



Computational detection of gene regulatory signals in human genome sequence

A dissertation submitted to the University of Cambridge
for the degree of Doctor of Philosophy

by

Aroul Selvam Ramadass

Trinity Hall, University of Cambridge

and

The Wellcome Trust Sanger Institute, Hinxton, Cambridge

July 2004

DECLARATION

This dissertation is my own work and contains nothing which is the outcome of the work done in collaboration with others, except as specified in the text and acknowledgements.

The work in this thesis is not substantially the same as any I have submitted for a degree or diploma or other qualification at any other university.

Aroul Selvam Ramadass

July 2004

SUMMARY

Transcription, the first step in gene expression, is initiated from a transcription start site and terminated some distance downstream of the cleavage site. In this thesis I attempt to identify and model different regulatory signals involved in the process of transcription, towards the development of a signal based *ab initio* gene predictor.

First I attempt to identify regulatory signals in the sequence downstream of the cleavage site that may be responsible for transcription termination. Base compositional analyses reveal no significant bias in the nucleotide composition. An investigation based on free-energy minimisation Zuker algorithm indicates the possibility of a secondary structure in the sequence downstream of the cleavage site. A probabilistic machine learning algorithm based on Bayes theorem and Generalised Linear Models, Eponine, used to scan for motifs, learns a model to classify termination sites from other sequences. The model captures a few multiplex signals that might be responsible for polymerase II pause and termination. An evaluation of this termination model against annotated human chromosomes shows that the model performs better than existing methods. However a significant number of predictions also appear near the annotated start site of genes. Approximately 10% of predictions lie within genes and their density is correlated with gene length and intron size. I propose two hypotheses to explain these anomalies and discuss results from recent experiments.

Splicing is now found to be interlinked temporally and spatially with transcription and I attempt to develop a donor and acceptor site model using Eponine. Comparisons of the models with annotated sites show the models have higher positional accuracy and perform comparably with existing programs, GeneSplicer and StrataSplice.

Like transcription, translation machinery is influenced to a great extent by regulatory signals and I investigate them by scanning for motifs around translation start and stop sites using Eponine. The start model learnt only the regulatory elements and not the coding potential of exons. Despite this it performs better than the existing program NetStart, although less well than the program ATGpr.

The availability of these models creates the possibility to build an *ab initio* gene prediction program based purely on gene regulatory signals.

ACKNOWLEDGEMENTS

I thank my supervisor, Tim Hubbard, for giving an opportunity to work on this project in The Wellcome Trust Sanger Institute. I very much appreciate and am grateful to his advice and support he has given during this project. I thank Thomas Down for his help and the *Eponine* trainer that facilitated this study. Raphaël Lepläe and Mathew Pocock were grateful and patient in introducing me to the programming concepts. I thank them for this and all the help they gave me during the years I worked with them. Discussing ideas with Samiul Hasan, Yen-Hua Huang and Bernard Leong were always stimulating and I thank them for making the work interesting.

This project would not have been possible without the hard work of annotation and curation by members of Ensembl, VEGA and Sanger human Chromosome 22 group. I appreciate them for this excellent job and making the data available in the public domain.

I thank Rajkumar Sasidharan for all the support and collaboration he rendered in the ‘Domain Insertion’ project mentioned in the appendix. During my years of stay in Cambridge, I made numerous friends and it would be unfair if I mention only few. I thank every one of them for making my life enjoyable and happy.

I am grateful to Cambridge Commonwealth Trust and The Wellcome Trust Sanger Institute for funding and Trinity Hall College, Cambridge, for academic support.

I thank Sudhakaran Prabakaran, Jane E Swatton and James M Carr for correcting the manuscript.

Finally I would like to extend my gratitude to my parents, brother and sister for their love and support all these years.

CONTENTS

| | |
|---|------------|
| DECLARATION | ii |
| SUMMARY | iii |
| ACKNOWLEDGEMENTS | iv |
| CONTENTS | v |
| LIST OF ABBREVIATIONS | x |
| LIST OF TABLES | xii |
| LIST OF FIGURES | xiv |
| INTRODUCTION | 1 |
| 1.1 Motivation | 1 |
| 1.2 An overview of gene structure | 3 |
| 1.3 Defining transcription termination | 4 |
| 1.4 Transcription termination in prokaryotes | 5 |
| 1.5 Transcription termination in eukaryotes | 6 |
| 1.5.1 Polymerase I transcription termination | 6 |
| 1.5.2 Polymerase III transcription termination | 8 |
| 1.5.3 Polymerase II transcription termination | 8 |
| 1.5.4 Computational detection of transcription termination signals ... | 12 |
| 1.5.5 Transcription termination models | 14 |
| 1.6 Splicing and transcription | 15 |
| 1.6.1 Splicing mechanism | 15 |
| 1.6.2 Roles of splicing | 19 |
| 1.6.3 Computational detection of splicing signals | 22 |
| 1.7 Transcription and translation | 23 |
| 1.7.1 Translation mechanism | 25 |
| 1.7.2 Computational detection of translation signals | 29 |

