42

Mycoplasma capricolum Genome Project

P. M. Gillevet, A. Ally, M. Dolan, E. Hsu, M. S. Purzycki, R. Overbeek, E.E. Selkov, S. Smith, C. Wang, and W. Gilbert

Background

The Mycoplasmas are very small, wall-less bacteria phylogenetically related to gram-positive Eubacteria such as *Bacillus subtilis* (*see also* Chapters 40 and 41). *Mycoplasma capricolum* is an example of one of the smallest of free-living organisms (Ryan and Morowitz, 1969) with a genome estimated to be between 724 kb (Poddar and Maniloff, 1989) and 1.1 megabases. As *M. capricolum* is a parasitic organism with a truncated metabolism and can be grown in a defined medium, much of its truncated physiology has been biochemically defined (Maniloff, 1992). The acquisition of the entire genome sequence of the organism will corroborate these classic biochemical studies and allow the complete elucidation and eventual modeling of its truncated metabolism. Furthermore, the comparative analysis of this metabolic network with larger metabolic networks from organisms such as *Haemophilus influenzae* would open the door to the unprecedented opportunity to begin to analyze the minimal set of fundamental genes involved in the process we call "life." (*see* Chapter 37).

The *Mycoplasma capricolum* genome project originated at Harvard University where we developed a novel DNA sequencing strategy termed Multiplex Genomic Walking. The first two years of the Harvard project were spent on sequencing technology development, as described below. A production line to sequence the genome of *Mycoplasma capricolum* was implemented during the third year of the project and resulted in the generation of over a million raw bases of data that assembled into contigs covering some 250,000 linear bases. The project was successful in developing the technology to directly sequence genomes approaching a million bases in size and defining the standard operating procedures, informatics support, appropriate process control and quality assurance to run an integrated production facility. Lastly, the project defined two technical limitations in the walking process which affected the overall throughput of the project, specifically the oligo failure rate and the autoradiographic signal strength.

Multiplex Genomic Walking

The Multiplex Genomic Walking technique reveals the DNA sequence of the organism directly, essentially by hybridization of a Southern blot (Ohara *et al.*, 1989). To produce these sequencing blots, the genomic DNA of the entire organism is completely digested with restriction enzymes, treated with the chemical sequencing reactions, electrophoresed through a sequencing gel, and transferred to a charged nylon membrane. Each genomic restriction digest is represented on the membrane by a set of sequence lanes. When such a membrane is probed with a labeled oligonucleotide, the resulting autoradiograph displays sequencing patterns in those lanes in which the oligonucleotide has hybridized near a restriction cut. Sequence can be read out, in both directions, from the position of the oligonucleotide, on one strand of the DNA.

Figure 42-1 is a schematic that summarizes the Multiplex Genomic Walking strategy. An oligo probe is selected based on a known starting sequence and hybridized to a membrane bearing the chemically sequenced DNA from genomic restriction enzyme digests. Hybridization of this probe to the membrane reveals a single sequence ladder in each restriction enzyme digest where the probe hybridizes near the end of the restriction fragment. Sequence ladders are read in the 5′ to 3′ direction away from the probe along one strand of the DNA when the probe hybridizes near the 5′ end of a restriction fragment. Conversely, when the probe hybridizes near the 3′-end of a fragment, sequence ladders are read "backwards" in the 3′ to 5′ direction. Thus a single probing produces several "Reads" (*see* Figure 42-1) or sequence fragments which assembled into two clusters, one cluster reading in the 3′ to 5′ direction, the other cluster reading in the 5′ to 3′ direction away from the probe.

A simple majority consensus is generated from these reads and the next oligo probe is chosen from the complementary strand for the subsequent hybridization. Figure 1 illustrates a walk proceeding from the 3′ end of the contig. In this case, the second hybridization will yield a sequence cluster reading in the 3′ to 5′ direction away from the probe that provides coverage over the first step on the opposite strand. Thus as the walk proceeds on a contig, the probing of alternate strands results in multiple coverage on both strands of DNA.

Production Cycle

The production process for Multiplex Genomic Walking depicted in Figure 42-2 involves a repetitive cycle of:

