1 If It Works, It Works: Pragmatic Molecular Discovery

OVERVIEW

In the Introduction, three distinct pathways toward finding a useful molecule were introduced. To start this chapter, more detail on the nature of serendipitous and empirical discovery is provided. Then, we survey how the marriage of empirical approaches and modern technology can be highly productive, and how this overlaps and leads to rational design. And there is more to rational design itself than may at first meet the eye.

THREE WAYS OF DISCOVERY

The Three Wise (*Serendipitous*) Men and Other Stories

Scientific progress is often viewed as an orderly path of advancement based on systematic testing of reasoned hypotheses. Certainly, this process is the underpinning of the scientific method, which has had a spectacular track record in unraveling and interpreting nature's mysteries. And when a scientific paper is prepared, it is necessary for reasons of economy and clarity to present the results in a developmentally logical manner. Yet in reality, the process of discovery is frequently far from such a straightforward, linear progression. This is especially so when a major advance is achieved, where initial observations are hard to reconcile with pre-existing conceptions. An experiment designed to answer a specifically posed question may yield totally unexpected results, which direct the worker into a new and perhaps revolutionary field.

This kind of process, fueled by a sizable element of chance, has come to be termed "serendipity," based on an ancient name for Sri Lanka in a Persian fairy tale "The Three Princes of Serendip." The protagonists in this story constantly make useful but accidental discoveries, prompting the eighteenth-century English earl Horace Walpole to coin the term as a result of a serendipitous event in his own life, and his chance familiarity with the Persian fable. In the latter half of the twentieth century, serendipity and its historical importance have gained a higher profile, although this may not always be acknowledged as a significant factor in scientific research. A quick search of abstracts (article summaries) on PubMed (the free online U.S. National Library of Medicine repository of published biomedical information) with "serendipity" or "serendipitous" as keywords reveals a total of around 1120 hits, a small number indeed, considering the size of database (over 18 million published articles). Throwing in the related word "fortuitous" yields an increased total score (around 2800), although still a

Searching for Molecular Solutions: Empirical Discovery and Its Future. By Ian S. Dunn

tiny portion of the total (these specific citation figures will change with time, but the proportionate use of these terms is unlikely to vary greatly). Part of the reason for this, of course, is the way scientific papers are typically prepared, as mentioned above. In reporting a series of linked findings that constitute a research article, citation of a chance event as having a major influence on the study will often appear inelegant, almost an embarrassment. So, while it is difficult to quantify, the number of acknowledged instances where chance has significantly influenced research progress is likely to be only the tip of the iceberg. To be sure, a fortuitous observation or event can only be useful if the researcher can correctly interpret it in the first place, and then has the capacity and will to follow it up. One is reminded of the famous quote from the great French chemist and microbiologist Louis Pasteur, "In the field of observation, chance favors the prepared mind."

It is hard to imagine that serendipity will ever cease to play a role in scientific and technological advancement. But before the conscious application of the scientific method, most human knowledge (such as it was) was obtained by hit-and-miss trials where chance was a major partner. For all human history until very recent times, the use and development of any sort of natural pharmaceutical has largely been a serendipitous process. Certainly, many tribal and traditional medicines do indeed contain potent and therapeutically useful drugs, but this information can be developed further in useful ways with modern technologies. These processes can have aspects of both empirical testing and rational design, so before proceeding further we should first draw some contrasts between the meanings of serendipitous, empirical, and rationally directed discoveries. The following section revolves around this theme and introduces a whimsical fellow traveler in the field of molecular discovery.

An Empirical Fable

A human progenitor (not necessarily *Homo sapiens*) becomes ill with an intestinal parasite. We can call her Lucy if you like, and consider her at least as an honorary Australopithecine, even if her cognitive endowment is a little more advanced. She wanders off from her band, most likely to succumb to the infection. At this point, for the sake of argument, we will assume that the other members of her group are not acquainted with the very concept of herbal remedies. While staggering through the scrub, Lucy notices an unusual leafy plant not previously familiar to her. For reasons we can only surmise (perhaps related to the effect of her illness itself on her better judgment), she eats about a dozen leaves. A short while later, she undergoes violent purging, but subsequently feels markedly improved. Making the cause-and-effect connection with the new plant, Lucy returns to her band and passes on her *serendipitous* observation (speaking Proto-World?).

Soon an epidemic of stomach troubles strikes the band, and thus having a number of patients and little to worry about from ethics committees, Lucy decides to test each sufferer with a different number of leaves from her plant. She is a very gifted individual and does this systematically, and on more than one occasion. The carefully interpreted results of her study show that on average, five leaves cure the affliction with the least number of side effects. Her *empirically* determined dosage is used by the band from that point on, but Lucy is a perfectionist and still feels that matters could be improved. She has no conception of how consuming a plant could cure an intestinal ailment, nor indeed of the existence of parasites (at least those not visible to the naked eye), or that different parasites might require different therapies. But in a flash of insight, she thinks of trying a large number of different plant leaves for their abilities to treat bowel illnesses. By so performing this *empirical* screening, she may discover promising new candidate plants, but her sample size of both new plants and patients is small. If she is extremely lucky, during her screen for treatments of a specific class of illness, she may also

discover *serendipitously* some useful agents for entirely different medical problems. Having a need to make sense of the universe, she may later invent explanations for the effects of her pharmacopoeia involving magic or spirits. These ad hoc stories become accepted wisdom and are passed on into the folklore of the band—but do not change the degree of efficacy of her treatments. Life is improved, but still far from perfect.

One night Lucy has a very strange dream, involving creatures similar to herself, yet physically different in subtle ways. One of them speaks to her, "This is how we can help you. We'll take samples of your plant and identify the chemical constituent with the potent anti-parasitic activity. With its structure in hand, we'll be able to test hundreds of analogs to characterize structure/function relationships, and in the end come up with a compound with greatly enhanced activity!" Another of the creatures pipes up, "We can do better than that! We'll focus on the parasite itself, and define which organism-specific proteins represent the best targets for therapies. With the protein crystal structures in hand, we'll use virtual docking software to design low molecular weight compounds as specific inhibitors!" Of course, Lucy understands not a word of any of this, and her memory of it is as fleeting as dreams usually are. All she remembers for a time is one of the creatures saying, "*Rationally* designed compounds will solve your problem . . ."

Productively Applying Empiricism

These distinctions between alternative avenues for discovery of new therapeutic agents are important in the context of what is presented in this book. It would certainly appear at first glance that the rational approach (so inaccessible to Lucy's people but often within reach by us) is "the way to go." There is much to say about this, and most of this falls within the territory of Chapter 9. By its very nature, pure serendipity has a chance or "wild-card" quality, which suggests that it will essentially be quite unpredictable as to when or how often it will rear its pretty head during the conduct of research. Also, serendipity is an interactive process in the sense that it requires both a fortunate observation and the correct interpretation of the data by the observer. This itself can span a spectrum from an extremely rare occurrence acting as a "lucky break," which most capable observers would seize upon (our fabled Lucy is favored by the Fates and also talented), to mundane events that serendipitously set off a chain reaction in the minds of only a small gifted set of individuals. What better example of the latter could one cite than the (possibly apocryphal) falling apple that very indirectly planted the Law of Universal Gravitation into Isaac Newton's head? This requirement for good fortune in both experimental results and the receptivity of experimenters themselves naturally renders serendipitous discovery impossible in principle to foresee. But an empirical approach, as we took pains to note in Lucy's Fable, can be applied in a systematic manner.