| | | |
|--|--|-----------|
| 1.8 | Objectives of this project | 31 |
| MATERIALS AND METHODS..... | | 32 |
| 2.1 | Introduction | 32 |
| 2.2 | Hidden markov models | 34 |
| 2.3 | Eponine | 36 |
| 2.4 | Modifying Eponine parameters..... | 40 |
| 2.4.1 | Distribution | 40 |
| 2.4.2 | Position weight matrix | 41 |
| 2.5 | Nussinov algorithm..... | 42 |
| 2.6 | Zuker algorithm..... | 44 |
| 2.7 | Biojava and Bioperl..... | 46 |
| 2.8 | Databases | 47 |
| 2.8.1 | ENSEMBL..... | 47 |
| 2.8.2 | VEGA..... | 48 |
| 2.8.3 | RefSeq | 49 |
| 2.9 | Other programs | 49 |
| 2.9.1 | ERPIN..... | 50 |
| 2.9.2 | GAZE..... | 51 |
| 2.10 | Concluding remarks..... | 53 |
| MODELLING TRANSCRIPTION TERMINATION SIGNALS | | 54 |
| 3.1 | Introduction | 54 |
| 3.2 | Datasets..... | 54 |
| 3.3 | Nucleotide composition analysis..... | 56 |
| 3.4 | Secondary structure analysis..... | 58 |
| 3.4.1 | Nussinov algorithm..... | 58 |
| 3.4.2 | Zuker algorithm..... | 60 |
| 3.5 | Eponine transcription termination model..... | 64 |

| | | |
|--|---|------------|
| 3.5.1 | Training the transcription termination model..... | 65 |
| 3.5.2 | Window size | 69 |
| 3.5.3 | Cross validation | 69 |
| 3.5.4 | Model refinement..... | 71 |
| 3.6 | Performance of the model..... | 74 |
| 3.7 | Positional accuracy of the model..... | 79 |
| 3.8 | Internal predictions | 82 |
| 3.9 | GO correlation | 85 |
| 3.10 | Predictions near annotated gene start sites..... | 88 |
| 3.11 | Hypotheses..... | 91 |
| 3.12 | Concluding remarks | 95 |
| MODELLING DONOR AND ACCEPTOR SITES | | 97 |
| 4.1 | Introduction | 97 |
| 4.2 | Datasets..... | 98 |
| 4.3 | Training the splice site models | 99 |
| 4.4 | Refining the models | 101 |
| 4.5 | Validating and testing the models | 108 |
| 4.6 | Position accuracy of the models | 110 |
| 4.7 | Comparison with other models | 112 |
| 4.8 | Concluding remarks..... | 117 |
| MODELLING TRANSLATION START AND STOP SITES..... | | 119 |
| 5.1 | Introduction | 119 |
| 5.2 | Datasets..... | 120 |
| 5.2.1 | Translation start model..... | 120 |
| 5.2.2 | Translation stop model..... | 122 |
| 5.3 | Training the translation models | 123 |
| 5.3.1 | Translation start model..... | 123 |

