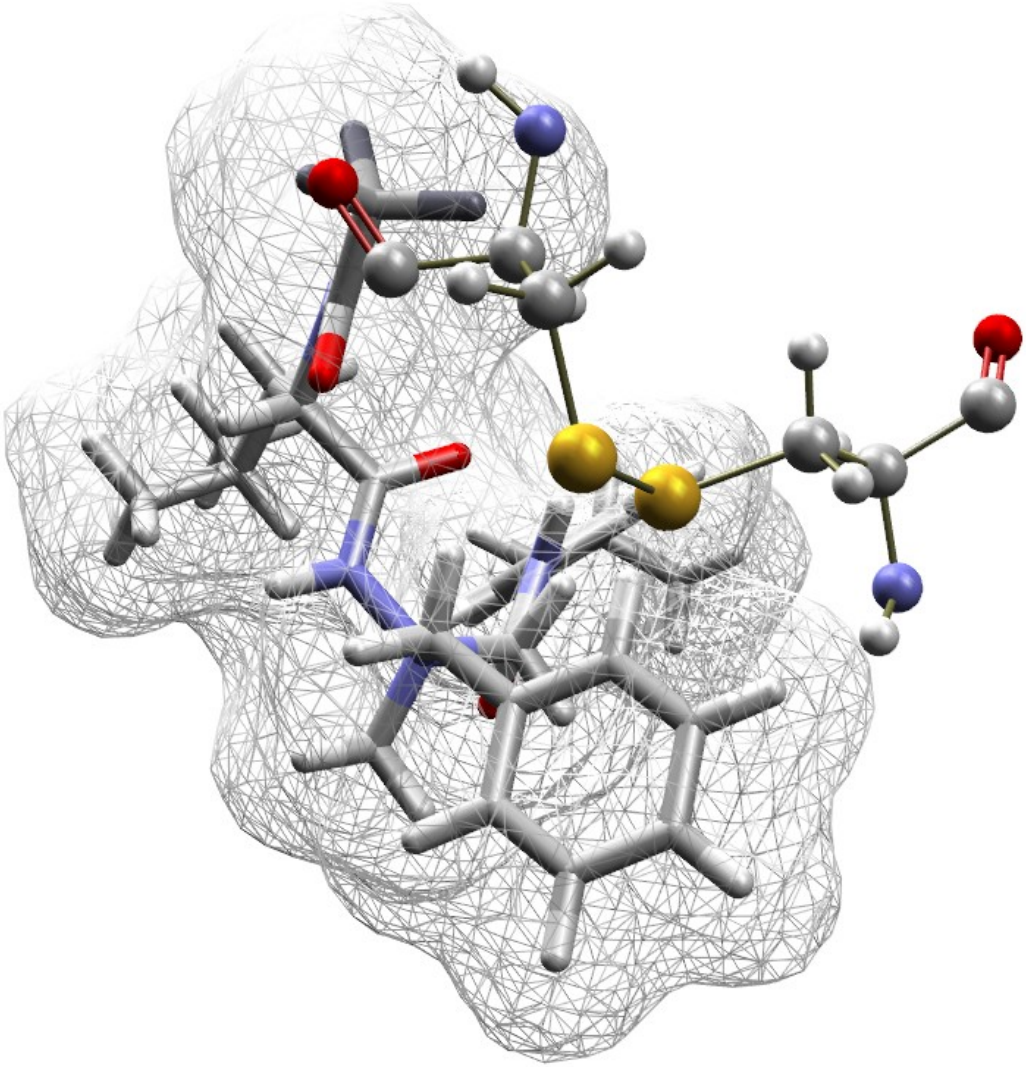
Introduction to Drug Discovery



www.molegro.com

1. Introduction to Drug Discovery

This chapter provides an overview of drug discovery and development, including:

- The workflow of drug discovery
- What technologies and methods are in use
- The number of potential customers and competitors
- What products Molegro plan for the drug discovery market

1.1 The Drug Discovery Process

Drug discovery and development is an expensive process due to the high R&D costs and extensive clinical testing. A recent study [DiMasi03] of 68 randomly selected drugs found that the average total capitalized R&D cost per drug was US$ 897 million[1]. The typical development time was 10-15 years.

The same study estimated that if a new drug was to be designed today - and approved in about 12 years - the total capitalized cost would be US$ 1.9 billion.

Two major phases are involved in creating a new drug:

- *The discovery phase*: First to identify what should be manipulated (i.e. a target protein), and second to find some suitable drug candidates that can alter the target.
- *Clinical testing*: Modern drug development requires extensive and expensive testing in order to obtain the necessary governmental approvals.

Table 3 summarizes the drug discovery and development process with a breakdown of the costs involved and the corresponding timeframe[2].

