The seventh layer of the clinical-genomics information infrastructure

A. Shabo
D. Dotan

Clinical genomics is an interdisciplinary field dealing with the use of genetic data in clinical practice, such as the prescription of drugs based on a patient's genetic profile. To support the integration of clinical genomics information with clinical practice and research, information technologies should enable the semantic association of patient-specific genetic data (e.g., gene variants) with the patient's phenotypic information. To create such genotype-phenotype associations effectively, information technologies should consider relevant data from patient medical records along with information from biomedical knowledge sources. In this paper, we describe how these challenges are addressed by Clinical Genomics Level Seven (CGL7), a set of Web services for clinical-genomics decision-support applications that follow the HL7® (Health Level 7®) Clinical Genomics standard for the representation and exchange of clinical genomics data.

INTRODUCTION

The goal of personalized medicine[1–3] depends on the ability to effectively associate personal genetic data with clinical data in order to support clinical decisions at the point of care for the individual patient. One of the challenges in achieving this goal is the extremely fragmented nature of health-care information,[4,5] which has resulted in incomplete patient data at the point of care.[6,7] For example, episodic data (such as discharge summaries) created in one hospital are typically not available in other hospitals. Medical records are handled by enterprise health-care information systems, which are centered on the needs of the health-care enterprise and on workflows, such as those related to billing and administration. The needs of the patient and the secondary uses of medical records, such as clinical research, clinical trials, and drug development, are not properly addressed by the current constellation of information in health care.[8] Adding personal genetic data into this welter of dispersed and dissimilar medical records makes the integration challenge much harder. This difficulty stems from a lack of standard genetic data representations and a lack of standard testing-method descriptions needed for quality determination.

The clinical genomics domain deals with the interrelation of clinical and genomic data at the level of distinct data items related to the individual patient. In contrast, much of the publicly available genomic information is still generic. For example, the human genome consists of the DNA sequences believed to be common to every human being. In fact, genetic variations such as single nucleotide polymorphisms (SNPs) occur about every thousand nucleotides and can be used to differentiate any two persons. Such SNPs often cause variations in the gene product and in the level of its expression, that is, the extent to which it generates the RNA that constructs the gene product. These variations often play a role in health conditions such as drug sensitivities, allergies, and diseases.[9]

The fusion of the patient's medical history and the most up-to-date medical knowledge lies at the heart of the vision of personalized medicine. The following example demonstrates our vision for patients in the upcoming personalized medicine era. Kobayashi et al.[10] reported on the case of a lung cancer patient who was responsive to the drug Gefitinib due to a certain mutation in the EGFR (epidermal growth factor receptor) gene and thus reached complete remission for two years. They then detected a second somatic mutation in the tumor tissue that led to a relapse and understood the influence of that mutation on the responsiveness to Gefitinib by using structural modeling and biochemical studies. If similar analysis could be done during the treatment of any patient, it would support clinicians in selecting the best possible treatment by giving them more information on the prognosis of each treatment alternative. In the long run, this approach could potentially serve as input to the drug development process, allowing drugs to be fine-tuned to the precise clinical-genomic condition of the patient.

In this paper, we describe our progress toward creating a layer of genotype-phenotype interrelations that current information technology (IT) solutions lack, and providing a higher-level infrastructure for decision-support applications. We first describe the newly created HL7** (Health Level 7**) Clinical Genomics (CG) standard,[11] through which patient-specific genetic data can be associated with the patient's clinical information and correlated with the most up-to-date scientific knowledge. The main design principle of this standard requires that genetic testing results be sent to the requesting health-care provider along with raw genomic data (e.g., the full sequencing data or expression levels) so that they can be saved in the patient's electronic health record (EHR) and additional interpretations can "bubble up" as new knowledge and data become available. We also introduce CGL7 (Clinical Genomics Level 7), a specialized middleware designed to become the technological enabler for HL7. We discuss CGL7's design and initial implementations, and give an example of a bubble-up strategy that utilizes the Online Mendelian Inheritance in Man (OMIM) database[12] to provide phenotype information pertinent to a specific patient. Finally, we describe several initial use cases and collaborations utilizing the clinical genomics layer implemented by CGL7.

Health Level 7

HL7 is an organization that focuses on the development of semantic standards for messaging and documentation in health care. These standards are widely used in hospitals worldwide. The HL7 Clinical Genomics Special Interest Group develops standards to enable the exchange of clinical and personalized genomic data between interested parties. The group focuses on routine use cases in health care such as detection of known mutations, while preparing the information infrastructure for more advanced cases, such as full sequencing and detection of somatic mutation or the use of gene expression methods.

The HL7 Clinical Genomics standard

Genomic data varies in its complexity and the extent to which it is used. Simple testing identifies genes and mutations; more complex assays include full DNA sequencing, RT-PCR (reverse transcription polymerase chain reaction)[13] for the expression level of a small number of genes,[14] and micro-arrays to identify the expression levels of vast numbers of genes in the individual.[15]

The place of genomic data sets within common clinical information constructs can be similar to that of other common health observations, but there are several characteristics that might distinguish it from typical observations such as blood pressure or potassium level[16]:

- The amount of data. Typically, a single genetic locus 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 genomic data. These are necessary for the receiver to interpret the data correctly and assess its level of reliability.
- The interpretation of the data. This is constantly evolving as new discoveries are made.
- The emerging common formats being used by bioinformatics communities, for example, the Bioinformatic Sequence Markup Language (BSML)[17] and the MicroArray and Gene Expression Markup Language (MAGE-ML).[18]
- The semantics of the genotype-phenotype relations, which are represented in a variety of ways, depending on the point of view (clinical research, pharmaceutical, or health care).