| | | |
|---|---|------------|
| 5.3.2 | Translation stop model..... | 128 |
| 5.4 | Validating and testing the models | 131 |
| 5.5 | Position accuracy of the models | 136 |
| 5.5.1 | Translation start model..... | 136 |
| 5.5.2 | Translation stop model..... | 137 |
| 5.6 | Comparison with other models | 138 |
| 5.7 | Concluding remarks | 139 |
| GENEPRED – AN <i>AB INITIO</i> GENE PREDICTOR..... | | 142 |
| 6.1 | Introduction | 142 |
| 6.2 | GAZE gene structure models | 143 |
| 6.3 | Eponine prediction models | 146 |
| 6.4 | Gene prediction with Eponine features | 147 |
| 6.5 | Tweaking GenePred gene prediction system | 154 |
| 6.5.1 | With Eponine translation models..... | 154 |
| 6.5.2 | Eponine Splice site predictions replaced with GeneSplicer predictions | 155 |
| 6.5.3 | Scaled down Eponine feature scores..... | 157 |
| 6.6 | Revisiting transcription termination predictions | 159 |
| 6.7 | Concluding remarks..... | 160 |
| CONCLUSIONS..... | | 166 |
| APPENDIX A: DOMAIN INSERTION..... | | 201 |
| A.1 | Introduction | 201 |
| A.2 | Types of domain insertions | 205 |
| A.3 | Nature and characteristics of domain insertions: Class level..... | 206 |
| A.3.1 | Size and function of domains involved in insertions | 206 |
| A.4 | Nature and characteristics of domain insertions: Fold and superfamily level..... | 208 |
| A.5 | Point of insertion..... | 210 |

| | | |
|---|--|------------|
| A.6 | Proximity of N- and C-termini in inserts..... | 210 |
| A.7 | Conclusions | 212 |
| APPENDIX B: PROTEIN EVOLUTION..... | | 214 |
| B.1 | Introduction | 214 |
| B.2 | Datasets..... | 217 |
| B.3 | Intermediate sequence search..... | 218 |
| B.4 | Structural homologs | 220 |
| B.5 | Clustering | 221 |
| B.6 | Orthology and paralogy | 229 |
| B.7 | Conclusions | 231 |
| APPENDIX C..... | | 233 |
| C.1 | Eponine transcription termination parameters..... | 233 |
| C.2 | GAZE gene structure models | 234 |

LIST OF ABBREVIATIONS

| | |
|------|--|
| ANN | Artificial Neural Network |
| BP | Branch Point |
| BPS | Branch Point Sequence |
| CATH | Class Architecture Topology Homologous superfamily |
| CF | Cleavage Factor |
| CPF | Cleavage and Polyadenylation Factor |
| CPSF | Cleavage and Polyadenylation Specificity Factor |
| CstF | Cleavage stimulation Factor |
| CTD | Carboxy Terminal Domain |
| DRE | Downstream Regulatory Element |
| EAS | Eponine Anchored Sequence |
| EST | Expressed Sequence Tags |
| FP | False Positives |
| GFF | General Feature Format |
| GLM | Generalized Linear Model |
| GO | Gene Ontology |
| HMM | Hidden Markov Model |
| ISS | Intermediate Sequence Search |
| MAZ | Myc-Associated Zinc finger protein |
| NMD | Non-sense Mediated Decay |
| NN | Neural Networks |
| PABP | Poly(A) Binding Protein |
| PAP | Poly(A) Polymerase |
| PC | Position Constraint |
| PDB | Protein DataBase |
| PE | Pause Elements |
| RMSD | Root Mean Square Deviation |
| ROC | Range Operating Characteristics |
| RVM | Relevance Vector Machine |
| SCFG | Stochastic Context Free Grammar |
| SCOP | Structural Classification Of Proteins |

| | |
|--------|------------------------------------|
| SD | Shine-Dalgarno |
| SLBP | Stem Loop Binding Protein |
| snRNPs | small nuclear Ribo-Nucleo Proteins |
| SR | Splice Regulatory |
| SSP | Secondary Structure Profile |
| SVM | Support Vector Machine |
| TFBS | Transcription Factor Binding Sites |
| TP | True Positives |
| URE | Upstream Regulatory Element |
| UTR | Un-Translated Region |
| WM | Weight Matrix |

LIST OF TABLES

- Table 1. Consensus motifs found in sequences between 50 and 2000 bases from cleavage site
- Table 2. Occupancy value for motifs detected in the transcription termination models.
- Table 3. Position constraints learnt while training chromosome 22 sequences
- Table 4. Coverage and accuracy values of transcription termination model along chromosome 20. ROC curve was constructed using these values.
- Table 5. Distribution of false positives within transcripts, exons and introns.
- Table 6. Occupancy value for motifs detected in the donor site models.
- Table 7. Occupancy value for motifs detected in the acceptor site models.
- Table 8. Occupancy value for motifs detected in the translation start site models.
- Table 9. Position constraints of translation start model learnt while training RefSeq cDNA sequences
- Table 10. Position constraints of translation stop model learnt while training RefSeq cDNA sequences
- Table 11. Occupancy value for motifs detected in the translation stop site models.
- Table 12. Performance of GenePred and GENSCAN in predicting VEGA annotated genes.
- Table 13. Performance of GenePred and GENSCAN in predicting VEGA annotated exons and splice sites.
- Table 14. Performance of GenePred constructed with translation start and stop features.
- Table 15. Performance of GenePred constructed with and without translation features along with GeneSplicer features instead of Eponine splice sites.
- Table 16. Performance of GenePred system constructed with and without translation after scaling down splice site and translation stop scores.

