

Cancer Genomics Object Model: An Object Model for Cancer Research Using Microarray

Yu Rang Park¹, Hye Won Lee¹, Sung Bum Cho¹, Ju Han Kim^{*12}

¹Seoul National University Biomedical Informatics (SNUBI) and ²Human Genome Research Institute, Seoul National University College of Medicine, Seoul 110-799, Korea

Email : kirarang@snu.ac.kr, hwon0615@snu.ac.kr, csb1749@snu.ac.kr, juhan@snu.ac.kr

Abstract: DNA microarray becomes a major tool for the investigation of global gene expression in all aspects of cancer and biomedical research. DNA microarray experiment generates enormous amounts of data and they are meaningful only in the context of a detailed description of microarrays, biomaterials, and conditions under which they were generated. MicroArray Gene Expression Data (MGED) society has established microarray standard for structured management of these diverse and large amount data. MGED MAGE-OM (MicroArray Gene Expression Object Model) is an object oriented data model, which attempts to define standard objects for gene expression. To assess the relevance of DNA microarray analysis of cancer research it is required to combine clinical and genomics data. MAGE-OM, however, does not have an appropriate structure to describe clinical information of cancer. For systematic integration of gene expression and clinical data, we create a new model, Cancer Genomics Object Model.

1. 서론

분자 생물학적인 기술과 로봇 공학의 발달로 인해 유전자 마이크로어레이 기술이 등장하게 되고, 이 기술은 생명과학 분야의 기초 연구는 물론 질병의 진단과 예측, 신약 개발 등 의학 분야에서 널리 응용되고 있다[1]. 유전자 마이크로어레이 실험으로 생성되는 데이터는 분석 방법이 다양하고 연구자마다 해석하는 방식이 다를 수 있기 때문에 데이터 교환을 위한 공통 기반이 필요하다[2]. 이를 위해 Microarray Gene Expression Data (MGED) 단체에서는 마이크로어레이 관련 정보의 표준화 작업을 진행시켜 왔다. Microarray Gene Expression Object Model(MAGE-OM)은 마이크로어레이 데이터 교환을 위한 포맷을 객체 모델로 디자인 한 것이다. 그리고 MAGE-OM을 XML로 구현한 것이 Microarray Gene Expression Markup Language (MAGE-ML)이다[3].

이러한 마이크로어레이를 이용한 암 연구가 증가하고 있으며, 마이크로어레이 실험에 관련된 논문의 30% 이상이 암 연구에 관련된 논문이다 [그림 1]. 일반적으로 암 연구에서 유전자 마이크로어레이는 유전자 발현 정보를 통한 암의 예후, 진행, 치료에 대한 반응 정도를 기준으로 새로운 분류법이나 암을 진단할 수 있는 종양 표지자를 찾는 데 많이 사용된다[1].

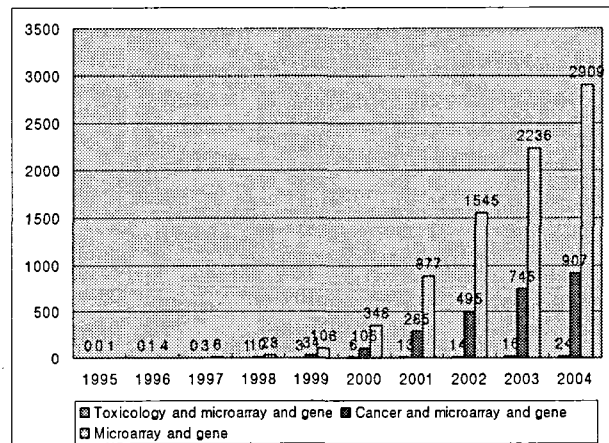


Figure 1. The growth in publication in cancer research using microarray. Searches were performed at PubMed using search string "Toxicology and Microarray and Gene" (green bar), "Cancer and Microarray and Gene" (pink bar) and "Microarray and Gene" (blue bar).

