Bioinformation
www.bioinformation.net

open access
Current Trends

Computational Epigenetics: the new scientific paradigm

Shen Jean Lim[1], Tin Wee Tan[1], Joo Chuan Tong[1,2,*]

[1]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597;[2]Data Mining Department, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01, Connexis, South Tower, Singapore 138632.Tel. 65-64082156; Shen Jean Lim - E-mail: jctong@i2r.a-star.edu.sg, * Corresponding Author

Received December 15, 2009; Revised January 13, 2009; Accepted January 21, 2009; Published January 23, 2010

Abstract:
Epigenetics has recently emerged as a critical field for studying how non-gene factors can influence the traits and functions of an organism. At the core of this new wave of research is the use of computational tools that play critical roles not only in directing the selection of key experiments, but also in formulating new testable hypotheses through detailed analysis of complex genomic information that is not achievable using traditional approaches alone. Epigenomics, which combines traditional genomics with computer science, mathematics, chemistry, biochemistry and proteomics for the large-scale analysis of heritable changes in phenotype, gene function or gene expression that are not dependent on gene sequence, offers new opportunities to further our understanding of transcriptional regulation, nuclear organization, development and disease. This article examines existing computational strategies for the study of epigenetic factors. The most important databases and bioinformatic tools in this rapidly growing field have been reviewed.

Keywords: epigenetic informatics, epigenetics, epigenomics, bioinformatics

Background:
Decoding the genomes of human and other model organisms have produced increasingly large volumes of data relevant for understanding natural selection, development and evolution, the causation of disease, and the interplay between genotypes and phenotypes during development. Collectively, this information reflects the current state of knowledge on the genetic and genome attributes of organisms. The huge amount of accumulated data represents a goldmine for the study of molecular evolution **[1]**,**[2]**, disease-specific mutations **[3]**, **[4]**, **[5]** and biodiversity measurements **[6]**, **[7]**. While much progress has been made in genomic research, increasing evidence have shown that the study of gene factors by itself is insufficient in explaining all aspects of heritable changes in phenotype, gene function or gene expression. It is now known that chemical modifications of DNA and histones can modify gene activity through alterations in chromatin structure that blocks or promotes transcriptional initiation **[8]**. Enzymes involved in this process include DNA methyltransferases, histone deacetylases, histone acetylases, histone methyltransferases and the methyl-binding domain protein MECP2 **[9]**.

The need to identify chemical modifications that can alter gene activity and expression has given rise to the field called epigenetics. This form of second-order genetics provides a whole new dimension to genes beyond the genome, and proposes a control system of genetic 'switches' for regulating gene expression. Epigenetics, first defined by Conrad Waddington in 1942, refers to the study of epigenesis, i.e., how genotypes give rise to phenotypes through programmed change **[10]**. At the heart of this new wave of research is the "study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" **[11]**. Recent data have shown that epigenetic modulations are essential in many developmental processes, including tissue formation, organ formation and allele-specific gene expression **[12]**. Changes in these normal epigenetic patterns can deregulate patterns of gene expression, resulting in adverse clinical outcomes **[13]**. Increasing evidence indicates that such mechanisms play important roles in psychiatric disorders **[14]**, obesity **[15]**, life experiences **[16]** and the etiology of diseases such as cancer **[17]**, schizophrenia **[18]**, Beckwith-Wiedemann syndrome **[19]** and Alzheimer's disease **[20]**.

Epigenetics is highly combinatorial in nature due to the array of diverse control elements. The human genome contains ~23,000 genes that are active in specific cells at precise moments. Cells control gene expression by wrapping DNA around clusters of core histone proteins to form nucleosomes **[21]**, which are then organized into chromatin. Changes to the structure of chromatin affect gene expression patterns: genes are inactivated when the chromatin is condensed, and they are expressed when chromatin is relaxed **[9]**. These dynamic chromatin states are controlled by DNA methylation **[22]**, histone modifications (e.g., methylation, acetylation, phosphorylation, sumoylation and ubiquitylation) **[23]**, **[24]** and DNA-binding proteins (e.g., polycomb and trithorax group proteins) **[25]**. Most of these epigenetic modification mechanisms have been shown to be regulated by non-coding RNAs (ncRNAs), such as microRNAs (miRNAs), small RNAs (guide RNAs, piRANS) and large RNAs, which play important roles in events including transposon activity and silencing, position effect variegation, X-chromosome inactivation and paramutation **[26]**.

With the rapid increase in the number of new modification sites being discovered each year, it has been suggested that post-translational modification may affect almost every solvent-accessible histone residue, allowing a high level of variability for signal transduction events **[21]**, **[27]**. This enormous combinatorial complexity **[28]** requires an extraordinarily large number of experiments, such as DNA methylation profiling, for systematic studies. Already, a number of large-scale initiatives have been established for the systematic mapping of epigenomic and related data. These include projects by the Alliance for the Human Epigenome and Disease (AHEAD) Task Force **[29]**, the ENCyclopedia Of DNA Elements (ENCODE) Project Consortium **[30]**, the Human Epigenome Project (HEP) Consortium **[31]** and the Highthroughput Epigenetic Regulatory Organisation In Chromatin (HEROIC) Project Consortium **[32]**. The huge quantity of experimental data generated by these and other projects requires appropriate bioinformatics infrastructure spanning general and specialist databases, basic bioinformatics tools and sophisticated algorithms for meaningful analysis, modeling and prediction of DNA-protein interactions. Pioneering efforts in the field of computational epigenetics have been reviewed by Bock and Lengauer **[33]**. Here, we review major tools and resources that have been developed in this rapidly growing field, with special analysis on the latest trends and future directions.

