Association Studies on Cervical Cancer Facilitated by Inference and Semantic Technologies: The ASSIST Approach

Pericles MITKAS [a,b,1], Vassilis KOUTKIAS [a], Andreas SYMEONIDIS [b], Manolis FALELAKIS [a,b], Christos DIOU [a,b], Irini LEKKA [a], Anastasios DELOPOULOS [a,b], Theodoros AGORASTOS [a] and Nicos MAGLAVERAS [a]
[a] *Aristotle University, Thessaloniki, GREECE*
[b] *CERTH, Informatics and Telematics Institute, Thessaloniki, GREECE*

Abstract. Cervical cancer (CxCa) is currently the second leading cause of cancer-related deaths, for women between 20 and 39 years old. As infection by the human papillomavirus (HPV) is considered as the central risk factor for CxCa, current research focuses on the role of specific genetic and environmental factors in determining HPV persistence and subsequent progression of the disease. ASSIST is an EU-funded research project that aims to facilitate the design and execution of genetic association studies on CxCa in a systematic way by adopting inference and semantic technologies. Toward this goal, ASSIST provides the means for seamless integration and virtual unification of distributed and heterogeneous CxCa data repositories, and the underlying mechanisms to undertake the entire process of expressing and statistically evaluating medical hypotheses based on the collected data in order to generate medically important associations. The ultimate goal for ASSIST is to foster the biomedical research community by providing an open, integrated and collaborative framework to facilitate genetic association studies.

Keywords. Association Studies, Integration of Biomedical Information, Semantic Technologies, Cervical Cancer, Biomedical Informatics

Introduction

Cervical cancer (CxCa) is the second leading cause of cancer-related deaths after breast cancer, for women between 20 and 39 years old [1]. Infection by the human papillomavirus (HPV) is considered as the central risk factor for CxCa [2]. However, it is unlikely to be the sole cause for developing cancer. Ongoing research investigates the role of specific genetic and environmental factors in determining HPV persistence and subsequent progression of the disease [3]. In this context, genetic association studies constitute a significant scientific approach that may lead to a more comprehensive and holistic insight on the origin of complex diseases, such as CxCa [4]. Genetic association studies aim to detect association between one or more genetic variants (e.g. polymorphisms) and a trait, which might be some quantitative characteristic, a discrete attribute, or a disease [5]. A genetic variant is genotyped in a

[1] Corresponding Author: Prof. Pericles Mitkas, Department of Electrical & Computer Engineering, Faculty of Engineering, Aristotle University, 54124, Thessaloniki Greece; Email: mitkas@auth.gr.

population for which phenotypic information is available (such as disease occurrence, or a range of different trait values). If a correlation is observed between genotype and phenotype, there is an association between the variant and the disease or trait [6].

Nevertheless, association studies are most of the times inconclusive, since the datasets employed are small, usually incomplete and of poor quality. In this regard, the ASSIST EU-funded research project aims to provide researchers with an integrated environment enabling association studies among genetic characteristics, environmental agents and viral factors, which can suggest pathogenetic mechanisms that will provide new markers of risk for CxCa. Its overall objective is to offer a new technological solution that will virtually unify multiple patient record repositories (physically located at different laboratories, clinics and/or hospitals) containing both genotypic and phenotypic data, thus, enabling researchers to utilize existing patient data from several clinics and perform research in a low-cost and time-efficient way.

In this paper, the overall system architecture and the underlying design conceptualization of ASSIST is presented, emphasising also on the medical data coding and the medical rules employed towards the unification of biomedical data and the semantic inference of medical knowledge. ASSIST's functionality is illustrated via an example usage scenario, while potential extensions and remarks on its virtue conclude the paper.

1. Rationale

The number of studies elaborating on phenotype-genotype associations for common diseases is rapidly increasing; however, several studies show variation in the underlying association between genotype and outcome between the populations studied, resulting in questionable findings. It is evident that reliable association studies require large sets of patient phenotypic and genotypic data, all provided in a structured format. In addition, significant drawbacks in current association studies are considered as the lack of standardisation in data collection, the lack of a standardised overall methodology, cost (mainly for genetic tests), time consumption and man power, failure to attempt replication of results, and the incorporation of 'disposable' study groups.

