Medical Informatics in a United and Healthy Europe
K.-P. Adlassnig et al. (Eds.)
IOS Press, 2009

doi:10.3233/978-1-60750-044-5-135

Relevance and Management of Methylation Data in Electronic Health Records

Georg GOEBEL [a,1], Karl P. PFEIFFER [a], Thomas SCHABETSBERGER [b], Claudius KALOZY [c], Heidi FIEGL [d], Karin LEITNER [e]
[a] *Department for Medical Statistics, Informatics and Health Economics, Medical University Innsbruck, Austria*
[b] *University of Health Sciences Medical Informatics and Technology, Research Division for eHealth and Telemedicine, Hall in Tirol, Austria*
[c] *Center of Excellence in Medicine and IT, Innsbruck, Austria*
[d] *Department of Gynaecology and Obstetrics, Medical University Innsbruck, Austria*
[e] *Tiroler Landeskrankenanstalten Gmbh, Innsbruck, Austria*

Abstract. In this paper we give an overview of challenges and chances of the integration of methylation data into Electronic Health Records. Perspectives of methylation data in terms of clinical relevance and characteristics of these data are shown. Among several standards OpenEHR, HL7 CG and LOINC are identified as starting point for the representation and communication of methylation data within the clinical context.

Keywords. hospital information system, epigenomics, DNA methylation, electronic health record, genetic services

1. Introduction

DNA methylation is a modification of the DNA that encodes for epigenetic information. It is the most common and best characterised epigenetic abnormality in human malignancies. DNA methylation is an intensively investigated epigenetic research topic and its relevance for the prognostic and predictive outcome has been shown in several clinical areas, especially in the genesis and therapeutic processing of cancer and (recently published) in psychiatric disorders [1]. Thus, the systematic integration of a patient's methylation profile or pattern into his/her EHR will become relevant in the future and may help in the diagnosis, prevention and treatment of patients with these disorders. Additionally, links between bio-(data)banks, which are very often built for research purposes, and clinical information systems need also to be addressed in this context. This fact was recently published as one among three main challenges for Medical Informatics in the context of Health Information Systems [2].

[1] Corresponding Author: Dr. Mag. Georg Goebel, Department of Medical Statistics, Informatics and Health Economics, Medical University of Innsbruck, Schoepfstrasse 41, A-6020 Innsbruck, Austria; E-mail: Georg.Goebel@i-med.ac.at.

2. Methods

After a short explanation of the term *DNA methylation* we analyse the perspectives for the use of epigenetic data in the clinical context. Secondly, we show characteristics of epigenetic data in biomedical information modelling and we analyse standards suitable for storage/processing of epigenetic information based on biobank functional needs [3]. Finally, we describe the potential utilisation of the selected standards for the representation of methylation data.

3. Results

3.1. Clinical Usage of DNA Methylation – State of the Art

DNA methylation is the addition of a methyl group to the carbon-5 position of cytosine residues. It is the only common covalent modification of human DNA and occurs almost exclusively at cytosines that are followed immediately by a guanine (CpG pairs). CpG pairs are underrepresented in the genome except for CpG islands (CGI), which are located in the promoter region of approximately half of all genes [4].

Currently no widely accepted standardised usage of DNA methylation data for diagnostics or therapeutic decision making could be found in the literature, whereas dozens of research publications have shown the clinical relevance of DNA methylation in cancer diseases. There is also increasing evidence that epigenetic mechanisms play a major role in the pathogenesis of psychiatric disorders [1]. Actually a few Phase I/II studies were identified, where different DNA-hypomethylating agents are tested in patients with advanced cancer [5, 6].

3.2. From Bench to Bedside: Perspectives for Clinical Usage of Methylation Biomarkers in the Future

DNA methylation has been considered a promising biomarker for several reasons [7].

First, unlike DNA mutations, methylation always occurs in defined regions (i.e., CpG islands) and can be detected using techniques with high sensitivity and high resolution. Second, hypermethylated DNA is associated with virtually every type of tumor, with each type of tumor apparently having its own signature of methylated genes, such as the methylation of *GSTP1* in prostate cancer, von Hippel-Lindau gene in renal cancer, the mismatch repair gene *MLH1* in colon cancer. In addition, some methylation aberrations occur early in cancer development.

