The development of modern biology has been motivated by a reductionist overview which is present even when not explicit. (Crick 1966) The understanding by most researchers is that the activities of organisms must be understood in the functioning of tissues and cells. Cells are to be understood in terms of the activities of organelles and macromolecules. Macromolecular function depends on the properties of monomers, the nature of polymerization and the details of the subsequent macromolecular folding. Underlying the understanding of all molecules both large and small is a basic doctrine of physics and chemistry which establishes our most fundamental understanding of the universe.

Our ability to comprehend life depends, within this context, on a complete molecular understanding of a living organism. The question "What is life?" is to be answered by a complete description of structure and function of a cell at the most reduced atomic and molecular level. Motivated by this kind of reasoning, a number of investigators began in the late 1950s and early 1960s to ask the question as to the smallest and simplest living cell. Since many small cells such as the rickettsia and chlamydia are obligate intercellular parasties using part of the host cell’s metabolic and genetic machinery, the search was sharpened to seek the smallest autonomous cell, one capable of replicating in nutrient medium devoid of any other organisms.

The criteria of life employed were cellularity and the ability to replicate such that a single cell could give rise to a clone. The presence of hereditary material and mutability were also considered necessary properties. The search focused on the smallest and simplest unicellular organisms that displayed the features outlined above.

Smallness was clearly a matter of cell size (volume) or cell mass. Simplicity was eventually defined in terms of genome size using the postulate that all the information necessary for a given species is present in its full genome. Genome size can be given in total mass of genetic DNA or more conveniently now-days in terms of the number of DNA base pairs in the double stranded chromosome.

Note that the search for smallness and simplicity is open ended for there may always be organisms as yet undiscovered which fit our criteria of life and are smaller and simpler than
those currently under study. Within these restrictions, the organisms of smallest size and smallest genome are all members of the genus *Mycoplasma*.

According to Bergeys Manual of Systematic Bacteriology Vol. 1 (Williams and Wilkins 1984) these organisms have a genome size of $5 \times 10^8$ daltons (800 kilobases). The smallest viable cells are $0.3\mu m$ in diameter and the cells are build of three organelles only: the cell membrane, the ribosome, and the characteristic procaryotic chromosome. The $G+C$ (guanine + cytosine) of the DNA of various strains range down to 24%.

The relation between size and atomicity is worth some brief elaboration. A sphere $0.3\mu m$ in diameter has a volume of $1.4 \times 10^{-14} cm^3$ and contains only a billion atoms. Most of this is water so the nonaqueous portion has only about 250 million atoms. At a pH of 7 a cell of this size will on the average contain less than one hydrogen ion. The lower limit of cell size embodies certain features based on the atomicity of matter.

A genome of 800 kilobases can code for a sequence of 266,000 amino acids – or about 400 different kinds of proteins. Thus a cell with all of the properties we usually ascribe to life at the cellular and molecular level is limited in terms of its complexity and lends itself to the possibility of complete computer modelling, not atom by atom, but in terms of a highly structured chemical kinetics. It is clear that if we are to understand higher forms of life at this strict reductionist level we must first understand the mycoplasma. Attempts at the complete modeling of this simplest system should provide the propaedeutic basis for moving on to higher organisms.

We consider next whether a cell could be much smaller or simpler than the mycoplasma. Here we rely on certain universal biochemical and molecular features that have characterized all cells that have been investigated. These include: presence of a lipoprotein membrane primarily structured of bilayers of amphiphiles, use of the central core of the universal metabolic chart for energy processing, and synthesis of informationally necessary small molecules, and a common method of macromolecule synthesis employing messenger RNA, tRNA, ribosomes, and a variety of synthetases as well as the energy supplying molecules to drive the synthesis. These housekeeping functions of a cell require at a minimum a genome at least half the size of the mycoplasma and perhaps larger. We are approaching the simplicity limit for life using the conventional chemistry that characterizes molecular biology.

A size limit of $0.3\mu m$ also seems near the lower limit. Again consider a cell of diameter $0.15\mu m$. The nonaqueous portion has about 31 million atoms, which is the order of 2,400,000 small molecules, or around 4000 total macromolecules. At that point the available hardware to perform the necessary functions tends to become limiting. A cell of $0.15\mu m$ could consist of a couple of ribosomes and a small loop of DNA surrounded by a membrane. There isn’t room for much more. Thus the theoretical minimal cells and the mycoplasma seem quite close in size and complexity. By working with this genus we appear to be studying the most reduced system to exhibit the properties of life. As such it represents an ideal object of study to test our understanding of what is life from a point of view of molecular biology.

