A. Christopher Previti
Dr. sc. hum.

Development of novel approaches to the prediction of CpG islands and their methylation status

Geboren am 8. August 1977 in Heidelberg
Diplom der Fachrichtung Biologie am 4. April 2002 an der Universität Heidelberg

Promotionsfach: DKFZ (Deutsches Krebsforschungszentrum)
Doktorvater: Prof. Dr. rer. nat. Sándor Suhai

CpG islands are important regulatory regions in vertebrate genomes and are involved in the epigenetic control of normal as well as disease-related gene expression, since the methylation of CpG islands located in the promoter region inhibits gene expression by making the chromatin inaccessible to the transcription machinery. This mechanism plays an important role in establishing the gene-expression patterns necessary for normal tissue differentiation and acts to protect the integrity of the genome against intragenomic parasites. Since both the experimental identification of CpG islands and the evaluation of their methylation status are costly and time-consuming, computational methods were developed for both of these tasks.

This thesis first introduces a novel CpG islands prediction algorithm, termed *CpGcluster* that was developed on the basis of the real distribution of CpGs within the human genome. This approach reflects the original concept of the CpG islands, which were initially described as clusters of CpGs within the generally CpG-depleted bulk DNA. Other CpG island prediction methods, on the other hand, employ a more indirect CpG island definition that does not reflect their most important characteristic, the high degree of CpG clustering. In contrast to these approaches, *CpGcluster* is able to adapt to the highly variable composition of vertebrate genomes and does distinguish sufficiently between true CpG islands and spurious *Alu*-retrotransposons, which have a similar sequence composition. Benchmarking experiments used to evaluate *CpGcluster* have shown that it demonstrates the highest specificity and overall accuracy in detecting experimentally identified CpG islands. Additionally, *CpGcluster* predicts the highest percentage of promoter-overlapping CpG islands and the lowest percentage of *Alu*-retrotransposons and the highest percentage of conserved *PhastCon* sequences.

Furthermore, *CpGcluster* readily captures small, functional CpG islands that are often missed by the other programs. A subset of these CpG islands is associated with the family of MAGE-genes, which become methylated during normal mammalian development and represent clear examples of tissue-specific genes that use DNA methylation as a primary mechanism for their regulation. They are part of a growing body of evidence showing that functional CpG islands do not necessarily always stay unmethylated under either normal or disease related circumstances, such as cancer.

Though a sizable fraction of the CpG islands may be affected, little is known about the linkage between tissue-specific CpG island methylation and other genomic attributes characterizing the DNA sequence composition and conformation, the presence of repetitive elements or evolutionary sequence conservation. Previous studies attempting to discover the CpG island features that discriminate between susceptibility and resistance to methylation have not separated between CpG islands with constitutive and tissue-specific patterns of

methylation. Therefore, in the second part of this thesis, unsupervised clustering methods were used to discover classes of CpG islands that are methylated in a tissue-specific manner using the experimental methylation data available from the Human Epigenome Project (HEP). These newly defined methylation classes included both uniformly methylated or unmethylated CpG islands, as well as CpG islands that were unmethylated exclusively in sperm. While it is highly probable that the uniformly unmethylated CpG islands have a role in the regulation of gene expression, since they are associated with promoter regions of genes, neither the uniformly methylated CpG islands nor those CpG islands lacking methylation solely in sperm showed this type of functionality. Though there are indications that the CpG islands of this latter class may be associated with the formation of nucleosomes due to their sequence structure and conformation, this theory still requires verification.

The features that best discriminated between uniformly methylated and unmethylated CpG islands were their degree of evolutionary conservation and the significance value assigned by *CpGcluster*, but the most conserved CpG islands were actually also most likely to be methylated. This is a surprising result that contradicts the conventional wisdom that the sequences with the highest regulatory potential are also the most conserved and points to a need for a more differentiated use of cross-species conservation for the detection of potential regulatory regions. These results further support the approach taken by *CpGcluster* in using a *p*-value to qualify the CpG islands, since a low *p*-value not only implies a high significance of the CpG island predictions, but also a lack of methylation and high potential for functionality. Furthermore, these novel classes of CpG islands were used to develop algorithms for the prediction of CpG island methylation that were able to outperform all current methods designed for this task.

Therefore, this work provides the basis for the future development of tools that can predict the methylation status of CpG islands across the genome in order to detect potentially valuable targets for large-scale methylation studies or biomarkers that are susceptible to methylation and could be used for early cancer detection.