REGULATION OF SXY IN *HAEMOPHILUS INFLUENZAE* AND OTHER
*PASTEURELLACEAE*

by

MILICA VOLAR

B.Sc., The University of Belgrade, 2004

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

(Zoology)

THE UNIVERSITY OF BRITISH COLUMBIA

December 2006

© Milica Volar, 2006
Abstract

Natural competence is ability of bacteria to take up exogenous DNA from their surroundings. *Haemophilus influenzae*, a member of the Pasteurellaceae family, tightly regulates competence development, inducing uptake only when cells are starved both for sugars and nucleotides. Two proteins are involved in activation of competence gene transcription: Sxy and CRP. While CRP is a global regulator of genes involved in sugar metabolism, Sxy regulates a small regulon composed primarily of competence genes. Competence levels are intimately connected with the abundance of Sxy protein in the cell, thus revealing the regulation of the *sxy* gene is an imperative for understanding how and why *H. influenzae* becomes naturally competent.

Previous studies found that strains carrying point mutations predicted to weaken *sxy* RNA secondary structure overproduce Sxy and are hypercompetent, while strains with mutations predicted to strengthen this structure, do not produce Sxy and are non-competent. This suggested that the secondary structure of *sxy* mRNA limits the gene’s expression in the wild type in competence non-inducing conditions. I used bioinformatics and biochemical techniques to investigate whether *sxy* is regulated via its RNA secondary structure. Nuclease mapping confirmed that *sxy* RNA folds into a stable secondary structure and that the point mutations in hyper- and non-competent *sxy* mutants weaken or strengthen this structure, respectively. I also examined whether the regulation of *sxy* is conserved in Pasteurellaceae by comparing the predicted secondary structures of RNA in *sxy* homologues. The results suggested that the RNA secondary structure is only important in regulation of *sxy* in *H. influenzae* strains and is not conserved in other
Pasteurellaceae. However, all of the examined species had long intergenic regions upstream of *sxy* homologues suggesting that additional regulatory element(s) is present. Based on the results in the thesis, I propose two models of *sxy* regulation: in one, *sxy* mRNA secondary structure inhibits translation initiation and in the other this structure allows binding of lignad(s) and acts as a riboswitch.
# TABLE OF CONTENTS

Abstract......................................................................................................................... ii
Table of contents.............................................................................................................. iv
List of tables................................................................................................................... vii
List of figures.................................................................................................................. viii
List of abbreviations...................................................................................................... ix
Acknowledgments.......................................................................................................... x

**CHAPTER 1: INTRODUCTION** .................................................................................. 1

1.1 Distribution of natural competence........................................................................... 1
1.2 Induction of competence development in *H. influenzae*.............................................. 1
1.3 Competence genes and regulation of competence in *H. influenzae*......................... 3
1.4 *sxy* as a competence gene and hypercompetence mutation isolation...................... 5
1.5 Regulation of *sxy*..................................................................................................... 8
1.6 RNA secondary structure could be regulating *sxy* expression ............................... 8
1.7 Specific objectives of the thesis.................................................................................. 11

**CHAPTER 2: MATERIALS AND METHODS** ......................................................... 13

2.1 General methods...................................................................................................... 13
  2.1.1 Strains, plasmids and culture methods................................................................. 13
  2.1.2 Template preparation.......................................................................................... 13

2.2 RNA methods......................................................................................................... 17
  2.2.1 RNA preparation............................................................................................... 17
2.2.2 RNA secondary structure mapping .............................................. 18

2.3 Bioinformatics analysis .................................................................... 19
  2.3.1 DNA sequence analyses ............................................................... 19
  2.3.2 sxy RNA secondary structure predictions .................................... 20
  2.3.3 Covariance analysis of sxy RNA in Pasteurellaceae ...................... 20
  2.3.4 Analysis of intergenic regions upstream of sxy in Pasteurellaceae ... 21

CHAPTER 3: RESULTS ............................................................................. 22

3.1 Phylogenetic comparative analysis in Pasteurellaceae ...................... 22
  3.1.1 Regions upstream of sxy gene in Pasteurellaceae ....................... 22
  3.1.2 Phylogenetic comparative analysis of sxy and its homologues ....... 26
  3.1.3 Covariance analysis in H. influenzae strains ................................. 32

3.2 RNA secondary structure mapping .................................................. 35
  3.2.1 sxy RNA folds into a stem loop structure ................................... 35
  3.2.2 Enzymatic mapping of sxy RNA supports computationally predicted structure ........................................ 43
  3.2.3 Secondary structure of sxy-1 mRNA ............................................ 46
  3.2.4 Secondary structure of sxy-7 mRNA .......................................... 53

CHAPTER 4: DISCUSSION .................................................................... 55

4.1 Co-variance analysis ....................................................................... 55
  4.1.1 Pasteurellaceae could carry a regulatory element in the region upstream of sxy homologues ................................................................. 55
  4.1.2 Regulation of sxy is probably not conserved in Pasteurellaceae .... 57
4.1.3 RNA secondary structure could be important in regulation of sxy in *H. influenzae* only...

4.2 RNA secondary structure mapping
4.2.1 5' RNA secondary structure may limit the expression of sxy
4.2.2 Is the described sxy RNA secondary structure present in *in vivo* conditions?

4.3 Possible mechanisms of sxy regulation
4.3.1 sxy expression could be regulated via translation initiation by sxy RNA secondary structure
4.3.2 sxy could be regulated by a riboswitch

4.4 Conclusion

References
Appendix
LIST OF TABLES

Table 2.1 Bacterial strains used in this study.................................................................14
Table 2.2 Plasmids used in this study.............................................................................15
Table 3.1 sxy homologues in Pasteurellaceae..............................................................23
Table 3.2. Specific cleavage sites for RNases used for in vitro mapping of sxy, sxy-1 and
sxy-7 RNA....................................................................................................................36
Table A.1. Accession numbers of annotated genomes used in the study.................74
LIST OF FIGURES

Figure 1.1 Model of competence development in *H. influenzae* ...........................................6

Figure 1.2 Transformation frequencies of wild type, *sxy*-1 and *sxy*-7 mutants ............................7

Figure 1.3 Region of *H. influenzae* KW20 chromosome with *sxy* and its upstream region .................9

Figure 2.1 Construction of pGEMsxy-7 ..................................................................................16

Figure 3.1 Comparison of intergenic regions upstream of *sxy* in Pasteurellaceae .......................25

Figure 3.2 ClustalW alignment of *sxy* and its homologues in Pasteurellaceae .........................28

Figure 3.3 RNA secondary structure predictions of *sxy* and its homologues in Pasteurellaceae ........31

Figure 3.4 Phylogenetic comparative analysis of five fully sequenced *H. influenzae* .................34

Figure 3.5 Sample analysis ......................................................................................................39

Figure 3.6 Nuclease mapping of *sxy* mRNA ............................................................................41

Figure 3.7 Secondary structure of *sxy* mRNA *in vitro* ............................................................44

Figure 3.8 Summary of base-pairing in *sxy* RNA secondary structure ........................................45

Figure 3.9 Nuclease mapping of wild type *sxy*, *sxy*-1 and *sxy*-7 RNAs ......................................49

Figure 3.10 Secondary structures of *sxy*, *sxy*-1 and *sxy*-7 RNA *in vitro* ...............................51

Figure 3.11 Summary of base-pairing in the *sxy*, *sxy*-1 and *sxy*-7 RNA secondary structures ..................52

Figure 4.1 Possible mechanisms of regulation of *sxy* expression ..................................................66

Figure A.1 Representative gel pictures .......................................................................................75

Figure A.2 Cleavage intensities of residues in *sxy*, *sxy*-1 and *sxy*-7 RNA .................................77
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amp</td>
<td>Ampicillin</td>
</tr>
<tr>
<td>ATP</td>
<td>adenosine 5'-triphosphate</td>
</tr>
<tr>
<td>BHI</td>
<td>brain heart infusion (rich culture medium)</td>
</tr>
<tr>
<td>bp</td>
<td>base pair</td>
</tr>
<tr>
<td>cAMP</td>
<td>3', 5' cyclic adenosine monophosphate</td>
</tr>
<tr>
<td>CRP</td>
<td>cAMP receptor protein</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DNase</td>
<td>deoxyribonuclease</td>
</tr>
<tr>
<td>E value</td>
<td>expected value (BLAST)</td>
</tr>
<tr>
<td>EtOH</td>
<td>Ethanol</td>
</tr>
<tr>
<td>MIV</td>
<td>&quot;M four&quot;, <em>H. influenzae</em> defined starvation medium</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>NAD</td>
<td>nicotinamide adenine dinucleotide</td>
</tr>
<tr>
<td>NaOAc</td>
<td>sodium acetate</td>
</tr>
<tr>
<td>NCBI</td>
<td>The National Center for Biotechnology Information</td>
</tr>
<tr>
<td>NTPs</td>
<td>ribonucleoside 5'-triphosphates</td>
</tr>
<tr>
<td>ORF</td>
<td>open reading frame</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>RBS</td>
<td>ribosome binding site</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>RNase</td>
<td>ribonuclease</td>
</tr>
<tr>
<td>sBHI</td>
<td>BHI supplemented with hemin and NAD</td>
</tr>
<tr>
<td>TIGR</td>
<td>The Institute for Genomic Research</td>
</tr>
<tr>
<td>UTP</td>
<td>uridine 5'-triphosphate</td>
</tr>
</tbody>
</table>
Acknowledgments

I would like to thank my supervisor, Rosemary Redfield for guiding and supporting me both morally and financially to think about beautiful processes that occur in the world of bacteria. I would also like to thank my supervisory committee George A. Mackie and Naomi Fast for expert guidance and for very open and useful conversations and suggestions.

Very special thanks to my lab mates, past and present: Andrew Cameron, Lindsay Wilson, Heather Maughan and Minghiu Yang with whom I shared many coffees in an attempt to understand sex in bacteria.

Many others also helped me with my project throughout my program, to whom I am very grateful: Arina Omer, Maria Zago, Janet Hankins and Brett McLeod, for teaching me the RNA techniques; Bryony Williams together with the rest of the Keeling lab, for teaching me molecular biology techniques.

I am deeply grateful to my families: both the one in Belgrade and the one in Waterloo, for sharing all of the good and bad moments that I went through in these two years. To my Bojan, who learned, with me, about sxy homologues, for his endless love and support. To my grandfather, Vojkan Pejovic, who taught me, among so many other things, to critically think, ask questions and nourish my imagination.

I owe special thanks to Bryony Williams and Andrea Basler for number of deep and interesting conversations throughout grad school and for being honest friends.

