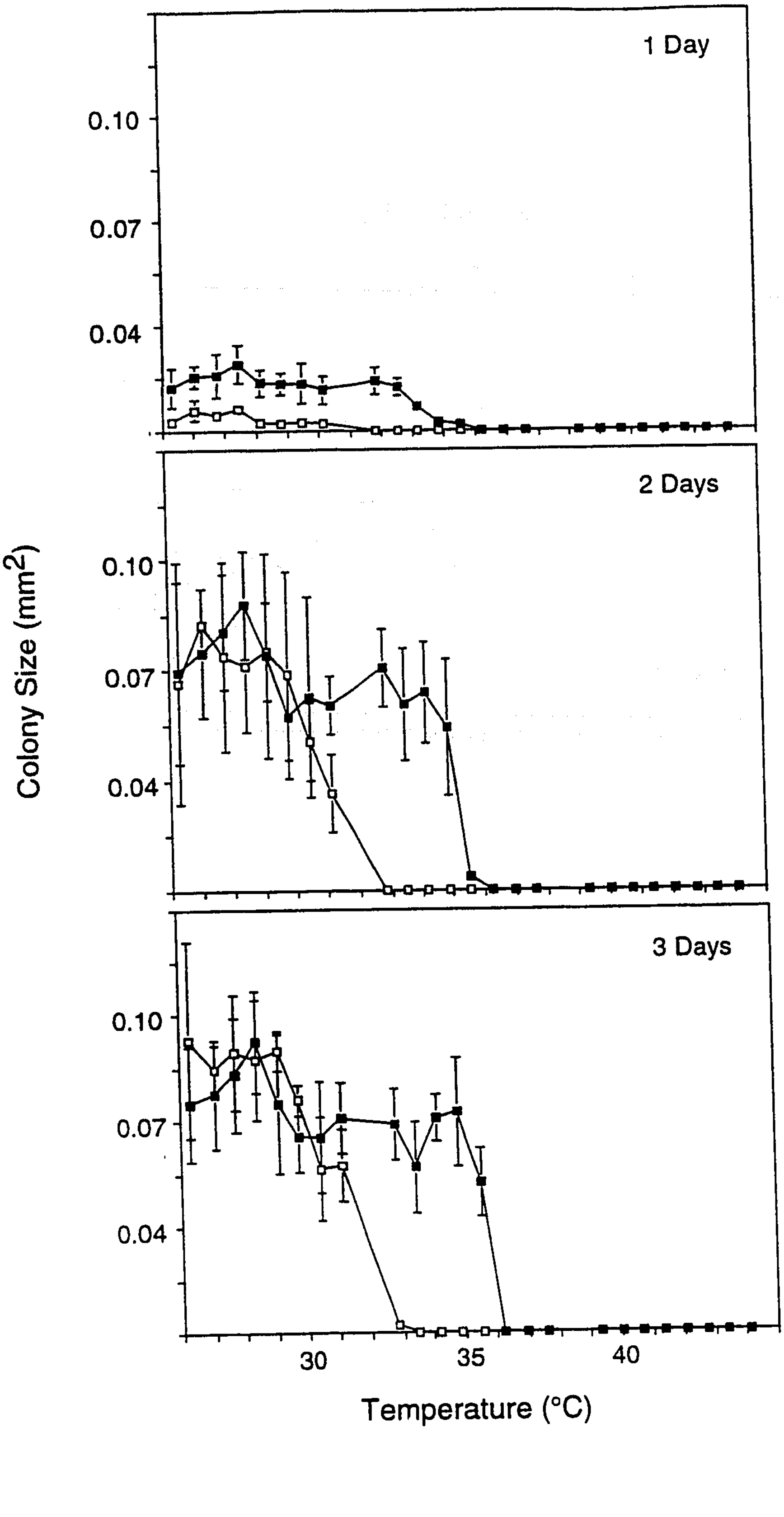

1 Day
2 Days
3 Days
0.10
0.07
0.04
Colony Size (mm^2)
30
35
40
Temperature (°C)

Figure 3.3

Fraction of cells which form colonies as a function of temperature. 200 cells are placed in each well and the fraction that form colonies is shown as a function of temperature (strain UKY499 (solid squares), ScTt (open squares)).

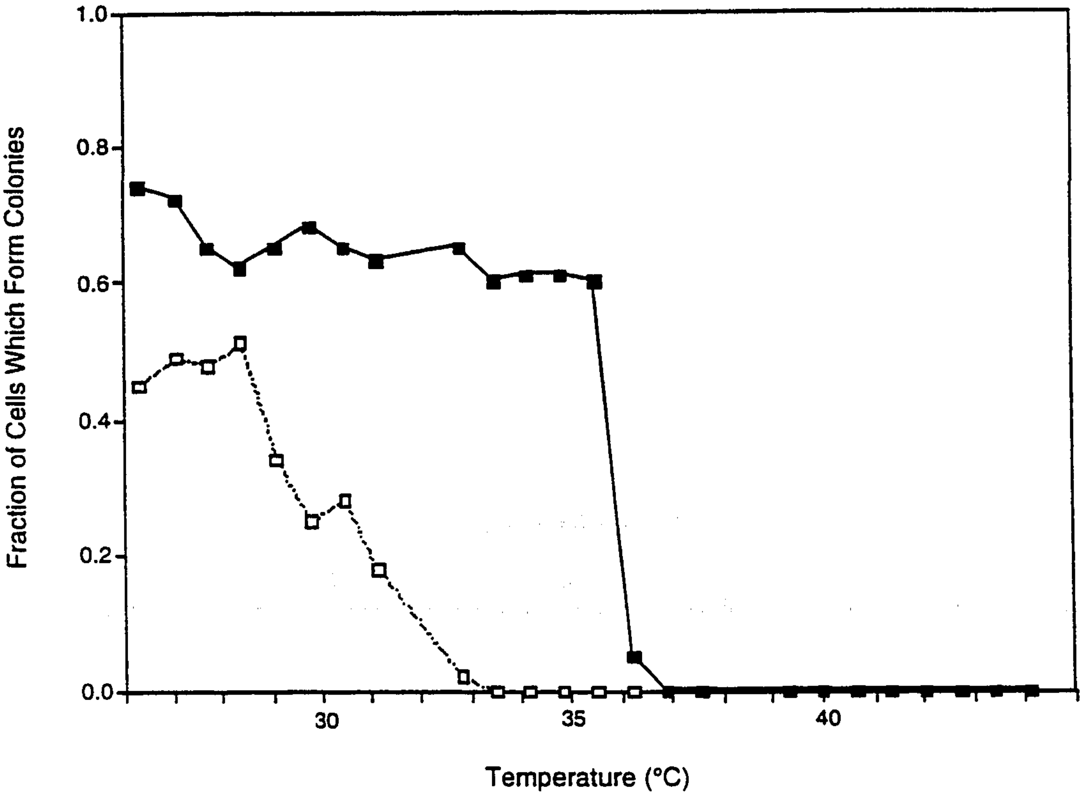
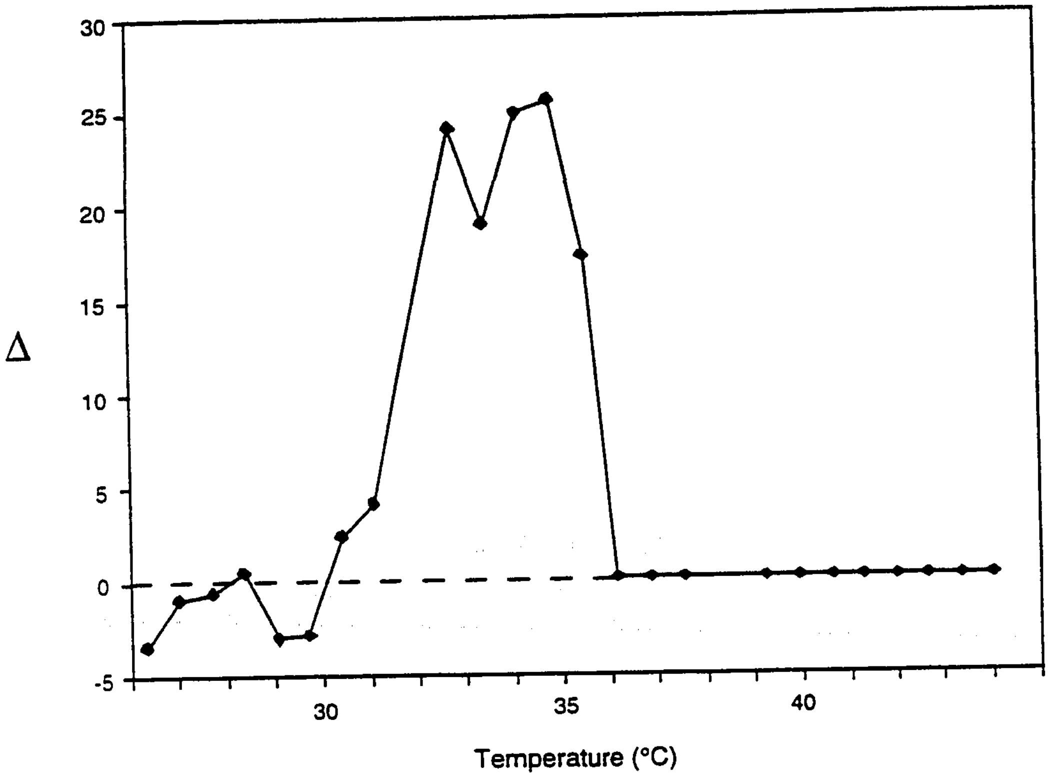

1.0
0.8
0.6
0.4
0.2
0.0
Fraction of Cells Which Form Colonies
30
35
40
Temperature (°C)

Figure 3.4

The product of the difference in colony area and difference in fraction of cells forming colonies (Δ) as a function of temperature. The average area of the ScTt colonies is subtracted from the average area of the UKY499 colonies and this difference is multiplied by the difference between the fraction of UKY499 cells that form colonies and the fraction of ScTt cells that form colonies. Δ shown multiplied by 1,000.


30
25
20
15
10
5
0
-5
Δ
30
35
40
Temperature (°C)

Chapter 4

OLIGONUCLEOTIDE MUTAGENESIS OF A UNIVERSALLY CONSERVED REGION IN *SACCAROMYCES CEREVISIAE* HISTONE H4

ABSTRACT

Using oligonucleotide cassette mutagenesis, a series of modifications were made to 3 arginines and 1 lysine residue (an arginine-arginine pair and an arginine-lysine pair) in a universally conserved region in the core of histone H4. These modified regions were examined for their viability in yeast cells in the context of both the *Saccharomyces cerevisiae* and *Tetrahymena thermophila* histone H4 sequences. For the first cassettes, each of the 4 positions were modified with 4 possible residues, generating a space of 256 possible sequences. From this space, 13 sequences were viable in the context of the *S. cerevisiae* histone H4 sequence and 16 were viable in the *T. thermophila* context. Of the 16 viable modifications in the *T. thermophila* context, all 16 were combinations of arginine and lysine only. Of the 13 sequences viable in the *S. cerevisiae* context, 5 were the result of arginine or lysine replacements only and 8 utilized other amino acids in combination with arginine or lysine. A second series of cassettes with only arginines and lysines at these four positions were generated in each context and confirmed the results of the first cassettes. These results suggest that: 1) a universally conserved region in histone H4 can be modified and still yield viable yeast cells, 2) the viability of modifications to a particular region of a protein are affected by other amino acid residues in the protein, and 3) in some positions, positively charged residues can be replaced by polar or non-polar residues in the core of histone H4.

INTRODUCTION

Histone H4 is one of the most conserved proteins known. From an alignment of all known histone H4 amino acid sequences (N=56), a universally conserved region between residues 29 and 49 is observed (2). No amino acid differences are found in this region. Such absolute conservation suggests that this region is critical for the function of histone H4.

The histone fold is known to be conserved throughout the nucleosomal histones H2A, H2B, H3, and H4 and is composed of a tandem duplication of a helix-loop-helix motif. The nucleosomal histones have only 15-20% amino acid sequence similarity with respect to one another (10). Examination of the nucleosomal core histone octamer at a resolution of 3.1Å suggests that the alpha helix 1 region of the histone fold motif is contained between residues 30 and 40 and the loop 1 region between residues 40 to 50 (1). The histone fold has been observed in other eukaryotic DNA binding proteins as well as in archaea (9). Maintenance of the histone fold is thought to be crucial to the structure and function of the histone proteins in the nucleosome.

Five arginines and one lysine reside within the highly conserved region between residues 29 and 49 (Figure 4.1). Positive residues such as arginine and lysine are thought to interact with the DNA as it winds around the nucleosome complex (Figure 4.2). Therefore, acceptable modification at these positions should require a similarly charged residue. For example, the arginine-lysine at positions 20 and 21 in human histone H4 is lysine-lysine in *Oxytricha* and has been completely reversed to lysine-arginine in *Tetrahymena pyriformis*. At positions 78 through 80 in human histone H4, the sequence lysine-arginine-lysine is observed, whereas this sequence is arginine-arginine-lysine in 13 of the 56 known histone H4 sequences including several plant, animal, and ciliate species. However, conservation of charge at these locations appears to be critical.

In contrast to the arginine-lysine pairs found within the core of the protein, lysine residues at positions 5, 8, 12, and 16 in the amino terminus of histone H4 are sensitive to arginine replacement. These residues are subject to a variety of post-translational modifications including reversible acetylation. They are known to be involved with both histone-DNA and histone-protein interactions (8). Replacement of the lysine at position 5 by arginine generated cells that grew normally and did not show a mutant phenotype. Simultaneous replacement of positions 5 and 8 or 5 and 12 with arginine also produced yeast cells with normal phenotypic characteristics. Replacement of arginine for lysine at position 16 generated cells that were phenotypically similar to wild-type yeast except for a decrease in mating efficiency. However, simultaneous replacement of all four lysine positions with arginine resulted in yeast that were not viable. Unexpectedly, simultaneous replacement of these same positions with glutamine produced viable cells; interestingly, these cells had a generation time 70% longer than wild-type and were sterile (8). These results suggest that in some combinations, positive residues can be replaced by negatively charged residues.

To test the permissible variation in the highly conserved core of histone H4 between residues 29 and 49, two arginines at positions 40 and 41 and the lysine and arginine residues at positions 45 and 46 were identified as candidates for mutagenesis. These positions sit at the end of helix 1 and the middle of loop 1 in the histone fold. Using cassette oligonucleotide mutagenesis of the histone H4 gene, these positions were modified in both the *S. cerevisiae* and *T. thermophila* histone H4 contexts and assayed their viability in *S. cerevisiae*. In terms of overall amino acid sequence similarity, *S. cerevisiae* and *T.*

thermophila histone H4 protein sequences differ by 21% and these differences are scattered throughout both the amino terminus and the core of the protein. Our first cassettes introduced equal molar amounts of A, T, C, or G into the second positions of each codon for the four specified amino acid positions creating a space of 256 possible histone H4 proteins in each context. A second series of cassettes introduced equal molar amounts of only A or G to the second position of each codon for these same amino acids, generating all 16 possible arginine-lysine combinations in each context. The results suggest that in each of these contexts, a different but overlapping set of alterations to the universally conserved arginine and lysine residues are viable in yeast. These results strongly indicate that positions outside of the region of modification are influential in determining the viability of these replacements.

MATERIALS AND METHODS

Plasmid Construction

In order to test modifications to the *S. cerevisiae* histone H4-2 and *T. thermophila* histone H4-1 genes, a specially constructed yeast strain (UKY403) was used. UKY403 was previously designed to permit a "glucose shift viability test" (7). Both of the endogenous histone H4 gene copies of strain UKY403 have been replaced with selectable markers and these cells contain the plasmid pUK421 with a *S. cerevisiae* histone H4-2 gene under control of the galactose promoter (*GAL1*) and a tryptophan marker (*TRP1*). The genotype of strain UKY403 is: *MAT***a** ade2-101(och) arg4-1 his3-Δ200 leu2-3 leu2-112 lys2-801(amb) trp1-Δ901 ura3-52 thr^- tyr^- Δhhf1[*HIS3*] Δhhf2[*LEU*2]/pUK421 (*CEN3 ARS1 TRP1 GAL1/HHF*2) (7).

The plasmid pUK499 containing the yeast histone H4-2 gene (*URA3, CEN3, ARS1, Amp*^{r}, and *Ori*) has been used previously for histone mutagenesis (5,7). A plasmid derived from pUK499 containing modifications to either the *S. cerevisiae* histone H4-2 gene or the *T. thermophila* histone H4-1 gene was constructed. Cloning of the *T. thermophila* gene into pUK499 and construction of the plasmid pScTt was described previously (5).

Construction of histone stuffer

Oligonucleotide cassette mutagenesis techniques can be used to add, delete, or substitute nucleotides in a known sequence of DNA (11). During this process, methods that can reduce the frequency of vectors containing unmodified DNAs are desired. Unmodified vector DNA containing a wild-type gene will result in false positives and distort the ratio of viable to non-viable transformants. A common method used to decrease the frequency of unmodified DNA is through the use of a "stuffer" DNA fragment. Using this method, a non-coding fragment of DNA is used first to replace the portion of the gene which is to be modified using oligonucleotide mutagenesis. With the fragment of non-coding DNA interrupting the gene of interest, the function of the protein product is rendered non-functional. Oligonucleotide cassettes can then be used to replace the stuffer in the gene. A percentage of these replacements will restore the original function of the gene and protein product. With this method, any colonies that are identified as viable are so because of the introduction of the oligonucleotide cassette.

Unfortunately, transformation with a vector containing stuffer increases the frequency of false negatives, distorting the ratio of viable to non-viable sequences. This possibility can be alleviated prior to transformation by cutting with a restriction enzyme restriction site unique to the stuffer. Restriction digestion will linearize any plasmid containing stuffer effectively removing the possibility of plasmid replication in the cell. To increase the probability that vector containing stuffer will be removed prior to transformation, unique sites for more than one restriction enzyme can be engineered into the stuffer.

For modification of the histone H4 genes, a "stuffer" was cloned into the pUK499 and pScTt plasmids. Initially, a 44-mer primer 5'-GCGGCCGCGGCGCGCCGTTTAA ACGGCCGGCCGCGGCCGCGGCC-3' was self annealed and extended via PCR amplification. The 72 base pair double stranded product contained multiple restriction sites for the restriction enzymes Not I (GCGGCCGC), Asc I (GGCGCGCC), Pme I (GTTTAAAC), and Fse I (GGCCGGCC). The PCR product was cleaned by using a QIAquick PCR purification kit (QIAGEN), polished with Mung Bean Nuclease to create blunt ends and cloned into the Nru I (TCGCGA) site in pBR322. Cloning of the PCR product was verified by subsequent restriction digestion analysis with Not I. Native pBR322 contains no Not I restriction site. The primers 5'-ATTCCTTGCGGCGGCGG-3' and 5'-GTAGGAGTTCCACAGGG-3' were used for PCR amplification of a 1430 bp fragment of plasmid pBR322 from positions 571 to 2001 respectively containing a Sal I site and Bsp EI site flanking the cloned octamer polylinker. This PCR product was inserted into pUK499 cut with Sal I and Bsp EI to create a "stuffer" in the center of the histone H4 gene.

Cloning of mutagenesis cassettes

pUK499 vector suitable for ligation was produced in microgram amounts following restriction digestions with Sal I and Eco RI enzyme. The products of these digestions were separated with 0.8% agarose gel electrophoresis in 0.5x TAE buffer. Following staining with ethidium bromide (0.5 μg/ml), the resulting vector (8 kb) was extracted from the gel and isolated using dialysis tubing by standard methods (11). This cleaved vector was incapable of self-ligation and was confirmed to produce no transformants in either bacteria or yeast suggesting that only vector free of stuffer was present. This vector was essentially missing the 3' half of the histone H4 gene.

Mutagenesis cassettes were developed as a series of primers (Integrated DNA Technologies) and were purified via HPLC prior to shipment from the manufacturer. Independent PCR amplification, using each of these primers in combination with the universal forward primer, provided the necessary 3' half of the histone gene, including modifications to the region of interest. PCR amplification was accomplished using either the *S. cerevisiae* histone H4 or *T. thermophila* histone H4 sequence previously cloned into the multiple cloning site of pUC19 as a template. The resulting 623 bp PCR product contained the entire 3' half of the histone gene (196 bp) as well as 427 bp of downstream flanking region. This product exactly duplicated the 3' half of the histone H4 gene missing from pUK499, except for the modifications that were introduced to the region of interest.

Mutagenesis cassettes

The first cassette "Sal-256", 5'-GCCAGCTATCCGTCGACTAGCTANAANAGG TGGTGTCANGANAATTTCTG-3' (Integrated DNA Technologies), altered four codons in the histone gene. At each of these positions, equal molar amounts of A, T, C, and G were incorporated (represented by N). This 50-mer cassette generated a spectrum of 256 possible histone proteins including the wild-type sequence. In this spectrum, each of the original arginines (positions 39, 40, and 45) is potentially replaced by lysine, threonine or isoleucine, while the lysine (position 44) is potentially replaced by arginine, threonine or methionine. The second primer, "Sal-RK", 5'-GCCAGCTATCCGTCGACTAGCTARA ARAGGTGGTGTCARGARAATTTCTG-3' (Integrated DNA Technologies), altered the same four positions in the histone gene. At each of these positions however, equal molar amounts of only A and G were incorporated (represented by R). This 50-mer cassette generated a smaller set of 16 possible histone proteins including of the wild-type sequence. Only arginine or lysine codons were generated at each position.

Transformation of Yeast

Transformation of yeast cells was accomplished with a standard protocol for high efficiency transformation (6) with slight modification. UKY403 yeast cells were grown in 20 ml glass flasks (Difco) in a shaking water bath at 28°C from a glycerol stock inoculate overnight to a concentration of approximately 2 x 10^7 cells/ml. The culture was harvested in a sterile 50 ml glass centrifuge tube (Corning) by centrifugation at 3,000 x g for 5 minutes. Cells were resuspended in 25 ml sterile dH_20 and centrifuged again, resuspended in 1.0 ml 100 mM LiAc and transferred to a 1.5 ml microcentrifuge tube for centrifugation at 12,000 x g for 15 seconds at room temperature. Cells were resuspended to a final volume of 500 µl in 100 mM LiAc. The cell suspension was then transferred in 50 µl aliquots to new microcentrifuge tubes and centrifuged to form a pellet. To each tube was added: 240 µl PEG (50% w/v), 36 µl 1.0 M LiAc, 25 µl single stranded salmon sperm carrier DNA (2 mg/ml), 50 µl dH_20, and plasmid DNA (5 µg). The mix was vortexed, incubated at 30°C for 30 minutes, and the tubes were then heat shocked in a water bath at 42°C for 25 minutes to induce transformation.

Transformation mixes were then centrifuged at 3,000 x g for 15 seconds, aspirated to remove the transformation mix and 1.0 ml of sterile dH_2O was used to gently resuspend the cell pellet. 200 μl of cells were transferred in a 10 fold dilution series on SG ura^- plates and incubated at 27°C for 3 days. In order to obtain high efficiencies, fresh yeast cells were used for each series of transformations.

Screening for Viable Yeast Colonies

Following transformation with each of the cassettes generating 256 possible proteins, colonies were grown on SG ura^- plates and 1000 colonies were individually transferred via toothpick to YPD plates and then subsequently to fresh SG ura^- plates. The plasmid pUK499 contains a uracil marker (*URA3*) and a modified histone H4 gene under control of the dextrose promoter. Therefore, growth on SG ura^- plates selects for the presence of the pUK499 plasmid due to the uracil requirement, while cells continue to produce wild type histone H4 from the histone H4 gene on plasmid pUK421 under control of the galactose promoter. Replica plating positive SG ura^- colonies to YPD media effectively switches off the galactose promoter leaving only the modified histone H4 expressed in the cells. This process is known as the "glucose-shift-viability assay" (5,7). Viable colonies on YPD were produced after two to three days of growth at 27°C. All colonies that grew on YPD media (YPD^+) also grew on SG ura^-. Glycerol stocks for each colony with modified histone were made by adding 800 µl of liquid cell culture to 250 µl of pre-sterilized glycerol. These tubes were stored at -70°C for future use.

PCR from yeast cells

PCR amplification of the histone H4 genes from yeast cells was accomplished using a standard protocol (3) with slight modification. Each yeast colony containing a modified histone H4 gene was grown in 3 ml of YPD media overnight at 28°C in a shaking water bath. Cultures were then resuspended with gentle stirring and a 50 µl aliquot was transferred to a 1.5 ml microcentrifuge tube. 100µl of a solution containing 2.5M LiCl, 50 mM Tris-HCl (pH 8.0), and 62.5mM EDTA was added to the cells, followed by an equal volume of phenol:chloroform and 0.2 g of Superbright glass beads (3M Company, St. Paul, Minnesota; type 100-5005). This mixture was vortexed for 2 minutes and centrifuged for 5 minutes at 12,000 x g in a desktop microcentrifuge. The supernatant was transferred to a fresh microcentrifuge tube and DNA was precipitated by the addition of 300µl of ice cold 95% ethanol with 300 mM sodium acetate on ice for 20 minutes. Samples were centrifuged 12,000 g at 4°C for 30 minutes and the resulting pellet was washed with cold 70% ethanol and dried under vacuum for 2 to 5 minutes. The DNA was resuspended in 20 µl of TE buffer.

The isolated DNA was cut with Not I enzyme, prior to PCR amplification of the histone H4 gene. pUK421 plasmid contains a Not I site in the *GAL1* promoter, whereas the plasmid pUK499 containing the modified histone genes does not. The Not I enzyme can be used to selectively digest the pUK421 internal to the primer sites used for PCR. This leaves the pUK499 plasmid as the only intact template for PCR amplification of the histone H4 gene.

PCR amplification of the histone H4 gene was performed by adding 1 µl of the extracted yeast DNA as template to 35.75µl of sterile dH_20. The PCR reaction was performed in a total volume of 50 µl, including: 5µl 10x PCR Buffer II (100 mM Tris-HCl, pH 8.3, 500 mM KCl), 3 µl of 25 mM $MgCl_2$, 4µl dNTP (10mM each dNTP), 0.5µl of each primer (0.4 pmoles/µl) "5'SCH4specific" 5'-CGGGCGCGAAATGCAGACCAG ACCAG-3', "3'SCH4specific" 5'-TAACAGTCTTTCTCTTGGCG-3', and 0.25µl (1.25 units) of AmpliTaq Gold™ (Perkin Elmer, Foster City, CA). All reactions were performed in thin walled 0.2 mL reaction tubes. Samples were subjected to a hot start for 10 minutes at 97°C, followed by 10 cycles of 30 seconds at 96°C, 30 seconds at 60°C, and 1 minute at 72°C, 25 cycles of 20 seconds at 94°C, 30 seconds at 60°C, and 1 minute at 72°, a 10 minute extension at 72°C and subsequent hold at 4°C. A 5µl aliquot of the PCR sample was checked on a 1.5% 0.5x TBE gel stained with ethidium bromide and visualized under ultraviolet light after electrophoresis at 100V for 80 minutes. PCR products were then cleaned using Microcon-100 microconcentrators (Amicon, Beverly, MA) following the manufacturers' instructions. 3µl of the cleaned PCR product was run on a 1.5% 0.5x TBE gel to determine the concentration.

PCR Sequencing Reactions

5 µl aliquots of the cleaned PCR product were used as templates for sequencing. The primer "5'H4" 5'-CGCTTAATTTATTCTTTTCTC-3' sits just upstream of the histone H4 gene on the pUK499 plasmid. This primer will hybridize only to the yeast histone H4 promoter on the plasmid containing the modified histone H4 gene. A 0.5µl (5 pmol) aliquot of 5'H4 primer was added to the mixture along with 4 µl of sequencing reaction mixture

provided by the UCLA Core Sequencing Facility and 0.5 µl of sterile dH_20 was added to each tube to bring the final volume to 10µl. All tubes were subjected to 25 cycles composed of: 10 seconds at 96°C, 5 seconds at 50°C, and 4 minutes at 60°C in a Perkin Elmer 2400 PCR machine (Perkin Elmer, Foster City, CA) using thin walled 0.2 mL reaction tubes. Following amplification, reaction products were transferred to a 1.5 mL microcentrifuge tube and 25µl of a 95% EtOH solution containing 300 mM sodium acetate was added. Samples were centrifuged at 12,000 g for 30 minutes; the pellet was washed with 70% ethanol and dried under vacuum. The resulting dried pellet was given to the UCLA Core Sequencing Facility for sequencing.

RESULTS

Amino acid replacements using the Sal-256 cassette in *T. thermophila* histone H4

Introduction of the Sal-256 cassette into the core of S. cerevisiae histone H4 yields 70 colonies out of 1069 colonies tested on dextrose media. Sequencing of 15 of 70 viable colonies from dextrose media generated a set of 11 different sequences. All of these sequences were determined to be combinations of arginine and lysine. No threonine, isoleucine, or methionine replacements were observed. Coupled with the fact that the percentage of viable colonies (6.5%) is close to the percentage expected with only arginine and lysine replacements (6.25%), this strongly suggests that all arginine and lysine replacements are viable in the *T. thermophila* context.

Amino acid replacements using the Sal-256 cassette in *S. cerevisiae* histone H4

Introduction of the Sal-256 cassette into the core of *S. cerevisiae* histone H4 yields 58 viable colonies out of 1,338 colonies tested on dextrose media. This percentage (4.3%) is less than would be expected if all 16 arginine and lysine only combinations functioned in the context of *S. cerevisiae* histone H4 (6.25%). The DNA sequences from the 50 colonies viable on dextrose media with Sal-256 cassette modifications in the *S. cerevisiae* context were analyzed.

From the 50 colonies, 13 different sequences were determined (Table 4.1). Of these 13 sequences, 5 (including the wild-type sequence) were composed of only arginine and lysine replacements, while the other 8 had predominately arginine and lysine replacements, but also included methionine, isoleucine, and/or threonine. Replacement of lysine for threonine was not observed in any viable protein. At the first and fourth modified positions (histone H4 positions 39 and 45), the wild type amino acid arginine was only replaced by lysine. The second and third modified positions (histone H4 positions 40 and 44) were more open to modification. At position 40, replacement of arginine with threonine is viable only when a lysine is also present in position 45. Similarly, replacement of arginine with isoleucine at position 40 is viable when positions 39 and 44 are replaced with lysine and methionine, respectively. At position 44, replacement of lysine with methionine occurs in 6 of the 13 histone H4 sequences known to be viable in yeast.

Modifications with the Sal-RK cassette

In order to confirm the surprising result that the set of arginine and lysine replacements in the *S. cerevisiae* context was different than the *T. thermophila* context, a second set of cassettes (Sal-RK) was tested. This cassette generated the set of 16 sequences with arginine and lysine replacements. With the Sal-RK cassettes in the *S. cerevisiae* context, 34 out of 100 colonies tested were viable on dextrose media. This value is in close agreement with our Sal-256 results. From the Sal-256 results, we expected 5 out of the 16 sequences (31.3%) to be viable with the Sal-RK cassette. With the Sal-RK cassette in the context of the *T. thermophila* histone H4, 98 out of 98 colonies tested were viable on dextrose media (100%). These results confirmed the analysis with the Sal-256 cassette and suggested that only arginines and lysines are viable in the context of the *T. thermophila* histone H4.

DISCUSSION

The very low rate of nonsynonymous substitution suggests that the evolution of histone H4 is highly constrained by the essential role it plays in the cell. Modifications to histone H4 that affect the structure and function of the protein would therefore be expected to be deleterious. This hypothesis can be tested by introducing a set of modifications into an absolutely conserved region and monitoring viability of yeast in a permissive laboratory setting.

It is reasonable to expect that more similar amino acids are (by charge or size) more easily replaced by one another. For instance, Dayhoff et al. (4) noted that of all of the amino acids, lysine most easily replaces arginine and arginine most easily replaces lysine. Both of these residues are positively charged and share similar shapes. Conservation of overall protein structure dictates that with single residue replacements, amino acids most similar to the wild-type residues are favored for replacement. The wild-type sequence for the region of interest contains an arginine-arginine pair closely followed by a lysine-arginine pair. These residues are believed to play an important role in the wrapping of negatively charged DNA around the nucleosome. Alignment of all known histone H4 proteins suggests that these residues are extremely conserved, as no alternative amino acids are known in naturally occurring histone H4 proteins at these locations.

At each of these amino acid positions, sets of new amino acid combinations were introduced through cassette mutagenesis of the histone H4 gene. The first cassette, Sal-256, generates 256 possible histone H4 proteins through a random modification of the second codon position for each of the four amino acid positions 39, 40, 44 and 45. The 256 sequence set contains all possible arginine and lysine combinations in addition to possible replacements by methionine and threonine at position 44 and isoleucine and threonine at positions 39, 40 and 45. *A priori*, it might be expected that due to similarities in charge and size, only arginines and lysines would function at each of these positions.

In the context of the *T. thermophila* histone H4 sequence, 16 of the 256 sequences generate viable yeast cells. All 16 of these sequences are arginine and lysine combinations only. These data agree with the hypothesis that positive residues most easily replace one another. In the context of the *S. cerevisiae* histone H4 sequence, however, 13 of the 256 sequences generate viable yeast. Of these 13 sequences that yield viable yeast, 5 were determined to contain only arginine and lysine combinations. One of these 5 was the wild-type sequence. The remaining 11 combinations of arginine and lysine only were non-viable in this context; this was an unexpected result with regards to the fact that all arginine and lysine combinations were viable in the *T. thermophila* context.

A more detailed analysis of the 5 viable arginine and lysine sequences in the *S. cerevisiae* histone H4 suggests that all of the 4 non-wild-type sequences (RKKK, KKKR, RKRK, and KKKK) each contain 2 or 3 differences with respect to the wild-type sequence

(Figure 4.3). It is surprising that some arginine and lysine only sequences that have only 1 or 2 changes with respect to the wild-type are not viable in this context. A hypothesis suggesting that the frequency of viable sequences increases when sequences are close to the wild-type is not supported by this data. Only 1 of the 13 cassette sequences that yield viable cells in *S. cerevisiae* histone H4 context has one amino acid difference with respect to the wild-type. Only 4 of the 16 cassette sequences that yield viable cells in *T. thermophila* histone H4 context have one difference with respect to the wild-type sequence. This data suggests that the changes at each position might be interactive.

Of the 13 viable sequences in the *S. cerevisiae* histone H4 context, 8 contained at least one position that was neither arginine or lysine (Table 4.1). This result is surprising relative to previous results demonstrating that the entire set of only arginine and lysine replacements was viable in the *T. thermophila* histone H4 context. Of these 8 sequences, 6 contained a methionine replacement at position 44. Only 1 sequence was found to contain a viable replacement of lysine by arginine at position 44. It is very surprising that the positively charged amino acid lysine was more easily replaced at this location by a non-polar methionine than a similar, positively charged, arginine. No threonine replacements at position 44 were viable, although 3 of the 8 sequences contained a threonine replacement at position 40. Only one sequence contained an isoleucine at position 40 and that sequence also has a methionine at position 44.

Positions 39 and 45 contain only arginine and lysine residues in roughly equal amounts in the cells that are viable from the Sal-256 cassette. However, arginine and lysine residues are not interchangeable in the context of *S. cerevisiae* histone H4. Among the sequences that yield viable cells from the Sal-256 cassette in the *S. cerevisiae* context, there are 3 sets which have identical internal positions (40 & 44) (Figure 4.4). These internal positions have only lysine replacements. The outer positions (39 & 45) for these 3 sets have four alternative arrangements of arginine and lysine (Figure 4.4). Functional sequences do not occur in equal abundance, suggesting a bias towards the replacement of lysine over arginine. The remainder of the non-wild-type sequences that are viable in the *S. cerevisiae* context require specific amino acid replacements at positions 39 and 45 (RKRK, RRMR, RTMK, KIMR).