Finding useful drugs from the environment is a very old human activity, which (as we have seen) can occur either purely serendipitously or by a directed empirical process. But the empirical approach can be harnessed in the laboratory to maximize its effectiveness. Rather than relying on what nature can provide in the environment, the essential technological innovation here is to produce an artificial large collection of variant molecular forms, commonly referred to as a "library." Again referring back to the fable of Lucy, the health problems of her band of protohumans stem from an intestinal parasite, which is therefore the central target for therapeutic intervention, whether or not empirical experimentalists are aware of the parasite's presence and effects. Identifying the parasite as the disease-causing agent is thus a very good start, but this will often require technologies not available until relatively recently, especially if the pathogen is invisible to the naked eye. (The existence of micro-organisms has been known since the time of Antonie van Leeuwenhoek, only a little over

300 years ago. This is a short span even compared with the duration of human recorded history, let alone the time since the arising of *Homo sapiens*.) If it is possible to grow the relevant parasitic organisms in the laboratory, it then becomes feasible to systematically screen as many drug candidates as possible to test which ones can kill the parasites or halt their growth. This empirical approach is a great advance over testing potential treatments directly on sick individuals, but it is still fairly cumbersome to the extent that each test involves a separate growth of the parasites treated with one candidate drug. To be sure, modern technologies have greatly streamlined this kind of process (whether searching for agents effecting parasite killing, or some quite distinct biological target) leading to "high-throughput" screening strategies.

Screening Versus Selection A short detour into semantics will be useful at this point. The words "selection" and "screening" occur throughout this volume, and although they have operational relationships, they are not at all synonymous. Some precise definitions are thereby in order, especially with respect to empirical molecular identification strategies. "Screening" involves a systematic evaluation of a (usually large) series of alternatives,* at a variety of possible levels ranging from molecules to cells to whole organisms. As such, a screen can be conducted with any type of testing mechanism, provided the assay that is used is informative toward the desired end property. Also, it is clearly important that specific positively-scored members from the available set of alternatives (a library of some kind) can be identified, isolated, and characterized. A direct selection process, on the other hand, allows a desired alternative to "pop out" from a large background without the need to plow through the evaluation of each alternative possibility. Biological selection exerted by natural processes is a fundamental aspect of natural evolution, and often defined as "differential survival." We will consider this further in the next chapter, but for the present purposes, an example taken from simple molecular cloning can help distinguish the selection/screening dichotomy.

Extrachromosomal loops of DNA with the ability to replicate are frequently found in bacteria, and these "plasmids" have been extensively used as vehicles for DNA cloning. Insertion of a foreign segment into a plasmid allows the replication of the novel DNA along with the rest of the plasmid *vector*. But how can you distinguish between the recombinant plasmid bearing the desired foreign sequence and the original vector alone, or plasmids that have recombined with some other spurious sequence? Consider if the desired extraneous DNA segment happened to encode and permit the expression of an enzyme that enables the bacterial host of the plasmid to escape killing by an antibiotic. One such enzyme is β-lactamase, which breaks down penicillins and thus allows bacteria producing it to survive in the presence of penicillin and other β-lactam antibiotics. Now, if β-lactamase itself was the target, one could laboriously screen numerous clones of bacteria for its expression by some assay that identified the appropriate enzyme activity (*in vitro* assays determining the rate of breakdown (hydrolysis) of appropriate β-lactam antibiotics). Though certainly possible, this would be rather foolish, since a vastly better approach is to use antibiotic resistance itself to "pull out" the clone of interest (Fig. 1.1). It is fairly obvious that only bacterial cells that possess the antibiotic resistance "marker" can grow in the presence of the specific antibiotic. Therefore, if the mixed population of bacterial clones is propagated along with the antibiotic, only those bearing the desired resistance gene can form colonies.

*While on the topic of semantics, it may be noted that instead of "alternatives," the word "candidates" could have been reasonably substituted in its ordinary usage. This was avoided, though, since "candidates" is often used to refer to a relatively small subset of possibilities identified through early rounds of library screening, rather than the whole library itself. A candidate molecule is thus on a molecular short list.

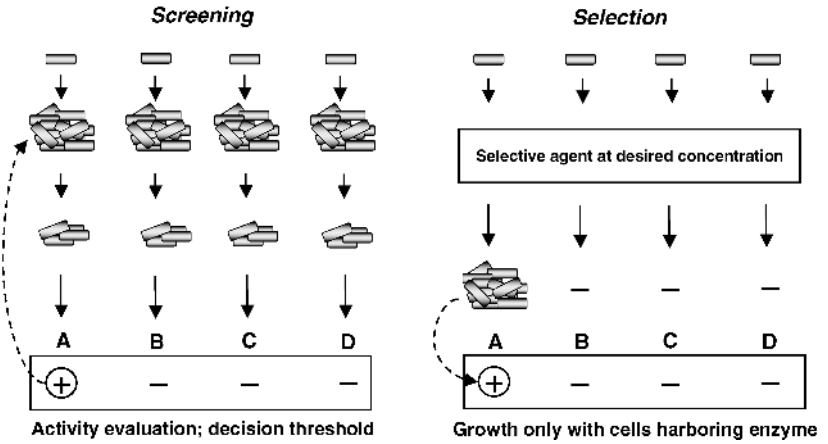

FIGURE 1.1 Screening versus selection. For screening, individual bacterial cells (represented by rods) grow into macroscopic colonies, from which samples can be taken (and repropagated if necessary) to allow evaluation of an activity. If measured activity surpasses a decided threshold, these data identify the corresponding colony bearing the desired genetic information encoding the enzyme (or other protein) of interest. If a selection process is applicable, it is exerted at the initial single cell level, and only cells expressing an appropriate enzyme or other protein (enabling growth in the presence of the selective agent) will survive and form colonies. For screening, the information from the assays allows colony identification, whereas the selection itself provides evidence that the desired gene product is present (dotted line arrows).

Of course, this is a very special case, and most cloned segments will not be so readily selected. A very common strategy is to ensue that the plasmid vector itself bears an antibiotic resistance marker, so that cells that have taken up a plasmid (whether the original parent or its derivative bearing a foreign DNA insert) can be readily selected from background of cells with no such plasmid. A foreign segment that is not directly selectable can then be identified through some screening process (often nucleic acid hybridization). In such cases, *select* for the plasmid, *screen* for the insert. This process is as equally applicable to eukaryotic cells as it is to bacteria.

These examples help to demonstrate the differences between selection and screening, but perhaps have still not quite pinpointed the essential distinction. We can make a better definition of selection as a process applicable *at the level of individual replicators* of any description, which allows specific replicators* within a population to be directly isolated, amplified, and identified as unitary entities. Let us explain this further by considering the above model of β-lactamase enzymes and antibiotics in the context of Fig. 1.1. In the screening model, we have started with a population of individual bacterial cells (individual replicators for our present purposes) and allowed them to grow into visible colonies, some of which have a plasmid encoding and expressing β-lactamase activity, and some of which do not. Assaying samples of each colony for β-lactamase levels will allow identification of the specific colony that is

*A "replicator" here is defined as a supramolecular unitary entity that carries both effector molecules and informational molecules, which enable its self-replication. We will see in Chapters 3–6 that selection is mediated at the *phenotypic* level, which is usually (but not exclusively) comprised of different molecules from the informational molecules carried by replicators. The phenotype is accordingly encoded by the latter informational molecules.

producing sufficient enzyme to exceed a prechosen threshold. In contrast, the selection model acts directly on bacterial cells at the outset, by only permitting the growth of cells expressing a high enough level of β-lactamase. Hence, in the selection model, individual replicators are "chosen" and amplified, which holds true for any biological selection process (including natural selection*).

An additional point flows from this: humans can in effect act as the "choosers" for selection, and this can potentially cause confusion between the levels where screening and selection operate. "Selective breeding" is a familiar term to most people, which conjures up images of dogs and other domestic animals, or many domesticated plants. We will touch upon this again at the beginning of Chapter 4, but let us examine the "selective" connotations of this a little more closely for the present purposes, with another metaphorical tale:

> An evolutionarily minded farmer with a herd of cows decides for obscure reasons to breed for a dark-coat color in his bovine charges. He systematically evaluates each animal and chooses the darkest subset of them for further use and correctly concludes that this undertaking is classifiable as a screening process. Then, he rather pointlessly announces to his cows, "Now, consider that in effect an environmental change has occurred, such that a pallid coat has become a definite low-fitness phenotype. *I* am in fact the relevant change in these circumstances!" He then sacrifices all cows except for his chosen dark-coat subset, and enables them to breed. Through this differential propagation, he concludes that he also has acted as the selective agent discriminating the "fittest" cows from the remainder.
>
> Selection can thus be an entirely natural process, but screening is a human activity that systematically evaluates a large number of alternatives for a desirable property. Artificial selection can be superimposed on such a screening process through its enabling of a human-based choice for differential replication. Nevertheless, the distinction between the processes is important, and here is a way to remember it:
>
> State clearly just what you mean
> (Because misuse is obscene)
> Is direct selection
> Your correct direction?
> Or is your intent to screen?

