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

.

List of Tables

List of Figures

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)

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).