Finally, I would like to thank Dragana Cvetkovic and Oliver Stojkovic who introduced me to the world of evolution and molecular biology and who encouraged me to continue thinking about these spectacular issues.
Chapter 1: Introduction

Bacteria have various ways to cope with changes in the natural environment. One of the changes is nutrient depletion, which induces natural competence development in a number of bacterial species, making them able to bind and take up exogenous DNA from their environment. This energetically expensive process is regulated in most bacteria (for a review see (14,32). Therefore, learning more about the signals that induce competence is important for understanding not only the mechanism of natural competence development, but also its function and evolution.

In *Haemophilus influenzae*, competence development is regulated by two proteins: Sxy and CRP. The focus of this thesis will be on the post-transcriptional regulation of *sxy* expression in *H. influenzae*. Additionally, comparative phylogenetic analysis was employed to investigate if the proposed regulatory mechanism is conserved in other species within the Pasteurellaceae family.

1.1 Distribution of natural competence

Natural competence is the ability of bacteria to take up exogenous DNA from their environment (14,32). DNA uptake is advantageous because it provides a cell with nucleotides, which are energetically expensive to synthesize *de novo*. It can also alter the cell’s genotype if the DNA escapes from degradation and recombines with the bacterial chromosome in a process called natural transformation.
Natural competence is broadly distributed across more than 40 Gram-positive and Gram-negative bacterial species, but the distribution is sporadic making its origin and evolution difficult to understand. Except for Neisseria gonorrhoeae, where natural competence is constitutive, the induction of competence in transformable bacteria is physiologically regulated (4). Mechanisms of natural competence are best understood in Haemophilus influenzae, Neisseria gonorrhoeae (both Gram-negative), Streptococcus pneumoniae and Bacillus subtilis (Gram-positive).

_H. influenzae_ is a member of the Pasteurellaceae family within the γ-proteobacteria subdivision. This clinically important species is a commensal in the human respiratory tract that causes wide range of diseases, including meningitis and eye, ear and lower respiratory tract infections. _H. influenzae_ becomes transformable under very specific conditions (15,29); hence, defining the signals that induce competence is crucial in understanding how and why this mechanism evolved.

### 1.2 Induction of competence development in _H. influenzae_

In _Haemophilus influenzae_ competence is tightly regulated. Since it requires synthesis of a range of new proteins, cells can afford to become competent only under certain environmental conditions where DNA uptake will be useful. In a laboratory environment, genes involved in competence development in _H. influenzae_ are usually induced by transferring exponentially growing cells from a nutritionally rich medium (usually sBHI) to a starvation medium (MIV) (for methods see (27)). This nutritional restriction results in an increase in transformation frequency from $\sim 10^8$ in non-induced cells to $\sim 10^3$.
in fully induced cells. An increase in transformation frequency can be also observed in cells entering the stationary phase when grown only in sBHI. This is explained by nutritional depletion of the medium and, eventually, starvation of the cells. In this case the transformation frequency is always much less than that seen after induction in MIV (usually \( \sim 10^2 \)).

With these observations in mind, it is important to answer which signals are sensed by cells when transferred to starvation medium. Also, what is the specific mechanism by which genes are induced in competence-inducing conditions.

In *H. influenzae*, two intracellular signals are required for competence development: a nucleotide signal and an energy signal (described in detail in following section). During an energy shortage, the production of cAMP in the cell increases, resulting in transcription of genes involved in sugar metabolism (38). The second regulatory signal, essential for maximal competence induction, is a limiting concentration of nucleotides in the cell. Macfadyen (16) showed that supplementing MIV with nucleotides prevents transcription of competence genes, indicating that purine ribonucleotide depletion induces competence development in *H. influenzae*.

1.3 Competence genes and regulation of competence in *H. influenzae*

Under competence-inducing conditions about 25 genes are significantly up-regulated in *H. influenzae*. These genes are organized into 13 transcriptional units whose transcription is activated by two proteins, CRP and Sxy (29).
Although essential for induction of competence genes, CRP (cAMP regulatory protein) is not a competence specific regulator. CRP is well described in *E.coli* where it acts as a global regulator of the genes involved in the cell's response to sugar depletion. In *H. influenzae*, when cAMP concentration rises in the cell, CRP is activated by binding cAMP. The resulting conformational change in CRP allows it to bind at CRP sites in the promoter regions of competence and sugar metabolism genes inducing their transcription (3,13).

*sxy*, on the other hand, is a competence-specific gene: mutants carrying *sxy* null mutations are non-transformable but fully viable indicating its only role is in natural competence. Conversely, constitutive expression of Sxy in *H. influenzae* results in elevated transformation levels under non-inducing conditions (33).

*sxy* is located between the *recA* and *rrnA* genes in the chromosome and encodes a small basic protein. Expressed early in competence development, Sxy induces transcription of competence genes involved in DNA binding, uptake and processing (29). Although it is speculated that *sxy* is an activator of transcription, the specific mechanism by which it acts is still unknown. Comparative analyses have shown that *sxy* is present in three γ-proteobacterial families: Pasteurellaceae, Vibrionaceae and Enterobacteriaceae (2), and that it plays essential roles in competence in some of the species examined (23). Figure 1.1 illustrates the steps in natural competence development in *H. influenzae*.

Production of Sxy protein is one of the first steps in competence induction (29) (see Figure 1.1, Step 1). In parallel, due to the lack of preferred sugars cAMP concentrations rise in the cell. CRP binds cAMP (Fig. 1.1, Step 2) and induces the transcription of competence genes by binding to CRP sites in the promoter regions (10). Recently, two
types of CRP sites with distinct consensus sequences have been described: CRP-N and CRP-S (29). Transcription of the genes that carry CRP-N sites in their promoter regions is activated by binding CRP (shown in red, Step 3), while competence genes that have CRP-S sites are transcribed only in the presence of both CRP and Sxy (shown in blue, Step 4) (3). Finally, the expression of competence genes in *H. influenzae* allows cells to bind, take up and process exogenous DNA (Step 5).

1.4 *sxy* as a competence gene and hypercompetence mutation isolation

*sxy* was first linked to natural competence in 1991. After performing EMS mutagenesis of *H. influenzae*, Redfield (1991) isolated first one (28) and later four additional mutant strains (unpublished) that were shown to be competent early in exponential growth, the phase when cells are usually not competent. All of the five point mutations were mapped to the region upstream of the *recA* gene. Since these strains were “always in the mood” for taking up exogenous DNA, the gene was named *sxy* and (after additional mutants were identified) the mutations were numbered from 1 to 5 (*sxy*-1 to *sxy*-5). The phenotypes of all five mutants are the same, but quite distinct from the wild type. When grown in sBHI medium the rate of spontaneous transformation is 100-1000 times higher in the mutants throughout growth (see Fig. 1.2). Due to the elevated transformation frequency the strains are referred to as hypercompetent mutants. Surprisingly, normal induction of competence (by transferring the cells in exponential phase to starvation medium (MIV)) produces the same transformation frequencies in wild type and hypercompetent mutants (Fig 1.2).
Figure 1.1. Model of competence development in *H. influenzae*. For specific details refer to text in section 1.3.
The distinct response of these strains depending on the different induction conditions raises questions about how the cells sense the signals from the environment, and how \textit{sxy} expression responds to those signals.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure1.png}
\caption{Transformation frequencies of wild type (KW20), \textit{sxy-1} and \textit{sxy-7} mutants. The cells were grown in sBHI and transferred to MIV to induce competence. The transformation frequencies were measured for the wild type (purple), \textit{sxy-1} mutant (pink) and \textit{sxy-7} mutant (blue) (adapted from L.A. Bannister’s PhD thesis (2)).}
\end{figure}
1.5 Regulation of \textit{sxy}

Mapping of the point mutations in \textit{sxy} hypercompetent mutants revealed that they are located in two distinct groups (Figure 1.3 A&B). Both the \textit{sxy}-1 and \textit{sxy}-2 mutations map to the beginning of the coding region: the \textit{sxy}–1 mutation results in an amino acid change (Val\textsubscript{19} \rightarrow Ile) while \textit{sxy}-2 is a silent mutation (Gln\textsubscript{17}). In contrast, the \textit{sxy}-3, 4 and 5 mutations were mapped to adjacent positions 36, 37 and 38 bp upstream of the \textit{sxy} start codon (Figure 1.3 C).

Zulty and Barcak (40) mapped the transcriptional start of \textit{sxy} RNA and characterized this gene's promoter region. Using primer extension analysis they showed that \textit{sxy} mRNA has a 51 nucleotide 5' untranslated region. Thus all the \textit{sxy} hypercompetence mutations alter the \textit{sxy} mRNA, but only \textit{sxy}-1 changes the protein sequence. To ensure that the hypercompetence phenotype is a result of substitutions in only the \textit{sxy} gene, the same mutations were produced by directed mutagenesis and transferred into the non-mutagenized background of \textit{H. influenzae} KW20, where they were found to give the same phenotype (2).

1.6 RNA secondary structure could be regulating \textit{sxy} expression

The position of the mutations, where most are found outside of the coding sequence, implied that the untranslated region of \textit{sxy} RNA has a regulatory role and suggested that \textit{sxy} could be regulated by secondary structure of its mRNA. The untranslated region was predicted (by eye) to base pair with the beginning of the coding region, forming a 16 bp
Figure 1.3. A. Region of the *H. influenzae* KW20 chromosome with *sxy* and its upstream region. B. Enlarged region with *sxy* untranslated region showing the position of point mutations and transcription initiation sites. Mutations causing hypercompetence are shown in red and mutations resulting in a non-competent phenotype are shown in blue. C. Predicted secondary structure of the 5’ end of *sxy* RNA including the positions of the point mutations, ribosomal binding site and the translational start (adapted from (2)).
stem (see Figure 1.3 B and C). Additionally, computer analysis, using software that predicts RNA folding based on the sequence (Mfold (20)), gave very similar results (2). Since sxy is an essential activator of competence genes and each of the hypercompetence substitutions was found to disrupt the base-pairing in the predicted stem region, it was hypothesized that destabilization of the stem causes overexpression of Sxy, producing the hypercompetent phenotype in the mutant strains. Similarly, in the wild type, it was hypothesized that sxy expression is limited during early log phase by the secondary structure of sxy mRNA (2).

To test if the stem in the secondary structure limits sxy expression, Laura Bannister created two additional mutant strains, sxy-6 and sxy-7 (Fig 1.3 B&C). The sxy-6 mutant was designed to restore pairing in the stem region by introducing compensatory mutations combining the sxy-3 and sxy-1 genotypes. The sxy-7 strain carries two mutations designed to form two additional base-pairs in the stem region, by replacing the 2 bp bubble (Fig 1.3 C). As reported in L. Bannister’s thesis (2), analysis of the sxy-6 and sxy-7 phenotypes revealed that they are non-transformable (Figure 1.2). Additionally, in competence-inducing conditions both of these strains have almost no sxy transcript (measured by Northern analysis) or protein (measured by β-galactosidase expression from a lacZ fusion) (2). This analysis provided direct genetic evidence that the stem in the mRNA secondary structure is biologically important in the regulation of sxy.