An analogy from physics may prove helpful. The complexities of spectroscopy were explored by an intense study of the spectrum of hydrogen, the formulations of the Bohr theory and the subsequent development of quantum mechanics. The theory was originally worked out for hydrogen, the simplest of all atoms. The theory of more complex atoms and molecules is
still based on the use of hydrogen wavefunctions. I want to suggest the possibility that the understanding of higher living forms compared to the understanding of mycoplasma may be analogous to the way in which more complex structures are understood in terms of our comprehension of the hydrogen atom.

The genus *Mycoplasma* is in the order Mycoplasmataceae in the class Mollicutes. The cells lack a cell wall and are surrounded by a plasma membrane. They are facultatively anaerobic and possess a truncated flavin-terminated electron transport chain devoid of quinones and cytochromes. Two species have been grown on chemically defined medium, Mycoplasma mycoides and its close relative Mycoplasma capricolum (Rodwell 1983).

Mycoplasma capricolum (Type strain ATC 27343). Pathogenic for goats, DNA ratio of G + C is 24.1 to 25.5%.

Mycoplasma mycoides var. mycoides (Type strain NCTC 10114). DNA ratio of G + C is 26.1 to 27.1%. Etiological agent of bovine pleuropneumonia.

Experiments could be undertaken to attempt to produce a more reduced genome than that found in the naturally occurring mycoplasma. Cells could be irradiated with ionizing radiation and plated out on very rich medium. Colonies could be picked and cloned, the DNA could then be purified, digested with restriction enzymes and analyzed by electrophoresis. Clones with deletions could be subjected to further irradiation until minimal genomes are obtained.

The G + C ratio of Mycoplasma capricolum and Mycoplasma mycoides provides some further considerations. In order to code for consensus protein using the most AU rich codon for each organism, we require coding DNA of a G + C content of 32.7%. This is shown in the analysis in Table I. This is very slightly modified in the AT rich mycoplasma by the use of UGA as a Tryptophan codon rather than as a stop codon (Muto et al 1986). The only simple way to account for the 25% ratio is the presence of high AT noncoding DNA. Even if this were pure AT it would have to account for 25% of the genome to yield the proper base ratios. This would considerably reduce the metabolic complexity of the organism. Alternatively, the proteins of Mycoplasma capricolum might have a very different amino acid composition when compared to the consensus value. Kawauchi et al (1982) reported a high number of basic proteins in M. capricolum. A predominance of lysine coded as AAA could lead to high AT. High phenylalanine, leucine, and isoleucine could also weight the distribution toward AT. At present the question of base ratio and amino acid composition remains unresolved.

Although only limited metabolic studies have been carried out on Mycoplasma capricolum, we are in a position to infer considerable information about the metabolism from the composition of the defined growth medium and the universality of the metabolic chart. The chart of intermediate metabolism appears to be universal throughout the biological domain so that the metabolic chart of any species is a subset of the master chart. All studies of mycoplasma indicate that they are perfectly conventional in terms of metabolic pathways.

A defined medium for Mycoplasma capricolum has been devised by Rodwell (1983) and is shown in Table II.

Kawauchi et al (1972) have analyzed whole cell proteins and ribosomal proteins of M.
capricolum using two dimensional polyacrylamide gel electrophoresis. The gels showed 355 spots, 202 in the acidic range and 153 in the basic range. In addition, 50 ribosomal proteins resembled those of bacteria.

Based on the information that we have been discussing, M. capricolum or some closely allied species now presents us with the opportunity of carrying out a series of experiments and analyses to test the completeness of molecular biology. After indicating the experiments and analysis we will discuss the advantages of mycoplasma for this type of study.

The mycoplasma genome is approximately 800,000 base pairs in length. Since each base pair contributes 2 bits to the overall information, the genetic message is $1.6 \times 10^6$ bits. About 3% of the genome has been sequenced by Muto and coworkers (1987) and this group is currently developing a restriction map. Given current technology, sequencing the entire genome could be carried out within a period of one year.

Sequence of the entire genome acquires meaning in so far as the genome can be indexed to gene products and the function of these products determined. (Anderson 1987) The primary gene products are tRNAs, rRNAs, and mRNAs. The secondary products are all the structured and functional proteins of the cell. Mycoplasma indexing can be carried out in the following way:

1. Sequence the entire genome. This presumably will represent the total hereditary information as no other genetic elements have been reported.

2. Grow up a large culture of cells. Break the cells open by osmotic lysis (or other techniques) and prepare three fractions: membranes, ribosomes, soluble proteins. There is a possible fourth fraction of DNA associated proteins. After appropriate treatment subject each fraction to electrophoretic analysis giving rise to approximately 400 spots. Each spot is indexed with respect to cell fraction, approximate size, and acid base dissociation behavior.

3. Extract the material from each spot and use N terminal analysis to sequence the first 6 terminal amino acids.

4. Translate the genome into an amino acid sequence and locate each indexed protein spot on the genome. Using information on size as well as stop codons, read out the amino acid sequence of each of the 400 spots. There may be some uncertainty in termination of sequences and size of spacer regions.