Data sources for Epigenetic research:
Large amount of data relevant for epigenetic research are available in scientific literature, molecular databases and case reports. Scientific literature serves as the primary source of data, providing high-level descriptions of biological entities and processes. As of January 2009, PubMed contained over 288,000 records related to epigenetic research. This information exists in the form of unstructured free text that makes the extraction of biologically meaningful information difficult. As the amount of electronically accessible textual material accumulates, the

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(7): 331-337 (2010)

quality of epigenetic research will be increasingly dependent on the ability to retrieve quality data to facilitate the discovery of new facts, interpretation of results, and design of experiments **[34]**.

The number and size of molecular databases have been increasing steadily. A total of 1,078 molecular biology databases are currently (March 2009) described in the *Nucleic Acids Research* online Molecular Biology Database Collection **[35]**. These include 3 nucleotide sequence databases, 60 databases on transcriptional regulatory sites and transcription factors, 65 databases on microarray data and gene expressions, and 114 databases on human genes and diseases. The international collaborative GenBank **[36]**, DNA Data Bank of Japan (DDBJ) **[37]** and EMBL **[38]** serve as worldwide repositories for nucleotide sequences of different origins.

A number of epigenetic databases have been reported. We have reviewed some of these databases **(Table 1 in supplementary material)**. DNA methylation databases are useful for studying the covalent modification of a cell's genetic material, particularly in the complex genomes of higher order vertebrates. Important sources of DNA methylation databases include MethDB **[39]**, MethPrimerDB **[40]** and MethyLogiX **[20]**, which contains information on DNA methylation genes and patterns across different species, individuals, tissue and cell types and phenotypes. Histone databases are important for research in the compaction and accessibility of eukaryotic and probably archaeal genomic DNA. The National Human Genome Research Institute (NHGRI)'s Histone Database **[41]**, **[42]** serves as a central data source for histones and histone fold-containing proteins. Cancer methylation databases are valuable for analyzing irregular methylation patterns that are correlated with various cancers. Major data sources include PubMeth **[43]** and MeInfoText **[44]**, which contains information on gene methylation profiles of specific cancer types. Online resources for cell-, disease-, organism- and stage-specific gene expression patterns are also available. The National Center for Biotechnology Information (NCBI)'s Gene Expression Omnibus (GEO) **[45]** serves as a central repository for high-throughput gene expression data. It also stores high-throughput functional genomic data such as genome copy number variations, chromatin structure, methylation status and transcription factor binding. The Gene Expression Nervous System Atlas (GENSAT) **[46]** provides information about the precise distributions of specific genes and proteins throughout brain development. StemBase **[47]** details gene expression data of stem cells and derivatives from rat, mouse and human. The Gene Normal Tissue Expression (GeneNote) database **[48]** contains complete gene expression profiles in healthy human tissues (bone marrow, brain, heart, kidney, liver, lung, pancreas, prostate, skeletal muscle, spinal cord, spleen and thymus) using the Affymetrix GeneChip HG-U95 set. The BloodExpress database **[49]** details information about mouse blood cell expression profiles, including both progenitors and terminally differentiated cells, derived from array experiments and independent studies. Such information allows for the identification of dynamic changes in gene expression during cell differentiation down the hematopoietic hierarchy. Other data sources exist and have been reviewed elsewhere **[50].**

Computational tools for Epigenetic research:
Numerous computational, mathematical and statistical methods, ranging from data mining, sequence analysis, molecular interactions, to complex system-level simulations, have been reported in the literature. Efforts have been channeled into the text mining of epigenetic information, though development in this field is still at an early stage. Current efforts are primarily focused on the extraction and analysis of DNA methylation patterns in various cancer types **[43]**, **[44]**. Traditional sequence analysis tools, such as ClustalW **[51]**, BLAST software suite **[52]**, BLAT (BLAST-Like Alignment Tool) **[53]** and MEGA (Molecular Evolutionary Genetics Analysis) **[54]**, allow for the inference of functional, structural, or evolutionary relationships between DNA or protein sequences. Such methods are employed in diverse applications, and have been applied to homology searches of ortholog candidates for the KEGG/GENES database **[55]**, predicting the secondary structures of histone deacetylases **[56]**, homology modeling of DNA methyltransferases **[57]**, and optimizing the activities of histone deacetylase inhibitors **[58]**.

Computational models have been used extensively to support various epigenome mapping initiatives such as chromatin immunoprecipitation (ChIP)-on-chip **[59]**, ChIP-seq **[60]** and bisulfite sequencing **[61]**. ChIP-on-chip is a microarray-based platform that allows the identification of DNA-protein binding sites on a genome-wide level **[59]**. The main computational tools that have been developed for ChIP-on-chip analysis are focused on the identification of ChIP enrichment sites. Examples include Chromatin ImmunoPrecipitation On Tiled arrays (ChIPOTle) **[62]**, TileMap **[63]** and Ringo **[64]**. ChIP-seq is a variant of ChIP-on-chip that uses high-throughput DNA sequencing for detecting differences between sample and control DNA **[60]**. Although such an approach requires minimal data processing and allows analysis to be made directly from sequence read counts **[65]**, a critical issue that needs to be resolved is the accurate mapping of short sequence reads to the reference genome. Algorithms that can identify regions of similarity between sequences such as BLASTN **[52]** and BLAT **[53]** are valuable for speeding up this process. Efforts are also channeled into the development of specialized algorithms for short-read assembly. Examples include QPalma **[66]** and AMOScmp **[67]**. Bisulfite sequencing **[61]** employs the use of bisulfite treatment of DNA to detect cytosine methylation patterns. Computational tools that focus on bisulfite sequencing include methods for data processing and quality assessment. The basic methods for bisulfite sequence analysis allow the quantitative measurement of cytosine methlylation levels **[68]**, estimating the effectiveness of bisulfite treatment **[68]**, and visualization of results **[69]**. Collectively, the developed algorithms enable the analysis of DNA methylation patterns of different tissue types **[70]** and the genome-wide comparison of histone modification sites identified by various epigenome mapping initiatives **[71]**.