Using RNA secondary structure mapping techniques, Bannister analyzed the folding of sxy and sxy-1 RNA in in vitro conditions. The goal was to test whether these RNAs form the predicted stem structure when folded, and to determine if the known mutations affect the folding. RNA secondary structure mapping with nucleases (reviewed in (9)) allows
identification of RNase sensitive sites in folded RNA, using enzymes with preferences for specific residue types and pairing states. Bannister used RNase CL3, T1 and T2 and these enzymes revealed unpaired As, Gs, Cs and Us in \textit{sxy} and \textit{sxy-1} RNA. The results showed that in both RNAs the secondary structure is similar to the one predicted by Mfold and that the predicted stem region forms in \textit{in vitro} conditions (2). Although genetic and protein expression data strongly suggest that \textit{sxy-1} mutation acts by destabilization of its secondary structure, mapping analysis \textit{did not} provide any evidence that the \textit{sxy-1} secondary structure differs from the wild type one. Further investigation of this issue via nuclease mapping is required.

1.7 Specific objectives of the thesis

The focus of this thesis is the role of RNA secondary structure in the regulation of the \textit{H. influenzae sxy} gene. Using phylogenetic comparative analysis, a common regulatory element was sought in the RNAs of \textit{H. influenzae sxy} and its homologues within the Pasteurellaceae family.

Additionally, the secondary structure \textit{sxy} RNA was examined by enzymatic mapping using RNases that differ in specificity from the ones previously used by Bannister. Two single-stranded RNases (RNase TI and RNase A) and an RNase specific for double-stranded regions were used in the mapping. To test whether the mutations affect RNA folding the secondary structures of a hypercompetent \textit{sxy-1} mutant and a non-competent \textit{sxy-7} mutant were determined.
Finally, whether the regulation of \textit{sxy} homologues in Pasteurellaceae is conserved and how the gene might be regulated is discussed.
Chapter 2: Materials and methods

2.1 General methods

2.1.1 Strains, plasmids and culture methods

The bacterial strains and plasmids used in this study are listed in Tables 2.1 and 2.2. All *H. influenzae* strains are derivatives of the original Rd strain (1). Methods for culturing *E.coli* and *H. influenzae* cells are described elsewhere (27). Ampicillin concentrations used for selection of resistant *E.coli* and *H. influenzae* strains were 100 μg/ml and 5.0 μg/ml, respectively.

2.1.2 Template preparation

Plasmids pGEMsxy, pGEMsxy-1 and pGEMsxy-7 were used as templates for *in vitro* preparation of RNAs. Plasmids pGEMsxy and pGEMsxy-1 were constructed by L.A. Bannister (2). Both were constructed by amplifying the full untranslated and coding sxy sequence (position -51 to +678 relative to the translational start) from genomic DNA of *H. influenzae* KW20 or the sxy-1 mutant strain (RR699). Amplicons were digested and cloned into the *ApaI* and *EcoRI* restriction sites of pGEM7 (Sigma), adjacent to the T7 promoter (see Fig 2.1.C). I constructed pGEMsxy-7 by first amplifying the sxy sequence from sxy-7 mutant strain RR854 (see Fig. 2.1), with *ApaI* and *EcoRI* sites introduced on the primers. The insert sequence started at the transcriptional start site (-51) and ended at position +272 in the coding region. The digested and purified sxy-7 insert was cloned into the *ApaI* and *EcoRI* sites of pGEM7.
### Table 2.1. Bacterial strains used in this study

<table>
<thead>
<tr>
<th>H. influenzae Rd</th>
<th>Genotype</th>
<th>Source or reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>KW20</td>
<td>Wild type</td>
<td>(37)</td>
</tr>
<tr>
<td>RR854</td>
<td>sxy-7</td>
<td>(2)</td>
</tr>
<tr>
<td><strong>E.coli</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DH5α</td>
<td>F- 80lacZΔM15 Δ(lacZYA-argF)U169 deoR recA1 endA1 hsdR17(rk-, mk+) phoA supE44 thi-1 gyrA96 relA1</td>
<td>(11)</td>
</tr>
<tr>
<td>JM109</td>
<td>F- 80lacZΔM15 Δ(lacZYA-argF)U169 deoR recA1 endA1 hsdR17(rk-, mk+) phoA supE44 thi-1 gyrA96 relA1 pGEM7-Zf- (Amp&lt;sup&gt;R&lt;/sup&gt;)</td>
<td>Promega</td>
</tr>
<tr>
<td>RR2077</td>
<td>F- 80lacZΔM15 Δ(lacZYA-argF)U169 deoR recA1 endA1 hsdR17(rk-, mk+) phoA supE44 thi-1 gyrA96 relA1 pGEMsxy-7 (Amp&lt;sup&gt;R&lt;/sup&gt;)</td>
<td>This study</td>
</tr>
</tbody>
</table>
Table 2.2. Plasmids used in this study

<table>
<thead>
<tr>
<th>Plasmid</th>
<th>Insert</th>
<th>Source or reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>pGEM7-Zf</td>
<td>pBR322 derivative (Amp(^R)) vector</td>
<td>Promega</td>
</tr>
<tr>
<td>pGEM(_{sxy})</td>
<td>(_sxy) sequence, insert including full untranslated region cloned into Apal and EcoRI sites of pGEM7 (Amp(^R))</td>
<td>(2)</td>
</tr>
<tr>
<td>pGEM(_{sxy-1})</td>
<td>(_sxy-1) sequence, insert including full untranslated region cloned into Apal and EcoRI sites of pGEM7 (Amp(^R))</td>
<td>(2)</td>
</tr>
<tr>
<td>pGEM(_{sxy-7})</td>
<td>truncated (_sxy-7) sequence (to position +272), insert including full untranslated region cloned into Apal and EcoRI sites of pGEM7 (Amp(^R))</td>
<td>This study</td>
</tr>
</tbody>
</table>
1) **pALTERsxyF** (Primer 13 from Bannister, 1999 (2))

\[ Apal \ 5'\text{AAAGGGCCCAGAGACTTTCTACTGACTC} \ 3' \]

2) **pALTERsxyR**

\[ EcoRI \ 5'\text{CATGAAATTCTGAATATCTGATCGAAGATGC} \ 3' \]

---

**Figure 2.1. Construction of pGEMsxy-7.**

**A.** Location and sequence of primers pALTERsxyF and pALTERsxyR used for sxy-7 amplification.

**B.** Cloning of sxy-7. The full UTR sequence (51 bp) together with the first 272 bp of the coding sequence of sxy-7 was amplified by PCR using chromosomal DNA as template. The product was digested with *Apal* and *EcoRI* (sites in the primers), purified and cloned into digested pGEM to be expressed under T7 promoter.

**C.** Position of T7 promoter in pGEMsxy plasmids and the sxy transcription start (arrow). Nucleotides that are underlined represent the left end of the sxy insert.
The insert sequences were confirmed by sequence analysis using Bannister's Primer 10 (2). Genomic DNAs were isolated by the phenol-chloroform method (31). Plasmid DNAs were isolated using a Qiagen plasmid mini kit.

2.2 RNA methods

2.2.1 RNA preparation

Wild type sxy, sxy-1 and sxy-7 RNAs were prepared by in vitro transcription (T7 MEGAscript T7 kit, Ambion) from plasmids linearized at position +272, resulting in 326 nt long run-off transcripts. Both pGEMSxy and pGEMSxy-1 were linearized with SnaBI (they have the full coding sequence cloned into the plasmid) while pGEMSxy-7 was linearized with EcoRI (since it only carries a partial coding sxy-7 sequence). The resulting RNAs were purified from the transcription mix, first by a DNase treatment using a DNA-Free Kit (Ambion) and next by a spin column (RNA easy kit, Qiagen). At this point each RNA sample was quantified by spectrophotometry, and quality was assessed both by agarose gel and A260/280 ratios. Next, the RNAs (~20pmol) were dephosphorylated in 100μl reactions at 37°C for 2 hours using 0.5 U of calf intestinal alkaline phosphatase (Roche). RNAs were extracted with an equal volume of phenol-chloroform and precipitated overnight at -20°C with 0.3M NaOAc and 2.5 volumes of 95% EtOH. After 30 minutes centrifugation at 3,780 x g at 4°C the pellets were rinsed with 200μl of 80% EtOH, centrifuged for 30 minutes (at 3,780 x g at 4°C), bench dried for 10 minutes and dissolved in 20μl nuclease-free water. Dephosphorylated RNAs (~10pmol) were 5’end-labeled in 50μl reactions using 20U of T4 polynucleotide kinase
(New England, BioLabs) and at least 20pmol of $\gamma^{32}$P ATP (6000 Ci/mmol, 250$\mu$Ci, GE Amersham). Labeling reactions were allowed to proceed for 1 hour at 37°C and then the enzyme was inactivated by a 10 minute incubation at 70°C. Unincorporated nucleotides were removed with a spin column (RNA easy kit, Qiagen), and RNA was eluted in 30$\mu$l nuclease-free water.

2.2.2 RNA secondary structure mapping

Prior to partial nuclease digestion, the RNAs were denatured for 5 minutes at 95°C and then allowed to refold for 15 minutes at 37°C. End-labeled RNAs (~2 pmols for each reaction) were then digested with the RNases A, VI and TI (Ambion). Both partially digested RNAs and control RNAs (ladders) were prepared following the manufacturers directions. The RNase concentration in digestions was considered to be optimal if the enzyme cut, on average, less than once per RNA molecule in the reaction. This was empirically determined by choosing conditions that left a substantial amount of undigested RNA in the well after the digestion. Prior to enzyme addition control samples were incubated for 5 minutes at 65°C in 1X Sequencing buffer which allowed denaturation of RNAs. The following enzyme concentrations were used for both partially digested and control sample preparations: for RNase A: 0.005U/$\mu$l, for TI: 0.05U/$\mu$l and for VI: 0.001U/$\mu$l.

After addition of enzyme, samples were incubated for 15 minutes at room temperature before 20$\mu$l of Inactivation/Precipitation buffer (Ambion) was added. An overnight precipitation at-20°C was followed by centrifugation at maximum speed (3,780 x g) at room temperature. The pellets were rinsed with 20$\mu$l of 80% EtOH, centrifuged at
maximum speed at room temperature, air-dried and resuspended in 7 μl of RNA loading
buffer (Ambion).

RNA fragments were resolved on 5%, 8% or 10% denaturing (8M urea) polyacrylamide
gels at 900V and 12mA. Gel dimensions were: 0.2mm thick and 400mm long. Alkaline
digested end-labeled RNA was used as a ladder to help in assigning the bands in the gels
to a specific residue in the RNA sequence.

Gels were visualized by Phospholmager and the intensities of the bands were quantified
using ImageQuant 5.2. The assigned intensities for the particular residues in sxy, sxy-1
and sxy-7 RNAs were normalized by the intensities of residue +27 (for residues in lanes
loaded with RNA partially digested with RNase A) and +29 (for residues in lanes loaded
with RNA partially digested with RNase T1). Three or more replicates were used in
comparisons of the same residue in RNAs of sxy, sxy-1 and sxy-7.