	Step	***Time (years)***	***Cost in million US$***	***Comments***
1	Identify disease			
2	Isolate protein	2-5		Identifying which proteins are responsible for the disease (e.g. by identifying genetic changes that can cause the disease)
3	***Finding the drug***	***2-5***	***400***	***Involves searching for compounds that interacts with the target protein and modifying these***
4	Pre-clinical test	1-3		Testing on animals

[1] Note that the authors use *capitalized* cost – meaning that the costs are adjusted to compensate for the number of the years they are bound in the investment. The *out-of-pocket* cost is approximately 60% of the capitalized cost. The numbers are in year 2000 US dollars.

[2] These figures are taken from [Wild03] and [Driscoll04].

	Step	*Time (years)*	*Cost in million US$*	*Comments*
5	Clinical Phase I	1-2		First human tests for safety: 10-100 healthy persons
6	Clinical Phase II	1-2		Testing for safety and effectivity: 50-500 persons from the disease group
7	Clinical Phase III	2-3	500	Large scale efficiency test: 200-2000 persons (typically several hospitals in different demographic regions)
8	FDA Approval	1-2		Obtain the necessary governmental approvals for drugs In America drugs must be approved by the US **F**ood and **D**rug **A**dministration
9	Post Approval (Phase IV)		100	Monitoring long-term effect Extending use of the drug to different classes of patients (such as children) Finding new therapeutic opportunities

Table 1. The drug discovery and development process.

Step 3 ("finding the drug") is of particular interest (for bioinformatics applications) - when the protein relevant to the disease has been identified and the search for a suitable candidate drug begins.

1.1.1 Finding the Drug

One way to find promising drug candidates is to investigate how the target protein interacts with randomly chosen compounds. This is done by using *compound libraries* which can contain more than a million synthetic and natural compounds. These libraries are then tested against the target protein – this is most often done in so called *high-throughput screening* facilities. Compound libraries are commercially available in sizes up to a million compounds. Big pharmaceutical companies also possess in-house compound libraries.

The most promising compounds obtained from the screening process are called *hits* – these are the compounds showing binding activity.

Some of these hits are then promoted to *lead compounds* – candidate structures which are further refined and modified in order to achieve more favorable interactions and less side-effects. These steps are summarized in Table 4.

	Step	*Time (years)*	*Cost Million US$*	*Comments*
1	Identifying hits	0.3 – 0.7	40	At this stage a lot of compounds are tested against the target protein. Compounds showing chemical activity against the target are promoted to *hits*. Screening a library with e.g. 1,000,000 compounds may result in 100-500 hits.
2	Hits-To-Lead			The hits are further examined in order to find some promising drug candidates – the so-called *leads*. Typically 1-3 lead compound series are found.
3	Lead Optimization	1.7 – 4.3	120	It is necessary to optimize the leads properties: i.e. toxicity, potency, binding strength, ... These modifications results in 10-500 lead variations sent for pre-clinical testing.

Table 2. The "drug finding" process in detail. Time & Cost figures are from [BCG01].

1.2 Drug Discovery Methods

Before computational drug discovery was introduced, drugs were discovered by chance in a trial-and-error manner. Not even the introduction of new technologies, such as high-throughput screening (HTS) that can experimentally test hundreds of thousands of compounds a day for activity against the target in question, have resulted in a more successful identification of promising drug candidates or reduced the R&D costs [Science2004]. Additionally, the use of HTS is very expensive (The market for HTS equipment and compounds had a 2003 global revenue of US$ 3095 million. This can be compared to the global drug finding R&D costs of US$ 6987 million) and companies need to purchase the synthesized compounds to be screened (if available at all). Moreover, in some cases HTS has failed whereas computational methods have been reported to succeed **[**Science2004**]**. Instead of actually performing the HTS experiments, the results can be computationally inferred or simulated – a process referred to as *Virtual Screening*. The computational methods used in the drug finding process are listed in Table 5 and explained in more detail in the following sections.

	Step	*Standard Method*	*Computational Alternative / Supplement*
1	Identifying hits	High Throughput Screening used to search large compound libraries.	• Molecular Docking • QSAR • Pharmacophore Mapping These methods can be used for Virtual Screening.
2	Hits-To-Lead	High Throughput Screening for validation runs. Laboratory experiments for confimation and investigation of binding properties.	• Molecular Docking • QSAR • Pharmacophore Mapping (PM) Molecular Docking can be used to investigate promising or modified ligand / target complexes individually. QSAR and PM can be used to obtain an understand why some leads bind to the target.
3	Lead Optimization	Laboratory experiments for testing the optimized leads.	

Table 3. Methods used in the "drug finding" process.

1.3 Computational Methods

Computational methods can be used to predict or simulate how a particular compound interacts with a given protein target. They can be used to assist in building hypotheses about desirable chemical properties when designing the drug and they can be used to refine and modify drug candidates.