In order to confirm the number of viable sequences with only arginine and lysine residues, a second cassette Sal-RK was constructed. This cassette altered the second position of each of the four codons only with an A or G and gave rise to only arginine and lysine replacements. Using the Sal-RK cassette, all 16 of the possible arginine and lysine combinations are viable in the *T. thermophila* histone H4 context. When using the *S. cerevisiae* context, the ratio of viable to non-viable yeast (31.3 %) suggests that only 5 of the 16 possible arginine or lysine combinations are viable. In both cases, evidence from the Sal-RK cassette supports the previous analysis with the Sal-256 cassette.

Curiously, all of the non-arginine or lysine replacements in the *S. cerevisiae* context occurred only in two positions (40 and 44). At positions 39 and 45, arginine was replaced only by lysine in a manner similar to the *T. thermophila* context. Positions 40 and 44 appear to be more free to vary than positions 39 and 45. A new set of oligonucleotide cassettes could be generated to test this hypothesis. Positions 39 and 45 could be held to either arginine or lysine and a wider spectrum of replacements could be inserted at positions 40 and 44. It would be interesting to determine if the arginine at position 40 can be replaced by methionine in a similar manner to the lysine at position 44. It would also be interesting to determine what (if any) other amino acids besides methionine, threonine, and isoleucine can be used to generate viable histone H4 proteins.

To more fully investigate the modifications of positions 40 and 44 in the Sal-256 cassette that yield viable cells, an alternate cassette could be generated. Ideally, this cassette would test isoleucine in position 44 and methionine in position 40 as well as regenerate the set of alternatives tested by the Sal-256 cassette. The use of codon ANR at positions 40 and 44 generates all five possible amino acids R,K,T,I,M in the ratios of 2:2:2:1:1 respectively. The use of codon ARA at positions 39 and 45 generates two possible amino acids (R & K). Due to the skewed representation of isoleucine and methionine at positions 40 and 44, these codons generate the equivalent of 256 sequences. 192 of these sequences have been previously interrogated using the Sal-256 cassette and, of that number, 13 different sequences are known to be viable in the context of *S. cerevisiae* histone H4. 64 of the 256 sequences have never been assayed for viability and contain methionine at position 40 and/or isoleucine at position 44. Therefore, the expected viability for this new cassette will range from 7.8% (only the known 20 [due to the skewed representation of M and I] viable sequences are viable using the new cassette) to 32.8% (all of the 64 new methionine and isoleucine sequences are viable in addition to the previous 20). The ratio of viability can easily be determined using the glucose shift viability assay. In the case where the viability is slightly above 5%, sequencing of 50 random viable clones should give a 95% probability of sequencing all viable sequences completely. In the case where the viability is closer to 30%, sequencing of 50 viable clones should give a wide variety of new sequences each of which should contain methionine or isoleucine replacements at positions 40 and 44 respectively.

Comparing the histone H4 proteins of *T. thermophila* and *S. cerevisiae* there are 21 amino acid differences, one deletion, and one insertion for a total of 23 differences. These differences are scattered throughout the protein (5). Despite this large difference, the *T.*

thermophila protein can successfully replace the *S. cerevisiae* histone H4 *in vivo* although yeast utilizing the *T. thermophila* histone have altered phenotypes (5). It is possible that the differences between these two wild-type histone contexts are responsible for the different sets of viable alterations observed between these sequences. Less latitude might be expected of yeast utilizing the *T. thermophila* histone H4 due to an already compromised phenotype. Potentially, this only explains the discovery of successful non-arginine or lysine residue replacements in the *S. cerevisiae* histone H4 context. However, the number of viable modified sequences in the *S. cerevisiae* context is fewer than in the *T. thermophila* context. The discovery that all arginine and lysine replacements are viable in the *T. thermophila* context but not in the *S. cerevisiae* context was most unexpected.

By using oligonucleotide cassette mutagenesis, a universally conserved region in histone H4 can be modified and yield viable yeast cells. At some positions in the protein, positively charged residues (arginine and lysine) can only be replaced by similarly positive residues in the context of both the *S. cerevisiae* histone H4 and *T. thermophila* histone H4 when assayed in yeast. However, at position 40 and 44, there exists a dramatic difference in the variety of amino acid replacement between the contexts of *S. cerevisiae* and *T. thermophila* histone H4. In the context of *T. thermophila* histone H4, only arginine and lysine residues are accepted. In the context of *S. cerevisiae* histone H4, a much wider variety of residues can be accepted. This suggests that, at some positions, positively charged residues can be replaced by polar or non-polar residues in the core of histone H4. *T. thermophila* and *S. cerevisiae* histone H4 proteins have 23 amino acid differences scattered in regions outside of the area that was the focus of this study. Therefore, it is likely that the viability of amino acid replacements to this region of histone H4 are to a large degree affected by the type and location other amino acid residues in the protein and their interactions in the maintenance of histone H4 structure and function.

REFERENCES

1. Arents, G., R. W. Burlingame, B.-C. Wang, W. E. Love, and E. N. Moudrianakis. 1991. The nucleosomal core histone octamer at 3.1Å resolution: A tripartite protein assembly and a left-handed superhelix. Proc. Natl. Acad. Sci. USA. 88:10148-10152.

2. Baxevanis, A. D. and D. Landsman. 1997. Histone and histone fold sequences and structures: a database. Nuc. Acids Res. 25(1):272-273.

3. Caldwell, G. A. and J. M. Becker. 1993. Rapid PCR1 Sequencing of plasmid DNA Directly from Colonies of *Saccharomyces cerevisiae*. Promega Notes Magazine. 44:6

4. Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. A Model of Evolutionary Change in Proteins. In Atlas of Proteins Sequence and Structure. (ed. Dayhoff, M.O.) National Biomedical Research Foundation, Washington, D. C. pp. 345-358.

5. Fogel, G. B. and C. F. Brunk. 1997. Expression of *Tetrahymena* histone H4 in yeast. Biochemica et Biophysica Acta. 1354:116-126.

6. Gietz, R. D. and R. H. Schiestl. 1995. Transforming yeast with DNA. Methods in Molecular and Cellular Biology. 5:255-269.

7. Kayne, P. S., U.-J. Kim, M. Han, J. R. Mullen, F. Yoshizaki and M. Grunstein. 1988. Extremely conserved histone H4 N-terminus is dispensable for growth but essential for repressing the silent mating loci in yeast. Cell. 55:27-39.

8. Megee, P. C., B. A. Morgan, B. A. Mittman, and M. Mitchell Smith. 1990. Genetic analysis of histone H4: Essential roll of lysines subject to reversible acetylation. Science. 247:841-845.

9. Ouzounis, C.A. and N. C. Kyrpides. 1996. The core histone fold: Limits to Functional Versatility. J. Mol. Evol. 43:541-542.

10. Ramakrishnan, V. 1995. The histone fold: Evolutionary questions. Proc. Natl. Acad. Sci. USA. 92:11328-11330.

11. Sambrook, J., E. F. Fritsch, and T. Maniatis. 1989. Molecular Cloning: A Laboratory Manual. Second edition. Cold Spring Harbor Laboratory Press.

Figure 4.1

Aligned sequences of histone H4 in *Saccharomyces cerevisiae* (Sc) and *Tetrahymena thermophila* (Tt). The amino acid sequences for each codon are given above and below the nucleotide sequences for each species. Bars between the sequences indicate nucleotide matches. The nucleotides modified to create a Sal I restriction site used for vector construction are shown. Boxes indicate the arginine and lysine residues at positions 39, 40, 44, and 45 which were modified in this study. Four arrows point to the second positions of the codons which were altered with either N (in the case of the Sal-256 cassette) or R (in the case of the Sal-RK cassette). Indications of helix (H) and loop (L) are given for the amino acids involved with the histone fold.

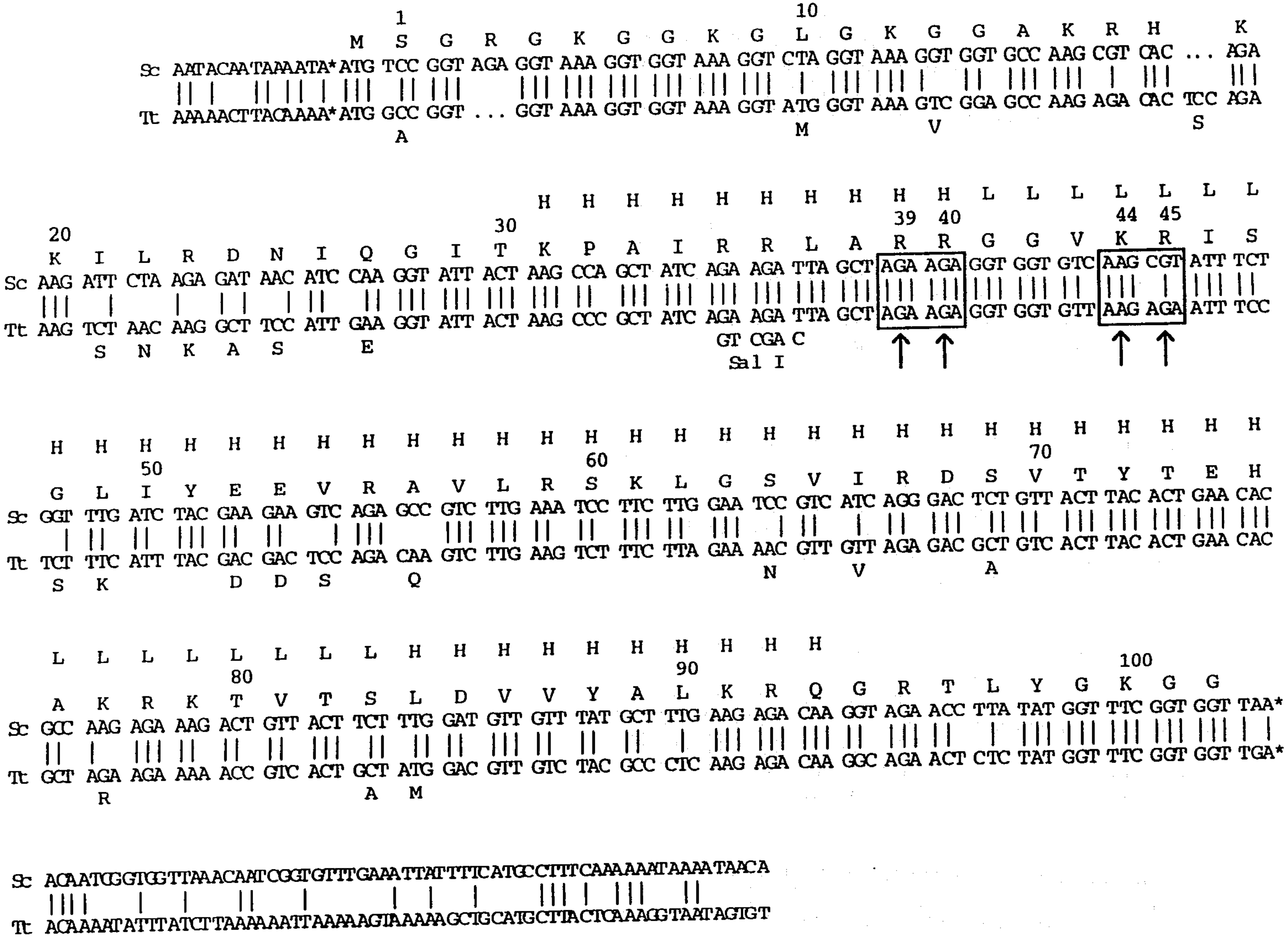
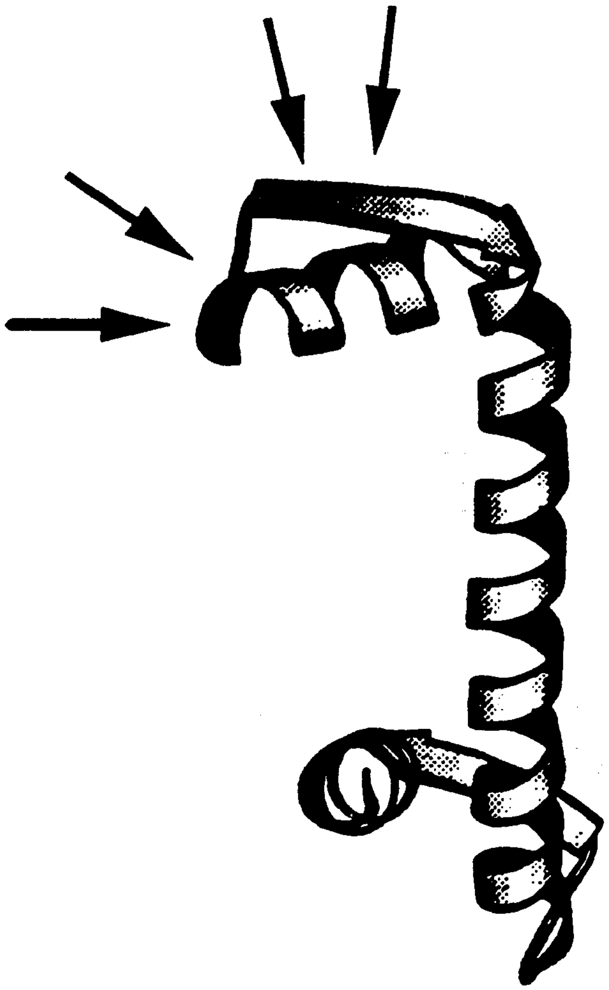

1 10
M S G R G K G G K G L G K G G A K R H K
Sc AATACAATAAAATA*ATG TCC GGT AGA GGT AAA GGT GGT AAA GGT CTA GGT AAA GGT GGT GCC AAG CGT CAC ... AGA
Tt AAAAACTTACAAAA*ATG GCC GGT ... GGT AAA GGT GGT AAA GGT ATG GGT AAA GTC GGA GCC AAG AGA CAC TCC AGA
A M V S
H H H H H H H H H H L L L L L L L
20 30 39 40 44 45
K I L R D N I Q G I T K P A I R R L A R R G G V K R I S
Sc AAG ATT CTA AGA GAT AAC ATC CAA GGT ATT ACT AAG CCA GCT ATC AGA AGA TTA GCT AGA AGA GGT GGT GTC AAG CGT ATT TCT
Tt AAG TCT AAC AAG GCT TCC ATT GAA GGT ATT ACT AAG CCC GCT ATC AGA AGA TTA GCT AGA AGA GGT GGT GTT AAG AGA ATT TCC
S N K A S E
GT CGA C
Sal I
H H H H H H H H H H H H H H H H H H H H H H H H H H H H
50 60 70
G L I Y E E V R A V L R S K L G S V I R D S V T Y T E H
Sc GGT TTG ATC TAC GAA GAA GTC AGA GCC GTC TTG AAA TCC TTC TTG GAA TCC GTC ATC AGG GAC TCT GTT ACT TAC ACT GAA CAC
Tt TCT TTC ATT TAC GAC GAC TCC AGA CAA GTC TTG AAG TCT TTC TTA GAA AAC GTT GTT AGA GAC GCT GTC ACT TAC ACT GAA CAC
S K D D S Q N V A
L L L L L L L L H H H H H H H H H H
80 90 100
A K R K T V T S L D V V Y A L K R Q G R T L Y G K G G
Sc GCC AAG AGA AAG ACT GTT ACT TCT TTG GAT GTT GTT TAT GCT TTG AAG AGA CAA GGT AGA ACC TTA TAT GGT TTC GGT GGT TAA*
Tt GCT AGA AGA AAA ACC GTC ACT GCT ATG GAC GTT GTC TAC GCC CTC AAG AGA CAA GGC AGA ACT CTC TAT GGT TTC GGT GGT TGA*
R A M
Sc ACAATGGGTGGTTAAACAATCGGTGTTTGAAATTATTTTCATGCCTTTCAAAAAATAAAATAACA
Tt ACAAAATATTTATCTTAAAAAATTAAAAAGTAAAAAGCTGCATGCTTACTCAAAGGTAATAGTGT

Figure 4.2

Tertiary structure of the histone fold from histone H4. Positions that were modified in this chapter are shown by arrows.



H4

Figure 4.3

Viable sequences from the set of 256 possible variants in both the *S. cerevisiae* context and *T. thermophila* context. Boxes indicate sequences found only when in the context of the *S. cerevisiae* histone H4. Circles indicate sequences common to both the *S. cerevisiae* and *T. thermophila* histone H4 contexts. The remaining sequences are found only in the context of the *T. thermophila* histone H4. Sequences are displayed with reference to the one letter amino acid codes for the wild-type amino acids (RRKR) at positions 39, 40, 44, and 45. Numbers at the top of each column indicate the number of differences with respect to the wild-type sequence. Solid lines indicate amino acid sequences that differ between each other by one amino acid replacement.

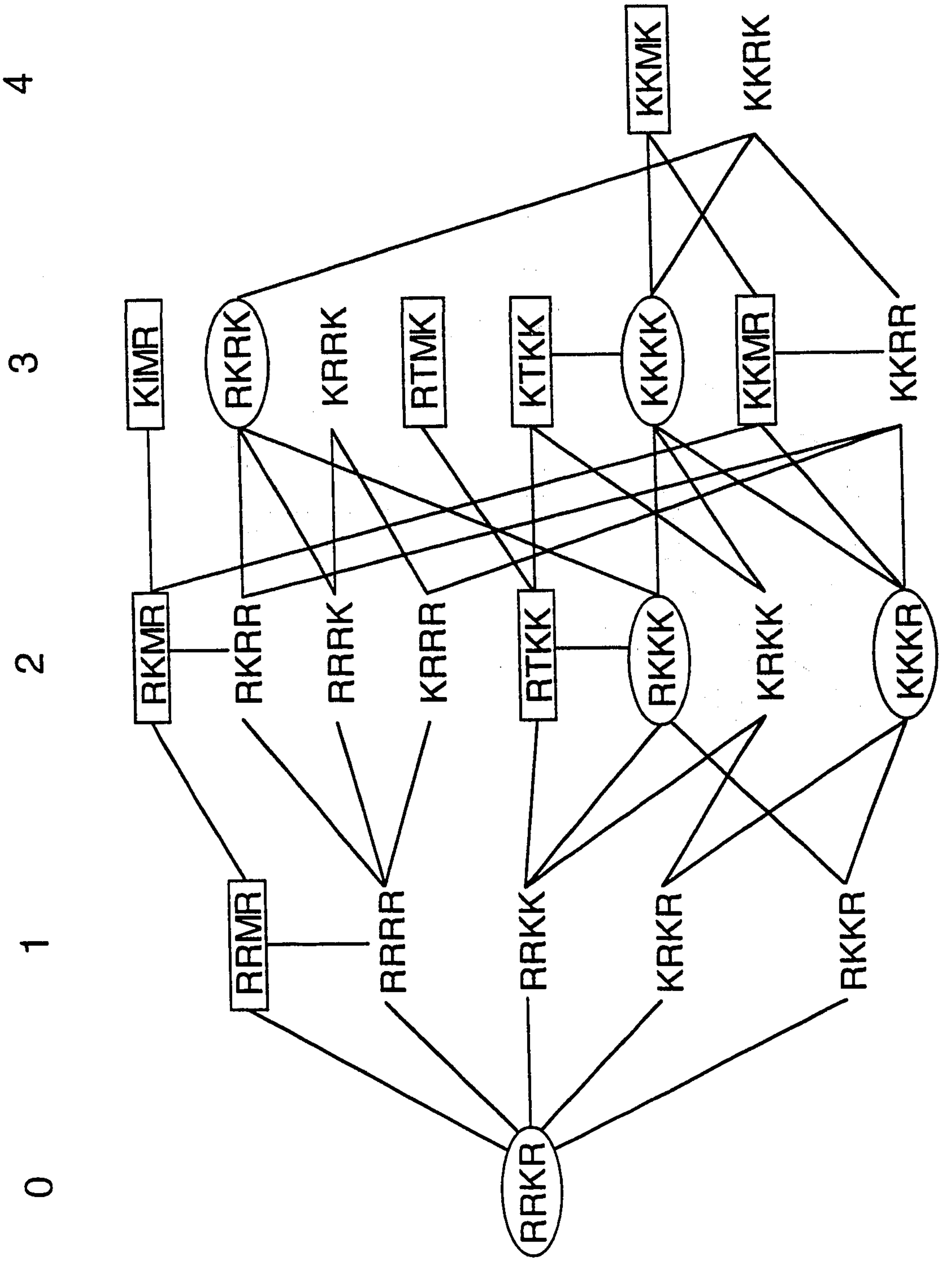

0
1
2
3
4
RRKR
RRMR
RRRR
RRKK
KRKR
RKKR
RKMR
RKRR
RRRK
KRRR
RTKK
RKKK
KRKK
KKKR
KIMR
RKRK
KRRK
RTMK
KTKK
KKKK
KKMR
KKRR
KKMK
KKRK

Figure 4.4

Three common sequence sets for amino acid replacements at positions 40 and 44 in the *S. cerevisiae* histone H4 context. Viable sequences are indicated by (+) whereas non-viable sequences are indicated by (-). The number of viable sequences is given in parenthesis for each of the sequence motifs.

		-**KM**-	(3)	-**KK**-	(3)	-**TK**-	(2)
K--K	(3)	K**KM**K	(+)	K**KK**K	(+)	K**TK**K	(+)
K--R	(2)	K**KM**R	(+)	K**KK**R	(+)	K**TK**R	(-)
R--K	(2)	R**KM**K	(-)	R**KK**K	(+)	R**TK**K	(+)
R--R	(1)	R**KM**R	(+)	R**KK**R	(-)	R**TK**R	(-)

Table 4.1

Viable sequences from the Sal-256 cassette mutagenesis in the *S. cerevisiae* histone H4 context. Amino acid positions are noted at the top of the table. Sequences are grouped by common amino acids listed on the left with the wild-type sequence (w.t.) given first. The number of amino acid differences with respect to the wild-type is noted for all sequences in the last column. The majority of viable sequences from this set have two or three differences with respect to the wild-type sequence.

pos.	39	40	44	45	# dif.
w.t.	R	R	K	R	0
RK	R	K	K	K	2
	K	K	K	R	2
	K	K	K	K	3
	R	K	R	K	3
RKM	R	R	M	R	1
	R	K	M	R	2
	K	K	M	R	3
	K	K	M	K	4
RKTIM	R	T	K	K	2
	K	T	K	K	3
	R	T	M	K	3
	K	I	M	R	3

R	6	1	1	5
K	6	7	5	7
T	0	3	0	0
I	0	1	0	0
M	0	0	6	0

Chapter 5

DERIVATION OF OLIGONUCLEOTIDE CASSETTES FROM AMINO ACID REPLACEMENT MATRICES

ABSTRACT

An accepted point mutation (PAM) matrix is used for the development of oligonucleotide cassettes for amino acid replacement in histone H4. Using a PAM matrix, a set of codons predicted to have a high probability of successful replacement were tested in five positions in a highly conserved region of *S. cerevisiae* histone H4. Our results suggest that this method can be used to efficiently search for novel functional proteins. Speculation on the predictive power of different PAM matrices is offered. The predictive power of any PAM matrix can be measured using yeast viability as a measure of success.

INTRODUCTION

Mutagenesis is an invaluable technique for the investigation of proteins. The ability to generate a variety of mutant proteins by inserting a variable cassette of amino acid sequences (cassette mutagenesis) adds to our knowledge of protein structure and activity and can be used for many related bioengineering purposes. Cassette mutagenesis has yielded important discoveries such as the isolation of antibody fragments with high affinities for arbitrary compounds (7,24); screening of peptide libraries for desirable properties (27,30,37); creation of drug-specific mutants for gene therapy (11); and identification and mapping of important residues in proteins such as herpes simplex virus type 1 thymidine kinase (10), glucoamylase (6), or human rhinovirus based HIV-1 immunogens (36).

Many strategies have been developed for the search of sequence space through mutagenesis (39). One potential strategy focuses on the replacement of amino acids at a single specific position in a protein. In this case, *R* amino acid replacements can be generated where *R* varies from 1 to 19 non-wild-type residues. This strategy is useful for locating critical and non-critical positions, such as the glutamic acid residue at position 461 in β-galactosidase (14) or the "helix clamp" in HIV-1 reverse transcriptase (8). Possible long range successful interactions of amino acid replacements are neglected by this approach. While this technique might demonstrate a number of viable replacements, the majority of the sequences are bound to be close to the wild-type sequence, as the search space is closely centered on the wild-type protein (29,31,32,35).

Another potential strategy is the replacement of *R* residues at *n* positions in the protein of interest (where *n* is larger than 1). The space of possible sequences in this set grows as a function of R^n giving a space that quickly becomes much too large to search exhaustively (18). When working with yeast (or most eukaryotic cells), high efficiency transformation rates of 10^4-10^5 cells/µg of DNA set an upper limit on the number of novel protein sequences that can be assayed *in vivo*. This transformation efficiency sets a limit on the size of the sequence space that can be investigated and therefore the number of positions and/or replacements at each position that can be simultaneously altered in a protein.

With this upper limit in mind, the researcher is forced to reduce the number of positions and/or replacements at each position to match the maximum transformation efficiency. The number of positions n can be chosen *a priori* by the researcher depending on the question which is being asked. In some cases, the researcher might be interested in understanding the potential variability in a highly conserved, critical region. In other cases, the researcher might be interested in the coupled interactions of non-conserved positions in the protein. Therefore, the number of positions (n) to be varied is determined by the researcher. Similarly, at each position, the number of amino acid replacements can be varied to match the desired purpose of the study and yet conform to the upper limit of transformation efficiency. In this case, the number of non-wild-type replacements, R_i, varies from 1 to 19 at each specified position j. This generates the relationship $N = \Pi R_j$, where N is the total number of sequences that are to be investigated and R_j is the number of replacements at each specific position. The researcher is left to decide two variables: the number and location of positions to be varied (n) and the set of replacements for each position to be varied (R_j). Both of these choices are generally determined in light of the question being asked. There are a variety of methods that can be used to select the best set of replacements at each position that is to be varied. The method described here utilizes an accepted point mutation (PAM) matrix as a basis for this selection for the modification of a highly conserved region in histone H4.

Modifications to a highly conserved region in *S. cerevisiae* histone H4

Studies on the evolution of histone H4 proteins in lower eukaryotes suggest that the amino acid sequences of *Saccharomyces cerevisiae* and *Tetrahymena thermophila* have substantially diverged with respect to the consensus sequence derived from all known histone H4 sequences (22). The histone H4 sequences of different species of *Tetrahymena* also show substantial divergence from each other (38). Using a system in which the endogenous histone H4 genes of *Saccharomyces* can be replaced by a histone H4 gene introduced to the cell, Fogel and Brunk (22) demonstrated that the *Tetrahymena* histone H4 can replace *Saccharomyces* H4 *in vivo* yielding viable yeast cells. This result suggests that a larger set of modified histone H4 sequences might also generate viable cells when used to replace the endogenous histone H4. Several viable modifications to an extremely conserved region in histone H4 were the subject of the previous chapter. An amino acid replacement cassette for the modification of a region of nine amino acid residues that is absolutely conserved in all histone H4 proteins between positions 38 and 46 was designed. Viable yeast containing modifications to this region should offer insights into the highly conserved nature of the histone fold (2,3,43). Within this region, four arginine and lysine residues at positions 39, 40, 44, and 45 were the focus of chapter 4. Assuming that these positions were potentially less free to vary due to charge restrictions (as was the case for modification of positions 39 and 45), we wished to alter the five residues at positions 38, 41, 42, 43, and 46 (Figure 5.1; Figure 5.2). A cassette was constructed that generated a wide spectrum of possible sequences with the highest predicted functionality using a PAM matrix. The results suggest that a large number of modifications to this region are viable and that the PAM matrix is a useful tool for the design of oligonucleotide cassettes. The predictive power of the PAM matrix might be tested in a series of additional experiments.

MATERIALS AND METHODS

A PAM matrix was originally defined as the matrix relating the replacement of one amino acid such that the new amino acid was successfully accommodated in the structure and function of the protein (17). A PAM matrix can be derived by either the common ancestor method (17) or through the use of a distance matrix (21). In the common ancestor approach, a phylogenetic tree is used to infer ancestral sequences and the matrix derived from the comparison of present day sequences to inferred ancestral sequences. In the use of a distance matrix, only present day sequences are compared. Regardless of the method, apparent amino acid replacements (A_{ij}) are tallied into a symmetrical pairwise exchange table (PET). The relative mutability of a given amino acid (m_j) is calculated as the number of observed changes for that amino acid divided by its frequency of occurrence in the PET.

With the apparent amino acid replacement information and relative mutabilities, a replacement probability matrix can be derived which gives the probability that a residue in column *j* of the matrix can replace the residue in row *i* in a specified unit of evolutionary time and yield a functional protein (28). Non-diagonal terms for the replacement probability matrix have the values:

$$M_{ij} = \lambda m_j A_{ij}/\Sigma A_{ij}$$

where A_{ij} is an element of the pairwise exchange table (Table 5.1), λ is a proportionality constant, and m_j is the relative mutability of amino acid *j* (Table 5.2). On-diagonal terms for the replacement probability matrix have the values:

$$M_{jj} = 1 - \lambda m_j$$

The probability of observing a change in a position containing any amino acid is proportional to the mutability of that amino acid. The same proportionality constant, λ, holds for all columns in the replacement probability matrix. For the 1 PAM, λ is set such that the average of the off-diagonal terms is equal to 1%. Matrix multiplication of the 1 PAM can be used to derive other evolutionary distances. An evolutionary distance produced by 250 matrix multiplications (250 PAM) is generally accepted as a useful balance between similar and divergent sequences.

The Jones PET91 and relative mutability index

Dayhoff et al. (17) used 72 sequence families consisting only about 1300 sequences to determine PAM matrices. With the recent dramatic increase in sequence information for protein data (12), Jones et al. (28) generated a more extensive update of the Dayhoff matrix using 16,130 protein sequences and 59,190 accepted point mutations from the SWISS-PROT database (5). The Jones "Pairwise Exchange Table 1991" (PET91; Table 5.1) is equivalent to the original pairwise exchange table (MDM78) derived by Dayhoff et al. (17). The PET91 gives information about the individual kinds of amino acid replacements and their frequency of occurrence in the database. Although many more sequences were used in the PET91, the PET91 was strikingly similar to the MDM78. The various amino acid replacement matrices have been compared previously (25,41,42). Although other replacement matrices exist, the PET91 matrix was chosen as a representative of a broad range of proteins that exist in the SWISS-PROT database.

The relative mutability of each amino acid was determined by dividing the number of replacements by the total number of occurrences for each specific amino acid in the PET91 database (28) (Table 5.2). The PET91 and the relative mutability index can be combined to generate an "accepted point mutation" or 1 PAM matrix (Table 5.3). The 1 PAM matrix gives the probability that a specified amino acid will be replaced by another specified amino acid after an evolutionary interval that gives an average of one amino acid replacement per 100 residues. The on-diagonal terms in the 1 PAM matrix are far bigger than the off-diagonal terms due to the lack of sufficient evolutionary divergence of proteins in the database. The 1 PAM provides the necessary information to simulate further evolutionary change through matrix multiplication. It is left for the researcher to choose the most appropriate PAM matrix in light of the questions being addressed. For instance, the 250 PAM matrix (250 matrix multiplications of the 1 PAM matrix) is commonly used for sequence alignment software and to determine the accuracy of programs to construct evolutionary trees.

In our construction of a suitable amino acid replacement matrix for oligonucleotide cassette design, a computer algorithm using FutureBasic II (STAZ Software Inc.) was developed for matrix multiplication. This algorithm was used to construct a PAM matrix that would be a good predictor of acceptable amino acid replacement in the highly conserved histone H4 protein. Matrix multiplication of the 1 PAM matrix from Jones et al. (28) was used until the time when the first off-diagonal term was about equal to its corresponding. This PAM matrix is referred to as a the "PAM critical." For the PET91 matrix and associated relative mutability index, this equality occurred after 125 matrix multiplications (125 PAM) (Table 5.4). In the 125 PAM matrix, the replacement of methionine by leucine (2194) is about equal to the retention of methionine (2203) (Table 5.4).

Construction of "optimal" oligonucleotide cassettes

Using this 125 PAM matrix, we selected amino acid replacements that would generate substantial variability and simultaneously should yield functional proteins with a high probability. This amino acid replacement cassette was referred to as the "best cassette." Alternatively, a cassette was generated from the Jones 125 PAM matrix that would have a great deal of variability and yield functional proteins with a relatively low probability, this amino acid replacement cassette was referred to as the "worst cassette." Together, these two cassettes could be used to test the predictive power of the PAM matrix.

For the construction of oligonucleotide cassettes using the Jones 125 PAM matrix, the complete set of possible triplet nucleotides (3,375) was computed using standard IUPAC nucleotides A, T, C, G, R, Y, M, K, S, W, H, B, V, D, N, where R = A or G; Y = C or T; M = A or C; K = G or T; S = C or G; W = A or T; H = A, C or T; B = C, G or T; V = A, C or G; D = A, G or T; and N = A, C, G or T (each nucleotide is present in equal amounts). Most of the codons code for multiple amino acids. The different type of amino acids replacements associated with each multiple amino acid codon and the frequency of occurrence for each amino acid was determined. For each amino acid replacement, the probability of replacement of the original amino acid by each of the amino acids specified by the replacement codon was calculated. The PAM score for the replacement codon is 10,000 times the weighted mean of the probability of replacement of the original amino acid by the various amino acids specified by the replacement codon.

The standard deviation of the PAM scores gives a measure of the closeness with which the PAM replacements are clustered. An "optimal" nucleotide substitution cassette is one that maximizes diversity as well as generates a high probability of amino acid replacement yielding functional proteins. For optimal nucleotide substitution, a low standard deviation of the PAM scores is desired. Dividing the PAM score by the standard deviation of the PAM scores (M/SD) for a replacement codon provides a metric to compare replacement codons. It is expected that replacement codons with high M/SD will yield a relatively high percentage of functional proteins.

Although each codon generates the same set of amino acids and relative frequencies of occurrence, the relative merit of each codon is dictated primarily by the wild-type amino acid at the position that is being modified. For instance, a codon that generates arginine and lysine residues in equal abundance will have a very high PAM score at positions 39 and 40 where the wild-type amino acid is arginine. However, using the same codon at position 38, where the wild-type amino acid is alanine, will generate a lower PAM score since it is less expected that arginine and lysine will replace alanine.

Construction of "optimal" oligonucleotide cassettes using 125 PAM

Microsoft Excel 5.0 was used to sort through the 3,375 possible codons and their associated amino acids. All triplets that gave a possibility of a termination codon were removed from the set of codons. The list was further condensed by removing codon triplets that produced a set of less than 7 possible amino acids leaving codons that would give a wide range of possible amino acid replacements. For each amino acid, the remaining codons were sorted by PAM score and the associated metric M*SD or M/SD to determine the most suitable codon(s) for the replacement of that amino acid. Even though each codon generated a specific set of amino acids and relative frequencies, the average PAM score for the entire set is predicated on the amino acid being replaced. Using this approach, the codons RBT, RMM, DYC, and DYT were selected to replace the amino acids alanine (position 38), glycine (positions 41 and 42), valine (position 43) and isoleucine (position 46), respectively. These codons will generate the largest variability and highest viability (Figure 5.3). The codon RBT generates the amino acids threonine, alanine, serine, glycine, isoleucine, and valine with equal probability. RMM generates lysine, asparagine, glutamic acid, alanine, aspartic acid, and threonine in unequal probabilities. With RMM, threonine and alanine are represented with probability 0.25, while all other amino acids are represented with probability 0.125. DYC and DYT generate the amino acids phenylalanine, serine, valine, alanine, isoleucine, and threonine all with equal probability. Using these four codons at the five specified positions a space of 6 x 8 x 8 x 6 x 6 or 13,824 possible sequences, of which one is the wild-type sequence.

