15

Similarity and Diversity in Chemical Design

Chapter 15 Notation

SYMBOL	DEFINITION
Matrices	
A	Dataset matrix ($n \times m$)
A_k	Rank k approximation to A
C	covariance matrix ($m \times m$), elements $c_{jj'}$
P_k	projection matrix
U	SVD factor of A ($n \times n$), contains left singular values
V	SVD factor of A ($m \times m$), contains right singular values; also eigenvector matrix of C
V_k	low-rank approximation to eigenvector matrix ($m \times k$)
Σ	SVD factor ($n \times m$), contains singular values
Σ_k	low-rank approximation to Σ
Vectors	
u_j	left singular value
v_j	right singular value
Xi	vector of compound i (components $Xi_1, Xi_2, \cdots Xi_m$)
$\hat{X}i$	scaled version of X_i
Yi	projection of X_i; also principal component of C
Scalars & Functions	
d_{ij}	intercompound distance ij in the projected representation
f, E	target optimization functions
l_{ij}	lower bounds on intercompound distance ij
u_{ij}	upper bounds on intercompound distance ij
m	number of dataset descriptors
n	number of dataset components

T. Schlick, *Molecular Modeling and Simulation: An Interdisciplinary Guide*,
Interdisciplinary Applied Mathematics 21, DOI 10.1007/978-1-4419-6351-2_15,

Chapter 15 Notation Table (continued)

SYMBOL	DEFINITION
N	number of variables
T_d	total number of distance segments satisfying a given deviation from target
α, β	scaling factors
δ	Euclidean distance (with upper/lower bounds u, l)
λ	eigenvalues
μ	mean value
ω	weights used in target optimization function
σ	singular values

> Every sentence I utter must be understood not as an affirmation but as a question.
>
> Niels Bohr (1885–1962).

15.1 Introduction to Drug Design

Following a simple introduction to drug discovery research, this chapter presents some mathematical formulations and approaches to problems involved in chemical database analysis that might interest mathematical/physical scientists. With continued advances in structure determination, genomics, and high-throughput screening and related (more focused) techniques, *in silico* drug design is playing an important role as never before. Thus, traditional structure-directed library design methods in combination with newer approaches like fragment-based drug design [496, 1447], virtual screening [453, 1179], and system-scale approaches to drug design [236, 278, 649] will form important areas of research.

For a historical perspective of drug discovery, see [7, 159, 335, 507, 589, 727, 772], for example, and for specialized treatments in drug design modeling consult the texts by Leach [709] and Cohen [254].

15.1.1 Chemical Libraries

The field of combinatorial chemistry was recognized by *Science* in 1997 as one of nine "discoveries that transform our ideas about the natural world and also offer potential benefits to society". Indeed, the systematic assembly of chemical building blocks to form potential biologically-active compounds and their rapid testing for bioactivity has experienced a rapid growth in both experimental and theoretical approaches (e.g., [640, 692, 1241]); see the editorial overview on combinatorial chemistry [207] and the associated group of articles. Two combinatorial chemistry journals were launched in 1997, with new journals since then, and a Gordon Research conference on Combinatorial Chemistry was created. The number of new-drug candidates reaching the clinical-trial stage is greater than ever.

Indeed, it was stated in 1999: “Recent advances in solid-phase synthesis, informatics, and high-throughput screening suggest combinatorial chemistry is coming of age” [151].

Accelerated (automated and parallel) synthesis techniques combined with screening by molecular modeling and database analysis are the tools of combinatorial chemists. These tools can be applied to propose candidate molecules that resemble antibiotics, to find novel catalysts for certain reactions, to design inhibitors for the HIV protease, or to construct molecular sieves for the chemical industries based on zeolites. Thus, combinatorial technology is used to develop not only new drugs but also new materials, such as for electronic devices. Indeed, as electronic instruments become smaller, thin insulating materials for integrated circuit technology are needed. For example, the design of a new thin-film insulator at Bell Labs of Lucent Technologies [333] combined an optimal mixture of the metals zirconium (Zr), tin (Sn), and titanium (Ti) with oxygen.

As such experimental synthesis techniques are becoming cheaper and faster, huge chemical databases are becoming available for computer-aided [159] and structure-based [41, 453, 1179, 1447] drug design; the development of reliable computational tools for the study of these database compounds is thus becoming more important than ever. The term *cheminformatics* (*chemical informatics*, also called *chemoinformatics*), has been coined to describe this emerging discipline that aims at transforming such data into information, and that information into knowledge useful for faster identification and optimization of lead drugs.

15.1.2 Early Drug Development Work

Before the 1970s, proposals for new drug candidates came mostly from laboratory syntheses or extractions from Nature. A notable example of the latter is Carl Djerassi’s use of locally grown yams near his laboratory in Mexico City to synthesize cortisone; a year later, this led to his creation of the first steroid effective as a birth control pill [323]. Synthetic technology has certainly risen, but natural products have been and remain vital as pharmaceuticals (see [666, 1006] and Box 15.1 for a historical perspective).

A pioneer in the systematic development of therapeutic substances is James W. Black, who won the Nobel Prize in Physiology or Medicine in 1988 for his research on drugs beginning in 1964, including histamine H_2-receptor antagonists. Black’s team at Smith Kline & French in England synthesized and tested systematically compounds to block histamine, a natural component produced in the stomach that stimulates secretion of gastric juices. Their work led to development of a classic ‘rationally-designed’ drug in 1972 known as *Tagamet* (cimetidine). This drug effectively inhibits gastric-acid production and has revolutionized the treatment of peptic ulcers.

Later, the term *rational drug design* was introduced as our understanding of biochemical processes increased, as computer technology improved, and as the field of molecular modeling gained wider acceptance. ‘Rational drug design’ refers to the systematic study of correlations between compound composition and its bioactive properties.