2.3 Bioinformatics analysis

2.3.1 DNA sequence analyses

The nucleotide sequences of sxy and its annotated homologues (H. ducreyi 35000HP,
Pasteurella multocida PM70 and Mannheimia succiniciproducens MBEL55E) were
obtained from The Institute For Genomic Research (TIGR, http://www.tigr.org).

Sequences of unfinished Actinobacillus succinogenes 130Z, A. pleuropneumoniae, A.
actinomycetemcomitans HK1651 and H. influenzae 86-028NP genomes were obtained
from The Laboratory for Genomics & Bioinformatics, University of Oklahoma
(http://microgen.ouhsc.edu/project_home.htm) and the unfinished sequence of H. somnus
129 PT was obtained from http://www.jgi.doe.gov. The sxy sequence of H. influenzae
strain 10810 was retrieved from the Sanger Institute (http://www.sanger.ac.uk/) while the sequences of other two completed H.influenzae strains (R2846 and 2866) were obtained from The University of Washington Genome Centre (http://www.genome.washington.edu/uwgc/).

In non-annotated genomes sxy homologues were detected by searching with Haemophilus influenzae Rd KW20 sxy amino acid sequence against nucleotide sequences using the BLAST search available on the NCBI web server. The amino acid and nucleotide sequence alignments of sxy and its homologues were done in ClustalX or ClustalW (35) using the default settings. The alignments were further refined by eye using MacClade (v 4.0 PPC).

2.3.2 sxy RNA secondary structure predictions

All RNA secondary structure predictions were done using Mfold (version 3.2) developed by Zuker et al (39)(20) (http://www.bioinfo.rpi.edu/applications/mfold/rna/). The sequence of sxy and its homologues used for this analysis contained the full UTR region (51nt) and 60 nt of the coding region. All Mfold parameters were kept at default values.

2.3.3 Covariance analysis of sxy RNA in Pasteurellaceae

Phylogenetic comparison of sxy and its homologues was performed by submitting the RNA sequences to Mfold (39)(20) and visually inspecting the resulting folded sequences for similarities in their secondary structures. sxy in H. influenzae KW20 strain served as a reference for choosing the beginning and end of RNAs to be folded in sxy homologues,
where information on length (or existence) of the untranslated region was not available. For all of the examined RNAs the same settings and conditions were used as for the \textit{sxy} secondary structure predictions, where +1 position represented the translational start annotated by the sequencing centre.

2.3.4 \textbf{Analysis of intergenic regions upstream of \textit{sxy} in Pasteurellaceae}

The sequences of the intergenic regions upstream of \textit{sxy} in annotated Pasteurellacean genomes were retrieved from TIGR. In non-annotated species open reading frames (ORFs) and non-coding (intergenic) regions were identified using the program Sequence Analysis (http://informagen.com/SA/). The parameters were left at default values except for the length of ORFs, where the minimum was set to 50 amino acids. The amino acid sequences of the ORFs found upstream of \textit{sxy} homologues were searched against the OMNIOME genome sequences using BLAST on the TIGR web server to determine gene identities and functions. Putative roles were determined for all ORFs except for the one in \textit{A. actinomycetemcomitans} where the available sequence was truncated.
Chapter 3: Results

3.1 Phylogenetic comparative analysis in Pasteurellaceae

3.1.1 Regions upstream of sxy gene in Pasteurellaceae

BLAST searches by L. Banister identified sxy homologues in several Pasteurellaceae, Enterobacteriaceae and Vibrionaceae species. More Pasteurellacean genomes have been sequenced since, and it was found that all of the species have sxy homologues (3). To test whether this gene shares a common regulatory mechanism in Pasteurellaceae, I compared predicted mRNAs secondary structures of sxy homologues. Table 3.1 lists the sequenced Pasteurellaceae species with their sxy homologues that were included in this study.

To determine if other Pasteurellaceae have a potential to carry a regulatory element in their RNAs (like the one present in H. influenzae KW20), I first examined intergenic regions upstream of sxy homologues in eight species from this family. Figure 3.1.A shows a comparison of the genome regions spanning sxy homologues and the closest upstream open reading frame. The lengths of the intergenic regions upstream of sxy in Pasteurellaceae range from 319 bp in H. influenzae to 1491 bp in P. multocida with the others all ranging from 407 bp to 607 bp.

The upstream genes are diverse: they are involved in different cellular processes such as recombination (recA in H. influenzae) and translation (fusA in A. succinogenes) or they code for ribosomal RNAs or proteins: rpsJ in H. ducreyi and A. pleuropneumoniae and 16SRNA in P. multocida. The transcription direction of the genes found upstream of sxy homologues is divergent to sxy, except in H. somnus where the genes are parallel. The
diversity in the upstream genes suggests that major rearrangements have occurred in this chromosomal region.

<table>
<thead>
<tr>
<th>Species</th>
<th>Gene</th>
<th>E value</th>
<th>Protein length (aa)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>H. influenzae</em> (<em>H. influen</em>)</td>
<td>HI0601</td>
<td>-</td>
<td>217</td>
</tr>
<tr>
<td><em>H. ducreyi</em> (<em>H. ducreyi</em>)</td>
<td>HD1985</td>
<td>1.2e-14</td>
<td>211</td>
</tr>
<tr>
<td><em>H. somnus</em> (<em>H. somnus</em>)</td>
<td>Hsom2370</td>
<td>2e-35</td>
<td>216</td>
</tr>
<tr>
<td><em>A. pleuropneumoniae</em> (<em>A. pleuro</em>)</td>
<td>Aple2132</td>
<td>5e-15</td>
<td>218</td>
</tr>
<tr>
<td><em>A. actinomycetemcomitans</em> (<em>A. actino</em>)</td>
<td>Not annotated</td>
<td>7e-34</td>
<td>211</td>
</tr>
<tr>
<td><em>A. succinogenes</em> (<em>A. succino</em>)</td>
<td>Not annotated</td>
<td>6e-37</td>
<td>219</td>
</tr>
<tr>
<td><em>P. multocida</em> (<em>P. multo</em>)</td>
<td>PM1558</td>
<td>2.5e-36</td>
<td>213</td>
</tr>
<tr>
<td><em>M. succiniciproducens</em> (<em>M. succini</em>)</td>
<td>MS2301</td>
<td>9.3e-30</td>
<td>217</td>
</tr>
</tbody>
</table>

Table 3.1. *sxy* homologues in *Pasteurellaceae*. This table shows species included in the study, the abbreviated species names that will be used throughout this thesis, each gene’s primary annotation, BLAST E values against *H. influenzae* *sxy* and the Sxy protein length.

The lengths of intergenic sequences upstream of *sxy* were compared to the average length of the intergenic regions in each Pasteurellaceae. The average intergenic lengths were determined by dividing the number of nucleotides in the total intergenic sequence by the number of coding genes in the genome (TIGR). It should be noted that this calculation
represents just a rough approximation of the average intergenic length since it did not consider operon organization of the genes and presence of RNA genes in the intergenic regions. The comparison showed that the average length of intergenic region in each Pasteurellaceae species is less than 150nt making the sequences upstream of sxy homologues unusually long and suggesting that a conserved regulatory element could be present in these regions.
Figure 3.1. Comparison of intergenic regions upstream of sxy in Pasteurellaceae.

View of the regions upstream of sxy and its Pasteurellaceae homologues. Numbers represent lengths of the intergenic and coding regions in bp. A. Intergenic regions upstream of sxy homologues in Pasteurellaceae. B. Region upstream of sxy in H. influenzae strains. The analysis was incomplete for H. influenzae R2846 and A. actinomycetemcomitans due to the end of contig (black bars).
3.1.2 Phylogenetic comparative analysis of sxy and its homologues

Phylogenetic comparative (covariance) analysis is one of the methods to infer RNA secondary structure by sequence comparisons among members of a set of functionally identical RNAs from related species. The analysis is based on the premise that RNAs that have identical function will fold in a similar secondary structure despite differences in their primary sequences. In the other words, if a substitution arises in a functionally important region of RNA, a mutation that restores base-pairing would be positively selected to preserve the secondary structure. These pairs of substitutions in the RNA sequences of related species are referred to as covarying residues. The first step in covariance analysis is to find a conserved motif from aligned sequences that can be later used as a reference for choosing the start of the RNA sequence to be computationally folded.

RNA secondary structure predictions for phylogenetic analysis can be made using two distinct classes of software. First, common RNA secondary structures can be obtained by using software that makes a prediction based on the alignment, inferring the possible pairing and taking into account the co-varying bases. This analysis, however, requires conserved motifs throughout the alignment (5). Alternatively, the predictions can be made for each sequence separately based on minimising free energy (ΔGs) of the sequence, using programs such as Mfold or RNADraw. These RNA secondary structure predictions (outputs) can be then compared for similarities. ΔGs represent an approximation of the predicted RNA secondary structure stability, where the structures predicted to be more stable have more negative values.
This particular analysis was based on the hypothesis that if the stem-loop structure present in *H. influenzae sxy* RNA is important in this gene's regulation it will be conserved in different Pasteurellaceae species. Since stem region(s) with multiple co-varying base-pairs are likely to exist in a native RNA structure, RNAs of *sxy* homologues should have a stem region that is very similar to the one present in *H. influenzae*, if their regulation is similar. More specifically, I would expect the stem region to be of a similar length and with few co-varying bases, if the hypothesis is correct.

In order to examine the level of conservation and to determine if there is a site (better than the translational start) that could be used as a reference for the downstream mRNA secondary structure predictions, the *sxy* amino acid and nucleotide sequences from Pasteurellacean genomes were aligned. The amino acid alignment showed that the Sxy sequence is not very conserved, especially at the beginning of the coding region (see Figure 3.2). However, several conserved motif blocks were present in the alignment suggesting that the function of the gene is conserved throughout Pasteurellaceae.

The nucleotide sequences upstream of *sxy* homologues weren't alignable and the sequences were too divergent to identify a single conserved motif that could be later used as a reference in the comparative secondary structure analysis.

Due to the absence of conserved motifs in the nucleotide alignment of *sxy* and its homologues, I decided to use software that produces predictions of individual secondary structures for the covariance analysis. The reference for choosing the length of sequences was the putative *sxy* mRNA secondary structure in *H. influenzae* where the genetic evidence strongly suggests that the -50 to + 60 region is biologically relevant in this gene regulation.
Figure 3.2. ClustalW alignment of sxy and its homologues in Pasteurellaceae. The colors show conserved amino acid groups in the sequence alignment with small hydrophobic residues shown in red, acidic residues in blue, basic in magenta and hydorxyl, amine and basic in green. Dots represent gaps in the sequence alignment.
All of the RNA sequences were 110 nucleotides long and the translation start was used as an anchor point. The RNA sequences were submitted to Mfold, a software that produces RNA secondary structure predictions with the lowest free energy basing them on the nucleotide sequence. This software usually provides more than one output with different thermodynamical stabilities predicted. For each of the RNAs, Mfold produced alternate structures with different ΔGs, except for *H. influenzae* where there was only one. The predicted secondary structures were compared visually, looking for the presence of a stem region similar to the one in *H. influenzae* in length and in position relative to the translation start.