The design underlying the CG standard addresses the challenges inherent in realizing the personalized medicine vision. The main characteristics of this design are the association of personal genetic data with clinical data using HL7 messages, the use of data representations such as MAGE-ML for gene expression data and BSML for sequencing data as the preferred formats for representing raw genomic data, and the use of the "encapsulate and bubble-up" conceptual workflow, which is described in the following subsection.

The "encapsulate and bubble-up" workflow

The *encapsulation* phase of the conceptual workflow of the CG standard includes the incorporation of raw genomic data received from sources like genetic laboratories into patient records, based on a predefined, constrained bioinformatics format. Constraining the extensive bioinformatics markup schemas is intended to omit portions such as the display elements in the BSML markup and others that seem irrelevant to clinical practice. It also ensures that the data refers to one patient only and thus makes it possible to cross-identify the patient identifiers in the genetic data with the patient identifiers in the clinical data.

The *bubble-up* phase is an iterative process wherein various genomic-oriented decision-support applications parse the encapsulated raw genomic data and make prominent those portions that seem to be most clinically significant to the patient's clinical history and current treatment plans, based on the most up-to-date knowledge available. The results of this phase are held in genotype-phenotype associations supported by the standard's specification.

The notion of bubbling up differs from the notion of summarization or annotation. Summarization implies that the data being summarized is thoroughly understood, whereas it is difficult to summarize clinical genomics data because most of the knowledge in this field is still unknown. Bubbling up is similar to gem mining activities; one does not always know what will be found, and the findings vary depending on the location. Annotation is the result of bubbling-up processes, whereby the data which is input to these processes is annotated and, more important, enriched through the creation of new or annotated genotype-phenotype associations.

The encapsulate and bubble-up workflow may lead to gradual distillation of the raw genomic data in the context of diagnosis and treatment provided to a specific patient at a specific time, while making the raw data available within the patient's medical records so that it may be possible to parse it again (for example, by alternative algorithms or when new knowledge becomes available). ***Figure 1*** shows how this workflow could be implemented in health care, with a focus on enterprise EHR systems accompanied by decision-support applications. It shows a workflow based on sequencing data. In the static phase of the workflow, encapsulation is performed based on a static predefined BSML schema. In the dynamic phase, clinically significant SNP data is bubbled up into HL7 `SequenceVariation` objects, and these objects are linked with clinical data from the patient EHR, thus enhancing the risk assessment process.

Models used in CGL7

The core model of the HL7 clinical genomics specifications is the `GeneticLocus` model, which consists of placeholders for various types of genomic data relating to a specific locus on the genome (e.g., DNA or RNA) including sequencing, expression, and proteomic data. Within the `GeneticLocus` model, existing bioinformatics markups were used to represent raw data received from genomic facilities, enabling the encapsulate and bubble-up workflow as described in the previous section. These markups were constrained to make them compliant with the HL7 design principles.

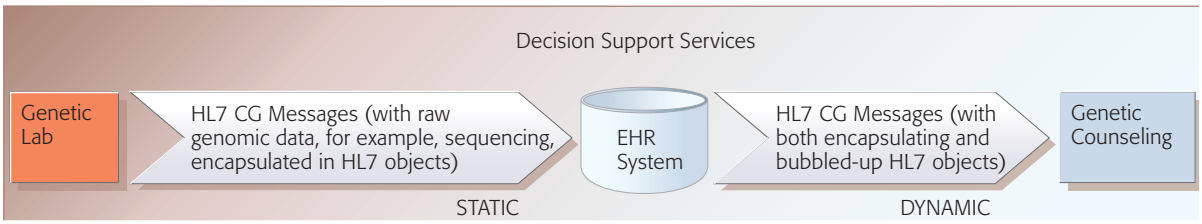

Figure 1
Encapsulate and bubble-up workflow

The `GeneticLoci` model describes a set of loci, such as a haplotype (several variations of the same chromosome), a genetic profile, and genetic testing results of multiple variations or gene expression panels. The `GeneticLoci` model uses the `GeneticLocus` model to describe each of these loci.

The `GeneticLocus` and `GeneticLoci` models evolved from the work of the HL7 clinical genomics group on the tissue-typing use case. Four information modules were identified in the clinical context of bone-marrow transplantation (BMT).[19] The first module involves the exchange of messages and documents in the BMT tissue typing use case, including entities such as the BMT ward, donor banks, tissue-typing laboratory, and others. The second information module focuses on the unique observations used in tissue-typing. This includes the individual tissue-typing observation and the matching observation, which indicates the level of matching between two individual tissue-typing observations (e.g., patient and donor). The rationale for this modularity is that the tissue-typing observation can be used in other tissue-typing use cases, such as paternity testing and forensic cases. The third module's focus is on the two individual haplotypes (from the two chromosomes), each consisting of several genotypes of the HLA (human leukocyte antigen) antigens, a key part of the human immune system. A haplotype can be described by using the `GeneticLoci` model. Finally, each of the alleles is described by using the core `GeneticLocus` model.

