REVIEWS

INTEGRATIVE GENOMICS – A BASIC AND ESSENTIAL TOOL FOR THE DEVELOPMENT OF MOLECULAR MEDICINE

JERZY OSTROWSKI

Department of Gastroenterology and Hepatology, Medical Center for Postgraduate Education and the Maria Skłodowska-Curie Memorial Cancer Center, Institute of Oncology, 5 Roentgena St., 02-781 Warsaw, Poland.

Abstract: Understanding the molecular mechanisms of disease requires the introduction of molecular diagnostics into medical practice. Current medicine employs only elements of molecular diagnostics, and usually on the scale of single genes. Medicine in the post-genomic era will utilize thousands of molecular markers associated with disease that are provided by high-throughput sequencing and functional genomic, proteomic and metabolomic studies. Such a spectrum of techniques will link clinical medicine based on molecularly oriented diagnostics with the prediction and prevention of disease. To achieve this task, large-scale and genome-wide biological and medical data must be combined with biostatistical analyses and bioinformatic modeling of biological systems. The collecting, cataloging and comparison of data from molecular studies and the subsequent development of conclusions create the fundamentals of systems biology. This highly complex analytical process reflects a new scientific paradigm called integrative genomics.

Keywords: molecular medicine, integrative genomics, high-throughput technologies, bioinformatics, system biology

Individual susceptibility to disease is based on genetic variance that determines the patient's defense and adaptive mechanisms to environmental factors, particularly the mechanisms on the molecular level. Therefore, molecular medicine should utilize genetic and molecular biomarkers indicative for increased risk of disease development. It should also employ new approaches in research on the molecular mechanisms underlying a disease, which may establish a more accurate molecular classification of diseases. This discipline may allow for optimization of the health care system by targeting the individual patient and population needs, which may increase therapeutic efficacy, decrease side effects and lower costs. "Empirical" treatment should employ the molecular classification of diseases, molecular subtypes and the individual's variations in drug metabolism, leading to the selection of the optimal drug to be administered at the right time and at the appropriate dosage. Thus, molecular medicine would personalize disease prevention, diagnosis and treatment.

Although a diseased phenotype usually results from the end-point of the molecular status of an organism, modern medicine employs diagnostics mostly based on anatomic and microscopic imaging and basic biochemical tests. These methods allow for disease diagnosis upon presentation of initial symptoms, while a disease in its early pre-symptomatic stages can be detected only by screening procedures performed in general populations. However, screenings are costly and difficult to organize on the population scale.

Prediction of disease risk is a basic expectation of preventive programs. Currently, gene-based diagnostic methods employed for the diagnosis of disease subtypes and individualized disease treatment are mostly applicable in oncology. The preventive aspect of molecular medicine should allow for the prediction of disease by measuring the probability of disease development in the future. The enrollment of healthy individuals with predicted risk of disease in appropriate medical preventive programs would increase the effectiveness of screening exams and greatly diminish healthcare costs. The ready availability of efficient and reliable genotyping methods allows for the use of molecular diagnostics and genetic counseling with monogenic diseases. Prediction of complex diseases is a much greater challenge, largely due to the limited number

* Corresponding author: phone: (48 22) 5462575; fax: (48 22) 6440209; e-mail: jostrow@warman.com.pl

of established guidelines or procedures that determine the impact of environmental factors on humans over the span of a lifetime.

Genetic information

The gene is the structural unit of genomic DNA and the functional unit of genetic information. According to the dogma of genetics, the flow of genetic information is unidirectional, whereby the genetic information encoded by genomic DNA is transcribed into RNA, which, in turn, is translated into the amino acid sequence of a protein. The human genome contains 20,000-25,000 genes that code for over 100,000 mature messenger RNAs as a result of alternative splicing of primary transcripts. Furthermore, newly synthesized proteins are post-translationally modified, resulting in several hundred thousand or even millions of functionally distinct macromolecules in the cell. The genetic information is amplified and executed by distinct macromolecules that represent its short-term storages, selectively provided at the time. The complex mechanisms that regulate gene expression utilize regulatory motifs of macromolecules, RNA and proteins, which are grouped in metabolic pathways and create heterogeneous functional modules. Thus, distinct levels of genetic networks underlie the large-scale and dynamic organization of a cell .

