Computational detection of non-coding RNAs in genomes

Yen-Hua Huang

This dissertation is submitted for the degree of Doctor of Philosophy

The Wellcome Trust Sanger Institute and Churchill College, Cambridge

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

The work in this thesis has not been submitted in whole, or in part, for a degree, diploma, or any other qualification at any other university.

Yen-Hua Huang

March 2008, Cambridge, UK

Abstract

Noncoding RNAs (ncRNAs) have become implicated in a variety of regulatory mechanisms as well as structural roles, suggesting that functional ncRNAs may be more prevalent in genomes than previously supposed. Nonetheless, *in silico* ncRNA finding is difficult, even though a mass of genome sequence is publicly available. Few computational approaches are really reliable for genome-wide ncRNA finding. This thesis is devoted to assessing available approaches and trying new solutions for finding ncRNAs in genomes.

In the first half of this thesis, reasons that may contribute to the slow progress of genome-wide ncRNA finding are explored. A comprehensive analysis on a genome-wide scale of the credibility of currently used signals for classifying ncRNAs is conducted. Two factors, conservation of ncRNAs in human-mouse syntenic regions and abundance of covariations between human-mouse synteny-conserved ncRNAs, are evaluated. The result reveals that current comparative-genomics-based methods may not be able to find ncRNAs effectively in mammalian genomes. In addition, possible genomic features that could distinguish real ncRNAs from pseudogenes are investigated. Two different criteria, distribution of bit scores and physical clustering in genomes, are applied to filter out tRNA pseudogenes and to enrich *bona-fide* tRNA genes. Physiological roles of the tRNA genes in human-mouse synteny-conserved clusters are discussed and the degradation patterns of tRNA pseudogenes are analyzed.

In the second half of this thesis, computational techniques are applied to model signals that may be potentially useful for genome-wide ncRNA finding. A sparse Bayesian learning algorithm, Eponine, is applied to model the transcription start sites of mammalian ncRNA genes that are transcribed by RNA polymerase III. In addition to modelling *cis*-regulatory elements for transcription, a new computational module, which extends the capability of

Eponine to learn motifs consisting of both primary sequences and RNA secondary structures, is created. The capability of this new module is demonstrated by applying it to analyze several known cases of ncRNA motifs. The strength and the weakness of applying this new computational approach for finding ncRNAs are discussed.

Acknowledgements

I thank everyone who has influenced me during this project. I am particularly grateful to my supervisor, Dr. Tim Hubbard, for giving me this opportunity to work on this project in the Wellcome Trust Sanger Institute. I greatly appreciate his support and advice, which have helped me extend many originally hopeless paths of this project.

Being a bioinformatics PhD student with a pure background of biomedical science, I encountered many difficulties in understanding the mathematics required for my project. I am so fortunate that I worked with Thomas Down. His guidance enabled me to apply machine-learning techniques to my project. I greatly benefited from the discussion with him and the development of the Eponine RNA-motif extension was based on the Eponine and related codes he has created.

I thank Aroul Selvam Ramadass for the stimulus he provided. I was therefore inspired to investigate human-mouse synteny-conserved tRNA gene clusters and to create the Eponine RNA-motif extension. Without the discussions with him, the related sections might not exist.

I thank Jenny Mattison, Thomas Down, Mutlu Dogruel, and Jing Su for reading and checking this manuscript.

I thank Andreas Prlic for his helps on my project and for making the work interesting.

I thank the Wellcome Trust Sanger Institute for the funding and Churchill College, Cambridge, for academic support.

Finally, I would like to express my best gratitude to my wife, Yu-Yen, who accompanies me with love through all the difficulties and joys these years.

Table of Contents

Declaration i

Abstract ii

Acknowledgements iv

Table of Contents v

List of Tables vii

List of Figures x

Chapter 1. Introduction 1

1.1. What are ncRNAs 2

1.2. RNA structures 6

1.2.1. RNA secondary-structure motifs 6

1.2.2. RNA tertiary structures 9

1.2.3. The dynamic aspect of RNA structures 11

1.2.4. The definition of “RNA motifs” used in this thesis 12

1.3. Prediction of RNA structures 12

1.3.1. Minimization of free energy (MFE) 13

1.3.2. Phylogenetic covariation analysis 16

1.3.3. Grammatical approaches for RNA sequence analysis 17

1.4. Current state of genome-wide ncRNA finding 25

1.4.1. Few statistically useful features for classifying ncRNAs 27

1.4.2. Assumptions made in previous work 28

1.4.3. Few appropriate data sets for training ncRNA-finding algorithms 29

1.5. Objectives of this project 30

Chapter 2. Constraints from comparative genomics for ncRNA finding 32

2.1. The conservation patterns of vertebrate ncRNAs 35

2.1.1. Materials and Methods 35

2.1.2. Evaluating different approaches for finding human-mouse synteny-conserved ncRNAs 42

2.1.3. Results 44

2.1.4. Discussions 55

2.2. Gene-order conservation of mammalian tRNA genes 59

2.2.1. Materials and methods 59

2.2.2. Results 67

2.2.3. Discussions 82

2.3. Summary 92

Chapter 3. Distinguishing functional ncRNAs from pseudogenes in mammalian genomes 96