최근 유전자 마이크로어레이 기반 암 연구에서 유전적 정보와 임상정보의 결합을 통해서 의미 있는 유전자를 찾아낸 성공적인 실험이 많이 이루어지고 있다[4][5]. 유전자 마이크로어레이를 사용하는 암 연구에서 실험 결과의 정확한 분석을 위해서는 유전적 정보와 임상 정보의 유기적인 결합이 매우 중요하다[6]. 그러나 MAGE-OM (Microarray Gene Expression Object Model)은 마이크로어레이 실험의 전반적인 과정이나 실험의 결과에 대한 구조화된 표현을 목적으로 하기 때문에 암에 관련된 임상정보를 표현할 수 있는 구조를 제공하고 있지 않다.

본 연구에서는 이러한 문제점을 착안하여 암 정보와 임상정보의 유기적인 결합을 지원하기 위해서 Cancer Genomics Object Model을 개발하고자 한다. 그러나 암은 다양한 요인에 의해서 여러 기관에서 발생하는 매우 복잡한 질병이다. 또한 암은 기관별 특성이 다양해서 암의 전반적인 정보를 정의하고 기술하는 것은 어려운 문제이다. 이러한 문제의 해결 방안으로 다양한 관점에서 암 정보를 표현을 목적으로 하는 여러 표준안을 기반으로 암 관련 임상 정보의 구조를 정의했다.

2. 관련 연구

2.1 MAGE-OM 확장을 통한 새로운 모델

최근 많은 연구에서 유전자 발현 자료와 단백질체학이나 대사체학에 관련된 자료의 결합을

통해 유전자 발현과 단백질 발현 등을 유기적으로 해석을 통해 생물학적인 새로운 발견을 하고 있다[7][8]. 이러한 동향에 발맞추어 객체 모델 기반으로 유전자 발현 자료와 단백질체나 대사체학 자료의 연계를 위해 MAGE-OM을 확장하여 새로운 모델을 개발한 여러 연구가 있다.

첫 번째는 National Center for Toxicogenomics (NCT)에서 만든 SysBio-OM이다. SysBio-OM은 MAGE-OM과 단백질 발현 자료를 표현하기 위한 표준 모델인 PEDRo 그리고 단백질 상호작용, 대사체학 자료등을 표현하는 새로운 모델이다[9]. 이 모델의 목적은 Chemical Effects in Biological Systems(CEBS)system에서 transcriptomics, proteomics 그리고 metabolomics 자료들의 상호 작용에 대한 연구를 가능하게 하는 것이다.

두 번째는 FGE-OM은 MAGE-OM 과 PEDRo 그리고 Glasgow university 에서 만든 단백질 발현 자료를 표현하기 위한 모델인 Gla-PSI를 기반으로 만들어졌다. 이 모델은 마이크로어레이와 two-dimensional gel electrophoresis (2-DE)과 mass spectrometry (MS)에 의한 protein identification과 같은 단백질체학과 관련된 자료를 표현하는데 초점이 맞춰져 있다[10].

2.2 임상 자료와 유전학적 자료의 통합

최근 많은 연구자들이 의학 분야에서 유전자 기반의 방법을 적용해야 하고 임상 실습에서 유전학적인 정보와 임상 정보를 통합해서 적용해야 한다는 것을 강조하고 있다[11], [12], [13]. 실제로 의료 분야에서 유전자 관련 정보를 사용하는 경우(예: 골수이식에서 세포 종류, 유방암 진단을 위해 BRCA 유전자의 발현 여부 측정)가 증가함에 따라 HL7 Clinical-Genomics SIG (Special Interesting Group)에서는 유전적 정보가 포함된 임상 정보를 교환하기 위한 기본 단위로 Genotype Shared Model을 제시하였다[14]. 그러나 Genotype Shared Model은 임상 정보와 유전학적 자료의 온전한 통합을 위한 모델이 아니라 개개인의 환자에 대한 임상 자료와 유전체 자료의 교환을 위해 만든 모델이다.

