1 The art and science of the drug discovery pipeline: History of drug discovery

William T. Loging

The disease condition is a standard in the lives of humans worldwide. As far back as the Roman era, early investigators pondered the reason for a wide array of disorders, such as typhoid and polio. In no place was this more evident than the bubonic plagues of the Middle Ages, which are reported as causing the deaths of more than 30% of the population in Europe (Alchon, 2003). As the observational science of biology grew, individual scientists increased their understanding of the underlying causes of human illness. In the nineteenth century, scientists like Louis Pasteur made significant contributions to the human understanding of microbiology and bacteriology. Pasteur's determination to understand human illness was born of the fact that several of his children did not survive to adulthood (Feinstein, 2008), a standard occurrence preceding the advent of twentieth-century medicine. Less than 200 years ago, prior to Pasteur's discoveries, it was coarsely thought that life spontaneously generated from inert materials (Farley and Geison, 1974); this thinking gave little value to the washing of hands and other hygienic procedures. However, additional discoveries quickly followed, such as those made by innovative physician scientists like Joseph Lister, who deduced ground-breaking procedures on aseptic treatment of patients.

The fundamentals of modern drug discovery can be found as an outline to the pioneering work of both Edward Jenner and Alexander Fleming. Although occurring more than 50 years ago, their contributions to combating human illnesses have collectively saved tens of millions of human lives; work that first started with an observation – one that caught their interest (Willis, 1997). Jenner observed that milkmaids were less susceptible to smallpox, and hypothesized that their immunity was due to contracting cowpox, an illness less virulent but related to smallpox. He successfully tested this hypothesis by inoculating human test subjects with cowpox and observed that this protected them from contracting smallpox (Barquet and Domingo, 1997). In the 1700s, it was estimated that nearly half a million people in Europe were killed by the Variola virus, the cause of smallpox (Behbehani, 1983). Due, in part, to Jenner's pioneering work on smallpox vaccination and immunity,

Bioinformatics and Computational Biology in Drug Discovery and Development, ed. W.T. Loging. Published by Cambridge University Press.

the World Health Organization declared in 1979 that smallpox was eradicated worldwide.

The path to Alexander Fleming's discovery of antibiotics began when he observed *Penicillium notatum* fungi growing in laboratory glassware (that he was about to wash after being on vacation), and noted the mechanisms of antibiotic zones. Fleming was not thinking about a "target" in the way that modern drug discovery scientists think, but he merely noticed that bacteria were not growing where some should have been. It is interesting to note that Fleming could have let this observation pass without notice, but luckily for the millions of people whose lives were saved by his discovery, he exhibited one of the main characteristics of an innovative scientist: he was inquisitive about what he observed. In fact, during his observation he was famously quoted as remarking "That's funny" (Brown, 2004). The road to commercial antibiotics, ones that patients could orally take to cure their bacterial infections, was not an easy one, but it did serve as a template for modern-day drug discovery. The creation of the drug discovery paradigm that was created in the twentieth century relied on inputs from multiple scientists as well as high-level strategic thinkers (Paul *et al.*, 2010). However, in the example of both Jenner's and Fleming's discoveries, once the impacts of their findings were noted as beneficial in alleviating the effects of human disease, it was reasonable to assume that treatments for other illnesses could be identified by applying similar innovative approaches. The idea that one could create a treatment that would cure or alleviate diseases instilled hope for other debilitating illnesses like polio and led to subsequent additional discoveries. Because of this, generations will never have to fear the words "smallpox" or "polio"; it is a fact that we too often take for granted in today's modern world.

The birth of computational biology

Biologists are often at a disadvantage when it comes to understanding the complex workings of the human body. Unlike engineering or other scientific fields, no complete blueprint exists for the complex, everyday workings of the human condition. One of the discrete advantages to our race is that we are not clones of each other; however, genetic variations often make it extremely difficult to account for a singular working model of a human being. This fact forces biologists to "learn as they go along" and requires them to rely on comparative biology for understanding the inner workings of normal and diseased states. Comparative biology is conducted by first observing a normal population of cells, tissues, or even a live organism and comparing their functions to those with disease. During the molecular biology revolution of the 1980s, scientists began generating more and more data about cellular machinery by using technics such as Southern blotting (Southern, 1975) and phage display. As these data were being generated, the standard biologist was becoming quickly inundated with higher amounts of information to sift through. The advent of gene sequencing also meant that researchers now had to measure,