The cell is the basic structural and functional unit of a living organism. It is built from millions of macromolecules, various small molecules and metabolites, and the cellular "parts list" is encoded by the cellular genome. Though each cell of the body contains the same complete set of genetic information, cells in multicellular organisms are assigned to specific and distinct structures and functions. This cellular diversity results from differential gene expression that adjusts the metabolism of each cell in response to the developmental state, adaptive processes and environmental factors. An individual's phenotype is comprised of the sum of the cell-specific, developmental-stage-specific and metabolism-related changes in gene expression, modified by inter-gene interactions. Consequently, altered interactions between macromolecules and metabolites result in aberrations in cellular signaling and regulatory pathways that may lead to a diseased phenotype.

The direction of genetic information flow from genotype to phenotype indicates that diseased phenotypes may reflect the molecular alterations arising from changes in genetic information caused by mutations and/or epigenetic changes. Studying of the role of these genomic alterations in the development of diseased phenotypes should consider both disease-specific mutations and polymorphisms and sequence-related gene expression profiling in populations. The development of nanotechnologies, including new sequencing methods, microarray- and mass spectrometry-based analytical platforms, will greatly further the research in this field.

The two draft versions of the human genome have been available for six years. However, individualized genomics requires faster and less expensive DNA sequencing technology. As the possibility of a $1000 genome (which means an approximate cost at which human genome sequence should be reached) is on the horizon, in just several years, an individual's genome sequence could contribute to a wide range of medical applications, including population studies on the genetic contribution to increased risk of diseases.

Despite these promising advances, even then we will be able to predict a disease risk that is associated rather with causative mutation of a single gene than with polymorphic status of genomic DNA. Multifactoral diseases are a result of many low penetrative genetic determinants which independently have little or no effect on the patient's phenotype. These molecular changes are relatively common in the general population and, based on the discrete selection of particular determinants, can have a cumulative effect on a particular individual. Environmental factors are crucial in the development of complex diseases. Thus, researchers will require methods that can test the contributions of both genetic and environmental factors on disease risk by measuring cellular metabolism on a genomic scale.

Modern biology can run large-scale and genome-wide experiments that provide lists of cellular components. To extract the biological meaning from these results, high-throughput data should be combined with intracellular biological networks. Although the assessment of large quantities of experimental data in the functional context is a challenging task, the integration of biological data and computational analyses, typical for both bioinformatics and systems biology, is believed to provide relevant functional information.

Integrative genomics and molecular medicine

A gene is associated with a specific function, subsets of interacting genes are connected to a particular functional module, and all cellular components fit together to form a living structure. Development of most human diseases results from the dysfunction of signaling pathways in large-scale

cellular networks, and therefore, in-depth analyses of disrupted cellular components and the related aberrant signaling may be a key factor in studies on the molecular basis of disease. However, such approaches are challenging, as they should relate static pictures of gene function provided by most of the high-throughput analytical platforms into cellular networks that represent concerted and dynamic interactions of macromolecules, small molecules and metabolites. Static description of the cellular "parts list" allows for limiting conclusions on biological mechanisms. Therefore, the fundamental question is how to predict biological networks from a simple map of genes, transcripts and proteins? This certainly requires a prediction process that will employ expert information extracted from the relevant numerous public and private databases. Such biocomputational integration is a domain of bioinformatics.

Many resources used in bioinformatics, such as the EntrezGene database, the UCSC genome browser and the EBI/Ensemble browser, are publicly available online. These allow for computational analyses, annotation and visualization of high-throughput biological data. The basic local alignment search tools (BLAST) or multiple alignment methods (ClustalW, FASTA) allow for sequence comparison to infer function from gene homology. Sequence alignment can localize DNA sequence to the genome region, while gene prediction programs, such as GENESCAN and AUGUSTUS, allow for identification of gene sequence by prediction of exonic or intronic structure. PromoterScan, PromoterInspector and Dragon Promoter Finder are examples of promoter prediction algorithms.

Bioinformatics tools link experimental data to external databases. Information of genome and mRNA sequences can be extracted from publicly available global-scale databases, like those deposited at repositories of the National Center for Biotechnology Information (GeneBank, NCBI), the European Molecular Biology Laboratory (EMBL) and the DNA Database of Japan (DDBJ). Databases like GenBank and SwissProt provide descriptions about the gene protein products. The Human Gene Mutation Database (HGMD) and the Online Mendelian Inheritance in Man (OMIM) mutation database are genotype and phenotype repositories, while the International HapMap Project, the dbSNP database and the Japanese Single Nucleotide Polymorphism database maintain catalogs of single nucleotide polymorphisms among different populations. Others, like Gene Ontology and Kyoto Encyclopedia of Genes and Genomes, give information about the functional data related to an individual gene by describing different informational aspects of the gene, such as biological process, molecular function and cellular localization of gene products.