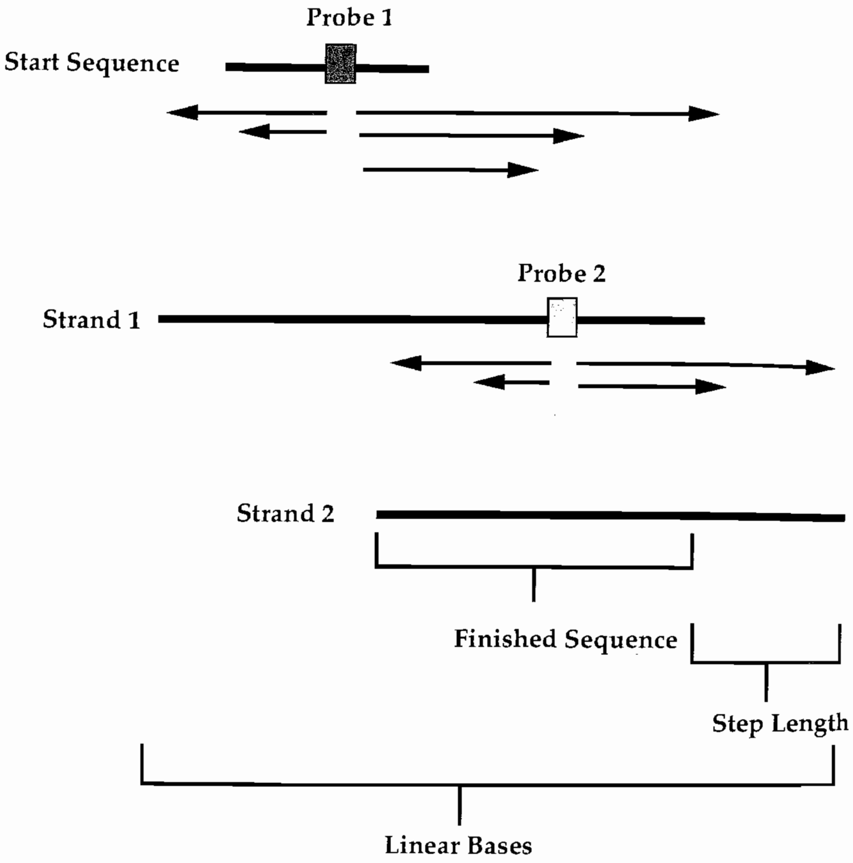


Figure 42-1. Multiplex Genomic Walking Strategy: An oligoprobe is selected based on a known sequence and hybridized to a genomic sequencing membrane. Sequence ladders read in the 5′ to 3′ direction away from the probe along one strand of the DNA when the probe hybridizes next to the 5′ restriction enzyme cut at position. Conversely, when the probe hybridizes near the 3′ restriction enzyme cut at position B, sequence ladders read "backwards" in the 3′ to 5′ direction from the position of the probe. A simple majority consensus is generated from these sequence reads and the next oligo probe is chosen from the complementary strand for the subsequent hybridization. This second hybridization yields sequence reads which extend the consensus in the 5′ to 3′ direction. We define Step Length as the difference in the lengths of Consensus 1 and Consensus 2 sequences. We define Finished sequence as those nucleotide positions in the contig where the majority consensus of both strands agree. It should be noted that the ends of each growing contig are represented by the consensus of only one strand (the region from the distal probe binding site to the end of the contig) and that these positions do not meet the criteria of Finished. Linear sequence is defined as the sum of the lengths of these individual single strand consensus sequences and the lengths of Finished sequences for all contigs.

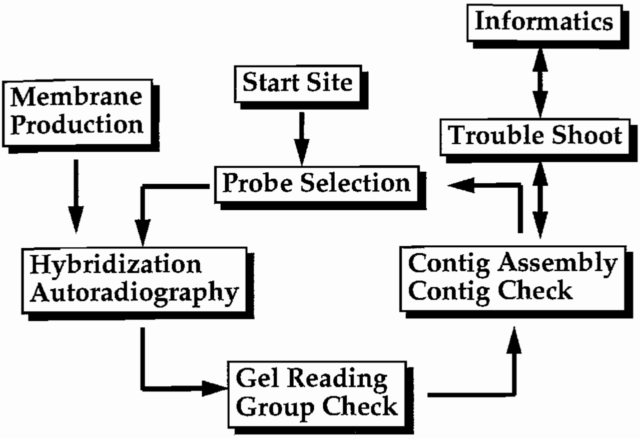


Figure 42-2. The Production Cycle: The repetitive process involved in Multiplex Genomic Walking is shown. It involves the synthesis of an oligonucleotide, the hybridization of that probe onto a membrane, washing and exposure of the membrane, stripping the membrane, reading the sequence, synthesizing a new oligonucleotide, and reprobing the membrane.

- hybridizing an oligo probe to a sequencing membrane,
- manual reading of the autorads and checking the raw data,
- assembling the contig and checking the assembly, and
- picking and synthesizing of new probes to continue the process.

Figure also illustrates the initial production of starting sequences, the membrane generation, the trouble shooting, and quality control functions of the process. The generation of each of these aspects are presented below.

Overall Process Tracking and Control

This component consists of the computational software and hardware to manage the sequence data, the bar coding, and the computerized database for the laboratory. These systems track all of the oligonucleotides, the membranes, and the sequence fragments as well as quality control information (Gillevet, 1993). The Genetic Data Environment (GDE), an X Windows based Graphic User Interface (GUI) was used for the maintenance of our internal database and automated data control systems (Smith *et al.*, 1994). This system allowed the seamless integration of a core multiple sequence editor with pre-existing external sequence analysis programs and newly developed programs into a single prototypic environment. A shell tool with the same expandable menu system as GDE was developed as an interface for technicians to input all tracking information. This model proved extremely flexible and not only allowed direct access to remote procedure calls on high-performance computers, and network-based servers, but was easily modified on a daily basis as the need arose. The latter point was critical in the development of a new technology where the needs of the system cannot predicted *a priori* (see http://uranus.gmu.edu/Mycoplasma/GDE.html).

The entire process was tracked by changing the status field associated with each individual probe. Every aspect of the walking process, from probe picking to data entry had a GUI to control it and allowed information entry in a systematic and consistent manner (Figure 42-3). Tools were developed to query the internal database, to change the status of probes, to assign membranes and probes for hybridization, and to enter data. Our internal database was a set of "tagged" fields that was linked with each of the sequences, whether that sequence corresponded to an oligonucleotide or a sequence fragment. If a field was not used by the program module being run it was not lost but was tracked by each module as an inert link. The concept of a "tagged field" allowed us to freely expand the system as the need arose. We again emphasize that this proved to be a crucial issue, as we constantly refined the walking process in the wet lab, and the informatics system was modified on a day-to-day basis.

