Bioinformatics and its Impact on Clinical Research Methods

Findings from the Section on Bioinformatics

E. Lang, Managing Editor for the IMIA-Yearbook Section on Bioinformatics
University of Applied Sciences Darmstadt, Dept. of Information and Kowledge Management, Darmstadt, Germany

Summary

Objectives: To summarize current excellent research in the field of bioinformatics.

Method: Synopsis of the articles selected for the IMIA Yearbook 2006.

Results: Current research in the field of bioinformatics clearly shows ongoing unification of experimental findings and clinical outcomes. Microarray data, gene sequences and clinical data are more and more perceived as different but related facets of one entity. Significant work is done in the area of text and data mining in order to bring together patient data and biochemical phenomena by means of ontologies. A strong trend in the clinical field is performance of exhaustive studies on DNA material derived from patients that suffer from diseases that are already known to be inherited. Examination of appropriate methods covers data and text mining, ontologies as well as machine learning and classification.

Conclusions: The best paper selection of articles on bioinformatics shows examples of excellent research on methods used for studying inherited diseases and their underlying genetic dispositions. Clinical studies, inclusion of experimental findings like microarray data, and of knowledge representation formats all lead to a better understanding the linkage between gene sequences, biological functions and clinical findings in the form of healthy state or physiological disorders.

Haux R, Kulikowski C, editors. IMIA Yearbook of Medical Informatics 2006. Methods Inf Med 2006; 45 Suppl 1: S104-6.

Keywords

Medical Informatics, International Medical Informatics Association, Yearbook, Bioinformatics.

Introduction

Clinical bioinformatics should focus on relating clinical practise and biochemical/genetic principles. The common problem in clinical bioinformatics is relating microscale findings derived from experimental data (microarray data, Serial Analysis of Gene Expression, linkage analysis, structure analysis methods as mass spectroscopy) and macroscale properties such as disease symptoms and metabolic pathways. At the moment, there are hardly any data that could fulfill the requirements on both scales: there are gene sequence and annotation data related to microscale phenomena on the one side, and on the other there are clinical studies describing the findings on the macroscale. Many approaches deal with methods trying to relate data of both types, and they mostly use ontologies and automatic extraction methods to process the vast amounts of data that can be found in this realm ([1-8]). Building bridges between experimental and clinical findings will open new possibilities in diagnostic and therapeutic efforts ([9]).

Best Paper Selection

The best paper selection of articles for the section 'bioinformatics' in the IMIA Yearbook 2006 reflects these trends and follows the tradition of previous yearbooks ([10,11]) in presenting examples of excellent research on methods used for biomedical text mining, automatic processing of gene sequence data, handling and screening gene data annotations as well as comprehensive clinical studies.

Five excellent, mostly multi-national, articles representing the research in four different continents were selected from four international peer reviewed journals in the fields of medicine, medical informatics, and bioinformatics. Table 1 presents the selected papers. A brief content summary of the selected best papers can be found in the appendix of this report.

Conclusions and Outlook

The best paper selection for the Yearbook section 'bioinformatics' clearly indicates that clinical bioinformatics has been established as a field where classical bioinformatics, with basic techniques as pattern recognition, data mining, and chemical structure analysis, and clinical applications meet in a fruitful way. Perhaps the findings of genetic analysis will influence clinical work in a way that is comparable to 19 th century's recognition of bacteria as the cause of infectious diseases. Further work will be necessary in refining ex-

Table 1 Best paper selection of articles for IMIA Yearbook of Medical Informatics 2006 in the section 'Bioinformatics'. The articles are listed in alphabetical order of the first author's surname.

Section Bioinformatics
▪ Baumgartner C, Bohm C, Baumgartner D, Marini G, Weinberger K, Olgemoller B, Liebl B, Roscher AA. Supervised machine learning techniques for the classification of metabolic disorders in newborns. Bioinformatics 2004; 20(17): 2985-96.
▪ Hauser ER, Crossman DC, Granger CB, Haines JL, Jones CJ, Mooser V, et al. A genomewide scan for early-onset coronary artery disease in 438 families: the GENECARD Study. Am J Hum Genet 2004; 75: 436-47.
▪ Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Using literature-based discovery to identify disease candidate genes. Int J Med Inform 2005; 74: 289-98.
▪ Koike A, Niwa Y, Takagi T. Automatic extraction of gene/protein biological functions from biomedical text. Bioinformatics 2005; 21(7): 1227-36.
▪ Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Research 2005; 33(5): 1544-1552.

perimental work, but also in carrying out more detailed analyses on patient data that have already been screened in order to identify disease gene candidates ([12], [13]). Text mining and relating text data to gene sequences will remain a challenge for the meantime as there are plethora of raw data waiting for analysis and classification ([1],[2],[5]). Today's work still focuses on putting together the two faces of the medal. Probably, the near future will show its value.