3.1. Are Rfam synteny-non-conserved tRNA genes functional? 101

3.1.1. Materials and methods 101
3.1.2. Results 102
3.1.3. Discussion 117
3.2. Clustering – a useful criterion for filtering out ncRNA pseudogenes? 119
3.2.1. Materials and methods 119
3.2.2. Results 120
3.2.3. Discussion 124
3.3. Summary 124
Chapter 4. Modelling functional elements associated with ncRNAs 127
4.1. Computational detection of transcription regulatory regions 128
4.1.1. Computational detection of over-represented motifs 130
4.1.2. Computational detection of functional sites 138
4.2. Modelling local RNA motifs 147
4.2.1. Available methods for finding consensus RNA motifs in sequences 148
4.2.2. Extending Eponine to include RNA structural motifs 152
4.3. Summary 162
Chapter 5. Modelling the transcription regulatory elements of mammalian RNA polymerase III genes 164
5.1. Modelling the transcription start sites of mammalian pol III type II genes 166
5.1.1. Materials and methods 168
5.1.2. Results 173
5.1.3. Discussion 189
5.2. Summary 191
Chapter 6. Finding RNA motifs in genomes 192
6.1. Using the Eponine RNA-motif extension 193
6.1.1. Modelling RNA-motifs of mammalian tRNAs 193
6.1.2. Modelling *rho*-independent transcription termination 204
6.1.3. Modelling pseudoknots 219
6.2. Discussions 222
6.2.1. Considerations of using the Eponine RNA-motif extension 222
6.2.2. Towards creating general EWR models of vertebrate ncRNAs 224
6.3. Summary 225
Chapter 7. Conclusions 227
Appendix A . Tables related to the investigation of tRNA-gene order conservation in mammalian genomes 230
Appendix B. The program sets written for this thesis 238
Reference 243

List of Tables

Table 2-1. Conservation of different classes of Rfam human ncRNAs in human-mouse syntenic regions 45

Table 2-2. Distribution of the human synteny-non-conserved ncRNAs in the regions corresponding to mouse finished contigs or UR-WGS-containing regions (regions with unfinished gaps in contig-base sequencing and regions from whole genome shotgun sequencing) 46

Table 2-3. Distribution of human synteny-conserved ncRNAs in the regions corresponding to mouse finished contigs or UR-WGS-containing regions (regions with unfinished gaps in contig-base sequencing and regions from whole genome shotgun sequencing) 46

Table 2-4. Numbers of the human-mouse synteny-conserved and the synteny-non-conserved ncRNAs in regions which have undergone different evolutionary events 50

Table 2-5. Numbers of the human-mouse synteny-conserved ncRNAs that contain various numbers of covariations 51

Table 2-6. Average numbers of bases involved in covariations per sequence of the human-mouse synteny-conserved ncRNAs and of the human-zebrafish orthologous ncRNAs 51

Table 2-7. Numbers of the human-mouse-zebrafish orthologous ncRNAs that contain various numbers of covariations 52

Table 2-8. Estimating sensitivities of ncRNA-finding algorithms by using the alignments of genomic sequences of human tRNA genes 55

Table 2-9. The statistics of clustered tRNA gene loci in the human, mouse, and opossum genomes 67

Table 2-10. The synteny conservation of clustered human tRNA gene loci 67

Table 2-11. Transitions of the anticodons of tRNA gene loci 70

Table 2-12. The statistics (aligned and inserted regions) of the human-mouse tRNA symbol alignments 71

Table 2-13. The statistics of the gene-order conservation of human and mouse tRNA gene clusters 71

Table 2-14. Relation of synteny-conservation of tRNA gene clusters and the quality of the mouse genome assembly FCS: finished contig sequence; CSN: unfinished contig sequence (with gaps); WGS: whole genome shotgun sequence 73

Table 2-15. Relation of synteny-conservation of non-clustered tRNA genes (singlets) and the quality of the mouse genome assembly 73

Table 2-16. Evolutionary origin of the insertions in the human-mouse tRNA symbol alignments 75

Table 2-17. Evolutionary origin of the deletions in the human-mouse tRNA symbol alignments76