store, and review long lists of CTAGs, the makings of the genetics code. A new field of science needed to be created in order to keep up with these large volumes of data. Fortunately for these researchers, the 1980s was the beginning of the computer age and created a serendipitous intersection in the two fields for which both biology and computer science came together in the form of bioinformatics and computational biology. Whether described as dry biology, *in silico* biology, conceptual biology or knowledge discovery (Weeber *et al.*, 2003), or electronic biology (eBiology; Loging *et al.*, 2007), bioinformatics and computational biology techniques have a key role in the application of electronic data in drug discovery.

The fields of bioinformatics and computational biology are, by nature, deemed to be high-throughput for a number of reasons. First, they make use of multiple well-known, large-scale and open-source data that are coupled with a wealth of established uses and literature. This lends naturally to the second reason: they are not often dependent on initial "wet-lab" investigations before they can begin. Third, protocols and workflows created by users within these categories are comprehensive enough in that they can often be repurposed to address demands from multiple teams investigating different diseases, rather than having to be applied to one distinctive field of study. This makes it possible to reuse the time and investment already created, along with the user knowledge in setting up these computational workflows across numerous drug discovery projects. Last, many of the pipelines described in this book can be rerun regularly, which provides a means for apprising new data and therefore continuous new findings. It is necessary to point out, and is often a reason why the promise of computational approaches can often be left unfulfilled, that these methodologies still eventually require input from human disease scientists. However, the features described within this book provide a broad overview to introduce the novice, as well as the experienced computationalist, to methods that can be implemented to provide a higher rate of return on investment, when compared to other, less high-throughput styles of computational biology used within drug discovery. Such practices can become a useful element to the *in silico* workbench of any drug discovery organization, as it is often noted that there are far more research and development projects to support than there are computational scientists. The subsequent chapters of this book will follow these forms of computational biology application along the drug discovery pipeline; from target identification, to small-molecule identification and optimization and ending with the evaluation of the physiological effects of a candidate drug in the clinical phases.

The drug discovery pipeline as assembly line

Biochemist Akira Endo is credited with the discovery of the statin, a class of drugs that has been shown to prevent cardiovascular disease. Investigating how one might combat high-cholesterol phenotypes in patients with familial hypercholesterolemia, following in Fleming's line of thinking, he hypothesized that fungi generate small molecules to protect themselves against other opportunistic organisms. He further

suggested that inhibiting cholesterol synthesis of the invading species could provide a selective advantage to the fungi, because it was known that fungi cell membranes contain ergosterol in place of cholesterol. Endo's research proved correct as he discovered the first statins from his studies on fungi (Endo *et al.*, 1976). The statin class went on to generate billions of dollars in revenue for the companies that market it. Endo never benefited financially from his discovery, despite the statins being among the most commonly prescribed medications worldwide (Simons, 2003). The sheer impact of Endo's work was highlighted by Nobel Prize Biochemists Michael S. Brown and Joseph Goldstein: "The millions of people whose lives will be extended through statin therapy owe it all to Akira Endo" (Landers, 2006). In 2005, sales of the statin class were estimated at $18.7 billion in the USA, of which atorvastatin was listed as one of the world's best-selling pharmaceuticals in history (Simons, 2003). I mention Dr. Endo's work for two reasons: first, to again document the tremendous impact that a single innovative scientist can have on the lives of millions of people, and second, to show that the success of Dr. Endo's approach subsequently led others to envision that such discoveries could be created *en masse*. As the statin class began generating billions in revenue, additional emphasis was placed on new discoveries, but the path to generate these life-saving drugs can be a long and arduous process; therefore, the mainstream drug discovery pipeline was relied upon to generate novel discoveries.

The term "drug discovery" itself is somewhat of a misnomer, as the vast amount of drugs that are brought to market are more often made than they are discovered. The historical paradigm of drug discovery processes, popularized from the early 1990s until today (or the writing of this book), has its foundation in the following:

Stage 1 Target Identification: Disease area scientists employ comparative biology methodologies to define a therapeutic target (which is often a protein that plays a role in the illness). The approach can range from obtaining a target idea from a publication to conducting complex protein or pathway screens within *in vitro* environments.