3. 재료 및 방법

Cancer Genomics Object Model을 만들기 위해서는 암 정보 표현을 위한 구성요소를 추출해야 한다. 그러나 암은 다양한 기관에서 다양한 원인으로 발생하는 매우 복잡한 질병으로써, 암의 한 가지 측면을 보고 암을 정보 표현을 위한 구조를 만드는 것은 적절하지 않다. 그래서 이 연구에서는 다양한 암 정보를 표현을 위한 프로젝트의 결과물을 기준으로 암 정보 기술을 위한 구성요소를 추출했다.

첫 번째로 암 등록 관련 표준 포맷인 North American Association of Central Cancer Registries (NAACCR)을 들 수 있다. NAACCR은 1987년에 중앙 암 등록 표준안에 대한 필요성에 따라 재정되었고, 모든 암의 공통적인 특성을 16개의

섹션으로 구분되는 435개의 자료 요소로 표현하였다[15]. 본 연구에서 NAACCR은 암의 공통적으로 사용되는 특징에 대한 정보를 구조화 하는데 사용되었다.

두 번째로 암 임상 시험에서 사용되는 구성 요소를 정의한 NCI Common Data Element (CDE)가 있다. CDE는 NCI에서 진행하고 있는 암 기반의 연구의 기반 구조를 만들기 위한 프로젝트인 caCORE의 일부이다[16]. CDE의 목적은 암 관련 정보를 파악하고 이를 제대로 표현할 수 있는 구조를 정의하는 것이다. CDE는 암이 발생한 기관별로 정의되어 있으며, 암을 기술하는데 필요한 사항들을 포함하고 있다[17]. 현재 CDE는 암 예방 임상 실험 및 암 관련 생물학적 표지 연구 등 다양한 암 관련 연구 분야에서 개발되고 있다.

세 번째로 암 조직의 병리학적 정보 교환을 위해 College of American Pathologists (CAP)에서 만든 Cancer Protocol이 있다. CAP Cancer Protocol은 암이 발생하는 기관별로 발행되며, 병리학적 정보를 표현하기 위한 서식, 체크 리스트 표 형식의 데이터로 구성되어 있다[18]. 본 연구에서는 CAP Cancer Protocol은 암의 조직 병리학적 정보를 구조화하는데 유용하게 사용되었다. 또한 기관별 특이적인 조직 병리학적 암의 특성을 구조화하는데 사용되었다.

위의 세 가지의 표준안을 바탕으로 암 정보를 표현하기 위한 data element를 추출하였고, 이를 MAGE-OM에 추가하여 UML을 이용하여 표현하였다(그림 2). UML은 큰 소프트웨어의 개발 시 시스템에 대한 구조를 디자인하거나 보여줄 때 사용되는 표준화된 표기법이다[19]. UML로 표현된 Object Model은 개발자가 복잡한 시스템의 구성 요소를 쉽게 이해할 수 있게 해주고, XML로의 표현과 database의 구현을 쉽게 할 수 있다는 장점이 있다.

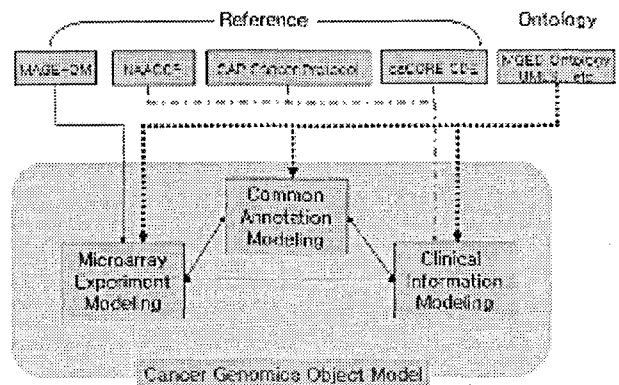


Figure 2. Architecture of Cancer Genomics Object Model

4. 결과

4.1 Overview of Cancer Genomics Object Model

Cancer Genomics Object Model은 마이크로어레이 실험, 임상 정보 그리고 공통 주석 부분으로 구성되어 있다. Cancer Genomics Object

Model은 모두 23개의 패키지로 구성된다[그림 3].

