

Can Gains and Losses of Transcription Factor Binding Sites be Related to What Occurs Elsewhere in the Regulatory Region?

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Life Sciences

2008

Christopher Robert Lockwood

R

Contents

1	INTRODUCTION AND LITERATURE REVIEW	17
1.1	THE FIELD OF RESEARCH CHOSEN	17
1.2	BACKGROUND INFORMATION ON REGULATION OF GENE EXPRESSION IN EUKARYOTES	18
1.2.1	Regulation by transcription factors	18
1.2.2	Other methods of gene regulation	23
1.2.3	More detailed studies of TFs and TFBSs	26
1.2.4	Regulatory DNA: features beyond conventional TFBSs	33
1.3	INFORMATION FROM THE LITERATURE RELATING TO THE EVOLUTION OF TRANSCRIPTION FACTOR BIND- ING SITES	34
1.3.1	Evidence of TFBS evolution in eukaryotes	34
1.3.2	Are unnecessary TFBSs removed by natural selection?	44
1.3.3	Correlated evolution: the combined effect of the evo- lution of several TFBS	46
1.3.4	“Turnover” of TFBSs	50
1.3.5	Must the evolution of TFs be studied as well as the evolution of TFBSs?	52
1.3.6	Alignment and analysis techniques for regulatory regions	55

1.3.7	Evolution of TFBSs in network research	61
1.4	SOME POSSIBLE SCENARIOS FOR HOW REGULATORY REGIONS MIGHT EVOLVE	64
1.5	DEFINITION OF “EVOLVED TFBS”	66
1.6	AIMS	67
2	EXAMINATION OF KNOWN REAL SITES (FROM THE TRANSFAC DATABASE)	68
2.1	INTRODUCTION	68
2.2	METHODS	69
2.2.1	Determining whether a TFBS is conserved or not: Out- line of method	69
2.2.2	Determining whether a TFBS is conserved or not: De- tailed method: the PWMs	71
2.2.3	Determining whether a TFBS is conserved or not: De- tailed method: Obtaining Data	80
2.2.4	Determining whether a TFBS is conserved or not: De- tailed method: Analysing Data	87
2.2.5	Distinguishing versions of outputs	106
2.2.6	Minor utilities	106
2.2.7	Generating a sample of fictional TFBSs	108
2.2.8	Using the samples of conserved and non-conserved TF- BSs	114
2.3	RESULTS	119
2.3.1	Problems using TransFac co-ordinates	119
2.3.2	Counting the proportion of TFBSs that are not conserved	119
2.3.3	Initial examination of several measures	122

2.3.4	Further examination of sequence conservation near each TFBS	128
2.3.5	Reserve List	133
2.4	DISCUSSION	133
3	EXAMINATION OF KNOWN REAL SITES (FROM CHIP-CHIP DATA)	136
3.1	INTRODUCTION	136
3.2	METHOD	137
3.2.1	The choice of TF	137
3.2.2	The PWM and scores	137
3.2.3	The initial list of genes regulated	142
3.2.4	The size of the upstream search region	142
3.2.5	The analysis of whether the TFBS was conserved . . .	143
3.2.6	"Second-best matches"	144
3.2.7	Fictional TFBSs	146
3.3	RESULTS	147
3.3.1	The proportion of TFBSs that are successfully located and the choice of p-value	147
3.3.2	Conservation of DNA surrounding conserved and diverged TFBSs	150
3.4	USING TRANSFAC DATA TO ESTIMATE ERRORS FROM CHIPCHIP DATA	152
3.4.1	Method	152
3.4.2	Results	153
3.5	SUMMARY AND DISCUSSION	155