How to be a Librarian Screening can be carried out on a small scale, or taken up to "high-throughput" levels (Fig. 1.2). But another means of empirical molecular identification exists, in the form of directly selecting a candidate from a library through a specific molecular binding interaction. Before looking more closely at this, we should first think about the nature of targets and probes, in relation to molecules of interest isolated from a library. It should be noted that a "target" molecule in this case is simply the "starting point" defined molecular structure for which an appropriate functionally interactive compound is sought. Again there are important semantic issues to take note of. In this terminology, a "probe" is essentially synonymous with a target, as in a statement of the type, "the target protein was used to probe the library for molecules which would bind to it." The pharmaceutical industry routinely refers to the choice of "drug targets," and the search for new ones. Choosing consistent word usage in this area is not trivial, since loose terminology may be confusing.†

*An exception would be selection putatively operating on groups of replicators rather than individuals, as with selection at the group or species level. This often controversial topic will also be noted in Chapter 2.

†By an alternative viewpoint, library "hits" (primary active candidates) might be seen as the targets for the screening process itself, but this is not the standard meaning of the word "target" in this context.

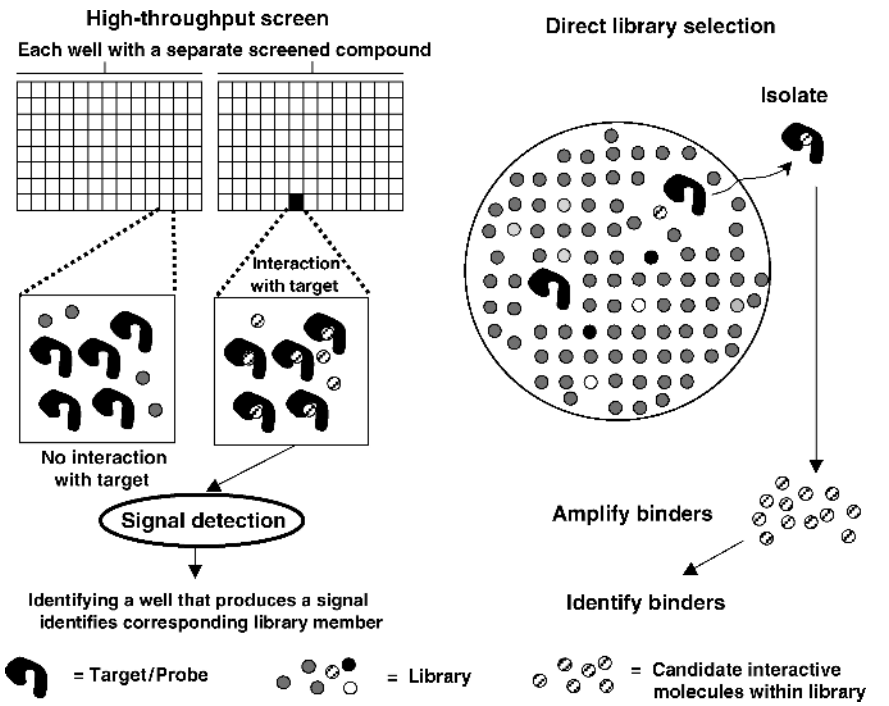

FIGURE 1.2 Depiction of empirical screening of large collections of candidate molecules (libraries) for interaction with a specific target molecule (or probe). In high-throughput screening, a large parallel series of tests are performed, physically separated in discrete wells each with the same target (which can be a single molecule or a complex system such as a mammalian cell). Each well is tested with a defined separate drug candidate from the library. Some form of positive read-out signal is necessary to recognize potential positive candidates, such that specific positive wells identify the corresponding candidate drug. In a "direct library selection," the library is treated with the target molecule. It is necessary to have the capability of separating bound candidate molecules, and then amplifying the bound fraction in order to obtain sufficient material for identification purposes.

"Target" for our purposes will therefore be defined in general as a specific molecule or system that serves as an objective, toward which one seeks another molecule that will modify the properties of the objective molecule or system in a desired manner. This usage is consistent with the original wording of Paul Ehrlich referred to in the Introduction, where a "magic bullet" (drug) is one's holy grail against disease. And bullets naturally must be aimed at a target, pathogenic microorganisms in Ehrlich's case. Successful "hits" within a library must by definition interact with the target system, and where the target is a defined molecule, the interactive library molecules may themselves be termed "ligands."

But for molecular discovery in general, both large and small molecules can be of interest as functional mediators, and either can act as targets or library members. For example, libraries of mutant variants of a single protein can be screened for gains in thermostability. The target in this case can be considered the original "wild-type" parent, variants of which are sought to exhibit improvements over the parental properties. But the library of mutants itself does not contain the original target, although obviously this parental protein is used as the yardstick by which

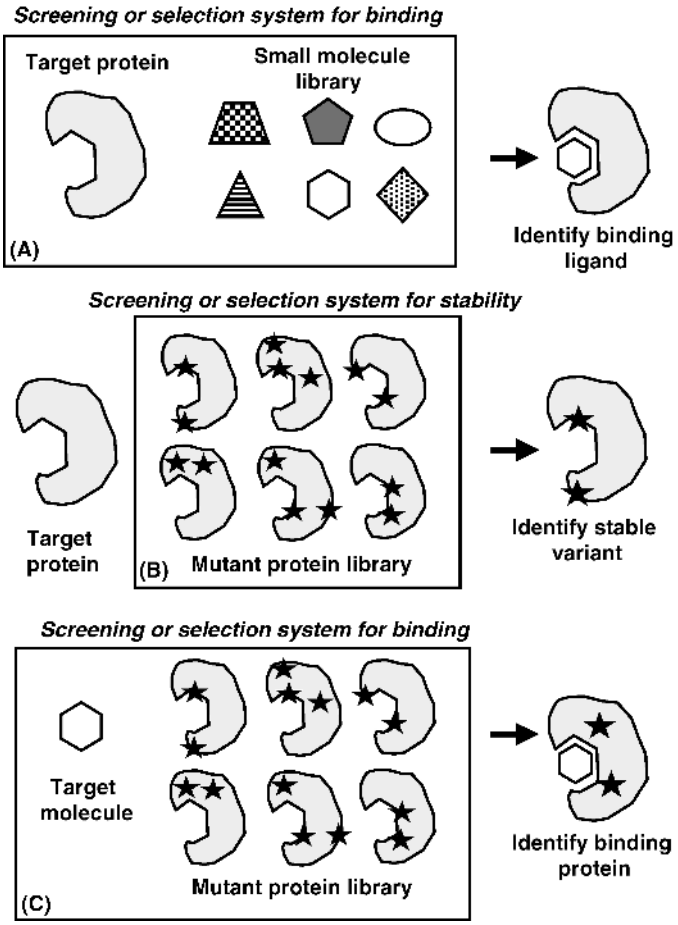

FIGURE 1.3 Invariant targets and diverse libraries, where the screening/selection systems A–C are boxed. (A) With the depicted small molecule library, the designated target is physically part of the system for identifying a binding ligand. (B) When a library of mutant variants of the target protein is screened or selected for stability, the target is not physically part of the library itself, but may be used as a reference for measurement purposes. (C) Searching for a binding protein in a library of mutants for a specific target molecule ligand. This is conceptually the same as seeking a specific antibody.

improvements are gauged. A library then is always a diversified collection of molecules, while the target is invariant. (Note, though, that invariance does not mean that the target must necessarily be a single molecular entity; whole cells or even whole simple organisms can be screened with molecular libraries.) Different types of target and screening arrangements are depicted in Fig. 1.3.