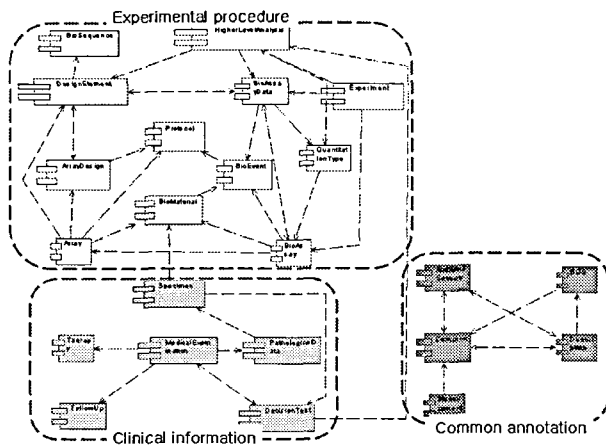


Figure 3. The relationship of Cancer Genomics Object Model packages

4.2 A general workflow of clinical examination in cancer

구조화된 암 정보의 표현을 위해서는 실제 임상에서 암을 진단, 치료에 대한 workflow에 대한 분석이 필요하다[그림 4]. 의사는 환자 정보, 사진, 청진, 촉진, 문진 등을 통해서 환자에 병명에 대한 일차적인 진단을 내리게 된다. 그 후 Lab, Image, Molecular, Pathological test 등의 결과를 통해 좀더 상세한 병명에 대한 결정을 내리게 되고, 병명 결정 후 환자의 치료 방법을 결정하게 된다. 환자에 대한 치료는 크게 surgical, radiological, chemotherapy가 이루어 질 수 있다. 이때 고형암인 경우에는 수술의 결과로 종양 조직을 얻게 되고 이것을 통해 2차적인 병리학적 검사를 거치게 된다. 모든 치료가 이루어진 후에 치료 후 경과(Post-treatment follow up)를 지켜보게 되는데, 이 결과에 따라서 새로운 치료 방법 결정이 이루어지게 된다. 모든 치료가 끝난 후에는 일정 기간동안 암의 재발여부를 추적하게 된다.

4.3 Microarray Experiment Packages

Microarray Experiment 패키지는 유전자 마이크로어레이 실험의 전반적인 설계에 대한 기술과 실험을 통해 얻어진 자료의 구조적인 표현을 위한 패키지이다. Microarray Experiment 패키지는 총 12개의 package로 구성되어 있다. 본 연구에서는 MAGE-OM의 Microarray experiment 부분을 참조하여 Cancer Genomics Object Model을 디자인하였다.

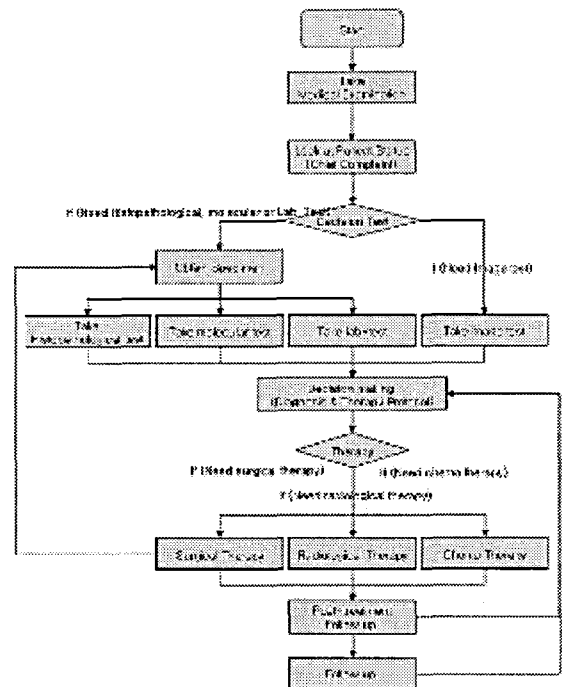


Figure 4 Workflow of clinical examination of cancer

4.4 Clinical Information Packages

Clinical Information 패키지는 MedicalExamination, Specimen, Pathologicaldata, DecisionTest, Therapy, FollowUp 과 같은 6개의 패키지로 구성되어 있다. 각각의 패키지는 암 정보 표현을 위한 NAACCR, NCI CDE, CAP cancer protocol등의 암 관련 표준안에서 추출한 구성 요소를 기준으로 만들어 졌다. 각각의 패키지들의 연관관계는 [그림 3]과 같다.