Testing "optimal" cassettes in *S. cerevisiae* histone H4

Optimal cassettes were tested for viability in the context of *S. cerevisiae* histone H4 using the same methods described in chapter 4 with the following exceptions. For PCR, the cassette "Jones-Sal" 5'-GCCAGCTATCCGTCGACTARBTAGAAGARMMRMM DYCAAGCGTDYTTCTGGT-3' was constructed (Integrated DNA Technologies). This primer was used in conjunction with the universal forward primer to PCR amplify the 3' half of *S. cerevisiae* histone H4 using pUC19 plasmid containing the full length histone H4 gene as a template. The Jones-Sal primer also introduces a Sal I site at the same position as described for primers in chapter 4. After ligation and transformation of plasmid containing the Jones-Sal cassette, yeast cells were initially grown on SG ura$^-$. Subsequent assays were made using dextrose (YPD) media.

Construction of a histone 32 PAM matrix

For the construction of the histone H4 specific pairwise exchange table, all complete known H4 sequences were downloaded from GenBank (Release 96.0) using BLAST (1). Redundant sequences from this set were eliminated, leaving 51 sequences that differed by at least one amino acid from every other sequence (Figure 5.4). By employing a user-derived phylogenetic tree in PAUP 3.1.1 (40), 22 of the 51 sequences were used to produce an additional 19 sequences representing ancestral node sequences on the tree (derived ancestral sequences) (Figure 5.5). Since histone H4 is highly conserved, contains no intervening sequences, and is a relatively small protein, the final set of 70 histone H4 sequences could be easily aligned by eye for the calculation of a pairwise exchange table (Table 5.5) and relative mutability index (Table 5.6).

The pairwise exchange table and relative mutability index were used to construct a 1 PAM specifically from the histone H4 sequence data (Table 5.7). Matrix multiplication software was used to derive the PAM critical for the histone H4 sequence data set (32 PAM; Table 5.8). In establishing a PAM critical, matrix multiplication was continued until the third amino acid, serine, had an off-diagonal term approximately equal to the on-diagonal term. Cysteine was the first amino acid for which the off-diagonal term was approximately equal to the on-diagonal term. However, there are very few cysteine residues in the histone H4 database and matrix multiplication was continued. Similarly, the off-diagonal term was approximately equal to the on-diagonal term for methionine relatively early in matrix multiplication. Multiplication was continued as methionine is also under represented in the histone H4 database (most of these methionine residues are initiation amino acids). The first well represented amino acid for which the off-diagonal term was approximately equal to the on-diagonal term was serine. At this point, matrix multiplication was halted and the resulting matrix was used as a PAM critical matrix. "Best" and "worst" cassettes were then calculated from the histone H4 amino acid replacement matrix for future analysis to test the predictive power of the histone H4 matrix relative to the PET91 matrix (Figure 5.3).

RESULTS

Amino acid replacement with the PET91 "best" cassette

To determine the percentage colonies containing a functional histone H4 sequences after transformation with the PET91 "best" cassette, 1,103 colonies growing on SG ura$^-$ media were transferred via sterile toothpick to YPD media. These cells were grown for 2 days at 28°C until and colonies formed. Of the 1,103 colonies, 101 grew successfully on YPD (9.2%). This ratio is surprisingly high with respect to the extreme conservation found in this region. Using the "best" codons derived from the PET91 matrix, the total number of sequences generated is 13,824. From the ratio of viability, 9.2% of these sequences (1,272 sequences) are functional in yeast.

Four of the colonies found to be viable on YPD media were chosen at random for PCR sequencing using the methods described in chapter 4. The wild-type amino acid sequence in this region is alanine (A), glycine (G), glycine (G), valine (V), and isoleucine

(I) at positions 38, 41, 42, 43, and 46 respectively. Each of the four colonies gave a different histone H4 sequences (Table 5.9). Based on this very limited amount of sequencing, it appears that of the five positions the isoleucine at position 46 is the most conserved. The glycine and valine at positions 42 and 43 are also found in three of the four sequences. Positions 38 and 41 appear most free to vary. At position 38, alanine is replaced by either glycine or valine and at position 41, glycine is replaced by threonine or alanine. Most striking is the successful replacement of valine by phenylalanine at position 43 in one of the four sequences.

Construction of other oligonucleotide cassettes

Construction of non-optimal codons for substitution follows the same approach as for optimal codons, but with the a requirement for low PAM scores instead of high PAM scores. The appropriate metric, in this case, is the mean PAM multiplied by the standard deviation (M*SD). Nucleotide substitutions generated under these conditions will provide a tight cluster of replacement amino acids that give low predicted PAM scores. Such replacements should yield a relatively low percentage of functional proteins. In selecting non-optimal nucleotide substitutions, termination codons (stops) and very improbable amino acid replacements should be avoided as they dramatically reduce the fraction of potentially functional protein sequences. The "worst" codons generated for histone H4 positions 38, 41, 42, 43, and 46 are shown in Figure 5.3. Selection of optimal and non-optimal cassettes for positions 38, 41, 42, 43, and 46 was based on the histone H4 amino acid replacement database. These codons are also shown in Figure 5.3.

DISCUSSION

Dayhoff and Eck (15) and Dayhoff et al. (16,17) suggested that the stability of proteins is less affected by replacement of amino acids with similar physico-chemical properties than by those with different physico-chemical properties. Feng et al. (20) also suggested that the most common evolutionary changes in proteins are replacements between amino acids of similar structure or between amino acids that can be exchanged by single-nucleotide substitutions. It has been generally demonstrated that amino acid positions that have a large effect on protein structure or function are conserved over evolutionary history. Positions that do not effect the protein structure are relatively free to vary and can accept a wider variety of non-wild-type amino acid replacements.

Delagrave and Youvan (19) suggested that an optimal mutagenesis strategy must simultaneously alter as many residues as possible while still generating the largest number of functional proteins. Non-functional proteins only serve to increase the number of sequences that must be analyzed before arriving at a functional modification. Several methods have been developed with this "optimal" search strategy in mind including combinatorial cassette mutagenesis (CCM), which uses all 20 amino acids at each position (33,34); target set mutagenesis (TSM) (23); recursive ensemble mutagenesis (REM) (18); and exponential ensemble mutagenesis (EEM) (19). Each of these strategies employ some type of simultaneous mutagenesis strategy to increase the probability of generating functional mutants while avoiding extensive random mutagenesis at only one location (19). Sequence space is too large to effectively search without a simultaneous variation of regions of interest.

In an effort to limit the search space, Arkin and Youvan (4) developed strategies based on Dayhoff and Eck (15) and Dayhoff et al. (16,17) that would yield amino acid replacements which were physico-chemically similar to the wild-type sequence. This strategy should limit the potential search space to amino acid replacements which would yield a high probability of functional modifications. Each set of nucleotide substitutions was considered "optimal" if the resulting nucleotide mixture maximized the probability of obtaining each of the amino acids in the desired set with equal probability (4). Such nucleotide substitution schemes could be developed for any position in the protein.

In an extension of this method, Goldman and Youvan (23) used 29 homologous light harvesting II (LHII) proteins as a database to construct a target set of seven positions for mutagenesis of the *Rhodobacter capsulatus* LHII protein. Specific mixtures of nucleotide substitutions were chosen to maximize the probability of the generating replacements previously determined to produce functional proteins. Mixing and matching combinations of replacements previously identified to be functional were used, thereby increasing the probability of producing new functional proteins. In comparison to conventional CCM techniques, Goldman and Youvan (23) demonstrated that this target set mutagenesis showed a 100-600 fold improvement in the yield of functional proteins when compared to random cassettes techniques (using [NNN] at each position).

In cases where no phylogenetic sequence data is available, recursive ensemble mutagenesis (REM) can be used to generate a "pseudophylogeny" (44). After one iteration of conventional CCM using [NNS] codons and sequence analysis of positive mutants, viable replacements can then be used as a target set in the same way phylogenetic information is used in TSM (18). The codon NNS generates all 20 amino acids using the genetic code and only gives the possibility of one stop signal (UAG). Exponential ensemble mutagenesis (EEM) uses a similar method to generate libraries with a high probability of function alterations (19). In EEM, non-overlapping groups of 5 to 6 amino acids are randomized in the first round using the codon [NNS] and functional mutants are discovered. The successful mutants from each set of 5 to 6 amino acids are then combined into larger non-overlapping groups and the positions re-mutagenized. Theoretically, entire proteins could be altered using this bottom-up approach by building up successful alterations. Stemmer et al. (39) criticized the EEM technique for its lack of ability to search for novel long range interactions due to the limits of small cassettes. Combinatorial multiple cassette mutagenesis (CMCM) (13) was developed to alleviate some of these concerns.

A key trade off to any successful search strategy of sequence space is to increase efficiency by increasing the number of functional proteins (4), while exploring as much variability as possible and increasing the search space to the upper limits of a transformational capacity. Any amino acid replacement matrix could be used to determine non-functional and functional modifications that have occurred over evolutionary history. To make the TSM method more "global" as a search strategy, the PET91 matrix derived by Jones et al. (28) was used in this study as an index of functional and non-functional replacements that have occurred over evolutionary history.

Viable modifications in *S. cerevisiae* histone H4

By using the Jones-Sal oligonucleotide cassette, it was determined that 9.2% of the sequences from this set are viable in the context of the *S. cerevisiae* histone H4. Using the codons for this optimal cassette, 13,824 potential sequences are generated. 9.2% or 1,272 of these sequences are viable in yeast. This is a suprisingly high number with respect to the absolute conservation of these residues over evolutionary history. However, this cassette was developed using knowledge of successful replacements in a wide variety of proteins (the PET91 matrix).

Limited sequencing of four viable colonies gave rise to four new sequences in the context of *S. cerevisiae* histone H4 (Table 5.9). The isoleucine at position 46 is conserved in all four sequences. However, it is quite surprising that the amino acids threonine, aspartic acid, and phenylalanine can replace the two glycines and one valine at positions 41 through 43. The glycine and valine at positions 42 and 43 are also found in three of the four sequences. Positions 38 and 41 appear most free to vary. It is interesting to note that positions 38 and 41 each have two successful replacements, positions 42 and 43 have one successful replacement and position 46 has no successful replacements. The fact that each of the four colonies sequenced has different replacements to this region and that multiple amino acid replacements are observed in single sequences suggests that our ratio is not biased by the unlikely inclusion of false positives. It will be interesting to characterize a larger set of these colonies for sequence and corresponding growth rate relative to yeast utilizing the wild-type histone H4.

Using "best" and "worst" cassettes derived from the PET91

Using the PET91, the "best" codons for the positions that we wish to vary in histone H4 were determined. A set of "worst" codons that would be predicted to generate the lowest PAM score for those positions were also determined. A comparison between these two cassettes should shed light on the predictive power of the PET91 matrix. The ratio of viable and non-viable colonies on dextrose media could be correlated to the average PAM score for each set of codons. It is expected that the "worst" codons will have a sufficiently different ratio to the "best" codons such that the difference in the ratios can be observed by testing 1,000 colonies.

A test for the predictive power of alternate PAM matrices

Cassettes derived with the PET91 matrix contain codons with PAM scores relative to a very large set of diverse proteins in the SWISS-PROT database. When making modifications to a specific protein, it is unclear whether such a "global" matrix will be a better or worse predictor of successful replacements than a matrix developed from sequences for the specific protein of interest. Using Dayhoff's common ancestor method, an aligned set of 70 histone H4 proteins was generated. This set was used to determine a PAM critical matrix that can be used for the determination of PAM scores and appropriate codons for amino acid replacement at any position.

Using cassettes derived from the H4 PAM critical, the percentage of the cassette sequences that yield viable cells can be compared with relative efficiency of amino acid replacement predicted from the PET91 matrix. In essence, this strategy tests the predictive ability of a "global" matrix (in the case of the Jones PAM matrix) relative to a "local" matrix (in this case histone PAM matrix). Since histone H4 is an extremely conserved protein, the frequency and type of amino acid replacements might be expected to be very different from the total set of proteins in the SWISS-PROT database.

Comparing cassettes with the largest PAM score differential

A PAM matrix can be used as a predictor of functional and non-functional alterations while designing cassettes for mutagenesis strategies. The predictive power of this and other amino acid matrices can be tested by constructing a series of oligonucleotide cassettes that introduce alterations as dictated by the matrix. The relative predictive power of two different matrices (for instance, the PET91 and the H4 PAM) can similarly be tested by using cassettes containing nucleotide substitution schemes that exploit the difference between codons predicted to be successful by one matrix and not successful by another. With a wide discrepancy in PAM scores, differences in the ratio of colony viability will be easy to determine.

Modifications in different histone H4 contexts

In chapter 4, a series of oligonucleotide cassettes were tested in both the context of the *S. cerevisiae* and *T. thermophila* histone H4. A surprising difference was discovered in the set of viable sequences between these two contexts. It would therefore be interesting to test cassettes derived with both the PET91 and H4 PAM in the context of both the *S. cerevisiae* and *T. thermophila* histone H4. The positions that were modified using the PET91 matrix in this chapter flank the regions that were modified in chapter 4. 9.2% of these PET91 derived sequences work in the context of the *S. cerevisiae* histone H4 protein. Similar modifications to the *T. thermophila* histone H4 context might be expected to give a different set of viable modifications as was the case for positions 39, 40, 44, and 45.

Conclusions

The use of a PAM matrix for the development of oligonucleotide cassettes was demonstrated. The PAM matrix can be used to develop cassettes that are predicted to have a high probability or a low probability of success in the protein. Using the PET91 matrix, a cassette predicted to have a high probability of successful replacement of five positions in a highly conserved region of *S. cerevisiae* histone H4 was tested. Our results suggest that this method can be used to efficiently search for novel functional proteins. We have also described techniques that can be used to test the predictive power of both the PET91 matrix and other matrices developed more specifically for the protein that is to be modified by using yeast viability ratios as a metric of success. For the histone H4 protein, it is unclear which matrix (PET91 or H4 PAM) will be more useful as a predictor of successful replacement. Future mutagenesis experiments will help to clarify these issues.

REFERENCES

1. Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. Journal of Molecular Biology. 215, 403-10.

2. Arents, G. and E. N. Moudrianakis. 1993. Topography of the histone octamer surface - repeating structural motifs utilized in the docking of nucleosomal DNA. Proc. Natl. Acad. Sci. USA. 90(22):10489-10493.

3. Arents, G. and E. N. Moudrianakis. 1995. The histone fold - a ubiquitous architectural motif utilized in DNA compaction and protein dimerization. Proc. Natl. Acad. Sci. USA. 92(24):11170-11174.

4. Arkin, A. P. and D. C. Youvan. 1992. Optimizing nucleotide mixtures to encode specific subsets of amino acids for semi-random mutagenesis. Bio/Technology. 10:297-300.

5. Bairoch, A. and A. Boeckmann. 1993. The SWISS-PROT protein sequence data bank, recent developments. Nucleic Acids Research. 21(13):3093-3096.

6. Bakir, U., P. M. Coutinho, P. A. Sullivan, C. Ford, and P. J. Reilly. 1993. Cassette Mutagenesis of *Aspergillus awamori* glucoamylase near its general acid residue to probe its catalytic and pH properties. Protein Engineering. 6(8):939-946.

7. Barbas, C. F. III, J. D. Bain, D. M. Hoekstra, and R. A. Lerner. 1992. Semisynthetic combinatorial antibody libraries: A chemical solution to the diversity problem. Proc. Natl. Acad. Sci. USA. 89:4457-4461.

8. Beard, W. A., D. T. Minnick, C. L. Wade, R. Prasad, R. L. Won, A. Kumar, T. A. Kunkel, and S. H. Wilson. 1996. Role of the "helix clamp" in HIV-1 reverse transcriptase catalytic cycling as revealed by alanine-scanning mutagenesis. Journal of Biological Chemistry. 271(21):12213-12220.

9. Binder, M., S. Ortner, B. Plaimauer, M. Fodinger, G. Wiedermann, O. Scheiner, and M. Duchene. 1995. Sequence and organization of an unusual histone H4 gene in the human parasite *Entamoeba histolytica*. Mol. Biochem. Parasitol. 71(2):243-247

10. Black, M. E. and L. A. Loeb. 1993. Identification of important residues within the putative nucleoside binding site of HSV-1 thymidine kinase by random sequence selection: analysis of selected mutants in vitro. Biochemistry. 32:11618-11626.

11. Black, M. E., T. G. Newcomb, H.-M. P. Wilson, and L. A. Loeb. 1996. Creation of drug-specific herpes simplex virus type 1 thymidine kinase mutants for gene therapy. Proc. Natl. Acad. Sci. USA. 93:3525-3529.

12. Bowie, J. U., R. Lüthy, and D. Eisenberg. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. Science. 253:164-169.

13. Crameri, A. and W. P. C. Stemmer. 1995. Combinatorial multiple cassette mutagenesis creates all the permutations of mutant and wild-type sequences. BioTechniques. 18(2):194-196.

14. Cupples, C. G., J. H. Miller, and R. E. Huber. 1990 Determination of the roles of Glu-461 in β-galactosidase (*Escherichia coli*) using site-specific mutagenesis. Journal of Biological Chemistry. 265(10):5512-5518.

15. Dayhoff, M. O. and R. V. Eck. (Eds.): Atlas of Protein Sequence and Structure, vol. 3, Silver Spring: National Biomedical Research Foundation. Washington D.C. 1968.

16. Dayhoff, M. O., R. V. Eck and C. M. Park. *in* Atlas of Protein Sequence and Structure Vol. 5. (ed. Dayhoff, M. O.), pp.89-99. National Biomedical Research Foundation. Washington D.C. 1972.

17. Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. *in* Atlas of Protein Sequence and Structure Vol. 5, Suppl 3. (ed. Dayhoff, M. O.) pp.345-358. National Biomedical Research Foundation. Washington D.C. 1978.

18. Delagrave, S., E. R. Goldman and D. C. Youvan. 1993. Recursive ensemble mutagenesis. Protein Engineering. 6(3):327-331.

19. Delagrave, S. and D. C. Youvan. 1993. Searching sequence space to engineer proteins: Exponential ensemble mutagenesis. Bio/Technology. 11:1548-1552.

20. Feng, D. F., M. S. Johnson and R. F. Doolittle. 1985. Aligning amino acid sequences: Comparison of commonly used methods. Journal of Molecular Evolution. 21:112-125.

21. Fitch, W. M. and E. Margoliash. 1967. Construction of phylogenetic trees. Science. 115:279-284.

22. Fogel, G. B. and C. F. Brunk. 1997. Expression of *Tetrahymena* histone H4 in yeast. Biochemica et Biophysica Acta. 1354:116-126.

23. Goldman, E. R. and D. C. Youvan. 1992. An algorithmically optimized combinatorial library screened by digital imaging spectroscopy. Bio/Technology. 10:1557-1561.

24. Gram, H., L.-A. Marconi, C. F. Barbas, T. A. Collet, R. A. Lerner and A. S. Kang. 1992. *In vitro* selection and affinity maturation of antibodies from a naive combinatorial immunoglobulin library. Proc. Natl. Acad. Sci. USA. 89:3576-3580.

25. Henikoff, S. 1996. Scores for sequence searches and alignments. Current Opinion in Structural Biology. 6:353-360.

26. Henikoff, S. and J. G. Henikoff. 1992. Amino acid substitution matrices from protein building blocks. Proc. Natl. Acad. Sci. USA. 89:10915-10919.

27. Houghten, R. A., C. Pinilla, S. E. Blondelle, J. R. Appel, C. T. Dooley and J. H. Cuervo. 1991. Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery. Nature. 354:84-86.

28. Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Computer Applications in the Biological Sciences. 8(3):275-282.

29. Kleina, L. G and J. H. Miller. 1990 Genetic studies of the lac repressor. XIII. Extensive amino acid replacements generated by the use of natural and synthetic nonsense suppressors. Journal of Molecular Biology. 212:295-318.

30. Lam, K. S., S. E. Salmon, E. M. Hersh, V. J. Hruby, W. M. Kazmiersky and R. J. Knapp. 1991. A new type of synthetic peptide library for identifying ligand-binding activity. Nature. 354:82-84.

31. Loeb, D. D., R. Swanstrom, L. Everitt, M. Manchester, S. E. Stamper, C. A. Hutchison. 1989. Complete mutagenesis of the HIV-1 protease. Nature. 340:397-400.

32. Markiewicz, P., L. G. Kleina, C. Cruz, S. Ehret, J. H. Miller. 1994. Genetic studies of the lac repressor. XIV. Analysis of 4000 altered *Escherichia coli* lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. Journal of Molecular Biology. 240:421-433.

33. Oliphant, A. R., A. L. Nussbaum, and K. Struhl. 1986. Cloning of random-sequence oligodeoxynucleotides. Gene. 44:177-183.

34. Reidhaar-Olson, J. F., J. U. Bowie, R. M. Breyer, J. C. Hu, K. L. Knight, W. A. Lim, M. C. Mossing, D. A. Parsell, K. R. Shoemaker and R. T Sauer. 1991. Random mutagenesis of protein sequences using oligonucleotide cassettes. Methods Enzymol. 208:564-587.

35. Rennell, D., S. E. Bouvier, L. W. Hardy, and A. R. Poteete. 1991. Systematic mutation of bacteriophage T4 lysozyme. Journal of Molecular Biology. 222:67-87.

36. Resnick, D. A., A. D. Smith, A. Zhang, S. C. Geisler, E. Arnold, and G. F. Arnold. 1994. Libraries of human rhinovirus-based HIV vaccines generated using random systematic mutagenesis. AIDS Research and Human Retroviruses. 10(suppl 2):S47-S52.

37. Roberts, B. L., W. Markland, A. C. Ley, R. B. Kent, D. W. White, S. K. Guterman, and R. C. Ladner. 1992. Directed evolution of a protein: Selection of potent neutrophil elastase inhibitors displayed on M13 fusion phage. Proc. Natl. Acad. Sci. USA. 89:2429-2433.

38. Sadler, L. A. and C. F. Brunk. 1992. Phylogenetic relationships and unusual diversity in histone H4 proteins within the *Tetrahymena pyriformis* complex. Molecular Biology and Evolution. 9(1):70-84.

39. Stemmer, W. P. C. 1995. Searching sequence space: Using recombination to search more efficiently and thoroughly instead of making bigger combinatorial libraries. Bio/Technology. 13:549-553.

40 Swofford, D. 1993. Phylogenetic analysis using parsimony (PAUP) ver. 3.1.1. Smithsonian Institution.

41. Tomii, K. and M. Kanehisa. 1996. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Engineering. 9(1):27-36.

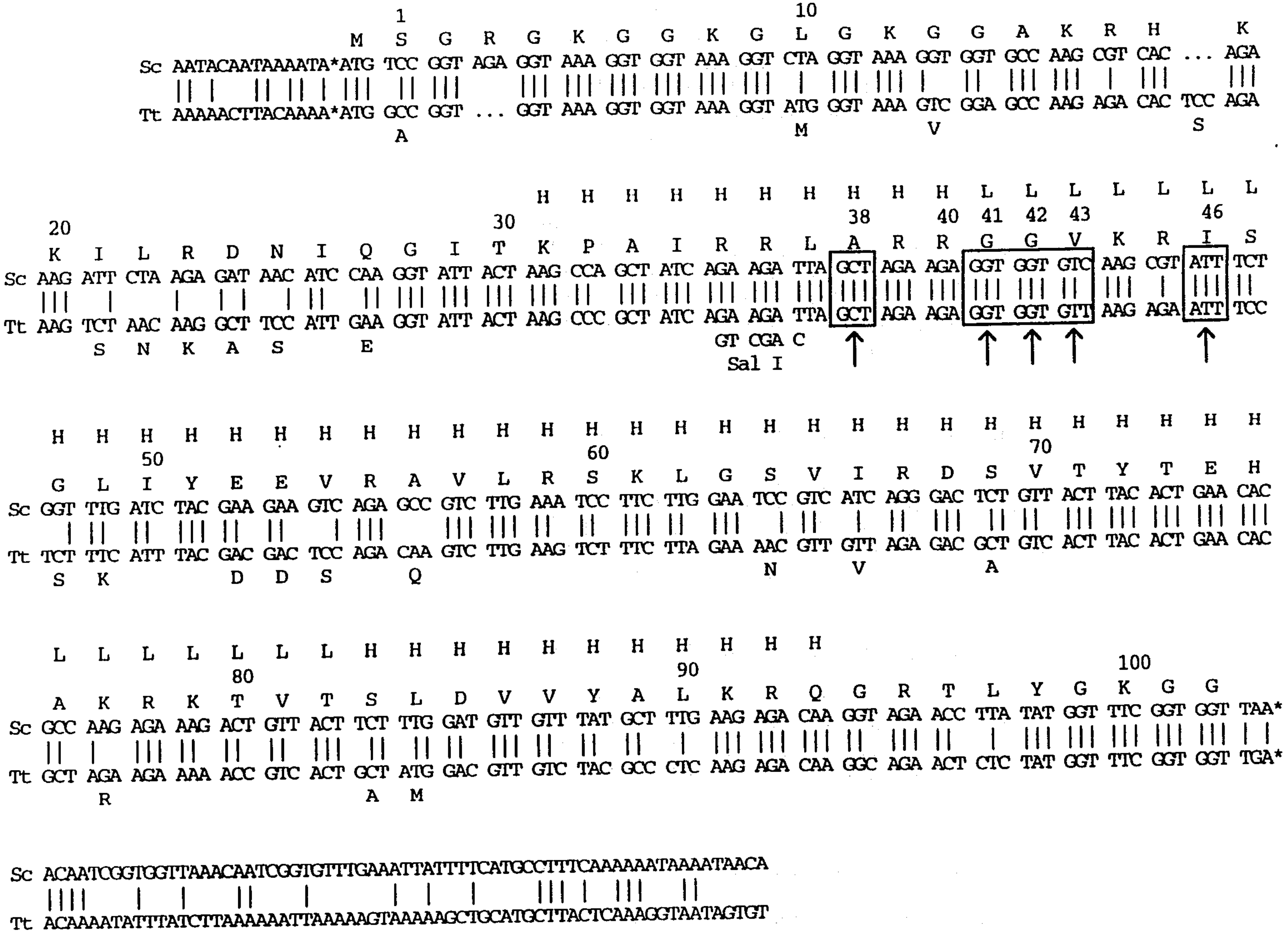
42. Vogt, G., T. Etzold, and P. Argos. 1995. An assessment of amino acid exchange matrices in aligning protein sequences: The twilight zone revisited. Journal of Molecular Biology. 249:816-831.

43. Wang, B.-C., J. Rose, G. Arents and E. N. Moudrianakis. 1994. The octameric histone core of the nucleosome - structural issues resolved. Journal of Molecular Biology 236(1):179-188.

44. Youvan, D. C. 1995. Searching Sequence Space. Bio-Technology. 13(8):722-723.

Figure 5.1

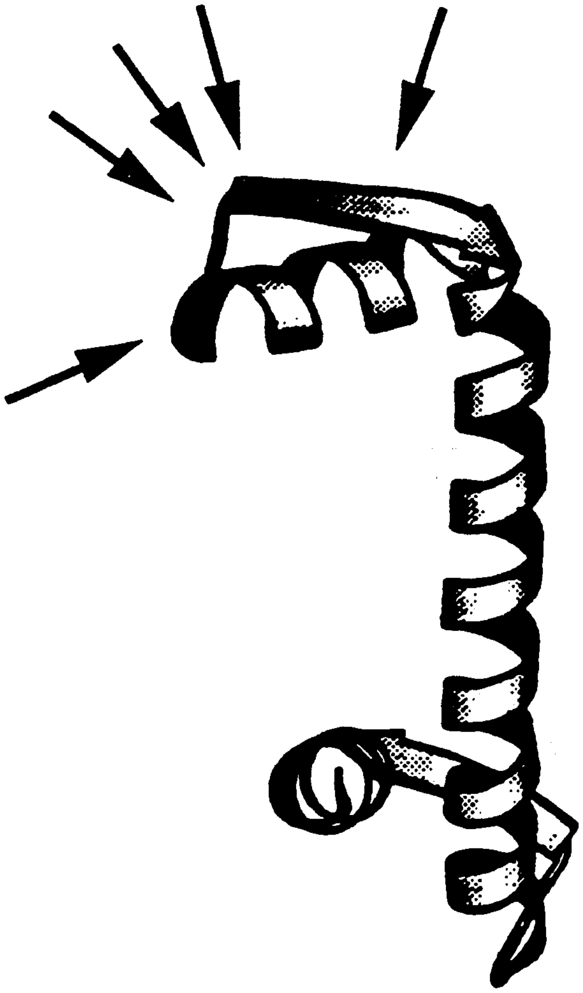
Histone H4 sequences for *Saccharomyces cerevisiae* (Sc) and *Tetrahymena thermophila* (Tt). The amino acid sequences for each codon are given above and below the nucleotide sequences for each species. Boxes indicate the alanine at position 38, glycines at positions 41 and 42, valine at position 43 and isoleucine at position 46 which were modified using a cassette derived from the Jones PET91 matrix. The nucleotides modified to create a Sal I restriction site used for vector construction are shown. Arrows indicate codons which were modified on the basis of highest average PAM score. Indications of helix (H) and loop (L) are given for the amino acids involved with the histone fold.



```
                        1                                       10
                    M   S   G   R   G   K   G   G   K   G   L   G   K   G   G   A   K   R   H       K
Sc AATACAATAAAATA*ATG TCC GGT AGA GGT AAA GGT GGT AAA GGT CTA GGT AAA GGT GGT GCC AAG CGT CAC ... AGA
   ||| |  || || | |||  || |||     ||| ||| ||| ||| ||| |||  |  ||| |||  |   || ||| |||  |  |||     |||
Tt AAAAACTTACAAAA*ATG GCC GGT ... GGT AAA GGT GGT AAA GGT ATG GGT AAA GTC GGA GCC AAG AGA CAC TCC AGA
                        A                                   M           V                       S

                                               H   H   H   H   H   H   H   H   H   H   L   L   L   L   L   L   L
                                                                                   38      40  41  42  43          46
    20                                         30
    K   I   L   R   D   N   I   Q   G   I   T   K   P   A   I   R   R   L   A   R   R   G   G   V   K   R   I   S
Sc AAG ATT CTA AGA GAT AAC ATC CAA GGT ATT ACT AAG CCA GCT ATC AGA AGA TTA GCT AGA AGA GGT GGT GTC AAG CGT ATT TCT
   |||  |       |  | |  | ||  || ||| ||| ||| ||| ||  ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||  |||  |  |||  ||
Tt AAG TCT AAC AAG GCT TCC ATT GAA GGT ATT ACT AAG CCC GCT ATC AGA AGA TTA GCT AGA AGA GGT GGT GTT AAG AGA ATT TCC
        S   N   K   A   S       E                                   GT CGA C  ↑           ↑   ↑   ↑           ↑
                                                                      Sal I

    H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H
            50                                      60                                      70
    G   L   I   Y   E   E   V   R   A   V   L   R   S   K   L   G   S   V   I   R   D   S   V   T   Y   T   E   H
Sc GGT TTG ATC TAC GAA GAA GTC AGA GCC GTC TTG AAA TCC TTC TTG GAA TCC GTC ATC AGG GAC TCT GTT ACT TAC ACT GAA CAC
     | ||  ||  ||| ||  ||    | |||     ||| ||| ||  ||  ||| ||  |||   | ||   |  ||  |||  ||  || ||| ||| ||| ||| |||
Tt TCT TTC ATT TAC GAC GAC TCC AGA CAA GTC TTG AAG TCT TTC TTA GAA AAC GTT GTT AGA GAC GCT GTC ACT TAC ACT GAA CAC
    S   K           D   D   S       Q                               N       V           A

    L   L   L   L   L   L   L   H   H   H   H   H   H   H   H   H   H   H
                    80                                  90                                  100
    A   K   R   K   T   V   T   S   L   D   V   V   Y   A   L   K   R   Q   G   R   T   L   Y   G   K   G   G
Sc GCC AAG AGA AAG ACT GTT ACT TCT TTG GAT GTT GTT TAT GCT TTG AAG AGA CAA GGT AGA ACC TTA TAT GGT TTC GGT GGT TAA*
   ||   |  ||| ||  ||  ||  |||  ||  || ||  ||| ||  ||  ||   |  ||| ||| ||| ||  ||| ||   |  ||| ||| ||| ||| ||| | |
Tt GCT AGA AGA AAA ACC GTC ACT GCT ATG GAC GTT GTC TAC GCC CTC AAG AGA CAA GGC AGA ACT CTC TAT GGT TTC GGT GGT TGA*
        R                       A   M

Sc ACAATCGGTGGTTAAACAATCGGTGTTTGAAATTATTTTCATGCCTTTCAAAAAATAAAATAACA
   ||||    |   |   ||   |       | |  |    ||| | |||  ||
Tt ACAAAATATTTATCTTAAAAAATTAAAAAGTAAAAAGCTGCATGCTTACTCAAAGGTAATAGTGT
```

Figure 5.2

Tertiary structure of the histone fold from histone H4. Positions that were modified in this chapter are shown by arrows.



H4

Figure 5.3

Codons derived using two amino acid replacement matrices (Jones PET91 and histone H4 PAM) for each of five positions in the histone H4 gene. The optimal codons, number of amino acids, and the average PAM value for that set of amino acids is given for each cassette ("best" or "worst") and for each specified amino acid. Average PAM values for the complete sequence is given below followed by the number of sequences produced by the codons.