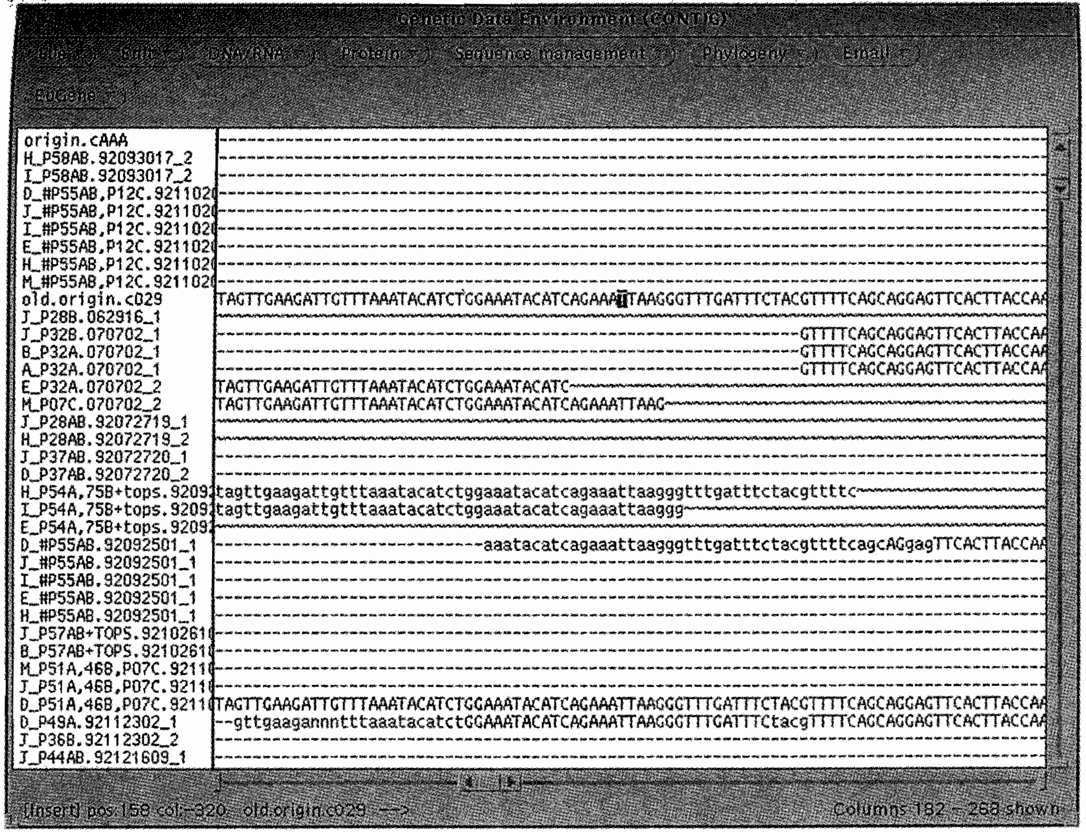
Production of Starting Sequences

Access to about 1,000 randomly arranged sequences around the *Mycoplasma* genome was needed in order to have a continuous supply of starting points to walk from and to statistically cover the entire genome. To develop that set of sequences, we cloned *M. capricolum* DNA into M13 clones and sequenced random clones with an ABI fluorescent sequencer. We sequenced some 1,505 random clones from the organism for random start points which produced 187,309 raw bases of DNA sequence.

Oligonucleotide Production

The Multiplex Genomic Walking strategy is based on oligonucleotide hybridization to sequencing membranes. Therefore, the efficiency of the method is dependent on both the rate and cost of oligonucleotide synthesis. To address these factors, we modified the cycles on the Millipore two-column "Cyclone" synthesizer to simplify and shorten the cycle time. These machines originally had an eight-minute cycle time in a two-column mode (producing about 12 oligos per 8 hour day). We eliminated the capping steps, shortened each chemical step, used TCA to speed the detritylation, and increased the gas pressure to increase the flow rate. We developed a 90-second cycle time, for the two-column mode, and thus were able to synthesize two 20-mers in 30 minutes.

The two-column Cyclone synthesizers were fitted with a RS232 serial board and connected to the Sun computer system via a serial line so that the programs and the oligonucleotide sequences were downloaded directly. The machines were



```
origin.cAAA            ---------------------------------------------------------------------------------------
H_P58AB.92093017_2     ---------------------------------------------------------------------------------------
I_P58AB.92093017_2     ---------------------------------------------------------------------------------------
D_#P55AB,P12C.921102   ---------------------------------------------------------------------------------------
J_#P55AB,P12C.921102   ---------------------------------------------------------------------------------------
I_#P55AB,P12C.921102   ---------------------------------------------------------------------------------------
E_#P55AB,P12C.921102   ---------------------------------------------------------------------------------------
H_#P55AB,P12C.921102   ---------------------------------------------------------------------------------------
M_#P55AB,P12C.921102   ---------------------------------------------------------------------------------------
old.origin.c029        TAGTTGAAGATTGTTTAAATACATCTGGAAATACATCAGAAATTAAGGGTTTGATTTCTACGTTTTCAGCAGGAGTTCACTTACCAA
J_P28B.062916_1        ---------------------------------------------------------------------------------------
J_P32B.070702_1        -------------------------------------------------------------GTTTTCAGCAGGAGTTCACTTACCAA
B_P32A.070702_1        -------------------------------------------------------------GTTTTCAGCAGGAGTTCACTTACCAA
A_P32A.070702_1        -------------------------------------------------------------GTTTTCAGCAGGAGTTCACTTACCAA
E_P32A.070702_2        TAGTTGAAGATTGTTTAAATACATCTGGAAATACATC--------------------------------------------------
M_P07C.070702_2        TAGTTGAAGATTGTTTAAATACATCTGGAAATACATCAGAAATTAAG----------------------------------------
J_P28AB.92072719_1     ---------------------------------------------------------------------------------------
H_P28AB.92072719_2     ---------------------------------------------------------------------------------------
J_P37AB.92072720_1     ---------------------------------------------------------------------------------------
D_P37AB.92072720_2     ---------------------------------------------------------------------------------------
H_P54A,75B+tops.92092  tagttgaagattgtttaaatacatctggaaatacatcagaaattaagggtttgatttctacgttttc--------------------
I_P54A,75B+tops.92092  tagttgaagattgtttaaatacatctggaaatacatcagaaattaaggg--------------------------------------
E_P54A,75B+tops.92092  ---------------------------------------------------------------------------------------
D_#P55AB.92092501_1    ----------------------------aaatacatcagaaattaagggtttgatttctacgttttcagcAGgagTTCACTTACCAA
J_#P55AB.92092501_1    ---------------------------------------------------------------------------------------
I_#P55AB.92092501_1    ---------------------------------------------------------------------------------------
E_#P55AB.92092501_1    ---------------------------------------------------------------------------------------
H_#P55AB.92092501_1    ---------------------------------------------------------------------------------------
J_P57AB+TOPS.9210261   ---------------------------------------------------------------------------------------
B_P57AB+TOPS.9210261   ---------------------------------------------------------------------------------------
M_P51A,46B,P07C.9211   ---------------------------------------------------------------------------------------
J_P51A,46B,P07C.9211   ---------------------------------------------------------------------------------------
D_P51A,46B,P07C.9211   TAGTTGAAGATTGTTTAAATACATCTGGAAATACATCAGAAATTAAGGGTTTGATTTCTACGTTTTCAGCAGGAGTTCACTTACCAA
D_P49A.92112302_1      --gttgaagannntttaaatacatctGGAAATACATCAGAAATTAAGGGTTTGATTTCtacgTTTTCAGCAGGAGTTCACTTACCAA
J_P36B.92112302_2      ---------------------------------------------------------------------------------------
J_P44AB.92121609_1     ---------------------------------------------------------------------------------------

[Insert] pos:158 col:-320  old.origin.c029 -->                                  Columns 182 - 268 shown
```