MedicalExamination 패키지는 의사에 의해 이루어지는 암 환자의 임상진료에 대한 정보를 의미하며, Clinical Information 부분에서 다른 5개의 패키지들과 밀접하게 연관된 중심 패키지라고 할 수 있다. MedicalExamination 패키지는 MedicalExamination class를 기준으로 PatientDemography, History, PhysicalExam, Diagnosis, Plan 등의 class로 구성되어 있다. Medical Examination은 환자의 신상정보, 과거력 그리고 사진, 청진, 촉진으로 구성되어 있는 Physical Exam, Diagnosis, 검사와 치료를 위한 Plan으로 구성되어 있다. 특히 History는 환자의 과거 질병력을 얻어내는 과정으로써, FamilyHistory, PastMedicalHistory, PresentIllness로 구성되어 있다. Diagnosis는 Pathological data와 Decision test의 결과를 통해 의사가 decision making을 하게 된다. 이와 같은 Medical Examination에 대한 상세 class diagram은 [그림 5]와 같다.

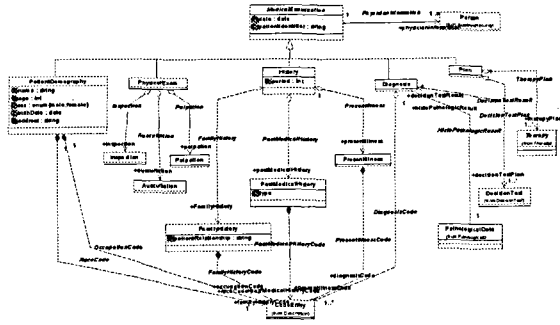


Figure 5. A class diagram of MedicalExamination

Specimen 패키지는 Clinical data와 Microarray data의 결합을 위해서 매우 중요한 패키지이다. Specimen 패키지는 수술이나 생검을 통해 얻어진 환자의 조직에 대한 자료를 구조화한 패키지이다. 조직은 기본적으로 Liquid, Solid의 두 가지 타입에 속하게 되며, 모든 조직은 조직을 얻게 된 날짜와 조직의 양에 대한 속성을 가지게 된다. 또한 조직을 설명하기 위해서는 조직을 모은 방법과 조직의 양을 재는 측량법에 대한 자료가 필요하다. 병리학적 검사, molecular test, lab-test, microarray은 이러한 조직을 사용하여 실험을 하게 되며, 이런 실험의 결과는 환자의 질병에 대한 정확한 판단을 내릴 수 있게 한다. Specimen package의 상세 class diagram은 [그림 6]과 같다.