Computational analysis of DNA methylation:
DNA methylation plays an important role in the regulation of genomic stability and cellular plasticity **[72]**. It is essential for normal cell development and is associated with numerous fundamental processes including genomic imprinting **[73]**, X-chromosome inactivation **[74]**, maintenance of repetitive elements **[75]** and carcinogenesis **[76]**. DNA methylation is mainly accomplished by the transfer of methyl groups from *S*-adenosyl-methionine to the 5 position of the cytosine pyrimidine ring in a reaction catalyzed by a DNA methyltransferase or methylase **[77]**. In mammals, four active DNA methyltransferases (DNMT) have been identified, namely DNMT1, 2, 3A and 3B **[78]**, **[79]**. DNMT1 is the most commonly found DNA methyltransferases in mammals, and predominantly methylates hemimethylated CpG di-nucleotides. DNMT2 has been identified as a DNA methyltransferases homolog that methylates cytosine-38 in the anticodon loop of aspartic acid transfer RNA instead of DNA **[80]**, while, DNMT3A and DNMT3B are de novo methyltransferases that act on both hemimethylated and unmethylated CpG sites **[78]**, **[79]**.

A variety of methods for the modeling and prediction of DNA methylation patterns have been reported. An example is the use of linear discriminant analysis and artificial neural networks (ANN) for the *classification of individual lung cancer cell lines* **[81]**. The use of support vector machines (SVM) for genomic mapping of methylation patterns for all 22 human autosomes has also been described **[82]**. In recent years, there has been increasing focus on the development of computational technologies that facilitates the prediction of protein methylation sites. These include procedures based on support vector machines (SVM), hidden Markov models (HMM), ANNs, naïve bayes, logistic regression, K-nearest neighbors and decision trees **[83]**, **[84]**, **[85]**. However, the implementation of such systems is difficult due to the lack of publicly available experimental data for model

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(7): 331-337 (2010)

construction. Many such systems are currently focused on arginine and lysine methylations as their mechanisms are currently the best understood and experimental data is most readily available **[84]**, **[85]**. Epigenome prediction pipeline that integrates DNA methylation, polymerase II preinitiation complex binding, histone H3K4 di- and trimethylation, histone H3K9/14 acetylation, DNase I hypersensitivity and SP1 binding has also been reported **[86]**. Currently, the relative value of such computational tools remains unknown. As experimental data becomes increasingly available, the usefulness of these technologies will become clearer and it should be expected that more integrative models will be made available and that current models will also be refined.

Computational analysis of histone modifications:

Histones are the main protein components of chromatin. They play a key role in the compaction and accessibility of eukaryotic and probably archaeal genomic DNA **[41]**, and are subject to a wide variety of post-translational modifications including methylation, acetylation, phosphorylation, sumoylation and ubiquitylation **[23]**, **[24]**. Covalent modifications of the histone proteins take place primarily within the histone amino-terminal regions that protrude from the surface of the nucleosome as well as the globular core region **[87]**, **[88]**. Histone modifications may affect chromosome function via two distinct mechanisms **[89]**. First, they may alter its electrostatic properties, resulting in a change in the histone structure or its DNA-binding activity. Second, they may generate binding surfaces for protein recognition modules, and help engage specific functional complexes to their relevant sites of action.

There is intense interest in the use of informatics for the analysis, modeling and prediction of histone modifications in DNA sequences. Cellular automata have been described for examining a variety of epigenetic modifications. For example, Sneppen and coworkers **[90]** reported the use of a simplified stochastic model to examine the conditions for bistability and heredity of nucleosome modification-based epigenetic memory. The team demonstrated that robust bistability required cooperativity of two or more modified nucleosomes in the modification reactions, and that nucleosomes occasionally stimulate modifications beyond their neighbor nucleosomes. Comparative genomics has taken advantage of new technologies to help identify histone marks and regulatory elements in higher eukaryotic genomes. Schübeler and colleagues **[91]** have performed a genome-wide comparison of chromatin structures in higher eukaryotes. Their work revealed the existence of a binary pattern of histone modifications among euchromatic genes, with active genes being hyperacetylated at H3/4 and hypermethylated at H3, and inactive genes being hypomethylated and deacetylated at the same locations. Roh *et al*. **[92]** reported a genome-wide mapping technique to determine the distribution of lysine-9/14-diacetylated histone H3 in human peripheral T cells. The team showed that this form of chromatin modification is correlated with active gene promoters and with regulatory elements associated with gene expression. In a follow-up study, the team extended their work to the genome-wide screening of conserved and non-conserved enhancers by histone acetylation patterns **[93]**.