The introduction of clinical (bio-)markers differs from the regulated and formalised process of therapeutic drug development [8]. Due to this issue, it is difficult to speculate about timelines concerning the acceptance of methylation biomarkers in the daily clinical routine. Nevertheless one can expect that information about a patient's DNA methylation status will be used in clinical workaday life like the use of other (genetic) biomarkers (e.g., HER/2, BRCA1/2).

3.3. Modelling of Methylation Data

Methylation data are representable in adequate XML-based data-models [9, 10], as their details range from ASCII streams in case of sequencing results to simple scalars (per genetic locus or region of CpG Islands) in case of qRT-PCR analysis. Additionally phenotype data and descriptive data about the patient and methods used for the analysis must be managed. In many research contexts very detailed extended information are generated (e.g., type of tissue, data about the whole experiment) to assure that experiments are completely reproducable (e.g., observer's name, primer information, etc.). There are several characteristics that might help to distinguish these data from typical clinical observations such as blood pressure or glucose levels:

- The amount of data.
  Typically, a single genetic locus or a region of CpG Islands is not sufficient.
- The complexity of the data.
  The DNA sequences (… AGCT…) need to be represented along with their variations, transcription outcome, and translation to proteins.
- Detailed descriptions of the methods used to obtain the data.
  These are necessary for the receiver to interpret the data correctly.
- The semantics of the genotype-phenotype relations, which are represented in a variety of ways, depending on the point of view (clinical research/trial, pharmaceutical, or health care).
- In contrast to genomic data methylation data of an individual may change by time, kind of biomaterial of a patient and other life-related circumstances.

3.4. Existing Standards [9, 11–14]

We identified several relevant standards for storage / processing of methylation data which can support semantics in a clinical context [11] and mapped them against six important domains identified by [3]; the domains were identified as major elements of an organizing bio-banking framework (please see Table 1). For the representation of epigenomic data the clinical and molecular dimensions were used as selection criteria together with suggestions of Sax and Schmidt [11].

Table 1. Relevant standards (rows) for storage / processing of methylation data are mapped against six important domains (columns) and reviewed for their usability in the context of methylation data.

Standard	Clinical	Sample	Molecular	Annotation	Analysis	Interpretation
MAGE-ML [14]		x	x	x	x	
CDISC	x					
HL7v3 CG	x		x			
HL7 CDA	x					
OpenEHR [15]	x		x	x		
LOINC	x		x			
SNOMED CT	x	x		x		x

Within these standards we found two relevant standards for representation of methylation data in an EHR: HL7 Clinical Genomics and OpenEHR [13]. Both approaches provide a domain-knowledge-based approach to the representation of clinical information.

Roughly spoken the OpenEHR approach uses archetypes to express clinical information. They are reusable, formal models of domain concepts and can be defined by the CEN- and ISO-standardised "archetype definition language" (ADL). Garde [13] has demonstrated how archetypes and templates may be used to facilitate the use of legacy health record and message data in an OpenEHR health record system, and output standardised messages and HL7 CDA documents. To our knowledge no archetype has been defined in the context of (epi-)genetic clinical information yet.

The HL7v3 standard is also based on object-oriented principles supported by its Reference Information Model (RIM) [12]. The HL7 Clinical Genomics SIG has been developing the HL7 Genotype model since 2003 [9]. The actual model encapsulates the following categories to represent genomic data:

- Locus / Allele
- Sequence / Proteomics
- Expression Data
- Sequence Variation
- Clinical Phenotype

With extensions the HL7 Genotype Model provides the representation of raw genetic and also epigenetic data and it allows wrapping the data and annotating them with metadata.

As relevant standards for the communication of epigenetic reports between organisations or clinical departments, LOINC and HL7 CDA were identified. Both standards are well known and have already been widely used in laboratory information systems and clinical information systems, respectively; a detailed explanation can be found at www.loinc.org and in [15].

