Closing the Genotype-phenotype Gap
Findings from the Section on Bioinformatics

Y. L. Yip, Section Editor for the IMIA Yearbook Section on Bioinformatics
Knowledge Management, Merck Serono International S.A., Geneva, Switzerland

Summary
Objectives: To summarize current excellent research in the field of bioinformatics.
Method: Synopsis of the articles selected for the IMIA Yearbook 2010.
Results: The selection process for this yearbook's section on Bioinformatics results in five excellent articles highlighting the progress made in advancing the understanding of genotype-phenotype relationship, and their concrete application in clinical settings. First, next generation sequencing techniques have allowed the discovery of an ever larger number of genetic variations at a greater resolution, and methods were developed to ensure accurate data analysis. Second, innovative approaches were applied to gene expression data to allow its link to a wider phenotypic spectrum and to enhance its use for disease understanding. Third, there is a notable trend in visualizing diseases as network rather than individual entities, and this has provided new insights for disease interpretation. The progress mentioned above is further aided by continual development in bio-ontologies which provide means for semantic, and thus phenotype, comparison.
Conclusions: The current literature showed a tightening link between genotype and phenotype, placing us one step closer to a better disease classification, patient stratification as well as the development of personalized medicine.

Keywords
Medical informatics, International Medical Informatics Association, yearbook, bioinformatics, genotype-phenotype relationships, bio-ontologies, semantic similarity

Yearb Med Inform 2010: 82-5

Introduction

Recent technological developments such as high-throughput arrays and next-generation sequencing have propelled genetic medicine to an exciting new era, characterized by the abundance of genomic data. In this context, the relatively scarce amount of phenotype data clearly makes the latter the bottleneck of any optimal exploitation of human genetic variation information. Recent survey of the literature highlights the awareness of this fact, and the progress made in using innovative approaches to close the genotype to phenotype gap. Realizing the inherent technical difficulties in obtaining high-throughput phenotypic data, these approaches rely on computational techniques which either extend the use of gene expression data beyond gene-disease association to a finer gene-phenotype association, or adopt a network- or semantic-based analysis strategy [1-4]. Indeed, ever since the first works describing the modular nature of human diseases and human disease network were published [5-6], the network view of diseases is gaining popularity and this has allowed novel interpretations of disease phenotypes. Different from gene sequences, the direct comparison between phenotypes is far from trivial as it often involves comparing textual descriptions. In this aspect, the continual development of biomedical ontologies and semantic similarity measures will certainly alleviate this difficult task [7].

In the 2009 Yearbook, it was noted that community-based data sharing solutions enabling more efficient multi-institutional collaborative works was gaining momentum [8], recent survey shows that these efforts and trends continue [9-12].

Best Paper Selection

The best paper selection of articles for the section 'bioinformatics' in the IMIA Yearbook 2010 follows the tradition of previous yearbooks [8,13] in presenting examples of excellent research in bioinformatics that are most relevant to medical informatics. As a result of a comprehensive review process, five articles were selected from international peer-reviewed journals in the fields of medicine, medical informatics, and bio-informatics.

The first paper reflects the advance made in the discovery of single nucleotide polymorphisms (SNP) using next-generation sequencing technology [14]. Two other papers exemplify the current research strategies in enriching and exploiting the phenome space using genomic data. Xu *et al.* described a computational method, PhenoProfiler, for predicting the quantitative phenotype information missing from a genomic dataset [1]. Hu *et al*., on the other hand, made use of the large amount of gene expression data to create a large-scale disease-drug network. The network was demonstrated to be effective for drug repositioning and drug target/pathway identification [2]. Schwarz *et al*. further extended the network ap-

proach to clinical data, and investigated the relationship between patient specific variables and the disease. Their approach may prove useful to improve diagnosis and better disease understanding [3]. The last selected paper offered a review and a classification on semantic similarity measures [7].

Table 1 presents the selected papers. A brief summary of the selected best papers can be found in the appendix of this report.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2010 in the section 'Bioinformatics'. The articles are listed in alphabetical order of the first author's surname.

Section Bioinformatics
▪ Hu G, Agarwal P. Human disease-drug network based on genomic expression profiles. PLoS One 2009;4(8):e6536.
▪ Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J. SNP detection for massively parallel whole-genome resequencing. Genome Research 2009;19:1124-32.
▪ Pesquita C, Faria D, Falcao AO, Lord P, Couto FM. Semantic similarities in biomedical ontologies. PLoS Computational Biology 2009; 5(7):e1000443.
▪ Schwarz E, Leweke FM, Bahn S, Lio P. Clinical bioinformatics for complex disorders: a schizophrenia case study. BMC Bioinformatics 2009;10 (Suppl 12):S6.
▪ Xu M, Li W, James GM, Mehan MR, Zhou XJ. Automated multidimensional phenotype profiling using large public microarray repositories. Proc. Natl. Acad. Sci USA 2009;106:12323-8.

