Data mining based on -omics

Hiro Takashi

Slide 1

I am Hiro Takashi of Kanazawa University, where I am responsible for data mining based on -omics.

Slide 2

Data mining based on -omics is based on the field of bioinformatics, which involves the use of computers, so I will give a basic description of bioinformatics and then provide examples of use of data mining based on -omics in medicine.

Slide 3-7

I will now describe the Human Genome Project, which heavily involved bioinformatics. The Human Genome Project was a project that started by the US Department of Energy and the National Institutes of Health (NIH) in October 1990 with a budget of $3 billion. Ultimately, 24 institutions in 6 countries participated in the project; 60% were in the US, 30% were in the UK, and 6% were in Japan (the third leading participant). Japan began participating in the project in 1991. The project yielded a rough draft of 2.86 billion base pairs, corresponding to 99% of the haploid human genome, in June 2000 and a more accurate official version in April 2003. What did the project reveal? Well, it merely revealed strings of bases. Thus, the question is how did those strings of bases lead to drug discovery and identification of biological phenomena.

Slide 8-14

As I explained a moment ago, the Human Genome Project revealed that the human genome has around 6 billion base pairs. This information is DNA. Information in DNA is conveyed by RNA. Extensive RNA information is referred to as the transcriptome. There are around 30,000 genes. Genetic information is ultimately turned into proteins that perform a function. Extensive protein information is referred to as the proteome. There are 100,000 to 200,000 proteins. Over the past few years, techniques for extensively examining genomes and identifying mRNAs and proteins have been developed and a vast amount of information has been compiled. Use of this information allows, as an example, elucidation of the causes of illness based on differences in the genomes of patients and healthy individuals. Moreover, comparing cancer tissue and normal tissue at the level of the transcriptome and proteome allows an examination of how normal tissue becomes diseased. In other words, examining various differences will allow the elucidation of the mechanisms of disease.

Slide 15-18

Such information has been obtained from humans and various other organisms. The suffix -omics is added to a term such as genome, transcriptome, or proteome to indicate the study of biology at a particular molecular level, and -omics refers to the field of study that encompasses all of those levels. A large collection of -omics data results in vast amounts of bioscience data. Humans would have difficulty visually analyzing vast amounts of bioscience data. Thus, the field of bioinformatics is crucial. Using computers to analyze vast amounts of bioscience data allows the consolidation and identification of important information from -omics. To me, gleaning findings from that mountain of data is a treasure hunt with a computer.

Slide 19-20

As I have explained thus far, one method of obtaining -omics data is known as a DNA chip or a microarray. Microarrays appeared in the late 1990s. Next-generation sequencers appeared in the 2000s and have currently garnered attention.

Slide 21-25

How are -omics data obtained? As an example, clinical data or samples of peripheral are collected during a blood medical checkup. DNA can be obtained from white blood cells in the blood. A sample of diseased tissue can be obtained, and vast amounts of bioscience data can be obtained using a next-generation sequencer.

Slide 26-27

As I just explained, genomic information is converted into proteins via RNA, and those proteins perform a function. The transcriptome can easily and simply be examined using a DNA chip, and this approach is suitable for diagnostic markers and therapeutic targets. What can be analyzed with a DNA chip can also be analyzed with a next-generation sequencer. DNA chips are often used because of our increased familiarity with them.

Slide 28-30

I will now explain DNA chips (microarrays). A microarray is a densely packed array of several thousand to 10,000 genes on a glass or silicon chip. If we look closer, each black dot is an immobilized DNA fragment corresponding to an individual gene. Use of such a chip allows large quantities of genetic information to be processed and analyzed at one time.

Slide 31

I have thus far explained how -omics data are obtained. I will now explain how those data are used in medicine.

Slide 32-38

I will start by explaining the Millennium Genome Project, which I was involved in while at the National Cancer Center. The Millennium Genome Project assembled healthy individuals and patients and it obtained and analyzed their medical charts and blood samples. The information obtained has been analyzed bioinformatically to identify genes causing diseases and to identify predictors of drug efficacy. Susceptibility to illness and drug efficacy and safety are being determined and new drugs are being developed. This will ultimately lead to personalized medicine, i.e. treatment tailored to an individual, to thus improve patient QOL.