4. Discussion

The role of aberrant DNA methylation in cancer and other human diseases has been persuasively argued, but causality is notoriously difficult to establish and thus diagnostic and therapeutic applicability must be based on more clinical evidence. For patient-centered data integration we will need information about used methods and laboratory meta-data as well as comprehensive methylation data [10]. Therefore we need strategies to link an individual's EHR to (epi-)genomic data stored in research databases to provide longitudinal knowledge about the patient and his/her conditions and outcome in the future. Genome-Wide Association Studies (GWAs) and pharmacogenomic clinical trials are already performed by using epigenomic information. Additionally new findings about biomarkers could be easier applied, if detailed information about one's epigenome is available.

In terms of data modelling HL7 CG and OpenEHR seem to be the most relevant standards. Within both approaches additional work has to be done, because no detailed representations of epigenetic data in the clinical context have been implemented up to now. In order to evaluate drawbacks and advantages of these approaches each standard must further be investigated.

On the communication level, an epigenetic report could be built according to the HL7 CDA standard. Following Sax and Schmidt in [11] information about the document type as well as the universal observation identifiers could be defined according to the LOINC standard in future.

5. Conclusions and Future Outlook

The integration of laboratory-results based on methylation or epigenetic analyses will be a necessity after the approval of epigenetic testing for clinical usage. One challenge will be the development and application of appropriate standards for representation and communication of the data. Especially OpenEHR and HL7 CG are useful approaches already available. A further task might be the linkage between the EHR and research-oriented (epi-)genomic databases, where semantic and functional issues as well as ethical and privacy concerns must be considered to ensure trust and acceptance of clinicians, researchers and last but not least of citizens and patients.

References

[1] Peedicayil, J. (2008) Epigenetic biomarkers in psychiatric disorders. *British Journal of Pharmacology* 155 (6): 795–796.
[2] Prokosch, H.-U., Ganslandt, T. (2009) Perspectives for medical informatics. Reusing the electronic medical record for clinical research. *Methods of Information in Medicine* 48(1):38–44.
[3] Lincoln, S. (2005) Report on Worldwide Biobank Summit II 2005, p44, http://www-03.ibm.com/industries/global/files/Biobanks.pdf.
[4] Antequera, F. and Bird, A. (1993) Number of CpG islands and genes in human and mouse. *Proceedings of the National Academy of Sciences of the United States of America* 90(24):11995–11999.
[5] Braiteh, F., Soriano, A.O., Garcia-Manero, G. et al. (2008) Phase I study of epigenetic modulation with 5-azacytidine and valproic acid in patients with advanced cancers. *Clinical Cancer Research* 14(19):6296–6301.
[6] Odenike, O.M., Alkan, S., Sher, D. et al. (2008) Histone deacetylase inhibitor romidepsin has differential activity in core binding factor acute myeloid leukemia. *Clinical Cancer Research* 14(21):7095–7101.
[7] Li, L.-C., Carroll, P.R., Dahiya, R. (2005) Epigenetic Changes in Prostate Cancer: Implication for Diagnosis and Treatment. *Journal of the National Cancer Institute* 97(2):103–115.
[8] Pepe M.S. (2001) Phases of Biomarker Development for Early Detection of Cancer. *Journal of the National Cancer Institute* 93(14):1054–1106.
[9] CGL7 Clinical Genomics Level 7, HL7 CG SIG, http://www.haifa.ibm.com/projects/software/cgl7.
[10] Goebel, G., Müller, H.M., Fiegl, H., Widschwendter, M. (2005) Gene methylation data—a new challenge for bioinformaticians? *Methods of Information in Medicine* 44(4):516–519.
[11] Sax, U., Schmidt, S. (2005) Integration of genomic data in Electronic Health Records—opportunities and dilemmas. *Methods of Information in Medicine* 44(4):546–550.
[12] Shabo, A. (2008) Integrating genomics into clinical practice: standards and regulatory challenges. *Current Opinion in Molecular Therapeutics* 10(3):267–272.
[13] Garde, S., Knaup, P., Hovenga, E.J.S., Heard, S. (2007) Towards semantic interoperability for electronic health records. *Methods of Information in Medicine* 46(3):332–343.
[14] Spellman, P.T., Miller, M., Stewart, J. et al. (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biology* 3(9):research0046.1–0046.9.
[15] Dolin, R.H., Alschuler, L., Boyer, S. et al. (2006) HL7 Clinical Document Architecture, Release 2. *Journal of the American Medical Informatics Association* 13(1):30–39.