Content

Abstract

1. Motivation 1

1.1 Thesis outline 2

1.2 Publication 4

1.3 Acknowledgements 4

2. Organisation of eukaryotic genomes 5

2.1 DNA as carrier of the genetic code 5

2.2 The structure of eukaryotic genes 7

Promoter 7

Transcription start site 7

Exons 9

Introns 9

2.3 Regulation of gene expression 9

2.3.1 Transcription factors 10

General TFs and the RNA polymerase II initiation complex 10

Gene specific transcription factors 12

Cellular control over TF activity 15

2.3.2 Chromatin modifications 16

Histone modifications 17

DNA methylation 18

2.4 Measuring gene expression 19

2.4.1 Microarray analysis 20

2.4.2 EST sequencing 20

Clustering ESTs and inferring tissue specificity of a gene 22

The GeneNest database 22

2.5 Measuring transcription factor binding 24

2.5.1 Small scale experiments 24

Promoter bashing 24

Electrophoresis mobility shift assay (EMSA) 24

DNase footprinting assay 24
2.5.2 Large scale approaches 25
SELEX 25
Protein binding arrays 25
ChIP-chip 26
ChIP-PET 27

3. Predicting TF Binding 29
3.1 Principles of TF binding site discovery 29
3.1.1 Deriving a consensus sequence 29
Basic site search 29
Consensus sequences 30
3.1.2 Deriving a TF binding motif 31
Pseudo counts 34
3.1.3 Statistical methods for finding binding sites in the genome 35
Requirement for a background model 35
Which score is large enough to indicate a binding site? 37
3.1.4 A biophysical model to predict TF binding energies 39
Statistical mechanical theory for TF-DNA interactions 40
3.1.5 From the biophysical approach to PWMs 45

3.2 Predicting regulating TF 49
3.2.1 Phylogenetic footprinting 50
3.2.2 oPOSSUM 51
3.2.3 PAP 52
3.2.4 CLOVER 53

4. The TRAP method 55
4.1 Deriving the TRAP model 55
4.1.1 Obtaining TF binding probabilities from mismatch energies 56
4.1.2 Deriving a binding measure for longer sequences 60
4.1.3 Deriving the TRAP parameters 61
How are predictions and experimental binding data related? 61
Experimental binding data 63

Utilized position frequency matrices 63
General parameter determination 63
General features of the parameter space 65
Optimal parameters choice derived from experimental data 69
Parameter choice in the absence of experimental data 72
Choosing the proper motif length 77
4.2 Results 79
4.2.1 Comparison of TRAP with hit-based methods 79
Comparison of achieved correlation 79
Comparison of ROC AUCs 80
4.2.2 Affinities are comparable between factors 85
4.2.3 Contributions from low affinity sites to ⟨N⟩ 87
4.2.4 Ranking of intergenic regions is robust 91
4.3. Discussion 91

5. Application of TRAP to higher eukaryotes 95
5.1. The *Drosophila eve* promoter 95
5.2. P53 binding predictions 96
5.3. Affinity ranking for the transcription factor SRF 98
TF ranking for SRF target genes 100

6. PASTAA – Predicting TFs associated with groups of genes 103
6.1 Statistical measures for detecting affinity enrichments 104
6.1.1 Testing for increased average affinity 105
Affinities tend to be log-normally distributed 106
Applying the z-score test 106
6.1.2 Testing for target gene enrichment 107
Cutoff selection for TF targets and tissue specific genes 109
6.2 Data and methods used for analyzing tissue specific gene sets 110
6.2.1 Sequence data 110
6.2.2 Defining tissue specific gene expression 110
EST data 110
GNF data 111
6.2.3 Assessing TF expression across tissues 111

6.3 Validation of the PASTAA approach 112

6.3.1 Performance on ChIP-chip and PBM data sets from yeast 112

Results obtained from the z-score test 113

Results obtained from the hypergeometric test 117

Comparison to results from ROC curve AUCs 122

6.3.2 Validation on individual vertebrate genes 122

6.3.3 PASTAA can account for ChIP data from vertebrates 125

6.4 Analysis of tissue specific promoters 127

6.4.1 Detection of regulatory regions for tissue specific genes 127

6.4.2 TFs preferentially associated with EST derived tissue sets 130

Tissue specific genes are associated with TATA box motifs 133

6.4.3 Tissues preferentially associated with a given TF 133

TFs are themselves over-expressed in their top ranking tissues 134

Exemplary analysis of eye specific target genes for CHX10 136

6.4.4 Distribution of functional binding signals across promoters 137

Windows with maximal affinity are located near the TSS 140

TF-tissue ranks are robust against changes in promoter size 142

6.4.5 Overlap between GNF and EST data 143

6.4.6 Comparison to oPOSSUM, PAP and Clover 145

6.5 Web implementation of PASTAA 147

6.6 Discussion 149

7. Discussion 153

Bibliography 159

Appendix A. The Boltzmann distribution 177

A.1 Derivation via the dominating energy distribution 177

A.2 Derivation via assuming contact with a heat reservoir 182

Appendix B. Zusammenfassung 185