The more information available concerning the biological system that one wishes to modify, the more favorable the chances of defining the ultimate functional target molecule(s), and in turn, the better the chances for designing an optimal screening process for candidate drugs. Accordingly, if one or more specific proteins of a parasite that are essential to its functioning are known, they can be used as targets for drug development. This can in principle

be either through rational design, as in Lucy's dream, or by "applied empiricism" where such target proteins are used to screen a molecular library in the laboratory. Very broadly, this can be done either by a screening assay with maximal possible speed and processing efficiency (hence *high-throughput*) or by a selection process, subject in both cases to the nature of the library itself. In Fig. 1.2, the principles of high-throughput screening versus direct library selection are contrasted. The basic difference between the two approaches concerns the means for identifying the specific candidate binding molecules. In the case of high-throughput screening, evaluation of each member of a library is done as a separate test, which requires devising some measurable assay for a positive response, whether this is killing of a parasite or a tumor cell, changing expression of a specific gene, or a huge range of other biological responses. Thus, the separate screening reactions for each library member can be performed in minute wells of special plates, and set up and assayed collectively by robotic mechanisms. The library chemical members (of whatever nature) are added to each plate as a pre-arrayed grid, such that a positive assay signal from a specific well automatically provides the grid position and identity of the library member.* Because each library member is separately assayed, the target system can be indefinitely complex provided an unambiguous read-out assay for the desired effect can be devised.

And what of the selection-based alternative? In such a process, a molecular target is mixed with the combined collection of molecules within the library, under conditions where specific molecular forms (if represented in the library in the first place) can interact with the target/probe. The underlying premise here is that a molecule interacting with the target will bind to it with significant affinity. (The functional consequences of such binding are another matter, but it is the binding itself that enables one to "pull out" candidate library molecules of interest.) For such a library-based selection to work effectively, some other fundamental requirements must also be met. The bound complexes between the target and candidate library molecules must be purified away from all other irrelevant library members. Then, the interactive library molecules must be identified, but this is generally not possible directly. Since library collections of molecules are large, any specific molecule constitutes a tiny fraction of the total range, and the amount that binds to the target is commensurately small. So an additional *amplification* step† is needed, where the bound library molecules are increased in number until such a stage where they are amenable to characterization. The necessary isolation of complexes between target and bound library molecules, and subsequent amplification of bound molecular candidates is also depicted in Fig. 1.2. Since this complete operation applies at the level of individual replicators that are amplified as unitary entities, we are entitled to indeed refer to it as a "selection process" by the earlier definition we have arrived at. In practice, the first pass of the target through the library will often yield a set of molecules highly enriched for true target-binding candidates but not yet free from extraneous library members. A second or third pass of the target through increasingly enriched libraries may accordingly be needed before useful candidate molecules can be evaluated.

Another way of looking at both of the library screening processes of Fig. 1.2 is to see them as the implementation of search algorithms for finding members of the library of interest that fulfill preset search criteria. A flowchart of the algorithm (Fig. 1.4) refers to a sequential evaluation operation, which would mirror a laborious one-by-one screen of a set of N compounds for useful activity. Such a process is similar to the pioneering chemotherapeutic experiments of Paul

*As we will see further in Chapter 8, this is really a special case of encoding the library content (by *spatial positioning*), but other means for library encoding also exist.

†Biological nucleic acids can be amplified by replication, and proteins indirectly amplified through encoding them with replicable nucleic acids. This is detailed in Chapters 5–7 and revisited in the final chapter.

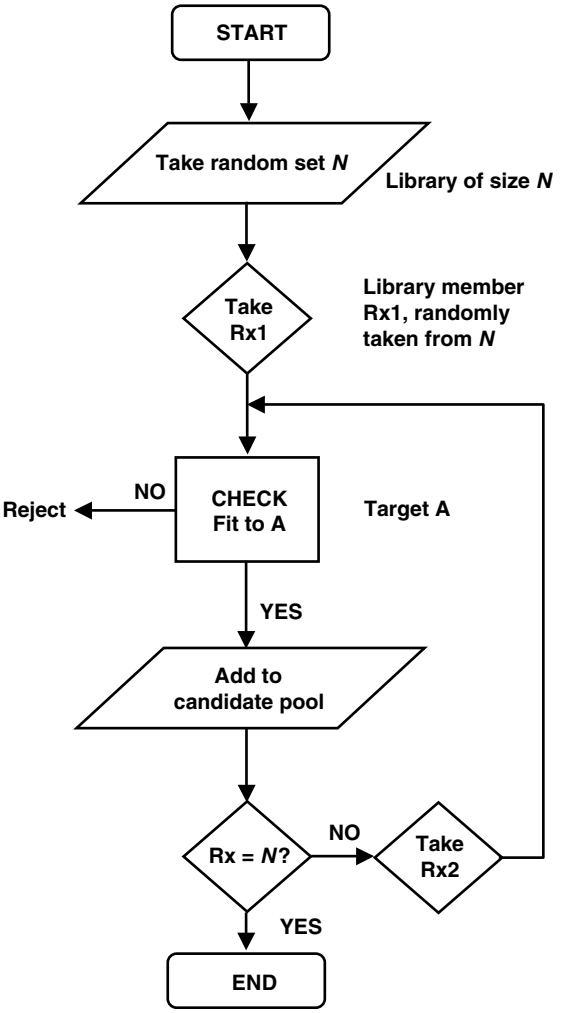

FIGURE 1.4 Empirical library screening as a search algorithm. "Check fit" refers to the process for assigning a potential "hit" (often binding of a library member to target A, but complex screening processes are possible). The algorithm is nondeterministic since the library of size *N* is taken randomly from a much larger molecular space.

Ehrlich referred to in the Introduction. Both the high-throughput screening and library selections of Fig. 1.2 side step this problem by engaging in extensive *parallel processing*. In the former case, each library member is screened separately, but as components of a very large array such that each library member is identifiable through some encoding process (spatially as in the plate grid example of Fig. 1.2). For direct selection from a library, the parallel processing is done in a single mixture, with the unbound "rejects" physically separated from those bound to the target (and thus satisfying the primary search criterion).

But this is still not the end of the story. Such a search process (by any strategy) really only constitutes a single round of evaluation, and in practice a workable solution is unlikely to

emerge directly. First, because there is always "noise" in the experimental operation, the first-round pool of candidates would need to be rescreened to confirm their correct status. Beyond this, the primary candidates typically serve as frameworks to generate secondary variant libraries based on the demonstrably useful first-pass molecules. Repeated rounds of screening or selection and identification of improved candidates (often under increasingly stringent conditions) is a way for cumulative beneficial changes to accrue, and is the essence of evolution, of which there is much more to say below and in later chapters.

Having considered these points, let us now think about libraries from a somewhat different stance . . ..

Demon-strating the Power of Empirical Screening and Selection The great nineteenth-century physicist James Clark Maxwell, famed for demonstrating the unity of electromagnetic phenomena, once imagined tiny "demons" that could sort atoms or molecules by virtue of their temperatures, and thereby reverse entropy. Although later physicists have shown the impossibility of this process even in principle, a looser version of Maxwell's demons can be used as a metaphor of sorts for a device or structure that is capable of performing some useful function on a nanoscale level. While reversing entropy is indeed a tall order, a molecule-sized demon could be proposed to perform a vast number of more modest but highly useful tasks. If a "task" is stripped down to "recognize a specific target molecule, and no other molecule, and bind to it in a specific way," then a demon becomes nothing more than a tool-key, a magic bullet, or an idealized drug molecule. But in order to help remind us of the "demonic meaning" in the context of this specific metaphor, perhaps an acronym for DEMON (Discovered Empirically, Molecules Of Note) would be useful. At the same time, another acronym (Don't Ever Molest Other Names) cautions us not to get carried away with this sort of thing.

