Basics of Bioinformatics and how it supports medical research

Simen Skogly Russnes

Abstract

This paper describes some of the basic influence bioinformatics has on the medical research field. The information here is mostly based on the paper Bioinformatics and Drug Discovery[3], and is an extremely simplified extract of the data presented by its author.

1. Introduction

Bioinformatics is simply put the tools developed through computer science, applied to the field of biology. These tools are in general made for systematic and easier handling of information, which is quite useful in biology where the data is abundant. The human genome for example consists of approximately 20.000 genes, each of which having an average length of about 8000 base pairs[2]. The genes are furthermore regulated by various mechanisms which turn them on and off depending on the kind of cell that is using the genome, and changes in the environment that may or may not trigger the need for a particular gene product (one environment may be one of high lactose presence, where proteins to break it down begin to be synthesized). Furthermore, genes may have relationships with other genes, and as this complexity intensifies, information management through bioinformatic systems prove increasingly useful.

2. Examples of use of bioinformatics in medicine

A protein is one type of a product of a gene, and can perform a wide variety of functions throughout the body of an organism.

Proteins and other molecules may work together, and very often do so in cellular pathways that trigger functionality in a cascade of intermolecular interactions[1]. Many pathways are similar across related organisms, and the most basic mechanisms are shared between almost all living organisms, albeit with specific variations and distinctions. The synthesis of proteins from DNA (or from RNA in some simpler organisms) is for example almost identical across all species. This relationship is the reason why it is possible to use model organisms such as mice to do experiments that may result in treatments which can work on humans.

There are however slight differences that make each organism unique, which again increases the amount of data required to keep track of the pathways, genes, and etc. that do exist.

2.1. Protein homology modeling

Two proteins that perform very similar functions in two different organisms are called homologs. Genome and protein databases make it simpler to find homologs through algorithms matching their composition[3]. These databases have been accumulating for many years, and now include vast amounts of detailed information of the inner workings of organisms in terms of proteins and their relationships. Identifying an unknown protein is now commonly done by looking for a previously identified protein that match in sequence or other more complex matching schemes which could give hints to the function and structure of the culprit. Before this technique of homology modeling was possible, new protein identification was often a much more time consuming and complicated procedure.

2.2. Killing a pathogen by destroying one of its functions

One way to destroy a pathogen (for example a virus or a bacteria in the body) is to destroy its function in some way. As explained, organisms have complex pathways that lead to the expression of a gene, for example through synthesis of a protein that performs a given function. These pathways involve many steps, which can be simplified as the following steps: The presence of a toxic molecule which signals the need of a protein for breaking down the toxin. This is followed by DNA reading machinery being activated on the gene site, where the DNA is then read and the little functional unit, the protein, is produced. The protein then typically has to be transferred to the location where it is needed, where it binds to the toxin and breaks it down

in some way. Finally, the protein begins to fall apart and is itself broken down. This is as stated of course a very simplified way of thinking about the process, and many more detailed steps are involved, like the need for a protein to be folded into a 3-dimensional structure, rather than simply remaining as a long string of amino acids. Proteins also sometimes come together to form larger protein complexes that work together as one machine, and in general, many small molecules are involved here and there to guide the process.

With that in mind, if you simply break one of these steps, by introducing a chemical that breaks down whatever is responsible for folding the protein, or anything of the like, the pathway would seize to function, and the pathogen no longer is able to break down the toxin, which then accumulates, and the pathogen dies.

Thus, although these pathways may be similar across related species, automated information systems that guide in maintaining an overview is of great benefit when looking for ways to knock out a pathway[3]. Looking for homologs in the host organism (i.e. the infected human) is also important, to make sure that the chemical breaking the pathogen doesn't also break the person.

2.3. Diseases caused by mutations breaking cellular functionality

A disease may be the cause of an infection, but it may also be the result of a mutation that destroys cellular function. If that is the case, then bioinformatics can help identify a paralogous gene or an alternative cellular pathway [3] which may work as a treatment strategy.

2.4. Protein 3D modeling

In addition to matching with existing proteins and genes in databases, tools for manually modeling proteins are also an important part of the bioinformatics toolset, such as the application PyMOL. PyMOL is a very popular Open Source program that can be used to model proteins and other molecules. It can be an important tool when looking for a protein homolog, or when trying to design a signal molecule drug to fit into a protein receptor. In fact, many of the available databases provide files that can be loaded into PyMOL allowing the user to view and modify, and possibly try to match one protein with another, or design a molecule that would fit into the active site of a receptor protein. See figures 1, 2, and 3 in appendix A for an example of a protein (Chain A of 2HU4) modeled in PyMOL with a signal molecule (Oseltamivir) bound to its active site.

References

[1] Bruce Alberts et al. "Cell Signaling". In: *Molecular Biology of the Cell*. 6th. Garland Science, 2014. Chap. 15.

[2] Niclas Jareborg, Ewan Birney, and Richard Durbin. "Comparative Analysis of Noncoding Regions of 77 Orthologous Mouse and Human Gene Pairs". In: *Genome Research* 9.9 (1999), pp. 815–824.

[3] Xuhua Xia. "Bioinformatics and Drug Discovery". In: *Current Topics in Medicinal Chemistry* 17.15 (2017), pp. 1709–1726.

A. Images

Figure 1. Protein 2HU4ChA with active site turned away

Figure 2. Protein 2HU4ChA with active site facing the viewer

Figure 3. Protein 2HU4ChA with active site facing viewer, having colored the surface of the protein by element