Table 17. Performance of transcription termination model with the support of GenePred prediction system.

Table 18. Performance of GENSCAN with and without GenePred in predicting VEGA annotated genes.

Table 19. Performance of GENSCAN with and without GenePred in predicting VEGA annotated exons.

Table 20. Nucleotide coverage by predictions of GenePred and GENSCAN.

Table 21. Coverage and accuracy of GenePred and GENSCAN for predictions offset by 1, 2 and 3 mega bases.

Table 22. Performance of GenePred and GENSCAN in identifying VEGA

Novel_transcripts and Putative genes. Coverage and accuracy for each annotation is given for GenePred with and without translation models. Each of these GenePred system is combined with either Eponine splice site or GeneSplicer features. Numbers in brackets shows the absolute values.

Table 23. Summary of performance of various versions of GenePred and GENSCAN in identifying VEGA annotated genes in human chromosome 20

Table 24. SCOP (1.61 release) classification statistics for chains in PDB_90 (April 2002 release)

Table 25. Distribution of inserted and parent domains at the SCOP class and fold level. The number of domains and the number of folds they come from is given for inserted and parent domains across the five different classes in the SCOP hierarchy. Percentage gives the number of folds contributing to insertions over total number of folds under the class.

LIST OF FIGURES

Figure 1. Schematic diagram showing (a) Typical gene structure of protein coding gene transcribed by RNA polymerase II (b) Matured RNA transcript with 5' cap and 3' poly(A) tail.

Figure 2. Rho-factor independent transcription termination in prokaryotes.

Figure 3. Rho-factor dependent transcription termination in prokaryotes.

Figure 4. Structure of RNA polymerase I terminators from yeast and mouse.

Figure 5. Schematic representation of 3'-end processing signals in human and yeast.

Figure 6. The nucleotide distribution of Y6 and R6 runs around transcription initiation and cleavage site. (a), (b) shows distribution in vertebrate mRNAs while (c) and (d) in mammals.

Figure 7. Nucleotide Distribution at Donor and Acceptor site analysed from 3,673 introns from human chromosome 22

Figure 8. Splicing mechanism where introns are spliced and exons are linked.

Figure 9. Translation initiation in eukaryotes

Figure 10. Translation termination mechanism mediated by release factors

Figure 11. Schematic diagram showing a section of HMM architecture

Figure 12. An example of Eponine model. (a) Position constraints along with Gaussian width and position. The nucleotide distribution in the weight matrices are represented as sequence logos. (b) Eponine model constructed from these constraints

Figure 13. RNA secondary structure features.

Figure 14. Four possible ways of extending a sub-optimal structure using Nussinov algorithm. (a) i unpaired (b) j unpaired (c) i, j pair (d) bifurcation.

Figure 15. Nucleotide composition spanning -200 to 2000 bases relative to the cleavage site

Figure 16. Nucleotide composition spanning -100 to 50 bases relative to the cleavage site

Figure 17. Averaged score values of sequences around cleavage site calculated using Nussinov algorithm

Figure 18. Averaged free energy values of sequences around cleavage site calculated using Zuker algorithm

Figure 19. Percentage of GC residues in the sequences around cleavage site

Figure 20. Percentage of GT residues in the sequences around cleavage site

Figure 21. Transcription termination model trained from chromosome 22 sequences

Figure 22. Transcription termination model trained from chromosome 20 sequences

Figure 23. Transcription termination model trained from chromosome 22 sequences with modified parameters

Figure 24. Schematic representation of hierarchical sequence model. Most promoter sequences have TATA box and CpG islands. Each promoter element can have specific DNA binding site for transcription factors to dock (represented as TFBS). Variations in each TFBS can in turn be modelled giving a hierarchical view of classifying different promoter types.

Figure 25. Schematic diagram showing criteria used for determining True positives (TP), False Positives (FP) and Ignored Predictions (IP).

Figure 26. ROC curve on transcription termination sites in chromosome 20 for Eponine model.

Figure 27. ROC curves on transcription termination sites for Polyadq, ERPIN and Eponine in comparison with random model.