Slide 39-42

The Millennium Genome Project performed 2 types of genetic analysis. The first was genetic analysis of the germline, i.e. analysis of innate genes. Based on those findings, adverse reactions to anticancer agents will be predicted, therapies to prevent cancer will be identified, and drugs will be discovered. The second type of analysis was the genetic analysis of diseased tissue. The particular nature of an illness will be determined, leading to its diagnosis and treatment. These techniques are not limited to cancer and can be applied to various illnesses. I will start of by describing genetic analysis of the germline.

Slide 43

Genetic analysis of the germline involves analysis of polymorphisms.

Slide 44-46

As I just explained, genomic information is converted into proteins via RNA, and those proteins perform a function. If base sequences in the genome differ even slightly, the functions of proteins can change and the amount of proteins can change.

Slide 47-50

Polymorphisms are differences between individuals. When there are individuals with a different genotype in a group of the same species of organisms, polymorphisms refer to the different genes and DNA sequences. Typically, a gene is considered common if it occurs at a frequency higher than 1% in the population while a gene occurring at a frequency lower than 1% is called a mutation.

Blood types are an example of a well-known polymorphism. The ABO blood group system was devised by Landsteiner in 1901 and is the most widely used blood group system. The blood type phenotypes are AB, A, B, and O. Many Japanese are type A while many whites are type O.

Slide 51-54

Another example involves types of earwax. We have long been aware of the dry and wet varieties of earwax. The gene that causes those varieties have already been identified. Adenine is substituted for guanine in ABCC11 on chromosome. Many Japanese have the dry variety while many whites have the wet variety. Such ethnic differences in polymorphisms are known. Other known polymorphisms are related to eye color and nicotine dependence.

Slide 55

I will now talk about the anticancer agent irinotecan as an example of personalized medicine based on polymorphisms.

Slide 56-63

Irinotecan is administered as a prodrug. Irinotecan is metabolized in the liver by a protein such as CES2, resulting in its active form called SN38. SN38 is then inactivated by a protein known as UGT1A1, resulting in SN38G. SN38G is then excreted from the body. SN38 has antitumor action and it causes adverse reactions such as diarrhea and neutropenia, so the concentration of SN38 in the blood must be controlled. There are polymorphisms in the promoter of the UGT1A1 gene. One polymorphism has 6 TA repeats while another has 7 TA repeats. UGT1A1 activity in a person with 7 TA repeats is lower than that in a person with 6 TA repeats, and the concentration of SN38 in the blood is known to be 2 times higher even when irinotecan is administered in the same dose. In other words, whether a person has 7 TA repeats for not is determined beforehand to regulate the dose of irinotecan, thus allowing personalized medicine. The Millennium Genome Project performed an exhaustive genomic analysis and found that polymorphisms in KCNQ5 (a channel-related gene) are correlated with adverse reactions. This is one of the topics I worked on while at the National Cancer Center.

Slide 64

I have thus far talked about genetic analysis of the germline. I will now talk about genetic analysis of somatic cells.

Slide 65-69

Personalized medicine based on an analysis of gene expression involves a search for biomarkers. These substances indicate whether turning a gene on or off is associated with the degree of malignancy. As an example, CCND1 is known to be a useful biomarker for diagnosis. Expression of the CCND1 gene is analyzed in various patients. If the level of expression is low, the patient can be given a good prognosis; if the level of expression is high, the patient can be given a bad prognosis. The reality,

however, is not that simple. Expression is determined by numerous biomarkers, and new biomarkers need to be identified.

Slide 70-73

I will now talk about joint research conducted with the National Cancer Center. A DNA chip was used to analyze samples of esophageal cancer and information on transcriptome gene expression was obtained. A filtering method (S2N', using a modified signal-to-noise ratio) that I developed was then used to rank genes associated with the degree of malignancy of esophageal cancer. We identified the genes KRT7 and FOXA1 as a result. We closely examined these genes and found that FOXA1 is a transcription factor that controls KRT7 and that FOXA1 controls a gene known as LOXL2.

Slide 74-76

The figure on the left depicts the survival curve depending on whether KRT7 expression is present or absent. KRT7-negative patients have a good prognosis while KRT7-positive patients have a poor prognosis. When expression of LOXL2 is silenced, the percent of infiltrating cancer cells, i.e. the metastatic potential of cancer, is inhibited. In other words, patients with high levels of KRT7 expression can be identified by examining KRT7 expression, and personalized medicine targeting LOXL2 can be provided. I have thus used examples of research in personalized medicine to explain data mining based on -omics data.