	Jones Best	Jones Worst	H4 PAM Best	H4 PAM Worst
Alanine (position 38)	RBT 6 0.7532	HKK/HKS 12 0.5266	MHK/MHS 12 0.3265	MNR 16 0.1702
Glycine (position 41,42)	RMM 8 0.6241	NWC/NWT 8 0.2191	MHK/MHS 12 0.1422	YNT 8 0.1739
Valine (position 43)	DYC 6 0.6131	VRV/VRD 18 0.2935	MNG 8 0.1471	YNC/YNT 8 0.1053
Isoleucine (position 46)	DYT 6 0.7222	RRV/RRD 12 0.2442	MNG 8 0.2551	YNC/YNT 8 0.1159
Average PAM	1.3×10^{-1}	1.8×10^{-3}	2.4×10^{-4}	6.3×10^{-5}
Total Sequences	13,824	165,888	110,592	65,536

Figure 5.4

Aligned set histone H4 sequences used to create a pairwise exchange table and relative mutability index specific for histone H4. All histone H4 sequences were downloaded from GenBank (Release 96.0) and common sequences combined to give 51 unique sequences. The highly divergent *Entameoba histolytica* sequence (9) listed as the last entry in the alignment was not used in subsequent analysis. For the user derived tree, only *T. paravorax* and *T. thermophila* as representatives for a much larger set of *Tetrahymena* histone H4 sequences. 19 internal nodes were derived using PAUP and are numbered 1-19. All known unique histone H4 sequences, including internal node sequences, were aligned by eye. Dashes represent gaps in alignment and X represents missing data from the sequences. The histone 1 PAM was derived from 70 unique sequences.

```
1        M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKVFLENVIRDAVTYCEHAKRKTVTAMDVVYALKRQGRTLYGFGG--
2        M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKVFLENVIRDAVTYTEHAKRKTVTAMDVVYALKRQGRTLYGFGG--
3        M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKVFLENVIRDAVTYTEHAKRKTVTAMDVVYALKRQGRTLYGFGG--
4        M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKIFLENVIRDAVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG--
5        M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRTVLKNFLENVIRDSVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG--
6        M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKIFLENVIRDAVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG--
7        M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKIFLENVIRDAVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG--
8        M-----SGRGKGG---KG--LGKG--GA-KRH-RKILRDNIQGITKPAIRRLARRGGVKRISAMIYEETRGVLKSFLESVIRDAVTYTEHAKRKTVTSLDVVYALKRQGRTLYGFGG--
9        M-----SGRGKGG---KG--LGKG--GA-KRH-RKILRDNIQGITKPAIRRLARRGGVKRISALIYEETRAVLKSFLESVIRDAVTYTEHAKRKTVTSLDVVYALKRQGRTLYGFGG--
10       M-----SGRGKGG---KG--LGKG--GA-KRH-RKILRDNIQGITKPAIRRLARRGGVKRISALIYEETRGVLKSFLESVIRDAVTYTEHAKRKTVTSLDVVYALKRQGRTLYGFGG--
11       M-----SGRGKGG---KG--LGKG--GA-KRH-RKILRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKIFLENVIRDAVTYTEHAKRKTVTALDVVYALKRQGRTLYGFGG--
12       M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKIFLENVIRDAVTYTEHAKRKTVTAMDVVYALKRQGRTLYGFGG--
13       M-----AG-GKGG---KG--MGKV--GA-KRHSRKSNKASIEGITKPAIRRLARRGGVKRISSFIYDDSRQVLKSFLENVVRDAVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG--
14       M-----AGRGKGG---KG--YGKV--GA-KRHTKKSLKETIEGITKPAIRRLARRGGVKRISSLIYEETRNVLKSFLENVIRDSVTYTEHAKRKTVTALDVVYALKRQGRTLYGFGG--
15       M-----AGRGKGG---KG--YGKV--GA-KRHTKKALRETIEGITKPAIRRLARRGGVKRISSLIYEETRAVLKGFLENVIRDSVTYTEHAKRKTVTALDVVYALKRQGRTLYGFGG--
16       M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKIFLENVIRDAVTYTEHAKRKTVTAMDVVYALKRQGRTLYGFGG--
17       M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKVFLENVIRDAVTYTEHAKRKTVTAMDVVYALKRQGRTLYGFGG--
18       M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKVFLENVIRDAVTYTEHAKRKTVTAMDVVYALKRQGRTLYGFGG--
19       M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKVFLENVIRDAVTYTEHAKRKTVTAMDVVYALKRQGRTLYGFGG--
E.cra    M-----PGRGKGGKGGKG--YGKV--GA-KRHTKKALRETILGVTKPAIRRLARRGGVKRISSLVYEETRAVLKGFLENVIRDSVTYTEHAKRKTVTALDVVYALKRQGKTLYGFGG--
E.cra    M-----PGRGKGGKGGKG--YGKV--GA-KRHAKKALRETILGVTKPAIRRLARRGGVKRISSLVYEETRAVLKGFLESVIRDSVTYTEHAKRKTVTALDVVYALKRQGKTLYGFGG--
G.cha    M-----APGGKGG---KG--MGKV--GA-KRHTKRTNKASIEGIT--------------------------------------------------------------------------
O.nov    M-----AGRGKVG---KG--YGKV--GA-KRHTKKSLKETIMGITKPAIRRLARRGGVKRISSLIYEETRNVLRSFLENVIRDSVTYTEHAKRKTVTALDVVYALKRQGRTLYGFGG--
T.asi    M-----VGG-KGG---KG--MGKV--GA-KRHSKRSNKASIEGITKPAIRRLARRGGVKRISSFTYDDSRQVLKSFLENVVRDAVTYTEHARRKTVTAMDVVYALKRQ-----------
T.set    M-----AGG-KGG---KG--MGKV--GA-KRHSKRSNKASIEGITKPAIRRLARRGGVKRISSFTYDDSRQVLKSFLENVVRDAVTYTEHARRKTVTAMDVVYALKRQ-----------
T.ell    M-----AGG-KGG---KG--MGKV--GA-KRHSRKSNKASIEGITKPAIRRLARRGGVKRISSFTYDDSRQVLKSFLENVVRDAVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG--
T.set    ----------------------KV--GA-KRHSRKSNKASIEGITKPAIRRLARRGGVKRISSFTYDDSRQVLKSFLESVVRDAVTYTEHARRKTVTAMDVVYALKRQ-----------
T.set    ----------------------KV--GA-KRHSRKSNKASIEGITKPAIRRLARRGGVKRISSFTYDDSRQVLKSFQENVVRDAVTYTEHARRKTVTAMDVVYALKRQ-----------
T.set    -----------------------------------------------------------------------------------------------------------------------
T.ell    M-----AGG-KGG---KG--MGKV--GA-KRHSRRSNKASIEGIT--------------------------------------------------------------------------
T.ell    ----------------------KV--GV-KRHSRKSNKASIEGITKPAIRRLARRGGVKRISSFTYDDSRQVLKSFLENVVRDAVTYTEHARRKTVTAMDVVYALKRQ-----------
T.ame    ----------------------KV--GA-KRHSKRSNKASIEGITKPAIRRLAKRGGVKRISSFTYDDSRQVLKSFLENVVRDAVTYTEHARRKTVTAMDVVYALKRQ-----------
T.mal    -----------------------------------------------------------------------------------------------------------------------
T.par    -----------------------------------------------------------------------------------------------------------------------
T.pat    M-----AGG-KGG---KG--MGKV--GA-KRHSRKSNKASIEGVT--------------------------------------------------------------------------
T.pyr    M-----PSG-KGG---KG--MGKV--GA-KRHARKSNKASIEGIT--------------------------------------------------------------------------
T.ros    M-----VGG-KGG---KG--MGKV--GA-KRHSKRSTKASIEGITKPAIRRLARRGGVKRISSFTYDDSRQVLKSFLENVVRDAVTYTEHARRKTVTAMDVVYALKRQ-----------
T.ros    M-----AGG-KGG---KG--MGKV--GA-KRHSKRSNKASIEGIIKPAIRRLARRGGVKRISSFTYDDSRQVLKSFLENVVRDAVTYTEHARRKTVTAMDVVYALKRQ-----------
A.lumb   M-----AGG-KGG---KG--MGKV--GA-KRHSKRSNKASIEGITKPAIRRLARRGGAKRISSFTYDDSRQVLKSFLENVVRDAVTYTEHARRKTVTAMDVVYALKRQ-----------
D.mel    M-----PSG-KGG---KG--MGKV--GA-KRHSKRSNKASIEGITKPAIRRLARRGGVKRISSFTYDDSRQVLKSFLENVVRDAVTYTEHARRKTVTAMDVVYALKRQ-----------
A.for    M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRSVLKVFLENVIRDAVTYCEHAKRKTVTAMDVVYALKHQGRTLYGFGG--
S.pur    M-----TGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKVFLENVIRDAVTYTEHAKRKTVTAMDVVYALKRQGRTLYGFGG--
S.sti    M-----SGRGKGG---KG--LGKG--GA-KRH-RKILRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKVFLENVIRDAVTYTEHAKRKTVTAMDVVYALKRQGRTLYGFGG--
P.liv    M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKVFLENVIRDAVTYCEHAKRKTVTAMDVVYALKRQGRTLYGFGG--
         M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKVFLENVIRDAVTYCEHAKRKTVTSMDVVYALKRQGRTLYGFGG--
         M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITLPAIRRL-------------------------------------------------------------------
```

```
S.pil   M-----SGRGKGG---KG--LGKG--GA-KRH-RKVIRDNIQGITKPATRRLARRGGVKRISGLIYEETRGVLKVFLENVIRDAVTYTEHAKRKTVTALDVVYALKRQGRTLYGFGG--
X.lae   M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKVFLENVIRDAVTYTEHAKRKTVTAMDVVYALKRQGRTLYGFGG--
M.mus   M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKVFLENVIRDAVTYTEHAKRKTVTAMDVVYRLKRQGRTLYGFGG--
H.sap   M-----SGOGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRXARRGGVKRISGLIYEETRGVLKVFLENVIRDAVTYTEHAKRKTVTAMDVVYALKRQGXTLYGFGG--
H.sap   M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIHGITKPAIRRL-RRGGVKRISGLIYEETRGVLKVFLENVIRDAVTYTEHAKRKTVTAMDVVYALKRQGRTLYGFGG--
T.aes   M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKIFLENVIRDAVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG--
T.aes   M-----SGRDKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKIFLENVIRDAVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG--
Z.may   M-----SGRGKGG---KG--LGK---GARKRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVRKIFLENVIRDAVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG--
L.tem   M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKIFLENVIRDAVTYTEHAXRKTVTAMDVVYALKRQGRTLYGFGG--
L.esc   M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKIFLENVIRDSVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG--
L.esc   M-----SGHGKRG---KG--LGKG--GA-KRH-KKVLRDNIQGITKPAIHRLARRGGVKRISCLIYEE---------------------------------------------------
O.sat   M-----SGRGKGG---RG--LGKG--GA-KRH-RKVFRDNIQGITKPAIRRLARRGGVKRISGLIYEETR-------------------------------------------------
O.sat   M-----SGRGLGG---KG--LGKG--GA-KRH-RKVLHDNIQGITKPAIRRLARRGGVKRISGLIYE----------------------------------------------------
B.cam   M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKIFLENVIRNAVTYTGHARRKTVTAMDVVYALKRQGRTLYGFGG--
B.cam   M-----SGRGKGG---KG--LGRG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKIFLEDVIRDAVTYTGHARRKTVTSMDVVY----------------
C.rei   M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRTVLKTFLENVIRDSVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG--
V.car   M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRTVLKNFLENVIRDSVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG--
P.sal   M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRSVLKVFLENVIRDAVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG--
E.nid   M-----SGRGKGG---KG--LGKG--GA-KRH-RKILRDNIQGITKPAIRRLARRGGVKRISAMIYEETRGVLKSFLESVIRDAVTYTEHAKRKTVTSLDVVYALKRQGRTLYGFGG--
E.nid   M-----SGRGKGG---KG--LGKG--GA-KRH-RKILRDNIQGITKPAIRRLARRGGVKRISAMIYEETRGVLKTFLEGVIRDAVTYTEHAKRKTVTSLDVVYALKRQGRTLYGFGG--
N.cra   M-----TGRGKGG---KG--LGKG--GA-KRH-RKILRDNIQGITKPAIRRLARRGGVKRISAMIYEETRGVLKTFLEGVIRDAVTYTEHAKRKTVTSLDVVYALKRQGRTLYGFGG--
S.pom   M-----SGRGKGG---KG--LGKG--GA-KRH-RKILRDNIQGITKPAIRRLARRGGVKRISALVYEETRAVLKLFLENVIRDAVTYTEHAKRKTVTSLDVVYSLKRQGRTTYGFGG--
S.car   M-----SGRGKGG---KG--LGKG--GA-KRH-RKILRDNIQGITKPAIRRLARRGGVKRISGLIYEEVRAVLKSFLESVIRDSVTYTEHAKRKTVTSLDVVYALKRQGRTLYGFGG--
P.chr   M-----SGRGKGG---KG--LGKG--GA-KRH-RKILRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKIFLENVIRDSVTYTEHAKRKTVTALDVVYALKRSGRTLYGFGA--
P.pol   M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISNTIYEETRGVLKTFLENVIRDAVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG--
P.pol   M-----SGRGKGG---KG--LGKG--GA-KRH-RKVLRDNIQGITKPAIRRLARRGGVKRISKTIYEETRGVLKTFLENVIRDAVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG--
E.his   MATDTGSGRGKGG---KGVTLGKGSKGA--------------GITKPAIRRLARRGGVKRINGAVYDETRNVLKQFLEQVIRDSVTYTEHAKRRTVTAMDVVYALKRQGRTLYG---YS
```

Figure 5.5

User derived tree used for the construction a pairwise exchange table. 19 ancestral node sequences (numbered 1-19) were determined from this analysis. The tree topology was based on a loose interpretation of phylogenies derived by SSU rRNA sequence information.

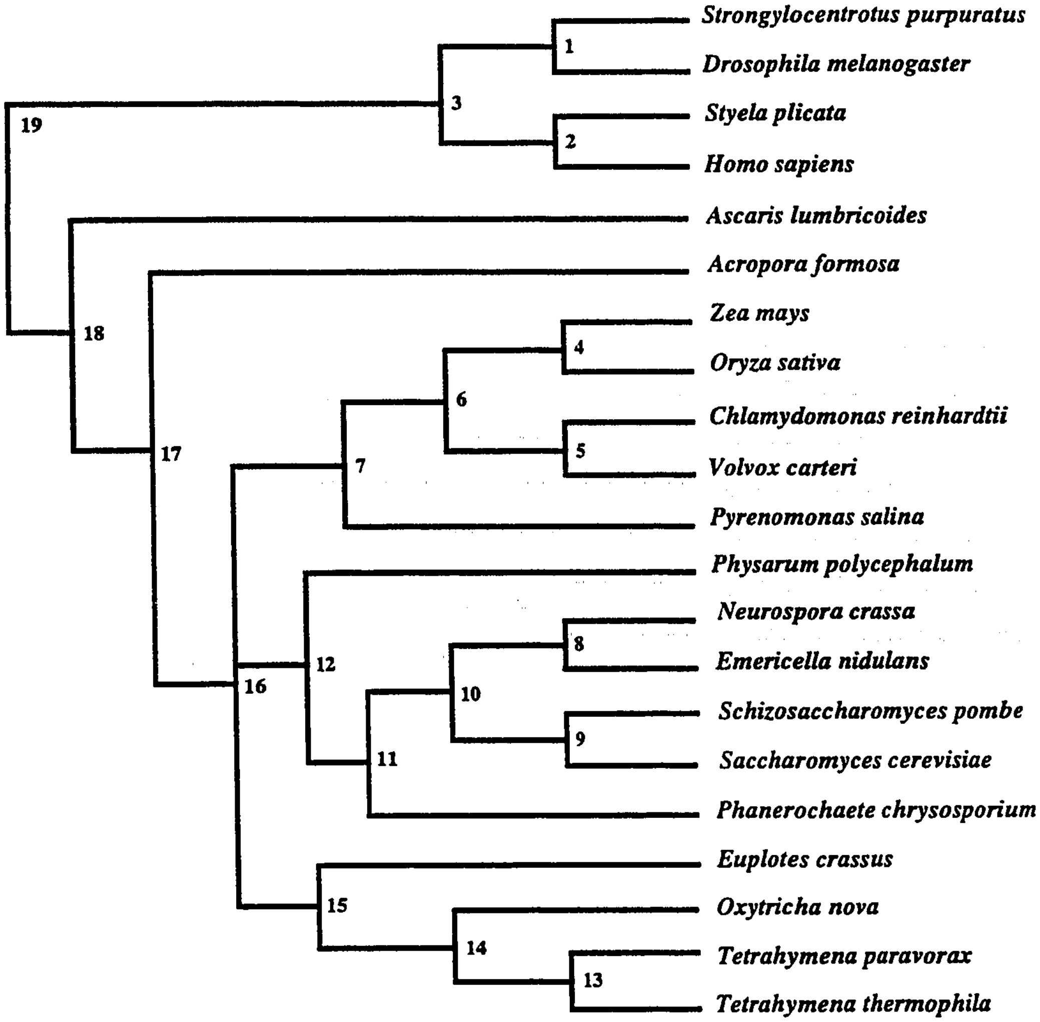

Strongylocentrotus purpuratus
Drosophila melanogaster
Styela plicata
Homo sapiens
Ascaris lumbricoides
Acropora formosa
Zea mays
Oryza sativa
Chlamydomonas reinhardtii
Volvox carteri
Pyrenomonas salina
Physarum polycephalum
Neurospora crassa
Emericella nidulans
Schizosaccharomyces pombe
Saccharomyces cerevisiae
Phanerochaete chrysosporium
Euplotes crassus
Oxytricha nova
Tetrahymena paravorax
Tetrahymena thermophila
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

Table 5.1. Pairwise exchange table (PET91) for the set of 59,190 accepted point mutations found in 16,130 protein sequences. Adapted from Jones et al. (28).

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	-																			
R	247	-																		
N	216	116	-																	
D	386	48	1433	-																
C	106	125	32	13	-															
Q	208	750	159	130	9	-														
E	600	119	180	2914	8	1027	-													
G	1183	614	291	577	98	84	610	-												
H	46	446	466	144	40	635	41	41	-											
I	173	76	130	37	19	20	43	25	26	-										
L	257	205	63	34	36	314	65	56	134	1324	-									
K	200	2348	758	102	7	858	754	142	85	75	94	-								
M	100	61	39	27	23	52	30	27	21	704	974	103	-							
F	51	16	15	8	66	9	13	18	50	196	1093	7	49	-						
P	901	217	31	39	15	395	71	93	157	31	578	77	23	36	-					
S	2413	413	1738	244	353	182	156	1131	138	172	436	228	54	309	1138	-				
T	2440	230	693	151	66	149	142	164	76	930	172	398	343	39	412	2258	-			
W	11	109	2	5	38	12	12	69	5	12	82	9	8	37	6	36	8	-		
Y	41	46	114	89	164	40	15	15	514	61	84	20	17	850	22	164	45	41	-	
V	1766	69	55	127	99	58	226	276	22	3938	1261	58	559	189	84	219	526	27	42	-

Table 5.2. Relative mutability index derived from the PET91 for the set of 20 amino acids. Adapted from Jones et al. (28). *Mutability relative to Alanine, which was arbitrarily assigned a mutability of 100 (28).

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Relative mutability*	100	83	104	86	44	84	77	50	91	103	54	72	93	51	58	117	107	25	50	98

Table 5.3. Mutation probability matrix for the evolutionary distance of 1 PAM derived from the PET91 matrix and associated relative mutability index. Values multiplied by 10,000. Adapted from Jones et al. (28).

Original Amino Acid

Replacement Amino Acid

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	98759	27	24	42	12	23	66	129	5	19	28	22	11	6	99	264	267	1	4	193
R	41	98962	19	8	21	125	20	102	74	13	34	390	10	3	36	69	38	18	8	11
N	43	23	98707	284	6	31	36	58	92	26	12	150	8	3	6	344	137	0	23	11
D	63	8	235	98932	2	21	478	95	24	6	6	17	4	1	6	40	25	1	15	21
C	44	52	13	5	99450	4	3	41	17	8	15	3	10	28	6	147	28	16	68	41
Q	43	154	33	27	2	98955	211	17	130	4	64	176	11	2	81	37	31	2	8	12
E	82	16	25	398	1	140	99042	83	6	6	9	103	4	2	10	21	19	2	2	31
G	135	70	33	66	11	10	70	99369	5	3	6	16	3	2	11	129	19	8	2	32
H	17	164	171	53	15	233	15	15	98867	10	49	31	8	18	58	51	28	2	189	8
I	28	12	21	6	3	3	7	4	4	98722	212	12	113	31	5	28	149	2	10	630
L	24	19	6	3	3	29	6	5	12	122	99328	9	90	101	53	40	16	8	8	117
K	28	334	108	14	1	122	107	20	12	11	13	99101	15	1	11	32	57	1	3	8
M	36	22	14	10	8	19	11	10	8	253	350	37	98845	18	8	19	123	3	6	201
F	11	3	3	2	14	2	3	4	11	41	230	1	10	99357	8	65	8	8	179	40
P	150	36	5	7	3	66	12	16	26	5	97	13	4	6	99278	190	69	1	4	14
S	297	51	214	30	44	22	19	139	17	21	54	28	7	38	140	98548	278	4	20	27
T	351	33	100	22	9	21	20	24	11	134	25	57	49	6	59	325	98670	1	6	76
W	7	65	1	3	23	7	7	41	3	7	49	5	5	22	4	21	5	99684	24	16
Y	11	12	30	23	43	10	4	4	134	16	22	5	4	222	6	43	12	11	99377	11
V	226	9	7	16	13	7	29	35	3	504	161	7	71	24	11	28	67	3	5	98772

Table 5.4. The 125 PAM PET91 matrix derived via matrix multiplication of the 1 PAM. Values are multiplied by 10,000.

Original Amino Acid

Replacement Amino Acid

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2552	345	741	750	337	596	839	1125	357	595	377	393	474	232	1135	1279	1304	92	214	867
R	157	3660	263	137	104	522	162	83	560	197	123	975	306	100	297	307	204	548	58	130
N	350	266	1537	917	87	379	535	378	698	185	108	532	149	116	252	565	442	95	208	161
D	425	156	1065	2472	54	609	1530	441	419	159	79	341	112	51	231	424	333	35	82	164
C	122	98	68	35	7168	37	34	48	93	110	28	37	38	46	120	270	134	20	244	159
Q	248	476	362	498	40	2483	887	157	925	140	197	384	236	57	354	223	204	52	63	138
E	491	194	650	1615	54	1154	2583	379	399	213	135	347	168	57	313	373	306	31	95	213
G	1136	238	793	804	206	401	666	4775	267	236	177	321	222	140	493	1088	600	64	94	447
H	140	451	594	303	91	803	259	83	3565	82	108	212	87	166	236	178	148	111	267	98
I	231	153	170	123	159	133	144	90	97	2494	563	169	608	394	101	164	348	40	157	1180
L	324	207	261	137	73	416	190	163	353	1364	5481	275	2194	988	288	222	361	359	341	1099
K	366	1895	1035	612	86	788	582	295	483	353	207	4403	929	95	383	614	682	154	137	240
M	79	97	58	34	18	102	43	30	43	254	388	185	2203	97	42	72	116	22	31	222
F	113	94	120	46	62	58	48	97	184	445	460	50	314	5386	47	141	128	285	1835	149
P	665	375	301	239	149	478	315	293	377	183	194	249	181	118	4168	610	422	54	64	239
S	1024	553	957	614	588	425	524	841	366	320	189	538	322	219	856	1993	1218	296	209	363
T	888	313	634	412	205	316	351	397	235	554	243	500	402	162	503	1027	2473	77	180	554
W	12	136	13	6	7	11	6	8	13	9	6	19	9	91	11	54	14	7418	93	5
Y	91	40	160	59	244	55	76	39	260	143	139	33	74	1379	36	101	99	196	5278	99
V	618	215	243	200	270	244	242	290	236	2048	809	207	937	276	308	332	609	51	200	3496

Table 5.5. Pairwise exchange table for the aligned histone H4 sequences shown in figure 5.4. Values multiplied by 10,000.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2754																			
R	2	5700																		
N	2	0	616																	
D	17	0	3	1164																
C	0	2	0	0	20															
Q	2	0	2	0	0	722														
E	5	0	0	73	0	49	1596													
G	74	6	17	2	2	14	1	6568												
H	0	8	0	0	0	2	0	0	858											
I	4	0	6	0	0	0	0	6	0	2560										
L	0	2	22	0	0	2	1	0	0	2	3086									
K	0	170	1	0	0	0	0	0	0	0	4	4502								
M	0	0	0	0	0	0	2	0	0	0	119	0	226							
F	0	0	0	0	0	0	0	0	0	0	21	0	1	824						
P	3	0	0	0	0	0	0	2	0	0	0	0	0	0	430					
S	171	0	76	0	0	3	0	53	0	39	1	0	0	0	2	908				
T	5	0	37	0	21	1	0	15	0	6	0	0	0	0	0	35	2822			
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Y	0	0	0	0	0	0	0	0	0	0	33	0	5	0	0	0	0	0	1652	
V	20	0	7	0	0	0	0	59	0	116	0	0	0	0	1	57	2	0	0	3198

Table 5.6. Index of relative mutability derived from the pairwise exchange table for histone H4 (Table 5.5). *Mutability relative to Alanine, which was arbitrarily assigned a mutability of 100.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Relative mutability*	100	32	220	76	537	94	76	37	12	65	63	37	361	26	18	326	41	0	22	76

Table 5.7. 1 PAM for the aligned histone H4 sequences shown in figure 5.4. Values are shown multiplied by 10,000.

Original Amino Acid

Replacement Amino Acid

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	9849	1	4	21	0	4	4	17	0	2	0	0	0	0	10	193	3	0	0	9
R	1	9951	0	0	113	0	0	1	14	0	1	55	0	0	0	0	0	0	0	0
N	1	0	9667	4	0	4	0	4	0	3	10	0	0	0	0	86	19	0	0	3
D	8	0	6	9885	0	0	64	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	1	0	0	8594	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0
Q	1	0	4	0	0	9857	43	3	4	0	1	0	0	0	0	3	1	0	0	0
E	3	0	0	88	0	93	9885	0	0	0	0	0	9	0	0	0	0	0	0	0
G	37	2	33	2	113	27	1	9944	0	3	0	0	0	0	7	60	8	0	0	26
H	0	2	0	0	0	4	0	0	9983	0	0	0	0	0	0	0	0	0	0	0
I	2	0	12	0	0	0	0	1	0	9901	1	0	0	0	0	44	3	0	0	51
L	0	1	42	0	0	4	1	0	0	1	9905	1	512	38	0	1	0	0	30	0
K	0	44	2	0	0	0	0	0	0	0	2	9943	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	2	0	0	0	55	0	9454	2	0	0	0	0	5	0
F	0	0	0	0	0	0	0	0	0	0	10	0	4	9961	0	0	0	0	0	0
P	2	0	0	0	0	0	0	0	0	0	0	0	0	0	9972	2	0	0	0	0
S	85	0	146	0	0	6	0	12	0	22	0	0	0	0	7	9506	18	0	0	25
T	3	0	71	0	1182	2	0	3	0	3	0	0	0	0	0	40	9937	0	0	1
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10000	0	0
Y	0	0	0	0	0	0	0	0	0	0	15	0	22	0	0	0	0	0	9966	0
V	10	0	14	0	0	0	0	13	0	64	0	0	0	0	4	64	1	0	0	9885

Table 5.8. 32 PAM matrix (PAM critical) for the histone H4 sequences shown in figure 5.4. 32 PAM was used as a PAM critical. This is the first time where a commonly represented amino acid (serine) has an off-diagonal term roughly equal to an on-diagonal term. Values are shown multiplied by 10,000.

Original Amino Acid

Replacement Amino Acid	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	6560	14	654	466	139	144	150	464	1	190	11	1	5	0	302	2532	176	0	0	341
R	26	8651	9	1	722	4	1	39	406	1	28	1504	14	2	1	9	19	0	1	2
N	169	1	3640	66	280	78	14	100	0	108	182	6	104	12	18	837	386	0	10	110
D	192	0	104	7122	4	209	1459	19	1	3	4	0	15	0	4	53	6	0	0	5
C	1	3	10	0	134	1	0	3	0	1	0	0	0	0	0	5	67	0	0	1
Q	38	1	79	133	16	6448	938	78	88	4	24	0	21	2	2	63	17	0	1	7
E	85	0	30	2002	3	2031	7240	20	12	1	27	0	119	2	1	27	3	0	2	3
G	1033	45	868	103	889	661	80	8435	5	204	16	4	6	1	231	1232	297	0	0	718
H	0	59	1	1	4	97	6	1	9456	0	0	5	0	0	0	0	0	0	0	0
I	170	0	377	6	80	12	2	83	0	7420	29	0	14	2	20	752	116	0	1	1213
L	12	16	756	10	18	98	49	8	1	35	8114	36	6807	1038	1	97	29	0	866	9
K	2	1195	38	0	78	1	0	3	27	0	51	8435	26	3	0	4	2	0	2	0
M	1	1	46	4	1	9	24	0	0	2	730	2	2109	82	0	5	1	0	111	0
F	0	0	13	0	0	2	1	0	0	1	267	1	196	8827	0	1	0	0	13	0
P	44	0	9	1	2	1	0	15	0	3	0	0	0	0	9151	46	2	0	0	14
S	1113	2	1426	57	264	107	21	243	1	369	40	1	18	2	141	2509	349	0	1	412
T	169	10	1438	13	7289	66	5	128	1	125	26	1	10	1	13	764	8439	0	1	75
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10000	0	0
Y	0	0	22	0	0	3	2	0	0	1	445	1	531	28	0	2	0	0	8987	0
V	385	1	477	14	76	29	5	364	0	1532	10	0	4	0	113	1061	87	0	0	7092

Table 5.9. Viable sequences determined for the Jones "best" cassette. Amino acid positions are noted at the top of the table and the wild-type (w.t.) amino acid sequence is given first. The number of differences with respect to the wild-type sequence is noted in the last column.

pos.	38	41	42	43	46	# dif.
w.t.	A	G	G	V	I	0
	G	G	G	V	I	1
	V	G	G	V	I	1
	A	A	G	V	I	1
	A	T	D	F	I	3

Appendix 1

A FIELD INVERSION GEL ELECTROPHORESIS REGIME FOR THE ENHANCED SEPARATION OF DNA FRAGMENTS BETWEEN 10 TO 50 KILOBASE PAIRS

ABSTRACT

Asymmetric field inversion gel electrophoresis was used to enhance the separation of DNA molecules in the 10 to 50 kb range. Comparisons to constant field gel electrophoresis are offered. A very short pulse of high electric field (22.5 V/cm) in the forward direction was followed by a long pulse of low electric field (1.25 V/cm) in the reverse direction. The integrated forward electric field was twice that of the integrated reverse electric field. With this regime, the separation of 10.1 kb and 9.1 kb DNA fragments was increased two fold and the separation of 24.0 kb and 10.1 kb fragments four fold relative to their separation under constant electric field gel electrophoresis. This FIGE method is an easy technique for the increased separation and subsequent isolation of similarly sized small endonuclease restriction fragments for mapping and cloning purposes.

INTRODUCTION

Standard agarose gel electrophoresis is commonly used to separate DNA fragments between 500 bp and 20 kb. Field inversion gel electrophoresis (FIGE) has been used to separate DNA fragments between 20 and 2,000 kb (2,3,7,8) but is most commonly applied to fragments larger than 100 kb. FIGE is useful for genome separation and mapping (12), the identification of bacteria and viruses (6,14), and yeast artificial chromosome projects (1) among others.

In some cases, it is important to promote the separation of similarly sized bands for increased resolution of restriction digest patterns for identification or extraction. Using standard gel electrophoresis, fragments over 15 kb fall into a range of limited mobility independent of the agarose percentage. A FIGE regime capable of promoting the separation of small fragments between 10 and 50 kb is presented. This method uses a very short pulse of high forward electric field followed by a longer pulse of low reverse electric field and can enhance the separation of bands 1 kb different in length (10.1 kb and 9.1 kb) 2 fold relative to separation using constant field gel electrophoresis.

MATERIALS AND METHODS

DNA Samples for Electrophoresis

A ladder of lambda phage DNA digested with Hind III was used as a length standard control for all runs (Life Technologies, Gaithersburg, MD, USA). For sample lanes, a set of bands with equal intensity were constructed including full length linearized lambda phage DNA (48.5 kb), lambda phage DNA cut with Xba I (24.5 kb and 24.0 kb), plasmid pDH4 DNA linearized with Eco RI (10.1 kb), and plasmid pUK499 DNA (9) linearized with *Eco* RI (9.1 kb). Plasmid pDH4 is a derivative of pUK499 that contains a 1.0 kb DNA insert. For ease of analysis, the two Xba I bands were treated as 24.0 kb. All enzyme digestions were performed with manufacturers procedures (Life Technologies, Gaithersburg, MD, USA).

Electrophoresis Equipment

All gels were made using UltraPURE agarose (Life Technologies, Gaithersburg, MD, USA) and 0.5x TBE buffer (45 mM Tris, 45 mM Boric acid, and 1.25 mM Na_2EDTA adjusted to pH 8.3). Gels were run in 0.5x TBE buffer for eight hours and then stained with ethidium bromide (0.5 µg/mL). A horizontal gel apparatus constructed from lucite with a distance 12.0 cm between electrodes and overall dimensions of 13.4 cm (length), 6.3 cm (width) was used for all experiments. During electrophoresis, the temperature was maintained at 23°C by recircularization of the buffer. For constant field gel electrophoresis, DNA fragments were separated in 0.8% and 0.6% 0.5x TBE gels at 1.7 V/cm were run in the same electrophoresis apparatus without field inversion. All relative band separations were measured after eight hours.

Asymmetric Field Inversion Gel Electrophoresis

An EC 452 power supply (E-C Apparatus Corp., St. Petersburg, Florida, USA) was used for all experiments. Pulsing of the forward and reverse electric fields was provided by a homemade electronic switcher between the power supply and the electrophoresis chamber. Pulse timing was generated by a 555 timer chip, which produced a square wave with a pulse time that could be switched between 0.4 and 0.7 ms. This square wave (as measured by an oscilloscope) drove a series of 7493 binary counters that multiplied the original pulse length by factors of two, producing twelve different pulse lengths varying from 0.4 ms to 1.434 seconds. The pulse generator was followed by a 7490 decade and a 7408 "AND" gate, which produced a 1:9 ratio of the forward pulse duration (T_+) to the reverse pulse duration (T_-). These time pulses were used to drive a photovoltaic relay (PVD 3354) which applied the forward power supply and reverse power supply to the gel.

This FIGE regime was predicated on a very short forward electric field pulse so that the larger DNA molecules did not become completely streamlined. The electronic switcher was capable of switching a maximum of 300 volts; thus, V_+ was limited to 270 volts to provide a reasonable margin of safety. To determine that band inversion was not affecting the results, all fragments were run in separate lanes as well as all in one lane relative to the lambda Hind III ladder. Band inversion was not detected (data not shown).

Evaluation of Relative Migration

Following electrophoresis, gels were photographed using GEL analysis software (version 2.0, Ultra Violet Products,) and migration distances were measured with a fluorescent ruler. The relative separation (Δ) of two DNA bands is defined as the separation between the bands divided by the mean migration distance of the bands:

$$\Delta = 2(m_1 - m_2)/(m_1 + m_2)$$

where m_1 and m_2 are the migration distances of the smallest band and the largest band, respectively.

RESULTS AND DISCUSSION

The net movement of DNA molecules in FIGE is determined by the electric field in the forward (E_+) and reverse direction (E_-), the time E_+ is applied (T_+) and E_- is applied (T_-), temperature, and gel concentration (8). The net migration is proportional to the ratio (R) of the integrated forward electric field to the integrated reverse electric field ($R = E_+ T_+ / E_- T_-$). R must be greater than one for net forward movement to occur. We used an R of two to provide a reasonable net migration of the DNA molecules based on previous studies (3,7). The ratio of V_+ to V_- was adjusted to be 1:18 so that the integrated forward electric field was twice the integrated reverse electric field. These FIGE parameters were similar to those offered previously (5,11,13). However our approach uses a much stronger E_+ pulse for a shorter time (on the order of 4 ms to 1 second). Previous reports tend to focus on methods of increasing the separation of large DNA molecules on the order of hundreds or thousands of kilobases. Using the parameters $E_+ = 22.5$ V/cm, $E_- = 1.25$ V/cm, $T_+ = 1/10$ cycle, $T_- = 9/10$ cycle, the relative separation (Δ) of the different DNA molecules was determined for various T_+ (Figure A1.1). Under these conditions, optimal separation of the 10.1 kb and 9.1 kb bands can be achieved with a T_+ of 32 ms giving a Δ of 0.40. In comparison with constant field gel electrophoresis on standard 0.8% or 0.6% agarose gels at the identical average forward electric field of 1.7 V/cm, the maximal separation of 10.1 kb from 9.1 kb fragments is only 0.19 and 24 kb fragment separation from 10.1 kb fragment is only 0.22.

The relative mobility as a function of T_+ was calculated for all fragments (figure A1.2). When T_+ is sufficient for the molecules to become completely streamlined with E_+, the net mobility of the DNA molecule is maximized. Therefore, increasing T_+ does not increase the net mobility. For example, with the 4.4 kb band, any forward pulse time greater than 64 ms has a maximum net mobility; whereas, at 64 ms the mobility of 48.5 kb DNA fragments is still increasing. The maximum net mobility of small DNA molecules is generally higher than the maximum net mobility of large DNA molecules (except in the case of very short T_+). A comparison of the relative separation between 48.5 kb fragments and all others in the set is shown in figure A1.3.