The strongest secondary structure was predicted for *H. influenzae* (ΔG=−23.06 kcal). Most of the other structures were predicted to be much weaker, with two of them having ΔGs close to −15 kcal (*M. succini* and *A. succino*) and the rest ΔGs with less than −10 kcal (see Figure 3.3). The first round of analysis included only the thermodynamically strongest predicted structures; however, because no structural similarities were seen the analysis was expanded to include the rest of the outputs. Additionally, the analysis included structures produced by changing the length of the submitted RNA sequence by sequentially adding nucleotides and looking for alternate similar structure that could rise. None of these analyses yielded an obvious common secondary structure. Although stem regions were present in several predicted structures, the lack of co-varying sites suggested that the RNA secondary structure and thus its function is not conserved in the sxy homologues.

Since Mfold predicted at least one secondary structure in all of the RNAs, I tested folding of 110 nucleotides of both random RNA sequence and the sequence from the *recA*
coding region. In both of the sequences, secondary structures are not expected to be important therefore these RNAs should be unstructured. However, Mfold predicted folding in both of the sequences with ΔGs even stronger than in sxy RNA, suggesting that this algorithm may not be the best choice for the phylogenetic comparative analysis.

To test whether the predicted foldings are the same using slightly different program algorithms, I compared the outputs of RNAs of sxy homologues produced by RNAdraw (http://www.rnadraw.com/), Vienna server (http://www.tbi.univie.ac.at/~ivo/RNA/) and Kinefold (http://kinefold.curie.fr/). Except for Kinefold, all of the other programs predicted the same structures as Mfold did. Kinefold secondary structure predictions, although slightly different, did not reveal any additional structural similarities in the examined RNAs (data not shown).
Figure 3.3. RNA secondary structure predictions of sxy and its homologues in Pasteurellaceae. Sequences of sxy and its homologues from eight fully sequenced Pasteurellaceae were folded in Mfold. All sequences extended from -50 to +60 where +1 is the translational start (red bar). Only the structures with the lowest ΔG predictions are shown.
1.1.3 Covariance analysis in *H. influenzae* strains

In the absence of any evidence for structural conservation of *sxy* homologues in other Pasteureallaceae species, *sxy* amino acid and nucleotide sequences of five available *H. influenzae* strains were aligned (*H. influenzae* KW20, R2846, R2866, 86-028NP and 10810). The *sxy* coding and the upstream intergenic sequences are nearly identical, with only six substitutions in the coding region (two of which are silent), and none in the first 50 nucleotides of the upstream sequence (see Figure 3.4A). The most divergent sequence is that of R2846.

The length of the intergenic regions upstream of *sxy* is 319 bp in strains KW20 and 10810, and about 370 bp in strains R2866, R2846 and 86-028NP. As shown in Figure 3.4B, all the strains except for R2846 had identical –50 to +60 sequence and thus were predicted to fold identically. Although mutations are present only in the R2846 *sxy* sequence, they do not necessarily alter *sxy* RNA folding in this strain: both mutations C_{39} \rightarrow U and A_{50} \rightarrow G conserve base-pairing, while the U_{27} \rightarrow G change is in a loop.

However, secondary structure produced by Mfold showed stem Ia+Ib to be shorter by one base pair. Additionally, the substitutions are predicted to change secondary structure of *sxy* RNA in this strain by increasing the number of base-pairs and producing an additional stem region. The ΔG in R2846 secondary structure was predicted to be stronger than in the rest of the examined strains. The very high sequence similarity and the restriction to only five *H. influenzae* strains makes this analysis uninformative in comparative phylogenetic analysis. Perhaps increasing number of sequenced genomes
and more reliable RNA secondary structure prediction algorithms may soon allow more detailed phylogenetic analysis of sxy in Pasteurellaceae.
Figure 3.4. Phylogenetic comparative analysis of five fully sequenced *H. influenzae* strains. A. Nucleotide alignment of *sxy* sequences (the untranslated and partial coding region) from different *H. influenzae* strains, showing the position of substitutions in strain R2846 (in grey). The amino acid sequence of KW20, R2866, 10810 and 86-028NP strains is shown below the alignment and the two changes in R2846 strain are indicated below. B. *sxy* RNA secondary structure predictions in *H. influenzae* strains. Sequences of *sxy* in five *H. influenzae* strains were folded in Mfold. The sequences extended from -50 to +60 where +1 is a translational start (red bars). The nucleotide substitutions in R2846 strain are shown in grey.
3.2 RNA secondary structure mapping

3.2.1 sxy RNA folds into a stem loop structure

Bannister used the sxy-6 and sxy-7 mutants to show that restoring base-pairing or creating additional base-pairing in the stem region of sxy RNA drastically reduces transformation frequency, suggesting that sxy is likely regulated by its RNA secondary structure. To test if and how sxy mRNA naturally folds, I used in vitro nuclease mapping of RNA secondary structure.

The method is based on RNase cleavage of specific residues in folded RNA molecules. Different RNases have different specificities with some cutting only double-stranded regions and the others cutting 3’ to specific single-stranded residues (see Table 3.1). Use of several different RNases in the analysis allows insight into the RNA secondary structure in in vitro conditions. For RNases that are single-strand specific, partial digestion of folded RNAs allows identification of single stranded nucleotides in the secondary structure while the digestion of denatured RNAs (control or ladder samples) provides information on all the nucleotides in the sequence that are recognized and cut by the RNase. For example, RNase TI in the control sample would cut all of the Gs in the sequence (since all of them are single-stranded in denatured RNA) while in a folded RNA this enzyme would cut only Gs that are single-stranded.

It should be noted that in vitro mapping analysis provides information on the average secondary structure of a thermally dynamic RNA population (rather than a set of RNA molecules in identical conformations). Interactions between residues involved in
secondary structure formation are dynamic, resulting in different RNA conformations. The data from mapping analysis for a certain residue actually represents the probability that the residue is base-paired under specific conditions and over the time of the digestion.

<table>
<thead>
<tr>
<th>RNase</th>
<th>Specific cleavage site</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3' of ss Cs and Us</td>
</tr>
<tr>
<td>VI</td>
<td>base-paired nucleotides</td>
</tr>
<tr>
<td>Tl</td>
<td>3' of ss Gs</td>
</tr>
</tbody>
</table>

Table 3.2. Specific cleavage sites for RNases used for in vitro mapping of sxy, sxy-1 and sxy-7 RNA.

RNA for these analyses is usually prepared by 5' end labeling with $^{32}$P, but other methods have been also employed (9). This is followed by a partial digestion of folded molecules and the resulting fragments are resolved on denaturing sequencing gels. The percentage of polyacrylamide used in gels depends on the size of the RNA fragments that are to be resolved (each gel type usually gives resolution of about 100 nucleotides).

An alkaline ladder, commonly used in nuclease mapping, is prepared by exposing end labeled RNA to alkaline conditions for a limited time which results in nonselective partial cleavage of all phosphodiester bonds and production of a band (in a gel) for every residue. This ladder is used to facilitate the assignment of the bands in the gels to the residues in the RNA sequence.
The decision on the residue status (base-paired or not) is made by comparing the band intensities in the experimental lane (i.e., conditions of limited digestion) to those in the ladder lane; this provides information on how strongly each of the sites in the RNA is cleaved (see Figure 3.5). For example, if the examined band is less intense in the experimental lane than in the control the corresponding residue in RNA is classified as weakly cleaved and is more likely to be base-paired. Classification of this type allows more objective analysis of this relatively subjective method.

Additionally, for sxy, sxy-1 and sxy-7 RNA the bands were quantified by ImageQuant and the intensities for the particular residues were plotted to aide in making comparisons (Fig. A.2 in Appendix).

RNases used in the mapping of sxy RNA secondary structure are shown in Table 3.1. The analysis provided information on single stranded (ss) Gs, ssCs, ssUs and base-paired nucleotides in the sequence. Other potentially informative enzymes are no longer commercially available (9). The gels were prepared to the resolve first 110 nucleotides at the 5' end of the sxy transcript. The analysis I summarize below is based on more than 10 digestions and over 30 gels (for each RNA species); only two gel pictures will be presented (additional representative gel pictures are presented in the Appendix section).

RNase T1 digestion.

RNase T1 specifically cuts at single-stranded Gs; thus, the stronger the cleavage in the folded RNA, the higher the proportion of that G residue that is unpaired in the population of RNA molecules. Lane 3 in Fig 3.6 shows the cleavage pattern for folded sxy RNA
digested with RNase T1. Lane 2 in the same figure is the T1 ladder, where unfolded sxy RNA was digested with the same amount of enzyme.
Figure 3.5. Sample analysis. Segment of a gel illustrating different cleavage intensities in the "structure lane". The decision on the nucleotide status (the right side of the figure) is made by comparing the band intensities in the ladder and structure lane.
In the denatured RNA (lane 2, Fig 3.6) RNase TI cut all of the Gs present in the sxy sequence, giving bands of very similar intensities. In folded RNA (relative to the control RNA) very strong RNase TI cleavage was seen at G_{10}, G_{8}, G_{7}, G_{+25} and G_{+29} and moderate cleavage was observed at G_{+11} and G_{+64}. Some of the sites (G_{+36}, G_{+3}, G_{+12}, G_{+13}, G_{+16}, G_{+31}, G_{+35}, G_{+42}, G_{+46}, G_{+51} and G_{+55}) were cleaved more weakly than in the control, suggesting that these positions are either double stranded or inaccessible in the folded RNA. The bands corresponding to residues +18, +58 and +59 were not present in the structure lane; however, corresponding bands were observed in the TI ladder lane, implying that they are being protected from cleavage in the folded RNA (Fig. 3.6).

**RNase A digestion.**

Pancreatic RNase A specifically cuts unpaired pyrimidines in folded RNA. Unlike RNase TI, which cleaved all of the Gs in the sxy sequence, RNase A was unable to cut some of the single-stranded Us and Cs in the denatured sxy RNA (see Fig 3.6, ladder lane 4). However, like RNase TI, the cleavage sites and the relative band intensities were reproducible over different digestions performed under the same conditions. Lane 5 in Figure 3.6. shows the cleavage pattern of sxy RNA partially digested by RNase A, while lane 4 in the same figure is an RNase A ladder. The following nucleotides were cut strongly in folded sxy RNA relative to the ladder: U_{-28}, U_{-23}, U_{-4}, U_{-23} and U_{-27}. Residues at sites U_{+1} and U_{+21} were moderately cut, while some sites were cut very weakly (U_{+46}, U_{+40}, U_{+19}, C_{+6}, U_{+6}, U_{+8}, U_{+19}, U_{+44}, U_{-23}, C_{+49}, U_{+53} and U_{+56}).
Figure 3.6. Nuclease mapping of *sxy* mRNA. 5' end-labeled *sxy* RNA was subjected to nuclease digestion (see section 2.2.2) and the resulting fragments were resolved on an 8% sequencing gel.