Although describing the `GeneticLocus` model in detail is beyond the scope of this paper, ***Figure 2*** demonstrates the way this model is built by showing a portion of the model related to the `Sequence` class. The `Sequence` class is associated with the `IndividualAllele` class, which in turn is associated with the entry point—the `GeneticLocus` class (not shown in the figure). The `value` attribute of the `Sequence` class can hold raw data in a bioinformatics markup format, such as BSML. There is a recursive association (`derivedfrom2`, shown at the bottom left) that allows the representation of a biological sequence derived from a source sequence (e.g., an mRNA sequence derived from a DNA sequence). An association with clinical phenotypes (`pertinentInformation`) allows genomic data to be linked to the clinical data that most likely resides in the patient medical records. `ClinicalPhenotype` in the model code stands for the phenotype model, whose classes are fully described elsewhere in the `GeneticLocus` model. Other associations relate to sequence variations (`derivedfrom3`) and other types of data not shown in this code, such as proteomic data. Optionally, the performer of the sequencing assay can be specified (`performer1`). The entire model can be found in the HL7 V3 package available from the HL7 site.[11]

The conventions used in this model are based on the HL7 V3 Reference Information Model (RIM).[20] All HL7 V3 specifications are derived from the RIM, thus ensuring a better level of semantic interoperability. In essence, these notations are examples of UML** (Unified Modeling Language**) models with constraints and extensions that reflect the RIM, which is represented as a regular UML model.[21]

Figure 3 shows a code segment taken from a sample `GeneticLocus` XML (Extensible Markup Language) instance. The code segment illustrates how it is possible to encapsulate sequencing data, including several variations found in the patient's sequence. In the figure an EGFR variation is bubbled up and associated with clinical phenotypes (lung cancer and responsiveness to the Gefitinib drug).

Another part of the HL7 CG specification is the `FamilyHistory` model, aimed at describing a pa-

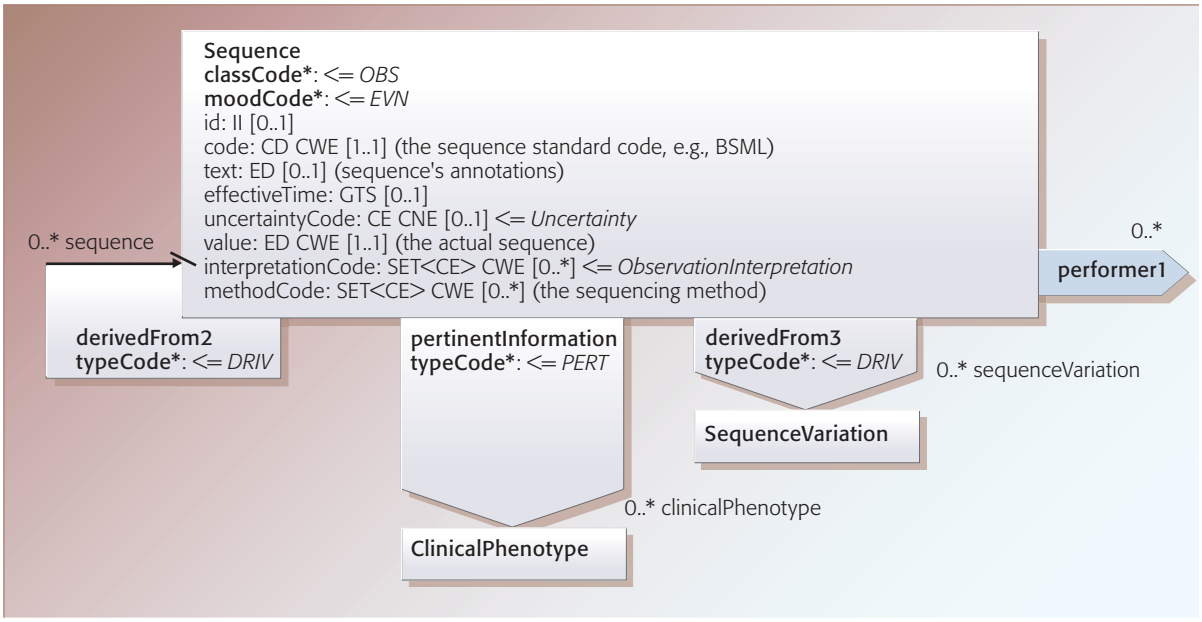

Figure 2
Portion of GeneticLocus model

tient's genetic family history (pedigree) with clinical and genomic data. It also includes a message interaction between two disparate pedigree applications, in which a person's pedigree is sent from one application to the other. This model utilizes the `GeneticLocus` model to carry the genomic data for the patient's relatives. Breast and ovarian cancers are among the use cases where there is a clear need to represent a patient's family history, because these diseases run in families. These cancers are associated with variations in the breast cancer genes. A complete family history can be assessed for potential risk,[22] supporting clinical decisions in genetic counseling and at the point of care.