4	SEQUENCE CHANGES EXPECTED TO BE ASSOCIATED WITH GAINS AND LOSSES OF TFBS	158
4.1	INTRODUCTION	158
4.2	OUTLINE OF MODELS	159
4.3	DIFFERENCES BETWEEN MODELS	162
4.3.1	Conceptual differences	162
4.3.2	Possible observable differences	162
4.3.3	Complications if TFBSs are subject to counterselection when they are lost	163
4.3.4	Time taken to disrupt a TFBS that is not under natural selection	167
4.4	SUMMARY	170
5	PHYLOGENETIC TREES: DISTINGUISHING GAINS FROM LOSSES OF TFBSs	171
5.1	INTRODUCTION	171
5.2	METHOD	172
5.2.1	Use of phylogenetic tree	172
5.2.2	Fictional TFBSs	174
5.2.3	Conservation of DNA surrounding the TFBS	175
5.2.4	TFBSs located by a Chip-Chip Experiment	175
5.3	RESULTS	176
5.3.1	Based on TFBSs from the TransFac Database	176
5.3.2	Based on TFBSs located by a Chip-Chip Experiment	179
5.4	DISCUSSION	182
5.4.1	Ratio of gains to losses	182

5.4.2	Ratio of gains to losses - comparison from the literature	182
5.4.3	Fictional TFBSs	185
5.4.4	Conservation of neighbourhood DNA	186
5.4.5	Discovery bias	186
5.4.6	Gains or losses of TFBSs not caused by mutations	187
5.4.7	Mistaken TFBSs from chip-chip data	188
5.4.8	A methodological issue: should losses and gains have equal weight?	188
5.4.9	Other problems with parsimony	189
6	SUMMARY AND DISCUSSION	190
6.1	SUMMARY	190
6.2	COMPARISON WITH WORK IN THE LITERATURE	193
6.2.1	The turnover model	193
6.2.2	Other studies	197
6.3	DISCUSSION	199
6.3.1	Possible mechanisms	199
6.3.2	Questions raised and implications	202
7	BIBLIOGRAPHY	205
A	APPENDIX: MUTATION RATES AT FOURFOLD SYNONYMOUS SITES	219
A.1	INTRODUCTION	219
A.2	METHOD	220
A.3	RESULTS	221
A.4	DISCUSSION	222

B APPENDIX: DETAILS OF COMPUTER CODE	223
B.1 INTRODUCTION	223
B.2 METHOD	223
B.2.1 PWMs	223
B.2.2 Obtaining a list of TFBSs	224
B.2.3 Identity of a Gene	225
B.2.4 Locating the precise TFBS	226
B.2.5 Alignment of sequences	226
B.2.6 Examination of alignments	227
B.2.7 Deciding if a TFBS was conserved or not	229
B.2.8 Examination of “neighbourhood” (DNA surrounding a TFBS)	229
B.2.9 Awkward alignments	229
B.2.10 Generating a conclusion	230
B.2.11 Minor utilities	230
B.2.12 Generating a sample of fictional TFBSs	231
B.2.13 The chip-chip data	231
B.2.14 Phylogenetic trees: distinguishing gains and losses of TFBSs	232
C APPENDIX: PARSIMONY, TFBS LOSS AND PHYLOGE- NETIC TREES	233
D SUMMARY: TFBSs in Vertebrates: Do They Evolve To- gether, Or Individually?	236

List of Tables

1	List of abbreviations	12
2.1	Points considered when choosing PWMs	74
2.2	Points considered when choosing PWMs, continued	75
2.3	TFs and matrices used - main list	76
2.4	TFs and matrices used - reserve list	77
2.5	Worked example of using a PWM to score a DNA sequence . .	79
2.6	Example of locating the precise TFBS	88
2.7	Example of locating the precise TFBS, cont.	89
2.8	Example illustrating alignment gaps against a TFBS	100
2.9	Simple example illustrating when a TFBS is “between-aligns” .	104
2.10	Numbers of fictional TFBSs - unfiltered	112
2.11	Numbers of fictional TFBSs - filtered	113
2.12	Simple example illustrating why species-stratification is desirable	116
2.13	Most common reasons for abandoning analysis of a TFBS before it was compared with any other species (Run F)	125
2.14	Number of TFBS comparisons (in Run F)	125
2.15	Similarities between conserved and diverged TFBSs	126
2.16	Real and fictional results that were similar	127