To clarify the different approaches to empirical screening, in a brief interlude, metaphorical "demonic" models for molecular libraries can illustrate the process of extracting molecules of interest from them. First, as a metaphor for the type of chemical library involved with high-throughput screening, think of a very large number of boxes, each with the same target sitting inside. The target is the object or group of objects that you want to modify by means of a molecular interaction, and this target may be a highly complex system in its own right. Let us visualize it as a number of balls, where each ball has a hole with a different specific shape. Previously, you have shown that if a hole in a ball is filled with a closely matching "key," then the ball changes in some way, but it is very hard to predict what the overall effect will be (especially since a changed ball can in turn influence other balls within the same box). Remember that the target itself is comprised of all such balls collectively in the same box (there might be thousands of balls per box defining one specific target), and you have many, many copies of these target-containing boxes. You are looking for a way to change this target in a particular manner; perhaps to make the balls jump up and down in unison, for example. This specific change in the target is actually the *signal* that will enable you to identify an agent that produces the effect that you are seeking. So you have your target and your aim; now you want to find something that will produce the results you want, and a very large and obliging library of demons is ready to help.

Somewhere within this vast demonic collection is the right one for the job, meaning that you have to figure out how to screen the library to find it. You can picture the demons any way you like, but each demonic individual carries a tool with a unique shape, which will be tested for its ability to fit into any hole that the demon finds. It is your hope, then, that a particular demon within the library will hold a tool that will fit a *specific* target ball, which in turn will cause *all* of the balls in the target box to jump up and down. (Other demons may have tools that fit different target balls, but without eliciting any trace of the desired effect. Also, although the vast majority

of demons will be irrelevant to your needs, it may be the case that more than one specific demon can trigger the same result that you are seeking.) So you arrange a huge number of your target boxes, and put one specific demon next to each box. Every demon has a unique number as well as its unique tool, so you place the demons in a set pattern alongside the boxes. When the demons jump into their target boxes on your command, if you know which box has the right response you will then also know which demon ("it is box 300 from the left and 2000 from the top . . . that means it is demon no. 60,000"). But there is a catch. The demons are not so obliging that they will yell out and tell you when the objective is achieved. They are not too bright, actually, since all they will do for you is try out their tools for a fit in a hole (although this they do very diligently). To make matters worse, they insist on closing the lids of the boxes when they jump in. So you are stymied unless you can come up with a way of telling independently when one of the demons has been successful. You realize that you can measure the jumping of the balls by the sound they create inside the box, and the more the balls within each box jump up and down, the greater the sound.

Inspired by this, you arrange a monitor on each box that will automatically measure the sound after the demons jump in and send the information back to you. In practice, you may not need absolutely 100% of the balls to move, perhaps 80% would be satisfactory. In any case, you may find a range of sound levels, where the vast majority of demons cause no sound to issue from their respective boxes whatsoever, but a small number produce varying sound levels. You could simply pick the demons that produced the loudest results and study them further. Even better, it might be possible to set up preliminary tests where you independently make the balls jump, and measure the loudness in order to calibrate the sound signals (from the target boxes after demon entry) with the numbers of jumping balls. But in any case, you have empirically "fished out" some candidate demons for the desired effect. You may not know how changing one ball directly could affect the majority of balls, but you surmise that there may be "master balls" that respond to an exact matching "key" to their holes (provided by specific demons) with a cascade of actions that ultimately affect most or all of the other balls. This you can study further; perhaps, it will lead to other ways of rationally changing the actions of the balls within your total target.

There may be additional properties of this target (or other unrelated targets) toward which you might want to search for useful modifying demons, but you find that there are practical limitations on how many target boxes you can use in your screening of a demon library. This in turn limits how many demons you can check, and the more the demons you can screen, the greater the chances of success. A friend makes an interesting suggestion to you. "Why screen only one demon per target box? You could make pools of say 10 random demons and have 10 of them jump into a box at once. From this you could identify promising pools of demons in the usual manner, and then split these pools up into individuals for rescreening to find the one that's really active. So if you used the same maximal number of target boxes as before, you could increase the number of screened demons 10-fold! This procedure is called sib-selection and it does work in some circumstances." You investigate this further, but find a potential problem. Some demons within the library can knock out the ability of the "master balls" to respond to the very demons you are searching for (they do this indirectly by binding to other balls that in turn modulate the "master balls"). So you might miss a positive signal from some demon pools by this kind of interference effect. You realize, though, that this problem would not apply for a very simple target box, namely one with only a single ball within.

This leads us to the second metaphor for the process of direct selection from a library. In this case the target is simpler, and the aim initially is to find demons that will bind tightly to it. (Binding in itself is required as a prerequisite for any functional changes to the target, and any such alterations can be investigated after binding demons are found.) This time instead of

having the target in separate boxes, you can simply imagine an enormous number of copies of a one-ball target (with the same kind of hole that can receive a specific key) bobbing around in a vast swimming pool. Each ball is free to move, but is firmly attached to the bottom of the pool by a pegged rope. You also have a demon library again with each demon armed with a unique tool, but with important differences. These demons are not individually numbered, but have the ability to multiply to form exact copies of themselves if they receive the right stimulus. This is very important because another strange property of these demons is that by themselves (as individual demons) they are invisible, but large quantities of perfect copies of specific demons can be collectively seen and identified. So you take this type of demon library and throw it onto the pool with the one-ball targets. They cannot be seen, but the demons energetically swim around trying out their tools for matches with the single type of target holes. Almost all the demons fail to find matches, but a tiny set of them (from the multitude within the library) are successful. Now, despite the size of the pool, you have the ability to rapidly drain all the water away, and anything not tied down is also drained away. You do this, and the only things remaining are the attached targets and any demons that may have found a good and strong hole match. Snap your fingers (or whatever stimulus is needed) and the demons multiply until you can see and identify them, and you have your candidate binders of the target.

In making the library comparisons as in Fig. 1.2, it should be noted that there are certain variations on these two overall themes that will be described in later chapters. If the biggest advantage of high-throughput screening is its ability to use high-complexity targets and complex screening assays, the great benefit of direct selection from a library is the sheer size of the collection of variant molecules that can be practicably evaluated. Whether biologically active agents are screened from environmental sources or artificial libraries, and regardless of the screening process itself, one of the most fundamental issues is the size and diversity of the total pool of molecules that is available. If an appropriate molecule for a given target is not represented within a molecular collection, then it is clear that no amount of sophisticated selection or screening will produce the desired molecular solution. What is an appropriate way of visualizing the nature of molecular libraries in general, in order to gain insight into both their strengths (diversity, size) and weaknesses (constraints on diversity, exclusion of useful compounds)? One way is to arrange molecules of various classes into mathematically defined multidimensional spaces. For the present purposes a metaphorical space in which all stable molecules are found can be used to illustrate some principles involved with real collections of molecules (natural or artificial libraries) at later points.

Dreaming of Pandemonium: A Universal Molecular Space

In the above "demonic library" explanations, the well-known metaphor of Maxwell's demons was extended to include useful chemical agents because in reality specific modification of molecular systems is best done with other molecules. And finding the correct molecular tool for a task may seem demonically difficult, even if one accepts that in principle such a tool exists. So any given "demon" in this sense is a specific grouping of atoms held together by well-understood chemical bonds, with sufficient stability under normal conditions as to be practically useful. Having defined such entities, can we imagine all *possible* molecular "demons," large or small, as a single vast set? If so, it would be hard to resist calling this pandemonium, following Milton's coinage for a place of all demons. Demons and pandemonium in general have been popular themes in various contexts. For example, the human mind has been modeled as the outcome of interactions between mental "demons" corresponding to various sensory and cognitive functions,[1] whose interplay in the arena of mind constitutes a pandemonic synthesis.

So our demons, as demons will, all reside in a vast pandemonium, which may invoke infernal associations. But being able to exploit this molecular sea-of-all-demons at will would be anything but hellish, since it would be an immense force for human progress. (Although evil applications of it could also be found given malevolent intent.) Then it is justifiable to think about this in more detail, to ensure that this molecular pandemonium, if not the original term, as a useful metaphor.