Evidently, this rather limited progress in the field is mainly due to the problems in unification and utilization of data stemming from several similar yet 'isolated' studies. Clearly, methodologies for data and system interoperability, assisted by the appropriate technologies, will alleviate some of the problems mentioned above. While standardisation of phenotypic data and clinical practices seems unrealistic, semantic annotation and transformations to widely agreed classification schemes may provide workable solutions. This approach constitutes the basis for ASSIST, which aims to be used primarily by biomedical researchers for the identification of new markers of risk, diagnosis and prognosis, and possibly treatment of CxCa via the implementation of case-control association studies.

In essence, three types of usage scenarios are supported in ASSIST, i.e. multi-criteria based data retrieval from the available medical archives, design and implementation of an association study using existing patient data from different medical institutions, and assessment of patient risk for development of CxCa based on similar available patient cases. More specifically, typical examples of use are:

- What percentage of women who have been diagnosed with *LCIN* (Low grade Cervical Intraepithelial Neoplasia) are smokers?

– Find CxCa subjects with *MTHFR* (MethyleneTetraHydroFolate Reductase) data.
– What is the association of the *MTHFR C/T* polymorphism and smoking with the development of CxCa?
– What is the risk for patients with low risk *HPV* infection and *MTHFR C/C* genotype to develop *LCIN*?

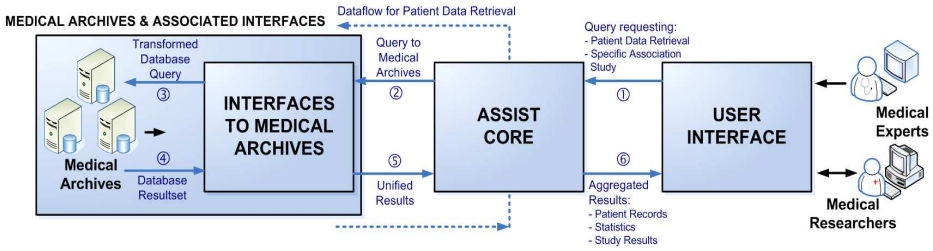

Figure 1. ASSIST framework overview and dataflow between ASSIST subsystems

2. Core Technologies and System Components

From a technical viewpoint, ASSIST aims at: i) the design of a unification mechanism of multiple patient record repositories at the syntactic and semantic level, ii) the conformance to regulations and local data access policies at each participating clinical site, iii) the specification of an automated and transparent data retrieval mechanism to support datasets definition for case-control association studies [5-6] and/or the evaluation of medical hypotheses to assess patient risk for development of CxCa, iv) the development of an inference mechanism capable of statistically evaluating medical hypotheses, and v) the design of expressive and friendly tools to perform association studies. In this regard, ASSIST follows a semantic approach [7], and resorts to medical data virtual unification and inferencing applied on real patient data. Figure 1 provides an overview of the system and the dataflow between its three subsystems, namely:

– The **Medical Archives and Associated Interfaces** subsystem. The Medical Archives that are available to ASSIST constitute research-oriented data repositories related to CxCa that ASSIST has been granted full access to. These repositories are not identical to the Hospital Information System that corresponds to each ASSIST site, but might contain a part of its hosted information, and contain only anonymised patient data, in full compliance to the legal national and EU medical research requirements and code of practice, that are generated by an anonymisation tool developed within the context of the project. ASSIST considers heterogeneous and legacy patient data repositories for integration. For each Medical Archive, the Associated Interfaces map the contained information to corresponding entities as defined in the knowledge model of ASSIST. The interfacing part between the Medical Archives and the rest of the ASSIST system ensures transparent and uniform access to patient data, confronting this way the semantic and syntactic heterogeneity of the data sources incorporated in ASSIST.
– The **ASSIST Core** subsystem. It constitutes the intermediate between the former subsystem and the User Interfaces subsystem. Based on the information provided by the Medical Archives and Associated Interfaces, the Core subsystem infers knowledge about patients [8], and offers retrieval (query answering) and data analysis (statistical) services, supporting the definition, execution and

management of association studies. It also incorporates mechanisms for handling uncertainty in data retrieval as well as query expansion and validity assessment, while it takes into account optimisation of response time with respect to the inference and retrieval procedures. Equally important, it provides the query schema to the User Interfaces subsystem and the result schema for the Associated Interfaces via its Medical Knowledge Base.

– The **User Interfaces** (UI) subsystem. It enables transparent and advanced access to the CxCa related data repositories incorporated in ASSIST. Specifically, it offers query expression as well as patient data and statistical results visualisation to the ASSIST end-users. It also incorporates modules for user profile management, i.e. a users' database, profile and session management modules, etc., as well as the means to search for previous association studies that have been performed via ASSIST. In general, the UI subsystem constitutes the front-end of the users for accessing the ASSIST services, offering a comprehensive and integrated environment for conducting association studies related research.