Box 15.1: Natural Pharmaceuticals

Though burdened by political, environmental, and economic issues, pharmaceutical industries have long explored unusual venues for disease remedies, many in remote parts of the world and involving indigenous cures. Micro-organisms and fungi, in particular, are globally available and can be reproduced readily. For example, among the world's 25 top-selling drugs in 1997, seven were derived from natural sources. Some notable examples of products derived from Nature are listed below.

- A fungus found on a Japanese golf course is being used by Merck to make the cholesterol lowering drug mevacor, one of the 25 top-sellers of 1997.

- A fungus found on a Norwegian mountain is the basis for another 1997 top-seller, the transplant drug *Cyclosporin*, made by Novartis.

- A fungus from a Pacific yew tree is also the source of the anticancer agent paclitaxel (taxol).

- The rosy periwinkle of Madagascar is the source of Eli Lilly's two cancer drugs vincristine and vinblastine, which have helped fight testicular cancer and childhood leukemia since the 1960s.

- A microbe discovered in a Yellowstone hot spring is the source of a heat-resistant enzyme now key in DNA amplification processes.

- Ocean salmon is a source for osteoporosis drugs (*Calcimar* and *Miacalcin*), and coral extracts are used for bone replacement.

- The versatile polymer chitosan, extracted from crab and shrimp shells, is a well known fat-binding weight-loss aid, in addition to its usage in paper additives, pool cleaners, cosmetics, and hair gels.

- The Artemisiam annua plant (also known as sweet wormwood), which grows in China, Vietnam, and some parts of the United States, provides the raw material for a malaria drug, artemisinin.

- Frog-skin secretions serve as models for development of painkillers with fewer side effects than morphine. This chemical secret, long exploited by Amazon rain forest tribesmen, is now being pursued with frogs from Ecuador by Abbott Labs.

- Marine organisms from the Philippines are being investigated as sources of chemicals toxic to cancer cells.

- The venomous lizard termed Gila monster inhabiting Phoenix, Arizona, may provide a powerful peptide, exenden, for treating diabetes, because it stimulates insulin secretion and aids digestion in lizards that gorge thrice-yearly.

- A compound isolated from a flowering plant in a Malaysian rainforest, calanolide A, is a promising drug candidate for AIDS therapy, in the class of non-nucleoside reverse transcriptase inhibitors.

- A protein from a West African berry was identified by University of Wisconsin scientists as 2000 times sweeter than sugar; sweeteners are being developed from this source to make possible sweeter food products by gene insertion.

- A natural marine product (ecteinascidin 743) derived from the Caribbean sea squirt *Ecteinascidia turbinata* was found to be an active inhibitor of cell proliferation in the late 1960s, but only recently purified, synthesized, and tested in clinical trials against certain cancers.
- A Caribbean marine fungus extract (developed as halimide) shows early promise against cancer, including some breast cancers resistant to other drugs.

One of the most challenging aspects of using natural products as pharmaceutical agents is a sourcing problem, namely extracting and purifying adequate supplies of the target chemicals. For example, biochemical variations within specifies combined with international laws restricting collection (e.g., of frogs from Ecuador whose skins contain an alkaloid compound with powerful painkilling effects) limit available natural sources. In the case of the frog skin chemical, this sourcing problem prompted the synthetic design of a new type of analgesic that is potentially nonaddictive [1006].

15.1.3 Molecular Modeling in Rational Drug Design

Since the 1980s, further improvements in modeling methodology, computer technology, as well as X-ray crystallography and NMR spectroscopy for biomolecules, have increased the participation of molecular modeling in this lucrative field. Molecular modeling is playing a more significant role in drug development [453, 496, 666, 772, 1179, 1301, 1376] as more disease targets are being identified and solved at atomic resolution (e.g., HIV-1 protease, HIV integrase, adenovirus receptor, protein kinases), as our understanding of the molecular and cellular aspects of disease is enhanced (e.g., regarding pain signaling mechanisms, or the immune invasion mechanism of the HIV virus), and as viral genomes are sequenced [529]. Indeed, in analogy to genomics and proteomics — which broadly define the enterprises of identifying and classifying the genes and the proteins in the genome — the discipline of *chemogenomics* [198] has been associated with the delineation of drugs for all possible drug targets.

As described in the first chapter, examples of drugs made famous by molecular modeling include HIV-protease inhibitors (AIDS treatments), SARS virus inhibitor, thrombin inhibitors (for blood coagulation and clotting diseases), neuropeptide inhibitors (for blocking the pain signals resulting from migraines), PDE-5 inhibitors (for treating impotence by blocking a chemical reaction which controls muscle relaxation and resulting blood flow rate), various antibacterial agents, and protein kinase inhibitors for metastatic lung cancer and other tumors [913]. See Figure 15.1 for illustrations of popular drugs for migraine, HIV/AIDS, and blood-flow related diseases.

Such computer modeling and analysis — rather than using trial and error and exhaustive database studies — was thought to lead to dramatic progress in the design of drugs. However, some believe that the field of rational drug design has not lived up to its expectations.

One reason for the restrained success is the limited reliability of modeling molecular interactions between drugs and target molecules; such interactions must be described very accurately energetically to be useful in predictions. Newer approaches consider multiple targets [278] and work in system-oriented approaches [649] to improve success.