Figure 42-3. The Genetic Data Environment: GDE was used to track the entire walking process by changing the status field associated with each individual probe. A typical window is depicted showing sequence fragments going in both directions from a probe site (see http://uranus.gmu.edu/Mycoplasma/GDE.html).

controlled by a barcode reader attached to a VT100 terminal allowing automated programming from a file of oligo sequences. Using our short cycle and two modified cyclone machines, 32 to 36 oligos were synthesized per day at an estimated reagent cost of $12.70 for a 20-mer oligo. A total of 3,782 total oligonucleotides were synthesized during the whole project and 2,443 of these were used as walking probes.

Choice of Restriction Enzymes and Chemistries: Mycoplasma capricolum

(California kid) strain ATCC #27343, is a very AT-rich and thus we optimized the restriction enzymes digests to produce appropriate fragment size distributions. As current gel technologies limit read lengths to the range of 400-500 nucleotides, the optimum strategy for Multiplex Genomic Walking is to use mixtures of enzymes that cut on the average every 600 to 1,000 bases. We used nine convenient sets of enzymes for the production effort which provided the ability to "walk" across all regions of the genome.

Sequencing the AT-rich genome of *M. capricolum* (70% AT) also required modification of standard chemical sequencing methods developed for DNA with a higher GC content. Six redundant chemical modification and cleavage reactions (G, G>A, A>C, C, C+T and T) were performed for each restriction enzyme digest to ensure accurate reading of the sequence ladder. Chemical sequencing reaction conditions were optimized to (1) accommodate the higher AT content of *M. capricolum* DNA, (2) enable maximum reproducibility of the chemical sequencing ladders generated from restriction fragments of about 500 base pairs, and (3) allow simultaneous processing of multiple samples (Dolan *et al.*, 1995).

Indirect Transfer Electrophoresis

Direct blotting electrophoresis technology enables the collection of DNA fragments on a nylon membrane concurrent with their size fractionation by denaturing gel electrophoresis. We developed a modification of direct blotting electrophoresis that we call "indirect transfer electrophoresis (ITE) to produce membranes that had sequence data resolved to about 350 bases. We used instruments from Betagen, a design close to that of the original Pohl and Beck device, in which the sequencing membrane is supported by a nylon mesh. Pall Biodyne plus, a positively charged nylon, was routinely used for all the production sequencing membranes in this project.

Resolution of the sequencing ladders was enhanced by altering the gel composition to include 40% formamide in both the gel and the lower buffer chamber. These gels ran about 30% faster than conventional urea gels and we routinely resolved about 30% more sequence on the membranes with this method. A further innovation was to remove the direct contact between the membrane and the gel. ITE gels were recessed approximately two millimeters from the edge of the glass plates and this region was filled with the formamide containing buffer. As the membrane passed across the ends of the glass plates, the DNA bands left the gel, migrating undisturbed through the liquid layer, and impinged on the nylon membrane. The lack of physical contact between the gel and the membrane prevented mechanical damage to the gel by the passage of the membrane, which can tear off gel fragments. Using this technology, we routinely made sequencing membranes from which we were able to read out to 300 to 350 bases from the restriction site.

Sequence Reading and Assembly

In this component, the contigs were assembled, proofread, and new oligonucleotides were predicted and queued for synthesis. The autoradiographs were read manually and the sequence information was entered directly into a GDE interface. The set of "reads" from a single oligonucleotide were assembled within the interface and double-checked against the autoradiograph. Subsequently, these sets of "reads" were assembled onto the contigs automatically by our system.

Our sequence assembly problem was much simpler than that of a shotgun sequencing project, since the Multiplex Genomic Walking strategy is a directed walk. We were able to keep track of the growing point on each contig along with the corresponding oligo that was probing that growing point by tracking a link between each probe, each membrane, and each film in the database. Therefore, we knew from which contig each sequence cluster had been derived and consequently to which contig it had to assemble to. As described above, a hybridization generated sequence fragments that assembled into two clusters, one going 5′ to 3′ away from the probe hybridization site and the other going 3′ to 5′ away from this site. These two clusters were assembled and then compared automatically to the sequences at the ends of the contig, to determine which sequence was the complement to existing sequence and which was *de novo* data that extended the contig across the chromosome. Subsequently, a new consensus sequence was determined for each contig.

New oligos were selected using a computer routine that identifies, in the new consensus, the last restriction site (from the enzymes set used in the project) in the forward direction and then selects a 20-base long sequence for the next oligonucleotide in the region just beyond the last restriction site. The oligo selection was subject to appropriate criteria of melting temperature and secondary structure. The list of new oligonucleotides to be synthesized was sent through the computer system to the oligonucleotide synthesizers automatically and inventoried in the database.

Troubleshooting

We determined a consensus sequence, and checked each growing point, so that after one or two cycles the only portion of the sequence that was ambiguous was that portion at or near the growing point which was not yet covered on the opposite strand. The sequence behind that was confirmed on both strands and considered completed. Mismatches between the two strands, and regions covered only by one strand, were color-flagged for identification. If a mismatch between the two strands could not be resolved by re-examination of the raw data or the coverage of the other strand was not obtained by the current probing, a new oligonucleotide was selected, synthesized, and hybridized to the membranes to obtain the additional sequence information to clarify these regions.

When the growing points overlapped or crossed each other, we treated those as independent sequences. We did not form a consensus initially between the two growing points, but rather checked the sequences for accuracy. Since the two growing points crossed a given region in different directions and using different oligonucleotides, they represented independent sequences of the same region, and the agreement between those sequences represents an independent test of the accuracy of the sequencing.