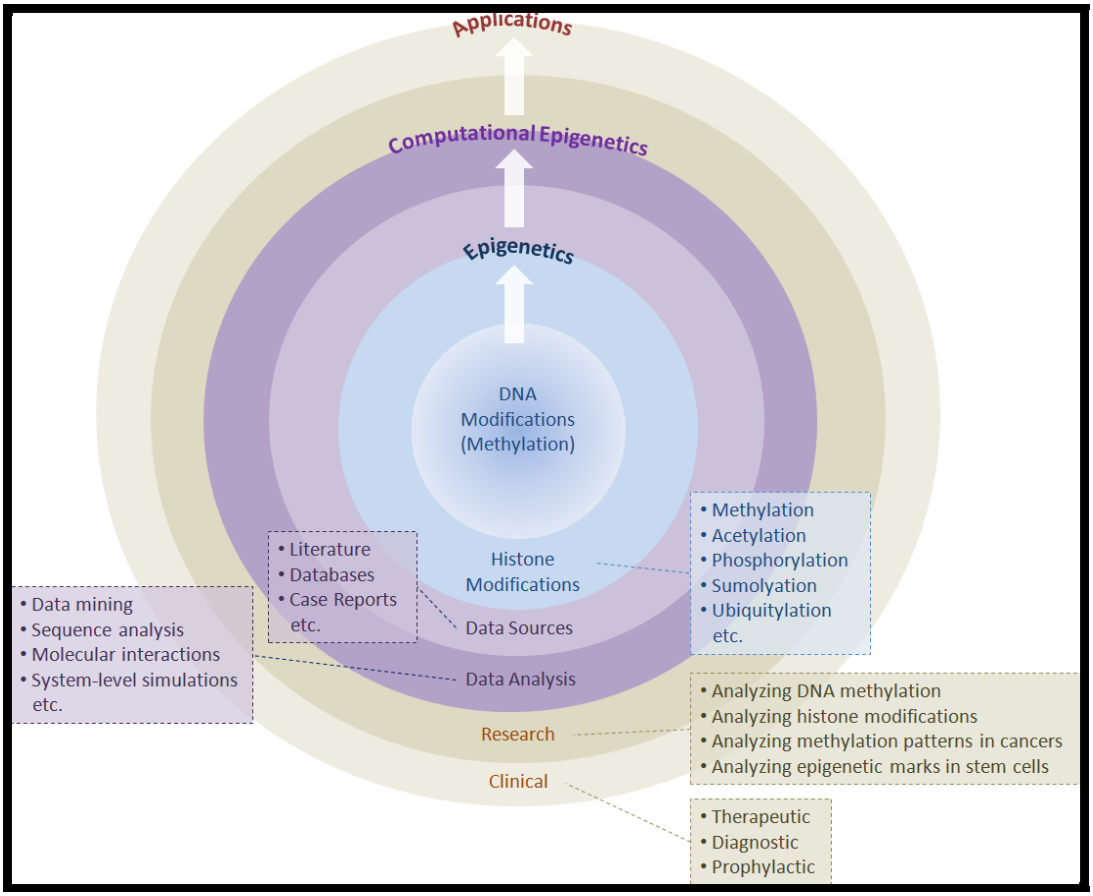

Figure 1: The computational epigenetics paradigm. Existing data sources from epigenetic-related experiments are analyzed with computational strategies and methods for the simulation and prediction of epigenetic patterns. Computational data analysis generates new hypothesis and knowledge for the development of therapeutic, diagnostic and prophylactic clinical tools.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(7): 331-337 (2010)

The design of machine-learning algorithms for locating histone-occupied as well as acetylation, methylation and phosphorylation positions in DNA sequences has also been well reported **[71]**, **[94]**, **[95]**, **[96]**, **[97]**. Some of these tools have also been applied to ChIP-on-chip and ChIP-seq datasets. An example is the use of HMMs to infer the states of histone modification changes at each genomic position based on ChIP fragment counts **[71]**. The use of wavelet analysis combined with HMMs for the discovery of activating and repressive histone modifications using selected ChIP-on-chip datasets from the ENCODE project was also reported **[98]**. These algorithms allow the screening of histone marks in large sets of protein sequences, such as those encoded by the complete genomes of higher complexity organisms. In recent years, a number of structure-based techniques, including quantitative structure-activity relationship (QSAR) analysis **[99]**, **[100]**, homology modeling **[101]** and molecular docking techniques **[102]**, for the design of epigenetic inhibitors were also described. Dynamic activities over the past two years have seen the development of at least five computational methods for the functional annotation of epigenetic factors **[96]**, **[103]**. These tools are particularly useful for the understanding of epigenetic events, both within specific cell types and in an evolutionary context.

Cancer informatics:

Cancer progression is a form of somatic evolution in which certain mutations provide cancer cells with a selective growth advantage. It is now known that DNA methylation patterns in cancers generally display more variation compared with that of normal tissues **[104]**. Several studies have shown that aberrant methylation occurs in a tumor type-specific manner **[105]**. A number of cancer epigenetic projects are currently underway to identify novel methylation patterns that correlate with progression to malignancy. An example is the CancerDip Consortium, a research initiative funded by the 7th Framework Programme for Research and Technological Development of the European Commission (FP7), which employs the use of Methyl-DNA immunoprecipitation (MeDIP) assays for identifying methylation patterns in different tumor types and the epigenetic machinery involved in establishing these abnormal patterns **[106]**. Genome-wide analysis of MeDIP data from colon (Caco-2) and prostate (PC3) cancer as well as several tumor cell lines have shown that tumor-specific methylated genes belong to distinct functional categories, possess common sequence motifs in their promoters and occur in clusters on chromosomes **[107]**. Abnormal DNA methylation within CpG islands is among the most frequent form of alterations in cancers. Experiments have now shown that entire CpG islands may become aberrantly methylated in cancer **[108]**, and is mechanistically linked to histone methylation **[109]**. Bock and coworkers **[110]** have performed a detailed analysis of inter-individual stability and variations of DNA methylation profiles among healthy individuals using linear regression models and the EpiGRAPH web service (http://epigraph.mpi-inf.mpg.de/WebGRAPH/). This work showed that CpG islands may act collectively as emergent and bistable epigenetic switches for maintaining a CpG-island-wide 'on' or 'off' state. An example for tumor class prediction in human cancers was reported by Olek and colleagues **[111]**, in which SVMs were trained to recognize the difference between T and B cell leukaemias and CD19+ B cells and CD4+ T cells obtained from healthy donors using a set of selected CpG sites. Other computational models for classifying cancer subtypes based on epigenetic marks were also reported. Specific examples include the use of SVMs for discriminating between acute lymphoblastic leukemia and acute myeloid leukemia **[112]**, as well as the use of Manhattan distance and average linkage algorithms for hierarchical cluster analysis of human colorectal tumors **[113]**.