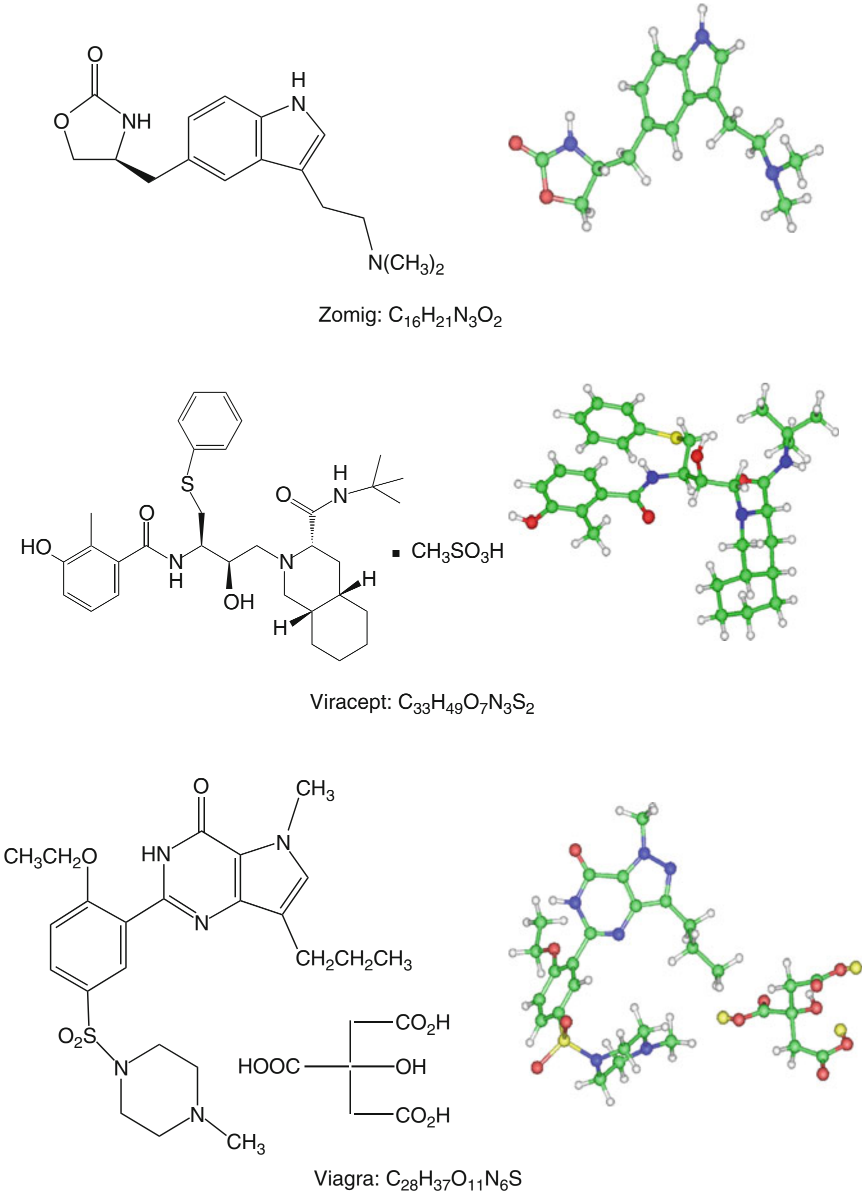
Another reason for the limited success of drug modeling is that the design of compounds with the correct binding properties (e.g., dissociation constants in the micromolar range and higher) is only a first step in the complex process of drug design; many other considerations and long-term studies are needed to determine the drug's bioactivity and its effects on the human body [1364]. For example, a compound may bind well to the intended target but be inactive biologically if the reaction that the drug targets is influenced by other components (see Box 15.2 for an example). Even when a drug binds well to an appropriate target, an optimal therapeutic agent must be delivered precisely to its target [999], screened for undesirable drug/drug interactions [1061], lack toxicity and carcinogenicity (likewise for its metabolites), be stable, and have a long shelf life.

The problems of viability and efficacy are even more important now with the increased development and usage of *biologics* or *biotherapeutics* — biological molecules like proteins derived from living cells and used as drugs — rather than small-molecule drugs. Such biologics, which include various vaccines, are typically administered by injection or infusion. Successful recent examples are Wyeth's *Enbrel* for rheumatoid arthritis, Genetech's *Avastin* for cancer, and Amgen's *Epogen* for anemia. Many large pharmaceutical companies are increasing their work on biologics because such drugs are more complex and expensive to replicate and hence much less vulnerable to the usual patent expiration which allows introduction of generics and thereby restricts the profits of the original manufacturers. However, the big challenge in biologics is dealing with the characteristic heterogeneity of such biological molecules and better understanding their mechanism of action related to the disease target and long-term effects.

15.1.4 The Competition: Automated Technology

Even accepting those limitations of computer-based approaches, rational drug design has avid competition from automated technology: new synthesis techniques, such as robotic systems that can run hundreds of concurrent synthetic reactions, have emerged, thereby enhancing synthesis productivity enormously. With "high-throughput screening", these candidates can be screened rapidly to analyze binding affinities, determine transport properties, and assess conformational flexibility.

Many believe that such a production *en masse* is the key to establishing diverse databases of drug candidates. Thus, at this time, it might be viewed that *drug design need not be 'rational' if it can be exhaustive*. Still, others advocate a more focused design approach, based on structures of ligands or receptors [453], fragment-based drug design [1447], or virtual screening approaches applied to smaller subsets of compounds [453, 1179].



Zomig: $C_{16}H_{21}N_3O_2$

Viracept: $C_{33}H_{49}O_7N_3S_2$

Viagra: $C_{28}H_{37}O_{11}N_6S$

Figure 15.1. Popular drug examples. Top: *Zolmitriptan* (zomig) for migraines, a 5-HT$_1$ receptor agonist that enhances the action of serotonin. Middle: *Nelfinavir Mesylate* (viracept), a protease inhibitor for AIDS treatment. Bottom: *Sildenfel Citrate* (viagra) for penile dysfunctions, a temporary inhibitor of phosphodiesterase-5, which regulates associated muscle relaxation and blood flow by converting cyclic guanosine monophosphate to guanosine monophosphate. See other household examples in Figure 15.3.