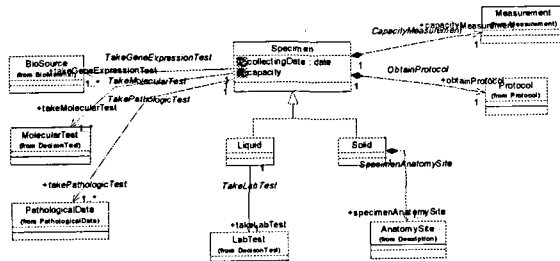


Figure 6. A class diagram of Specimen

PathologicalData Package는 수술이나 생검을 통해 얻어진 조직을 사용해서 병리학적 검사를 한 뒤에 나온 자료를 구조화 시킨 패키지이다. 암에서 병리학적 자료는 암의 경과나 환자의 치료에 영향을 주는 매우 중요한 요인이다. 병리학적 검사를 거친 자료는 일반적으로 종양에 대한 primary site, histology type, histology grade, tumor size, lymph node involvement, 그리고 distance metastasis 정보는 모든 암에 상관없이 기술한다. 부가적으로 Primary Site와 Distance Metastasis 등은 해부학적인 위치에 대해서 기술해야 한다. 그러나 암은 기관 특이적인 특성이 많이 존재하기 때문에 이 패키지에서는 organ specific factor를 통해서 이를 해결하였다. Tumor Size는 종양의 크기를 재는 방법과 단위에 대한 정보가 필요하다. 수술을 통해 나온 종양 조직과 다르게 생검을 통해 얻어진 조직은 병리학적 검사를 통해 얻을 수 있는 정보는 Primary Site와 Histology 외에는 존재하지 않는다. PathologicalData에 대한 상세 class diagram은

[그림 7]과 같다.

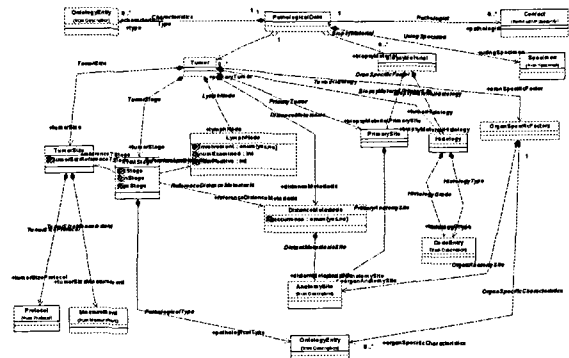


Figure 7. A class diagram of PathologicalData Package

DecisionTest Package는 임상에서 환자의 병명을 결정할 때 사용되는 여러 가지 검사 방법에 대한 패키지이다. DecisionTest는 MolecularTest, ImageTest, LabTest등의 세 가지 검사 방법으로 구성되고, 이중 Molecular Test는 clinical data와 microarray data의 연결고리로 사용된다. 예를 들어 p53과 같은 암 발생과 관련이 있는 유전자의 발현을 보는 것도 molecular test로 볼 수 있다. ImageTest는 MRI나 x-ray film을 보고 의사가 판독한 내용을 구조화 시킨 것이다. LabTest는 다양한 feature에 대해서 기술하기 때문에 OntologyEntry를 통해서 구조화 시켰다. DecisionTest에 대한 상세 class diagram은 [그림 8]과 같다.

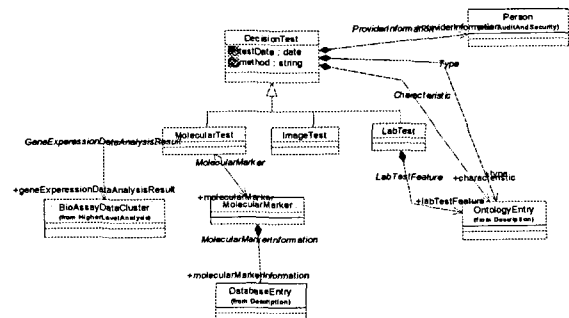


Figure 8. A class diagram of DecisionTest package

Therapy Package는 MedicalExamination을 통해 진단을 내린 후에 환자 치료에 대한 부분을 구조화 하였다. 환자에 대한 치료는 기본적으로 surgical, radiological, chemo therapy로 구성되며, 모든 치료는 치료 기간 (startDate, endDate)과 치료 방법 (method), 치료 목적 (purpose), 치료를 시행한 의사에 대한 정보, 치료 후 경과에 대한 기술이 필요하다. Surgical therapy는 surgical margin과 surgical location 대한 해부학적 정보가 필요하고, 이 Surgical therapy를 통해서 환자의 조직을 얻을 수 있다. Therapy에 대한 상세 class diagram은 [그림 9]와 같다.