In many circumstances people need to think about large sets of objects that vary over a wide range, but with specific rules for the smallest discrete changes that can be applied to change object A into another object. This second object must by definition be closely related to the original object A, since it only differs by a basic "quantum" unit of change. Around the original A object, a "cloud" or neighborhood of such closely related entities thus exist.[2] If the rules determining transition from one object to another can be applied successively, such "near-A" objects could be modified in turn, continuing *ad infinitum*. Now, if it were only possible to modify each object in one discrete way at a time, there would be a linear transition from objects A → B → C → D, and so on. But B, C, D, and onwards in turn can be part of a branching transition series, and so likewise can each new object arising from the B, C, D pathway, and so on. Clearly a one-dimensional depiction of these branching transitional series will not suffice, but even before we decide how many dimensions are required, it can be seen that the series of objects are being arrayed into a "space" based on their relationships with each other. A fundamental aspect of this depiction is that each point or "node" represents a specific unique object in the artificial space, and the same object cannot be found at any other point in this space. And then what if multiple ways of discrete modification of the objects are allowed rather than just one? New transition series vector lines would have to originate from our original starting point of object A, radiating out and continuously branching into this theoretical *N*-dimensional space from each new object node.

Chemists have given many terms to theoretical spaces where the "objects" are organic molecules, including "chemical space," "design space," "diversity space," "structure space," "topological structure space," and "chemogenomic knowledge space."[3] Considerable effort has been devoted to defining these and allied spatial constructs with mathematical precision, although they may still fall short of the precision of abstract mathematical vector spaces.[4] "Shape space" has been a useful theoretical construct for the modeling of antigen–antibody interactions.[5] These studies aim to give practical guides as to the minimal number of compounds required to cover a maximal amount of diversity, or theoretical space volume.

As an explanatory aid for this book, it will be useful to invoke a universal, all-inclusive molecular space, although this is certainly not (nor could ever be) a precise mathematical construct. It is intended rather as an instructive metaphor to illustrate some real features of large molecular collections (libraries) that enable screening or selection for specific molecules with desired properties. The tag "universal" is simply meant to distinguish this mental conception from other molecular spaces with defined restricted ranges. For example, "structure space" has been used specifically in a protein folding context[6–8] as opposed to protein or DNA sequence spaces. Let us term our particular construct "OMspace," for Overarching Molecular space. Om (or Aum) in Hindu tradition denotes the ultimate reality (which includes, but is not limited to, the observable physical universe), so a universal molecular space should fit within "Om" through the rather broad mandate of the latter.

Spatial models of varying mathematical precision have been frequently used in a number of fields, especially those related to chemical sciences. Given this widespread usage, perhaps is it justifiable to refer to this "spatial categorization" as a special "meme," or mental construct that tends to replicate itself by "infectiously" spreading from mind to mind.[9] Indeed the originator of memes has himself used "animal space" as a convenient way of depicting evolutionary

transitions in the animal kingdom.[10] All computer networks constituting the Internet are very often seen as a spatial array (cyberspace), but another information-based and potent meme has some relevance to molecular spaces. In this I refer to the tale of the Universal Library of Jorge Luis Borges[11] (the "Library of Babel"), where not only all books that have ever been published are represented, but also all *possible* books including all imperfect versions of real literature and oceans of gibberish are present as a single copy.* So the complete works of William Shakespeare are there, along with all the mistake-ridden variants completed by those tireless hypothetical monkeys locked in a room and typing away at random until the final works emerge. By relevance to OMspace, I do not mean that the ideal version of *Searching for Molecular Solutions* is contained within the Borges Library, although what you are reading surely falls far short of the perfect edition somewhere in the Library's metaphorical vastness. Of course, this could be said about any written work; perhaps even Shakespeare's plays could be improved by the judicious insertion of a word or two here and there. This precise issue, in fact, has been leveled as a criticism of the Library—how could you ever know when you had pulled out the perfect edition of anything; by what standards could you judge it? For the Library by definition contains not only all possible accurate knowledge, but also all possible falsehoods and red herrings.

We can nonetheless compare OMspace and the Borges Library if we consider them both as universal repositories of information. In the case of the former (or in a real physical molecular library), the information is revealed through the means for "interrogating" the library, as with a target/probe molecule that is used to try to identify an interactive binding molecule. (Thus, in effect, we are seeking the information specifying which library molecules will act as efficient ligands with the target/probe.) A potent message of the Library of Babel is indeed that all knowledge can be distilled into a problem of selection,[12] if one can but ask the right question, or deploy the appropriate "probe." A thought experiment could accordingly be devised where the Borges Library is searched for a technical book precisely specifying (in words) the composition and structure of the ideal molecule that interacts with a desired target. However, again there would be a huge (infinite?) number of books with suboptimal solutions to the binding problem, and indeed a vast range of books with wrong "solutions" (those specifying molecules that do not complete the desired task of high-affinity interaction with the target probe). The "right questions" to ask in this instance would remain obscure. The crucial difference between pulling a candidate molecule from OMspace and picking a structure described within the Borges library lies in the availability of an objective standard for assigning high-level function (if not perfection) for the former. In other words, in the case of OMspace, the "right question" is embodied by the target molecule used to "interrogate" the universal library.†

This process is simply the pragmatic evaluation of the performance of the candidate molecule for its ability to bind and modify the properties of the target/probe. In the real world, a molecule isolated from a physical library (a minute subsection of OMspace) would very rarely be the ideal form; much more probably it would constitute a useful *lead* for further refinement. Optimizing such lead molecules can be approached by systematic chemical modifications, or by reiteratively probing a specialized library whose members vary around a central theme based on the information obtained via the original compound. Thus, reaching into a molecular space and extracting a reasonable approximation to the desired solution is a major achievement. The take-home message is that while all knowledge may reduce to a selection process as a matter of

*Within Borges' story, the size of each book is set within a specific boundary. Although thus not infinite, the number of Library volumes is nonetheless of such a vast magnitude as to make the label "hyperastronomical" meaningless. For the purposes of this rumination, we can consider the Library contents as effectively unlimited.

†Of course, as an imaginary universal set, OMspace must also encompass the target molecule itself, and all other molecules that interact with it.

theoretical principle, it is often hard to empirically apply this dictum to real-world problems. But when seeking a functional molecule, empirical strategies, embodied by real molecular libraries, are both possible, logical, and potent.

MODES OF MOLECULAR DISCOVERY

A process that results from "blind chance" without any predictive basis is hard to rely on, and serendipity, as we have seen, falls into this category. Empirical discovery is also often thought of as a chance-based process, but the reality is far more complex than this. The major differences between serendipitous and empirical ways of finding useful molecules are shown in Table 1.1, also contrasted with rational approaches. In some cases, systematic empirical screening will almost always yield useful molecules, although by definition their exact nature is not known at the outset. The determining factors are the size of the molecular library to be screened, the desired properties of the sought-after molecule, and the nature of the screening process. An analogy (pursued in more detail in Chapter 3) may be made with natural selection, where the raw material for selection (genetic variation) may be random, but the process of cumulative selection itself is certainly not. Cumulative repetitive selection for molecular function by laboratory "directed evolution" is an analogous and parallel process where the selective pressure is determined by the experimenter. Since artificial molecular library size is

TABLE 1.1 Comparison of the Three Pathways to Molecular Discovery

	Molecular Discovery Mode		
Feature	Serendipitous	Empirical	Rational
Principle of discovery	Chance	Experimentation	Knowledge
Operational algorithms	None	Nondeterministic	Deterministic/ nondeterministic
Raw material for molecular discovery	Local environment/ unspecified	Specified by experimenter[a]	Precisely defined
Amenability of the *discovery process* to optimization and development	None	Highly optimizable	Optimization inherent in the design process
Prior knowledge of *target molecule*	Not required	Not required[b]	Required, at detailed structural and/or system level
Prior knowledge of chemical nature of *discovered molecule*	None	Limited to class of molecules screened or selected[c]	Predicted in advance
Specific structure of *discovered molecule*	Not known in advance	Not known in advance	Predicted/designed in advance of synthesis or expression[d]

[a] The experimentalist determines whether a complex system or a single-molecule target is to be screened.
[b] Not required for empirical screening *per se* (e.g., complex systems such as whole cells can be empirically screened for compounds affecting their viabilities or morphologies). However, a defined single-molecule target can also be effectively subjected to many forms of empirical screening methods.
[c] For example, if a peptide library is screened, a successful ligand found within this library is obviously a peptide.
[d] Although generally as a lead molecule requiring rounds of optimization.

a definable entity, it constitutes one factor contributing to the probability of success of empirical screening.