Another convincing argument for the focused design approach is that the amount of synthesized compounds is so vast (and rapidly generated) that computers will be essential to sort through the huge databases for compound management and applications. Such applications involve clustering analysis and *similarity* and *diversity sampling* (see below), preliminary steps in generating drug candidates or optimizing bioactive compounds.

This information explosion explains the resurrection of computer-aided drug design and its enhancement in scope under the new title **combinatorial chemistry**, affectionately endorsed as ‘the darling of chemistry’ [1376].

15.1.5 Chapter Overview

In this chapter, a brief introduction into some mathematical questions involved in this discipline of chemical library design is presented, namely *similarity and diversity sampling* for ligand-based drug design. Some ideas on cluster analysis and database searching are also described. This chapter is only intended to whet the appetite for chemical design and to invite mathematical scientists to work on related problems.

Because medicinal chemistry applications are an important subfield of chemical design, this last chapter also provides some perspectives on current developments in drug design, as well as mentioning emerging areas such as pharmacogenomics of personalized medicine and biochips (see Boxes 15.3 and 15.4).

15.2 Problems in Chemical Libraries

Chemical libraries consist of compounds (known chemical formulas) with potential and/or demonstrated therapeutic activities. Most libraries are proprietary, residing in pharmaceutical houses, but public sources also exist, like the National Cancer Institute’s (NCI’s) 3D structure database.

Both target-independent and target-specific libraries exist. The name ‘combinatorial libraries’ stems from the important combinatorial problems associated with the experimental design of compounds in chemical libraries, as well as computational searches for potential leads using concepts of **similarity** and **diversity** as introduced below.

15.2.1 Database Analysis

In broad terms, two general problem categories can be defined in chemical library analysis and design:

Database systematics: analysis and compound grouping, compound classification, elimination of redundancy in compound representation (dimensionality reduction), data visualization, etc., and

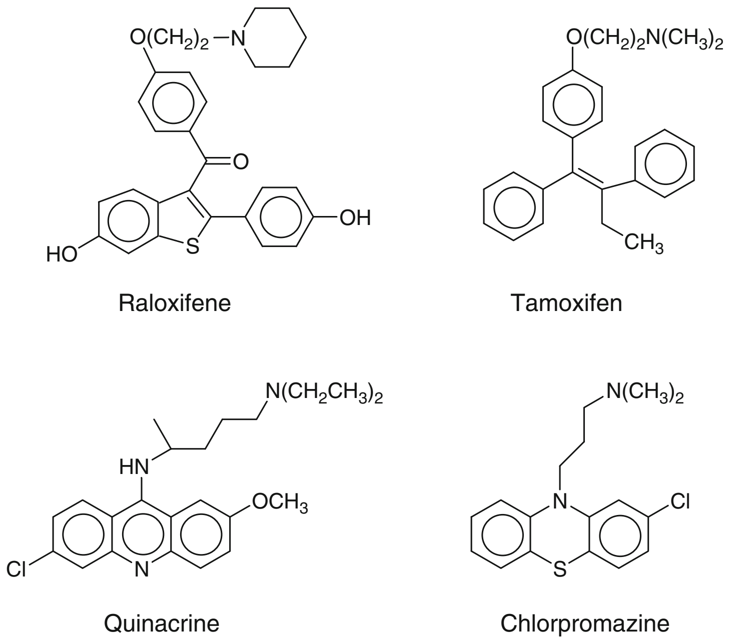


Figure 15.2. Related pairs of drugs: the antiestrogens raloxifene and tamoxifen, and the tricyclic compounds with aliphatic side-chains at the middle ring quinacrine and chlorpromazine.

Database applications: efficient formulation of quantitative links between compound properties and biological activity for compound selection and design optimization experiments.

Both of these general database problems involved in chemical libraries are associated with several mathematical disciplines. Those disciplines include multivariate statistical analysis and numerical linear algebra, multivariate nonlinear optimization (for continuous formulations), combinatorial optimization (for discrete formulations), distance geometry techniques, and configurational sampling.

15.2.2 Similarity and Diversity Sampling

Two specific problems, described formally in the next section after the introduction of chemical descriptors, are the similarity and diversity problems.

The *similarity* problem in drug design involves finding molecules that are 'similar' in physical, chemical, and/or biological characteristics to a known target compound. Deducing compound similarity is important, for example, when one drug is known and others are sought with similar physiochemical and biological properties, and perhaps with reduced side effects.

One example is the target bone-building drug *raloxifene*, whose chemical structure is somewhat related to the breast cancer drug *tamoxifen* (see Figure 15.2) (e.g., [1093]). Both are members of the family of *selective estrogen receptor modulators* (SERMs) that bind to estrogen receptors in the breast cancer cells and exert a profound influence on cell replication. It is hoped that raloxifene will be as effective for treating breast tumors but will reduce the increased risk of endometrial cancer noted for tamoxifen. Perhaps raloxifene will also not lose its effectiveness after five years like tamoxifen.

Another example of a related pair of drugs is *chlorpromazine* (for treating schizophrenia) and *quinacrine* (antimalarial drug). These tricyclic compounds with aliphatic side chains at the middle ring group (see Figure 15.2) were suggested as candidates for treating Creutzfeldt-Jakob and other prion diseases [677].

Because similarity in structure might serve as a first criterion for similarity in activity/function, similarity searching can be performed using 3D structural and energetic searches (e.g., induced fit or ‘docking’ [41, 818]) or using the concept of molecular descriptors introduced in the next section, possibly in combination with other discriminatory criteria.