Figure 2 illustrates the overall architecture of ASSIST, detailed at the modules level, as well as the data exchange means between subsystems.

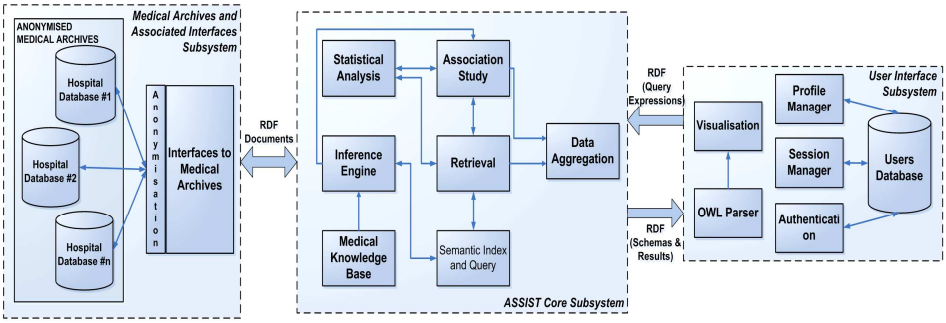

Figure 2. ASSIST overall system architecture (detailed at modules level) and data workflow

3. Medical Data Coding and Medical Rules

Currently, ASSIST integrates patient data originated from three gynaecology clinics located in Belgium, Germany and Greece, resulting in more than a thousand records. Each data repository has its own autonomous schema and captures patient information related to CxCa, according to the local clinical procedures followed. ASSIST virtually unifies these heterogeneous data sources syntactically, by defining a common data schema, and semantically via a domain ontology for CxCa that encapsulates the appropriate medical knowledge to model the disease towards inferring its severity index via the available diagnostic and therapeutic information (cytology, colposcopy, histology and histopathology), the HPV infection results, the genetic profile based on putative polymorphisms for CxCa, as well as lifestyle and personal information. As far as the examinations are concerned, a uniform coding has been elaborated for each one, so that the underlying heterogeneity in each clinic site is addressed. This uniform classification of exam results is then employed in the context of medical rules for the inference part of the system. Figure 3 illustrates the heterogeneity of example CxCa-related biomedical data handled by ASSIST, focusing on the various coding schemes for cytology and the proposed ASSIST common coding (virtual unification part).

ASSIST considers polymorphisms of *p53*, *MTHFR*, *CYP1A1*, *CYP2E1*, *GSTM1* and *GSTT1* genes as genetic markers on the basis of their involvement in the defence against viral infections and tumor growth, as well as the number of published articles that have reported positive correlation between CxCa and genetic markers. ASSIST handles data on HPV types, as well as lifestyle and personal quality attributes, e.g. smoking, use of contraceptives, sexual activity, etc. to support the conduction of association studies.

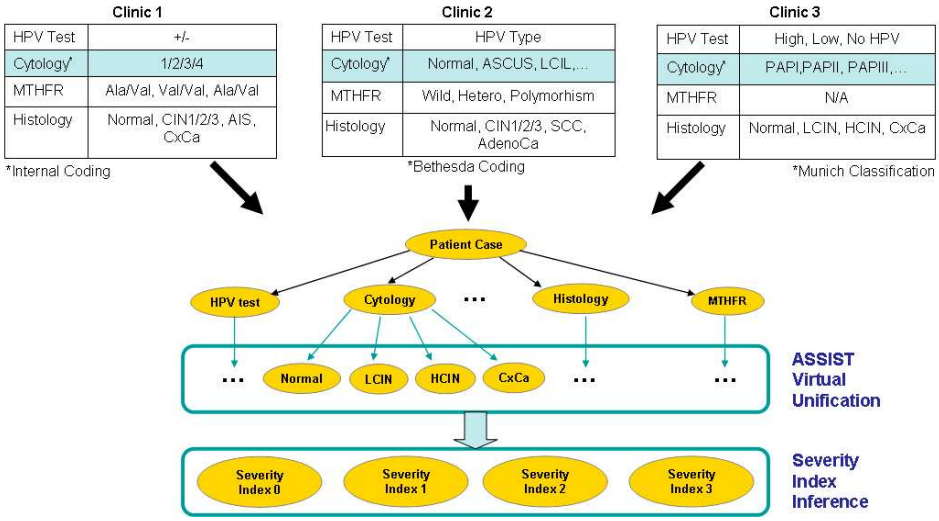

Figure 3. Heterogeneity in biomedical data related to CxCa and ASSIST virtual unification

4. Usage Scenario

From a user viewpoint, association studies involve a rather complex multi-step procedure. In these steps, several parameterisations take place, while there are also several pre/post-conditions that have to be fulfilled in order to move from one step to the subsequent one. ASSIST provides a controlled environment to ensure the validity of users' choices and offers guidance in any operation that users want to perform.

