Prediction and Analysis of Nucleosome Positioning in Genomic Sequences

Samiul Hasan

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

April 2003

Wolfson College, University of Cambridge

and

The Wellcome Trust Sanger Institute,

Hinxton, Cambridgeshire

Abstract

A nucleosome is the resultant structure formed when 1.6 left-handed turns of DNA (~146 bp) are wound around a basic complex of histone proteins (the histone octamer). Nucleosomes occur naturally and ubiquitously in all eukaryotic genomes; the histone proteins themselves are highly conserved in eukaryotes. Experimental evidence suggests that specific DNA sequences may exhibit high or low nucleosome-forming tendencies compared to random DNA. This could mean that nucleosomes, whose positions are influenced by the underlying DNA sequence, can in turn govern the accessibility of regulatory DNA sequences such as transcription initiation and replication origin sites. This forms the need to search for evidence of nucleosome positioning and consequently build models to predict and investigate such locations.

One theory suggests that DNA sequences, which are intrinsically "curved", can position nucleosomes. In a previous study, using "cyclical" hidden-markov models, it had been suggested that a 10 periodic occurrence of the [VWG] motif could have such an effect and could help nucleosomes to be positioned in human exons. This work was extended in this thesis. 60% of human genomic sequences were seen to be covered in apparently weak 9-10 bp periodic patches of [CWG]. [CWG]-dense regions were seen to alternate with regions which were rich in [W] motifs in human. However, the pattern was not the same in mouse.

Another theory suggests that highly flexible or highly rigid DNA sequences may favour or disfavour nucleosome formation respectively. The locations of such patterns were investigated in human sequences using the wavelet technique. This approach identified confined periodic patterns (in the range of 80-200 bp) of rigidity in human genomic sequences; the patterns themselves were, however, mainly consequences of alu repeat-clustering. However, the same analysis in the mouse genome indicated that such a mechanism for positioning nucleosomes was not conserved and therefore unlikely.

A different approach to model nucleosomes was to train weighted DNA matrices from experimentally-mapped nucleosome datasets. This technique gave some encouraging results (one model showing 100% accuracy at 40% coverage), but was restricted by the limited size of the datasets.

Overall the conclusion is that there is some evidence for sequence specific nucleosome positioning, but that more experimental data is needed to build and evaluate practical and predictive computational models.

Preface

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 is not substantially the same as any I have submitted for a degree or diploma or other qualification at any other University.

Samiul Hasan,

14/07/2003

Acknowledgements

I would firstly like to thank my supervisor, Dr Tim Hubbard, for giving me the opportunity to carry out my PhD research at the Sanger Institute. The nucleosome prediction problem was challenging and I definitely needed all the help I got. I am grateful for all his invaluable advice and for all that I have learnt as a result. I am also indebted to Thomas Down and Matthew Pocock for their courteous advice and help, especially for solving a number of difficult programming issues in my project. Thanks also to Aroul Selvam Ramadass, Raphael Leplae and Yen-Hua Huang for their generous help and for frequent discussions relating to this project.

A special thanks to Dr. Akilesh Eswaran for helping me with computer-related problems on numerous occasions. I am also grateful to Jason Dowling, Diego Di Bernardo, Arnaud Kerhornou, Arundhati S Ghosh, Mahesh Menon, Ashish Tuteja, Maushumi Guha, Giota Mitrou, Andreas Heger, Kevin Howe, the Arrelanos, Visalakshi Kannan, Antonella Ferrecchia, Gabriella Rustici, Ranjeeva Ranasingh, Waqas Awan, Jayne Vallance and to the many other friends I made during my time at Cambridge for their friendship and courtesy and for helping me to enjoy my time here even more. Finally, I would also like to express my gratitude to my family in particular to my parents, my brother, my uncles Tariq and Safiul, Dr Karim, Fariha, Sofia and Mowli for their kind words and acts of encouragement during this time.

Ambiguity Codes for DNA as specified by the Convention of the International Union of Pure and Applied Chemistry[1]

IUPAC Code	Meaning	Complement
A	A	T
C	C	G
G	G	C
T/U	T	A
M	A or C	K
R	A or G	Y
W	A or T	W
S	C or G	S
Y	C or T	R
K	G or T	M
V	A or C or G	B
H	A or C or T	D
D	A or G or T	H
B	C or G or T	V
N	G or A or T or C	N

[1] Cornish-Bowden, A. (1985). Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res* 13, 3021-30.

Prediction and Analysis of Nucleosome Positioning in Genomic Sequences

List of Figures

List of Tables