Table 2-18. Local-duplication associated insertions in the human-mouse tRNA symbol alignments 82
Table 3-1. Numbers of the human low-scoring tRNA genes which are more similar to either the human nuclear tRNA genes or the human mitochondrial tRNA genes........ 109
Table 3-2. Comparison between types of anticodons of yeast and the human tRNAs 122
Table 3-3. Comparison between types of anticodons of yeast and mouse tRNAs........ 123
Table 4-1. Performance of different algorithms for three hairpins of 168 human tRNAs........ 157
Table 5-1. The TFs and the TFBSs associated with three distinct types of eukaryotic pol III genes 167
Table 5-2. The training and test data sets for creating an EAS model for pol III type II TSSs .. 171
Table 5-3. Ratios of MIRs in different predictions for pol III type II genes on human chromosomes 11 and 13 182
Table 5-4. Ratios of MIRs in the predictions made models 1 and 2 for pol III type II genes on human chromosomes 11 and 13 186
Table 5-5. The synteny conservation of the non-tRNA pol III type II signals on human chromosomes 11 and 13 188
Table 5-6. Distributions of the synteny-conserved pol III type II promoter signals in intronic and exonic regions 189
Table 6-1. The training and test data sets for modelling the human tRNAs 194
Table 6-2. The trained parameters of an anchored RNA structural model for mammalian tRNAs by using the stringent mode for locating local hairpins 196
Table 6-3. The trained parameters of an anchored RNA structural model for mammalian tRNAs by using the fast mode for locating local hairpins 197
Table 6-4. The trained parameters of the EAS mixed model presented in Figure 6-2........ 199
Table 6-5. The high-scoring false positives predicted by using the mixed model of human tRNAs 203
Table 6-6. The trained parameters of an EAR model for *bacillus rho*-independent transcription terminators 209
Table 6-7. Comparison of the performance of different algorithms in finding *rho*-independent transcription terminators in *B. subtilis* 212
Table 6-8. The trained parameters of an EWR model for *bacillus rho*-independent transcription terminators 216
Table 6-9. The trained parameters of an EWR model for pseudoknots in 3' UTRs of viral genes 221
Table 6-10. The execution time for training the EAR and the EWR models of tRNAs and rho-independent transcription terminators 223
Table 6-11. The execution time for using the EAR model of *rho*-independent transcription terminators to scan the genomes of *B. subtilis* and *E. coli* respectively 223

Table A 1. Lookup table of anticodon types and the tRNA-gene symbols 230
Table A 2 The start and end coordinates of the tRNA gene clusters in the human genome (assembly NCBI 36)..231
Table A 3. The start and end coordinates of the tRNA gene clusters in the mouse genome (assembly NCBI M36). 232
Table A 4. The synteny conservation of clustered human tRNA gene loci in the mouse genome 234
Table A 5. The synteny conservation of non-clustered human tRNA gene loci (singlets) in the mouse genome..237

Table B 1. Functions of the program sets written for this thesis...238
Table B 2. Number of lines and file sizes of the program sets written for this thesis................241

List of Figures

Figure 1-1. Organization of repeating units in RNA and DNA respectively. 3

Figure 1-2. Elements of RNA secondary structures...7

Figure 1-3. The cloverleaf-like secondary structure of a tRNA...9

Figure 1-4. Non-nested base pairs in a pseudoknot 10

Figure 1-5 Two representations of the pairwise correlations in an RNA molecule with two non-interlaced hairpins...20

Figure 1-6. A crossing interaction that may be found in RNA tertiary structures.........................24

Figure 2-1. Physical relations of human and mouse synteny-conserved ncRNAs to UBRHPs-bound syntenic regions 40

Figure 2-2. A multi-sequence secondary-structure alignment generated by cmalign...................41

Figure 2-3. Synteny-conservation ratios and average copy numbers for different categories of human ncRNAs (mapped by Rfam)...48

Figure 2-4. The procedure of identifying the syntenic tRNA gene clusters in mammalian genomes 61

Figure 2-5. Different types of tRNA gene-order conservation 65

Figure 2-6. The conservation pattern of the human tRNA gene clusters 4.1.36 and its syntenic cluster in the mouse genome (see next page)...68

Figure 2-7. Summary of the synteny conservation of human and mouse tRNA gene loci...........72

Figure 2-8. The conservation pattern of human tRNA gene cluster 3.1.42 and its syntenic clusters in the mouse and opossum genomes...80

Figure 2-9. the synteny conservation of human non-clustered tRNA gene loci in the syntenic regions of other mammalian genomes 85

Figure 2-10. The structural alignment of a human tRNA gene locus and its syntenic (but degraded) counterpart in the mouse genome 87

Figure 2-11. The structural alignment of a mouse tRNA gene locus and its syntenic (degraded) counterpart in the human genome...88