This FIGE regime presents a high E_+ for a duration too rapid for long DNA molecules to assume a highly streamlined conformation. Following the forward pulse, the DNA molecules are reoriented into a conformation that has low mobility in the forward direction. The low E_- does not produce significant streamlining in the reverse direction and minimized the backward migration of the DNA molecules. With T_- much longer than T_+, the DNA molecules assumed a steady state conformation during the reverse pulse.

The time required for a DNA molecule of a specific size to become streamlined in this situation can be estimated from figure 2. When T_+ becomes too short to allow the DNA molecule to fully streamline in the forward direction, the net mobility of the molecule is decreased. Apparently, the 48.5 kb DNA molecules take longer than 1.4 seconds to become substantially streamlined, while 4.4 kb strands are fully streamlined in 64 ms or less under these conditions. Maximum separation between any two bands occurs at a T_+ for which there is greatest difference in mobilities. Comparing figures A1.2 and A1.3, it is apparent that the optimal separation between two DNA molecules occurs when T_+ is just short enough to begin to lower the mobility of the smaller DNA molecule. Thus for maximum separation, neither of the DNA molecules achieves a fully streamlined conformation in the forward direction. Surprisingly, this applies even when the DNA molecules are an order of magnitude different in length (4.4 kb versus 48.5 kb).

One major potential drawback of FIGE is the phenomenon of band inversion. In this situation, high molecular weight DNA molecules move faster than low molecular weight DNA molecules. Under standard FIGE conditions, band inversion is associated with DNA fragments on the order of megabases. By running the fragments of the DNA ladder in separate lanes as well as together in a single lane, potential band inversion could be observed. No band inversion was detected for the fragment sizes and conditions shown here. The speeding up of large molecules as T_+ becomes much smaller than their optimum as is the case with larger DNA molecules in standard FIGE conditions was not observed.

This FIGE regime significantly enhances the separation of DNA fragments under 50 kb relative to constant field gel electrophoresis. Similar regimes offered previously focus on the separation of DNA molecules 200 to 2200 kb in length (10) or use with polyacrylamide gels (7). Carle and Olson (3) reported a two to three percent increase in the separation of DNA fragments smaller than 15 kb using FIGE. Others have attempted to increase the resolution of lambda phage DNA restriction fragments using FIGE; however separation of fragments smaller than 15 kb was not substantially increased relative to constant field conditions (4). DNA molecules in the 5 kb to 50 kb size range are particularly useful for restriction fragment analysis, extraction, and subsequent cloning. Relative to the FIGE regime described here, constant field electrophoresis generally performs poorly in this range. Even for DNA molecules of very similar size (10.1 kb and 9.1 kb), separation can be enhanced several fold over conventional gel electrophoresis when using this FIGE regime.

REFERENCES

1. Burke, D. T. and M. V. Olson. 1991. Preparation of Clone Libraries in Yeast Artificial-Chromosome Vectors. *in*: Guthrie, C. and G. R. Fink (Eds.), Guide to Yeast Genetics and Molecular Biology. Academic Press, San Diego. pp. 251-270.

2. Carle, C. F., M. Frank and M. V. Olson. 1986. Electrophoretic separations of large DNA molecules by periodic inversion of the electric field. Science. 232:65-68.

3. Carle, C. F. and M. V. Olson. 1986. Electrophoretic separations of large DNA molecules. *in*: Lerman, L. S. (Ed.), DNA Probes: Applications in Genetic and Infectious Disease and Cancer. Cold Spring Harbor Laboratory Press. New York. pp. 93-95.

4. Daniels, D. L., C. H. Olson, R. Brumley, and F. R. Blattner. 1990. Field inversion gel electrophoresis applied to the rapid multi-enzyme restriction mapping of phage lambda clones. Nuc. Acids Res. 18:1312.

5. Denko, N., G. Amato, B. Peters, T. D. Stamato. 1989. An Asymmetric Field Inversion Gel Electrophoresis Method for the Separation of Large DNA Molecules. Anal. Biochem. 178:172-176.

6. Green, M., K. Barbadora, S. Donabedian, and M. J. Zervos. 1995. Comparison of field inversion gel electrophoresis with contour-clamped homogeneous electric field electrophoresis as a typing method for *Enterococcus faecium*. J. Clin. Microbiol. 33:1554-1557.

7. Heller, C. and S. Beck. 1992. Field inversion gel electrophoresis in denaturing polyacrylamide gels. Nuc. Acids Res. 20:2447-2452.

8. Heller, C. and F. M. Pohl. 1989. A systematic study of field inversion gel electrophoresis. Nuc. Acids Res. 17:5989-6003.

9. Kim, U.-J., M. Han., P. Kayne, and M. Grunstein. 1988. Effects of histone H4 depletion on the cell cycle and transcription of *Saccharomyces cerevisiae*. EMBO. J. 7:2211-2219.

10. Lalande, M., J. Noolandi, C. Turmel, J. Rousseau, and G. W. Slater. 1987. Pulsed-field electrophoresis: application of a computer model to the separation of large DNA molecules. Proc. Natl. Acad. Sci. USA. 84:8011-8015.

11. Noolandi, J. and C. Turmel. 1992. Preparation, Manipulation, and Pulse Strategy for One-Dimensional Pulsed-Field Gel Electrophoresis (ODPFGE). *in*: Methods in Molecular Biology, Vol. 12: Pulsed-Field Gel Electrophoresis. Humana Press. Totowa, New Jersey. pp. 73-103.

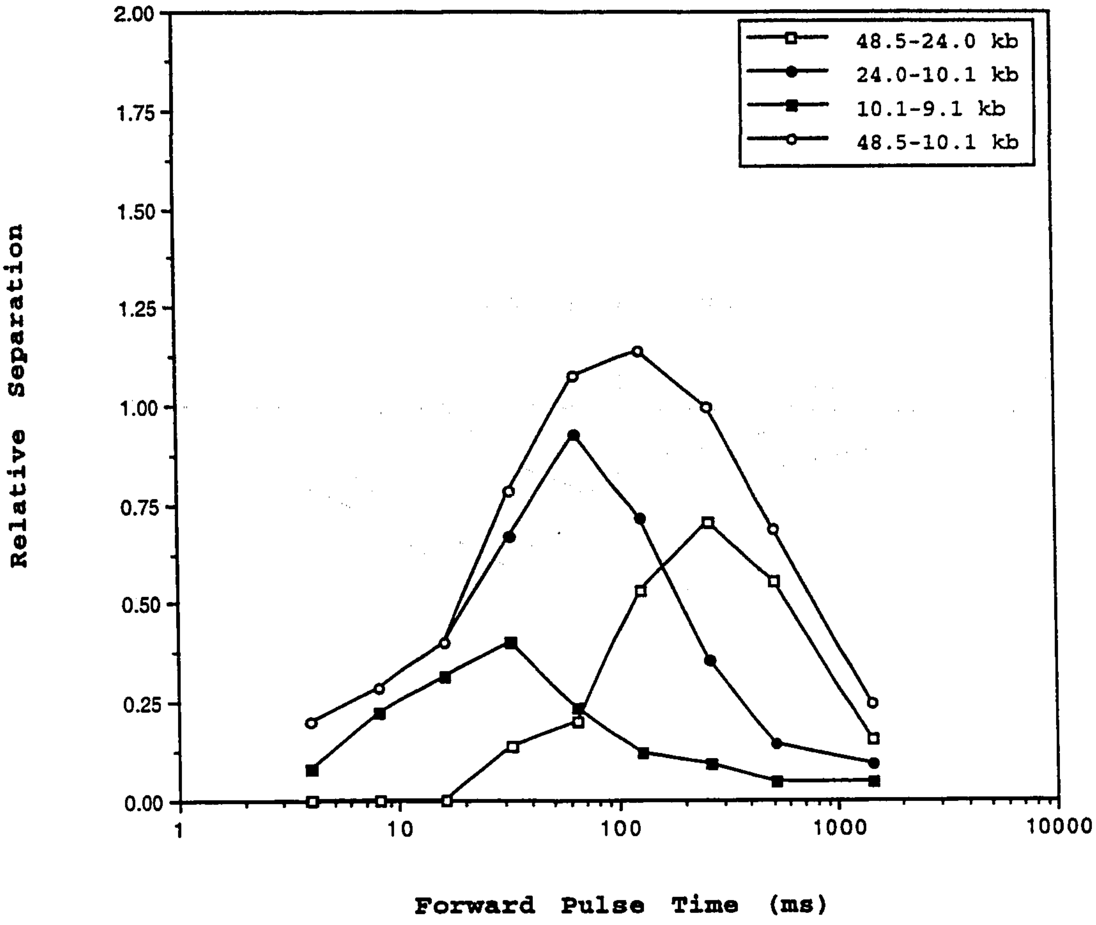
12. Sherman, F. and P. Wakem. 1991. Mapping Yeast Genes. in: Guthrie, C. and G. R. Fink (Eds.). Guide to Yeast Genetics and Molecular Biology. Academic Press, San Diego. pp. 38-57.

13. Turmel, C. and J. Noolandi. 1993. Effect of one-dimensional pulsed-field gel electrophoresis on linear and circular DNA. Electrophoresis. 14:304-312.

14. Zhang, X., S. Efstathiou, and A. Simmons. 1994. dentification of novel herpes simplex virus replicative intermediates by field inversion gel electrophoresis: implications for viral DNA amplification strategies. Virology. 202:530-539.

Figure A1.1

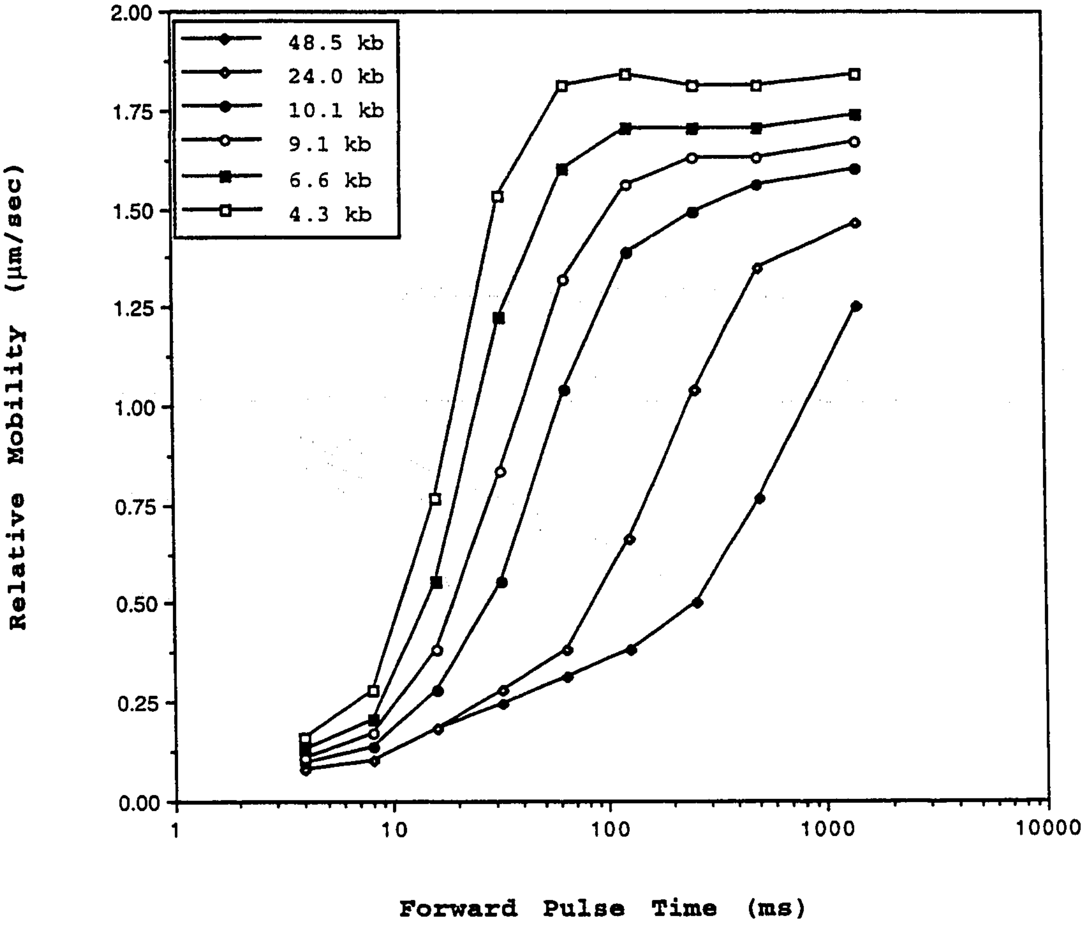
Relative separation of DNA fragments as a function of the forward pulse time using FIGE. Relative separation (Δ) varies from zero, where the two bands have identical migration distances to two, where the larger band does not migrate at all. FIGE conditions included a forward electric field of 22.5 V/cm , reverse electric field of 1.5 V/cm, forward pulse time 1/10 of the total cycle, and reverse pulse time for 9/10 of the total cycle. Forward pulses ranged from four milliseconds to slightly over one second.



2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Relative Separation
1
10
100
1000
10000
Forward Pulse Time (ms)
48.5-24.0 kb
24.0-10.1 kb
10.1-9.1 kb
48.5-10.1 kb

Figure A1.2

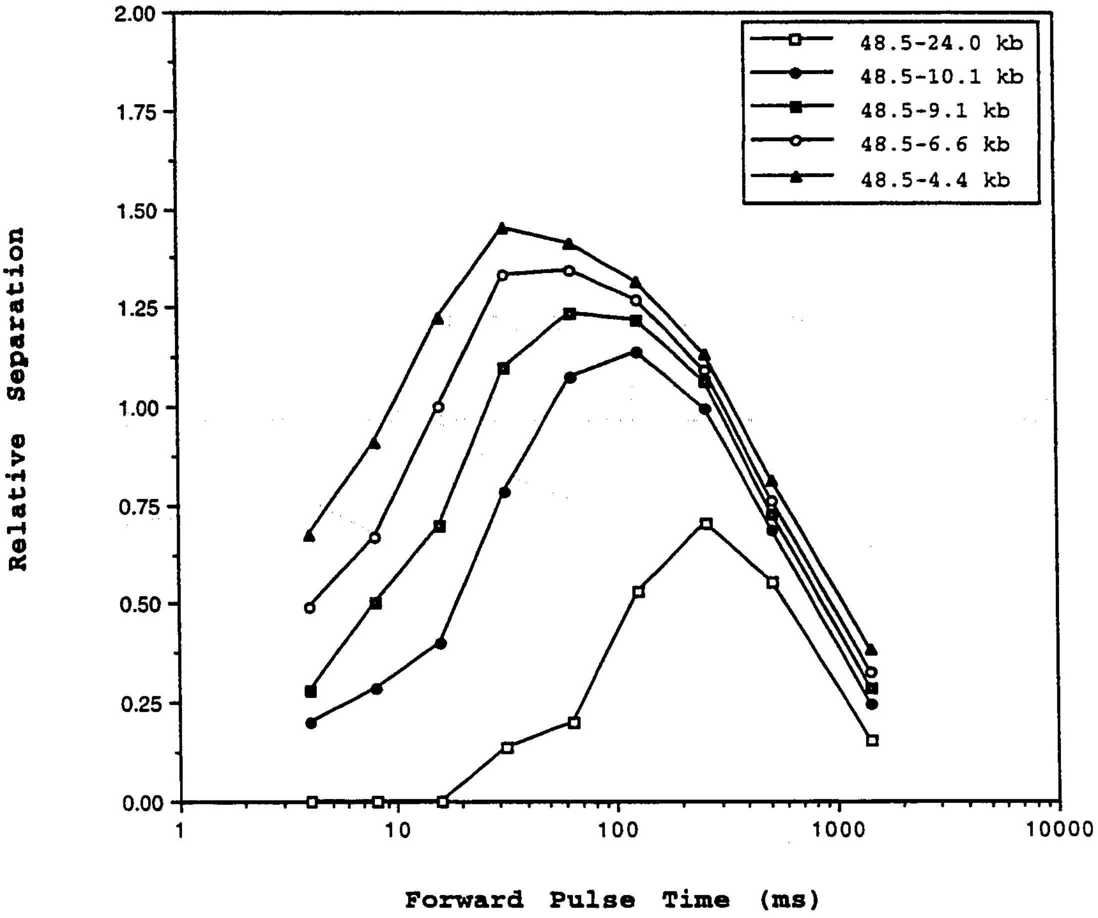
Relative mobility of DNA molecules as a function of the forward pulse time using FIGE. Smaller molecules make the transition to a streamlined configuration in a much more rapid time than do larger molecules. The maximum separation between any two bands occurs when there is the greatest difference in relative mobility.



2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
1
10
100
1000
10000
Relative Mobility (μm/sec)
Forward Pulse Time (ms)
48.5 kb
24.0 kb
10.1 kb
9.1 kb
6.6 kb
4.3 kb

Figure A1.3

Relative separation of 48.5 kb DNA fragments compared to various smaller DNA fragments as a function of the forward pulse time. Appropriate forward pulse times for maximal separation decrease as a function of the size of the smaller fragment being separated.



2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
1
10
100
1000
10000
Relative Separation
Forward Pulse Time (ms)
□ 48.5-24.0 kb
● 48.5-10.1 kb
■ 48.5-9.1 kb
○ 48.5-6.6 kb
▲ 48.5-4.4 kb

Appendix 2

PROKARYOTIC GENOME SIZE AND SSU rRNA COPY NUMBER: ESTIMATION OF MICROBIAL RELATIVE ABUNDANCE FROM A MIXED POPULATION

INTRODUCTION

In recent years, PCR amplification of prokaryotic small subunit ribosomal RNA (SSU rRNA) genes has generated a more complete understanding of prokaryotic ecology and evolution. PCR has been used to effectively determine the presence or absence of an organism from a sample. However, the relative abundance of organisms in a mixed population is not necessarily reflected in the amount of PCR products. The use of an appropriate internal standard in competitive PCR will allow the determination of the amount of SSU rRNA sequences for a specific taxa or group present in a complex sample. The amount of the SSU rRNA sequences can then be used to determine the relative abundance of the specific taxa or group in the original complex. Such determinations require knowledge of genome sizes for the organisms in the sample and SSU rRNA copy number for the specific taxa or group of interest. Estimates for the genome size of 303 prokaryotic taxa and estimates for SSU rRNA copy number of 108 taxa collected from the literature are reviewed. This information, in combination with quantitative PCR, can be used to calculate the relative abundance of specific prokaryotic taxa or groups in a given sample. Such measurements are most valuable in applied and environmental microbiology.

QUANTIFICATION OF BACTERIAL POPULATIONS USING SSU rRNA

Competitive PCR

Molecular techniques have recently provided a means for the identification of organisms independent of the ability to culture them in the laboratory (8,9,168,169,229). PCR amplification (156) of the small subunit rRNA (SSU rRNA) genes, has dramatically improved the ability to characterize prokaryotes present in a sample. An appropriate choice of PCR primers allows for a wide variety of SSU rRNA sequences to be specifically amplified from a complex sample (33). If universal primers are used, almost all SSU rRNA sequences from a mixed sample can be PCR amplified. However, if a set of specific PCR primers are chosen, only the SSU rRNA sequences of a small group or a single species of prokaryote will be PCR amplified. Although these methods are powerful and can determine the presence or absence of an organism, they do not always provide a measure of relative abundance of the organisms. There is no guarantee that all SSU rRNA sequences will be PCR amplified with equal efficiency.

Competitive PCR (154,176,182), using specific primers and the amplification of an internal standard, can be used to determine the amount of targets present in a sample. If the internal standard has an identical sequence to that of the target except for the modification of a unique restriction site (usually only 2-3 base pair differences) then during PCR amplification, the internal standard will be amplified in exactly the same manner as the target (15,83,151). When a known amount of internal standard is mixed with a DNA sample, the ratio of target sequence to internal standard sequence can be determined by analyzing the PCR product after digestion with the restriction endonuclease unique to the internal standard following amplification. The ratio of internal standard to target in the PCR product is identical to the ratio of internal standard to target in the original sample. Knowing the amount of internal standard added to the sample, the amount of target in the original sample can easily be calculated. If the internal standard sequence and the target sequence differ by the modification of a unique restriction site in the target to another unique restriction site in the internal standard, then the PCR product may be analyzed with each of the unique restriction enzymes. This produces complementary information and significantly increases the accuracy of determining the ratio of internal standard to target in the PCR product.

Relative Abundance of Taxa

It is of major interest to determine the relative abundance of a specific organism in an original sample. In order to calculate relative abundance of a specific organism in an original sample, one must know the amount of target SSU rRNA sequences in the original sample, the copy number of SSU rRNA genes in the target organism, and the weighted average of the genome sizes for all organisms in the original sample.

The copy number of SSU rRNA genes and the genome size for many prokaryotes are available in the literature. With recent advancements in DNA sequencing technology, this information is accumulating rapidly as is the number of completely sequenced prokaryotic genomes. A variety of methods including field inversion gel electrophoresis, pulsed field gel electrophoresis, hybridization, and subsequent C_0t curve analysis have been used to estimate genome size. By surveying the literature, genome size estimates for 303 prokaryotes were gathered (Table A2.1). There are fewer estimates for SSU rRNA gene copy number to be found in the literature (108 included in this survey). These estimates of SSU rRNA gene copy number are the result of restriction digestion, gel electrophoresis, and Southern hybridization using an SSU rRNA probe. Estimates of both genome size and SSU rRNA gene copy number are available in the literature for 89 organisms (Table A2.1). In order to determine the relative abundance of a particular prokaryote in a population distribution, this information is a crucial resource.

Summary

This compilation of genome size and SSU rRNA gene copy number provides a guide for making a reasonable estimate of genome size and SSU rRNA gene copy number for a wide variety of organisms. Even if the SSU rRNA gene copy number for a specific taxa of interest is not given in the literature, a reasonable estimate can often be made from the copy numbers of closely related taxa. Similarly, although the exact composition of taxa in a sample is seldom known, a reasonable estimate for the average (weighted arithmetic mean) of the genome sizes can be determined from the genome sizes of taxa likely to occur in the sample. We have displayed taxa for which genome sizes and SSU rRNA gene copy number are available as a taxonomic tree to facilitate making these types of estimations. The data presented here allows one to make an estimation of the relative abundance of a taxa in a sample based on the abundance of SSU rRNA present in the sample.

GENOME SIZE AND SSU rRNA COPY NUMBER ESTIMATES

Taxa Names

A list of organisms identified by genus and species was developed from the literature. These organisms were arranged taxonomically according to the RDP SSU rRNA Prokaryotic Taxonomic List (145) (Table A2.1, Figure A2.1). Organisms not explicitly found in the RDP database were grouped with the most appropriate taxa (32). These taxa are identified by a "?" symbol in the last digit of their RDP taxonomic number. Genus and species names were cross-checked to insure appropriate placement (100,201). Synonymous names were placed in parenthesis in Table A2.1.

Genome and SSU rRNA Estimates

For species with multiple estimates of genome size in the literature, an average and standard deviation was recorded (Table A2.1). In the event that the complete genome sequence for an organism was known, the exact number of base pairs was listed instead of any previous estimates. SSU rRNA copy number estimates were collected from the literature and organized in a spreadsheet (Microsoft Excel 5.0). In cases where estimates of SSU rRNA copy number varied for a particular genus (an infrequent situation), the most commonly cited and/or most recent estimate was used for Table A2.1. If SSU rRNA copy number varied from strain to strain for a particular species, the range was included in Table A2.1.

Phylogenetic Analysis

The majority of SSU rRNA sequences were downloaded from the RDP database in a pre-aligned format. Sequences downloaded from GenBank Release 104.0 (17) were aligned with the RDP database. This set of aligned SSU rRNA sequences was used to create a user-derived tree for the set of 89 organisms having estimates for both genome size and SSU rRNA copy number using PAUP 3.1.1 (214). The user-derived tree was developed using the RDP grouping numbers as a guide (Figure A2.1). Species identifiers at the tips of each branch are listed in Table A2.1. The genome size (Mb) and SSU rRNA copy number for each organism is provided as a part of the species identifier for ease of comparison on the tree.

Genome Size Analysis

Genome size estimates for 303 prokaryotes following the RDP taxonomy are given in Table A2.1, including mean and standard deviation where a range of values appeared in the literature. The variation in genome size estimates for a number of taxa including several *Pseudomonas* species, *Desulfovibrio vulgaris*, *Mycobacterium avium*, *Clostridium acetobutylicum*, *Lactococcus lactis*, *Streptococcus agalactiae*, and *Bacillus cereus* was quite large. For *Mycoplasma* species and related taxa of RDP grouping 2.16.4, the standard deviation of reported genome size in the literature was very low. A histogram of genome size over all taxonomic groupings is displayed in Figure A2.2. The distribution roughly spans one order of magnitude with a mean and standard deviation of 3,604 kb ± 1,997 kb. However, 78% of the genomes fall between 0.6 kb and 4.8 kb. 25% of the genomes fall under 2.0 kb and 75% are under 4.8 kb. The genome of *Mycoplasma genitalium* G-37 is completely sequenced and is the smallest genome reported with only 580,070 bp (76). Cyanobacterial taxa form a tail of large genome organisms, with the genome of *Scytonema* sp. estimated to be the largest at 11,593 kb (94).

Relationship of Genome Size and SSU rRNA Copy Number

The relation between genome size and SSU rRNA copy number for 107 organisms is shown in Figure A2.3. The genome size and SSU rRNA copy number might be expected to be positively correlated. However, no correlation is apparent ($y = 2267.5 * x$; $r^2 = 0.128$). The organisms with 10 to 12 SSU rRNA genes (the largest number observed) have genome sizes less than 5 Mb, whereas organisms with the largest observed genomes (10 to 11.6 MB) have only 6 SSU rRNA genes.

SSU rRNA Copy Number Analysis

A histogram of SSU rRNA copy number for 103 taxa is given in Figure A2.4. The mean SSU rRNA copy number is 3.8 with a standard deviation of 2.9. SSU rRNA copy number ranges from 1 (over 10% of the taxa) to 12 in some strains of *Bacillus cereus* (105). The relationships between the 89 taxa with a value for both genome size and SSU rRNA copy number, are displayed in a user-derived tree, constructed following RDP taxonomic classification (Figure A2.1). In most cases, genome sizes and copy numbers were very similar within the major taxonomic groupings. For instance, all Archaea have similar genome sizes (2449 ± 357 kb) and are restricted to two or fewer SSU rRNA genes.

Summary

Genome sizes for a total of 303 prokaryotes are presented in this review. It is clear that the methods used to determine each estimate have varying degrees of accuracy. In particular, it appears that older citations generally have a higher variance in genome size estimation, perhaps as a result of methods used (data not shown). For ease of calculation, all citations were treated with equal merit in making estimates of genome size averages (Table A2.1).

Over 40% of the organisms surveyed have 1 or 2 SSU rRNA genes. Curiously, no organisms with a copy number of 5 were found in this survey. This anomaly could be the result of a small sample size. However, the abundance of organisms with 6 and 7 SSU rRNA genes suggests that organisms with 5 SSU rRNA genes might be precluded by some type of selection pressure on SSU rRNA gene copy number. It is particularly interesting that some *Azomonas* and *Bacillus* species have strain dependent differences in copy number.

The user-derived tree is based on RDP taxonomy and general relationships of groups. This type of tree may offer insights into the evolution of both genome size and SSU rRNA copy number. The tree is also useful for choosing neighboring values for the estimation of copy number or genome size where the values for a specific organism are not available in the literature. Using the tree, a few potential anomalies can be seen in relation to rRNA copy number estimation for the 89 taxa available in the literature. For instance, *Aquifex pyrophilus* is reported to contains 6 copies of SSU rRNA (199), however, all neighboring taxa to *A. pyrophilus* have only 2 SSU rRNA copies.

METHODS OF ESTIMATING BACTERIAL RELATIVE ABUNDANCE

An Estimation Method

It is of major interest to determine the relative abundance of a target organism in an environmental or complex laboratory sample. The amount of target SSU rRNA sequences in a sample can be determined by quantitative PCR with specific primers and an internal standard. In addition, the number of SSU rRNA genes in the target organism and weighted average of the genome sizes for the other organisms in the sample must be known in order to determine the relative abundance of a target organism.

Using quantitative PCR with specific primers and an internal standard, it is possible to measure the amount of SSU rRNA for a specific target in a given sample (15, 83). In cases where SSU rRNA copy number and/or genome size are unknown, estimates can be made using averages from neighboring organisms or groups based on the RDP taxonomy shown in Figure A2.1. The closest relatives are likely to have similar genome sizes and SSU rRNA copy numbers to the taxa of interest.

As an example, to determine the amount of *Bacillus subtilis* present in an environmental sample, a known amount of internal standard DNA is added to a known amount of total genomic DNA and the mixture is PCR amplified with a set of primers specific for *B. subtilis* SSU rRNA. From the amount of internal standard sequence in the

PCR product, the amount of *B. subtilis* rDNA present in the original sample can be calculated. Assume that measurement indicates 20 pg of *B. subtilis* rDNA is present in a total of 100 ng of genomic DNA. From the literature we know that *B. subtilis* has 10 SSU rRNA genes. From our knowledge of the organisms present in the sample, assume that we estimate the weighted average of the genome sizes in the sample to be 3.6×10^6 bp. From this weighted average it can be calculated that 100 ng of genomic DNA represents 2.5×10^7 organisms. The *B. subtilis* specific SSU rRNA PCR product is about 1200 bp in length, thus 20 pg of this product corresponds to 1.5×10^7 SSU rRNA genes. Since each *B. subtilis* has 10 SSU rRNA genes, 20 pg of this product corresponds to 1.5×10^6 *B. subtilis.* In this case, the sample of 100 ng of genomic DNA represents 2.5×10^7 organisms of which 1.5×10^6 are *B. subtilis.* Thus, in this sample, *B. subtilis* make up 6% of the organisms.

Summary

Estimation of the *in situ* relative abundance of a specific prokaryotic taxa or group from a sample is a crucial step in microbial ecology studies. This is especially true in terms of monitoring temporal variance of bacterial population distributions from environmental or medical samples. Competitive PCR amplification provides a measure of the amount of taxa or group specific SSU rRNA genes present in the sample. Knowledge of genome sizes and SSU rRNA copy numbers is required to determine the relative abundance of the specific organisms present in the sample. With future advances in genome sequencing, estimates of genome sizes and SSU rRNA copy numbers will become more numerous and more precise. In turn, these offer even more data from which to make measurements of prokaryotic relative abundance in situations where it was previously only possible to identify either the presence or absence of prokaryotic taxa.

REFERENCES

1. Afseth, G., Y. Y. Mo, and L. P. Mallavia. 1995. Characterization of the 23S and 5S rRNA genes of *Coxiella burnetii* and identification of an intervening sequence within the 23S rRNA gene. J. Bacteriol. 177:2946-2949.

2. Allardet-Servent, A., G. Bourg, M. Ramuz, M. Pages, M. Bellis, and G. Roizes. 1988. DNA polymorphism in strains of the genus *Brucella*. J. Bacteriol. 170:4603-4607.

3. Amikam, D., G. Glaser, and S. Razin. 1984. Mycoplasmas (*Mollicutes*) have a low number of rRNA genes. J. Bacteriol. 158:376-378.

4. Amikam, D., S. Razin, and G. Glaser. 1982. Ribosomal RNA genes in *Mycoplasma*. Nucleic Acids Res. 10:4215-4222.

5. Andresson, O. S. and O. H. Fridjonsson. 1994. The sequence of the single 16S rRNA gene of the thermophilic eubacterium *Rhodothermus marinus* reveals a distant relationship to the group containing *Flexibacter*, *Bacteroides*, and *Cytophaga* species. J. Bacteriol. 176:6165-6169.

6. Bancroft, I., C. P. Wolk, and E. V. Oren. 1989. Physical and genetic maps of the genome of the heterocyst-forming cyanobacterium *Anabaena* sp. strain PCC-7120. J. Bacteriol. 171:5940-5948.

7. Baril, C., J. L. Herrmann, C. Richaud, D. Margarita, and I. Saint Girons. 1992. Scattering of the rRNA genes on the physical map of the circular chromosome of *Leptospira interrogans* serovar *icterohaemorrhagiae*. J. Bacteriol. 174:7566-7571.

8. Barns, S. M., C. F. Delwiche, J. D. Palmer, and N. R. Pace. 1996. Perspectives on archael diversity, thermophily and monophyly from environmental rRNA sequences. Proc. Natl. Acad. Sci. USA 93:9188-9193.

9. Barns, S. M., R. E. Fundyga, M. W. Jeffries, and N. R. Pace. 1994. Remarkable archael diversity detected in a Yellowstone National Park hot spring environment. Proc. Natl. Acad. Sci. USA 91:1609-1613.

10. Bathe, B., J. Kalinowski, and A. Puehler. 1996. A physical and genetic map of the *Corynebacterium glutamicum* ATCC 13032 chromosome. Mol. Gen. Genet. 252:255-265.

11. Bautsch, W. 1988. Rapid physical mapping of the *Mycoplasma mobile* genome by two-dimensional field inversion gel electrophoresis techniques. Nucleic Acids Res. 16:11461-11467.

12. Bautsch, W. 1993. A *Nhe*I macrorestriction map of the *Neisseria meningitidis* B1940 genome. FEMS Microbiol. Lett. 107:191-198.

13. Bautsch, W., D. Grothues, and B. Tummler. 1988. Genome fingerprinting of *Pseudomonas aeruginosa* by two-dimensional field inversion gel electrophoresis. FEMS Microbiol. Lett. 52:255-258.

14. Baylis, H. A. and M. J. Bibb. 1988. Organisation of the ribosomal RNA genes in *Streptomyces coelicolor* A3(2). Mol. Gen. Genet. 211:191-196.

15. Becker-André, M. and K. Hahlbrock. 1989. Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a *P*CR *A*ided *T*ranscript *T*itration *A*ssay (PATTY). Nucleic Acids Res. 17:9437-9448.

16. Bensaadimerchermek, N., J. C. Salvado, C. Cagnon, and S. Karama. 1995. Characterization of the unlinked 16S rDNA and 23S-5S rRNA operon of *Wolbachia pipientis*, a prokaryotic parasite of insect gonads. Gene 165:81-86.

17. Benson, D. A., M. Boguski, D. J. Lipman, and J. Ostell. 1994. GenBank. Nucleic Acids Res. 22:3441-3444.

18. Bentley, R. W. and J. A. Leigh. 1995. Determination of 16S ribosomal RNA gene copy number in *Streptococcus uberis, S. agalactiae, S. dysgalactiae, and S. parauberis*. FEMS Immun. Med. Microbiol. 12:1-7.

19. Bercovier, H., O. Kafri, and S. Sela. 1986. *Mycobacteria* possess a surprisingly small number of ribosomal RNA genes in relation to the size of their genome. Biochem. and Biophys. Res. Comm. 136:1136-1141.

20. Berger, F., G. Fischer, A. Kyriacou, and B. Decaris. 1996. Mapping of the ribosomal operons on the linear chromosomal DNA of *Streptomyces ambofaciens* DSM40697. FEMS Microbiol. Lett. 143:167-173.

21. Birkelund, S. and R. S. Stephens. 1992. Construction of physical and genetic maps of *Chlamydia trachomatis* serovar L2 by pulsed-field gel electrophoresis. J. Bacteriol. 174:2742-2747.

22. Blattner, F. R., G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao. 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277:1453-1474.

24. Bobovnikova, Y., W-L. Ng, S. Dassarma, and N. Hackett. 1994. Restriction mapping the genome of *Halobacterium halobium* strain NRC-1. Syst. and Appl. Microbiol. 16:597-604.