Stage 2 Lead Identification: The identified target is formulated into a functional high-throughput assay and screened against large chemical/biological libraries, often composed of millions of potential candidates, for identifying probable small molecule or antibody inhibitors.

Stage 3 Lead Optimization: Any "hits" from the lead identification stage are then passed through an optimization phase for assessment of the chemical and/or biologic states that make the candidate drug-like. Lead candidates are then adapted, modified, and improved upon in order to progress the drug frontrunner to the preclinical safety and efficacy testing.

Stage 4 Preclinical Assessment: If the candidate meets all the criteria to pass previous stages, it is progressed into rounds of preclinical safety and expanded efficacy assessments. This stage is often, depending on the disease and possible organism model, where the candidate is administered to lower forms of mammals, such as rodents or non-human primates.

Stage 5 Clinical Phases: Lastly, the potential drug is used in human clinical applications, starting with safety testing Phase I with a small group of human subjects. If successful, the candidate is then moved into continued efficacy testing in actual patients who are affected by the disease, in a smaller (Phase II) and subsequently broader population (Phase III). Success is measured in how safe the drug is, as well its level of efficacy, in human patients.

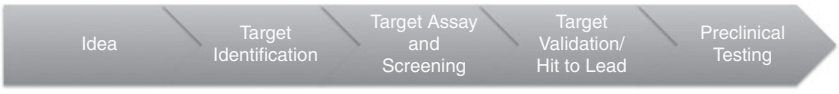

Figure 1.1 An example of the standard drug discovery pipeline.

Idea to targets

New drug targets are often ascertained from vast amounts of work conducted mainly in the field of biology. As mentioned before, these data provide investigators with a specific area of intervention that would affect the progression of a disease. This is highlighted where several research organizations have used different methods, whereas some researchers obtained target ideas by solely reading scientific publications. As stated before, when the concept of target identification became more prevalent, scientists began relying on non-publication-based methods that utilized the large amounts of data generated from methods such as SAGE (serial analysis of gene expression) and other leading techniques (Velculescu *et al.*, 1995), such as genes chips, and then to next-generation sequencing (NGS), where we currently stand today. New target work and comparative biology often go hand in hand, and no other field has benefited more from this marriage than the field of oncology (Jiao *et al.*, 2010). Standard experimental design collates normal, as well as cancerous, tissue that is measured from two separate states using several of the techniques which will be reviewed in this book. The results often lead researchers to potential proteins that may play a role in the disease; for example, oncogenes that control cell growth are often found to be overamplified in the cancerous experimental arm compared to normal, whereas tumor suppressors are often under-represented. Once researchers have a protein of ascertainable function and it has been verified through public knowledge or internally validated work, then researchers can search for molecular inhibitors (small molecules or proteins) that can inhibit the activity of this protein, therefore bringing the system back into balance with the hopes of curing disease. In the early 2000s, sequencing of genetic material was considered "high-throughput" as 30,000 base pairs could be processed in a time frame of about 3 months. In fact, when the human genome project was completed in 2001, it took 10 years to complete, as well as costing more than a billion dollars (Collins and Galas, 1993). Consider now the current speed at which genetic material is sequenced using technology like that of NGS. These applications can be applied to comparative biology studies of both genes (DNAseq) and gene transcripts (RNAseq). For example, Illumina now markets their HighSeq X Series NGS machine as having the ability to sequence the entire genome of a human being in less than one week, with cost ranges in the area of $1000 (www.illumina.com). This rapid advancement in genome sequencing technology, where speed is increased and cost is reduced, is nearly unprecedented within human history progression and has even been presented to surpass that