Quality Control

The issue of quality control was a far greater problem that initially realized at the beginning of this project. The question of how reproducibly things were done and whether people actually knew, or recorded exactly how they did each experiment, turned out to be a much larger problem than anticipated. It was not just that records must be kept, but there must be a way of assuring that people actually did the experiments exactly the way they recorded them. These are classic issues of "General Manufacturing Process" management and without these checks one could not troubleshoot the experiment if it went wrong or be guided clearly on the development path. We developed the use of bar codes and sign-offs in SOP's "Standard Operating Protocols" (SOPs) to provide quality assurance. A large fraction of our effort went into these problems of control and assurance to verify that the machines have worked correctly or to learn why a particular procedure had stopped working.

Data Analysis

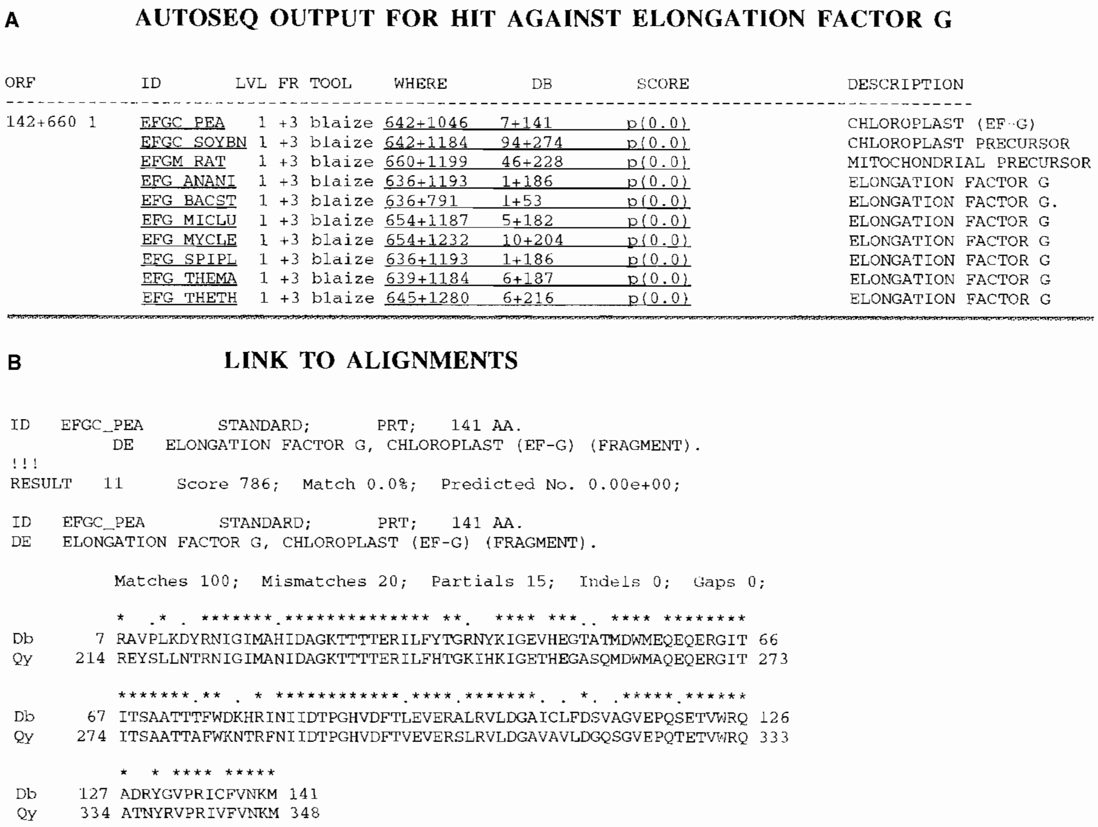
We collaborated with Peer Bork on the initial analysis of the *M. capricolum* data using the Genquiz developed at EMBL (Bork *et al.*, 1995). The assembled contigs were subjected to Blastx searches and the Blastx output was automatically parsed to check for frameshifts and artificial stop codons using the program "Frameshift". All possible open reading frames (ORFs) longer than 10 amino acids were predicted and translated. The requirement for recording ORFs within the contigs was the presence of start and stop codons; in terminal fragments only start (C-terminus) or stop codons (N-terminus) had to occur. BLAST homology searches was carried out on the resulting 1845 ORFs. DNA sequence databases were then screened for non-coding elements such as RNAs and internal repeats and all results were stored in a relational database for further statistical evaluation (see http://uranus.gmu.edu/Mycoplasma/EMBL.FunctionTable.html).

We collaborated with Ross Overbeek, Terry Gaasterland, and Natalia Maltsev to develop a browsing engine called AUTOSEQ in an attempt to incorporate heuristics into the primary identification problem. This semi-automated tool was developed to sort through the output of various search engines and putative identification of each orf is presented in a html browser (Figure 42-4a) which links to the alignments supporting the assertion (Figure 42-4b). We used the initial Genquiz identifications to set the cutoff parameters for the various tools and succeeded in automating the primary identification considerably (see http://www.mcs.anl.gov/home/gaasterl/autoseq/Reports/mcj1995/SUMMARY-mcj1995.html). The latest version of this tool called MAGPIE (see Chapter 45) has an improved user interface and an updated browsing engine (see http://www.mcs.anl.gov/home/gaasterl/magpie.html).

Evgeni Selkov and his group have worked for many years developing the database of Enzymes and Pathways which is a collection of data encoding all facts relating to enzymology and metabolism. Ross Overbeek and his group at Argonne National Laboratory have been working with Dr. Selkov, helping him to prepare the collection for distribution from within his integrated system of Phylogeny, Metabolism and Alignments (PUMA). In 1994, P.M. Gillevet proposed to examine the question of what could be learned about the metabolism of an organism from the sequence data. The first organism that was selected was *Mycoplasma capricolum* with an estimated third of the genome sequenced (see http://www.mcs.anl.gov/home/compbio/PUMA/Production/ReconstructedMetabolism/reconstruction.html).