25. Boehringer Mannheim GmbH Biochemica, 1993. Guide to Genome Analysis and Mapping, Boehringer Mannheim GmbH Biochemica, Mannheim 31, Germany.

26. Borges, K. M. and P. L. Bergquist. 1993. Pulsed-field gel electrophoresis study of the genome of *Caldocellum saccharolyticum*. Cur. Microbiol. 27:15-19.

27. Borges, K. M. and P. L. Bergquist. 1993. Genomic restriction map of the extremely thermophilic bacterium *Thermus thermophilus* HB8. J. Bacteriol. 175:103-110.

28. Bourgeois, P. L., M. Lautier, M. Mata, and P. Ritzenthaler. 1992. Physical and genetic map of the chromosome of *Lactococcus lactis* subsp. *lactis* IL1403. J. Bacteriol. 174(21):6752-6762.

29. Bourgeois, P. L., M. Mata, and P. Ritzenthaler. 1989. Genome comparison of *Lactococcus* strains by pulsed-field gel electrophoresis. FEMS Microbiol. Lett. 59:65-70.

30. Bourget, N., J.-M. Simonet, and B. Decaris. 1993. Analysis of the genome of the five *Bifidobacterium breve* strains plasmid content, pulsed-field gel electrophoresis, genome size estimation, and *rrn* loci number. FEMS Microbiol. Lett. 110:11-20.

31. Bourke, B., P. Sherman, H. Louie, E. Hani, P. Islur, and V. Chan. 1995. Physical and genetic map of the genome of *Campylobacter upsaliensis*. Microbiol. 141:2417-2424.

32. Brock, T. D., M. T. Madigan, J. M. Martinko, and J. Parker. 1994. Biology of Microorganisms, Prentice Hall, New Jersey.

33. Brunk, C. F., E. Avaniss-Aghajani, and C. A. Brunk. 1996. A computer analysis of primer and probe hybridization potential with bacterial small-subunit rRNA sequences. Appl. and Env. Microbiol. 62:872-879.

34. Bult, C. J., O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J.-F. Tomb, M. D. Adams, C. I. Reich, R. Overbeek, E. F. Kirkness, K. G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N. S. M. Geoghagen, J. F. Weidman, J. L. Fuhrmann, D. Nguyen, T. R. Utterback, J. M. Kelley, J. D. Peterson, P. W. Sadow, M. C. Hanna, M. D. Cotton, K. M. Roberts, M. A. Hurst, B. P. Kaine, M. Borodovsky, H.-P. Klenk, C. M. Fraser, H. O. Smith, C. R. Woese, and J. C. Venter. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii* Science. 273:1058-1073.

35. Canard, B. and S. T. Cole. 1989. Genome organization of the anaerobic pathogen *Clostridium perfringens*. Proc. Natl. Acad. Sci. USA 86:6676-6680.

36. Canard, B., B. Saint-Joanis, and S. T. Cole. 1992. Genomic diversity and organisation of virulence genes in the pathogenic anaerobe *Clostridium perfringens*. Mol. Microbiol. 6:1421-1429.

37. Carlson, C. R., A. Gronstad, and A.-B. Kolstø. 1992. Physical maps of the genomes of three *Bacillus cereus* strains. J. Bacteriol. 174:3750-3756.

38. Carlson, C. R., T. Johansen, and A.-B. Kolstø. 1996. The chromosome map of *Bacillus thuringiensis* subsp. *canadensis* HD224 is highly similar to that of the *Bacillus cereus* type strain ATCC 14579. FEMS Microbiol. Lett. 141:163-167.

39. Carlson, C. R., T. Johansen, M.-M. Lecadet, and A.-B. Kolstø. 1996. Genomic organization of the entomopathogenic bacterium *Bacillus thuringiensis* subsp. *berliner* 1715. Microbiol. 142:1625-1634.

40. Carlson, C. R. and A.-B. Kolstø. 1993. A complete physical map of a *Bacillus thuringiensis* chromosome. J. Bacteriol. 175:1053-1060.

41. Carlson, C. R. and A.-B. Kolstø. 1994. A small (2.4 Mb) *Bacillus cereus* chromosome corresponds to a conserved region of a larger (5.3 Mb) *Bacillus cereus* chromosome. Mol. Microbiol. 13:161-169.

42. Chang, N. and D. E. Taylor. 1990. Use of pulsed-field agarose gel electrophoresis to size genomes of *Campylobacter* species and to construct a *SalI* map of *Campylobacter jejuni* UA580. J. Bacteriol. 172:5211-5217.

43. Charlebois, R. L., J. D. Hofman, L. C. Schalkwyk, W. L. Lam, and W. F. Doolittle. 1989. Genome mapping in halobacteria. Can. J. Microbiol. 35:21-29.

44. Charlebois, R. L., L. C. Schalkwyk, J. D. Hofman, and W. F. Doolittle. 1991. Detailed physical map and set of overlapping clones covering the genome of the archaebacterium *Haloferax volcanii* DS2. J. Mol. Biol. 222:509-524.

45. Charles, H., G. Condemine, C. Nardon, and P. Nardon. 1997. Genome size characterization of the principal endocellular symbiotic bacteria of the weevil *Sitophilus oryzae*, using pulsed field gel electrophoresis. Ins. Biochem. Mol. Biol. 27:345-350.

46. Chen, H., I. M. Keseler, and L. J. Shimkets. 1990. Genome size of *Myxococcus xanthus* determined by pulsed-field gel electrophoresis. J. Bacteriol. 172:4206-4213.

47. Chen, H., A. Kuspa, I. M. Keseler, and L. J. Shimkets. 1991. Physical map of the *Myxococcus xanthus* chromosome. J. Bacteriol. 173:2109-2115.

48. Chen, X. and W. R. Widger. 1993. Physical genome map fo the unicellular cyanobacterium *Synechococcus* sp. strain PCC 7002. J. Bacteriol. 175:5106-5116.

49. Cheng, H. P., and T. G. Lessie. 1994. Multiple replicons constituting the genome of *Psudomonas cepacia* 17616. J. Bacteriol. 176:4034-4042.

50. Chiaruttini, C. and M. Milet. 1993. Gene organization, primary structure and RNA processing anaylsis of a ribosomal RNA operon in *Lactococcus lactis*. J. Mol. Biol. 230:57-76.

51. Choudhury, S. R., R. K. Bhadra, and J. Das. 1994. Genome size and restriction fragment length polymorphism analysis of *Vibrio cholerae* strains belonging to different serovars and biotypes. FEMS Microbiol. Lett. 115:329-334.

52. Choy, K.-T., J.-R. Wu, F.-S. Wen, and Y.-H. Tseng. 1995. Genome size and physical map for *Xanthomonas campestris* pv. *campestris*, p. 503. *In* Abstracts of the 95th General Meeting of the American Society for Microbiology 1995. American Society for Microbiology, Washington, D.C.

53. Claus, H., H. Rötlich, and Z. Filip. 1992. DNA fingerprints of *Pseudomonas* spp. using rotating field electrophoresis. Microb. Rel. 1:11-16.

54. Cocks, B. G., L. E. Pyle, and L. R. Finch. 1989. A physical map of the genome of the *Ureaplasma urealyticum* 960T with ribosomal RNA loci. Nucleic Acids Res. 17:6713-6719.

55. Cohen, A., W. L. Lam, R. L. Charlebois, W. F. Doolittle, and L. C. Schalkwyk. 1992. Localizing genes on the map of the genome of *Haloferax volcanii* one of the Archaea. Proc. Natl. Acad. Sci. USA 89:1602-1606.

56. Daniel, P. 1995. Sizing of the *Lactobacillus plantarum* genome and other lactic acid bacterial species by transverse alternating field electrophoresis. Curr. Microbiol. 30:243-246.

57. Daniel, P., E. De Waele, and J.-N. Hallet. 1993. Optimisation of transverse alternating field electrophoresis for strain identification of *Leuconostoc oenos*. Appl. Microbiol. Biotech. 38:638-641.

58. Davidson, B. E., N. Kordias, M. Dobos, and A. J. Hillier. 1996. Genomic organization of lactic acid bacteria. Antonie Van Leeuwenhoek 70:161-183.

59. Dempsey, J. A., A. B. Wallace, and J. G. Cannon. 1995. The physical map of the chromosome of a serogroup A strain of *Neisseria meningitidis* shows complex rearrangements relative to the chromosomes of the two mapped strains of the closely related species *N. gonorrhoeae*. J. Bacteriol. 177:6390-6400.

60. Dempsey, J. A., J. York, and J. G. Cannon. 1993. Characterization of the genome of *Acetobacter xylinum* by pulsed field gel electrophoresis, abstr. H-64, p. 201. *In* Abstracts of the 93rd General Meeting of the American Society for Microbiology 1993. American Society for Microbiology, Washington, D.C.

61. Dempsey, J. A. F., W. Litaker, A. Madhure, T. L. Snodgrass, and J. G. Cannon. 1991. Physical map of the chromosome of *Neisseria gonorrhoeae* FA1090 with locations of genetic markers, including *opa* and *pil* genes. J. Bacteriol. 173:5476-5486.

62. Devereux, R., S. G. Willis, and M. E. Hines. 1997. Genome sizes of *Desulfovibrio desulfuricans*, *Desulfovibrio vulgaris*, and *Desulfobulbus propionicus* estimated by pulsed-field gel electrophoresis of linearized chromosomal DNA. Curr. Microbiol. 34:337-339.

63. Dutt, D. and S. Krawiec. 1994. Genome sizes of two *Thiobacillus* spp. p.234. *In* Abstracts of the General Meeting of the American Society for Microbiology 1994. American Society for Microbiology, Washington, D.C.

64. Ely, B., T. W. Ely, C. J. Gerardot, and A. Dingwall. 1990. Circularity of the *Caulobacter crescentus* chromosome determined by pulsed-field gel electrophoresis. J. Bacteriol. 172:1262-1266.

65. Ely, B. and C. J. Gerardot. 1988. Use of pulsed-field-gradient gel electrophoresis to construct a physical map of the *Caulobacter crescentus* genome. Gene 68:323-333.

66. Eremeeva, M. E., V. Roux, and D. Raoult. 1993. Determination of genome size and restriction pattern polymorphism of *Rickettsia prowazekii* and *Rickettsia typhi* by pulsed field pel electrophoresis. FEMS Microbiol. Lett. 112:105-112.

67. Farinha, M. A. 1996. A physical map of the genome of *Branhamella catarralis*, abstr. H-46, p. 491. *In* Abstracts of the 96th General Meeting of the American Society for Microbiology 1996. American Society for Microbiology, Washington, D.C.

68. Fitz-Gibbon, S. Personal Communication.

69. Fleischmann, R., M. Adams, O. White, R. Clayton, E. Kirkness, A. Kerlavage, C. Bult, J.-F. Tomb, B. Dougherty, J. Merrick, K. McKenney, G. Sutton, W. Fitzhugh, C. Fields, J. D. Gocayne, J. Scott, R. Shirley, L.-I. Liu, A. Glodek, J. Kelley, J. Weidman, C. Phillips, T. Spriggs, E. Hedblom, M. Cotton, T. Utterback, M. Hanna, D. Nguyen, D. Saudek, R. Brandon, L. Fine, J. Fritchman, J. Fuhrmann, N. S. Geoghagen, C. Gnehm, L. McDonald, K. Small, C. Fraser, H. Smith, and J. Venter. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496-512.

70. Fonstein, M., E. Koshy, T. Nikolskaya, P. Mourachov, and R. Haselkorn. 1995. Refinement of the high-resolution physical and genetic map of *Rhodobacter capsulatus* and genome surveys using blots of the cosmid encyclopedia. EMBO J. 14:1827-1841.

71. Fonstein, M., S. Zheng, and R. Haselkorn. 1992. Physical map of the genome of *Rhodobacter capsulatus* SB 1003. J. Bacteriol. 174:4070-4077.

72. Forsyth, M. H., J. G. Tully, T. S. Gorton, L. Hinckley, S. Frasca, H. J. Van Kruiningen, and S. J. Geary. 1996. *Mycoplasma sturni* sp. nov., from the conjunctiva of a European starling (*Sturnus vulgaris*). Int. J. Syst. Bacteriol. 46:716-719.

73. Franklin, R. A. and L. Daneo-Moore. 1994. Molecular analysis of the *Steptococcus gordonii* chromosome, abstr. H189, p. 233. *In* Abstracts of the 94th General Meeting of the American Society for Microbiology, 1994. American Society for Microbiology, Washington, D.C.

74. Franklin, R. A. and L. Daneo-Moore. 1995. Molecular analysis of the *Streptococcus gordonii* chromosome, abstr. H-65, p. 503. *In* Abstracts of the 95th General Meeting of the American Society for Microbiology, 1995. American Society for Microbiology, Washington, D.C.

75. Fraser, C. M., S. Casjens, W. M. Huang, G. G. Sutton, R. Clayton, R. Lathigra, O. White, K. A. Ketchum, R. Dodson, E. K. Hickey, M. Gwinn, B. Dougherty, J.-F. Tomb, R. D. Fleischmann, D. Richardson, J. Peterson, A. R. Kerlavage, J. Quackenbush, S. Salzberg, M. Hanson, R. van Vugt, N. Palmer, M. D. Adams, J. Gocayne, J. Weidman, T. Utterback, L. Watthey, L. McDonald, P. Artiach, C. Bowman, S. Garland, C. Fujii, M. D. Cotton, K. Horst, K. Roberts, B. Hatch, H. Smith, and J. C. Venter. 1997. Genomic sequence of a lyme disease spirochaete, *Borrelia burgdorferi*. Nature. 390:580.

76. Fraser, C. M., J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A.R. Kerlavage, G. Sutton, J. M. Kelley, J. L. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, J. M. Merrick, J.-F. Tomb, B. A. Dougherty, K. F. Bott, P.-C. Hu, T. S. Lucier, S. N. Peterson, H. O. Smith, C. A. Hutchinson III, J. C. Venter. 1995. The minimal gene complement of *Mycoplasma genitalium*. Science. 270:397-403.

77. Frutos, R., M. Pages, M. Bellis, G. Roizes, and M. Bergoin. 1989. Pulsed-field gel electrophoresis determination of the genome size of obligate intracellular bacteria belonging to the genera *Chlamydia*, *Rickettsiella*, and *Porochlamydia*. J. Bacteriol. 171:4511-4513.

78. Gaher, M., K. Einsiedler, T. Crass, and W. Bautsch. 1996. A physical and genetic map of *Neisseria meningitidis* B1940. Mol. Microbiol. 19:249-259.

79. Garnier, T., B. Canard, and S. T. Cole. 1991. Cloning, mapping, and molecular characterization of the rRNA operons of *Clostridium perfringens*. J. Bacteriol. 173:5431-5438.

80. Gasc, A.-M., L. Kauc, P. Barraille, M. Sicard, and S. Goodgal. 1991. Gene localization, size, and physical map of the chromosome of *Streptococcus pneumoniae*. J. Bacteriol. 173:7361-7367.

81. Gazumyan, A., J. J. Schwartz, D. Liveris, and I. Schwartz. 1994. Sequence analysis of the ribosomal RNA operon of the Lyme disease spirochete, *Borrelia burgdorferi*. Gene 146:57-65.

82. Genome Therapeutics Corp. Personal Communication.

83. Gilliland, G., S. Perrin, K. Blanchard, and H. F. Bunn. 1990. Analysis of cytokine mRNA and DNA: detection and quantitation by competitive polymerase chain reaction. Proc. Natl. Acad. Sci. USA 87:2725-2729.

84. Ginard, M., J. Lalucat, B. Tummler, and U. Romling. 1997. Genome organization of *Pseudomonas stutzeri* and resulting taxonomic and evolutionary considerations. Int. J. Syst. Bacteriol. 47:132-143.

85. Glass, J., E. Lefkowitz, G. Cassell, and E. Chen. 1997. Genome sequencing of *Ureaplasma urealyticum*, p. 283. *In* Abstracts of the 97th General Meeting of the American Society for Microbiology 1997. American Society for Microbiology, Washington, D.C.

86. Gorton, T. S., M. S. Goh, and S. J. Geary. 1995. Physical mapping of the *Mycoplasma gallisepticum* S6 genome with localization of selected genes. J. Bacteriol. 177:259-263.

87. Grothues, D., and B. Tümmler. 1987. Genome analysis of *Pseudomonas aeruginosa* by field inversion gel electrophoresis. FEMS Microbiol. Lett. 46:419-422.

88. Hacioglu, E., H. Basim, and R. Stall. 1996. Rarely cutting restriction endonucleases useful for determining genome size and physical map of the chromosome of *Xanthomonas axonopodis* pv. *vesicatoria*. Phytopathology 86:S77-S78.

89. Hantman, M. J., S. Sun, P. J. Piggot, and L. Daneo-Moore. 1993. Chromosome organization of *Streptococcus mutans* GS-5. J. Gen. Microbiol. 139:67-77.

90. Hartmann, R. K. and V. A. Erdmann. 1989. *Thermus thermophilus* 16S rRNA is transcribed from an isolated transcription unit. J. Bacteriol. 171:2933-2941.

91. Hartmann, R. K., H. Y. Toschka, N. Ulbrich, and V. A. Erdmann. 1986. Genomic organization of rDNA in *Pseudomonas aeruginosa*. FEBS Lett. 195:187-193.

92. Hartmann, R. K., N. Ulbrich, and V. A. Erdmann. 1987. An unusual rRNA operon constellation: in *Thermus thermophilus* HB8 the 23S/5S rRNA operon is a separate entity from the 16S rRNA operon. Biochemie 69:1097-1104.

93. He, Q. and L. J. Shimkets. 1993. Construction of a physical and genetic map of the *Myxococcus xanthus* chromosome with a YAC library, abstr. H-63, p. 201. *In* Abstracts of the 93rd General Meeting of the American Society for Microbiology. American Society for Microbiology, Washington, D.C.

94. Herdman, M. 1985. The evolution of bacterial genomes, p.37-68. *In* T. Cavalier-Smith (ed.), The evolution of genome size, John Wiley & Sons Ltd., Chichester.

95. Herdman, M., M. Janvier, R. Rippka, and R. Y. Stanier. 1979. Genome size of cyanobacteria. J. Gen. Microbiol. 111:73-85.

96. Herrero, A. and C. P. Wolk. 1986. Genetic mapping of the chromosome of the cyanobacterium *Anabaena variabilis*. J. Biol. Chem. 261:7748-7754.

97. Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B. C. Li, and R. Herrmann. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res. 24:4420-4449.

98. Holloway, B. W., M. D. Escuadra, A. F. Morgan, and R. Saffery. 1992. The new approaches to whole genome analysis of bacteria. FEMS Microbiol. Lett. 100:101-105.

99. Holloway, B. W. and A. F. Morgan. 1986. Genome organization in *Pseudomonas*. Ann.Rev.Microbiol. 40:79-105.

100. Holt, J. G. 1992. Stedman's Bergey's bacteria words. Williams & Wilkins, Baltimore.

101. Honeycutt, R. J., M. McClelland, and B. W. S. Sobral. 1993. Physical map of the genome of *Rhizobium meliloti* 1021. J. Bacteriol. 175:6945-6952.

102. Huber, I. and S. Selenska-Pobell. 1994. Pulsed-field gel electrophoresis-fingerprinting, genome size estimation and *rrn* loci number of *Rhizobium galegae*. J. Appl. Bacteriol. 77:528-533.

103. Hubner, A., R. Edelstein, and W. Hendrickson. 1996. Genome organization of *Burkholderia cepacia* AC1100, p. 493. *In* Abstracts of the 96th General Meeting of the American Society for Microbiology 1996. American Society for Microbiology, Washington, D.C.

104. Jean, A. S. and R. L. Charlebois. 1996. Comparative genomic analysis of the *Haloferax volcanii* DS2 and *Halobacterium salinarium* GRB contig maps reveals extensive rearrangement. J. Bacteriol. 178:3860-3868.

105. Johansen, T., C. R. Carlson, and A.-B. Kolstø. 1996. Variable numbers of rRNA gene operons in *Bacillus cereus* strains. FEMS Microbiol. Lett. 136:325-328.

106. Kaneko, T., T. Matsubayashi, M. Sugita, and M. Sugiura. 1996. Physical and gene maps of the unicellular cyanobacterium *Synechococcus sp.* strain PCC6301 genome. Plant Mol. Biol. 31:193-201.

107. Kaneko, T., S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirosawa, M. Sugiura, S. Sasamoto, T. Kimura, T. Hosouchi, A. Matsuno, A. Muraki, N. Nakazaki, K. Naruo, S. Okumura, S. Shimpo, C. Takeuchi, T. Wada, A. Watanabe, M. Yamada, M. Yasuda, and S. Tabata. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. 30:109-136.

108. Kauc, L. and S. H. Goodgal. 1989. The size and a physical map of the chromosome of *Haemophilus parainfluenzae*. Gene. 83:377-380.

109. Kieser, H., T. Kieser, and D. Hopwood. 1992. A combined genetic and physical map of the *Streptomyces coelicolor* A3(2) chromosome. J. Bacteriol. 174:5496-5507.

110. Kim, E., H. Kim, S.-P. Hong, K. H. Kang, Y. H. Kho, and Y.-H. Park. 1993. Gene organization and primary structure of a ribosomal RNA gene cluster from *Streptomyces griseus* subsp. *griseus*. Gene 132:21-31.

111. Kim, J.-R., B.-S. Kang, J.-H. Ko, J.-S. Park, S.-J. Kim, G.-H. Bai, T.-H. Chung, K.-S. Nam, Y.-K. Choi, I.-S. Choe, T.-W. Chung, Y.-C. Lee, and C.-H. Kim. 1996. Genomic heterogeneity in clinical strains of *Mycobacterium tuberculosis*, *M. terrae* complex, *M. gordonae*, *M. avium-intracellulare* complex and *M. fortuitum* by pulsed-field gel electrophoresis. J. of Biochem. Mol. Biol. 29:569-573.

112. Kim, N. W., H. Bingham, R. Khawaja, H. Louie, E. Hani, K. Neote, and V. L. Chan. 1992. Physical map of *Campylobacter jejuni* TGH9011 and localization of 10 genetic markers by use of pulsed-field gel electrophoresis. J. Bacteriol. 174:3493-3498.

113. Klein, A. and M. Schnorr. 1984. Genome complexity of methanogenic bacteria. J. Bacteriol. 158:628-631.

114. Klenk, H. P., R. A. Clayton, J.-F. Tomb, O. White, K. E. Nelson, K. A. Ketchum, R. J. Dodson, M. Gwinn, E. K. Hickey, J. D. Peterson, D. L. Richardson, A. R. Kerlavage, D. E. Graham, N. C. Kyrpides, R. D. Fleischmann, J. Quackenbush, N. H. Lee, G. G. Sutton, S. Gill, E. F. Kirkness, B. A. Dougherty, K. McKenney, M. D. Adams, B. Loftus, S. Peterson, C. I. Reich, L. K. McNeil, J. H. Badger, A. Glodek, L. Zhou, R. Overbeek, J. D. Gocayne, J. F. Weideman, L. McDonald, T. Utterback, M. D. Cotton, T. Spriggs, P. Artiach, B. P. Kaine, S. M. Sykes, P. W. Sadow, K. P. D'Andrea, C. Bowman, C. Fujii, S. A. Garland, T. M. Mason, G. J. Olsen, C. M. Fraser, H. O. Smith, C. R. Woese, and J. C. Venter. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature. 390:364-370.

115. Kolstø, A.-B., A. Gronstad, and H. Oppegaard. 1990. Physical map of the *Bacillus cereus* chromosome. J. Bacteriol. 172:3821-3825.

116. Konai, M., R. F. Whitcomb, J. G. Tully, D. L. Rose, P. Carle, J. M. Bove, R. B. Henegar, K. J. Hackett, T. B. Clark, and D. L. Williamson. 1995. *Spiroplasma velocicrescens* sp. nov., from the vespid wasp *Monobia quadridens*. Int. J. Syst. Bacteriol. 45:203-206.

117. Krueger, C. M., K. L. Marks, and G. M. Ihler. 1994. *Bartonella bacilliformis* genome size estimate and preliminary macrorestriction map, p. 233. *In* Abstracts of the 94th General Meeting of the American Society for Microbiology 1994. American Society for Microbiology, Washington, D.C.

118. Krueger, C. M., K. L. Marks, and G. M. Ihler. 1995. Physical map of the *Bartonella bacilliformis* genome. J. Bacteriol. 177:7271-7274.

119. Kuendig, C., C. Beck, H. Hennecke, and M. Goettfert. 1995. A single rRNA gene region in *Bradyrhizobium japonicum*. J. Bacteriol. 177:5151-5154.

120. Kundig, C., H. Hennecke, and M. Gottfert. 1993. Correlated physical and genetic map of the *Bradyrhizobium japonicum* 110 genome. J. Bacteriol. 175:613-622.

121. Kunst, F., N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, P. Bessieres, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S. C. Brignell, S. Bron, S. Brouillet, C. V. Bruschi, B. Caldwell, V. Capuano, N. M. Carter, S. K. Choi, J.-J. Codani, I. F. Connerton, N. J. Cummings, R. A. Daniel, F. Denizot, K. M. Devine, A. Düsterhöft, S. D. Ehrich, P. T. Emmerson, K. D. Entian, J. Errington, C. Fabret, E. Ferrari, D. Foulger, C. Fritz, M. Fujita, Y. Fujita, S. Fuma, A. Galizzi, N. Galleron, S.-Y. Ghim, P. Glaser, A. Goffeau, E. J. Golightly, G. Grandi, G. Guiseppi, B. J. Guy, K. Haga, J. Haiech, C. R. Harwood, A. Hénaut, H. Hilbert, S. Holsappel, S. Hosono, M.-F. Hullo, M. Itaya, L. Jones, B. Joris, D. Karamata, Y. Kasahara, M. Klaerr-Blanchard, C. Klein, Y. Kobayash P. Koetter, G. Koningstein, S. Krogh, M. Kumano, K. Kurita, A. Lapidus, S. Lardinois, J. Lauber, V. Lazarevic, S.-M. Lee, A. Levine, H. Liu, S. Masuda, C. Mauël, C. Médique, N. Medina, R. P. Mellado, M. Mizuno, D. Moestl, S. Nakai, M. Noback, D. Noone, M. O'Reilly, K. Ogawa, A. Ogiwara, B. Oudega, S.-H. Park, V. Parro, T. M. Pohl, D. Portetelle, S. Porwollik, A. M. Prescott, E. Presecan, P. Pujic, B. Purnelle, G. Rapoport, M. Rey, S. Reynolds, M. Rieger, C. Rivolta, E. Rocha, B. Rocha, M. Rose, Y. Sadaie,

T. Sato, E. Scanian, S. Schleich, R. Schroeter, F. Scoffone, J. Sekiguchi, A. Sekowska, S. J. Seror, P. Serror, B.-S. Shin, B. Soldo, A. Sorokin, E. Tacconi, T. Takagi, H. Takahashi, K. Takemaru, M. Takeuchi, A. Tamakoshi, T. Tanaka, P. Terpstra, A. Tognoni, V. Tosato, S. Uchiyama, M. Vandenbol, F. Vannier, A. Vassarotti, A. Viari, R. Wambutt, E. Wedler, H. Wedler, T. Weitzenegger, P. Winters, A. Wipat, H. Yamamoto, K. Yamane, K. Yasumoto, K. Yata, K. Yoshida, H.-F. Yoshikawa, E. Zumstein, H. Yoshikawa, and A. Danchin. 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature. 390:249-256.

122. Ladefoged, S. A. and G. Christiansen. 1992. Physical and genetic mapping of the genomes of five *Mycoplasma hominis* strains by pulsed-field gel electrophoresis. J. Bacteriol. 174:2199-2207.

123. Lamfrom, H., A. Sarabhai, and J. Abelson. 1978. Cloning of *Beneckea* genes in *Escherichia coli*. J. Bacteriol. 133:354-363.

124. Le Bourgeois, P., M. Lautier, L. van den Berghe, M. J. Gasson, and P. Ritzenthaler. 1995. Physical and genetic map of the *Lactococcus lactis* subsp. *cremoris* MG1363 chromosome: comparison with that of *Lactococcus lactis* subsp. *lactis* IL 1403 reveals a large genome inversion. J. Bacteriol. 177:2840-2850.

125. Le Bourgeois, P., M. Lautier, M. Mata, and P. Ritzenthaler. 1992. Physical and genetic map of the chromosome of *Lactococcus lactis* subsp. *lactis* IL1403. J. Bacteriol. 174:6752-6762.

126. Leblond, P., M. Redenbach, and J. Cullum. 1993. Physical map of the *Streptomyces lividans* 66 genome and comparison with that of the related strain *Streptomyces coelicolor* A3(2). J. Bacteriol. 175:3422-3429.

127. Leonardi, M., M. Hobbs, and T. Kawula. 1995. *Haemophilus ducreyi* 35000 genome size estimate, p. 25. *In* Abstracts of the 95th General Meeting of the American Society for Microbiology 1995. American Society for Microbiology, Washington, D.C.

128. Leong-Morgenthaler, P., C. Ruettener, B. Mollet, and H. Hottinger. 1990. Construction of the physical map of *Lactobacillus bulgaricus*. FEMS Microbiol. Rev. 87:19.

129. Leth Bak, A., C. Christiansen, and A. Stenderup. 1970. Bacterial genome sizes determined by DNA renaturation studies. J. Gen. Microbiol. 64:377-380.

130. Lezhava, A., T. Mizukami, T. Kajitani, D. Kameoka, M. Redenbach, H. Shinkawa, O. Nimi, and H. Kinashi. 1995. Physical map of the linear chromosome of *Streptomyces griseus*. J. Bacteriol. 177:6492-6498.

131. Liesack, W. and E. Stackebrandt. 1989. Evidence for unlinked *rrn* operons in the planctomycete *Pirellula marina*. J. Bacteriol. 171:5025-5030.

132. Lin, N.-T., C.-H. Lo, and Y.-H. Tseng. 1997. Sequence and chromosomal location of *rrn* operons of *Xanthomonas campestris* pv. *campestris*, p.286. *In* Abstracts of the

97th General Meeting of the American Society for Microbiology 1997. American Society for Microbiology, Washington, D.C.

133. Lin, N.-T. and Y.-H. Tseng. 1997. Sequence and copy number of the *Xanthomonas campestris* pv. *campestris* gene encoding 16S rRNA. Biochem. and Biophys. Res. Comm. 235:276-280.

134. Lin, W.-J. and E. Johnson. 1995. Genome analysis of *Clostridium botulinum* type A by pulsed-field Gel electrophoresis. Appl.and Env. Microbiol. 61:4441-4447.

135. Linton, D., F. E. Dewhirst, J. P. Clewley, R. J. Owen, A. P. Burnens, and J. Stanley. 1994. Two types of 16S rRNA gene are found in *Campylobacter helveticus*: analysis, applications and characterization of some of the intervening sequence found in some strains. Microbiol. 140:847-855.

136. Liu, S.-L., A. Hessel, and K. E. Sanderson. 1993. Genomic mapping with I-*Ceu* I an intron-encoded endonuclease specific for genes for ribosomal RNA in *Salmonella* spp. *Escherichia coli* and other bacteria. Proc. Natl. Acad. Sci. USA 90:6874-6878.

137. Liu, S.-L. and K. E. Sanderson. 1995. The chromosome of *Salmonella paratyphi* A is inverted by recombination between *rrnH* and *rrnG*. J. Bacteriol. 177:6585-6592.

138. Liu, S.-L., A. Hessel, and K. E. Sanderson. 1993. The *Xba*I-*Bln*I-*Ceu*I genomic cleavage map of *Salmonella typhimurium* LT2 determined by double digestion, end labelling, and pulsed-field gel electrophoresis. J. Bacteriol. 175:4104-4120.

139. Liu, S.-L. and K. E. Sanderson. 1992. A physical map of the *Salmonella typhimurium* LT2 genome made by using *Xba*I analysis. J. Bacteriol. 174:1662-1672.

140. Liu, S.-L., A. Hessel, H. Y. Cheng, and K. E. Sanderson. 1994. The *Xba*I-*Bln*I-*Ceu*I genomic cleavage map of *Salmonella paratyphi* B. J. Bacteriol. 176:1014-1024.

141. Lopez-Garcia, P., J. P. Abad, C. Smith, and R. Amils. 1992. Genomic organization of the halophilic archaeon *Haloferax mediterranei*: physical map of the chromosome. Nucleic Acids Res. 20:2459-2464.

142. Lopez-Garcia, P., A. St.Jean, R. Amils, and R. L. Charlebois. 1995. Genomic stability in the archaeae *Haloferax volcanii* and *Haloferax mediterranei*. J. Bacteriol. 177:1405-1408.

143. Lortal, S., A. Rouault, S. Guezenec, and M. Gautier. 1997. *Lactobacillus helveticus*: Strain typing and genome size estimation by pulsed field gel electrophoresis. Curr. Microbiol. 34:180-185.

144. MacDonald, M. and T. Melton. 1996. The organization and copy number of the rRNA genes of Azotobacteracea, abstr. H-24, p. 487. *In* Abstracts of the 96th General Meeting of the American Society for Microbiology 1996. American Society for Microbiology, Washington, D.C.

145. Maidak, B. L., G. J. Olsen, N. Larsen, R. Overbeek, M. J. McCaughey, and C. R. Woese. 1997. The RDP (Ribosomal Database Project). Nucleic Acids Res. 25:109-111.

146. Majumder, R., S. Sengupta, G. Khetawat, R. K. Bhadra, S. Roychoudhury, and J. Das. 1996. Physical map of the genome of *Vibrio cholerae* 569B and localization of genetic markers. J. Bacteriol. 178:1105-1112.

147. Maniloff, J. and S. Poddar. 1989. Genome sizes of *Mycoplasma* species, abstr. G-29, p.153. *In* Abstracts of the 89th General Meeting of the American Society for Microbiology 1989. American Society for Microbiology, Washington, D.C.

148. Marin, I., R. Amils, and J. P. Abad. 1997. Genomic organization of the metal-mobilizing bacterium *Thiobacillus cuprinus*. Gene 187:99-105.

149. Matsuda, M., Y. Asami, T. Miyazawa, T. Sugawara, M. Kumano, Y. Isayama, and M. Honda. 1994. Estimation of the size of the genome of *Taylorella equigenitalis* by crossed-field gel electrophoresis. Vet. Res. Comm. 18:99-102.

150. Mellado, E., M. T. Garcia, J. J. Nieto, S. Kaplan, and A. Ventosa. 1997. Analysis of the genome of *Vibrio costicola*: Pulsed-field gel electrophoretic analysis of genome size and plasmid content. Syst. and Appl. Microbiol. 20:20-26.

151. Michaux, S., J. Paillisson, M. J. Carles-Nurit, G. Bourg, A. Allardet-Servent, and M. Ramuz. 1993. Presence of two independent chromosomes in the *Brucella melitensis* 16M genome. J. Bacteriol. 175:701-705.

152. Michel, E. and P. Cossart. 1992. Physical map of the *Listeria monocytogenes* chromosome. J. Bacteriol. 174:7098-7103.

153. Miyata, M., W. Lu, and T. Fukumura. 1991. Physical mapping of the *Mycoplasma capricolum* genome. FEMS Microbiol. Lett. 79:329-334.

154. Möller, A. and J. K. Jansson. 1997. Quantification of genetically tagged cyanobacteria in baltic sea sediment by competitive PCR. BioTechniques 22:512-518.

155. Moreira, L., M. F. Nobre, I. Sa-Correia, and M. S. D. Costa. 1996. Genomic typing and fatty acid composition of *Rhodothermus marinus*. Systematic and Applied Microbiology. 19:83-90.