Stem cell informatics:

Stem cells are unspecialized cells that can either renew themselves through mitotic cell division or undergo differentiation into more specialized cells **[114]**. Two classes of mammalian stem cells are available: 1) embryonic stem (ES) cells, which are blastocyst-derived, pluripotent cells that can differentiate into all cell types except the extra-embryonic tissue **[114]**, **[115]**, and 2) adult stem cells, which act as a repair system for the body, replenishing specialized cells and regenerating damaged tissues **[116]**. Recent studies have shown that DNA methyltransferases **[117]** and Polycomb/Trithorax group response elements (PRE/TRE) **[118]**, **[119]** possess epigenetic signatures that are important for the differentiation of both human ES cells and germ line stem cells. Of particular interest is the revelation that stem cells are the target cells for cancer, and epigenetic changes may occur long before they are distinguishable as tumor cells **[120]**. By unraveling the nature of epigenetic modifications, it is expected that this will lead to improved culture and differentiation technologies, as well as next-generation drugs that can directly manipulate stem cells in patients.

Bioinformatic analysis of epigenetic marks in stem cells is at its formative stages. Analyses of up- and down-regulated gene clusters provide valuable information on the effect of exogenous control on ES cell state in human. Stanford and colleagues **[121]** have recently performed temporal expression microarray analyses of ES cells after the initiation of commitment and integrated these data with known genome-wide transcription factor binding. This work revealed a repressive model of ES cell maintenance, and helped define the regulatory balance that is needed for maintaining ES cell state. Ringrose and coworkers **[122]** performed an analysis of PRE/TREs in the *Drosophila melanogaster* genome and defined the sequence criteria that distinguish PRE/TREs from non- PRE/TREs. Using a series of weighted motifs, the team identified 167 candidate PRE/TRE sequences, which map to genes involved in development and cell proliferation. Position-specific matrices for predicting *cis*-regulatory elements were also reported, and used for studying PRE/TREs in *Drosophila melanogaster* **[123]**.

Conclusion:

Realizing the full benefits of the informatics revolution will require significant advances in the efficiency of which new data is discovered, processed, interpreted and made accessible to researchers. With the huge amount of epigenetic-related experimental data generated by high throughput methodologies, the future will witness a shift towards the computational epigenetics paradigm **(Figure 1)**. With the paradigm shift, one crucial issue lies in effective data annotation and management. Currently, a centralized repository for epigenetic-related data is still lacking. Resources like such will greatly facilitate computational studies on epigenetics. Another challenge in the field of computational epigenetics lies in the efficient processing of experimental data, which includes normalization and interpretation of data across various experiments from different research groups.

To date, computational algorithms that model different aspects of epigenetic modifications **[81]-[86]** and disease **[107]-[110]** have been described. On the other hand, cellular automata have also been proposed for exploring a variety of epigenetic modifications **[90]**. With the explosion in the number, variety and sophistication of resources and analysis tools, the challenge lies in integrating the strengths and not the weaknesses of each approach. The next few years will see increased interest in the use of cluster computing, central (cloud) computing and distributed systems for large-scale epigenetic data analysis and screening. Computing grid technologies harnessing the resources of multiple computers in a network have been developed rapidly to solve high-throughput scientific research problem **[124]**. On the other hand, cloud computing technologies, which offers scalable resources on demand, have emerged in recent years to complement the rate of data output and drive the rate of data analysis and knowledge discovery **[125]**. The different bioinformatic and mathematical modeling approaches, in combination with advances in computational infrastructures, clearly could lead to improved understanding of epigenetic modifications at multiple levels of complexity, from the sub-cellular molecular level, to the cellular and systems level, and

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(7): 331-337 (2010)

beyond. More importantly, research efforts in computational analysis, identification and classification of variations in epigenetic modifications contribute to further understanding of epigenetic-associated diseases and consequently, the design of relevant diagnostic, therapeutic and prophylactic tools. One exciting possibility, based on the highly combinatorial nature of epigenetics, is the cataloguing of individuals' genome and epigenome and the development of personalized or more specific drugs with lower toxicity and less side effects, paving the way for personalized medicine and a new era of "personal"-omics **[126]**.