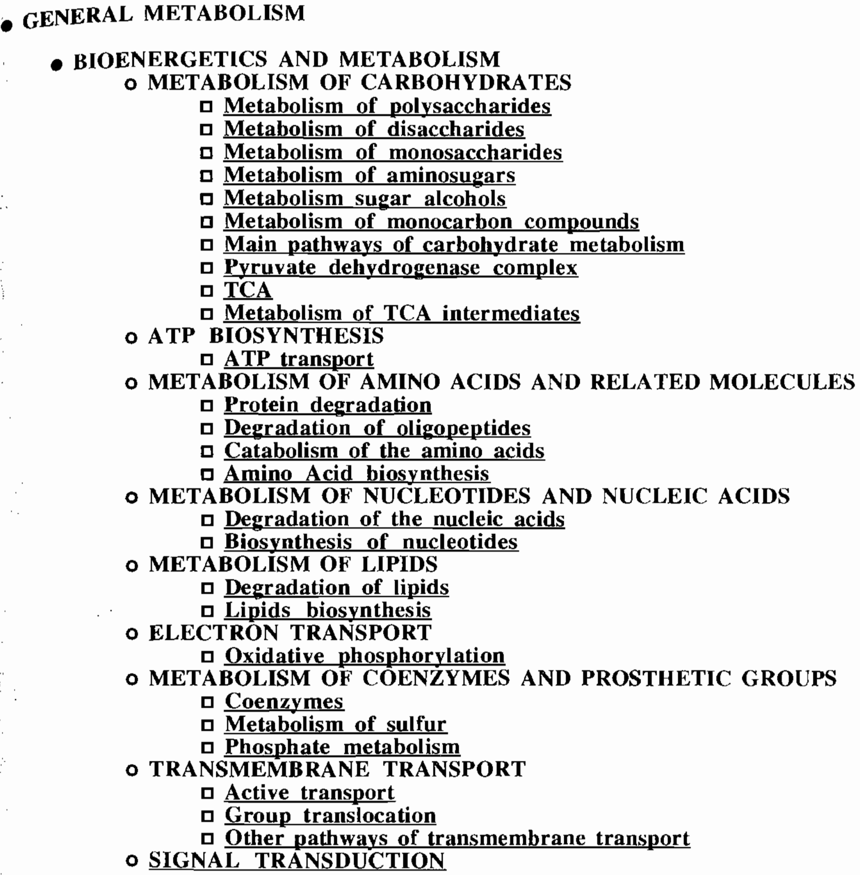
Once one has arrived at an estimated set of metabolic pathways (see Figure 42-5), it is possible to use EMP to derive a set of enzymes that have not yet been located, but are implied by the pathways (or, more precisely, one can produce a list of implied enzymes and a list of possibly occurring enzymes). This list can be used to support a more sensitive analysis of the new sequence data. That is one takes all existing versions of these predicted enzymes and searches the new sequence data for similarities with lower thresholds (in contrast to the initial searches, which took new sequence and searched entire repositories for similarities.

Recently, the PUMA system has been expanded into a new technology that translates sequenced genome information into a consistent model of metabolic and functional organization for a bacterial cell. This technology is based on "What Is There" (WIT), which uses a sequenced genome, detected similarities, and a collection of about 2000 metabolic and functional charts from the Database on Enzymes and Metabolic Pathways (EMP) to construct a model of cellular organization represented as a set of pathway and function diagrams and assertions connecting specific regions of sequence to roles in the diagrams. This model is



```
A          AUTOSEQ OUTPUT FOR HIT AGAINST ELONGATION FACTOR G

ORF          ID          LVL FR TOOL    WHERE        DB         SCORE              DESCRIPTION
----------------------------------------------------------------------------------------------
142+660 1    EFGC PEA     1 +3 blaize 642+1046   7+141        p(0.0)             CHLOROPLAST (EF-G)
             EFGC SOYBN 1 +3 blaize 642+1184   94+274       p(0.0)             CHLOROPLAST PRECURSOR
             EFGM RAT     1 +3 blaize 660+1199   46+228       p(0.0)             MITOCHONDRIAL PRECURSOR
             EFG ANANI    1 +3 blaize 636+1193   1+186        p(0.0)             ELONGATION FACTOR G
             EFG BACST    1 +3 blaize 636+791    1+53         p(0.0)             ELONGATION FACTOR G.
             EFG MICLU    1 +3 blaize 654+1187   5+182        p(0.0)             ELONGATION FACTOR G
             EFG MYCLE    1 +3 blaize 654+1232   10+204       p(0.0)             ELONGATION FACTOR G
             EFG SPIPL    1 +3 blaize 636+1193   1+186        p(0.0)             ELONGATION FACTOR G
             EFG THEMA    1 +3 blaize 639+1184   6+187        p(0.0)             ELONGATION FACTOR G
             EFG THETH    1 +3 blaize 645+1280   6+216        p(0.0)             ELONGATION FACTOR G

B                   LINK TO ALIGNMENTS

ID   EFGC_PEA       STANDARD;      PRT;   141 AA.
        DE   ELONGATION FACTOR G, CHLOROPLAST (EF-G) (FRAGMENT).
!!!
RESULT   11     Score 786;  Match 0.0%;  Predicted No. 0.00e+00;

ID   EFGC_PEA       STANDARD;      PRT;   141 AA.
DE   ELONGATION FACTOR G, CHLOROPLAST (EF-G) (FRAGMENT).

        Matches 100;  Mismatches 20;  Partials 15;  Indels 0;  Gaps 0;

            *  .* .  *******.************** **.  **** ***.. **** ********
Db       7 RAVPLKDYRNIGIMAHIDAGKTTTTERILFYTGRNYKIGEVHEGTATMDWMEQEQERGIT 66
Qy     214 REYSLLNTRNIGIMANIDAGKTTTTERILFHTGKIHKIGETHEGASQMDWMAQEQERGIT 273

            *******.** . * ************.****.*******. . *. .*****.******
Db      67 ITSAATTTFWDKHRINIIDTPGHVDFTLEVERALRVLDGAICLFDSVAGVEPQSETVWRQ 126
Qy     274 ITSAATTAFWKNTRFNIIDTPGHVDFTVEVERSLRVLDGAVAVLDGQSGVEPQTETVWRQ 333

            *  * **** *****
Db     127 ADRYGVPRICFVNKM 141
Qy     334 ATNYRVPRIVFVNKM 348
```

Figure 42-4. Autoseq: A semi-automated tool was developed to sort through the output of various search engines. A putative identification of a orf is presented in a html browser (a) with links to the alignments supporting the assertion (*b*).