CGL7 architecture and services

To promote the encapsulate-and-bubble-up conceptual workflow described in the previous section, we developed the CGL7 pilot platform, which provides uniform access to various strategies for this workflow, as well as common services such as an API (application programming interface) for message parsing, manipulation, persistency, and serialization. It allows developers of decision-support applications to define and use new strategies easily and to test them in different combinations and usage scenarios. In this section, we describe the architecture of CGL7, as well as some of our preliminary strategies.

CGL7 is implemented using Web Services technology standards. This allows CGL7 to be rapidly integrated into different client applications, such as doctors' workbenches, medical portals, and decision-support applications. In addition, the reliance on Web Services standards allows rapid creation of cooperative cross-institutional research efforts. This corresponds with the vision of service-oriented architecture (SOA), which is the cornerstone of many recent health-care IT efforts that attempt to integrate legacy applications with newer systems.

Many of the use cases for CGL7 involve multistep processes. An example of the use of encapsulation and two different bubble-up strategies is shown in *Figure 4*. Because HL7 clinical genomics instances can be relatively large, especially when they contain encapsulated data, it is important to avoid the back-and-forth transfer of files between client and server. To this end, CGL7 is provided as a stateful, session-based service. In the example depicted in Figure 4, the encapsulation result—an HL7 `GeneticLocus` instance—is kept in the service session memory and is then used in later steps. As the scenario proceeds, the stored instance is modified, by adding variant-level phenotype information, for example. Finally, the result is returned at the end of the second bubble-up step as an enriched HL7 `GeneticLocus` instance. The decision of whether to retrieve the

```
<GeneticLocus>
        <individualAllele moodCode="EVN">
            <text>EGFR receptor gene</text>
            <value code="EGFR"/>
                <sequence moodCode="EVN">
                    <value mediaType="text/xml">
                        <bsml:Sequences>
                            <bsml:Sequence id="seq1" molecule="dna" title="EGFR..."
                                length="5616">
                                <bsml:Seq-data>
                                gcgcggccgc agcagcctcc gccccccgca cggtgtgagc gcccgacgcg
                                ccggagtccc gagctagccc cggcggccgc cgccgcccag accggacgac
                                    ...
                                </bsml:Seq-data>
                            </bsml:Sequence>
                        </bsml:Sequences>
                        <bsml:Isoforms>
                            <bsml:Isoform-set>
                                <bsml:Isoform id="variation1" seqref="seq1" location="2240"
                                    change="" replaces="cctcttcatg cgaaggcg"/>
                                <!-- possibly more isoform tags denoting other variations -->
                            </bsml:Isoform-set>
                        </bsml:Isoforms>
                    </value>
                    <sequenceVariation moodCode="EVN">
                        <code code="DNA.MUTATION"/>
                        <text>A somatic mutation in the active site of the EGFR receptor gene
                                is found in about 10% of non-small cell lung cancer tumors</text>
                        <value xsi:type="CE" code="131550.0001" displayName="18-BP DEL,
                                NT2240" codeSystemName="OMIM"/>
                        <interpretationCode code="DELETERIOUS"/>
                            <clinicalPhenotype classCode="ORGANIZER">
                                    <observationGeneral>
                                        <code/>
                                        <statusCode/>
                                        <effectiveTime value="20010101"/>
                                        <value xsi:type="CE" code="D2-F1007"
                                                codeSystemName="SNOMED CT"
                                                displayName="Non-small cell lung cancer"/>
                                    </observationGeneral>
                            </clinicalPhenotype>
                            <clinicalPhenotype classCode="ORGANIZER">
                                    <observationGeneral>
                                        <text>Iressa (gefitinib) responder</text>
                                        <effectiveTime value="20010101"/>
                                        <value xsi:type="CS" code="gefitinib-responder"/>
                                    </observationGeneral>
                            </clinicalPhenotype>
                            <associatedProperty>
                                <code code="TYPE"/>
                                <text>
                                    <reference value="#variation1"/>
                                </text>
                                <value xsi:type="CV" code="SNP"/>
                            </associatedProperty>
                            <associatedProperty>
                                <code code="REFERENCE"/>
                                <value xsi:type="URL"
                                       value="http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?
                                       db=nucleotide&val=41327737"/>
                            </associatedProperty>
                    </sequenceVariation>
                </sequence>
        </individualAllele>
</GeneticLocus>
```

Encapsulated data

Bubble-up data

Figure 3
Sample code segment from GeneticLocus XML instance

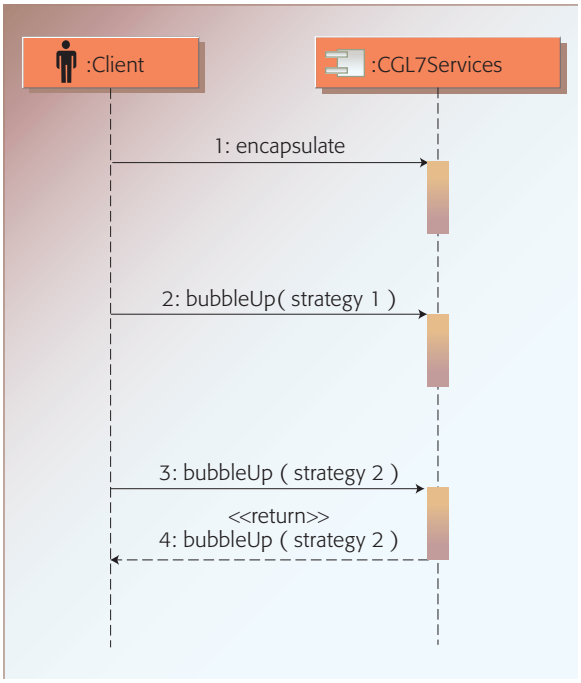

Figure 4
Sample sequence diagram for CGL7 use case

result at each step is made by simply specifying the value of a boolean parameter. Similarly, if the client does not pass an input instance, the stored document is used.