Computational Methods can also be used to automate repetitive tasks such as searching large compound databases. *Virtual Screening* (VS) is a general term for computational methods that use computers to screen a database of virtual drug candidates (called *compounds*) to identify promising candidates (*leads*). This can be seen as an alternative to perform laboratory experiments or to perform HTS.

The main advantages of computational methods compared to laboratory (wet-lab) experiments are:

- Low costs, no compounds have to be purchased externally or synthesized by a chemist.
- It is possible to investigate compounds that have not been synthesized yet.
- Conducting HTS experiments is expensive and VS can be used to reduce the initial number of compounds before using HTS methods.
- Huge chemical search space. The number of possible virtual molecules available for VS is much higher than the number of compounds presently available for HTS.

1.3.1 Molecular Docking

When the structure of the target is known (available), usually from X-ray crystallography, the most commonly used virtual screening method is molecular docking. Molecular docking can also be used to test possible hypotheses before conducting costly laboratory experiments.

Molecular docking programs try to predict how a drug candidate binds to a protein target without performing a laboratory experiment. Molecular docking software consists of two core components:

- A *search algorithm* (sometimes called an *optimization* algorithm). The search algorithm is responsible for finding the best conformations of the ligand[3] and protein system. A conformation is the position and orientation of the ligand relative to the protein. In *flexible* docking a conformation also contains information about the internal flexible structure of the ligand – and in some cases about the internal flexible structure of the protein. Since the number of possible conformations is extremely large, it is not possible to test all of them, therefore sophisticated search techniques have to be applied. Examples of some commonly used methods are Genetic Algorithms and Monte Carlo simulations.
- An *evaluation function* (sometimes called a *score function*). This is a function providing a measure of how strongly a given ligand will interact with a particular protein. *Energy force fields* are often used as evaluation functions. These force fields calculate the energy contribution from different terms such as the known electrostatic forces between the atoms in the ligand and in the protein, forces arising from deformation of the ligand, pure electron-shell repulsion between atoms and effect from the solvent in which the interaction takes place.

It is not possible to guarantee that the search algorithm will find the same solution as the true natural process, but more efficient search algorithms will be more likely to find the true solution (if the evaluation functions properly reflect the natural processes).

Figure 2 shows a two-dimensional sketch of the docking between a protein target (black color) and a drug candidate (grey color). The cave-like structure in the target is the active site (also called the binding pocket) upon which the drug attaches.

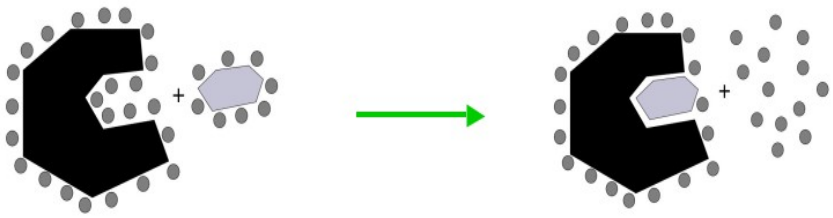


Figure 1. An example of a drug candidate (grey color) binding to a target (black color). The small filled circles represent solvent (water) molecules.

Metaphorically, the active site of the protein can be viewed as a lock, and the ligand can be thought of as a key. In this picture, molecular docking is the process of testing whether a given key fits a particular lock. This picture is slightly oversimplified due to the fact that neither the ligand nor the protein are completely rigid structures. Their shapes are somewhat flexible and may adapt to each other.

[3] A *ligand* is a small, drug-like molecule.

1.3.2 Quantitative Structure-Activity Relationships

As mentioned in the previous paragraph it is necessary to know the geometrical structure of both the ligand and the target protein in order to use molecular docking methods. QSAR (Quantitative Structure-Activity Relationships) is an example of a method which can be applied regardless of whether the structure is known or unknown.

QSAR tries to formalize what is experimentally known about how a given protein interacts with some tested compounds. As an example, it may be known from previous experiments that the protein under investigation shows signs of activity against one group of compounds, but not against another group (see Table 6).

In terms of the lock and key metaphor, we do not know what the lock looks like, but we do know which keys work, and which do not. QSAR can be considered as the method of trying to build a model for why some keys work and others do not.

In order to build a QSAR model for deciding why some compounds show sign of activity and others do not, a set of *descriptors* are chosen. These are assumed to influence whether a given compound will succeed or fail in binding to a given target. Typical descriptors are parameters such as molecular weight, molecular volume, and electrical and thermodynamical properties. When the QSAR models have been constructed (and validated) they can be used to test a number of compounds to see whether they are appropriate drug candidates for the protein under investigation (virtual screening) or not.