Consider a researcher wants to identify potential association of the MTHFR polymorphism with smoking in developing CxCa. Following an authentication procedure, she is granted access to the ASSIST services and selects to design an association study among the available options. For this operation, ASSIST provides her with a guidance wizard consisting of consecutive steps, i.e.: i) Data Retrieval, ii) Data Validation, iii) Genetic Description, and iv) Association Test. Specifically, she first defines the features of interest in terms of both study factors (smoking, MTHFR polymorphism) and inclusion criteria (severity of CxCa) and performs data retrieval. To ensure rich query expressiveness, a tree-like hierarchy was conceptualized and implemented in the ASSIST UI for the description of the incorporated concepts related to CxCa and patient data consisting of categories, such as *Severity of Cervical Neoplasia*, *Diagnostic Information*, *Therapeutic Intervention*, *Lifestyle*, *Personal Profile*, *Patient Profile*, *Genetic Marker* and *Specimen Availability*, and relevant subcategories that further specify the corresponding concept. The retrieved dataset can be viewed either in detail, or through statistical measures and histograms (e.g. race and

age distributions). Next step involves data validation against allele and haplotype frequencies of different populations by accessing reference polymorphism databases (e.g. HapMap). Moreover, the Hardy-Weinberg equilibrium is extracted for the case and control groups and the allele, genotype frequencies are calculated for the subject dataset in the succeeding step (Genetic Description). The results of the statistical analysis (e.g. logistic regression) are presented in the Association Tests step and the clinical significance is evaluated through the corresponding p-values.

5. Conclusion

Lately, there is an increasing research interest in genetic association studies. Several projects are under development, aiming to provide researchers with more powerful tools for investigating associations among various types of clinical and genetic data [9]. The International HapMap project (http://www.hapmap.org/), being one of the most ambitious ones, aims to develop a haplotype map of the human genome and provide it as a public resource that will help researchers find genes associated with human diseases, responses to drugs and environmental factors. ASSIST moves in a parallel course and, upon successful completion, its platform aspires to function as an IT tool enabling association studies linked for example with CxCa research by establishing a collaborative environment and allowing any medical group active in this area to use its facilities and/or contribute their own data/results. Following a generic design, the ASSIST system may be expanded in terms of its underlying knowledge model in order to facilitate genetic association studies for other diseases, e.g. colon cancer and cardiovascular diseases.

Acknowledgment

This work was supported in part by the IST-2004-027510 project, entitled “Association Studies Assisted by Inference and Semantic Technologies (ASSIST)”, funded by the Commission of the European Community (CEC).

References

[1] S.H. Landis et al., Cancer statistics, 1999. CA Cancer, J Clin **49**(1) (1999), 8–31.
[2] J.M.M. Walboomers et al., Human papillomavirus is a necessary cause of invasive cervical cancer worldwide, J Pathol **189**(1) (1999), 12–9.
[3] T. Agorastos et al., Human papillomavirus testing for primary screening in women at low risk of developing cervical cancer. The Greek experience, Gynaecol Oncol **96**(3) (2005), 714–20.
[4] J.N. Hirschhorn, K. Lohmueller, E. Byrne, K. Hirschhorn, A comprehensive review of genetic association studies, Genet Med **4**(2) (2002), 45–61.
[5] H.J. Cordell, D.G. Clayton, Genetic association studies, Lancet **366**(9491) (2005), 1121–31.
[6] J.N. Hirschhorn, M.J. Daly, Genome-wide association studies for common diseases and complex traits, Nat Rev Genet **6**(2) (2005), 95–108.
[7] G. Antoniou, F. van Harmelen, A Semantic Web Primer, MIT Press, 2004.
[8] V. Kashyap, T. Hongsermeier, Can Semantic Web technologies enable translational medicine?, In: C.J.O. Baker CJO, K.H. Cheung (Eds.). Semantic Web: Revolutionizing knowledge discovery in the life sciences, Springer, 2007.
[9] M. Dugas, Impact of integrating clinical and genetic information, In Silico Biol **2**(3) (2002), 383–91.