Data integration with biological pathways

Summary

One of the reasons we study biology is to find cures for diseases. Some diseases are harder to study than others. Take for example cancer, a disease that has a big impact on public health, and consequently a lot of time and money is invested in researching it. But in spite of decades of research, we still do not understand enough of what goes wrong in a cancer cell, and we cannot always cure it.

Diseases such as cancer, and type II diabetes is another example, are what we might call *complex diseases*. The reason that they are so difficult to understand is that so many factors are involved. Inheritance, nutrition and lifestyle all play a role. There is not just one thing that goes wrong. A single inherited cancer gene is not enough to cause cancer, but it increases the risk. A single milkshake is not enough to cause diabetes, but a lifetime of overconsumption certainly does not help.

Complex diseases involve many genes, proteins and biochemicals. To better understand this, researchers have been trying, and succeeding, to measure anything and everything inside cells. In modern experiments, the activity of tens of thousands of genes can be measured at the same time. With all these measurements it is possible to completely characterize the state of a cell or tissue sample.

With so many measurements being done, it is very important to have good tools to analyze them. This is where bioinformatics comes in. Very broadly speaking, you could say that the goal of bioinformatics is to analyze biological data using computers. And when we measure different types of molecules, such as proteins and transcripts, we do not just have to analyze the data, we have to combine the different types in a single analysis. This is what we might call *data integration*.

In this thesis I have looked specifically at a set of analysis methods based on *biological pathways*. What is a pathway? Processes in the cell often form chains of reactions. For example, sugar molecules that are used as fuel for cells are not consumed all at once. This process takes place in several steps. First, a couple of enzymes prepare the molecule to be broken down, by adding chemical groups to energize it. When the sugar molecule is loosened up enough, another enzyme comes in and chops it in two. The two halves are then further processed by other enzymes still until there is nothing left. All these steps together form a pathway. Pathways are easiest to explain using *pathway diagrams*. Pathway diagrams are a little bit like the blue prints of cellular machinery. They clarify the roles of components and make their relations understandable.

By taking these two pieces, namely *pathway diagrams* and *data integration* together, we can formulate the main goal of this thesis, which is as follows:

Goal: develop methods to integrate multiple types of experimental data and visualize them on pathway diagrams.

PathVisio

Pathway diagrams are easy to make. In the example given above, you could draw the sugar molecules and the enzymes as circles, and the chemical reactions as lines connecting the circles.

Pathway diagrams can be drawn with pencil and paper, but it is better to draw them using a computer. An electronic pathway diagram in a computer could be automatically linked to online databases that contain information about gene sequences, protein structures, chemical reactions and so on. Digitized pathway diagrams could be converted with the press of a button to a format suitable for a PowerPoint presentation or a journal publication. A collection of pathways could be indexed and searched quickly and easily.

To deal with pathways in software, we created a tool called PathVisio. PathVisio is first and foremost a drawing program for pathway diagrams.

Pathway drawing programs are not new. One of the first programs in this area was called GenMAPP. GenMAPP has many good features, such as a user-friendly graphical interface, and the ability to automatically link pathways to large experimental datasets.

But the biggest problem with GenMAPP was that it was written in an old fashioned programming language that is not adapted to the age of the Internet. That made it hard to implement new ideas and support new experimental methods. For example, GenMAPP was heavily focused on transcript measurements, and was not very suitable for other types of data. This was a reason to start the development of PathVisio.

Pathway diagrams have been created since the early beginnings of the field of biochemistry, often using pencil and paper. But there is no general agreement on the symbols used in pathway diagrams. Some diagrams use T-bars (a short perpendicular bar at the end of a line) to indicate negative feedback, other diagrams use an arrow with a minus sign next to it, yet others use a red color. The lack of standardization discourages researchers to make really complex and detailed diagrams, because other researchers will not understand the symbology without a lot of explanation. Because we are gaining more and more detailed knowledge about pathways, there is a need for a good, agreed upon way to put that knowledge down in a diagram.

One of the standard notations that is currently being pushed as a standard is called molecular interaction maps (MIM). PathVisio was the first bioinformatics tool to support MIM notation.