The *diversity* problem in drug design involves delineating the most diverse subset of compounds within a given library. Diversity sampling is important for practical reasons. The smaller, representative subsets of chemical libraries (in the sense of being most ‘diverse’) might be searched first for lead compounds, thereby reducing the search time; representative databases might also be used to prioritize the choice of compounds to be purchased and/or synthesized, similarly resulting in an accelerated discovery process, not to speak of economic savings.

Box 15.2: Treatments for Chronic Pain

Amazing breakthroughs have been achieved recently in the treatment of chronic pain. Such advances were made possible by an increased understanding of the distinct cellular mechanisms that cause pain due to different triggers, from sprained backs to arthritis to cancer.

What is Pain? Pain signals start when nerve fibers known as nociceptors, found throughout the human body, react to some disturbances in nearby tissues. The nerve fibers send chemical pain messengers that collect in the dorsal horn of the spinal cord. Their release depends on the opening of certain pain gates. Only when these messengers are released into the brain is pain felt in the body.

Natural Ammunition. Fortunately, the body has a battery of natural painkillers that can close those pain gates or send signals from the brain. These compensatory agents include endorphins, adrenaline, and serotonin (a peptide similar to opium). Many painkillers enhance or mimic the action of these natural aids (e.g., opium-bases drugs such as *morphine*, *codeine*, and *methadone*). However, these opiates have many undesirable side effects.

Painkiller Targets. To address the problem of pain, new treatments are targeting *specific* opiate receptors. For example, Actiq, developed by Anesta Corp. for intense cancer pain, is a lozenge placed in the cheek that is absorbed quickly into the bloodstream, avoiding the gut. Other pain relievers include a class of drugs known as COX-2 inhibitors, like Monsanto's *Celebrex* (celecoxib) and Merck's *Vioxx*, which relieve aches and inflammation with fewer stomach-damaging effects. They do so by targeting only one (COX-2) of two enzymes called cyclo-oxegenases (COX), which are believed to cause inflammation and thereby trigger pain.

While regular non-steroidal anti-inflammatory drugs (NSAIDs, like *Aspirin*, *Ibuprofen*, and *Naproxen*) and others available by prescription attack both COX-1 and COX-2, COX-1 is also known to protect the stomach lining; this explains the stomach pain that many people experience with NSAIDs and the pain relief without the side effects that COX-2 inhibitors can offer.

Modern pain treatment also involves compounds that stop pain signals before the brain gets the message, either by intercepting the signals in the spinal cord or by blocking their route to the spine. Evidence is emerging that a powerful chemical called 'substance P' can be used as an agent to deliver pain blockers to receptors found throughout the body; an experimental drug based on this idea (marketed by Pfizer) has proven effective at easing tooth pain.

15.2.3 Bioactivity Relationships

Besides database systematics, such as similarity and diversity sampling, the establishment of clear links between compound properties and bioactivity is, of course, the heart of drug design. In many respects, this association is not unlike the protein prediction problem in which we seek some target energy function that upon global minimization will produce the biologically relevant, or native, structure of a protein.

In our context, formulating that 'function' to relate sequence and structure while not ignoring the environment might even be more difficult, since we are studying small molecules for which the evolutionary relationships are not clear as they might be for proteins. Further, the bioactive properties of a drug depend on much more than its chemical composition, three-dimensional (3D) structure, and energetic properties. A complex orchestration of cellular machinery is often involved in a particular human ailment or symptom, and this network must be understood to alleviate the condition safely and successfully.

A successful drug has usually passed many rounds of chemical modifications that enhanced its potency, optimized its selectivity, and reduced its toxicity. An example involves obesity treatments by the hormone *leptin*. Limited clinical studies have shown that leptin injections do not lead to clear trends of weight loss in people, despite demonstrating dramatic slimming of mice. Though not a quick panacea in humans, leptin has nonetheless opened the door to pharmacological manipulations of body weight, a dream with many medical — not to speak of monetary — benefits. Therapeutic manipulations will require an understanding of the complex mechanism associated with leptin regulation of our appetite, such as its signaling the brain on the status of body fat.

Box 15.2 contains another illustration of the need to understand such complex networks in connection with drug development for chronic pain. These examples clearly show that *lead generation*, the first step in drug development, is followed by *lead optimization*, the challenging, slower phase.

In fact, this complexity of the molecular machinery that underlies disease has given rise to the subdisciplines of *molecular medicine* and *personalized medicine* (see Boxes 15.3 and 15.4), where DNA technology plays an important role. Specifically, DNA chips — small glass wafers like computer chips studded with bits of DNA instead of transistors — can analyze the activities of thousands of genes at a time, helping to predict disease susceptibility in individuals, classify certain cancers, and to design treatments [400].

For example, DNA chips can study expression patterns in the tumor suppressor gene *p53* (the gene with the single most common mutations in human cancers), and such patterns can be useful for understanding and predicting response to chemotherapy and other drugs. DNA microarrays have also been used to identify genes that selectively stimulate metastasis (the spread of tumor cells from the original growth to other sites) in melanoma cells.

Besides developments on more personalized medicine, which will also be enhanced by a better understanding of the human body and its ailments, new advances in drug delivery systems may be important for improving the rate and period of drug delivery in general [1304].

Box 15.3: Molecular and Personalized Medicine

Pauling's Groundwork. Molecular medicine seeks to enhance our therapeutic solutions by understanding the molecular basis of disease. Linus Pauling lay the groundwork for this field in his seminal 1949 paper [977] which demonstrated that the hemoglobin from sickle cell anemia sufferers has a different electric charge than that from healthy people. This difference was later explained by Vernon Ingram as arising from a single amino acid difference [590]. These pioneering works relied on electrophoretic mobility measurements and fingerprinting techniques (electrophoresis combined with paper chromatography) for peptides.