At each step of the process, the client developer specifies the name of the bubble-up or encapsulation strategy to be performed. These strategies are implemented as Java** classes and are stored in a strategy registry (see ***Figure 5***). This allows new classes to be registered and used without changes to the core CGL7 code base.

In-memory graphs

To manipulate the clinical genomics elements effectively, for instance in the encapsulation and bubble-up strategies, the elements are represented in memory as Java objects. This is advantageous because it makes object graphs easy to compute for various needs. They can be manipulated, annotated, compared, searched, and so forth. Furthermore, message graphs can be combined to create instances of other message types that have common elements, for example, creating an HL7 report message to public health agencies based on components of HL7 clinical genomics and clinical document instances.

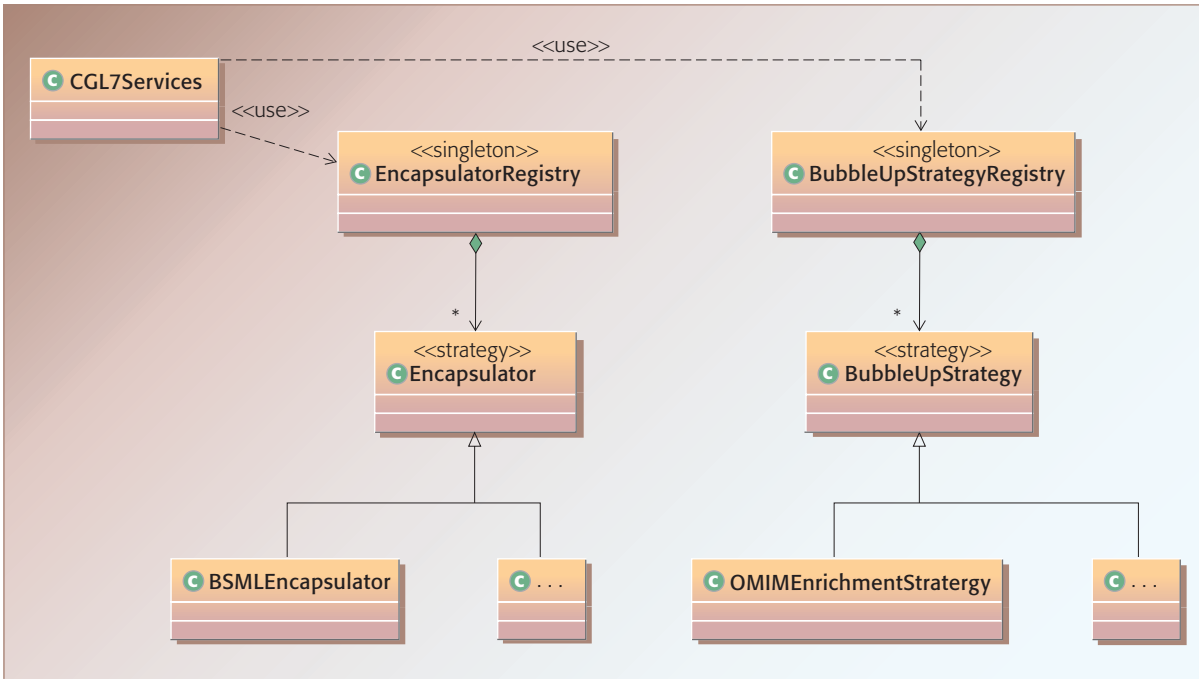

Figure 5
Sample class diagram for CGL7 use case

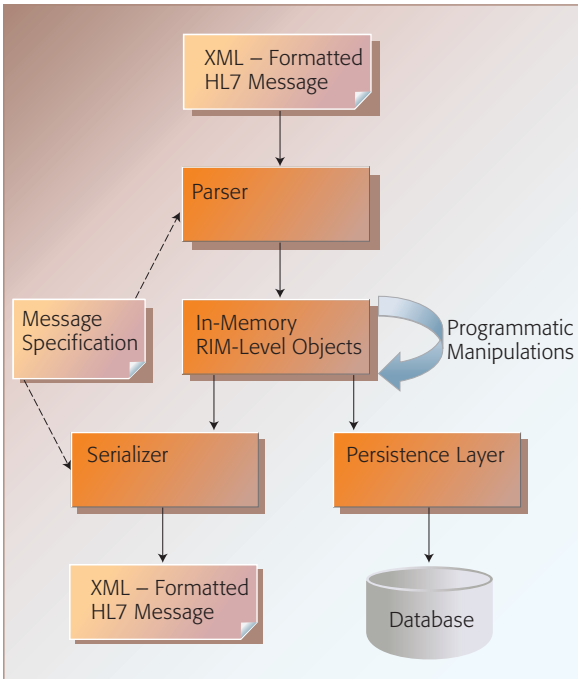

Figure 6
Block diagram of CGL7 API usage

In CGL7, the parsing of XML instances to create Java objects and vice versa is done by using a special-purpose open-source API created by members of the HL7 community. ***Figure 6*** shows the basic usage of this API. The XML instance is loaded by the parser, which also takes an XML description of the message specification (in HL7's Message Interchange Format [MIF]). This specification introduces the different domain-specific entities, such as those described previously for the clinical genomics domain, and describes each entity in terms of its derivation from the base HL7 RIM. Using this information, the parser knows which RIM-level classes to instantiate for any domain-level XML entry that it encounters. Likewise, the serializer of the HL7 API uses this information to correctly serialize the in-memory objects into legal XML, including checking validity and filling out default values as defined in the MIF.

Persistence

The HL7 Java API also provides support for saving ("persisting") the HL7 XML instances in a relational database. It does so by using the Hibernate[23] object-relational mapping technology. CGL7 uses this capability to store the current `GeneticLocus` document upon client request, query the database for stored documents, and retrieve them. The two latter functions may be used by client developers as well as by strategy developers, for example, to create bubble-up strategies that look for cases similar to the bubbled-up instance at hand. This feature could be used in a case-based reasoning application that attempts to suggest solutions for a given case, based on similar cases that are stored in a case database along with information on the outcome of each case.

CGL7 API

In the HL7 Java API, all HL7 elements are represented as RIM-level classes. This makes it difficult to program domain-level applications because the domain-level classes and their relationships (called "clones" in HL7 terminology) are not represented by their own name as defined by the HL7 model itself. For example, both the `GeneticLocus` and `Sequence` classes are represented as `Observation` objects, following their RIM base class, which is `Observation`. The relationship between these classes, which is called `COMPONENT4` in the clinical genomics domain and whose RIM base class is `ActRelationship`, cannot be semantically distinguished from other component-type `ActRelationships`.

CGL7 provides a layer that allows strategy developers to work with domain-specific classes instead of working with the RIM-level classes provided by the RIM API. This layer provides a wrapper for the RIM API classes and supports the different domain-specific classes and relationships used in the clinical genomics domain. In addition, by using XMLBeans technology,[24] it is possible for application developers to get Java representations of several bioinformatics payload formats, such as BSML, MAGE, and OMIM.