A newer system is called Systems Biology Graphical Notation (SBGN). The two are not completely separate: SBGN is a newer system that has copied several features of MIM.

A community of biologists is discussing and improving both notation standards. SBGN came out after the first version of PathVisio, so it is not yet supported. (The intention is that PathVisio will also support SBGN some day).

There is not a single bioinformatics application that is suitable for all research questions, or can handle all types of data. A well known bioinformatics application is Cytoscape, which, because of its plug-in system, has attracted a large group of bioinformatics developers. Cytoscape has features that PathVisio does not and vice versa. Therefore it is important that the two programs interact. We have enabled interaction by developing a plug-in in Cytoscape that can read pathways created in PathVisio. Transferring data from PathVisio to Cytoscape is a simple matter of copy and paste. Then Cytoscape can be used to improve the pathway, for example using a literature search plug-in, a feature that is not available in PathVisio.

WikiPathways

We have a collection of pathways, which were originally created for GenMAPP, but which are also suitable for PathVisio. This collection would be most useful if it was always complete and up-to-date with the latest research developments. But keeping it up-to-date is a tremendous amount of work. New discoveries are being made all the time, and it is nearly impossible for a small group of people to keep up with all the new developments. Thus we faced the problem that our pathway collection was doomed to lag behind constantly.

In an attempt to solve this problem, we developed WikiPathways. WikiPathways is a website, inspired by the successful WikiPedia project, where any researcher can come and contribute pathway information. In this way the load of creating pathways and checking them against the scientific literature can be shared between a large group of people. We call this *community curation.*

On WikiPathways, each pathway has its own page. On such a page, you see a diagram of the pathway, and below that lists of genes, metabolites and literature references, plus links to information in other bioinformatics databases. Each pathway can be downloaded in several formats. But most importantly, just below the diagram there is a big “edit” button. After you click that button, you can edit the pathway directly on the website.

We expect that if more people join WikiPathways, the quantity and quality of the pathway collection will improve. Therefore we tried to make the website easy to use. For example, although we like standard graphical notations such as MIM (see above), we do not force people to use that out of fear that they would be put off by the extra complexity. Instead, pathways can be drawn in any style.

The pathway history page is another feature that is intended to lower the barrier for newcomers. WikiPathways remembers all old versions of each pathway. A newcomer may be afraid to edit a pathway out of fear of accidentally messing it up. But because it is always possible to go back to an old version, there is no danger of doing permanent damage. The absence of this danger removes an important psychological barrier to contributing. On the pathway history page, a visitor can see exactly how a pathway changed over time. Old and new versions of a pathway are shown side-by-side, and elements of the pathway are colored green if they have been added, red if they have been deleted and yellow if they have been modified.

WikiPathways is an experiment. The number of pathways, and also the number of users, is growing slowly but steadily, but it is too early to know if the wiki approach really gives good results in the long run.

BridgeDb

I have mentioned already several times the possibility to link pathways to online databases. To make that link work, a gene on a pathway must have an identifier. That identifier then points to a record from one of the common bioinformatics databases, such as Entrez Gene from the USA or Ensembl from the UK.

Because there are so many different databases, you can choose from many identifiers. A record in database X describing gene G may be related to a record in database Y describing the same gene. Both databases contain almost the same information, but they use very different identifiers. To be able to put data from the two together, it is necessary to translate identifiers from database X to identifiers from database Y. This problem is called the *identifier mapping problem*, and this problem occurs every time two sets of data from different origins are being integrated.

The identifier mapping problem is often very messy. There are dozens of solutions out there, but they are very disorganized. For example, some work only for certain species – tough luck if you are studying an uncommon organism. You might encounter an online tool that does identifier mapping very well, but has a maximum of one thousand identifiers.

In an attempt to organize the existing identifier mapping tools, BridgeDb was created. BridgeDb is an application programming interface (API) that connects identifier mapping services on one side, to bioinformatics tools on the other side. Thus, a bioinformatics tool developer no longer has to choose which identifier mapping service to use – by incorporating BridgeDb all of them can be used together. With BridgeDb we took existing tools and organized them better so that they could be used more effectively, and applied in more situations.