Disease Simulations. A modern incarnation of molecular medicine involves conducting virtual experiments by computer simulation with the goal of developing new hypotheses regarding disease mechanisms and prevention. For example, scientists at Entelos Inc. (Menlo Park, California) are simulating cell inflammation caused by asthma to try to learn how blocking certain inflammation factors might affect cellular receptors and then to identify targets for steroid inhalers.

From SNPs to Tailored Drugs. Another significant current trend in medicine is personalized medicine, the tailoring of drugs to individual genetic makeup. User-specific drugs have great potential to be more potent and to eliminate adverse side effects experienced by some individuals. *Pharmacogenetics* is the field of studying how genetic factors influence drug response. Its newer sibling *pharmacogenomics* involves using genomics to describe individual responses to drugs. Pharmacogenomics (also abbreviated as Pgx or pgx) has become possible with the advent of microarray technology (e.g., [544, 1174]): these make possible large-scale genome-wide analyses to test thousands of genes for related activity with a specific drug. Developing tailored diets and vitamins based on individual responses to diet (determined in part by one's genes) is another growing field called *nutritional genomics* or *nutrigenomics*.

Specifically, the drug tailoring idea is based on identifying the small variations in people's DNA where a single nucleotide differs from the standard sequence. These mutations, or individual variations in genome sequence that occur once every couple of hundred of base pairs, are called single-nucleotide polymorphisms known as SNPs (pronounced "snips"). The presence of SNPs can be signaled visually using DNA chips or biochips, instruments of fancy of the biotechnology industry (See [400] and Box 1.4 of Chapter 1). Other genomic factors besides SNPs also serve as distinguishing factors in pharmacogenomics studies.

Pharmacogenetics gained momentum in April 1999 when eleven pharmaceutical and technology companies and the Wellcome Trust announced a genome *mapping* consortium for SNPs. The consortium's goal is to construct a fine-grained map of order 300,000 SNPs to permit searching for SNP patterns that correlate with particular drug responses. Efforts are ongoing, and many companies have specialized in this area. Pharmacogenomics now receives considerable attention both from the professional medical circles and the popular press. It has potential to markedly improve medical intervention, reduce hospitalization costs, and alleviate human suffering by increasing the efficacy and decreasing adverse effects in the drug treatment of various human diseases.

Some notable examples of success of pharmacogenomics include the genotype-based dosaging of the blood thinning drug *Warfarin*; administration of *Abacavir* (an RT inhibitor) to HIV patients; *Herceptin* treatment for HER2-positive breast cancer patients; and *Gleevac* and other cancer drugs for individual cancer patients. See Box 15.4 for details of some of these drugs.

Directed drugs are also under development to treat or diagnose diabetics, neurological diseases like Alzheimer's, prostate cancer, and ailments requiring antibiotics. Though there are many hurdles to this new field, not to mention possible financial drawbacks of genotyping, it is hoped that some benefits of cost savings in prescriptions and hospitalizations for adverse drug effects could be realized in the not-too-distant future [588].

Box 15.4: Examples of Successes in Pharmacogenomics

Warfarin. *Warfarin* is the "darling" of pharmacogenomics because international collaborations by the International Warfarin Pharmacogenetics Consortium and the Pharmacogenetics Research Network have led to development of a dosing algorithm [591].

This is a milestone in the evolution of drug prescription from trial and error to exact science [591]. Warfarin is an anti-coagulation agent given to patients with risk of heart disease. However, adverse effects can be catastrophic since the patient may bleed to death. Practical experience has shown that reactions to the drug vary widely from person to person. But why? Pharmacogenomics analyses revealed that a patient's response to Warfarin depends on the presence of two genes encoding two proteins: CYP2C9 which metabolizes warfarin, and VKORC1 which recycles vitamin K and affects clotting factors. Certain genotypes make reaction much more sensitive. In 2007, FDA modified Warfarin labels to highlight the potential relevance of genetic information to prescribing decisions.

Abacavir. *Abacavir* is a guanosine reverse-transcriptase inhibitor used as an anti-retroviral treatment against infections of HIV. However, 5 to 8% of the white population develops adverse side effects, namely toxic skin reaction. In 2002, it was discovered that the HLAB*5701 gene variant is highly associated with this hyper sensitivity. Genotyping has thus been used to effectively reduce the number of such adverse reactions. Genetic testing for sensitivity to abacavir is now widely used.

Herceptin. *Herceptin* (Trastuzumab) is an antibody used to treat breast cancer. Studies have shown that Herceptin is effective for patients with over-expression of the human epidermal growth factor receptor HER2, which occurs in invasive breast carcinomas. Herceptin has now been approved by the FDA for patients with invasive breast cancer that over expresses HER2.

Codeine for Breast-Feeding Mothers. *Codeine* is a painkiller often prescribed to help women with post-delivery pain. Codeine is metabolized into morphine, but it was generally considered to be safe for breast-feeding mothers. In 2005, a 13-day old male baby in Toronto who was breastfed by a codeine-treated mother died of a morphine overdose [673]. Investigations revealed that the mother was an "ultra-metabolizer" of codeine, and this led to an unusually high amount of morphine in the baby. Studies have shown that the metabolism of codeine is related to the CYP2D6 gene. Subsequently, genetic testing for this variant has been suggested for mothers who want to breastfeed and receive codeine for post-delivery pain [1375]. Alternatively, breast feeding can be avoided, reduced, and/or the level of morphine in the neonate monitored carefully to prevent unnecessary deaths.