Conclusions and Outlook

An initiative for a Human Phenome Project was suggested as early as 2003 [15]. The progress made in this domain is however clearly not comparable to that in the field of genomics. Phenotype data is complex, and its complexity is not only at the technical handling level, but also at the semantic representation level. The best paper selection for the Yearbook section 'Bioinformatics' shows the current progress made in this area, especially in terms of informatics solution and analysis strategies. These advances should render feasible the systematic and rigorous quantitative analysis of disease phenotypes. By closing the genotype and phenotype gap, a greater insight into clinical manifestation of genome abnormalities, the interplay of different genes in complex diseases, as well as molecular mechanisms of pathophysiology should be gained. In this regard, large collaborative efforts, such as the European 7th Framework project Gen2Phen (www.gen2phen.org), should further accelerate the process.

Acknowledgement
I would like to acknowledge the support of Martina Hutter and the reviewers in the selection process of the IMIA Yearbook.

References

1. Xu M, Li W, James GM, Mehan MR, Zhou XJ. Automated multidimensional phenotype profiling using large public microarray repositories. Proc Natl Acad Sci USA 2009;106:12323-8.
2. Hu G, Agarwal P. Human disease-drug network based on genomic expression profiles. PLoS One 2009;4(8):e6536.
3. Schwarz E, Leweke FM, Bahn S, Lio P. Clinical bioinformatics for complex disorders: a schizophrenia case study. BMC Bioinformatics 2009; 10(Suppl 12):S6.
4. Sadreyev RI, Feramisco JD, Tsao H, Grishin NV. Phenotypic categorization of genetic skin diseases reveals new relations between phenotypes, genes and pathways. Bioinformatics 2009;25(22):2891-6.
5. Oti M, Brunner HG. The modular nature of genetic diseases. Clin Genet 2007;71:1-11.
6. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL: The human disease network. Proc Natl Acad Sci U S A 2007; 104:8685-90.
7. Pesquita C, Faria D, Falcao AO, Lord P, Couto FM. Semantic similarities in biomedical ontologies. PLoS Computat Biol 2009,5(7): e1000443.
8. Yip YL. Accelerating knowledge discovery through community data sharing and integration. Findings from the Yearbook 2009 Section in Bioinformatics. Yearbook of Medical Informatics 2009;117-20.
9. Das S, Girand L, Greem T, Weitzman L, Lewis-Bowen A, Clark T. Building biomedical web communities using a semantically aware content management system. Brief Bioinform 2008;10(2): 129-38.
10. Janevski A, Kamalakaran S, Banerjee N, Varadan V, Dimitrova N. PAPAyA: a platform for breast cancer biomarker signature discovery, evaluation and assessment. BMC Bioinformatics 2009 10(Suppl 9):S7.
11. Liu X, Wu J, Wang J, Liu X, Zhao S, Li Z, et al. WebLab: a data-centric, knowledge-sharing bioinformatics platform. Nucleic Acids Res 2009;37: W33-W39.
12. Kelder T, Pico AR, Hanspers K, van Iersel MP, Evelo C, Conklin BR. Mining biological pathways using WikiPathways Web Services. PLOS One 2009;4(7):e6447.
13. Yip YL. The promise of systems biology in clinical applications. Findings from the Yearbook 2008 Section in Bioinformatics. Yearbook of Medical Informatics 2008;102-4.
14. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J. SNP detection for massively parallel whole-genome resequencing. Genome Res 2009;19: 1124-32.
15. Freimer N, Sabatti C. The Human Phenome Project. Nature Genetics 2003;34:15-21.
16. Li R, Li Y, Kristiansen K, Wang J. SOAP: Short oligonucleotide alignment program. Bioinformatics 2008;24:713–4.

Correspondence to:
Dr. Yum Lina Yip
Knowledge Management
Merck Serono S.A.
9 Chemin des Mines
Geneva, Switzerland
Tel: +41 22 414 3937
Fax: +41 22 414 3059
E-mail: lina.yip.sonderegger@merckserono.net

Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2010, Section Bioinformatics*

Hu G, Agarwal P
Human disease-drug network based on genomic expression profiles
PLoS One 2009;4(8): e6536

Disease-drug relationships are useful for the understanding of disease mechanisms, and for accelerating the drug discovery pipeline. The traditional views of diseases and drug actions are reductionist and have many limitations. In this paper, the authors took advantage of the rapid accumulation and availability of whole genome gene expression data to generate a large-scale disease-drug network. The main assumption of the approach is that effects of many diseases and drugs can be characterized to some extend by their gene expression. Therefore, diseases and drugs can be related by directly matching their induced expression profiles. The authors used human GEO datasets to generate human disease and drug genomic profiles, and developed an automatic process to systematically compare and analyze them. The resultant network includes 645 disease-disease, 5008 disease-drug and 164373 drug-drug relationships. The network was demonstrated to be valuable for revisiting disease classification, drug repositioning, identifying potential drug side-effect as well as deconvoluting drug targets or pathways. The method has the advantages of being scalable and allowing network update when more gene expression data are available. It suffers however from a high false-negative rate.

* The complete papers can be accessed in the Yearbook's full electronic version, provided that permission has been granted by the copyright holder(s).

Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J
SNP detection for massively parallel whole-genome resequencing
Genome Research 2009;19:1124-32

The next-generation sequencing technologies offer ultrahigh throughput at a reduced cost. They are thus particularly suitable for carrying out genetic variation studies which are important in determining risk to certain diseases, and response to drugs and the environment. While a universal standard for defining SNP detection accuracy is available for traditional Sanger sequencing techniques, novel methods for accurate SNP detection using next-generation sequencing techniques are still rare. In this paper, the authors described a method for the consensus calling and SNP detection for the massively parallel sequencing-by-synthesis Illumina GA technology. The method was developed primarily to handle the consensus assembly and SNP detection of one haploid or diploid genome with a known reference sequence, and took into account data quality, alignment, and experimental errors commonly seen in the Illumina sequencing technology. By using a Bayesian statistical method, a single quality score was derived to measure the accuracy of each nucleotide position in the consensus sequence. The quality of this method was evaluated using the high-quality Asian genome resequencing data, and was shown to have a very low false call rate at any sequencing depth. The method also offered excellent genome coverage in high-depth data, making it very useful for SNP detection at any sequencing depth. The methodology and the developed software described have been integrated into the previously described Short Oligonucleotide Alignment Program (SOAP) package [16].

Pesquita C, Faria D, Falcao AO, Lord P, Couto FM
Semantic similarities in biomedical ontologies
PLoS Computational Biology 2009;5(7): e1000443

Ontologies have gained importance in recent years in biomedical research as they are able to provide the formalism and common terminology necessary for researchers to describe, compare and share their results. The comparison between concepts, or entities annotated with those concepts is called semantic similarity. In this paper, the authors reviewed a number of recently described semantic similarity measures in the context of Gene Ontology. They also proposed a classification of these measures according to the strategies employed: node-based (which rely on the term themselves) versus edge-based (which rely on the structure of the ontology), and pairwise versus groupwise in situations where sets of concepts are being compared. Comparative assessment studies and examples of applications to biomedical research were presented. More importantly, the authors provided advices to researchers on how to choose the approach most suitable for their studies and how to best benefit from semantic similarity measures. As biomedical ontologies evolve towards increased coverage, formality and integration, the authors pointed out the importance of the development of gold standard corpora that would allow the effective comparison of semantic similarity measures. It is expected that semantic similarity measures will gain more relevance and become essential in biomedical research.

Schwarz E, Leweke FM, Bahn S, Lio P
Clinical bioinformatics for complex disorders: a schizophrenia case study
BMC Bioinformatics 2009;10(Suppl 12):S6

Clinical bioinformatics has the key goal to simultaneously exploit clinical and basic research data to improve patient care. In this study, the authors used Schizophrenia, a complex psychiatric

disorder with a broad spectrum of different clinical manifestations, to demonstrate how clinical bioinformatics approach could possibly improve diagnosis and better disease understanding. Their approach consisted in capturing and representing information available in complex diseases, such as clinical, laboratory, genetic and molecular profiling data, in networks. They then applied graph theoretical procedure to investigate the relationship between patient specific variables and the disease. The method was able to provide an estimate of the heterogeneity of the population of schizophrenia, and identify a subgroup of patients featuring abnormalities in a serum primary fatty acid network. The stability of this molecular network was compared in an extended dataset between schizophrenia and affective disorder patients and was found to be more stable in the latter. The new insights gained in this paper may prove useful in complex disease sub-classification and aid in the development of personalized medicine.

Xu M, Li W, James GM, Mehan MR, Zhou XJ

Automated multidimensional phenotype profiling using large public microarray repositories

Proc Natl Acad Sci USA 2009;106:12323-8

While the aim of modern genetics is to link genotype to phenotype, the paucity of phenotype data as compared to genomics data makes the former the bottleneck of the process. Indeed, it is difficult to quantify phenotypes in a high-throughput manner due to its complexity. Most phenotypic data are also qualitative rather than quantitative. This article described „PhenoProfiler," a computational method for predicting the quantitative phenotype information missing from a genomic dataset. The principle of PhenoProfiler is that similar genomic patterns are likely to be associated with similar phenotypic patterns. Thus, one can supplement the missing phenotypic information in a given genomics dataset with traits in other well-characterized datasets. In particular, this method associates each sample of a given dataset with the relative intensity of a specific phenotype trait. „Phenotype profile" thus refers to the quantitative measures of samples across the whole dataset. The method was applied to 587 human microarray datasets, covering >14,000 microarray samples. It was shown that the predicted phenotype profiles were highly consistent with known phenotype descriptions. The method offers several advantages, including the power to predict multiple phenotype profiles for a particular dataset thus facilitate the analysis of complex diseases and treatment design; the ability to provide quantitative phenotype description and to extrapolate phenotype profiles beyond provided classes. It also allows the detection of confounding phenotype factors that could otherwise bias biological inferences. Finally, since the method can be applied to cross-platform microarray data, its power and usefulness will only be enhanced with the continued accumulation of genomics data that provide the variety and phenotypes to be profiled.