- **GENERAL METABOLISM**
  - **BIOENERGETICS AND METABOLISM**
    - o **METABOLISM OF CARBOHYDRATES**
      - □ **Metabolism of polysaccharides**
      - □ **Metabolism of disaccharides**
      - □ **Metabolism of monosaccharides**
      - □ **Metabolism of aminosugars**
      - □ **Metabolism sugar alcohols**
      - □ **Metabolism of monocarbon compounds**
      - □ **Main pathways of carbohydrate metabolism**
      - □ **Pyruvate dehydrogenase complex**
      - □ **TCA**
      - □ **Metabolism of TCA intermediates**
    - o **ATP BIOSYNTHESIS**
      - □ **ATP transport**
    - o **METABOLISM OF AMINO ACIDS AND RELATED MOLECULES**
      - □ **Protein degradation**
      - □ **Degradation of oligopeptides**
      - □ **Catabolism of the amino acids**
      - □ **Amino Acid biosynthesis**
    - o **METABOLISM OF NUCLEOTIDES AND NUCLEIC ACIDS**
      - □ **Degradation of the nucleic acids**
      - □ **Biosynthesis of nucleotides**
    - o **METABOLISM OF LIPIDS**
      - □ **Degradation of lipids**
      - □ **Lipids biosynthesis**
    - o **ELECTRON TRANSPORT**
      - □ **Oxidative phosphorylation**
    - o **METABOLISM OF COENZYMES AND PROSTHETIC GROUPS**
      - □ **Coenzymes**
      - □ **Metabolism of sulfur**
      - □ **Phosphate metabolism**
    - o **TRANSMEMBRANE TRANSPORT**
      - □ **Active transport**
      - □ **Group translocation**
      - □ **Other pathways of transmembrane transport**
    - o **SIGNAL TRANSDUCTION**

Figure 42-5. Functional Classes of the Metabolic Network of *M. capricolum:* The putative identifications made with Genquiz and Autoseq were used to infer the set of metabolic pathways that exist in the organism. The results are accessed via an active html display that has links to the individual enzymes in each pathway along with the alignments associated with that inference (see http://www.mcs.anl.gov/home/compbio/PUMA/Production/ReconstructedMetabolism/reconstruction.html).

an attempt to reconcile the sequence data, the phylogenetic context, and the phenotypic and biochemical knowledge of the sequenced organism. An effort has been initiated to organize what is known of the metabolism for a number of the organisms for which substantial sequence has been released to the research community (see http://uranus.gmu.edu/WIT/wit.html). Several examples of the prediction from the metabolic reconstruction have been proven correct and it is hoped that further refinement of the system will enhance its robustness. Finally, as all theoretical prediction from the reconstruction must be confirmed experimentally, we are in the process of developing tools to identify critical enzymatic steps in the metabolic network that can be biochemically verified in the wetlab.

Summary

We accumulated over a million raw bases (1,039,095) of *Mycoplasma capricolum* sequence during the project at Harvard that assembled into a quarter of a million linear bases (267,686 bp). We have analyzed 372 non-overlapping contigs covering 214,528 base pairs and identified 220 open reading frames in the organism (Bork *et al.*, 1995). Only 61 frameshifts and aberrant stop codons were identified in 103,000 bases contained in the analyzed orfs indicating the error rate of our finished data is less than 10^{-3}. The identification of 220 distinct proteins revealed the minimum number of proteins encoded by the 372 contigs. At the DNA level, numerous matches with tRNA, rRNA and snRNA-like sequences were found. The current analysis of the *Mycoplasma capricolum* genome can be found on the World Wide Web at "http://uranus.gmu.edu/myc-collab.html" and the reconstruction of the metabolic network can be found at "http://www.mcs.anl.gov/cgi-bin/overbeek/Production/selkov__recon.cgi?Mycoplasma%20capricolum+evi dence."

The 220 distinct proteins represent nearly half of the total number of about 500 proteins expected in *M. capricolum* (Muto, 1987). Furthermore, we identified about 35% of the known infrequently occurring restriction sites in the organism in the 215,000 bases that were analyzed indicating that the size of the *M. capricolum* genome is on the order of 765 kb and that we sequenced close to a third of the genome (Bork *et al.*, 1995).

Technical Problems with Multiplex Genomic Walking

Two related technical problems with the process as implemented at Harvard were encountered. The first was a high failure rate of hybridization, that is 40% of the probes either failed to hybridize or gave very weak, unreadable signal. The hybridization failures were due to picking oligos in inaccurate sequence at the end of the growing contig such that the oligos failed to hybridize. This problem will plague any directed sequencing approach that picks probes or primers on single stranded coverage and may be unavoidable, that is one may have to accept this inherent failure rate. Picking probes on better quality regions (multiple coverage) would help but in our technique there was a negative tradeoff between the rate of stepping forward and picking the probe from confirmed multiple covered sequence. Specifically, the step size was decreased when probe selection was from regions of multiple coverage which are further in from the end of the growing contig.

Many of the oligonucleotides that gave weak signal were synthesized from accurate sequence in retrospect and it is still undetermined why they failed to produce stronger signals. There was no canonical secondary structure involved in these probes (potential probes are checked for stability, hairpins and self complementarity before they are made) but the majority do have a high T content and it is hypothesized that these failures are due to non-canonical secondary structures (non-Watson-Crick base pairing).

The second technical problem was related to the overall signal intensity of the autorads. Signal strength was variable with many reads having only a signal/noise ratio of about three. This issue dictated that the gels be read by hand and probably contributed to the above hybridization failure rate. Apparently other factors, in addition to those criteria addressed in our probe selection programs, are components in the efficiency of hybridization and the generation of signal when oligonucleotide probes are hybridized to charged nylon membranes. The membranes could have been exposed longer to increase the signal as the membranes were only exposed overnight in the production process but this would have led to a severe disruption of the production process. We are presently looking into alternative detection methodologies to alleviate this bottleneck in the strategy.