In summary, the CGL7 pilot platform takes care of parsing, serializing, and persisting domain-level representations, thus allowing strategy developers to implement genuine encapsulate-and-bubble-up strategies for evolving clinical-genomics decision-support applications.

Example of a bubble-up strategy

As an example of realizing the bubble-up concept, we developed the "OMIM enrichment" bubble-up strategy. Using data on a patient's genetic makeup (i.e., what variants the patient has for certain genes), the goal of this strategy is to find information about the phenotypes (disorders, clinical traits, drug responsiveness, etc.) that correspond to these

variants, as reported in the medical and scientific literature. Such information is summarized in Web-based resources such as OMIM.

Given an input `GeneticLocus` instance, the OMIM enrichment strategy searches for the OMIM entry that describes this particular genetic locus. That entry typically contains several paragraphs of locus-level information, as well as allele-specific information for known alleles. The strategy then adds the locus-level information to the root element (`GeneticLocus`) as an associated `KnownClinicalPhenotype`. Next, allele-specific information regarding each of the patient's variants is extracted from the OMIM entry and added as new `KnownClinicalPhenotype` elements to the `IndividualAllele` or `SequenceVariation` elements that represent the alleles. Finally, each locus-level entry that also relates to one of the patient's variants is referenced in the allele-level `KnownClinicalPhenotype`.

Genes and variations are identified in OMIM with a proprietary coding system, which is different from the nomenclature standards defined by the Human Genome Organization (HUGO)[25] for genes, and by the Human Genome Variation Society (HGVS)[26] for variations. Therefore, before OMIM entries are retrieved and allele-level information is extracted, the standard names are translated to OMIM gene and variation identifiers. The strategy adds these translations to the code elements of the locus' alleles and variations using the standard HL7 construct of translation subelements.

Sample use cases

The following CGL7 use cases were developed together with health-care providers whose representatives have been actively participating in the design of the CG standard and its underlying encapsulate-and-bubble-up workflow.

DNA resequencing

The Harvard Partners Center for Genetics and Genomics (HPCGG) provides resequencing services to physicians affiliated with Partners Healthcare (a nonprofit integrated health system located in Boston) who treat patients with non-small-cell lung cancer, provided that all previous genetic sequences of the patient's tumor tissue were stored in a special database. The resequencing results are then compared with previous sequences of the EGFR gene of the patient. This enables accurate reporting of any changes that have occurred since the last sequencing, in particular the somatic variations developed during the course of the disease.

In the existing information flow, the HPCGG receives genetic test orders from referring physicians in Partners Healthcare and sends back the test results annotated by its geneticists to the referring physicians. The HPCGG would like to complement these annotations with additional annotations based on the most up-to-date knowledge available in publicly available external reference databases, papers, ontologies, and so forth. This can be accomplished using CGL7 as follows:

1. The HPCGG sets up a session with the CGL7 Web services and defines the knowledge sources to use.
2. The HPCGG sends the patient's genetic test results to CGL7 as an HL7 clinical genomics XML instance.
3. CGL7 accesses the enterprise electronic medical record system to get a more complete medical history of the patient and to guide the bubbling-up process.
4. CGL7 outputs a bubbled-up HL7 clinical genomics XML instance with relevant information from publicly available resources such as OMIM.
5. The HPCGG sends the enriched XML instance to the applications used by the end users (geneticists, referring physicians, etc.) to facilitate better decision support.