Figure 28. Prediction density for transcription termination model along chromosome 20.

Figure 29. Prediction density along chromosome 20 for (a) ERPIN and (b) Polyadq.

Figure 30. Internal predictions per 100 kb of gene sequence in chromosome 20 for Eponine model.

Figure 31. Internal predictions per transcript in chromosome 20 for Eponine model.

Figure 32. Internal predictions of Eponine model in introns of chromosome 20 (a) Number of internal predictions versus intron length (b) Number of internal predictions normalised over intron length.

Figure 33. Prediction densities for transcription termination model near chromosome 20 annotated gene start sites. (a) Density of predictions in the same strand as of the gene (b) Density of predictions in the reverse strand as of the gene.

Figure 34. Prediction densities for two other transcription termination models near chromosome 20 annotated gene start sites (a), (c) Densities of predictions in the same strand as of the gene. (b), (d) Densities of predictions in the reverse strand as of the gene.

Figure 35. Prediction densities near chromosome 20 and 6 annotated gene start sites. (a) Density of predictions in the same strand as of the gene in chromosome 6 predicted by Eponine (b) Density of predictions in the reverse strand as of the gene in chromosome 6 predicted by Eponine (c) Density of predictions in the same strand as of the gene in chromosome 20 predicted by ERPIN (d) Density of predictions in the reverse strand as of the gene in chromosome 20 predicted by ERPIN (e) Density of predictions in the same strand as of the gene in chromosome 20 predicted by Polyadq (f) Density of predictions in the reverse strand as of the gene in chromosome 20 predicted by Polyadq.

Figure 36. Donor site model trained from SpliceDB sequences

Figure 37. Acceptor site model trained from SpliceDB sequences

Figure 38. Nucleotide Distribution at (a) Donor and (b) Acceptor site from chromosome 22 sequences

Figure 39. Donor site model trained from chromosome 22 sequences and donor site weight matrix

Figure 40. Position constraints of donor site model learnt while training chromosome 22 sequences

Figure 41. Acceptor site model trained from chromosome 22 sequences and acceptor site weight matrix

Figure 42. Position constraints of acceptor site model learnt while training chromosome 22 sequences

Figure 43. ROC curve for Eponine donor site model on chromosome 20 dataset

Figure 44. ROC curve for Eponine acceptor site model on chromosome 20 dataset

Figure 45. Prediction density for donor site model relative to annotated sites

Figure 46. Prediction density for acceptor site model relative to annotated sites

Figure 47. ROC curves on donor sites in chromosome 20 for Eponine, GeneSplicer, StrataSplice and Weight matrix.

Figure 48. ROC curves on acceptor sites in chromosome 20 for Eponine, GeneSplicer, StrataSplice and Weight matrix.

Figure 49. Exon coverage and accuracy on donor sites in chromosome 20 for Eponine, GeneSplicer, StrataSplice and Weight matrix.

Figure 50. Exon coverage and accuracy on acceptor sites in chromosome 20 for Eponine, GeneSplicer, StrataSplice and Weight matrix.

Figure 51. Prediction densities for StrataSplice and GeneSplicer donor and acceptor site predictions (a), (b) Densities for StrataSplice and GeneSplicer donor site predictions relative to the annotated site respectively (c), (d) Densities for StrataSplice and GeneSplicer acceptor site predictions relative to the annotated site respectively

Figure 52. Translation start model trained from (a) chromosome 22 (b) chromosome 20 genomic sequences

Figure 53. Translation start model trained from RefSeq cDNA sequences

Figure 54. Translation stop model trained from chromosome 22 cDNA sequences

Figure 55. ROC curve on human RefSeq cDNA dataset for Eponine translation start site predictions in same, opposite and both strands

Figure 56. ROC curve for translation start model on human RefSeq, mouse RefSeq and mouse Riken cDNA datasets (a) Predictions in the same strand (b) Predictions in both strands

Figure 57. ROC curve on translation stop sites in human RefSeq cDNAs for Eponine model

Figure 58. Prediction densities for translation start site model relative to annotated initiation sites (a), (b) Density of predictions in the same strand for human RefSeq and mouse RIKEN data respectively (c), (d) Density of predictions in both strands for human RefSeq and mouse RIKEN data respectively

Figure 59. Prediction density for translation stop model relative to annotated stop sites

Figure 60. ROC curves for Eponine, NetStart and ATGpr on human translation start sites in RefSeq cDNA sequences without (a) and with (b) strand information

Figure 61. Schematic representation of the gene models used for predicting genes from features in the forward strand. Reverse complementations of the forward strand rules are used for reverse strand gene predictions. (a) Simple gene model without translation models and thus no protein information. (b) Gene model with translation features. Any intron within 5' UTR region are not modelled. Based on these gene structures, candidate genes are predicted on both strands at the same time.