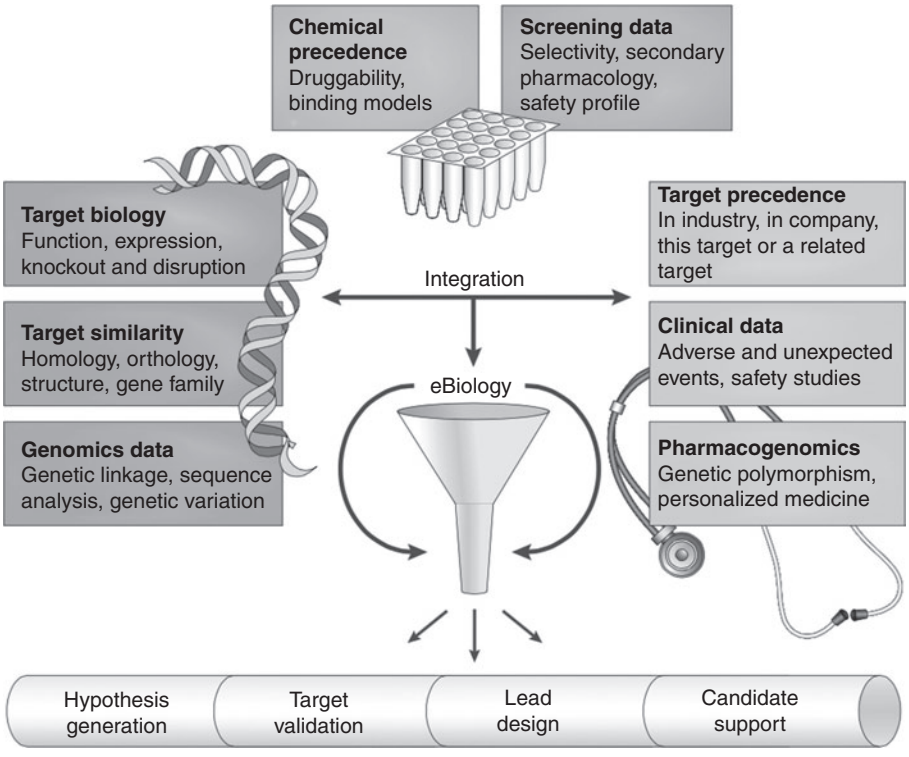

Figure 1.2 Breakdown on DDP by phase. From Loging et al. *Nature Reviews Drug Discovery* 2007.

of Moore's Law. With the large volumes of data that are generated from such high-capacity experiments, computational biology approaches will continue to be in demand for the drug discovery scientist.

Lead identification

Once a single, or multiple set, nominated target protein has been identified, the target is moved to lead identification, where suitable small-molecule (or antibody) inhibitors are identified through a biological drug design assay and the use of large chemical libraries. The target is often placed in a high-throughput set of experiments, often referred to as "screening," where millions of small molecules are interrogated for their ability to inhibit the function of the target protein in a reporter assay. Lead identification is a key decision-making stage for the research organization conducting it and can easily generate billions of data points, which require significant computational resources performed by both computational chemistry as well as computational biology experts. The druggability of the protein target via a small-molecule or antibody inhibitor is essential for the success of the discovery program. Computational biology plays a major role in the large-scale analyses of

these new protein sequences and their druggability by structural homology searching (Lui and Rost, 2002), as well as data interpretation support from computationally driven docking experiments of small molecule candidates. Molecular modeling techniques are typically applied to this area, as these approaches include nuclear magnetic resonance (Hajduk *et al.*, 2005) and x-ray crystallography (An *et al.*, 2004). Computational biology pipelines can provide immediate impact such as by reviewing more than 18,000 crystal structures from the Protein Data Bank. Again, those researchers who can link together computational pipeline aspects will be those who are most successful; for example, Cheng and colleagues (2007) created a technique for a druggability calculation that joins together multiple approaches that generate a number of molecular descriptors into a single algorithm. These methods use established data from orally available drugs, along with correlation-specific data such as ligand molecular mass along with protein-binding pocket surface area; key physicochemical properties are then linked together with structural analyses in order to obtain a scoring of druggability. Not only can this approach be applied specifically to this example, but it can also be run computationally across the large sets of crystal structure data, providing a high-throughput approach to target assessment. As mentioned earlier, a small-molecule approach may not be the only desired endpoint of a drug discovery program; antibody drug generation discovery projects also benefit from such methodologies through employing additional data such as epitope placement, cellular placement and plasma availability of proteins. Comparable to chemical approaches that often use large libraries of small molecules, the antibody engineering provides additional methods of drugging a specific protein that have led to the identification of drugs that were ultimately approved for use in humans.