Compound	Descriptors				Activity against the investigated protein
	Weight	Volume	Dipole	...	
A1	0.1	124	0.2	...	**10.9**
A2	0.3	125	1.2	...	**8.5**
A3	0.5	234	2.3	...	**7.2**
N1	0.2	576	1.2	...	**0.0**
N2	0.3	123	2.1	...	**0.1**
N3	0.5	432	2.1	...	**0.2**
U1	0.2	123	3.4	...	**???**
U2	0.1	345	3.1	...	**???**
...	...	...	...	...	**???**

Table 4. A QSAR example. In the table some compounds (A1, A2, A3) have experimentally shown activity against the investigated protein. Another group (N1, N2, N3) has not shown any (or low) activity. From these experimental findings, a QSAR model can be built describing how the descriptors (weight, volume, ...) relate to the activity. This can be used to determine whether an untested compound (U1, U2, ...) would show any activity against the target protein.

QSAR models may also offer some *interpretability* – that is they may provide understandable descriptions of why some compounds show activity and others do not, e.g.: 'heavy, low volume compounds with a small dipole moment show signs of activity'.

3D QSAR is a variant of the classical QSAR method (sometimes referred to as 2D QSAR), in which the descriptors are based on the 3D structure of the compounds instead of macroproperties like weight and volume.

1.3.3 Pharmacophore Mapping

Where QSAR focused on a set of descriptors like electrostatic and thermodynamic properties, Pharmacophore Mapping is a geometrical approach. A pharmacophore can be thought of as a 3D model of characteristic features of the binding site of the investigated protein. It may describe properties like: “In region A a positive charge is needed, in region B a hydrogen donor, region C may not be occupied...” and so on. The steps involved in creating a pharmacophore are shown in Table 7. Figure 3 displays a pharmacophore model – the spheres indicate regions where a certain feature (like positive charge) is required. A pharmacophore can also be thought of as a *template*, a partial description of a molecule where certain blanks need to be filled.

	Step	***Comment***
1	Find a number of ligands known to interact with the target	While the 3D structure of the target protein may be unknown, the 3D structures of a number of compounds known to react with the target may be known.
2	Find similarities between the ligands	These ligands can be compared in order to reveal matching structural features, such as the location of a positive charge or the location of a hydrogen donor.
3	Create pharmacophore	The pharmacophore is a formalized description of the shared features found in the previous step.
4	Use pharmacophore for virtual screening	A virtual library of compounds can now be tested to see if they fit the pharmacophore model and hence are potential drug candidates.

Table 5. Steps in constructing and using a pharmacophore model.

Like QSAR models, pharmacophores can be built without knowing the structure of the target. This can be done by extracting features from compounds which are known experimentally to interact with the target in question. Afterwards, the derived pharmacophore model can be used to search compound databases (libraries) thus screening for potential drug candidates that may be of interest. Figure 3 shows an example of a pharmacophore derived from two molecules.

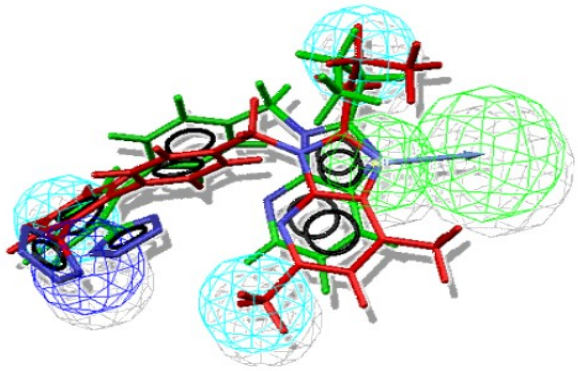


Figure 2. An example of a pharmacophore model. Spheres highlight important chemical groups that are present in both molecules (red and green sticks).

References

BCG01 *A REVOLUTION IN R&D: The Impact Of Genomics*
The Boston Consulting Group 2001.
http://www.bcg.com/publications/files/genomics.pdf

DiMasi03 The price of innovation: new estimates of drug development costs
JA DiMasi, RW Hansen, HG Grabowski, Journal of Health Economics, 2003
http://www.cptech.org/ip/health/econ/dimasi2003.pdf

Driscoll04 *A Virtual Space Odyssey.*
Cath O'Driscoll, 2004.
http://www.nature.com/horizon/chemicalspace/background/odyssey.html

Science2004 *The Many Roles of Computation In Drug Discovery*
William L. Jorgensen
Science Vol 303 19 March 2004 (p 1813)
http://saturn.med.nyu.edu/research/sb/PSB04/Week12/DrugDesign.pdf

Wild03 *New Technology in Drug Discovery. 2003*
Powerpoint slides by David J. Wild, Ph.D, University of Michigan
http://www-personal.engin.umich.edu/~wildd/mfg501/lecture.ppt