Family history

Another example of the use of CGL7 is the family history interoperability project, a joint project of the IBM Research Laboratory in Haifa, Massachusetts General Hospital, and the University of Massachusetts at Lowell. The goal of this project is to enable the exchange of family history data by using the HL7 standard. The data is created by different applications as well as by patients, using tools such as the United States Surgeon General's family history Web tool.[27] The project uses the family history specification of the CG domain, which represents a patient's pedigree. Each node on the pedigree represents a relative of the patient, with clinical and genomic information represented by the basic model (i.e., the `GeneticLocus` specification). The desired role of CGL7 is to enrich the pedigree and to compare a given pedigree to other pedigrees in a

clinical data repository containing family history information. Finding similar pedigrees and looking at the medical history of patients can help assess the risk for breast cancer patients and aid in selecting the optimal treatment.

Tissue typing

In tissue typing for bone-marrow transplantation (BMT), it is desirable to have full sequencing of the HLA alleles of the patient and the potential donors. This helps ensure that the match between their tissue types will best fit the goals of the transplantation. In the past, the goal was always to find the perfect match (e.g., from an identical twin). However, in new BMT treatments, such as the mini-transplant procedure,[28] there is evidence that mismatching certain HLA alleles could result in a clinical benefit for the patient. In these new treatments, the patient's immune system is not destroyed completely as in the traditional BMT procedure, but merely suppressed. Meanwhile, the bone marrow donation is augmented with the donor's own lymphocytes, which are able to fight the tumor cells during the GVM (graft-versus-malignancy) effect period. In research collaboration with the Bone Marrow Transplantation Center and the Tissue Typing Laboratory of the Hadassah University Hospital, CGL7 has served to bubble up those matches and mismatches needed for a certain treatment, such as mini-transplantation, while taking into account the HLA full sequencing data of the patient and donor and the patient's medical history.

CONCLUSION

This paper discusses possible ways to introduce a semantic layer of genotype-phenotype associations into clinical genomic repositories. Current IT solutions for clinical genomic repositories store clinical and genomic data in a side-by-side fashion, correlated only by the patient identifier. This situation overlooks the interrelations between distinct data items on both sides.

Current clinical genomic repositories contain patient-specific data, but are targeted mainly at clinical research. As we move into the era of personalized medicine, these repositories will increasingly serve the clinical practice by serving as an infrastructure for clinical decision-support applications at the point of care. The new clinical genomic repositories should also be connected with the most up-to-date knowledge sources, such as publicly available reference databases, ontologies, and terminology references for life sciences and health care. In addition, they should be connected to the operational EHR systems used by health-care providers, where the most complete medical history of the patient can be found.

The challenge of personalized medicine is to fuse available scientific knowledge with patients' histories and treatment goals. In this paper, we have described CGL7, a pilot middleware aimed at closing the gaps encountered in the current clinical genomics infrastructure. CGL7 is compliant with the newly created HL7 Clinical Genomics standards as well as their underlying workflow paradigm (encapsulate and bubble up), which allows the genotype-phenotype relationships to be dynamically created as new data and knowledge become available. In addition, the HL7 standard enables clinical genomics data to be embedded in the patient's longitudinal and cross-institutional EHR—the ultimate source of data in tomorrow's personalized medicine practice.

CITED REFERENCES

1. G. Ruano, "Quo Vadis Personalized Medicine?" *Personalized Medicine* **1**, No. 1, 1–7 (2004).
2. R. L. Davis and M. J. Khoury, "The Journey to Personalized Medicine," *Personalized Medicine* **2**, No. 1, 1–4 (2005).
3. E. S. Vazquez, "Personalized Therapy: An Interdisciplinary Challenge," *Personalized Medicine* **1**, No. 1, 127–130 (2004).
4. A. S. Rothschild, L. Dietrich, M. J. Ball, H. Wurtz, H. Farish-Hunt, and N. Cortes-Comerer, "Leveraging Systems Thinking to Design Patient-Centered Clinical Documentation Systems," *International Journal of Medical Informatics* **74**, No. 5, 395–398 (2005).
5. D. Lawrence, *From Chaos to Care: The Promise of Team-Based Medicine*, Perseus Publishing, Cambridge, MA (2002).
6. *To Err Is Human*, L. T. Kohn, J. M. Corrigan, and M. S. Donaldson, Editors, Institute of Medicine, Washington, DC (1999).
7. *Crossing the Quality Chasm*, L. T. Kohn, J. M. Corrigan, and M. S. Donaldson, Editors, Institute of Medicine, Washington, DC (2001).
8. A. Shabo, "A Global Socio-Economic-Medico-Legal Model for the Sustainability of Longitudinal Electronic

Health Records," *Methods of Information in Medicine* **45**, No. 3, 240–245 (2006).