Advantages of Technology

The project at Harvard has proven that by repeated probing of the same set of membranes one can walk around the genome of a small organism. The simple repetitive process involved in sequencing then is to synthesize an oligonucleotide, hybridize it onto a membrane, wash and expose the membrane, strip the membrane, read the sequence, synthesize a new oligonucleotide, and reprobe the membrane. The membranes were reused many times; the present set of membranes were hybridized around 70 times, with no diminution of signal strength. The repetitive process was very simple, and the ultimate rate of sequencing depended on the number of membranes that could be handled at once, and the length of the read achieved from each probing of a membrane.

A major advantage of this technology, especially as it applies to microorganisms with genomes less than a million bases, is that the organism's genome is sequenced the organism directly. This avoids ambiguities due to artifacts that could be introduced in the process of cloning and simplifies the closure problem in that there are no "unclonable" regions to analyze. A second advantage is that both DNA strands are examined directly, and thus the sequence is verified and the presence of modified bases, such as methylated C residues is readily observed. Finally, because this is a linear walking procedure, the computer assembly of the sequence is a straightforward one and does not involve the great complexities that arise with shotgun sequencing of very large organisms.

In conclusion, the technology presented in this report can be used to directly sequence small bacterial genomes or entire YACS or cosmids. Furthermore, the rationale for the Multiplex Genomic Walking will be applicable to novel sequencing methods now being developed using new fragment separation techniques and more sensitive detector systems.

Acknowledgments

We would like to thank all members of the Harvard Genome Lab, F. Barton, S.E. Brenner, R. Clark-Whitehead, N. Douglas, L. Marquez, B. Richter, J. Sartell, and J. Williams, for their efforts over the course of the *Mycoplasma capricolum* sequencing project to make this new technology a reality. We would like to thank the members of the Argonne team, T. Gaasterland and N. Maltsev, who were critical to the development of Autoseq. Finally, we would like to thanks all the members of Selkov team who have developed EMP and laid the foundation for the work on metabolic reconstruction. This work was supported by NIH Grant R01 HGD0124.

References

Bork, P., C. Ouzounis, G. Casari, C. Sander, M. Dolan, W. Gilbert, and P. M. Gillevet. 1995. Exploring the *Mycoplasma capricolum* genome: A parasite reveals its physiology. *Mol. Microbiol.* 16:955–967.

Dolan, M., A. Ally, M. S. Purzycki, W. Gilbert, and P. M. Gillevet. 1995. Large Scale Genomic Sequencing: Optimization of Genomic Chemical Sequencing Reactions. *BioTechniques* 19(2):264–273.

Gillevet, P. M. 1993. Integration of the Wet Lab and Data flow in Multiplex Genomic Walking. *Proceedings of the Second International Conference of Bioinformatics, Supercomputing and Complex Genome Analysis*, A. Hua, ed. World Scientific Publishing Co. River Edge, NJ.

Maniloff, J., R. N. McElhaney, L. R. Finch, and J. B. Baseman, eds. 1992. *Mycoplasma: Molecular Biology and Pathogenesis*, American Society for Microbiology, Washington DC.

Muto, A., F. Yamao, and S. Osawa. 1987. The genome of *Mycoplasma capricolum. Progr. in Nucl. Ac. Res.* 34:28–58.

Ohara, O., R.L. Dorit, and W. Gilbert. 1989. Direct genomic sequencing of bacterial DNA: The pyruvate kinase I gene of *Escherichia coli. Proc. Natl. Acad. Sci. USA* 86:6883–6887.

Poddar, A. K., and J. Maniloff. 1989. Determination of microbial genome sizes by two-dimensional denaturing gradient gel electrophoresis. *Nucl. Ac. Res.* 8:2889–2895.

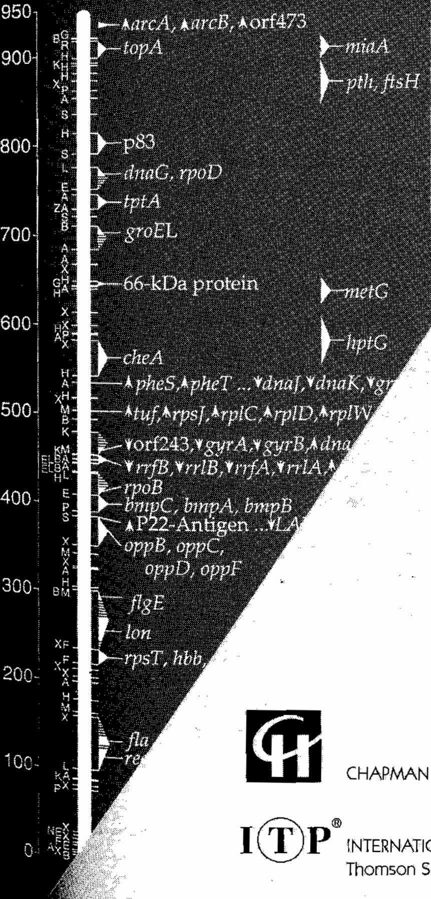
Ryan, J. L., and H. J. Morowitz. 1969. Partial purification of native rRNA and tRNA cistrons from Mycoplasma sp. (Kid). *Proc. Natl. Acad. Sci. USA* 63(4):1282–1289.

Smith, S. W., R. Overbeek, C. R. Woese, W. Gilbert, and P. M. Gillevet. 1994. The Genetic Data Environment (GDE): An Expandable Graphic Interface for Manipulating Molecular Information. *CABIOS* 10 (6):671–675.

CCCATTCTTCCTTTATGGC
TCCTGATTTTTGTTGGGA
GACACCTACTTCAACACC
AGTTTCTGGTGTTCAGAA

Bacterial Genomes

Physical Structure and Analysis



EDITED BY

Frans J. de Bruijn
Michigan State University

James R. Lupski
Baylor College of Medicine

George M. Weinstock
University of Texas Medical School

CHAPMAN & HALL

I(T)P® INTERNATIONAL THOMSON PUBLISHING
Thomson Science

New York • Albany • Bonn • Boston • Cincinnati • Detroit • London
Madrid • Melbourne • Mexico City • Pacific Grove • Paris • San Francisco
Singapore • Tokyo • Toronto • Washington