How then do we contrast this kind of molecular discovery with rational design? In essence, this hinges on foreknowledge of a target molecule or system and the technological ability to use this knowledge in order to make structural predictions. The key point distinguishing true rational design from an empirical strategy can thus be highlighted with an operational definition: Successful rational design can *specify a lead molecule with desired functional properties in advance of its actual physical realization, through processing of relevant structural information*, where the "physical realization" refers to synthesis or genetic expression. No matter what the power of an empirical strategy, its ability to predict the outcome of an experimental search is limited (Table 1.1). There is more to be said about this definition for rational design, which we will come to shortly.

At first glance, the categorization within Table 1.1 would suggest that molecular discovery can be cleanly rendered into a discrete tripartite arrangement. But although these discovery modes are quite distinguishable in their broad characteristics as shown, between the empirical and rational domains lies a gray area of detail, and this we should enter before moving on. This will be the opportunity to further note some of the features that distinguish these modes of identifying useful molecules.

The Borderland of Rational Design

Unlike serendipity, purely empirical discovery can be viewed as a nondeterministic search algorithm* (Fig. 1.4), since an arbitrary physical library (either from natural or from artificial sources) can be regarded as a very small random subsection of a vastly larger set of all possible molecules (ultimately OMspace). Nondeterminism in this context indicates simply that at certain decision points multiple different ways of continuing are possible, and a different answer will be delivered if the entire process is repeated. In contrast, with deterministic algorithms, the decisions and outcomes are precisely determined, such that processing of the same input data will always deliver the same output. Note that an empirical screening process performed on a defined library should produce an unequivocal result (each library member can be reproducibly evaluated as useful or not); it is the arbitrary sampling from a much larger set that results in the formal label of nondeterminism. This remains the case when the larger set itself is very far from a random sampling from the universal OMspace of molecules, as is certainly the case within the world of natural bioproducts. (Some of the factors influencing the molding of this "natural molecular space" are raised in Chapter 8.)

An empirical process can also involve rediversification and reiteration of the screening or selection rounds, in which it becomes an adaptive or evolutionary process, though still nondeterministic (Fig. 1.5). But here it is necessary to make a point that will occur repeatedly: an empirical procedure is in itself a rational pathway to molecular discovery, based on knowledge available at the time of action. In "pure" empirical sampling of natural molecular space, a completely arbitrary sample picking is made, but in the modern world there are many rational factors that channel such choices in very directed ways. An empirical approach can thus proceed logically without any pre-existing information (as the systematic screen instigated by

*To use Lucy's earlier screening project as an example, it is assumed that she randomly picked leaves from a variety of plants until she had a prechosen number. But this sample itself is taken randomly from a much larger set, so although the evaluation of the chosen sample is definitive, the outcome from the entire search would vary if she repeated the entire process, hence its nondeterministic nature.

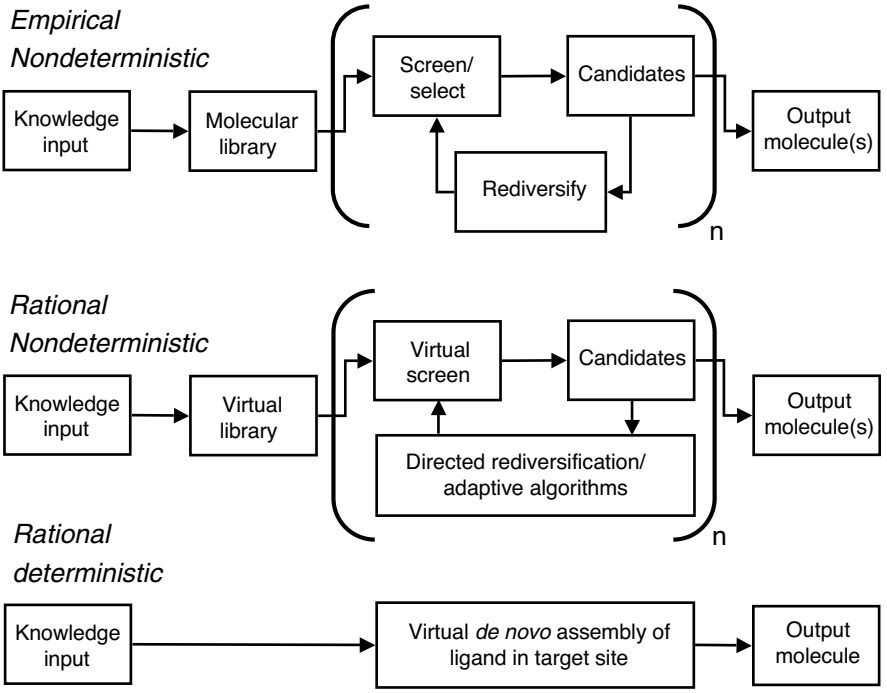

FIGURE 1.5 Schematics for empirical nondeterministic, rational nondeterministic, and rational deterministic molecular discovery processes. The *n* reiterations refer to evolutionary processes. A "pure" empirical process would have no knowledge input into the constitution of the molecular library, but the process becomes "semirational" as the pre-existing knowledge fund steadily increases. For computational rational nondeterministic processes, the rediversification itself can be undirected or knowledge based, or via an adaptive optimization algorithm. In such cases the nondeterministic status owes to alternative (non-predetermined) program steps, rather than the nature of the outcome. The idealized rational deterministic process as shown directly refers to small molecules, but can also apply in principle to *de novo* design of functional proteins or other macromolecules. This single-step representation represents an ideal for this design class; in practice, multiple assessments would be required.

Lucy above), or it can be progressively guided by increasingly refined models of the molecular problem that requires a specific molecular solution.

There are many examples of this that can be illustrative. Consider a situation where it is recognized that blocking the action of a natural hormone or mediator would be therapeutically beneficial. In this scenario, while the structure of the hormone is familiar, the mechanism by which it exerts its effects (its receptor and downstream signaling) is quite obscure. Given this starting point background information, a rational pathway is to use the structure of the known hormone to construct a series of chemical analogs (a small chemical library) one or more of which might act as an *antagonist* of the natural hormone's receptor and signaling. The assembled members of this chemical library could be screened serially or *en masse* by high-throughput methods as above: the efficiency is greatly different between such alternatives, but not the general principle. Yet the pathway followed is not as simple as indicated so far, since initial screenings may well provide continuing information that can be used to further guide the screening process. For example, a particular substituent at one site of the molecule of the above hypothetical hormone might have low but measurable activity as an antagonist. This could

in turn direct a focus upon this region, or on the specific substituent involved. Since the initial chemical analog library in such cases is far from a random collection, and candidate compounds in turn are modified for improvement in a directed fashion, the entire procedure is clearly not purely empirical, but neither is it strictly rational according to the definition given above. This gray area could be referred to as "guided empiricism," or (as most commonly seen) "semirational," perhaps an instance of glass-half-empty versus glass-half-full stances. (From another point of view, semirational might seem comparable to "semipregnant," but we will not be so pedantic.) Regardless of semantics, semirational strategies can greatly shorten the pathway toward identifying a useful drug, and in consequence there is very often constant feedback between available information and rational decisions made for molecular discovery approaches.*