Up-to-date information about current and future issues of the IMIA Yearbook is available at http://iig.umit.at/yearbook/.

Acknowledgement

We greatly acknowledge the support of Martina Hutter and of the reviewers in the selection process of the IMIA Yearbook.

References

1. Gaizauskas R, Demetriou G, Artymiuk PJ, Willett P. Protein Structures and Information Extraction from Biological Texts: The PASTA System. Bioinformatics 2003; 19(1): 135–43.
2. Ivanciuc O, Schein CH, Braun W. Data mining of sequences and 3D structures of allergenic proteins. Bioinformatics 2002; 18(10): 1358–64.
3. Ohno-Machado L, Vinterbo S, Weber G. Classification of gene expression data using fuzzy logic. Journal of Intelligent and Fuzzy Systems 2002; 12:19-24.
4. Petricoin EF, Liotta LA. Proteomic analysis at the bedside: early detection of cancer. Trends in Biotechnology 2002; 20 (12 Suppl): S30-4.
5. Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics 2002; 18(8): 1124–32.
6. Wallqvist A, Rabow AA, Shoemaker RH, Sausville EA, Covell DG. Linking the growth inhibition response from the National Cancer Institute's anticancer screen to gene expression levels and other molecular target data. Bioinformatics 2003; 19(17): 2212–24.
7. Sharma R, Maheshwari JK, Prakash T, Dash D, Brahmachari SK. Recognition and analysis of protein-coding genes in severe acute respiratory syndrome associated coronavirus. Bioinformatics 2004; 20(7): 1074–80.
8. Ein-Dor L, Kela I, Getz G, Givol D, and Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 2005; 21(2): 171-8.
9. Maojo V, Martin-Sanchez F. Bioinformatics: towards new directions for public health. Methods Inf Med 2004; 43: 208-14.
10. Knaup P, Ammenwerth E, Brandner R, Brigl B, Fischer G, Garde S, Lang E, Pilgram R, Ruderich F, Singer R, Wolff AC, Haux R, Kulikowski C. Towards clinical bioinformatics: advancing genomic medicine with informatics methods and tools. Methods Inf Med 2004; 43: 302-7.
11. Bott OJ, Ammenwerth E, Brigl B, Knaup P, Lang E, Pilgram R, et al. The challenge of ubiquitous computing in health care: technology, concepts and solutions. Findings from the IMIA Yearbook of Medical Informatics 2005. Methods Inf Med 2005; 44: 473-9.
12. Baumgartner C, Bohm C, Baumgartner D, Marini G, Weinberger K, Olgemoller B, et al. Supervised machine learning techniques for the classification of metabolic disorders in newborns. Bioinformatics 2004; 20(17): 2985-96.
13. Hauser ER, et al. A genomewide scan for early-onset coronary artery disease in 438 families: the GENECARD Study. Am J Hum Genet 2004; 75: 436-47.
14. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Using literature-based discovery to identify disease candidate genes. Int J Med Inform 2005; 74: 289-98.
15. Koike A, Niwa Y, Takagi T. Automatic extraction of gene/protein biological functions from biomedical text. Bioinformatics 2005; 21(7): 1227-36.
16. Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Res 2005; 33(5): 1544-52.

Correspondence to:
Elke Lang
University of Applied Sciences Darmstadt
Information and Knowledge Management
Haardtring 100
64295 Darmstadt
Germany
E-mail: lang@iuw.h-da.de

Appendix: Content Summaries of Selected Best Papers, Section Bioinformatics*

Baumgartner C, Bohm C, Baumgartner D, Marini G, Weinberger K, Olgemoller B, Liebl B, Roscher AA.