Lanes 1, 2 and 4 were loaded with ladders: lane 1- alkaline ladder, lane 2- RNase TI ladder and lane 4 - RNase A ladder. Lanes 3, 5 and 6 are “experimental” lanes loaded with *sxy* RNA partially digested with: 3: RNase TI, 5: RNase A and 6: RNase VI. The numbers on the left are base positions in the *sxy* sequence relative to the translational start; the sequence itself is given on the right.
RNase VI digestion.

RNase VI, unlike RNases T1 and A, is structure specific but not sequence specific, preferentially cleaving at any base-paired or stacked nucleotides. No ladder is shown in Fig 3.6 because this enzyme does not cut denatured RNA. The lack of a ladder makes classification of the sites recognized by this enzyme more subjective than for samples digested with RNases T1 or A. However, the RNase VI digest results were reproducible. Lane 6 in Figure 3.6. shows an RNase VI cleavage pattern. Several segments in folded sxy RNA were cut with RNase VI to various extents: U₄₃ to A₃₆, U₃₃ to A₂₇, U₁₉ to C₁₅, U₁₂ to U₅, A₁₉ to U₁₁ and C₄₉ to U₄₁. RNase VI shows strong specificity for nucleotides at positions -32 and -29. Bands with moderate intensity corresponded to nucleotides at positions -41, -31, 30, -29, -28, -16, -15, +21 and +49 while the rest of the above-mentioned nucleotides were weakly cut (see Fig.3.6). RNase VI also cut folded sxy RNA at positions -22, +34 and +49. The first two residues were cut very weakly while the last one was moderately cut.

3.2.2 Enzymatic mapping of sxy RNA supports its computationally predicted structure

Figures 3.7 and 3.8 present summaries of the data produced by nuclease mapping, superimposed on the thermodynamically most favorable structure predicted by Mfold. Overall, the enzymatic mapping provided evidence that the sxy transcript folds into a structure very similar to the one predicted by Mfold and that stem Ia+Ib, predicted to regulate sxy expression is present in in vitro conditions.
Figure 3.7. Secondary structure of sxy mRNA in vitro. 5' end-labeled sxy RNA was partially digested with RNase TI, RNase A or RNase VI in separate reactions. Resulting fragments were resolved on 5%, 8% and 10% sequencing gels. The residues were categorized as strong, moderate or weak cut sites, depending on the corresponding band intensities in the gels. The data was superimposed on the Mfold-predicted secondary structure of sxy RNA.
Evidence that the residue is

**Strongly base paired**

**Partly base paired**

**Single stranded**

Figure 3.8. Summary of base-pairing in sxy RNA secondary structure.

The cleavage results from different RNase digestions were summarized and superimposed on the Mfold-predicted secondary structure of sxy RNA.
Although the experimental data for most of the residues support the predicted structure, there are several discrepancies between the theoretical and experimental foldings. The RNase VI cleavage pattern supports the Mfold prediction that U$_{28}$ and the flanking residues in stem Ib are base-paired, but moderately-strong cleavage at by RNase A suggests that this residue is single-stranded. These contradictory data imply that this position is very dynamic in the secondary structure and that it is probably single-stranded in most of the molecules in the RNA population.

Although predicted to be unpaired, both U$_{+8}$ and G$_{+42}$ are only weakly cleaved by RNase A and TI respectively. This may be due to the position of these residues in a very short stretch of single-stranded nucleotides between two long stems, which may be physically inaccessible to the RNases. In the predicted secondary structure stem region I is interrupted by a small bubble with four unpaired residues (C$_{32}$, U$_{31}$, U$_{+49}$ and C$_{+50}$).

Although predicted to cut all four, RNase A very weakly cuts only two of these residues (U$_{+49}$ and C$_{+50}$) suggesting that these positions are mostly base-paired. Pairing of these bases is also supported by RNase VI cleavage pattern: residues U$_{31}$ and U$_{+49}$ were cut moderately and C$_{32}$ was cut strongly by RNase VI.

### 3.2.3 Secondary structure of sxy-1 mRNA

The sequence of sxy-1 differs from the wild type sequence by a single G to A transition at position +55. This change was hypothesized to increase Sxy expression by disrupting base-pairing in the stem region. The Mfold prediction shows that the substitution in sxy-1 would affect its secondary structure only by opening a small bubble in the stem region Ia.
This, however, is predicted to strongly decrease the stability of the stem Ia+Ib, changing its ΔG from -13.3 (in wild type sxy RNA) to -7.3 kcal.

To check whether sxy-1 RNA folds in the same manner as the wild type RNA, I mapped ~110 bases at the 5' end of the transcript. The sxy-1 mutation is predicted to make residues at positions -38 and +55 single-stranded and RNase A is predicted to detect this change by cutting residue C₃₈ in the folded sxy-1 RNA.

The analysis of sxy-1 RNA revealed a very similar cleavage pattern to that of wild type sxy RNA. Except for the three sites, G₃₆, G₄₅₆ and G₄₅₉, which were cut very weakly by RNase T₁, no wild type bands were missing (Fig. 3.9). However, when the relative band intensities were compared, certain residues appeared more sensitive to digestion in sxy-1 RNA. It should be noted that the partial digestions for both RNA types were performed under the same conditions and at the same time to minimize introduction of experimental variation.

Below, I discuss all of the sites whose cleavages differ in sxy and sxy-1 RNAs (see Figure 3.9). Residues at positions +51, +58 and +59 were more susceptible to RNase T₁ in sxy-1 RNA than in the wild type, while those at -28, -19, +6, +8, +44, +49, +53 and +56 were more susceptible to RNase A digestion.

It should be noted that the residues positioned in the loops (see Fig. 3.9. - for example residues G₈, G₇ or G₄₂₃) were cleaved with the same intensity both in sxy and sxy-1 RNA. This provides additional confidence that the differences in cleavage intensities for some sites in sxy and sxy-1 RNA are a reflection of actual structural changes rather than just a result of different conditions.
Nuclease mapping did not provide any data on the additional bubble (residues -32, -31, 
+48 and +49) predicted to form in folded sxy-1 RNA due to the mutation (see Fig. 3.10). 
Pairing between the nucleotides at -38 and +55 position couldn’t be directly detected 
because of the specificity of enzymes used in this analysis. Although C_{38} was predicted 
to be unpaired in sxy-1 RNA, this position was never cut (not cut in wild type sxy either) 
in any RNase A digested samples (not even in the denatured RNA). 
Overall, nuclease mapping of sxy-1 indicated that its RNA folds in a structure very 
similar to the one present in sxy RNA. Although the experimental data failed to produce 
evidence for the predicted additional bubble introduced by the point mutation, it suggests 
that the stem I in sxy-1 RNA secondary structure is generally less stable than the one in 
the wild type (see Fig. 3.11).
Figure 3.9. Nuclease mapping of wild type sxy, sxy-1 and sxy-7 RNAs. 5’ end-labeled RNAs were subjected to nuclease digestion (see section 2.2.2) and the resulting fragments were resolved on an 8% sequencing gel. Lanes 1, 2 and 6 were loaded with ladders: lane 1- alkaline ladder, lane 2- RNase TI ladder and lane 6 - RNase A ladder. Lanes 3, 4 and 5 were loaded with RNAs partially digested with RNase TI: lane 3- wt sxy RNA, lane 4- sxy-1 RNA and lane 5- sxy-7 RNA. Lanes 7, 8 and 9 were loaded with RNAs partially digested with RNase A: lane 7- wt sxy, lane 8- sxy-1 and lane 9- sxy-7. The numbers on sides are base positions in the sxy sequence relative to the translational start.
Figure 3.10. Secondary structures of sxy, sxy-1 and sxy-7 RNA in vitro.

5' end-labeled RNAs were partially digested with RNase T1 or RNase A in separate reactions. The resulting fragments were resolved on 5%, 8% or 10% sequencing gels. The residues were categorized as strong, moderate or weak cut sites, depending on the corresponding intensities in the gels. The data from seven independent digestions (for each enzyme) were summarized and superimposed on the Mfold-predicted secondary structures of sxy, sxy-1 and sxy-7 RNA. The mutated sites in sxy-1 and sxy-7 RNA are circled.

<table>
<thead>
<tr>
<th>RNase</th>
<th>Strong</th>
<th>Moder.</th>
<th>Weak</th>
<th>Not cut</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1 (ssGs)</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>x</td>
</tr>
<tr>
<td>A (ssC &amp; ssUs)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

Cleavage relative to the ladder lanes
Figure 3.11. Summary of base-pairing in the \( sxy \), \( sxy-1 \) and \( sxy-7 \) RNA secondary structures.

The cleavage results from different RNase digestions were summarized and superimposed on the Mfold-predicted secondary structures of the three RNA species.

<table>
<thead>
<tr>
<th>Evidence that the residue is</th>
</tr>
</thead>
<tbody>
<tr>
<td>□ Strongly base paired</td>
</tr>
<tr>
<td>□ Partly base paired</td>
</tr>
<tr>
<td>▶ Single stranded</td>
</tr>
</tbody>
</table>

\( sxy \) wt

\( sxy-1 \)

\( sxy-7 \)
3.2.4 Secondary structure of sxy-7 mRNA

The sxy-7 sequence carries two substitutions at positions -32 and -31, designed to increase base-pairing in the stem region by eliminating the bubble (2). Mfold predicts that the resulting perfect 18bp long stem formation will decrease ΔG in this region from -13.3 to -20.6 kcal. To check whether sxy-7 folds into the predicted structure, I used enzymatic mapping of sxy-7 RNA using the same conditions used for the wild type and sxy-1 RNAs (see sections 2.2.2 for details). Both RNase TI and A are predicted not to cut the stem region in the folded sxy-7 RNA, and the control (denatured) sxy-7 RNA should have an additional cut site (G₃₂) when compared to the wild type sxy RNA.

Mapping data of sxy-7 RNA showed that its secondary structure is very similar to the one present in the wild type sxy. As predicted G₃₂ was cut in the control and not cut in the folded RNA suggesting that this residue is base-paired (data not shown; see Fig A1 in Appendix). C₄₉, predicted to base pair with G₃₂, was only weakly cut suggesting that pairing does occur under in vitro conditions. Unfortunately, experimental data were not available for both A₃₁ and U₄₈ sites due to specific features of RNases used in the analysis. Residue U₄₈, predicted to base-pair with A₃₁, was not cut by RNase A, even in denaturing conditions.