Lead optimization

As a candidate drug progresses through the DDP, researchers must focus on the off-target selectivity effects, in addition to focusing on the binding of the molecule to its primary target. Naturally, a number of *in silico* approaches have been created that utilize additional data that go beyond protein homology searches. For example, investigators have utilized not only chemical data, but also genomic information to address kinase selectivity (Birault *et al.*, 2006). Further, *in silico* techniques can utilize added data such as predicted *in vivo* drug metabolites, and assist in understanding the physiological effects of known drugs. Whereas the lead optimization phase is characterized as making a drug candidate more drug-like, how this is accomplished is dependent upon several established approaches, such as the analysis of the structure–activity relationship (SAR) data. SARs are the features by which changes in the drug candidate may affect its ability to bind to and modulate the activity of its target protein. Again, intuitive scientists continue to rework the established drug discovery paradigm, in which researchers proved utility in the analysis of large SAR databases (Nettles *et al.*, 2006). One of the scientific areas that

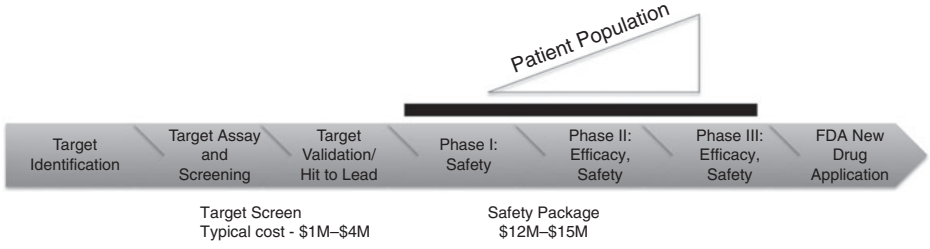

Figure 1.3 The clinical phases.

drug discovery researchers earnestly pursue is the understanding of small-molecule off-target activity.

A straightforward process would be to screen the entire human proteome with the compound in question; however, economic and technological requirements would make such an approach unfeasible. However, an *in silico* protocol that utilizes a broad panel of the druggable genome that is representative of major protein families can provide the interaction potential of the small molecule in question. This illustration panel of proteins is not limited to just the primary target, and creates a profile "fingerprint" of pharmacologic activity for a given compound. Fliri *et al.* (2005) provide a practical demonstration of how a druggable proteome cross-section can be analyzed to produce a pharmacologic fingerprint for a small molecule. These patterns can then be compared to a larger screening drug test set that numbers in the thousands. In fact, the database need not be confined to drugs; a wide array of additional molecules, such as natural products, can also be used to obtain comparative pharmacologic data. These data, termed biospectra profiles, have also been shown to be useful in functional activity prediction (agonist versus antagonist). This is useful because the small molecules brought out of high-throughput screening in lead identification do not normally provide this information (Fliri *et al.*, 2005). Biospectra can also lead researchers to understand how physiological effects are associated within secondary pharmacology of existing drugs, an approach that naturally lends itself to drug candidate disease repositioning (Campillos *et al.*, 2008). These approaches will be discussed in a relevant case study later in this book.

Preclinical and clinical phases

Perhaps the most critical stages of the drug discovery process are the progression through the preclinical and clinical phases before finally being submitted to appropriate governmental agencies for approval. As mentioned earlier, the candidate will likely enter into *in vivo* testing for the first time in this phase, being dosed to a wide array of mammals and even higher organisms, such as non-human primates. This is often where a critical hand-off also can take place, as the candidate is