Understanding the cellular complexity in regards to the molecular status of an organism requires models built on all the levels of phenotype organization by the translation of steady-state analyses into comprehensive models of molecular interactions. Currently used methods identify and quantify individual elements rather than functional interactions and dynamics among all the elements of the highly structured systems. Systems biology represents a new discipline that focuses on the systematic study of complex interactions in biological systems to understand the biology of behaviors. "*Systems biology...is about putting together rather than taking apart, integration rather than reduction. It requires that we develop ways of thinking about integration that are as rigorous as our reductionist programmes, but different.... It means changing our philosophy, in the full sense of the term*" (Denis Noble).

Both bioinformatics and systems biology use computational analyses to integrate experimental data and the bibliographic knowledge deposited in the databases, and such an analytical process reflects a paradigm of recently established integrative genomics.

Integrative genomics allows for in-depth translational research in clinical practice, which constitutes integrated molecular medicine by developing and validating practical application of interdisciplinary collaboration between practitioners and researchers. Working together, they should establish new diagnostic molecular methods of increased specificity and sensitivity. If so, molecular medicine can be defined as health care that employs genetic and molecular profiles of the individual for the identification of disease susceptibility and its molecular subtypes; this, in turn, should introduce individualized prevention and therapy of disease. However, several barriers limit the development of molecular medicine. Of them, the lack of evidence of its clinical utility seems to be the most challenging. Therefore, before partnerships among representatives of physicians, experimental biologists and computational biologists will be evaluated, the practical use of new molecular technologies as well as their realities and limitations must be demonstrated to clinicians. Moreover, researchers need to develop less time- and labor-intensive platforms before they may be beneficial for clinicians.

Modern medicine defines symptoms as a set of disturbances that are usually related to the state of classical physiology. Molecular medicine will combine the current clinical diagnostics with the methods from experimental molecular and cellular biology to provide a novel spectrum of molecularly oriented "symptoms" addressing all aspects of molecular alterations. In consequence, the molecular nosology of disease will be far more detailed than that provided by modern medicine. Thus, while debate on the organization of integrated molecular medicine is necessary before its introduction into clinical practice, molecular medicine will and should establish fundamentals of personalized medicine that will use practical information from alterations within the flow of genetic information.

Acknowledgment

SUGGESTED READING

1. Deverka P.A., Doksum T., Carlson R.J.: Intergrating molecular medicine into the US healthcare system: opportunities, barriers, and policy challenges. Clin. Pharmacol. Ther. 82, 427 (2007).

2. Gingeras T.R.: Origin of phenotypes: Genes and transcripts. Genome Res. 17, 682 (2007). Schadt E.E., Sieberts S.K.: Moving toward a system genetics view of disease. Mamm. Genome 18, 389 (2007).

3. Hood L., Heath J.R., Phelps M.E., Lin B.: Systems biology and new technologies enable predictive and preventive medicine. Science 306, 640 (2004).

4. Kitano H.: Computational systems biology. Nature 420, 206 (2002).

5. Kitano H.: Systems biology: a brief overview. Science 295, 1662 (2002).

6. Markowetz F., Troyanskaya O.G.: Computational identification of cellular networks and pathways. Mol. BioSyst. 3, 478 (2007).

7. Massoud T.F., Gambhir S.S.: Integrating noninvasive molecular imaging into molecular medicine: an evolving paradigm. Trends Mol. Med. 13, 183 (2007).

8. Ostrowski J.: Molecular medicine of the future – applications and pitfalls Acta Pol. Pharm. 62, 329 (2006).

9. Porta M., Hernandes-Aguado I., Lumbreras B., Crous-Bou M.: "Omnic" research, monetization of intellectual property and fragmentation of knowledge: can clinical epidemiology strengthen integrative research? J. Clin. Epidemiol. 60, 1220 (2007).

10. Robinson C.V., Sali A., Baumeister W.: The molecular sociology of the cell. Nature 40, 973 (2007).

11. Tan K., Tegner J., Ravasi T.: Integrated approaches to uncovering transcription regulatory networks in mammalian cells. Genomics 91, 219 (2007).

12. Teufel A., Krupp M., Weinmann A., Galle P.R.: Current bioinformatics tools in genomic biomedical research (Review). Int. J. Mol. Med. 17, 967 (2006).

13. Woodcock J.: Molecular medicine: how, what, and when? Clin. Pharmacol. Ther. 82, 376 (2007).