156. Mullis, K. B. and F. A. Faloona. 1987. Specific synthesis of DNA *in vitro* in a polymerase catalyzed chain reaction. Meth. Enzymol. 155:335-350.

157. Munson, M. A., L. Baumann, and P. Baumann. 1993. *Buchnera aphidicola* (a prokaryotic endosymbiont of aphids) contains a putative 16S rRNA operon unlinked to the 23S rRNA-encoding gene: sequence determination, and promoter and terminator analysis. Gene. 137:171-178.

158. Murray, B. E., K. V. Singh, R. P. Ross, J. D. Heath, G. M. Dunny, and G. M. Weinstock. 1993. Generation of restriction map of *Enterococcus faecalis* OG1 and

investigation of growth requirements and regions encoding biosynthetic function. J. Bacteriol. 175:5216-5223.

159. Naterstad, K., A.-B. Kolstø, and R. Sirevag. 1995. Physical map of the genome of the green phototrophic bacterium *Chlorobium tepidum*. J. Bacteriol. 177:5480-5484.

160. National Center for Genome Resources. Personal Communication.

161. Neumann, H., A. Gierl, J. Tu, J. Leibrock, D. Staiger, and W. Zillig. 1983. Organization of the genes for ribosomal RNA in Archaebacteria. Mol. Gen. Genet. 192:66-72.

162. Newnham, E., N. Chang, and D. Taylor. 1996. Expanded genomic map of *Campylobacter jejuni* UA580 and localization of 23S ribosomal rRNA genes by I-*Ceu*I restriction endonuclease digestion. FEMS Microbiol. Lett. 142:223-229.

163. Noll, K. M. 1989. Chromosome map of the thermophilic archaebacterium *Thermococcus celer*. J. Bacteriol. 171:6720-6725.

164. Nuijten, P. J. M., C. Bartels, N. M. C. Bleumink-Pluym, W. Gaastra, and B. A. M. van der Zeijst. 1990. Size and physical map of the *Campylobacter jejuni* chromosome. Nucleic Acids Res. 18:6211-6214.

165. Ojaimi, C., B. E. Davidson, I. Saint Girons, and I. G. Old. 1994. Conservation of gene arrangement and an unusual organization of rRNA genes in the linear chromosomes of the lyme disease spirochaetes *Borrelia burgdorferi*, *B. garinii* and *B. afzelii*. Microbiol. 140:2931-2940.

166. Okahashi, N., C. Sasakawa, N. Okada, M. Tamada, M. Yoshikawa, M. Tokuda, I. Takahashi, and T. Koga. 1990. Construction of a *Not*I restriction map of the *Streptococcus mutans* genome. J. Gen. Microbiol. 136:2217-2223.

167. Old, I. G., J. MacDougall, I. Saintgirons, and B. E. Davidson. 1992. Mapping of the genes on the linear chromosome of the bacterium *Borrelia burgdorferi*: possible locations for its origin of replication. FEMS Microbiol. Lett. 99:245-250.

168. Olsen, G. J., D. J. Lane, S. J. Giovannoni, and N. R. Pace. 1986. Microbial ecology and evolution: A ribosomal approach. Ann. Rev. Microbiol. 40:337-365.

169. Pace, N. R., D. A. Stahl, D. J. Lane, and G. J. Olsen. 1986. The analysis of natural microbial populations by ribosomal RNA sequences. Adv. Microbiol. Ecol. 9:1-55.

170. Pang, H. and H. H. Winkler. 1993. Stable RNA concentration, ribosome concentration, and the copy number of 16S ribosomal RNA gene in *Rickettsia prowazekii*, p.109. *In* Abstracts of the 93rd General Meeting of the American Society for Microbiology 1993. American Society for Microbiology, Washington, D.C.

171. Pang, H. L. and H. H. Winkler. 1993. Copy number of the 16S rRNA gene in *Rickettsia prowazekii*. J. Bacteriol. 175:3893-3896.

172. Park, J. H., J.-C. Song, M. H. Kim, D.-S. Lee, and C.-H. Kim. 1994. Determination of genome size and a preliminary physical map of an extreme alkaliphile, *Micrococcus sp.* Y-1, by pulsed-field gel electrophoresis. Microbiol. 140:2247-2250.

173. Patel, A. H., T. J. Foster, and P. A. Pattee. 1989. Physical and genetic mapping of the protein-A gene in the chromosome of *Staphylococcus aureus* 8325-4. J. Gen. Microbiol. 135:1799-1807.

174. Pattee, P. A. 1993. The genetic map of *Staphylococcus aureus*, p.489-496. *In* J.A. Hoch and R. Losick (ed.), *Bacillus subtilis* and other gram-positive bacteria: biochemistry, physiology and molecular genetics, American Society for Microbiology, Washington, D.C..

175. Philipp, W., S. Poulet, K. Eiglmeier, L. Pascopella, V. Balasubramanian, B. Heym, S. Bergh, B. Bloom, W. R. Jacobs, and S. Cole. 1996. An integrated map of the genome of the tubercle bacillus, *Mycobacterium tuberculosis* H37Rv, and comparison with *Mycobacterium leprae*. Proc. Natl. Acad. Sci. USA. 93:3132-3137.

176. Piatak, M., K. -C. Luk, B. Williams, and J. D. Lifson. 1993. Quantitative competitive polymerase chain reaction for accurate quantitation of HIV DNA and RNA species. BioTechniques. 14:70-80.

177. Prevost, H., J. F. Cavin, M. Lamoureux, and C. Divies. 1995. Plasmid and chromosome characterization of *Leuconostoc oenos* strains. Am. Jour. of Enol. and Viticult. 46:43-48.

178. Pyle, L. E., L. N. Corcoran, B. G. Cocks, A. D. Bergemann, J. C. Whitley, and L. R. Finch. 1988. Pulsed-field electrophoresis indicated larger-than-expected sizes for *Mycoplasma* genomes. Nucleic Acids Res. 16:6015-6025.

179. Pyle, L. E., R. Taylor, and L. R. Finch. 1990. Genomic maps of some strains within the *Mycoplasma mycoides* cluster. J. Bacteriol. 172:7265-7268.

180. Rainey, P., I. Thompson, and N. Palleroni. 1994. Genome and fatty acid analysis of *Pseudomonas stutzeri*. Int. Jour. of System. Bacteriol. 44:54-61.

181. Ratnaningsih, E., S. Dharmsthithi, V. Krishnapillai, A. Morgan, M. Siclair, and B. W. Holloway. 1990. A combined physical and genetic map of *Pseudomonas aeruginosa* PAO. J. Gen. Microbiol. 136:2351-2357.

182. Revillion, F., L. Hornez, and J.-P. Peyrat. 1997. Quantification of *c-erb*B-2 gene expression in breast cancer by competitive RT-PCR. Clin. Chem. 43:2114-2120.

183. Robertson, J. 1993. Genome sizes of *Ureaplasma urealyticum* appear related to serovar, p. 167. *In* Abstracts of the 93rd General Meeting of the American Society for Microbiology 1993. American Society for Microbiology, Washington, D.C.

184. Robertson, J., R. Sherbourne, T. Pirnak, and J. Wong. 1995. *Mycoplasma penetrans*: ultrastructure, genome size, and generation time, p. 304. *In* Abstracts of

the 95th General Meeting of the American Society for Microbiology 1995. American Society for Microbiology, Washington, D.C.

185. Robertson, J. A., L. E. Pyle, G. W. Stemke, and L. R. Finch. 1990. Human ureaplasmas show diverse genome sizes by pulse-field electrophoresis. Nucleic Acids Res. 18:1451-1455.

186. Rodley, P., U. Roemling, and B. Tuemmler. 1995. A physical genome map of the *Burkholderia cepacia* type strain. Mol. Microbiol. 17:57-67.

187. Romling, U., D. Grothues, W. Bautch, and B. Tummler. 1989. A physical map of *Pseudomonas aeruginosa* PAO. EMBO J. 8:4081-4089.

188. Romling, U. and B. Tummler. 1991. The impact of two-dimensional pulsed-field gel electrophoresis techniques for the consistent and complete mapping of bacterial genomes: refined physical map of *Pseudomonas aeruginosa* PAO. Nucleic Acids Res. 19:3199-3206.

189. Rouhbakhsh, D. and P. Baumann. 1995. Characterization of a putative 23S-5S rRNA operon of *Buchnera aphidicola* (endosymbiont of aphids) unlinked to the 16S rRNA-encoding gene. Gene 155:107-112.

190. Roussel, Y., C. Colmin, J.-M. Simonet, and B. Decaris. 1993. Strain characterization genome size and plasmid content in the *Lactobacillus acidophilus* group hansen and mocquot. J. of Appl. Bacteriol. 74:549-556.

191. Roussel, Y., M. Pebay, G. Guedon, J.-M. Simonet, and B. Decaris. 1994. Physical and genetic map of *Streptococcus thermophilus* A054. J. Bacteriol. 176:7413-7422.

192. Roux, V. and D. Raoult. 1993. Genotypic identification and phylogenetic analysis of the spotted fever group *Rickettsiae* by pulsed-field gel electrophoresis. J. Bacteriol. 175:4895-4904.

193. Salama, S., E. Newnham, N. Chang, and D. Taylor. 1995. Genome map of *Campylobacter fetus* subsp. *fetus* ATCC 27374. FEMS Microbiol. Lett. 132:239-245.

194. Sanderson, K. E., A. Hessel, and S.-L. Liu. 1995. The chromosomes of *Salmonella typhi* and *S. paratyphi* C are rearranged in comparison with those of *E. coli* K-12, *S. typhimurium* LT2 and other *Salmonella* spp. Journal of Cellular Biochemistry 19A:116.

195. Schmidt, K. D., B. Tummler, and U. Romling. 1996. Comparative genome mapping of *Pseudomonas aeruginosa* PAO with *P. aeruginosa* C, which belongs to a major clone in cystic fibrosis patients and aquatic habitats. J. Bacteriol. 178:85-93.

196. Schwartz, J. J., A. Gazumyan, and I. Schwartz. 1992. rRNA gene organization in the lyme disease spirochete *Borrelia burgdorferi*. J. Bacteriol. 174:3757-3765.

197. Sechi, L. A., F. M. Zuccon, J. E. Mortensen, and L. Daneo-Moore. 1994. Ribosomal RNA gene (*rrn*) organization in enterococci. FEMS Microbiol. Lett. 120:307-313.

198. Sensen, C., H.-P. Klenk, R. Singh, G. Allard, C. Chan, Q. Liu, S. Penny, F. Young, M. Schenk, T. Gaasterland, W. Doolittle, M. Ragan, and R. Charlebois. 1996. Organizational characteristics and information content of an archaeal genome: 156 kb of sequence from *Sulfolobus solfataricus* P2. Mol. Microbiol. 22:175-191.

199. Shao, Z., W. Mages, and R. Schmitt. 1994. A physical map of the hyperthermophilic bacterium *Aquifex pyrophilus* chromosome. J. Bacteriol. 176:6776-6780.

200. Sitzmann, J. and A. Klein. 1991. Physical and genetic map of the *Methanococcus voltae* chromosome. Mol. Microbiol. 5:505-513.

201. Skerman, V. B. D., V. McGowan, and P. H. A. Sneath. 1989. Approved lists of bacterial names (amended edition). American Society for Microbiology, Washington, D.C..

202. Smith, C. L. and C. R. Cantor. 1987. Purification, specific fragmentation, and separation of large DNA molecules, p.449-467. *In* R. Wu (ed.), Methods in Enzymology, Volume 155, Academic Press, San Diego.

203. Smith, D. R., L. A. Doucette-Stamm, C. Deloughery, H. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert, D. Harrison, L. Hoang, P. Keagle, W. Lumm, B. Pothier, D. Qiu, R. Spadafora, R. Vicaire, Y. Wang, J. Wierzbowski, R. Gibson, N. Jiwani, A. Caruso, D. Bush, H. Safer, D. Patwell, S. Prabhakar, S. McDougall, G. Shimer, A. Goyal, S. Pietrokovski, G. M. Church, C. J. Daniels, J. Mao, P. Rice, J. Nölling, and J. N. Reeve. 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: functional analysis and comparative genomics. J. Bacteriol. 179:7135-7155.

204. St Jean, A., B. Trieselmann, and R. Charlebois. 1994. Physical map and set of overlapping cosmid clones representing the genome of the archaeon *Halobacterium sp.* GRB. Nucleic Acids Res. 22:1476-1483.

205. Stanley, J., C. Jones, A. Burnens, and R. J. Owen. 1994. Distinct genotypes of human and canine isolates of *Campylobacter upsaliensis* determined by 16S rRNA gene typing and plasmid profiling. J. Clin. Microbiol. 32:1788-1794.

206. Stanton, T. B. and R. L. Zuerner. 1993. *Serpulina hyodysenteriae* genome map, abstr. H-58, p. 200. *In* Abstracts of the 93rd General Meeting of the American Society for Microbiology 1993. American Society for Microbiology, Washington, D.C.

207. Stettler, R., G. Erauso, and T. Leisinger. 1995. Physical and genetic map of the *Methanobacterium wolfei* genome and its comparison with the updated genomic map of *Methanobacterium thermoautotrophicum* Marburg. Arch. Microbiol. 163:205-210.

208. Stibitz, S. and T. L. Garletts. 1992. Derivation of a physical map of the chromosome of *Bordetella pertussis* Tohama I. J. Bacteriol. 174:7770-7777.

209. Suwanto, A. and S. Kaplan. 1989. Physical and genetic mapping of the *Rhodobacter sphaeroides* 2.4.1 genome: presence of two unique circular chromosomes. J. Bacteriol. 171:5850-5859.

210. Suwanto, A. and S. Kaplan. 1989. Physical and genetic mapping of the *Rhodobacter sphaeroides* 2.4.1 genome: genome size, fragment identification, and gene location. J. Bacteriol. 171:5840-5849.

211. Suzuki, T., T. Mori, Y. Miyata, and T. Yamada. 1987. The number of ribosomal RNA genes in *Mycobacterium lepraemurium.* FEMS Microbiol. Lett. 44:73-76.

212. Suzuki, Y., Y. Ono, A. Nagata, and T. Yamada. 1988. Molecular cloning and characterization of an rRNA operon in *Streptomyces lividans* TK21. J. Bacteriol. 170:1631-1636.

213. Suzuki, Y., K. Yoshinaga, Y. Ono, A. Nagata, and T. Yamada. 1987. Organization of rRNA genes in *Mycobacterium bovis* BCG. J. Bacteriol. 169:839-843.

214. Swafford, D. 1993. Phylogenetic Analysis Using Parsimony (PAUP), Smithsonian Institution, Washington, D.C.

215. Tabata, K. and T. Hoshino. 1996. Mapping of 61 genes on the refined physical map of the chromosome of *Thermus thermophilus* HB27 and comparison of genome organization with that of *T. thermophilus* HB8. Microbiol. 142:401-410.

216. Tanskanen, E. I., D. L. Tulloch, A. J. Hillier, and B. E. Davidson. 1990. Pulsed-field gel electrophoresis of *Sma*I digests of lactococcal genomic DNA, a novel method of strain identification. Appl. and Env. Microbiol. 56:3105-3111.

217. Taschke, C., M.-Q. Klinkert, J. Wolters, and R. Herrmann. 1986. Organization of the ribosomal RNA genes in *Mycoplasma hyopneumoniae*: The 5S rRNA gene is separated from the 16S and the 23S rRNA genes. Mol. Gen. Genet. 205:428-433.

218. Taylor, D. E., N. Chang, N. S. Taylor, and J. G. Fox. 1994. Genome conservation in *Helicobacter mustelae* as determined by pulsed-field gel electrophoresis. FEMS Microbiol. Lett. 118:31-36.

219. Taylor, D. E., M. Eaton, W. Yan, and N. Chang. 1992. Genome maps of *Campylobacter jejuni* and *Campylobacter coli.* J. Bacteriol. 174:2332-2337.

220. Tenreiro, R., M. A. Santos, H. Paveia, and G. Vieira. 1994. Inter-strain relationships among wine leuconostocs and their divergence from other *Leuconostoc* species, as revealed by low frequency restriction fragment analysis of genomic DNA. Journal of Applied Bacteriology 77:271-280.

221. The Institute for Genetic Research. Personal Communication.

222. Thong, K. L., S. D. Puthucheary, and T. Pang. 1997. Genome size variation among recent human isolates of *Salmonella typhi.* Res. Microbiol. 148:229-235.

223. Tigges, E. and F. C. Minion. 1994. Physical map of the genome of *Acholeplasma oculi* ISM1499 and construction of a Tn4001 derivative for macrorestriction chromosomal mapping. J. Bacteriol. 176:1180-1183.

224. Tomb, J.-F., O. White, A. R. Kerlavage, R. A. Clayton, G. G. Sutton, R. D. Fleischmann, K. A. Ketchum, H. P. Klenk, S. Gill, B. A. Dougherty, K. Nelson, J. Quackenbush, L. Zhou, E. F. Kirkness, S. Peterson, B. Loftus, D. Richardson, R. Dodson, H. G. Khalak, A. Glodek, K. McKenney, L. M. Fitzegerald, N. Lee, M. D. Adams, E. K. Hickey, D. E. Berg, J. D. Gocayne, T. R. Utterback, J. D. Peterson, J. M. Kelley, M. D. Cotton, J. M. Weldman, C. Fujii, C. Bowman, L. Watthey, E. Wallin, W. S. Hayes, M. Borodovsky, P. D. Karp, H. O. Smith, C. M. Fraser, and J. C. Venter. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori.* Nature. 388:539-547.

225. Tu, J. and W. Zillig. 1982. Organization of rRNA structural genes in the archaebacterium *Thermoplasma acidophilum.* Nucleic Acids Res. 10:7231-7245.

226. Tulloch, D. A., L. R. Finch, A. J. Hillier, and B. E. Davidson. 1991. Physical map of the chromosome of *Lactococcus lactis* subsp. *lactis* DL11 and localization of six putative rRNA operons. J. Bacteriol. 173:2768-2775.

227. VanWezel, G. P., M. J. Buttner, W. Vijgenboom, and L. Bosch. 1995. Mapping of the genes involved in macromolecular synthesis on the chromosome of *Streptomyces coelicolor* A3(2). J. Bacteriol. 177:473-476.

228. Vary, P. S. 1993. The genetic map of *Bacillus megaterium*, p.475-481. *In* J.A. Hoch and R. Losick (ed.), *Bacillus subtilis* and other gram-positive bacteria: biochemistry, physiology and molecular genetics, American Society for Microbiology, Washington, D.C.

229. Ward, D. M., R. Weller, and M. M. Bateson. 1990. 16S rRNA sequence reveal numerous uncultured microorganisms in a natural community. Nature. 345:63-65.

230. Ward-Rainey, N., F. A. Rainey, E. M. H. Wellington, and E. Stackebrandt. 1996. Physical map of the genome of *Planctomyces limnophilus*, a representative of the phylogenetically distinct planctomycete lineage. J. Bacteriol. 178:1908-1913.

231. Weil, M. D. and M. McClelland. 1989. Enzymatic cleavage of a bacterial genome at a 10-base-pair recognition site. Proc. Natl. Acad. Sci. USA. 86:51-55.

232. Wickman, T. and J. D. Wall. 1994. Analysis of the *Desulfovibrio desulfuricans* G201 genome using pulsed-field gel electrophoresis, abstr. H-194, p. 234. *In* Abstracts of the 94th General Meeting of the American Society for Microbiology 1994. American Society for Microbiology, Washington, D.C.

233. Wilkinson, S. R. and M. Young. 1993. Wide diversity of genome size among different strains of *Clostridium acetobutylicum.* J. Gen. Microbiol. 139:1069-1076.

234. Williamson, D. L., J. R. Adams, R. F. Whitcomb, J. G. Tully, P. Carle, M. Konai, J. M. Bove, and R. B. Henegar. 1997. *Spiroplasma platyhelix* sp. nov., a new mollicute with unusual morphology and genome size from the dragonfly *Pachydiplax longipennis*. Int. J. System. Bacteriol. 47:763-766.

235. Wong, K. K. and M. McClelland. 1992. A *Bln*I restriction map of the *Salmonella typhimurium* LT2 genome. J. Bacteriol. 174:1656-1661.

236. Yamagishi, A. and T. Oshima. 1990. Circular chromosomal DNA in the sulfur-dependent archaebacterium *Sulfolobus acidocaldarius.* Nucleic Acids Res. 18:1133-1135.

237. Yan, W. and D. E. Taylor. 1991. Sizing and mapping of the genome of *Campylobacter coli* strain UA417R using pulsed-field electrophoresis. Gene. 101:117-120.

238. Ye, F., F. Laigret, J. C. Whitley, C. Citti, L. R Finch, P. Carle, J. Renaudin, and J.-M. Bove. 1992. A physical and genetic map of the *Spiroplasma citri* genome. Nucleic Acids Res. 20:1559-1565.

239. Young, D. B. and S. T. Cole. 1993. Leprosy, tuberculosis, and the new genetics. J. Bacteriol. 175:1-6.

240. Zhang, Y. H., Y. Takahashi, and M. Fukunaga. 1993. Organization of ribosomal RNA genes in *Borrelia burgdorferi* sensu lato isolated from *Ixodes ovatus* in Japan. Microbiol. Immunol. 37:909-913.

241. Zuccon, F. M., L. A. Sechi, and L. Daneo-Moore. 1993. Physical and genetic map of *Enterococcus hirae* ATCC 9790, p.201. *In* Abstracts of the 93rd General Meeting of the American Society for Microbiology 1993. American Society for Microbiology, Washington, D.C.

242. Zuerner, R., J. L. Hermann, and I. Saint Girons. 1993. Comparison of genetic maps for two *Leptospira interrogans* serovars provides evidence for two chromosomes and intraspecies heterogenity. J. Bacteriol. 175:5445-5451.

243. Zuerner, R. L. 1991. Physical map of chromosomal and plasmid DNA comprising the genome of *Leptospira interrogans.* Nucleic Acids Res. 19:4857-4860.

244. Zuerner, R. L. and T. B. Stanton. 1994. Physical and genetic map of the *Serpulina hyodysenteriae* B78-T chromosome. J. Bacteriol. 176:1087-1092.

Figure A2.1

User derived tree reflective of RDP Prokaryotic Taxonomic List groupings. The 89 taxa have values for both genome size and SSU rRNA copy number and are in Table A2.1. See Table A2.1 for definitions of species identifiers. Numbers in parentheses give genome size (bp x 10^5) and SSU rRNA copy number.

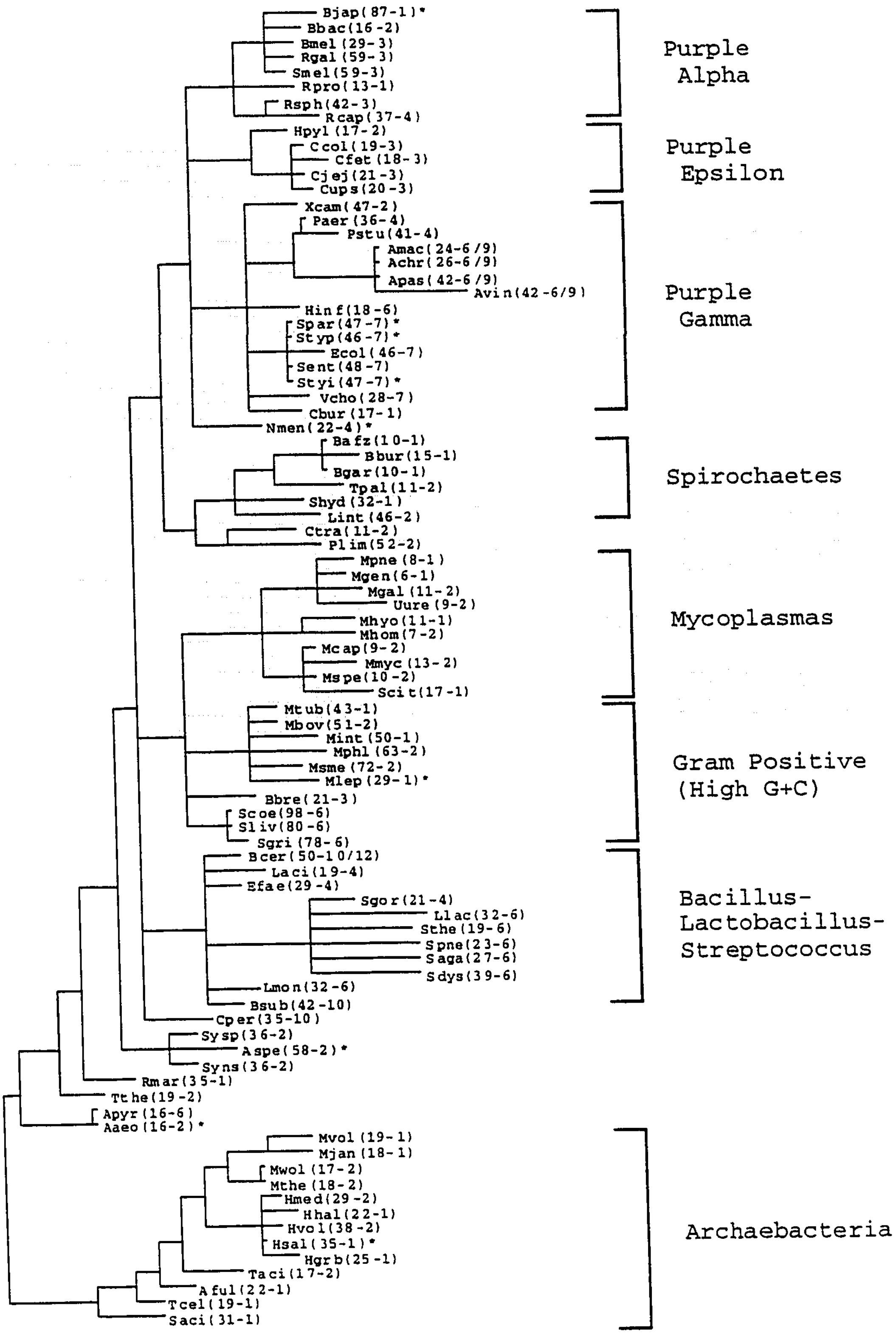

Bjap(87-1)*
Bbac(16-2)
Bmel(29-3)
Rgal(59-3)
Smel(59-3)
Rpro(13-1)
Rsph(42-3)
Rcap(37-4)
Hpyl(17-2)
Ccol(19-3)
Cfet(18-3)
Cjej(21-3)
Cups(20-3)
Xcam(47-2)
Paer(36-4)
Pstu(41-4)
Amac(24-6/9)
Achr(26-6/9)
Apas(42-6/9)
Avin(42-6/9)
Hinf(18-6)
Spar(47-7)*
Styp(46-7)*
Ecol(46-7)
Sent(48-7)
Styi(47-7)*
Vcho(28-7)
Cbur(17-1)
Nmen(22-4)*
Bafz(10-1)
Bbur(15-1)
Bgar(10-1)
Tpal(11-2)
Shyd(32-1)
Lint(46-2)
Ctra(11-2)
Plim(52-2)
Mpne(8-1)
Mgen(6-1)
Mgal(11-2)
Uure(9-2)
Mhyo(11-1)
Mhom(7-2)
Mcap(9-2)
Mmyc(13-2)
Mspe(10-2)
Scit(17-1)
Mtub(43-1)
Mbov(51-2)
Mint(50-1)
Mphl(63-2)
Msme(72-2)
Mlep(29-1)*
Bbre(21-3)
Scoe(98-6)
Sliv(80-6)
Sgri(78-6)
Bcer(50-10/12)
Laci(19-4)
Efae(29-4)
Sgor(21-4)
Llac(32-6)
Sthe(19-6)
Spne(23-6)
Saga(27-6)
Sdys(39-6)
Lmon(32-6)
Bsub(42-10)
Cper(35-10)
Sysp(36-2)
Aspe(58-2)*
Syns(36-2)
Rmar(35-1)
Tthe(19-2)
Apyr(16-6)
Aaeo(16-2)*
Mvol(19-1)
Mjan(18-1)
Mwol(17-2)
Mthe(18-2)
Hmed(29-2)
Hhal(22-1)
Hvol(38-2)
Hsal(35-1)*
Hgrb(25-1)
Taci(17-2)
Aful(22-1)
Tcel(19-1)
Saci(31-1)
Purple Alpha
Purple Epsilon
Purple Gamma
Spirochaetes
Mycoplasmas
Gram Positive (High G+C)
Bacillus-Lactobacillus-Streptococcus
Archaebacteria

Figure A2.2

Histogram of known and estimated genome sizes for prokaryotic organisms found in the literature (N=303). The distribution has a mean of 3,604 kb and standard deviation of 1,996 kb. The data do not follow a normal distribution (X^2 =25.19).

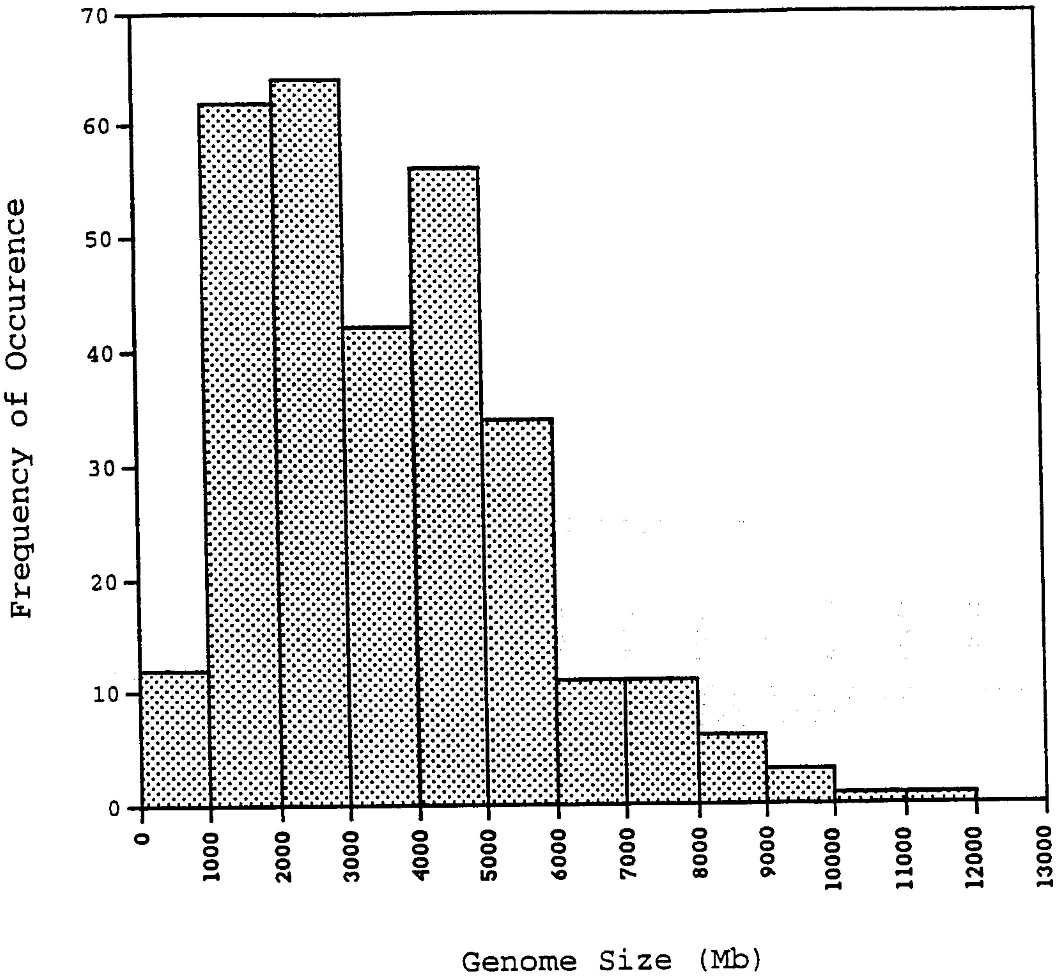
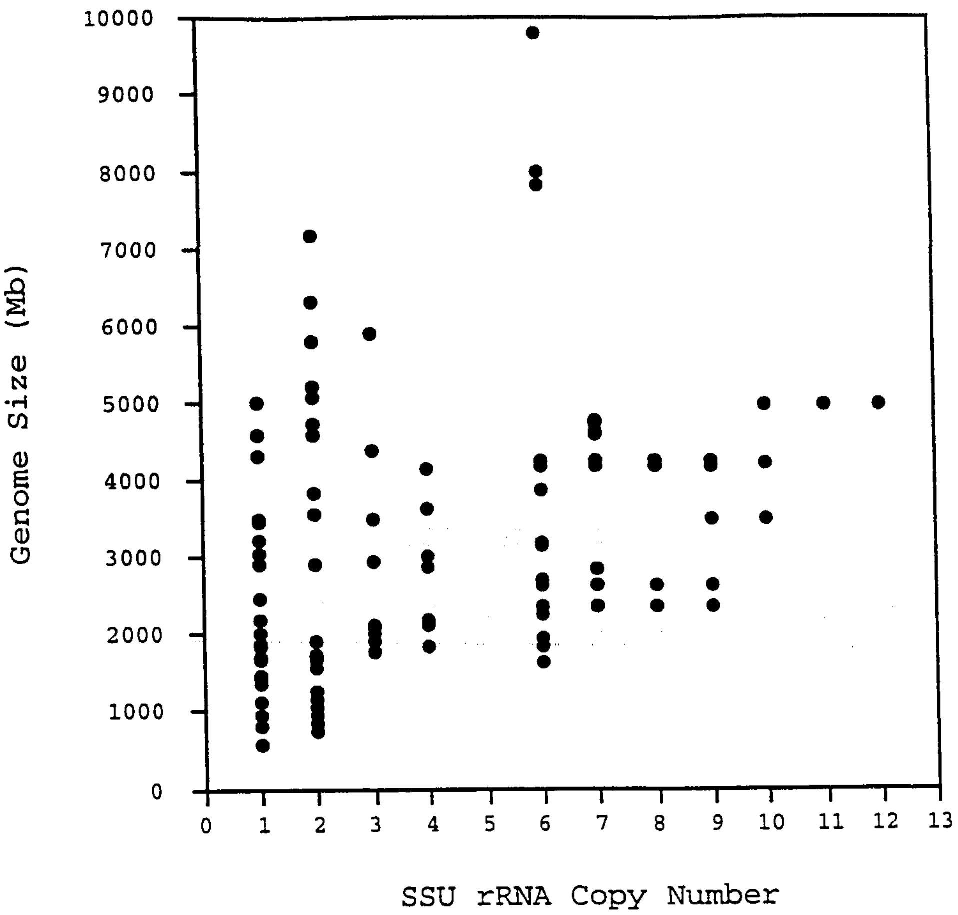

70
60
50
40
30
20
10
0
Frequency of Occurence
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
Genome Size (Mb)

Figure A2.3

Distribution of prokaryotic genome size versus SSU rRNA copy number for 103 taxa. For taxa with more than one SSU rRNA copy number reported in the literature, each value was used in the histogram, bringing the total number of entries to 120. Increased genome size shows virtually no correlation with increased copy number ($y = 2268 * x$; $r^2 = 0.128$).