References:

[1] A J Webster *et al.*, *Science* (2003) **301:** 478 [PMID:12881561]
[2] A D Cutter, S. Ward, *Mol Biol Evol* (2005) **22:** 178-188 [PMID:15371532]
[3] E Swanton *et al.*, *Proc Natl Acad Sci U S A* (2005) **102:** 4342-4347 [PMID:15753308]
[4] M Nishimura *et al.*, *Nat Med* (1999) **5:** 164-169 [PMID:9930863]
[5] F S Seibert *et al.*, *J Biol Chem* (1996) **271:** 15139-15145 [PMID:8662892]
[6] S A Smith, M. J. Donoghue, *Science* (2008) **322:** 86-89 [PMID:18832643]
[7] J W Sahl *et al.*, *Appl Environ Microbiol* (2008) **74:** 6444-6446 [PMID:18757573]
[8] J Tost, *Horizon Scientific Press, Norwich, UK* (2008)
[9] D Rodenhiser, M. Mann, *CMAJ* (2006) **174:** 341-348 [PMID:16446478]
[10] C H Waddington, *Allen & Unwin, London* (1957)
[11] V E A Russo *et al.*, *Cold Spring Harbor Laboratory Press, Woodbury* (1996)
[12] J Tremblay, P. Hamet, *Metabolism* (2008) **57 Suppl 2:** S27-31 [PMID:18803962]
[13] A P Feinberg, B. Tycko, *Nat Rev Cancer* (2004) **4:** 143-153 [PMID:14732866]
[14] A Petronis *et al.*, *Mol Psychiatry* (2000) **5:** 342-346 [PMID:10889541]
[15] J C Mathers, *Nutrition Bulletin* (2005) **30:** 6-12
[16] E B Keverne, J. P. Curley, *Front Neuroendocrinol* (2008) **29:** 398-412 [PMID:18439660]
[17] V F Chekhun, *Exp Oncol* (2008) **30:** 170 [PMID:19009718]
[18] R P Sharma, *Schizophr Res* (2005) **72:** 79-90 [PMID:15560954]
[19] S Rossignol *et al.*, *J Med Genet* (2006) **43:** 902-907 [PMID:16825435]
[20] S C Wang *et al.*, *PLoS ONE* (2008) **3:** e2698 [PMID:18628954]
[21] C L Peterson, M. A. Laniel, *Curr Biol* (2004) **14:** R546-551 [PMID:15268870]
[22] J Bender, *Annu Rev Plant Biol* (2004) **55:** 41-68 [PMID:15725056]
[23] T Jenuwein, C. D. Allis, *Science* (2001) **293:** 1074-1080 [PMID:11498575]
[24] D Nathan *et al.*, *Proc Natl Acad Sci U S A* (2003) **100:** 13118-13120 [PMID:14597707]
[25] T Mahmoudi, C. P. Verrijzer, *Oncogene* (2001) **20:** 3055-3066 [PMID:11420721]
[26] F F Costa, *Gene* (2008) **410:** 9-17 [PMID:18226475]
[27] B M Turner, *Nat Cell Biol* (2007) **9:** 2-6 [PMID:17199124]
[28] S Henikoff, *Proc Natl Acad Sci U S A* (2005) **102:** 5308-5309 [PMID:15811936]
[29] P A Jones, R. Martienssen, *Cancer Res* (2005) **65:** 11241-11246 [PMID:16357125]
[30] ENCODE Project Consortium, *Science* (2004) **306:** 636-640 [PMID:15499007]
[31] V K Rakyan *et al.*, *PLoS Biol* (2004) **2:** e405 [PMID:15550986]
[32] http://www.heroic-ip.eu
[33] C Bock, T. Lengauer, *Bioinformatics* (2008) **24:** 1-10 [PMID:18024971]
[34] M Krallinger, A. Valencia, *Genome Biol* (2005) **6:** 224 [PMID:15998455]
[35] M Y Galperin, *Nucleic Acids Res* (2008) **36:** D2-4 [PMID:18025043]
[36] D A Benson *et al.*, *Nucleic Acids Res* (2002) **30:** 17-20 [PMID:11752243]
[37] Y Tateno *et al.*, *Nucleic Acids Res* (2002) **30:** 27-30 [PMID:11752245]
[38] C Kanz *et al.*, *Nucleic Acids Res* (2005) **33:** D29-33 [PMID:15608199]
[39] V Negre, C. Grunau, *Epigenetics* (2006) **1:** 101-105 [PMID:17965614]
[40] F Pattyn *et al.*, *BMC Bioinformatics* (2006) **7:** 496 [PMID:17094804]
[41] S A Sullivan *et al.*, *Nucleic Acids Res* (2000) **28:** 320-322 [PMID:10592260]
[42] S Sullivan *et al.*, *Nucleic Acids Res* (2002) **30:** 341-342 [PMID:11752331]
[43] M Ongenaert *et al.*, *Nucleic Acids Res* (2008) **36:** D842-846 [PMID:17932060]
[44] Y C Fang *et al.*, *BMC Bioinformatics* (2008) **9:** 22 [PMID:18194557]
[45] T Barrett *et al.*, *Nucleic Acids Res* (2009) **37:** D885-890 [PMID:18940857]
[46] N Heintz, *Nat Neurosci* (2004) **7:** 483 [PMID:15114362]
[47] C J Porter *et al.*, *Methods Mol Biol* (2007) **407:** 137-148 [PMID:18453254]
[48] O Shmueli *et al.*, *C R Biol* (2003) **326:** 1067-1072 [PMID:14744114]
[49] D Miranda-Saavedra *et al.*, *Nucleic Acids Res* (2009) **37:** D873-879 [PMID:18987008]
[50] A D Baxevanis, *Curr Protoc Bioinformatics* (2006) **Chapter 1:** Unit 1 1 [PMID:18428753]
[51] J D Thompson *et al.*, *Nucleic Acids Res* (1994) **22:** 4673-4680 [PMID:7984417]
[52] S F Altschul *et al.*, *J Mol Biol* (1990) **215:** 403-410 [PMID:2231712]
[53] W J Kent, *Genome Res* (2002) **12:** 656-664 [PMID:11932250]
[54] S Kumar *et al.*, *Brief Bioinform* (2008) **9:** 299-306 [PMID:18417537]
[55] M Itoh *et al.*, *Bioinformatics* (2005) **21:** 912-921 [PMID:15509606]
[56] L Aravind *et al.*, *Science* (1998) **280:** 1167
[57] E V Koudan *et al.*, *J Biomol Struct Dyn* (2004) **22:** 339-345 [PMID:15473707]
[58] H Park, S. Lee, *J Comput Aided Mol Des* (2004) **18:** 375-388 [PMID:15662999]
[59] M J Buck, J. D. Lieb, *Genomics* (2004) **83:** 349-360 [PMID:14986705]
[60] T S Mikkelsen *et al.*, *Nature* (2007) **448:** 553-560 [PMID:17603471]
[61] P Hajkova *et al.*, *Methods Mol Biol* (2002) **200:** 143-154 [PMID:11951649]
[62] M J Buck *et al.*, *Genome Biol* (2005) **6:** R97 [PMID:16277752]
[63] H Ji, W H Wong, *Bioinformatics* (2005) **21:** 3629-3636 [PMID:16046496]
[64] J Toedling *et al.*, *BMC Bioinformatics* (2007) **8:** 221 [PMID:17594472]
[65] A Barski *et al.*, *Cell* (2007) **129:** 823-837 [PMID:17512414]