2.17	Differences between conserved and diverged TFBSs	128
2.18	Conservation of sequence flanking conserved and diverged TF- BSs	131
2.19	Conservation of sequence flanking fictional TFBSs	131
2.20	Conservation of flanking sequence, for every combination of species	132
2.21	Conservation of DNA around TFBSs: from Reserve List . . .	134
3.1	PWM for HNF4	139
3.2	Neighbourhood conservation by data source	150
3.3	Neighbourhood conservation: pseudo chip-chip	153
5.1	Gains and losses: TransFac data	177
5.2	Gains and losses: fictional data	178
5.3	Gain/loss significance test	178
5.4	Neighbourhood conservation, TransFac data, species-stratified analysis	180
5.5	Gains and losses: chip-chip data	181
5.6	Gains and losses: fictional chip-chip data	181
A.1	Mutation frequencies at synonymous sites	221

List of Figures

1.1	Two models for how enhancers operate	20
1.2	Estimating numbers of Zeste TFBSs	41
1.3	Finding a “seed” for a Blastz alignment	59
1.4	Gene duplication followed by gain and loss of TFBSs	63
1.5	Gene duplication followed by loss of TFBSs	63
1.6	Possible ways a regulatory region might evolve	65
2.1	Histogram of binding scores for Crx TFBSs	81
2.2	Flowchart showing how orthologous genes were obtained from XenoRef	86
2.3	Flowchart showing how a TFBS would be located precisely . . .	90
2.4	Criteria for deciding if a TFBS is not conserved	96
2.5	Flowchart showing how “neighbourhood” was examined	102
2.6	Flowchart showing “between-aligns” sequence analysis	105
2.7	Flowchart showing how to obtain an overall conclusion	107
2.8	Flowchart showing how fictional TFBSs were generated	109
2.9	Fictional CREB TFBSs, unfiltered	111
2.10	Fictional CREB TFBSs, filtered	114
2.11	Locations of TFBS: two estimates compared	120

2.12	Proportion of TFBSs that are not conserved	123
2.13	Proportion of TFBSs that are not conserved, by structure . .	124
2.14	Conservation of DNA flanking TFBSs	130
3.1	Outline of ChIP method	138
3.2	Histogram showing HNF4 binding sites	141
3.3	Histogram of p-values for chip-chip data	149
3.4	%id of neighbourhood: variations with p-value	155
4.1	A region of low “site-density”	160
4.2	A region of high “site-density”	160
4.3	Sequence conservation expected in lost and gained TFBSs . .	164
5.1	The eight species used, in an evolutionary tree	173
5.2	Examples showing apparent gains and losses	177
6.1	Turnover and other models	195
6.2	Another way a regulatory region might evolve	201
C.1	Example involving human, rabbit and mouse	234

Table 1: List of abbreviations

CGI	Common gateway interface
ChIP	Chromatin immuno-precipitation
chip-chip	Chromatin immuno-precipitation followed by genechip (microarray) analysis
ER	Estrogen receptor
GERP	Genomic Evolutionary Rate Profiling
HLH	Helix-loop-helix
HTH	Helix-turn-helix
LM2	Long motif 2
MCS	Multi-species conserved sequences
miRNA	Micro RNA
MRP	Mitochondrial ribosomal proteins
mya	Million years ago
NR	Nuclear receptor
PTRR	Putative transcriptional regulatory regions
PWM	Position weight matrix
RISC	RNA induced silencing complex
SUMO	Small ubiquitin-related modifier
TBA	Threaded blockset aligner
TF	Transcription factor
TFBS	Transcription factor binding site
TSS	Transcription start site

ABSTRACT

Background: This thesis asks whether a transcription factor binding sites (TFBSs) tend to evolve together with other TFBSs located in nearby DNA. Other studies suggest this occurs, but the present study asks if it occurs frequently.