Genome Size (Mb)
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
SSU rRNA Copy Number

Figure A2.4

Histogram of SSU rRNA copy numbers for the set of 103 taxa with estimated copy number. The mean rRNA copy number is 3.8 with a standard deviation of 2.9. The data are not normally distributed

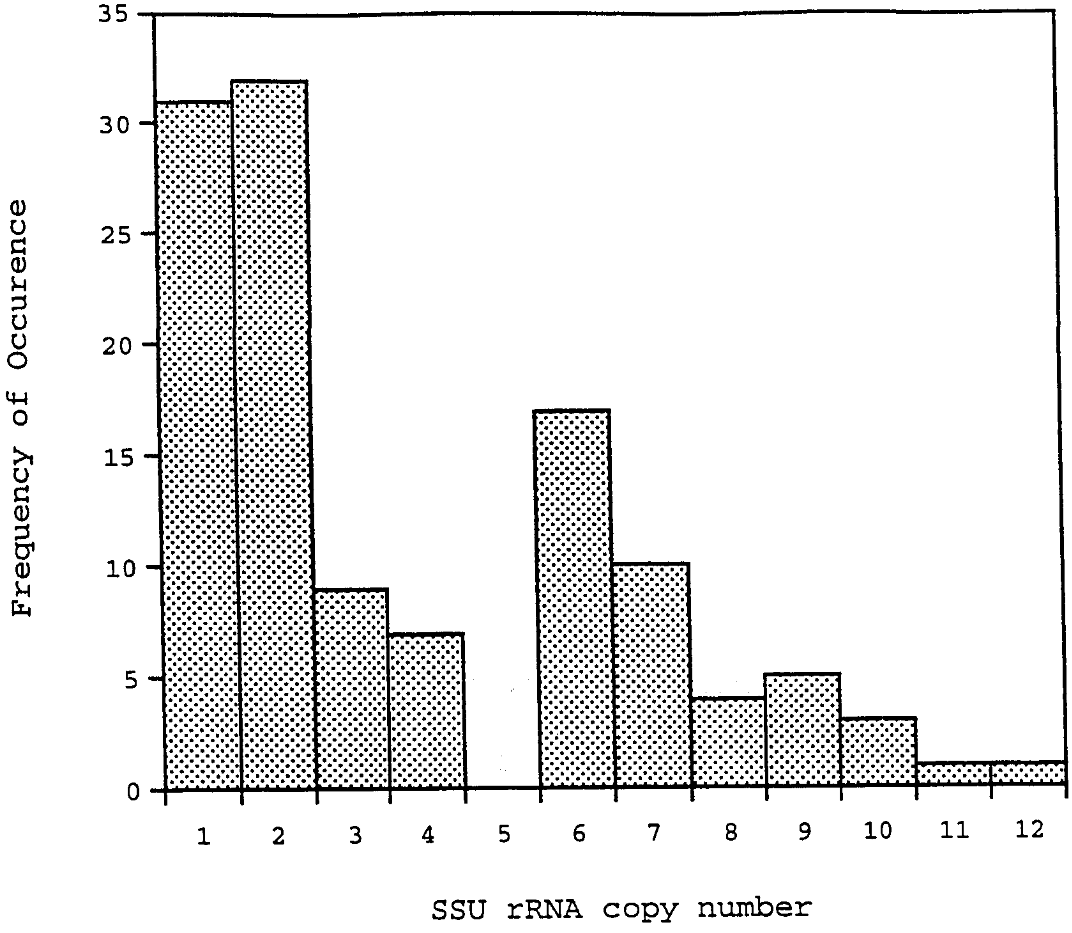

35
30
25
20
15
10
5
0
Frequency of Occurence
1
2
3
4
5
6
7
8
9
10
11
12
SSU rRNA copy number

Table A2.1 Genome sizes (N=303) and SSU rRNA copy numbers (N=103) for bacterial taxa listed according to taxonomic ranking using the Ribosomal Database Project numbering system. Genome sizes are given in terms of kilobase pairs (kb). For species with more than one genome size estimate, an average and standard deviation are shown. Identifiers are used on the brach tips in Figure A2.1. An asterisk (*) denotes taxa without a complete SSU rRNA sequence in RDP (see Figure A2.1). Complete genomes are denoted with a (†) symbol and are given in exact base pairs.

RDP	Taxa	Identifier	Genome Size (kb)	Ref.	SSU rRNA copy number	Ref.
1	**Archaebacteria**					
1.1	**Euryarchaeota**					
1.1.1	*Methanococcus jannaschii*	Mjan(17-2)	1664974†	34	2	34
1.1.1	*Methanococcus thermolithotrophicus*		1100	113		200
1.1.1	*Methanococcus voltae*	Mvol(18-1)	1840±57	113,200	1	82,204
1.1.2	*Methanobacterium thermoautotrophicum*	Mthe(18-2)	1751377†	204	2	207
1.1.2	*Methanobacterium wolfei*	Mwol(17-2)	1729	207	2	207
1.1.2	*Methanobrevibater arboriphilicus*		1800	113		
1.1.3.2.1	*Methanosarcina barkeri*		1100	113		
1.1.3.3	*Halobacterium halobium*	Hhal(22-1)	2200	24	1	161
1.1.3.3	*Halobacterium salinarium*	Hsal(35-1)*	3500±866	104,160,221	1	104
1.1.3.3	*Halobacterium sp. GRB*	Hgrb(25-1)	2470	205	1	205
1.1.3.3	*Halobacterium volcanii*		4000	43		
1.1.3.3	*Haloferax (Halobacterium) mediterranei*	Hmed(29-2)	2900	141,142	2	141,142
1.1.3.3	*Haloferax volcanii*	Hvol(38-2)	3820±613	44,55,104, 142	2	44,55,104
1.1.3.4	*Thermoplasma acidophilum*	Taci(17-2)	1700	221	2	225
1.1.3.5	*Archaeoglobus fulgidus*	Aful(22-1)	2178400†	114	1	114
1.1.4	*Pyrococcus furiosus*		2100	160,221		
1.1.?	*Pyrococcus horikoshii*		2000	160		
1.1.?	*Pyrococcus shinkaj*		2000	221		
1.1.4	*Thermococcus celer*	Tcel(19-1)	1890	163	1	161
	1.1 Average		2197±816			
1.2	**Crenarchaeota**					

1.2.?	*Crenarchaeum symbiosum*		2500	160	1	161
1.2.?	*Desulfurococcus Hvv3*				1	161
1.2.?	*Desulfurococcus mucosis*					
1.2.1	*Pyrobaculum aerophilum*		2223	68,221,		
1.2.1	*Sulfolobus acidocaldarius*	Saci(31-1)	3050	236	1	161
1.2.1	*Sulfolobus solfatricus*		3025±35	160,198,221	1	161
1.2.1	*Thermoproteus tenax*				1	161
1.2.2	*Thermofilum pendens*					
	1.2 Average		2700±406			
	1.0 Average		2449±357			
2	**Bacteria**					
2.1	**Thermophilic Oxygen Reducers**					
2.?	*Aquifex aeolicus*	Aaeo(16-2)*	1551328'	68,221	2	68
2.1	*Aquifex pyrophilus*	Apyr(16-6)	1620	199	6	199
	2.1 Average		1586±49			
2.2	**Thermotogales**					
2.2	*Thermotoga maritima*		1800	221		
2.4	**Green Non-Sulfur Bacteria and Relatives**					
2.4.2	*Deinococcus radiodurans*		3118±167	94,221		
2.4.2.1	*Thermus thermophilus*	Tthe(19-2)	1905±233	27,215	2	27,90,92
	2.4 Average		2512±858			
2.7	**Flexibacter-Cytophaga-Bacteroides Phylum**					
2.7.1.1.2	*Porphyromonas gingivalis*		2200	160,221		
2.7.1.2.?	*Flavobacterium balustinum*		4838	94		
2.7.1.2.?	*Flavobacterium breve*		5040±217	94		

2.7.1.2.5	*Flavobacterium group IIb*		5291	94		
2.7.1.2.5	*Flavobacterium group IIf*		2832	94		
2.7.1.2.5	*Flavobacterium group IIk*		4061	94		
2.7.1.2.?	*Flavobacterium meningosepticum*		4434	94		
2.7.1.2.?	*Flavobacterium odoratum type I*		4790±861	94		
2.7.2.8	*Rhodothermus marinus*	Rmar(35-1)	3450	155	1	5
	2.7 Average		4104±1065			
2.8	**Green Sulfur Bacteria**					
2.8	*Chlorobium tepidum*		2100	159		
2.10	**Planctomyces and Relatives**					
2.10.1	*Pirellula marina*				2	131
2.10.1	*Planctomyces limnophilus*	Plim(52-2)	5200	230	2	230
2.10.2	*Chlamydia psittaci*		1227±316	77,94		
2.10.2	*Chlamydia trachomatis*	Ctra(11-2)	1048±328	21,77,94,221	2	21
	2.10 Average		2492±2347			
2.11	**Cyanobacteria**					
2.11.1.1	*Oscillatoria sp.*		5672±264	94,95		
2.11.1.2	*Dermocarpa sp.*		5844±272	94,95		
2.11.1.2	*Dermocarpella sp.*		5217±243	94,95		
2.11.1.2	*Gloeocapsa sp.*		5010±238	94,95		
2.11.1.2	*Gloeothece sp.*		8022±373	94,95		
2.11.1.2	*Myxosarcina sp.*		5530±257	94,95		
2.11.1.2	*Pleurocapsa sp.*		5139±238	94,95		
2.11.1.2	*Prochloron sp.*		5809	94,95		
2.11.1.2	*Spirulina sp.*		3964±185	94,95		
2.11.1.2	*Synechocystis sp. strain PCC 6803*	Syns(36-2)	3573470'	107	2	107
2.11.1.3	*Anabaena sp.*	Aspe(58-2)*	5793±458	6,94,95	2	6
2.11.1.3	*Calothrix sp.*		8956±537	94,95		
2.11.1.3	*Chlorogloeopsis sp.*		7395±344	94,95		
2.11.1.3	*Chroococcidiopsis sp.*		6101±1083	94,95		

2.11.1.3	*Cylindrospermum sp.*		9290±431	94,95		
2.11.1.3	*Fischerella sp.*		6639±314	94,95		
2.11.1.3	*Nodularia sp.*		5233±243	94,95		
2.11.1.3	*Nostoc sp.*		7973±373	94,95		
2.11.1.3	*Scytonema sp.*		11593±539	94,95		
2.11.1.4	*Chamaesiphon sp.*		5793±275	94,95		
2.11.1.5	*Synechococcus sp.*	Sysp(36-2)	3547±936	48,94,95,106	2	48
2.11.1.?	*Xenococcus sp.*		6251±291	94,95		
2.11.2.3	*Pseudanabaena sp.*		4778±222	94,95		
2.11.3	*Gloeobacter violaceus*		4353	94		
	2.11 Average		6149±1915			
2.13	**Spirochaetes and Relatives**					
2.13.1	*Serpulina hydoysenteriae*	Shyd(32-1)	3200	206,244	1	244
2.13.2.2	*Treponema denticola*		3000	222		
2.13.2.2	*Treponema pallidum* subsp. *pallidum*	Tpal(11-2)	1137961'	222	2	217
2.13.2.4	*Borrelia afzelli*	Bafz(10-1)	948	165	1	159
2.13.2.4	*Borrelia burgdorferi*	Bbur(15-1)	1463725'	75	1	75,81,167, 196,240
2.13.2.4	*Borrelia garinii*	Bgar(10-1)	953	165	1	165
2.13.3.1	*Leptospira borgpetersenii*		4900	242		
2.13.3.1	*Leptospira interrogans*	Lint(46-2)	4590±119	7,242,243	1 or 2	7,242
	2.13 Average		2524±1630			
2.14	**Purple**					
2.14.1	**Alpha Subdivision**					
2.14.1.1.4	*Acetobacter aceti*		4061	94		
2.14.1.1.?	*Acetobacter rancens*		2929	94		
2.14.1.1.4	*Acetobacter xylinum*		2200	60		
2.14.1.1.4	*Gluconobacter oxydans*		2540	94		
2.14.1.2.4	*Rickettsia bellii*		1660	192		

2.14.1.2.4	*Rickettsia helvetica*		1397	192		
2.14.1.2.4	*Rickettsia massiliae*		1370	192		
2.14.1.2.?	*Rickettsia melolonthae*		1720	77		
2.14.1.2.4	*Rickettsia prowazekii*	Rpro(13-1)	1329±391	66,94,221	1	170,171
2.14.1.2.?	*Rickettsia rickettsia*		1634	94		
2.14.1.2.4	*Rickettsia typhi*		1441±435	66,94		
2.14.1.2.?	*Rickettsiella grylli*		2100	77		
2.14.1.2.7	*Wolbachia pipientis*				1	16
2.14.1.2.7	endosymbiont of *Sitophilus oryzae*		3000	45		
2.14.1.3	*Paracoccus denitrificans*		4369	94		
2.14.1.3	*Rhodobacter capsulatus*	Rcap(37-4)	3715	70,71	4	70,71
2.14.1.3	*Rhodobacter sphaeroides*	Rsph(42-3)	4180±311	71,209,210	3	209
2.14.1.6	*Zymomonas mobilis*		2023	94		
2.14.1.8	*Caulobacter crescentus*		3911±130	6,65,94,221		
2.14.1.9.1	*Bradyrhizobium japonicum*	Bjap(87-1)*	8700	119,120	1	119
2.14.1.9.5	*Hyphomicrobium sp.*		4045	94		
2.14.1.9.5	*Rhodomicrobium vannielii*		3398	94		
2.14.1.9.?	*Thiobacillus acidophillus*		3400	63		
2.14.1.9.?	*Thiobacillus cuprinus*		3800	148		
2.14.1.9.6	*Thiobacillus novellus*		3500	63		
2.14.1.9.8	*Agrobacterium sp.*		5502	94		
2.14.1.9.8	*Rhizobium galegae*	Rgal(59-3)	5892	102	3	102
2.14.1.9.?	*Rhizobium trifollii*		5081	94		
2.14.1.9.8	*Sinorhizobium meliloti*	Smel(59-3)	5928±809	94,101	3	101
2.14.1.9.9	*Bartonella bacilliformis*	Bbac(16-2)	1565±31	117,118	2	117
2.14.1.9.10	*Brucella abortus*		2600	2		
2.14.1.9.10	*Brucella melitensis*	Bmel(29-3)	2925±460	2,151	3	2
	2.14.1 Average		3288±1694			
2.14.2	**Beta Subdivision**					
2.14.2.1	*Chromobacterium violaceum*		7598±354	94,95		
2.14.2.?	*Neisseria catarrhalis*		1576	95		
2.14.2.?	*Neisseria crassa*		2710±126	94,95		
2.14.2.1	*Neisseria gonorrhoeae*		2001±197	61,94,95,202,221		
2.14.2.1	*Neisseria meningitidis*	Nmen(22-4)*	2188±212	59,78,94,221	4	59
2.14.2.1	*Neisseria sicca*		2346	94		

2.14.2.?	*Neisseria subflava*		2346	94		
2.14.2.?	*Alcaligenes eutrophus*		7443	94		
2.14.2.2.3	*Burkholderia cepacia*		8125±35	103,186		
2.14.2.2.5	*Alcaligenes faecalis (odorans)*		3285	94		
2.14.2.2.5	*Bordetella pertussis*		3700	208		
2.14.2.2.5	*Taylorella equigenitalis*		1682	149		
2.14.2.2.6	*Variovorax (Alcaligenes) paradoxus*		7686	94		
2.14.2.6	*Nitrosococcus sp.*		3317	94		
2.14.2.6	*Nitrosomonas sp.*		2265	94		
	2.14.2 Average		3885±2463			
2.14.3	**Gamma Subdivision**					
2.14.3.?	*Xanthomonas axonopodis*		4500	88		
2.14.3.4	*Xanthomonas campestris*	Xcam(47-2)	4725	52	2	132,133
2.14.3.?	*Xanthomonas pelargonii*		4790	94		
2.14.3.7.1	*Coxiella burnetii*	Cbur(17-1)	1683	94	1	1
2.14.3.7.2	*Legionella pneumophila*		4073±39	94,221		
2.14.3.10.1	*Acinetobacter calcoaceticus*		2654	94		
2.14.3.10.1	*Moraxella (Branhamella) catarrhalis*		1886±262	67,94		
2.14.3.10.1	*Moraxella osloensis*		2346	94		
2.14.3.10.?	*Azomonas macrocytogenes*	Amac(24-6/9)*	2350	144	6 to 9	144
2.14.3.10.?	*Azotobacter agilis*		2832	94		
2.14.3.10.?	*Azotobacter chroococcum*	Achr(26-6/9)*	2642±878	94,144	6 to 9	144
2.14.3.10.?	*Azotobacter paspali*	Apas(42-6/9)*	4180	144	6 to 9	144
2.14.3.10.?	*Azotobacter vinelandii*	Avin(42-6/9)*	4237±500	144	6 to 9	144
2.14.3.10.3	*Pseudomonas aeruginosa*	Paer(36-4)	3639±1637	13,53,87,95, 181,187,188, 195,222	4	91,195
2.14.3.10.?	*Pseudomonas cepacia*		7000	49		
2.14.3.10.?	*Pseudomonas facilis*		4531	94		
2.14.3.10.?	*Pseudomonas flava*		5016	94		
2.14.3.10.3	*Pseudomonas fluorescens*		4921±1453	53,94,95,99		
2.14.3.10.?	*Pseudomonas oleovorans*		6121	95		
2.14.3.10.?	*Pseudomonas palleronii*		4531	94		
2.14.3.10.?	*Pseudomonas piscicida*		6375	94		
2.14.3.10.?	*Pseudomonas pseudoflava*		6553	94		

2.14.3.10.3	*Pseudomonas putida*		4114±801	94,98		
2.14.3.10.?	*Pseudomonas putrefaciens*		5453	94		
2.14.3.10.?	*Pseudomonas rubescens*		5210	94		
2.14.3.10.?	*Pseudomonas saccharophila*		5663	94		
2.14.3.10.?	*Pseudomonas solanacearum*		5242±433	94,98		
2.14.3.10.3	*Pseudomonas stutzeri*	Pstu(41-4)	4137±769	53,84,94,95,180	4	84
2.14.3.10.?	*Pseudomonas trifolii*		6100	94		
2.14.3.12	*Shewanella putrefaciens*		4500	222		
2.14.3.13.1	*Salinivibrio (Vibrio) costicola*		2382	150		
2.14.3.13.2	*Vibrio cholerae*	Vcho(28-7)	2843±321	51,146,221	7	146
2.14.3.13.2	*Vibrio harveyi*				1	123
2.14.3.13.3	*Vibrio metschnicovii*		3540±165	94,95		
2.14.3.15.1	*Buchnera aphidicola*				1	157,189
2.14.3.15.2	*Escherichia coli K-12*	Ecol(46-7)	4639221'	22	7	22
2.14.3.15.?	*Salmonella enteritidis*	Sent(48-7)	4746	136	7	136
2.14.3.15.2	*Salmonella paratyphi*	Spar(47-7)*	4718±96	136,137,140	7	136,140
2.14.3.15.2	*Salmonella pullorum*		4442±217	94,95		
2.14.3.15.2	*Salmonella typhi*	Styi(47-7)*	4727±254	136,194,221	7	136
2.14.3.15.2	*Salmonella typhimurium*	Styp(46-7)*	4605±319	136,138,139,202, 235	7	136,235
2.14.3.15.?	*Citrobacter amalonaticus*		4595	94		
2.14.3.15.?	*Citrobacter diversus*		4693	94		
2.14.3.15.3	*Citrobacter freundii*		4288	94		
2.14.3.15.3	*Erwinia herbicola*		4935	94		
2.14.3.15.3	*Erwinia uredovora*		5275	94		
2.14.3.15.3	*Serratia marcescens*		7490±997	94,95		
2.14.3.15.5	*Yersinia pseudotuberculosis*		5875±273	94,95		
2.14.3.15.?	*Klebsiella ozaenae*		3698±172	94,95		
2.14.3.15.?	*Klebsiella pneumoniae*		4142	94		
2.14.3.15.?	*Klebsiella rubiaceurum*		4660	94		
2.14.3.15.?	*Shigella boydii*		3722	94		
2.14.3.15.?	*Shigella dysenteriae*		4466	94		
2.14.3.15.?	*Shigella flexneri*		4142	94		
2.14.3.15.?	*Shigella sonnei*		3275±152	94,95		
2.14.3.15.?	*Enterobacter aerogenes*		4207	94		
2.14.3.15.?	*Enterobacter cloacae*		4207	94		
2.14.3.15.?	*Proteus morganii*		3061	95		
2.14.3.15.6	*Proteus vulgaris*		3275±152	94,95		
2.14.3.16.1	*Haemophilus parainfluenzae*		2340	108		

2.14.3.16.?	*Haemophilus actinomycetemcomitans*	(*Actinobacillus*)	2200	160,221		
2.14.3.16.2	*Haemophilus aegyptius*		1833±85	94,95		
2.14.3.16.2	*Haemophilus ducreyi*		1785	127		
2.14.3.16.2	*Haemophilus influenzae Rd KW20*	Hinf(18-6)	1830137'	69	6	69
2.14.3.16.3	*Pasteurella multocida*		1770±82	94,95		
	2.14.3 Average		4137±1365			
2.14.4	**Delta Subdivision**					
2.14.4.1	*Desulfovibrio desulfuricans*		2600±707	62,232		
2.14.4.1	*Desulfovibrio gigas*		1764	94		
2.14.4.1	*Desulfovibrio vulgaris*		2723±1241	62,94		
2.14.4.?	*Bdellovibrio bacteriovorans*		2092±130	32,94		
2.14.4.?	*Bdellovibrio starrii*		2663±125	94		
2.14.4.?	*Bdellovibrio stolpii*		2341±100	32,94		
2.14.4.4	*Bdellovibrio W*		2136	89		
2.14.4.5	*Desulfobulbus propionicus*		3700	62		
2.14.4.6	*Myxococcus xanthus*		9452±2	46,47,93		
	2.14.4 Average		3275±2380			
2.14.5	**Epsilon Subdivision**					
2.14.5.1	*Helicobacter mustelae*		1700	218		
2.14.5.1	*Helicobacter pylori 26695*	Hpyl(17-2)	1667867'	224	2	224
2.14.5.2	*Campylobacter coli*	Ccol(19-3)	1926±585	42,94,219,237	3	219
2.14.5.2	*Campylobacter fetus*	Cfet(18-3)	1785±992	42,94,193	3	193
2.14.5.2	*Campylobacter helveticus*				2	135
2.14.5.2	*Campylobacter jejuni*	Cjej(21-3)	2067±765	42,94,112,162, 164	3	112,162
2.14.5.2	*Campylobacter laridis*		1451	42		
2.14.5.2	*Campylobacter upsaliensis*	Cups(20-3)	2000	31	3	205
	2.14.5 Subdivision Average		1800±215			

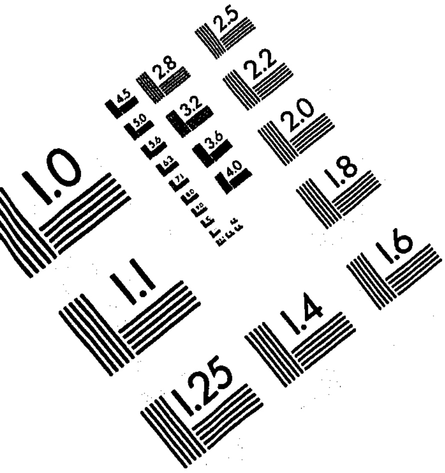
	2.14 Average		3277±907			
2.16	**Gram-positive**					
2.16.1	**High G+C Subdivision (39)**					
2.16.1.4	*Streptomyces ambofaciens*				6	20
2.16.1.4	*Streptomyces coelicolor*	Scoe(98-6)	9801±2547	94,109	6	14,227
2.16.1.4	*Streptomyces griseus*	Sgri(78-6)	7800	130	6	110
2.16.1.4	*Streptomyces lividans*	Sliv(80-6)	8000	94	6	126,212
2.16.1.4	*Streptomyces rimosus*		10599	94		
2.16.1.6.2	*Bifidobacterium breve*	Bbre(21-3)	2100	30	3	30
2.16.1.7.?	*Micrococcus flavus*		4061	95		
2.16.1.7.3	*Micrococcus luteus*		4466	94		
2.16.1.7.?	*Micrococcus lysodeikticus*		4273	95		
2.16.1.7.3	*Micrococcus sp Y-1*		4061	172		
2.16.1.12.1.1	*Nocardia asteroides*		4531	94		
2.16.1.12.1.?	*Nocardia caviae*		5178	94		
2.16.1.12.1.1	*Nocardia corynebacteroides*		3560	94		
2.16.1.12.1.?	*Brevibacterium ammoniagenes*		3000	95		
2.16.1.12.1.?	*Corynebacterium (Brevebacterium) ammoniagenes*		3204	94		
2.16.1.12.1.2	*Corynebacterium diphtheriae*		1942	94		
2.16.1.12.1.2	*Corynebacterium glutamicum*		2998±120	10,94		
2.16.1.12.1.?	*Corynebacterium (Brevebacterium) liquefaciens*		2913	94		
2.16.1.12.1.?	*Corynebacterium minutissimum*		2265	94		
2.16.1.12.1.2	*Corynebacterium renale*		1942	94		
2.16.1.12.1.?	*Corynebacterium (Brevebacterium) vitarumen*		1942	94		
2.16.1.12.2	*Mycobacterium avium*		5838±988	94,222		
2.16.1.12.2	*Mycobacterium bovis*	Mbov(51-2)	5065	94	2	213
2.16.1.12.2	*Mycobacterium chelonei*		4045	94		
2.16.1.12.2	*Mycobacterium farcinogenes*		7023	94		
2.16.1.12.2	*Mycobacterium fortuitum*		5000	111		
2.16.1.12.2	*Mycobacterium gastri*		6796	94		
2.16.1.12.2	*Mycobacterium gordonae*		7395	94,111		
2.16.1.12.2	*Mycobacterium intracellulare*	Mint(50-1)	5016	111	1	19
2.16.1.12.2	*Mycobacterium kansasii*		6197	94		
2.16.1.12.2	*Mycobacterium leprae*		2104	94,175		

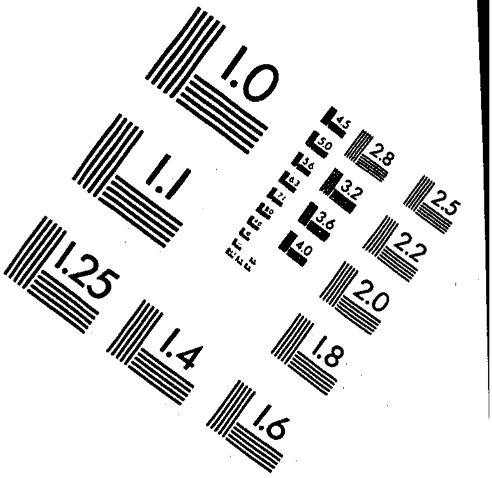
2.16.1.12.2	*Mycobacterium lepraemurium*	Mlep(29-1)*	2913	94	1	211
2.16.1.12.2	*Mycobacterium marinum*		5825	94		
2.16.1.12.2	*Mycobacterium phlei*	Mphl(63-2)	6311	94	2	19
2.16.1.12.2	*Mycobacterium scrofulaceum*		5639	94		
2.16.1.12.2	*Mycobacterium smegmatis*	Msme(72-2)	7168	94	2	19
2.16.1.12.2	*Mycobacterium stercoides*		6149	94		
2.16.1.12.2	*Mycobacterium tuberculosis*	Mtub(43-1)	4323±153	94,111,173,175,222	1	19
2.16.1.12.2	*Mycobacterium vaccae*		4045	94		
2.16.1.12.2	*Mycobacterium xenopi*		5324	94		
	2.16.1 average		6734±3276			
2.16.2	**Clostridia and Relatives**					
2.16.2.1.1	*Clostridium stercorarium*		3000	239		
2.16.2.1.1	*Clostridium thermocellum*		3500	239		
2.16.2.2	*Caldocellum saccharolyticum*		2780	26		
2.16.2.4.1	*Clostridium botulinum*		4039	134		
2.16.2.4.2	*Clostridium pasteurianum*		3900	239		
2.16.2.4.2	*Clostridium tyrobutyricum*		2500	239		
2.16.2.4.6	*Clostridium acetobutylicum*		4800±1476	222,233,239		
2.16.2.4.7	*Clostridium perfringens*	Cper(35-10)	3505±182	35,36,239	9 or 10	35,79
2.16.2.5	*Clostridium difficile*		3200	239		
	2.16.2 average		3469±706			
2.16.4	**Mycoplasmas and Relatives**					
2.16.4.1	*Mycoplasma arginini*				2	3
2.16.4.1	*Mycoplasma flocculare*		890	185		
2.16.4.1	*Mycoplasma hominis*	Mhom(7-2)	733±27	122,185	2	122
2.16.4.1	*Mycoplasma hyopneumoniae*	Mhyo(11-1)	1105±50	178,185	1	217
2.16.4.1	*Mycoplasma mobile*		780	11		
2.16.4.?	*Mycoplasma sturni*		870	72		
2.16.4.1	*Mycoplasma synoviae*		900	178		

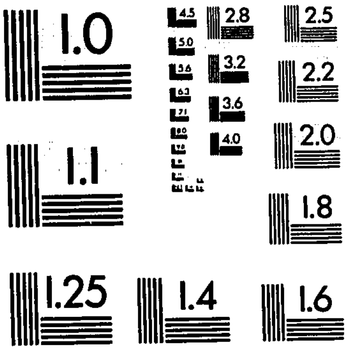
2.16.4.2	*Mycoplasma gallisepticum*	Mgal(11-2)	1052±3	86,178	2	3
2.16.4.2	*Mycoplasma genitalium G-37*	Mgen(6-1)	580070'	76	1	76
2.16.4.2	*Mycoplasma iowae*		1280	178		
2.16.4.2	*Mycoplasma penetrans*		1358	184		
2.16.4.2	*Mycoplasma pneumoniae M129*	Mpne(8-1)	816397'	97	1	97
2.16.4.2	*Ureaplasma urealyticum*	Uure(9-2)	858±99	19,54,85,183, 221	2	3,54
2.16.4.3	*Mycoplasma capricolum*	Mcap(9-2)	940±305	147,153	1 or 2	3,4
2.16.4.3	*Mycoplasma mycoides*	Mmyc(13-2)	1261±56	178,179,185	2	179
2.16.4.3	*Mycoplasma sp. PG-50*	Mspe(10-2)	1040	179	2	179
2.16.4.3	*Spiroplasma citri*	Scit(17-1)	1707±103	94,238	1	238
2.16.4.?	*Spiroplasma platyhelix*		780	234		
2.16.4.?	*Spiroplasma velocicrescens*		1480	116		
2.16.4.4	*Acholeplasma axanthum*				2	3
2.16.4.?	*Acholeplasma granularum*				2	3
2.16.4.4	*Acholeplasma laidlawii*		1649±57	185		
2.16.4.4	*Acholeplasma oculi*		1633	223		
	2.16.4 average		1086±335			
2.16.5	**Bacillus-Lactobacillus-Streptococcus Subdivision**					
2.16.5.?	*Lactococcus cremoris*		2600	29		
2.16.5.1	*Lactococcus lactis*	Llac(32-6)	3183±1481	28,29,58,94, 124,216,226	6	125
2.16.5.1	*Streptococcus agalactiae*	Saga(27-6)	2689±1232	94,95	6	50
2.16.5.1	*Streptococcus bovis*		5631	94		
2.16.5.1	*Streptococcus dysgalactiae*	Sdys(39-6)	3883	94	6	18
2.16.5.?	*Streptococcus faecium*		8091	87		
2.16.5.1	*Streptococcus gordonii*	Sgor(21-4)	2120	73,74	4	74
2.16.5.1	*Streptococcus mutans*		2145±28	89,166		
2.16.5.1	*Streptococcus pneumoniae*	Spne(23-6)	2267±74	80,94,221	6	80
2.16.5.1	*Streptococcus pyogenes*		1986±66	94,95,221		
2.16.5.?	*Streptococcus raffinolactis*		4126	94		
2.16.5.1	*Streptococcus sanguis*		2300	29		
2.16.5.1	*Streptococcus thermophilus*	Sthe(19-6)	1943±440	29,191	6	18
2.16.5.?	*Streptococcus uberis*		3398	94		
2.16.5.2.1	*Leuconostoc oenos*		1684±316	57,177,220		

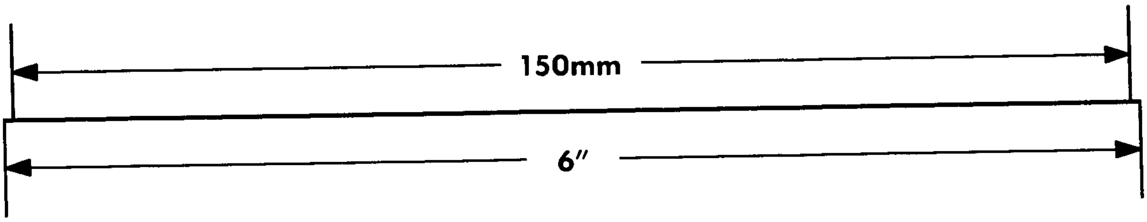
2.16.5.2.2	*Lactobacillus acidophilus*	Laci(19-4)	1850	190	4	190
2.16.5.2.2	*Lactobacillus delbrueckii*		2300	128		
2.16.5.2.2	*Lactobacillus gasseri*		2020	185		
2.16.5.2.2	*Lactobacillus helveticus*		1925	143		
2.16.5.2.3	*Lactobacillus casei*		2071	94		
2.16.5.2.?	*Lactobacillus cremoris*		2600	29		
2.16.5.2.?	*Lactobacillus lactis*		2460±89	29		
2.16.5.2.3	*Lactobacillus plantarum*		2259±769	56,94		
2.16.5.2.3	*Pediococcus acidilactici*		1560	56		
2.16.5.2.3	*Pediococcus pentosaceus*		1200	56		
2.16.5.3.1	*Enterococcus faecalis*	Efae(29-4)	2875±555	29,94,95,221	4	158
2.16.5.3.1	*Enterococcus hirae*				6	197,241
2.16.5.4	*Carnobacterium divergens*		3200	56		
2.16.5.6	*Listeria monocytogenes*	Lmon(32-6)	3150	152	6	152
2.16.5.8.1	*Caryophanon latum*		1861	94		
2.16.5.8.1	*Caryophanon tenue*		1570	94		
2.16.5.9	*Bacillus anthracis*		4355±203	94,95		
2.16.5.9	*Bacillus cereus*	Bcer(50-10/12)	4959±1213	37,41,94,105, 115	10 to 12	99
2.16.5.?	*Bacillus polymyxa*		4309±200	94,129		
2.16.5.9	*Bacillus thuringiensis*		5533±1305	38,39,40		
2.16.5.10.?	*Staphylococcus albus*		1697	95		
2.16.5.10.2	*Staphylococcus aureus*		2583±365	25,94,95,173, 174,231		
2.16.5.11	*Bacillus subtilis*	Bsub(42-10)	4214810'	121	10	121
2.16.5.12	*Bacillus megaterium*		4584±121	94,228		
	2.16.5 average		2978±1432			
	2.16 average		3567±2348			

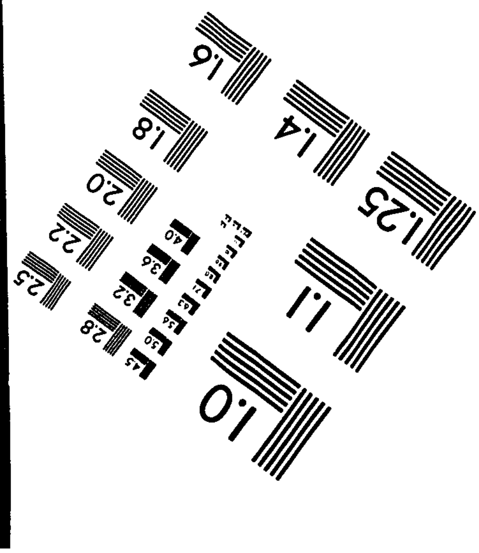
IMAGE EVALUATION
TEST TARGET (QA-3)











APPLIED IMAGE . Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989