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(7): 331-337 (2010)

[66] F De Bona *et al.*, *Bioinformatics* (2008) **24:** i174-180 [PMID:18689821]
[67] M Pop *et al.*, *Brief Bioinform* (2004) **5:** 237-248 [PMID:15383210]
[68] J Lewin *et al.*, *Bioinformatics* (2004) **20:** 3005-3012 [PMID:15247106]
[69] L A Boyer *et al.*, *Cell* (2005) **122:** 947-956 [PMID:16153702]
[70] Human Epigenome Consortium *et al.*, (2003)
[71] H Xu *et al.*, *Bioinformatics* (2008) **24:** 2344-2349 [PMID:18667444]
[72] R L Adams, *Biochem J* (1990) **265:** 309-320 [PMID:2405840]
[73] E Li *et al.*, *Nature* (1993) **366:** 362-365 [PMID:8247133]
[74] D C Kaslow, B. R. Migeon, *Proc Natl Acad Sci U S A* (1987) **84:** 6210-6214 [PMID:3476942]
[75] G Liang *et al.*, *Mol Cell Biol* (2002) **22:** 480-491 [PMID:11756544]
[76] P A Jones, *Oncogene* (2002) **21:** 5358-5360 [PMID:12154398]
[77] R L P Adams, R.H. Burdon, *CRC Press, Boca Raton, FL* (1983) **1:** 119-144
[78] K D Robertson *et al.*, *Nucleic Acids Res* (1999) **27:** 2291-2298 [PMID:10325416]
[79] S Xie *et al.*, *Gene* (1999) **236:** 87-95 [PMID:10433969]
[80] M G Goll *et al.*, *Science* (2006) **311:** 395-398 [PMID:16424344]
[81] A M Marchevsky *et al.*, *J Mol Diagn* (2004) **6:** 28-36 [PMID:14736824]
[82] R Das *et al.*, *Proc Natl Acad Sci U S A* (2006) **103:** 10713-10716 [PMID:16818882]
[83] M Bhasin *et al.*, *FEBS Lett* (2005) **579:** 4302-4308 [PMID:16051225]
[84] H Chen *et al.*, *Nucleic Acids Res* (2006) **34:** W249-253 [PMID:16845004]
[85] K M Daily *et al.*, *IEEE, CIBCB 2005, San Diego, California, USA, November 2005* (2005) 475-481
[86] C Bock *et al.*, *PLoS Comput Biol* (2007) **3:** e110 [PMID:17559301]
[87] P A Grant, *Genome Biol* (2001) **2:** REVIEWS0003 [PMID:11305943]
[88] A Vaquero *et al.*, *Sci Aging Knowledge Environ* (2003) **2003:** RE4 [PMID:12844523]
[89] M Iizuka, M. M. Smith, *Curr Opin Genet Dev* (2003) **13:** 154-160 [PMID:12672492]
[90] I B Dodd *et al.*, *Cell* (2007) **129:** 813-822 [PMID:17512413]
[91] D Schubeler *et al.*, *Genes Dev* (2004) **18:** 1263-1271 [PMID:15175259]
[92] T Y Roh *et al.*, *Genes Dev* (2005) **19:** 542-552 [PMID:15706033]
[93] T Y Roh *et al.*, *Genome Res* (2007) **17:** 74-81 [PMID:17135569]
[94] T H Pham *et al.*, *IEEE International Conference on Bioinformatics and Bioengineering, Boston, MA* (2007) 959-966
[95] K J Won *et al.*, *BMC Bioinformatics* (2008) **9:** 547 [PMID:19094206]
[96] D Miranda-Saavedra, G. J. Barton, *Proteins* (2007) **68:** 893-914 [PMID:17557329]
[97] I Kouskoumvekaki *et al.*, *SAR QSAR Environ Res* (2008) **19:** 167-177 [PMID:18311642]
[98] R E Thurman *et al.*, *Genome Res* (2007) **17:** 917-927 [PMID:17568007]
[99] N Dessalew, *Med. Chem. Res.* (2007) **16:** 449-460
[100] D C Juvale *et al.*, *Org Biomol Chem* (2006) **4:** 2858-2868 [PMID:16855733]
[101] Y C Lin *et al.*, *Cancer Res* (2008) **68:** 2375-2383 [PMID:18381445]
[102] E E Angeles *et al.*, *Lett. Drug Des. Discov.* (2005) **4:** 282-286
[103] N D Trinklein *et al.*, *Genome Res* (2007) **17:** 720-731 [PMID:17567992]
[104] G Strathdee, R. Brown, *Expert Rev Mol Med* (2002) **4:** 1-17 [PMID:14987388]
[105] T R Golub *et al.*, *Science* (1999) **286:** 531-537 [PMID:10521349]
[106] F V Jacinto *et al.*, *Biotechniques* (2008) **44:** 35, 37, 39 passim [PMID:18254377]
[107] I Keshet *et al.*, *Nat Genet* (2006) **38:** 149-153 [PMID:16444255]
[108] P W Laird, *Hum Mol Genet* (2005) **14 Spec No 1:** R65-76 [PMID:15809275]
[109] E Eden *et al.*, *PLoS Comput Biol* (2007) **3:** e39 [PMID:17381235]
[110] C Bock *et al.*, *Nucleic Acids Res* (2008) **36:** e55 [PMID:18413340]
[111] P Adorjan *et al.*, *Nucleic Acids Res* (2002) **30:** e21 [PMID:11861926]
[112] F Model *et al.*, *Bioinformatics* (2001) **17 Suppl 1:** S157-164 [PMID:11473005]
[113] D J Weisenberger *et al.*, *Nat Genet* (2006) **38:** 787-793 [PMID:16804544]
[114] M Stojkovic *et al.*, *Reproduction* (2004) **128:** 259-267 [PMID:15333777]
[115] L Duplomb *et al.*, *Stem Cells* (2007) **25:** 544-552 [PMID:17095705]
[116] M Raff, *Annu Rev Cell Dev Biol* (2003) **19:** 1-22 [PMID:14570561]
[117] M Bibikova *et al.*, *Genome Res* (2006) **16:** 1075-1083 [PMID:16899657]
[118] L A Boyer *et al.*, *Nature* (2006) **441:** 349-353 [PMID:16625203]
[119] X Chen *et al.*, *Science* (2005) **310:** 869-872 [PMID:16272126]
[120] A P Feinberg *et al.*, *Nat Rev Genet* (2006) **7:** 21-33 [PMID:16369569]
[121] E Walker *et al.*, *Cell Stem Cell* (2007) **1:** 71-86 [PMID:18371337]
[122] L Ringrose *et al.*, *Dev Cell* (2003) **5:** 759-771 [PMID:14602076]
[123] T Fiedler, M. Rehmsmeier, *Nucleic Acids Res* (2006) **34:** W546-550 [PMID:16845067]
[124] V Talukdar *et al.*, *Biotechnol J* (2009) **4:** 1244-1252 [PMID:19579217]
[125] A Rosenthal *et al.*, *J Biomed Inform* (2009) [PMID:19715773]
[126] F F Costa, *BioForum Europe* (2009) **13:** 29-31
[127] R J Roberts *et al.*, *Nucleic Acids Res* (2005) **33:** D230-232 [PMID:15608184]
[128] K Gendler *et al.*, *Nucleic Acids Res* (2008) **36:** D298-302 [PMID:17942414]
[129] A Shipra *et al.*, *Bioinformatics* (2006) **22:** 2940-2944 [PMID:17021159]