Results: Starting from a database of known TFBSs, a computer analysis produced a sample of TFBSs that seem to be conserved (eg within a human-mouse comparison), and another sample of TFBSs that seem to have diverged. The conserved TFBSs were flanked by DNA which was well conserved, whereas the diverged TFBSs were flanked by DNA which was less well conserved. The difference was typically 11%. The thesis considers if this difference could be produced by faulty data, or by TFBSs that have no effect on fitness, but shows this is unlikely by analysing a set of fictional TFBSs. Two possible explanations are: (i) correlated evolution, in which the loss of a TFBS is accompanied by the loss of several other TFBS within 50 bases; or (ii) a site-density effect, where the probability that a TFBS is lost/gained varies with the number of TFBSs in nearby DNA. To decide between these, a method was devised and implemented; it required gain-of-TFBS to be distinguished from loss-of-TFBS. This produced tentative evidence that “losses” are flanked by DNA that is more highly conserved than the DNA flanking “gains”. Such a result is difficult to explain using a “turnover” model or a “site-density” model, but can be explained by a “correlated-evolution” model.

Conclusions: It was found that “correlated-evolution” best explained the data, but this was a tentative conclusion, given the statistical significance levels. If true, the implication is that a common event in TFBS evolution is the simultaneous loss of several nearby TFBSs.

Declaration: no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning, except for the reuse of some Perl modules (altered as required), originally written during the candidate’s Master in BioInformatics course (University of Exeter)

- i. The author of this thesis (including any appendices to this thesis) owns any copyright in it (the “Copyright”) and he has given The University of Manchester the right to use such Copyright for any administrative, promotional, educational and/or teaching purposes.
- ii. Copies of this thesis, either in full or in extracts, may be made **only** in accordance with the regulations of the John Rylands University Library of Manchester. Details of these regulations may be obtained from the Librarian. This page must form part of any such copies made.
- iii. The ownership of any patents, designs, trade marks, and any and all other intellectual property rights except for the Copyright (the “Intellectual Property Rights”) and any reproductions of copyright works, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property Rights and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property Rights and/or Reproductions.
- iv. Further information on the conditions under which disclosure, publication and exploitation of this thesis, the Copyright and any Intellectual Property Rights and/or Reproductions described in it may take place is available from the Vice-President and Dean of the Faculty of Life Sciences.

ACKNOWLEDGEMENTS

Professor Steve Oliver was the supervisor for this PhD, and my thanks go to him for taking on the responsibility of supervising, giving me advice, and reading so many pages of draft reporting. However, initially Dr Erich Bornberg-Bauer was supervisor before he left for Germany. Prof Richard Reece was advisor.

The BBSRC supported this PhD financially.

The following projects provided publicly available genome sequence data that has been used during this PhD:

Chicken genome: (Hillier et al., 2004)

Cow genome: The Bovine Genome Project,
<http://www.hgsc.bcm.tmc.edu/projects/bovine/>

Dog genome: The Broad Institute, and Agencourt Bioscience

Human genome: The Human Genome Sequencing Consortium

Mouse genome: (MGSC, 2002)

Opossum genome: The Broad Institute

Rat genome: The Rat Genome Sequencing Consortium

Rhesus genome: Baylor College of Medicine Human Genome Sequencing Center, and the Rhesus Macaque Genome Sequencing Consortium,

<http://www.hgsc.bcm.tmc.edu/projects/rmacaque/>

THE AUTHOR

The author obtained an honours degree in physics (first class) from the University of Warwick. He subsequently researched road accident statistics at the Transport Research Laboratory.

His bioinformatics career started by obtained a Masters degree in BioInformatics (with Distinction) from the University of Exeter. He afterwards continued the masters research project by staying at the University of Exeter as a research assistant for a year, amounting to about 18 months research work on that project altogether. Two peer-reviewed papers were based entirely on this research work (Lockwood and Frayling, 2003) (Lockwood et al., 2003), and he also co-authored other publications from the University of Exeter (Bingham et al., 2003) (Pearson et al., 2004) (Lockwood and Frayling, 2002).