Figure 3-1. Comparison of the gene structures of a retrotransposed protein gene and a hypothetical retrotransposed ncRNA that contain internal promoters..99

Figure 3-2. Distributions of Rfam bit scores of tRNA genes of different categories 103

Figure 3-3. Distributions of numbers of the non-canonical base pairs in human tRNA genes... 106

Figure 3-4. Distributions of Rfam bit scores of tRNA genes of human-numt, Rfam-human, and tRNAscanSE tRNA genes.. 111

Figure 3-5. Distribution of identities of human numt-tRNAs and human non-tRNA numt-seqs in 80-90 percent identity regions to the human mitochondrial genome.. 114

Figure 3-6. Patterns of substitution in the human numt-tRNAs and in the human non-tRNAs embedded in regions with different percent identities to the human mitochondrial genome..... 115

Figure 3-7. Distribution of mutation numbers along human numt-tRNAs 116
Figure 3-8. Distribution of the Rfam bit scores of the human U6-like sequences identified by Rfam 4.1 118
Figure 3-9. The human low-scoring tRNA genes are enriched with non-clustered ones 121
Figure 4-1. The transcription initiation of mammalian tRNA genes is regulated by A and B boxes 138
Figure 4-2. How to calculate the score of a CSBF consisting of three PWMs and associated position distributions 146
Figure 4-3. Two modes (algorithm A: the stringent mode and algorithm B: the fast mode) for finding local hairpins for windowed regions 156
Figure 5-1. Separation of the sequence identity distributions between intra-group and inter-group sequences of tRNA genes. 170
Figure 5-2. An EAS model for pol III type II promoters (naïve training) 174
Figure 5-3. The sequence logos of position-constrained motif matrices of the naïve EAS model (Figure 5-2) for pol III type II promoters 175
Figure 5-4. Comparison between the sequence logos of the 8th-22nd positions of VARNA1s (left) and tRNAs (right) 175
Figure 5-5. Comparison between sequence logos of the presumable internal promoter regions of VARNA1s (left) and tRNAs (right) (after adjusting anchoring points of VARNA1s) 176
Figure 5-6. An EAS pol III type II promoter model (after adjusting the anchoring points of VARNA1s) (model 1) 177
Figure 5-7. The sequence logos of position-constrained motif matrices of model 1 (Figure 5-6) 178
Figure 5-8. C-A plots of model 1 and model 2 on the test data set 180
Figure 5-9. Intersection of the tRNA predictions made by different approaches (tRNAscanSE, eufindtRNA, the EAS pol III type II promoter model: model 1) for human chromosome 11 181
Figure 5-10. Intersection of the tRNA predictions made by different approaches (tRNAscanSE, eufindtRNA, the EAS pol III type II promoter model: model 1) for human chromosome 13 182
Figure 5-11. An EAS pol III type II model (using MIRs as negative training sequences) (model 2) 184
Figure 5-12. The sequence logos of position-constrained motif matrices of model 2 (Figure 5-11) 186
Figure 5-13. Intersection of the tRNA predictions made by different approaches (tRNAscanSE, eufindtRNA, the EAS pol III type II promoter model: model 2) for human chromosome 11 186
Figure 5-14. Intersection of the tRNA predictions made by different approaches (tRNAscanSE, eufindtRNA, the EAS pol III type II promoter model: model 2) for human chromosome 13 187
Figure 6-1. Two Eponine anchored RNA structural models for mammalian tRNAs 198
Figure 6-2. An Eponine anchored and mixed (primary-sequence and RNA structural) model .. 199

Figure 6-3. The sequence logos of position-constrained motif matrices in the Eponine EAS mixed model presented in Figure 6-2 and Table 6-4 200
Figure 6-4. Comparison of performances among models trained by different modes for classifying human tRNA genes from random genomic sequences 201
Figure 6-5. Preparation of a set of unanchored sequences that contain *rho*-independent transcription terminators at random positions 207
Figure 6-6. An EAR model for *rho*-independent transcription terminators 209
Figure 6-7. The sequence logos of the position-constrained motif matrices presented in Figure 6-6 and Table 6-6 210
Figure 6-8. Comparison between the C-A plots of the mixed, the structure-only, and the primary-sequence-only models of *rho*-independent transcription terminators 211
Figure 6-9. An EWR model for *rho*-independent transcriptional terminators 215
Figure 6-10. The sequence logos of position-constrained motif matrices presented in Table 6-8 and Figure 6-9 217
Figure 6-11. Comparison of the C-A plots of an EAR mixed model and an EWR model for *rho*-independent transcription terminators 218
Figure 6-12. An EWR model for the 3’ UTRs of viral genes 221