Edited by P. Shapshak
Lim *et al.*, Bioinformation, 4 (7) 331-337 (2010)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(7): 331-337 (2010)

Supplementary material

Table 1. Some epigenetic and related databases reviewed in this article**.**

Database	Description	URL	Ref
MethDB	Contains information on 19,905 DNA methylation content data and 5,382 methylation patterns for 48 species, 1,511 individuals, 198 tissues and cell lines and 79 phenotypes.	http://www.methdb.de	**[39]**
PubMeth	Contains over 5,000 records on methylated genes in various cancer types.	*www.pubmeth.org/*	**[43]**
REBASE	Contains over 22,000 DNA methyltransferases genes derived from GenBank.	http://rebase.neb.com/rebase/rebase.html	**[127]**
MeInfoText	Contains gene methylation information across 205 human cancer types.	http://mit.lifescience.ntu.edu.tw/	**[44]**
MethPrimerDB	Contains 259 primer sets from human, mouse and rat for DNA methylation analysis.	medgen.ugent.be/methprimerdb/	**[40]**
The Histone Database	Contains 254 sequences from histone H1, 383 from histone H2, 311 from histone H2B, 1043 from histone H3 and 198 from histone H4, altogether representing at least 857 species.	http://genome.nhgri.nih.gov/histones/	[**42**]
ChromDB	Contains 9,341 chromatin-associated proteins, including RNAi-associated proteins, for a broad range of organisms.	http://www.chromdb.org/	**[128]**
CREMOFAC	Contains 1725 redundant and 720 non-redundant chromatin-remodeling factor sequences in eukaryotes.	http://www.jncasr.ac.in/cremofac/	**[129]**
The Krembil Family Epigenetics Laboratory	Contains DNA methylation data of human chromosomes 21, 22, male germ cells and DNA methylation profiles in monozygotic and dizygotic twins.	http://www.epigenomics.ca	–
MethyLogiX DNA methylation database	Contains DNA methylation data of human chromosomes 21 and 22, male germ cells and late-onset Alzheimer's disease.	http://www.methylogix.com/genetics/database.shtml.htm	**[20]**

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(7): 331-337 (2010)