Supervised machine learning techniques for the classification of metabolic disorders in newborns.

Bioinformatics 2004; 20(17): 2985-96

Huge amounts as well as complexity of screening data imply the necessity for automatic processing of experimentally derived data. [12] have examined data from the Bavarian newborn screening programme which served to test all newborns for about 20 inherited metabolic disorders. Machine learning techniques were expected to reveal novel patterns in high-dimen-

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

sional metabolic data and find classification rules with high discriminatory power. The study was performed with six different machine learning techniques and the examples of two metabolic disorders, phenylketonuria and medium-chain acyl-CoA dehydrogenase deficiency. [12] show that their results could achieve all known metabolic pattern findings and indicate some novel patterns that could lead further in the understanding of newborn metabolism.

Hauser ER, Crossman DC, Granger CB, Haines JL, Jones CJ, Mooser V, McAdam B, Winkelmann BR, Wiseman AH, Muhlestein JB, Bartel AG, Dennis CA, Dowdy E, Estabrooks S, Eggleston K, Francis S, Roche K, Clevenger PW, Huang L, Pedersen B, Shah S, Schmidt S, Haynes C, West S, Asper D, Booze M, Sharma S, Sundseth S, Middleton L, Roses AD, Hauser MA, Vance JM, Pericak-Vance MA, Kraus WE.

A genomewide scan for early-onset coronary artery disease in 438 families: the GENECARD Study.

Am J Hum Genet 2004; 75: 436-47

Coronary artery disease (CAD) risk is increased by a family history which clearly shows a genetic influence. [13] have performed a study based on patient data from several countries. DNA samples from families in which more than two siblings were affected in young years were collected in order to identify genetic factors of CAD by link analysis. They processed over 1000 samples from more than 400 families, defining three phenotypic subsets depending on CAD multiplicity on the families, absence of type 2 diabetes and occurrence of atherogenic dyslipidemia. Analysing genotypes for 395 microsatellite markers led to some findings concerning relevant chromosomes and regions. Two regions met the criteria for genome wide significance, while a region on chromosome 3q13 is linked to early-onset CAD.

Hristovski D, Peterlin B, Mitchell JA, Humphrey SM.

Using literature-based discovery to identify disease candidate genes.

Int J Med Inform 2005; 74: 289-98

Detection of disease candidate genes has become a standard technique in clinical bioinformatics applied to data from large numbers of patients showing common symptoms. [14] deals with this challenge as well as [16]. Their approach is an interactive literature-based biomedical discovery support system for discovering potentially meaningful relations from a given starting concept to other concepts extracted from MEDLINE as a reservoir of concept candidates. Reduction of start and target sets is made by integration of background knowledge about the chromosomal location of the starting disease and of the chromosomal location of the candidate genes based on knowledge extracted from LocusLink and HUGO.

Koike A, Niwa Y, Takagi T.

Automatic extraction of gene/protein biological functions from biomedical text.

Crit Care Med 2004; 32: 1306-9

Biological function is a key feature for relating gene sequence annotation with descriptions of clinical meaning. [15] try to establish the link between analytical findings performed using high throughput analysis methods and textual descriptions of biological functions. Gene, protein, or family function can be recognized using GO as an ontology for evaluating co-occurrence or collocation similarities and for application of rule-based techniques. [15] have built a tool for generating automatically functional annotations of genes and therefore open the way for transformation of biological function descriptions from unstructured text to highly formalized annotation formats. Their results will help to enrich gene sequence data by adding annotations to gene sequence databases.

Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA.

Integration of text- and data-mining using ontologies successfully selects disease gene candidates.

Nucleic Acids Research 2005; 33(5): 1544-52

Identification of disease gene candidates is difficult if multiple contributing genes cannot be detected unambiguously due to low penetrance. [16] propose an approach to overcome the problem rising from hundreds of candidates found in screening results with low specificity. Additional filtering processes have to be defined and applied in order to reduce the initial candidate set to a set of relevant candidates. The criterion applied is the gene's expression profile. Expression data are extracted using the eVOC anatomical ontology for integration of text-mining and data-mining techniques. The approach has been verified by application to a candidate gene set containing a low amount of known disease genes. The test set could be reduced to about 60% while still containing almost all of the disease genes.