Open-source and Open access

A very important aspect that returns in all of the projects described in this thesis, PathVisio, WikiPathways and BridgeDb, are the principles of open access. Scientists never start from scratch: they always use the body of work of thousands of years of science before them. So an important academic ideal is the free sharing of information between scientists.

In this thesis, there are three types of information that are important to science: pathway diagrams, software and publications. All three types of information are important and should adhere to the basic principles of free sharing. All manuscripts are (or will be) published in open access journals. The source code of software is available under a so called open-source license. And access to the pathway data in WikiPathways is governed by creative commons licensing.

Putting it all together

The goal of this thesis is to *develop methods integrate multiple types of experimental data and visualize them on pathway diagrams*. To meet that goal, we have created PathVisio, a tool to edit pathways. We built WikiPathways, a database for community-curated pathways. And we have developed BridgeDb, a system for mapping identifiers. Now finally, all the pieces of the puzzle can be put together and applied to a scientific study.

In this particular study, we look at the effects of long-term food deprivation in mice. Mice were not fed for several days (for a mouse, three days without food is equivalent to a whole month for humans). and the state of the cells in the small intestine was characterized using two technologies: microarrays on the one hand, which measures the amount of transcripts, and 2D Gel electrophoresis on the other hand, which measures the amount of proteins. These two datasets have been published before, but now they were combined for the first time.

Because we are integrating two diverse datasets, identifier mapping was very important to complete this study. For the protein data, Uniprot identifiers were used, for the microarray data, Agilent identifiers were used. The pathways have a mixture of identifiers that is predominantly Entrez Gene. Each type of identifiers had its own problems. For example, Agilent identifiers are not available everywhere, and we had to use sequence alignment to find the right mapping. The protein identifiers had a couple of mistakes in them that could only be fixed manually. However, because of the flexibility of BridgeDb, all these problems could be solved.

So what was the result of this analysis? Overall, starvation has an effect on two processes; cell turnover and energy metabolism. Both effects can be easily explained. The gut is normally a very active site of cell growth. Cells in the inner lining of the intestine con-

stantly divide, grow and then die off. Naturally, maintaining constant growth costs a lot of energy, and this results in the reduction of the cell cycle pathway (which is necessary for the growth of new cells), and the apoptosis pathway (which is necessary for cell death). The second process that is affected is energy metabolism. Of course, the body as a whole tries to save energy as much as possible, but there are important differences between organs. The gut for example, stores fat when it is abundant, and during a period of starvation, this fat is exported from the gut to the rest of the body. The glycolysis pathway is reduced, and lypolysis and gluconeogenesis pathways are activated.

After carefully comparing transcripts and proteins, a couple of interesting things can be noticed. Transcripts are the precursors for proteins, so we can normally assume that if the amount of transcript increases, the protein also increases. In this study, the average correlation between transcripts and proteins is not as high as you might expect (Spearman rank correlation of 0.21). Although transcripts are the precursors for proteins, there are several processes in the cell that could affect the correlation between the two.

The picture is very dependent on which gene you look at. Some genes have very good correlation between the transcript and the protein, others do not. For elongation factor 2 (Eef2), the protein expression is reduced, but the transcript clearly increases. For ferritin (heavy and light chain), the transcript is reduced but the protein is increased. Another interesting case is that of triose phosphate isomerase 1 (Tpi1), an important enzyme in the glycolysis. Analysis of the data shows that a truncated version of the enzyme becomes more abundant under starvation conditions, which would be consistent with the observation that the glycolysis pathway is reduced in activity. There is not sufficient data to provide an explanation for all these phenomena, but it is clear that these proteins are regulated by something other than transcription rate under conditions of starvation.

Conclusion

The title of this thesis is “Data Integration with Biological Pathways”. Integration of data is currently one of the main problems in bioinformatics. In this thesis, integration of information occurs at several levels. Pathway diagrams can be used at one level, to integrate various bits of information, such as protein interactions, cellular locations, gene identifiers and literature references. At another level, identifier mapping services are used to integrate datasets from various sources. And finally, experimental data can be integrated with pathway diagrams to create visualizations that make the data easier to interpret.