- Suppl:469-73.
- [3] Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, et al. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* 2002 Aug 23;3(9):RESEARCH0046.
- [4] Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med.* 2002 Jun 20;346(25):1937-47.
- [5] Berns A. gene expression in diagnosis. *Nature.* 2000 Feb 3;403(6769):491-2.
- [6] Dugas M, Schoch C, Schnittger S, Kohlmann A, Kern W, Haferlach T, Uberla K. Impact of integrating clinical and genetic information. *In Silico Biol.* 2002; 2(3): 383-91
- [7] Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet.* 2001;2:343-72.
- [8] Waters M, Boorman G, Bushel P, Cunningham M, Irwin R, Merrick A, Olden K, Paules R, Selkirk J, Stasiewicz S, Weis B, Van Houten B, Walker N, Tennant R. Systems toxicology and the Chemical Effects in Biological Systems (CEBS) knowledge base. *EHP Toxicogenomics.* 2003 Jan;111(1T):15-28.
- [9] Xirasagar S, Gustafson S, Merrick BA, Tomer KB, Stasiewicz S, Chan DD, Yost KJ 3rd, Yates JR 3rd, Sumner S, Xiao N, Waters MD. CEBS object model for systems biology data, SysBio-OM. *Bioinformatics.* 2004 Sep 1;20(13):2004-15. Epub 2004 Mar 25.
- [10] Jones A, Hunt E, Wastling JM, Pizarro A, Stoeckert CJ Jr. An object model and database for functional genomics. *Bioinformatics.* 2004 Jul 10;20(10):1583-90. Epub 2004 May 14.
- [11] Altman RB. Bioinformatics in support of molecular medicine. *Proc AMIA Symp.* 1998;:53-61
- [12] Sander C. Genomic medicine and the future of health care. *Science.* 2000; 17:287(5460): 1977-8
- [13] Kohane IS. Bioinformatics and clinical informatics: the imperative to collaborate. *J Am Med Inform Assoc.* 2000; 7(5): 512-6
- [14] Health Level 7. Genotype Shared Model. Version 0.9, HL7 Web site. Available at: <http://www.hl7.org/special/Committees/clingenomics/clingenomics.htm>. Accessed Jan 29, 2005.
- [15] Seiffert JE. Development and use of the North American Association of Central Cancer Registries standards for cancer registries. *Top Health Inf Manage.* 1997 Feb;17(3):35-44.
- [16] Covitz PA, Hartel F, Schaefer C, De Coronado S, Fragoso G, Sahni H, Gustafson S, Buetow KH. caCORE: a common infrastructure for cancer informatics. *Bioinformatics.* 2003 Dec 12;19(18):2404-12.
- [17] Warzel DB, Andonaydis C, McCurry B, Chilukuri R, Ishmukhamedov S, Covitz P. Common data element (CDE) management and deployment in clinical trials. *AMIA Annu Symp Proc.* 2003;:1048.s
- [18] Leslie KO, Rosai J. Standardization of the Surgical Pathology Report: Formats, Templates, and Synoptic Reports. *Seminars in Diagnostic Pathology* 1994.; 11(4):253-7.
- [19] Rumbaugh, J., Jacobson, I. and Booch, G. The Unified Modeling Language Reference Manual. Addison-Wesley, Reading, MA, USA.
- [20] Russo G, Zegar C, Giordano A. Advantages and limitations of microarray technology in human cancer. *Oncogene.* 2003 Sep 29;22(42):6497-507.
- [21] European Bioinformatics Institute (EBI). MIAME/Tox. Draft August 2003, EBI Web Site. Available at: 21. <http://www.ebi.ac.uk/microarray/Projects/tox-nutri/index.html>. Accessed Jan 29, 2005.
- [22] Dolin RH, Alschuler L, Beebe C, Biron PV, Boyer SL, Essin D, Kimber E, Lincoln T, Mattison JE. The HL7 Clinical Document Architecture. *J Am Med Inform Assoc.* 2001 Nov-Dec;8(6):552-69.
- [23] Yu Rang Park, B.S., Ji Yeon Park, M.S., Se Young Kim, M.S., Hwa Jeoung Seo, Ph.D., Jung Ae Lee, M.S., Ju Han Kim M.D., Ph.D. Integrating Microarray Gene Expression-Object Model and Clinical Document Architecture using Common Data Element. *The Korean Society of Medical Informatics* 2004;10(Suppl 1):s41-s45.