Except for residues U₂₈, U₁₈ and G₄₂, all of the sxy-7 mapping data support the predicted structure. These three residues are cleaved with the same intensity in the wild type RNA (Fig. 3.11), suggesting that the substitutions did not affect this region of the sxy-7 structure. Finally, the data confirm that under in vitro conditions the stem region Ia+b exists in the sxy-7 RNA and that except for U₂₈ all of the nucleotides are paired in this
region (see Fig. 3.11). This implies that the stem Ia+b in sxy-7 RNA is stronger than stem Ia+Ib present in wild type sxy supporting the hypothesis that RNA secondary structure limits this gene’s expression.

Overall, RNase mapping data of sxy, sxy-1 and sxy-7 provided evidence that their RNAs fold, corresponding to Mfold predictions, into a similar stem-loop structure in in vitro conditions. Additionally, my data suggest that the point mutations present in sxy-1 and sxy-7 affect the stability of their RNA secondary structures rather than changing the way they fold.
Chapter 4: Discussion

4.1 Co-variance analysis

4.1.1 Pasteurellaceae could carry a regulatory element in the region upstream of sxy homologues

The comparative analysis I have done shows that sxy has homologues in all sequenced Pasteurellacean species. In addition, phenotype of the strain with compensatory mutations designed by combining sxy-1 and sxy-3 hypercompetence mutations in H. influenzae strongly implies that sxy expression is limited by its RNA secondary structure (2). If the regulation of sxy and its homologues is conserved in Pasteurellaceae, we would expect upstream intergenic regions to be similar length to the 5' untranslated region in H. influenzae sxy RNA and to carry a similar regulatory element. Surprisingly, my analysis showed that intergenic regions upstream of sxy homologues are much longer than that of each corresponding average (intergenic region) in every one of the species, with their lengths exceeding 500 nt in most of the species (in H. influenzae the average is about 120 nt long and in other annotated Pasteurellaceae is less than 150). The intergenic regions upstream of sxy homologues are very likely true non-coding regions since no tRNAs, rRNAs or sRNAs were predicted in these regions and high number of stop codons is present throughout the sequences.

Long intergenic regions are unusual for bacterial chromosomes, whose genome compactness is essential for rapid cell division and DNA replication (19,26). Selective
pressure on bacterial chromosomes makes the genomes compact with minimal length and number of non-coding regions (19). Considering the unusual lengths of the intergenic regions upstream of sxy and its homologues it is important to ask whether their length and structure can indicate a role in regulation and reveal more about the function of these regions?

One explanation for the length of the intergenic regions is that Pasteurellacean sxy genes carry one or more additional regulatory element(s) upstream of the gene. Interestingly, sxy in H. influenzae and two other Pasteurellaceae is predicted to include CRP sites in its promoter regions ((40), A. Cameron, personal communication). However, none of these sites has a strong CRP consensus binding sequence, and no experimental evidence is yet available. Testing the expression of Sxy in a strain with the putative CRP site deleted (without otherwise affecting the sxy 5’ RNA untranslated sequence) would show if this portion of the intergenic region is important in sxy regulation.

Another alternative explanation is that the intergenic regions upstream of sxy in Pasteurellaceae actually carry non-coding RNAs (ncRNA) in their sequences. This recently described class of RNA is involved in regulation of gene expression by complementary binding to transcripts (8). ncRNAs are found in all three domains of life but despite the availability of many sequenced genomes, the number and diversity of ncRNAs is still poorly understood. Due to the lack of both experimental evidence and reliable algorithms for detecting ncRNAs, it is believed that most of them are still unidentified. For example, in the E.coli genome over 60 ncRNAs have been experimentally detected, but comparative analysis predicts more than 200 ncRNAs in this genome, suggesting that most of them are still uncharacterized (21). Although ncRNAs
could account, in part, for the length of the intergenic regions upstream of \textit{sxy}
homologues in Pasteurellaceae, there is still neither bioinformatic nor experimental
evidence that they exist in the \textit{H. influenzae} genome.

4.1.2 Regulation of \textit{sxy} is probably not conserved in Pasteurellaceae

Since \textit{sxy} has homologues in three gamma-proteobacterial families and is essential for
competence development in both \textit{H. influenzae} and \textit{V. cholerae} (28), it is parsimonious to
expect that this gene is regulated in the same manner in all of the Pasteurellacean species.
To address this question, I looked for conserved motifs in the alignments of \textit{sxy} and its
homologues and compared predicted secondary structures of their RNAs. The
phylogenetic comparative analysis revealed no common secondary structure in the RNA
of \textit{sxy} homologues in Pasteurellaceae and in the following paragraphs I discuss some
possible implications.

One explanation is that RNAs of \textit{sxy} homologues can assume functionally similar
secondary structure in the presence of a RNA binding protein. Although RNAs of \textit{sxy}
homologues are predicted to fold in different manners it is possible that an RNA binding
protein binds to the untranslated regions of the RNAs and organizes them into a common
secondary structure to modulate \textit{sxy} expression. Some of the RNA binding proteins, such
as Hfq or CsrA are known to affect gene expression by changing their abundance in the
cell in growth depending manner (30, 34, 36). Although both \textit{hfq} and \textit{crsA} are present in
*H. influenzae* genome, there is no experimental evidence that they play role in regulation of the *sxy* gene.

Another explanation is that structures like that of *H. influenzae sxy* RNA are not present in the RNAs of other Pasteurellacean simply because the same mode of regulation is not conserved. The regulation of natural competence is very fluid and different taxa have found different ways to regulate competence (14). It is possible that other Pasteurellaceae species have altered the regulation of *sxy* expression to adapt to their specific environmental conditions. If this is true, the lack of common structures in the RNA of *sxy* homologues would suggest that the specific stem-loop structure in *H. influenzae* evolved recently in this species to regulate *sxy* expression.

Finally, it is possible that the method that I used in looking for a common structure in Pasteurellaceae was too crude to determine the fine structural similarities that might be present in RNAs of *sxy* homologues. One of the limitations of the analysis was in the usage of individual structures produced by Mfold and visual inspection for similarities. Mfold is commonly used for RNA secondary structure predictions, basing those on free energy minimisation of the RNA folding, so that it produces the most stable structure based on the sequence. This is good for general RNA secondary structure predictions; however, for covariance analysis this might not be the optimal choice, because the algorithm doesn’t anticipate information from covarying sites. A better choice would be one of the software packages such as RNAAlign (5) or RAGA (24) that combine comparative analysis with thermodynamic considerations. However, these require the presence of conserved motifs throughout the input sequences, which is lacking in the alignment of *sxy* and its homologues.
Another limitation of the analysis is that, except for *H. influenzae*, there is no experimental evidence that *sxy* homologues have significant untranslated regions in their RNA. Phylogenetic comparative analyses are usually done by comparing the secondary structure predictions of the aligned RNA sequences from related species, where the length of the untranslated region was previously experimentally determined. In this study, however, the comparative analysis was based only on characteristics of *H. influenzae* *sxy* RNA, with the working assumption that the RNAs of *sxy* homologues would have untranslated regions of similar lengths. Mapping the transcriptional start positions of *sxy* homologues in Pasteurellaceae, starting with *A. pleuropneumoniae* or *A. actinomycetemcomitans* (both shown to be naturally competent) would provide a better basis for the phylogenetic comparative analysis.

4.1.3 RNA secondary structure could be important in regulation of *sxy* in *H. influenzae* only

In the absence of a common RNA secondary structure in Pasteurellaceae phylogenetic comparative analyses were repeated using *sxy* sequences of five *H. influenzae* strains. Nucleotide alignment revealed very high sequence conservation, which is insufficiently informative for this type of analysis. However, comparison of the *sxy* RNA secondary structures from different *H. influenzae* strains to the corresponding transformation frequencies suggests that the regulatory mechanism could be shared among the competent strains. Preliminary data (H.Maughan, personal communication) showed that both *H. influenzae* R2866 and 86-028NP are competent, which is consistent with the identity of their *sxy* sequences to the one in KW20. The strain R2844, is not transformable (H.Maughan, personal communication) and has two substitutions that
change the amino acid sequence. The nucleotide substitutions present in the R2846 sequence are predicted (by Mfold) to increase base-pairing in the mRNA which results in a structure with longer stem regions and thus more stable secondary structure in this strain (see figure 3.4). Since these long double-stranded regions precede the translational start, it is likely that the secondary structure limits the translation of the sxy mRNA, resulting in a non-competent phenotype.

Alternatively, the nucleotide substitutions present in the R2844 sequence do not necessarily alter the folding of sxy RNA in this strain: two of the substitutions conserve base-pairing and the third is in a positioned in a loop (Fig 3.4) allowing RNA to assume the same secondary structure as in KW20. In this scenario, sxy RNA of the non-competent R2844 strain would have the same secondary structure as other competent H. influenzae strains suggesting that the substitutions present in R2844 could act by disrupting the Sxy protein function. This, however, is very unlikely since there is only one mutation that causes a significant amino acid change (Gln17 -> Arg) in the sxy sequence. In addition, this substitution is located in the beginning of the coding sequence, the region without conserved motifs in the Pasteurellaceae sxy alignment (see Fig 3.2).

Alternatively, non-competent phenotype in R2846 can be explained by the fact that its sequence is very diverged from that in KW20 and other competent strains (the mean sequence diversity is over 3% (22)). Therefore, the lack of competence in R2844 could also be due to mutations in other competence genes. Comparing expression profiles of KW20 and R2846, using microarray analysis could reveal which gene(s) is (are) responsible for non-transformable phenotype.
4.2 RNA secondary structure mapping

4.2.1 5’ RNA secondary structure may limit the expression of sxy

The RNase mapping analysis supported the hypothesis that the expression of sxy is limited by its RNA secondary structure. The majority of the experimental data from all three enzymes used supports the structure predicted by Mfold suggesting that the stem region Ia + Ib does form in in vitro conditions.

Although the most of the data agrees with the predicted structure, discrepancies do exist. Residues C_{32}, U_{31} and C_{+49}, predicted to form a small bubble in the stem region, yet are all moderately-strongly cut by the double-strand specific RNase VI. Likewise, residues G_{+42} and U_{+8}, predicted to be single-stranded, are only weakly cut by the single-strand specific RNases TI and A. Contradictory evidence also exists for residues at positions -28 and +21. Both of these residues are cleaved both with RNase VI (ds specific) and RNase A (specific for ss pyrimidines).

Despite these discrepancies, the experimental data do not necessarily contradict the computationally predicted structure. RNase VI has specificity for double-stranded regions in RNA, but it also cuts single-stranded nucleotides in stacked conformations or interacting in tertiary structures (9), which could explain the cleavage at positions -28, -31, -32, +21, +48 and +49. However, in previous results by L.Bannister, residue C_{+49} was cut both in folded sxy and sxy-1 RNA by RNase CL3 (specific for single stranded Cs). These contradictory data illustrate one of this method’s limitations, where modest variation of reaction conditions may result in an alteration to the digestion pattern.
Residues U+8 and G+42, are both positioned between two long double-stranded regions in the predicted structure and it is likely that steric hindrance (physical blockage) makes these nucleotides inaccessible to the RNase A or T1 digestion.