9. M. Sean, "Bioinformatics Approaches and Resources for Single Nucleotide Polymorphism Functional Analysis," *Briefings in Bioinformatics* **6**, No. 1, 44–56 (2005).
10. S. Kobayashi, T. J. Boggon, T. Dayaram, P. A. Jänne, O. Kocher, M. Meyerson, B. E. Johnson, M. J. Eck, D. G. Tenen, and B. Halmos, "EGFR Mutation and Resistance of Non–Small-Cell Lung Cancer to Gefitinib," *New England Journal of Medicine* **352**, No. 8, 786–792 (2005), http://content.nejm.org/cgi/content/full/352/8/786.
11. Health Level Seven (HL7), http://www.hl7.org.
12. OMIM—Online Mendelian Inheritance in Man, Johns Hopkins University, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM.
13. S. A. Bustin, "Quantification of mRNA Using Real-Time Reverse Transcription PCR (RT-PCR): Trends and Problems," *Journal of Molecular Endocrinology* **29**, No. 1, 23–39 (2002).
14. Oncotype DX Breast Cancer Assay, Genomic Health, http://www.genomichealth.com/oncotype/default.aspx.
15. J. Quackenbush, "Microarray Data Normalization and Transformation," *Nature Genetics* **32**, Supplement, 496–501 (2002).
16. P. Knaup, E. Ammenwerth, R. Brandner, B. Brigl, G. Fischer, S. Garde, E. Lang, R. Pilgram, F. Ruderich, R. Singer, A. C. Wolff, R. Haux, and C. Kulikowski, "Towards Clinical Bioinformatics: Advancing Genomic Medicine with Informatics Methods and Tools," *Methods of Information in Medicine* **43**, No. 3, 302–307 (2004).
17. Bioinformatic Sequence Markup Language (BSML), http://www.bsml.org.
18. MicroArray and Gene Expression Markup Language (MAGE-ML), http://www.mged.org/Workgroups/MAGE/mage-ml.html.
19. E. D. Thomas, "Bone Marrow Transplantation: A Review," *Seminars in Hematology* **36**, No. 4, 95–103 (1999).
20. HL7 Java Special Interest Group—Reference Information Model (RIM) API, http://www.hl7.org/Library/Committees/java/apidocs/org/hl7/rim/package-summary.html.
21. J. A. Lyman, K. Scully, S. Tropello, J. Boyd, J. Dalton, S. Pelletier, and C. Egyhazy, "Mapping from a Clinical Data Warehouse to the HL7 Reference Information Model," *American Medical Informatics Association Annual Symposium Proceedings* (2003), p. 920, http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=1480241&blobtype=pdf.
22. J. L. Jones, K. S. Hughes, M. Howard-McNatt, D. B. Kopans, R. H. Moore, S. S. Hughes, N. Y. Lee, C. A. Roche, N. Siegel, M. A. Gadd, B. L. Smith, and J. S. Michaelson, "Evaluation of Hereditary Risk in a Screening Mammography Population," *Clinical Breast Cancer* **6**, No. 1, 38–44 (2005).
23. Hibernate, Relational Persistence for Java and .NET, Red Hat Middleware, http://www.hibernate.org.
24. XMLBeans, Apache XML Project, http://xmlbeans.apache.org.
25. The Human Genome Organisation. http://www.hugo-international.org.
26. HGVS—Human Genome Variation Society, http://www.hgvs.org.
27. U. S. Surgeon General's Family History Initiative, United States Department of Health and Human Services, http://www.hhs.gov/familyhistory.
28. A. M. Carella, R. Champlin, S. Slavin, P. McSweeney, and R. Storb, "Mini-allografts: Ongoing Trials in Humans," *Bone Marrow Transplantation* **25**, No. 4, 345–350 (2000).

Accepted for publication August 20, 2006.
Published online December 12, 2006.

Amnon Shabo (Shvo)
IBM Research, Haifa Research Laboratory, Haifa University, Mount Carmel, Haifa 31905, Israel (shabo@il.ibm.com). Dr. Shabo is a research staff member involved in various IBM health-care and life-sciences projects. He specializes in health-informatics, health-care, and life-sciences standards. He is a co-chair of the Clinical Genomics Special Interest Group in HL7, as well as its modeling facilitator and primary contributor. He is also a co-editor of the HL7 CDA (Clinical Document Architecture) Release 2 and of the CCD (Continuity of Care Document), a joint effort of HL7 and ASTM (American Society for Testing and Materials) in applying the ASTM CCR (Continuity of Care Record) to the CDA. Dr. Shabo specializes in longitudinal and cross-institutional electronic health records (EHRs) and was a coauthor of the mEHR (EHR for mobile citizens) proposal made by a consortium of 19 partners to the European Commission's Sixth Framework Programme, based on his vision of independent health record banks for addressing the challenge of lifetime EHR sustainability.

Dolev Dotan
IBM Research, Haifa Research Laboratory, Haifa University, Mount Carmel, Haifa 31905, Israel (dotan@il.ibm.com). Mr. Dotan has a B.Sc. degree in computer science and bioinformatics and an M.Sc. degree in computer science from the Technion-Israel Institute of Technology. His areas of interest include bioinformatics, information integration, service-oriented architecture, ontologies, the semantic Web, visual languages, and model-driven development. ■