passed from the research side of the organization to the development side, which is responsible for the actual creation of the candidate in large-enough qualities as required within human-based trials in a clinical setting. This critical area has recently been under the spotlight for its strategic placement within the drug discovery process and has led to the generation of a new conceptual field referred to as "translation medicine." Many researchers had already been formulating opinions on the methods required to "translate" the drug candidate from the research arm to the development and clinical stages, and the information needed to make such a transition successful. Do the safety and efficacy noted in non-human models fully extrapolate into what will happen when the drug reaches the clinic? What data are required early on in the process in order to predict success? These are several of the areas of focus for the translation medicine scientist, whose main role is to provide scientific assurances that a specific drug program will ultimately be successful in the clinic – and therefore obtain approval by the appropriate governmental agencies. Several computational biology tools and methodologies have been created to specifically assist the translational medicine scientist in dealing with the large amounts of disease-based information. These data span a wide array of genomics, as well as even EMR (electronic medical records) that allow for meaningful comparisons between those of normal and diseased patient populations. Often, a drug candidate is found not to be efficacious in the clinic – then what? Many companies park their frontrunner in a Phase II or Phase III graveyard where the candidates often languish, never to be utilized in experiments, or in the extreme, never even talked about again, within their respective company. In the later chapters of this book, we will review strategies to conduct "repositioning experiments," providing insight into other possible indications in which researchers can test their candidates. Prior to the year 2000, researchers appeared to focus more on the primary indication that brought the program forward with little thought of conducting repositioning experiments. However, repositioning projects success stories, such as Pfizer's sildenafil, proved that effectively replaced drug candidates could generate billions of dollars in revenue. In the early 2010s, additional computational approaches – as well as entire companies devoted to the subject – were created (Ashburn and Tho, 2004). Along with examples of successfully placed drug programs, it proves that often the information that may draw connections between drug target and indication can go unnoticed by seemingly expert staff and that computational approaches can provide insight at a level not previously noted within the scientific field (Loging *et al.*, 2011).

An example of when drug discovery works: *PCSK9* and the rise of genomics-era drugs

In the early 2000s, scientists began to report on patients who had a family history of low "bad" or low-density lipid (LDL) cholesterol (Levy, 2015). These serum chemistry levels of LDL did not change regardless of diet or exercise. Comparative

Low Density Lipid Receptors (LDL-R) in the liver regulate cholesterol levels in the blood. The PCSK9 protein lowers LDL-R levels and thus increases a patient's cholesterol levels. By inactivating PCSK9, LDL-R activity is increased and in turn lowers blood cholesterol levels.

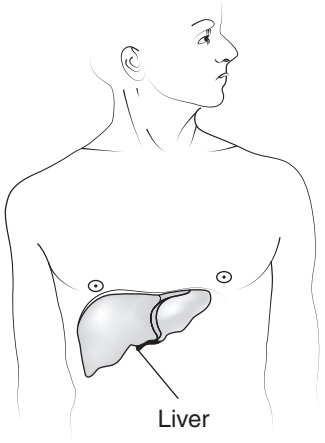

Figure 1.4 PCSK9 protein regulates cholesterol levels in humans.

biology experiments were with these individuals against other groups of patients who had familial high level of LDL. It's interesting to note that these patients with extremely low LDL also had family histories of no existence of heart disease, whereas the opposite was true for those patients with high LDL. This provided an excellent opportunity for comparing the two populations and utilizing the newly introduced approaches of whole-genome screening. It took more than 5 years, but researchers discovered that one of the major differences between the two extreme populations was a gene called *PCSK9* (Jialal and Patel, 2015). Once the gene was identified, the function of the newly associated cardiovascular gene had to be elucidated. By conducting additional computational and wet genomics studies, it was later hypothesized that *PCSK9* regulates the levels of LDL receptors in patients, a process that oversees how the liver cleanses bad cholesterol out of the blood. In these studies, the patients with low LDL contained mutations in *PCSK9* that caused a loss of function in the gene, whereas the patients with high bad cholesterol had carried *PCSK9* mutations, which amplified the function of the coding protein. It is now known that *PCSK9* acts as a protein marker that influences the recycling of the LDL receptor on the liver's surface. Due to the size of the binding interface between the LDL receptor and *PCSK9*, the vast majority of small molecules tested did not provide inhibitory utility, therefore driving researchers to employ biotherapeutic drug discovery approaches that have led to several antibody candidates, which are on track for Food and Drug Administration (FDA) approval by mid to late 2015 (Sabatine *et al.*, 2015). Preliminary data from Phase III studies of anti-PCSK9 antibodies has given clinical signals of not only reducing LDL levels but also reducing the chance of heart attack or stroke in those patients administered this drug candidate (Bloom *et al.*, 2014). The reason why I bring this point up is the very important observation that drug discovery should not be limited to the direct applications of the step-by-step pipeline, such as idea to target first, as highlighted in the example of *PCSK9*. The pipelined process was reversed slightly