Data-related Challenges of Genotype - Phenotype studies

Sax U, Mohammed Y
Abteilung Medizinische Informatik, CIOffice Forschungsnetze, Bereich Humanmedizin, Universität Göttingen, Deutschland
usax@med.uni-goettingen.de

Introduction and Question

Genome wide association studies are becoming more and more interesting for the biomedical community. As genotyping constantly gets cheaper[1, 2], many formerly phenome-related projects around complex diseases consider genotyping within the next couple of years.

Beyond the indisputable opportunities of these studies there are quite some challenges to be faced. A) where do we get sufficient case numbers for these studies, B) where do we get high quality patient material to be genotyped, C) where do we get the corresponding high quality phenotype data[3, 4], D) how can we homogenize the heterogeneous data sources[5] and E) how do we deal with the well-known privacy problems of these genotype-phenotype studies[6, 7]?

Materials and Methods

The above questions are raised on the background of two related BMBF-funded projects, namely the competence network for congenital heart disease (CHD)[8] as an interesting phenotype resource and MediGRID[5], dealing with the homogenization of heterogeneous data sources and the related privacy aspects.

Results

In Germany and world wide exist many disease related data collections in the form of registries or biomaterial collections. Whereas the registries for most complex diseases with sufficiently high prevalence in the target population are usually well organized, the world of biomaterial banks is likewise scattered due to ethical and legal reasons. This might be leveraged by the TMF biomaterial working group[9].

As genotyping prices constantly go down, the threshold for the entry in genotyping projects gets lower. But any sequencing or SNP-analysis does only get its value by the proper phenotype annotation. The phenotype could either be captured via expensive clinical trials, via using phenotype data from hospital data bases - taking into account the quality uncertainty or using phenotype data from disease-related registries.

The challenge starts as soon as genotype and related phenotype data from different data sources are available for further analysis.

We have to deal with several new data types, not being standardized. Genomic data tends to be more structured than phenotype data, as the Bioinformatics community is open source and XML based. Phenotype data is kept mostly in traditionally “hand carved”, non-compatible Information systems. Structured document approaches run since some years[10, 11] with moderate success. But for association studies not only the data formats have to be homogenized, more importantly the content has to be homogenized. Ontology-approaches using UMLS[12] had some success recently[13]. Finally new privacy models have to be developed and consented, as usually association studies tend the threaten privacy[6]. The TMF privacy working group is working on that challenge as well[14].

Discussion

Given the necessity to capture both environment and genomic state of a patient and their interaction, clinical information systems have to be redesigned. While genotyping seems to be automatable easily, this is not the case for clinical information. More integration work on terminologies and ontologies is to be done.

Researchers from medical informatics, bioinformatics and epidemiology will have to collaborate much more intensively than they formerly did. One of the main problems may be the different vocabulary and the different background of these researchers. Sustainable collaborations would give German Biomedical Informatics a competitive edge in the community.

This work was supported by the D-Grid Project MediGRID, funded by the Federal Ministry of Education and Research (BMBF), FKZ 01AK803H, and by the competence network for congenital heart disease (AHF), funded by the Federal Ministry of Education and Research (BMBF), FKZ 01G10210.

References

[1] Westphal, S.P. *Race for the $1000 genome is on*. 2002 [cited; Available from: http://www.newscientist.com/article.ns?id=dn2900.
[2] Greeley, M. *Nothing ventured ... Two Cents on the '$1,000 Genome'*. 2006 [cited 09.04.2006]; Available from: http://www.bio-itworld.com/archive/081303/ventured.html.
[3] Powell, J. and I. Buchan, *Electronic health records should support clinical research.* J Med Internet Res, 2005. **7**(1): p. e4.
[4] Dumitru, R.C. and O. Rienhoff, *Challenges to Patients Medical Records Supporting Clinical Research – Data Quality*, in *1st International Conference on Information Communication Technologies in Health*. 2003: Samos island, Greece.
[5] MediGRID. *Medical Grid Computing*. 2005 [cited; Available from: www.medigrid.de.
[6] Lin, Z., A.B. Owen, and R.B. Altman, *Genetics. Genomic research and human subject privacy*. Science, 2004. **305**(5681): p. 183.
[7] Malin, B.A., *An Evaluation of the Current State of Genomic Data Privacy Protection Technology and a Roadmap for the Future.* J Am Med Inform Assoc, 2004.
[8] KN-AHF. *Kompetenznetz Angeborene Herzfehler*. 2005 [cited 21.09.2005]; Available from: www.kompetenznetz-ahf.de.
[9] TMF-AG_Biomaterial. 2006 [cited 09.04.2006]; Available from: http://www.tmf-ev.de/site/DE/int/AG/BMB/container_bmb.php.
[10] HL7, *HL7 Receives ANSI Approval of Three Version 3 Specifications Including CDA, Release 2*. 2005.
[11] Dolin, R.H., et al. *HL7 Clinical Document Architecture (Release 2.0)*. 2004 [cited February 9, 2005]; Committee Ballot #3; Aug 03,2004:[Available from: http://hl7.org/library/Committees/structure/CDA.ReleaseTwo.CommitteeBallot03.Aug.2004.zip.
[12] Bodenreider, O., *The Unified Medical Language System (UMLS): integrating biomedical terminology*. Nucleic Acids Res, 2004. **32 Database issue**: p. D267-70.
[13] Butte, A.J. and I.S. Kohane, *Creation and implications of a phenome-genome network.* Nat Biotechnol, 2006. **24**(1): p. 55-62.
[14] TMF. *AG Datenschutz*. 2006 [cited 16.03.2006]; Available from: http://www.tmf-ev.de/site/DE/int/AG/DS/container_ag_ds.php.