Figure 62. Genes predicted by linking Eponine models using GenePred compared with annotations available in the forward stand. Annotations from VEGA, ENSEMBL, EST

transcripts, UNIGENE and Human cDNAs are shown as tracks along with GENSCAN predictions (both on masked and unmasked sequence). The comparison is possible with the ENSEMBL ContigView which can load predictions from external source as DAS tracks.

Figure 63. Genes predicted by linking Eponine models using GenePred compared with annotations available in the reverse strand. Annotations from VEGA, ENSEMBL, EST transcripts, UNIGENE and Human cDNAs are shown as tracks along with GENSCAN predictions (both on masked and unmasked sequence). This figure is reproduced from ENSEMBL ContigView viewer.

Figure 64. Genes predicted by linking Eponine models using GenePred compared with annotations available in both strands. VEGA annotations are shown as black bars. The region covered by a bar includes all the alternative transcripts of a gene. GenePred predictions are given in red color. The figure also shows GENSCAN predictions and ENSEMBL annotations in different tracks.

Figure 65. Pictorial representation of (a) split and (b) fused predictions in comparison with annotation. (c) Few annotated genes have internal genes in the same strand. Predictions matching these genes are ignored while calculating accuracy. Annotations are given in black while predictions are drawn in red.

Figure 66. Venn diagram showing the coverage of GenePred and GENSCAN.

Figure 67. Domain insertion in Escherichia coli enzyme RNA 3'-terminal phosphate cyclase (PDB 1qmhA). The E. coli enzyme RNA 3'-terminal phosphate cyclase consists of two domains, of which one is contained within the other. The parent domain (residues 5-184, 280-338, coloured purple) consists of three repeated folding units; each unit has two α -helices and a four-stranded β -sheet. The folding unit resembles the C-terminal domain of bacterial translation initiation factor 3 (IF3). Between an α -helix and a β -

strand of the third IF3-like repeat of the parent domain, there is a smaller inserted domain (residues 185-279, coloured red). Although the inserted domain has the same secondary structural elements as the parent domain, it has different topology and a different fold. Insert resembles the fold observed in human thioredoxin.

Figure 68. Schematic representation of types of domain insertions observed in protein structures. (a) Single insertion (e.g., 1qmhA). (b) Nested insertion (e.g., 1a6dA). 'insert1 N' and 'insert1 C' represent the N- and C-terminus of insert, respectively. (c) Two-domain insertion (e.g., 1zfyA). (d) Three-domain insertion (e.g., 1dq3A).

Figure 69. (a) Domain length distribution for all domains in the non-redundant set of proteins (PDB_90). (b) Domain length distribution for parent domains.

Figure 70. (a) Proportion of residues in parent and insert domains in parent-insert combinations. (b) Point of insertion in parent domain. Insert position is given as a fraction of total length of parent domain.

Figure 71. Schematic diagram showing performance of different sequence comparison methods. The filled circle represents the query sequence used in the database search and the open circles represent family members. The distance between two circles represents some arbitrary distance.

Figure 72. Comparison of alignments of two distant proteins with and without intermediates. (a) Alignment of the two domain produced by FASTA 3.3. (b) The progressive alignment generated by including one intermediate. (c) The progressive alignment generated by including two intermediates.

Figure 73. Consensus sequences derived for the four SCOP protein group in monodomain cytochrome c family

Figure 74. Consensus of consensus for sequences in monodomain cytochrome c family

Figure 75. Flow chart describing steps used in clustering and visualisation of data.

Figure 76. Cluster map of cytochrome c superfamily

Figure 77. Cluster map of P-loops superfamily

Figure 78. Cluster map of cytochrome c superfamily with demarcation of SCOP
superfamily, family and protein levels

Figure 79. Cluster map of P-loops superfamily with demarcation of SCOP superfamily,
family levels

Figure 80. A cluster produced by the automated method for cytochrome c superfamily

Figure 81. A cluster produced by automated method for P-loops superfamily

Figure 82. Topology diagram for adenylate kinase

Figure 83. Topology diagram for cytochrome c proteins