When enough information is available, "true" rational design becomes a possibility. Making a prediction for molecular solution is inherent in a rational design scheme, where a target molecule's structure, combined with high-level chemical understanding, allows a specific molecule to be designed from "scratch," which will fulfill the desired functional properties. But there are levels of prediction, too, ranging from broad generalities to the precise and rock hard. Prediction levels span the type of molecular discovery involved, where a purely empirical process can only make a trivial specification based on the type of molecules screened (Table 1.1). In between, there are cases such as the above hormone scenario, where of course the prediction is that the sought-after compound will be a chemical analog of the natural active molecule. Still, an important issue here is that the latter prediction is restricted in chemical space, but still imprecise and potentially requiring evaluation of a very large number of possible analogs with chemically diverse substituents. A rigorous ideal for rational design would demand the specification of a final desired molecule in advance of its actual physical realization, but in the real world this strict definition is a little too demanding. What if the design process came up with less than 10 possibilities, one of which proved ideal after testing each in turn? It might seem unfair and indeed "irrational" to insist that this process should still be labeled empirical discovery, or even just "semirational." A better compromise might be then to consider rational design as the knowledge-based prediction of a desired molecular solution within a narrow range of alternatives, but this only shifts the focus onto a definition of "narrow," a likely can of worms in this context. (Where would one draw the line? Obviously not 10^7, but 200? 20 or less?) But a simple concept from conventional drug discovery can be brought in here—that of the "lead" compound noted earlier, which *leads* toward a structurally related final optimal objective. (There are parallels between this and Darwinian evolution, which we will come across later.) The essential concept is that in any technology, it is virtually impossible to move in a single jump to an ideal form. Progression from a prototype is the only practical way in reality. In a likewise fashion, optimal drugs for human use rarely come ready-made, and require considerable "tuning."† So, while it is unreasonable to insist that rational design should in one swoop be capable of pulling out an *ultimate* molecular solution like a rabbit out of a hat, if it is truly rational it should be able to point directly to a useful lead compound for further function-based optimization.

More and more we find an increasing impact of computational approaches on not only rational design (not surprisingly) but also the design of optimal libraries for various (otherwise empirical) screening requirements. This leads to an interesting scenario that also bears upon the

*The example of the development of the histamine H2 receptor antagonist cimetidine is a case in point here, noted in Chapter 9.

†Some drugs originating as natural products have entered the market with little of no modification from their initial status as leads, but in such cases the "optimization" has been achieved by natural evolution over very long time periods.

distinctions between empirical and rational molecular discovery. Computer-aided design can be used for candidate drug modeling, and *virtual screening* or "docking" software can evaluate and help distinguish between likely positive design contenders. But consider taking this to a high level where both a single target macromolecule (such as a protein, with a specific binding pocket) and a very large library of compounds are screened virtually for interactions of suitable energetics. In such circumstances, the spatial and structural characteristics of each library member are rapidly modeled for their binding to the target protein without any preconceptions. In effect, this hypothetical "blind" virtual screening is transferring a physical library screen with real proteins and candidate chemical ligands into an *in silico* surrogate. If the computational screen (all done without anyone venturing into a laboratory) yields a candidate subsequently proven to be the correct choice, then is this a rational design? After all, it has entirely flown from the results of human knowledge, ingenuity, and technological sophistication. Yet does it not in essence remain the same process, if it was indeed performed in this manner? Before processing the entered "virtual library," no prediction of a lead compound can be made, so is this not "electronic empiricism?" In one sense, perhaps, but we should not forget that the key factor enabling rational design in the first place is information, and accurate spatial and structural computer modeling of library and target requires a very large amount of information indeed. And most notably, this informational requirement completely distinguishes the allegedly "empirical" computational screening from its real-world counterpart, where even rudimentary knowledge of the target protein structure is not necessary in principle for fully empirical evaluation of compound libraries. Possession of the required information, and the means for processing it in order to accomplish the electronic library screen, therefore places it into the domain of rational design.

An ideal for rational molecular discovery could be viewed as a deterministic process: data are acquired, sophisticated processing by previously designed algorithms is instituted, and a set of lead molecules is generated. The processing step here could involve virtual library screening, if the library members are chosen solely based on rational criteria, and is deterministic if reiteration of each step in the program will result in the same final output. For small molecule discovery, the virtual library can be rationally designed from the characteristics of the target protein binding pocket itself, and then used for virtual screening. Real implementation of virtual library screening also involves prefiltering of compound libraries based on "drug-likeness" or other relevant criteria, which we will look in more detail in Chapter 8. In comparison with the considerations of empirical discovery at the beginning of this section, as semirational design converges toward the fully rational, the sampling of universal OMspace becomes increasingly ordered and nonrandom, until an ultimate end point of molecular definition is attained and the search algorithm becomes deterministic.

But in practice, much current use of computational evolutionary algorithms and other "adaptive" optimization strategies also qualify as rational design, since this approach requires sufficient information to enable its practical implementation. Genetic algorithms[13] and genetic programs[14] are strategies that exploit evolutionary principles for optimizing computational problem solving and have application in the area of molecular discovery.[14,15] (These and other "adaptive algorithms" are considered in a little more detail in Chapter 4.) Rational design itself can then be viewed as the implementation of either deterministic or nondeterministic algorithms (Table 1.1; Fig. 1.5). Note, though, that pursuing either class of rational algorithms may return a design solution, but this is not necessarily the same thing as finding an ideal global optimum for the problem of interest. This point we will pursue further in Chapter 9.

As noted above, rational design is enabled only if sufficient knowledge regarding the target molecule or molecular system has been accumulated, but a few more words on this are in order. "Relevant structural information" will usually refer to the target(s), but consider the following

scenario as an extension of the above hormone receptor antagonist semirational design: If a large range of ligands for a receptor were described, along with their precise biological effects, it may become possible to rationally model an antagonist even without the target structure available. In effect, the database of relationships between ligand structure and their activities provides an indirect surrogate model for the receptor binding pocket, and application of this ligand-based principle is recognized as a *de novo* ("new," or first-principles) design approach (Chapter 9). But how does this stack up as truly rational design, given that the target in this scenario has not been structurally defined? This is where we can return to the "relevant" part of the above definition. If the ligand information is detailed enough, the model for the binding pocket could approach perfection, but it cannot provide such detail for the rest of the receptor, which might be a very large and multifunctional structure. Therefore, it cannot in principle be as strong a level of rational design as when the entire receptor structure is accurately defined. While we have seen that the transition between purely empirical and rational design includes a span of semirational strategies, even when rational design is attained (by our earlier definition), it too can be seen as a range of achievements, rather than as a single edifice. In considering the "strength" of rational design, for most purposes one should not stop at the level of a single target molecule, but continue into the entire complex milieu in which the target is located, which we will pursue further in Chapter 9.

The utility of nondeterministic computational design, though designated as "rational" through its knowledge-based implementation (Fig. 1.5), nonetheless mirrors empirical design carried out in an analogous manner with real molecules and (possibly undefined) targets. This is not to suggest that empirical screening has always been carried out in such a structured and logical manner, but even Lucy's plant screening project we noted earlier could indeed fit the empirical process chart if we substituted "samples" for molecules and "confirm/refine" for "rediversify" in Fig. 1.5. It should be noted, though, that while early empirical screens could be conducted logically and systematically, they lacked the ability to reiteratively accumulate change. Consequently, though organized and systematic, such early approaches could not be termed evolutionary, which the empirical process of Fig. 1.5 essentially is. Modern molecular discovery methods routinely exploit evolutionary principles, and although using advanced technologies, the output of these pathways is not determined in advance. As a consequence, they fall within the empirical side of the spectrum of empirical–rational design.

Before we return to rational design in Chapter 9, we will examine many areas of molecular discovery that can be labeled as empirical by such criteria, even though (as we have seen) there is much overlap between these two areas. Pragmatically speaking, as suggested by this chapter's title, if a promising molecular candidate for a given task or problem is found initially through a chance-based event, then obviously it will be pursued enthusiastically regardless of its origins. But for the time being, let us focus on empirical strategies in more detail. To start with, it is hard to go past the incredible range and diversity of natural biomolecules and biosystems as inspirations for "blind" molecular discovery, as the next chapter will pursue.