Overall, the sxy-1 RNA mapping data supported the hypothesis that the point mutation in sxy-1 mutant has a destabilizing effect on the secondary structure and, therefore, results in sxy overexpression. Surprisingly, the data showed that the effect of the point mutation is very broad and that it weakens the stem Ia+Ib in sxy RNA globally. This global destabilization of Ia+Ib stem could facilitate translation of sxy RNA explaining the increased levels of Sxy protein and the hypercompetent phenotype in the sxy-1 mutant. This point will be discussed in greater detail in section 4.3.1.

Mapping data for sxy-7 RNA also supported the predicted secondary structure. However, except for evidence of G32:C+49 base-pairing, it provided no direct additional data that this region is more strongly paired than it is in the sxy RNA secondary structure. Is the effect of this single additional base-pairing strong enough to fully inhibit sxy expression and therefore competence induction? My analysis indirectly showed that the residues flanking the G32:C+49 bp are paired (i.e. resistant to single-stranded-specific RNases), suggesting that the mutations may act by stabilizing stem Ia+Ib in general. It is possible that the longer and more stable stem region present in sxy-7 RNA limits translation and therefore prevents competence. In addition, Bannister’s expression data showed that in the sxy-7 mutant both sxy RNA and protein levels are drastically decreased (2).

To test further if sxy expression is limited by its RNA secondary structure, directed mutagenesis of the stem Ia+Ib could be performed to determine which residues are the most important for sxy regulation. Also nuclease mapping of sxy RNA secondary
structure in other hypercompetent and non-competent sxy mutants should be tested to confirm that the hypercompetence mutations act by global destabilization of stem Ia+Ib.

4.2.2 Is the described sxy RNA secondary structure present in in vivo conditions?

Although RNase mapping provides insight into the status of each residue in folded RNA, it also has obvious limitations. The results of an in vitro RNase mapping analysis are known to depend on reaction conditions (9). For example, a change in pH, temperature or divalent ion concentration in the reaction can alter the RNA cleavage pattern. Additionally, cleavage at a specific position in RNA can result in a secondary structure change and promote new, artifactual cuts in the same molecules. These “secondary” cuts, although weaker than the primary ones, can only be identified by comparing cleavage patterns of 5’ and 3’ labeled RNA molecules (9). With these concerns in mind one can ask if the sxy RNA structure mapped in vitro by RNases is real, if it exists in vivo, and to what extent the reaction conditions affect the folding of RNA.

Several lines of evidence support that sxy RNA folds similarly to the predicted structure. First, restoration and addition of base-pairs in the predicted stem Ia +Ib of sxy RNA results in a non-competent phenotype (2), providing very strong genetic evidence that the stem forms in vivo and has a function in sxy regulation. Additionally, the nuclease mapping done with 7 different enzymes, (both by L.Bannister and me), all supported the predicted sxy secondary structure and confirmed that the stem Ia+Ib exists. Although RNase mapping data in general fits well to the predicted sxy RNA secondary structure it would seem possible that an alternate structure exists. However, upon closer examination,
the alternate structures predicted by Mfold at both 25 and 37°C did not fit the data as well as the thermodynamically most favorable one (produced at 37°C, data not shown).

4.3 Possible mechanisms of \textit{sxy} regulation

How is \textit{sxy} regulated? We know that addition of ribonucleotides (or their derivatives) to the medium decreases transformation frequency (16) by directly affecting \textit{sxy} expression (A. Cameron, personal communication). The secondary structure of \textit{sxy} RNA is likely to limit the translation of the \textit{sxy} message: mutations that weaken the structure strongly increase Sxy expression and have only small impact on the RNA levels. Additionally, the strains carrying mutations that introduce base-pairing in the \textit{sxy} RNA secondary structure produce almost no transcript or Sxy protein (A. Cameron, personal communication, (2)). In the following sections, I discuss two possible mechanisms of \textit{sxy} regulation in \textit{H. influenzae}.

4.3.1 \textit{sxy} expression could be regulated via translation initiation by \textit{sxy} RNA secondary structure

Translational control via RNA secondary structure alteration is a common mechanism of gene regulation in bacteria (6). Gene expression in this case is limited by sequestering the translation initiation region via RNA secondary structure (7). The translation initiation region, which includes the ribosome binding site (RBS) and the translational start codon, is bound by ribosomes in unstructured (single-stranded) mRNA and is subsequently translated. Alternatively, protection of the ribosome binding site and start codon by RNA
secondary structure prevents ribosomes from binding and initiating translation of the RNA.

It is possible that nucleotide concentration in the cell directly regulates sxy expression by changing the rate of transcription elongation and accessibility of the ribosomal binding site or the start codon in sxy RNA. In this model, under non-inducing conditions (early log phase in sBHI), the high abundance of nucleotides in the cell allows steady transcription elongation and rapid progression of RNA polymerase (Fig 4.1 A). This fast elongation of sxy message allows forming of a strong secondary structure that blocks the ribosome from binding and initiating the translation. Under the inducing conditions (late log phase in sBHI), transcription elongation may proceed at a lower rate or stall due to nucleotide depletion in the cell (Fig 4.1 B). Pausing of RNA polymerase due to low concentration of the precursors could prevent formation of the secondary structure in the sxy transcript and allow a ribosome to bind the RBS in sxy RNA.

One piece of evidence that supports this model is an increase of both sxy RNA and protein abundance in hypercompetent sxy mutants. In these strains, destabilization of RNA secondary structure may allow ribosome binding and initiation of translation even in non-inducing conditions. It has been previously shown experimentally that the translational efficiency is directly related to the stability of RNA secondary structure in the region of RBS and the translation start codon (6). de Smit (1994) showed that a single point mutation in the RNA sequence can drastically alter its secondary structure, thus, the protein expression.
Figure 4.1. Possible mechanisms of regulation of sxy expression. For details refer to text in Section 4.3.
4.3.2 *sxy* could be regulated by a riboswitch

Riboswitches are cis-regulatory elements found in untranslated regions of RNAs. (18). They control gene expression by binding small molecules (ligands) with high specificity without protein interaction. Depending on the concentration of the ligand in the cell, RNA folds into alternate structures affecting the transcription or translation of the downstream gene (25). Nine distinct riboswitch classes have been described and each has distinct consensus sequence, ligand specificity and highly conserved secondary structure throughout different bacterial taxa (17).

Four putative riboswitch sequences have been detected in the *H. influenzae* genome by bioinformatics analysis but none of them are similar to the *sxy* RNA nucleotide sequence or secondary structure (www.sanger.ac.uk/cgi-bin/Rfam/). However, production of both *sxy* mRNA and protein is directly limited by purine nucleotides (A. Cameron, personal communication) and *sxy* RNA folds into a complex stem-loop structure suggesting that *sxy* expression could be riboswitch-regulated.

In this model *sxy* RNA would directly sense a change of ligand (i.e. nucleotide) concentration, and the binding would alter its secondary structure and its expression. In non-inducing conditions (early log phase in sBHI ) high ligand concentration would result in formation of translation-inhibiting *sxy* RNA secondary structure (Fig 4.1 C), while in the inducing conditions, the reduced concentration of the ligand would allow folding into an alternate structure that allows translation (Fig 4.1 D).

This model can be tested by performing RNase mapping analysis of *sxy* RNA secondary structure both in the presence and absence of different potential ligands. Alteration of the
secondary structure can be detected by comparing the results from these two conditions. Since nucleotides cannot be transported into the cell without their dephosphorylation in the periplasm (12), a range of concentrations of purine nucleosides and bases should be tested first.

Although this model provides a reasonable explanation of \(\text{sxy}\) regulation it also has some limitations. First, my analysis suggested that RNAs in \(\text{sxy}\) homologues are lacking both common secondary structure and conserved motifs in their sequences. Additionally, the \(\text{sxy}\) sequence does not match any of the concensus sequences of described riboswitch classes. However, currently used bioinformatics programs for riboswitch detection also have some limitations. These programs identify new riboswitch sequences throughout different genomes by comparing them to the conserved motifs and secondary structures of the known riboswitches classes. Thus, it is possible that some of riboswitch sequences have recently evolved to accommodate specific life styles of bacteria but due to the current methods for riboswitch detection none have been yet recognized.

**Conclusion**

To summarize, the RNA mapping analysis that I did provided support that \(\text{sxy}\) expression is limited by its RNA secondary structure. Additionally, the phylognetic comparative analysis suggested that the RNA secondary structure is important only in regulation of \(\text{sxy}\) in \(H. \text{influenzae}\) strains and that is not conserved in other Pasteurellaceae. Further experiments should be conducted to determine the precise regulation and role of this major competence regulator.
References:


Appendix

<table>
<thead>
<tr>
<th>Species</th>
<th>Accession number</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>H. influenzae</em> KW20</td>
<td>L42023.1</td>
</tr>
<tr>
<td><em>H. influenzae</em> 86-028NP</td>
<td>CP000436.1</td>
</tr>
<tr>
<td><em>H. ducreyi</em></td>
<td>AE017143.1</td>
</tr>
<tr>
<td><em>H. somnus</em></td>
<td>CP000436.1</td>
</tr>
<tr>
<td><em>P. multocida</em></td>
<td>AE004439.1</td>
</tr>
<tr>
<td><em>M. succiniciproducens</em></td>
<td>AE016827.1</td>
</tr>
</tbody>
</table>

*Table A.1. Accession numbers of annotated genomes used in the study*
Figure A.1 Representative gel pictures. End-labeled sxy, sxy-1 and sxy-7 RNAs were partially digested with RNases and the resulting fragments were resolved in a sequencing gel. A. Mapping of sxy RNA secondary structure. Lanes 1, 2 and 3 were loaded with ladders: 1-alkaline ladder, 2 & 3-TI ladder. Lanes 4 and 5 were loaded with folded sxy RNA partially digested with RNase TI (lane 4) and RNase VI (lane 5). B. Mapping of sxy-1 and sxy-7 RNA secondary structures. Lanes 1-6 were loaded with partially digested sxy-1 RNA and lanes 7-15 were loaded with partially digested sxy-7 RNA. Lanes 1,4,5,7,8,12 & 13 represent sequencing ladders: 1, 8 & 9-TI ladder, 4,5,12 & 13 RNase A ladder and 7-alkaline ladder. Samples partially digested with RNase TI are in lanes 2,3,10 & 11 and samples partially digested with RNase A are in lanes 6, 14 & 15.
Cleavage intensities of residues in *sxy*, *sxy*-1 and *sxy*-7 RNA secondary structure

Figure A.2. Relative cleavage intensities for residues in *sxy*, *sxy*-1 and *sxy*-7 RNA.