15.3 General Problem Definitions

15.3.1 The Dataset

Our given dataset of size n contains information on compounds with potential *biological activity* (drugs, herbicides, pesticides, etc.). A schematic illustration is presented in Figure 15.3. The value of n is large, say one million or more. Because of the enormous dataset size, the problems described below are simple to solve in principle but extremely challenging in practice because of the large associated computational times. Any systematic schemes to reduce this computing time can thus be valuable.

Chemical Library ($n >> m$)

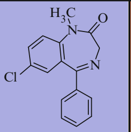
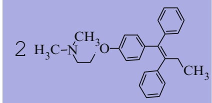
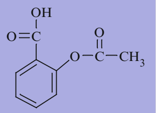
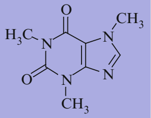
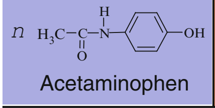
Compound ($i = 1,\ldots, n$)		Vectorial Descriptors ($k = 1,\ldots, m$)				Biological Targets ($j = 1,\ldots,m_B$)			
		Xi_1	Xi_2	$\cdot\cdot\cdot$	Xi_m	Bi_1	Bi_2	$\cdot\cdot\cdot$	Bi_{mB}
1	 Valium	0.873	0.763	...	0.531	0	1	...	0
2	 Tamoxifen	0.912	0.131	...	0.834	0	0	...	1
3	 Aspirin	0.763	0.214	...	0.533	0	0	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	 Caffeine	0.925	0.237	...	0.742	1	0	...	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	 Acetaminophen	0.347	0.279	...	0.846	1	1	...	0

Figure 15.3. A chemical library can be represented by n compounds i (known or potential drugs), each associated with m characteristic descriptors ($\{Xi_k\}$) and activities $\{Bi_j\}$ with respect to m_B biological targets (known or potential).

15.3.2 The Compound Descriptors

Each compound in the database is characterized by a vector (the *descriptor*). The vector can have real or binary elements. There are many ways to formulate these descriptors so as to reduce the database search time and maximize success in generation of lead compounds.

Conventionally, each compound i is described by a list of **chemical descriptors**, which may reflect *molecular composition*, such as atom number, atom connectivity, or number of functional groups (like aromatic or heterocyclic rings, tertiary aliphatic amines, alcohols, and carboxamides), *molecular geometry*, such as number of rotatable bonds, *electrostatic properties*, such as charge distribution, and various *physiochemical measurements* that are important for bioactivity.

These descriptors are currently available from many commercial packages like Molconn-X and Molconn-Z (Hall Associates Consulting, Qincy, MD). Descriptors fall into many classes. Examples include:

2D descriptors — also called molecular connectivity or topological indices — reflecting molecular connectivity and other topological invariants;

binary descriptors — simpler encoded representations indicating the presence or absence of a property, such as whether or not the compound contains at least three nitrogen atoms, doubly-bonded nitrogens, or alcohol functional groups;

3D descriptors — reflecting geometric structural factors like van der Waals volume and surface area; and

electronic descriptors — characterizing the ionization potential, partial atomic charges, or electron densities.

See also [8] for further examples.

Binary descriptors allow rapid database analysis using Boolean algebra operations. The MolConn-X and MolConn-Z programs, for example, generate topological descriptors based on molecular connectivity indices (e.g., number of atoms, number of rings, molecular branching paths, atoms types, bond types, etc.). Such descriptors have been found to be a convenient and reasonably successful approximation to quantify molecular structure and relate structure to biological activity (see review in [6]). These descriptors can be used to characterize compounds in conjunction with other selectivity criteria based on activity data for a training set (e.g., [322, 582]). The search for the most appropriate descriptors is an ongoing enterprise, not unlike force-field development for macromolecules.

The number of these descriptors, m, is roughly on the order of 1000, thus much smaller than n (the number of compounds) but too large to permit standard systematic comparisons for the problems that arise.

Let us define the vector Xi associated with compound i to be the row m-vector

$$\{Xi_1, Xi_2, \ldots, Xi_m\}\,.$$

Our dataset $\mathcal{S}$ can then be described as the collection of n vectors

$$\mathcal{S} = \{X1, X2, X3, \ldots, Xn\} ,$$

or expressed as a rectangular matrix $A_{n \times m}$ by listing, in rows, the m chemical descriptors of the n database compounds:

$$A = \left(\begin{array}{cccccc} X1_1 & X1_2 & \cdots & \cdots & \cdots & X1_m \\ X2_1 & X2_2 & \cdots & \cdots & \cdots & X2_m \\ \vdots & & & \cdots & & \\ \vdots & & & \cdots & & \\ \vdots & & & \cdots & & \\ \vdots & & & \cdots & & \\ \vdots & & & \cdots & & \\ Xn_1 & Xn_2 & \cdots & \cdots & \cdots & Xn_m \end{array} \right) . \tag{15.1}$$

In practice, this rectangular $n \times m$ matrix has $n \gg m$ (i.e., the matrix is long and narrow), where n is on the order of millions and m is several hundreds.

The compound descriptors are generally *highly redundant*. Yet, it is far from trivial how to select the “principal descriptors”. Thus, various statistical techniques (principal component analysis, classic multivariate regression; see below) have been used to assess the degree of correlation among variables so as to eliminate highly-correlated descriptors and reduce the dimension of the problems involved.

15.3.3 Characterizing Biological Activity

Another aspect of each compound in such databases is its *biological activity*. Pharmaceutical scientists might describe this property by associating a simple *affirmative* or *negative* score with each compound to indicate various areas of activity (e.g., with respect to various ailments or targets, which may include categories like headache, diabetes, protease inhibitors, etc.).