5. Compare the sequence of each spot with all known amino acid sequences and identify as many functions as possible. Computer programs are available for this analysis.

6. Index as many other spots as possible, using:
   - A list of necessary enzymes from metabolic considerations
   - Tests of enzymatic function. From an experimental point of view this is a very well understood but difficult part of the program. The unknown spots will have to first be purified under non-denaturing conditions and then tested for enzymatic function in a variety of conditions. Dennis Pollack has a program at Ohio State University devoted to a study of the metabolic enzymes of various mycoplasma.
7. Test all possible metabolic feedback control loops and all transport functions.

8. Formulate the total functional map of the organism including all transport steps, all metabolic steps, and all control steps. Investigate any missing information suggested by the preceding mapping.

By successive application of the preceding, it should be possible to: formulate a complete mapping of the genome into coding and noncoding regions, identify the proteins associated with each gene, determine the complete functional graph of the organism.

Owing to the small size of the organism, the analysis should be quite complete. Consider again a cell with $250 \times 10^6$ atoms. This corresponds to about 30,000 total macromolecules or an average of 75 macromolecules of each kind per cell. A protein present at only a few molecules per cell should give a detectable spot on a two dimensional pattern where the average spot corresponds to a cellular content of molecules. Indeed for a variety of reasons it seems likely that all or almost all enzymes are usually expressed in the small genome mycoplasma. The correspondence between the observed and predicted number of protein spots is an argument in favor of this view.

Once the genome and metabolic pathways have been established one can formulate a computer model of the organism consisting of compartments and fluxes between compartments. A compartment consists of a molecule present within the cell and is characterized by concentrations and locations. The fluxes between compartments are governed by enzyme kinetics modified by control steps. A complete model of this type should represent the entire functioning of the cell. The existence of such a model allows an overall computer analog of the cells and opens up a great deal of theoretical analysis of cellular function.

Using defined medium, labelled components can be added one at a time, intermediates can be followed, and a great deal of kinetic information can be obtained about the cells. Thus a constant interplay can be established between the behavior of the model and the behavior of the system.

Since Mycoplasma capricolum fulfills all of our requirements for being a living organism, a complete theoretical model which exhibited all the behavior of the cell would constitute a reductionist understanding of the essential features of life.

This is not to suggest that all this could happen immediately. It seems likely that the computer model and laboratory experiments on cells will develop inconsistencies and there will be a period of back and forth where the model suggests experiments and experiments modify the model. However this should successively narrow in on our understanding of the essential features of procaryotic life.

One of the motivations for looking for a system where the genome can be completely indexed is to provide guidance for problems that will arise in the sequencing and understanding of large size genomes such as that of Homo sapiens. A simple system that is open to complete analysis should provide a background of useful information and techniques.

In discussing the complete indexing of a genome Escherichia coli and Saccharomyces cervisiae have been suggested. The following will argue for the validity of the mycoplasma project on its own merits quite apart from whether the other species are done in parallel.
1. The mycoplasma genome is 1/6 the size of E coli and 1/20 the size of yeast. At this limiting value almost every coded function must be of direct value to the life of the cell. The yeast and coli will have many proteins difficult to assign function because they are used for the fitness of these organisms in their various niches. The number of unidentified indexed spots will be much smaller in mycoplasma.

2. The cell structure is much simpler in mycoplasma for it has no internal organelles and, lacking a cell wall and periplasmic space, it is less complex than walled procaryotes.

3. It has been argued that no genetic data is available on M. capricolum. While this is true the proposed experiments will rapidly index all proteins and RNAs to the genome and thus provide a complete genetic system.

4. A very major advantage of the mycoplasma will come at the modelling level. A 400 variable system is many orders of magnitude easier to deal with than a 2400 variable system. The restriction of most function to catalysis, and structure to ribosomes and membranes also provides a considerably reduced system. If as suspected almost all proteins are constitutive, a further simplification is provided.

Mycoplasma is thus being suggested as a test case for: completeness of molecular biology, the problems of indexing a genome, and understanding life at the molecular level. Compared to other projects involving complex genomes, the cost of this project should be quite reasonable. Sequencing the genome, indexing the spots, and identifying most of the spot’s function could probably be done in five years at a cost of 2.5 million dollars. The modelling work could proceed at about 100,000 dollars per year. Until the problems are better defined it is difficult to estimate the cost of the laboratory studies of cell physiology, biochemistry, and enzyme kinetics, but these are not expensive lines of research. For a relatively modest budget it will be possible to pursue the very deep biological question of probing the nature of life at the atomic and molecular level.

1 References

- Anderson, Norman, 1987, Personal communication
- Muto, A., Personal communication