Drugs may enhance/activate (e.g., *agonists*) or inhibit (e.g., *antagonists, inhibitors*) certain biochemical processes. This bioactivity aspect of database problems is far less quantitative than the simple chemical descriptors. Of course, it also requires synthesis and biological testing for activity determination. Studies of several drug databases have suggested that active compounds can be associated with certain ranges of physiochemical properties like molecular weight and occurrence of functional groups [451].

For the purpose of the problems outlined here, it suffices to think of such an additional set of descriptors associated with each compound. For example, a matrix $B_{n \times m_B}$ may complement the $n \times m$ database matrix A; see Figure 15.3. Each row i of B may correspond to measures of activity of compound i with respect to specific targets (e.g., binary variables for active/nonactive target response).

The ultimate goal in drug design is to find a compound that yields the desired pharmacological effect. This quest has led to the broad area termed SAR, an acronym for Structure/Activity Relationship [709]. This discipline applies various statistical, modeling, or optimization techniques to relate compound properties to associated pharmacological activity. A simple linear model, for example, might attempt to solve for variables in the form of a matrix $X_{m \times m_B}$, satisfying

$$AX = B\,. \tag{15.2}$$

Explained more intuitively, SAR formulations attempt to relate the given compound descriptors to experimentally-determined bioactivity markers. While earlier models for 'quantitative SAR' (QSAR) involved simple linear formulations for fitting properties and various statistical techniques (e.g., multivariate regression, principal component analysis), nonlinear optimization techniques combined with other visual and computational techniques are more common today [448]. The problem remains very challenging, with rigorous frameworks continuously being sought.

15.3.4 The Target Function

To compare compounds in the database to each other and to new targets, a quantitative assessment can be based on common structural features. Whether characterized by topological (chemical-formula based) or 3D features, this assessment can be broadly based on the vectorial chemical descriptors provided by various computer packages. A target function f is defined, typically based on the *Euclidean distance* function between vector pairs, δ, where

$$f(Xi, Xj) = \delta_{ij} \equiv \|Xi - Xj\| = \sqrt{\sum_{k=1}^{m} (Xi_k - Xj_k)^2}\,. \tag{15.3}$$

Thus, to measure the similarity or diversity for each pair of compounds Xi and Xj, the function $f(Xi, Xj)$ is often set to the simple distance function δ_{ij}. Other functions of distance are also appropriate depending upon the objectives of the optimization task.

15.3.5 Scaling Descriptors

Scaling the descriptor components is important for proper assessment of the score function [1372]. This is because the individual chemical descriptors can vary drastically in their magnitudes as well as the variance within the dataset. Subsequently, a few large descriptors can overwhelm the similarity or diversity measures. For example, actual descriptor components of a database compound may look like the following:

```
11.0000  0.6433 4.5000  0.0833 150.2200  8.4831  0.0159 -1.0000 113.2239 ..
 1.000   0.2917 0.5000  0.0000  40.0000  7.2566  0.0801  1.0000 782.7121 ..
-8.0000  0.2081 0.5000  0.0186  80.0000  0.0000  0.0017  1.0000  62.2016 ..
 2.0000  0.0000 2.5000 -0.9010   0.0000  1.3867  0.2500  1.0000 120.0030 ..
 0.0000  0.0000 3.0000  0.0326   0.0000 -4.3984  0.1759  1.0000  11.2189 ..
80.0000 -0.0442 6.0000  0.7002 210.0000 -1.9784  0.0026 -1.0000 370.3473 ..
-5.0000 -0.1491 0.0000  0.0000  10.0000  9.0909  0.1641  1.0000  98.2782 ..
-1.0000  0.5427 4.5000  0.8963  35.0000  2.0061  0.0720  1.0000 119.8090 ..
17.0000 -0.3209 0.5000  0.0803   0.0000  9.4765  0.0000 -1.0000  11.7011 ..
19.0000  0.2690 1.0000 -0.3420  90.0000  0.0000  0.0000 -1.0000 201.0180 ..
 0.0000  0.0000 0.0000  0.2000  40.0000  9.1702  0.0429 -1.0000  23.2423 ..
 4.0000  0.3061 0.5000  0.6670  10.0000  2.3820  0.0023  1.0000   0.0000 ..
 4.0000  0.7702 1.5000  0.1870   0.0000  0.0000  0.7290  1.0000   0.0000 ..
 1.0000 -0.1134 1.5000  0.3356  40.0000  0.0000  0.7782 -1.0000 314.6658 ..
 0.0000  0.0000 0.0000  0.7842   0.0000 -6.1659  0.0000  1.0000  85.2285 ..
 3.0000  0.0000 0.0000  0.2382  75.0000  4.2276  0.1260  1.0000   7.2854 ..
15.0000  0.3479 4.0000  0.0034   0.0000  0.5152  0.3018  1.0000 280.8721 ..
 7.0000  0.6945 3.5000  0.4552   0.0000  3.5315  0.3065 -1.0000   0.0000 ..
    .       .      .       .        .       .       .       .        .   ..
    .       .      .       .        .       .       .       .        .   ..
```

Clearly, the ranges of individual descriptors vary (e.g., 0 to 1 versus 0 to 1000). Thus, given no chemical/physical guidance, it is customary to scale the vector entries before analysis. In practice, however, it is very difficult to determine the appropriate scaling and displacement factors for the specific application problem [1372]. A general scaling of each Xi_k to produce $\hat{X}i_k$ can be defined using two real numbers α_k and β_k, for $k = 1, 2, \ldots, m$, termed the *scaling* and *displacement* factors, respectively, where $\alpha_k > 0$. Namely, for $k = 1, 2, \ldots, m$, we define the scaled components as

$$\hat{X}i_k = \alpha_k \, (Xi_k - \beta_k), \qquad 1 \leq i \leq n \, . \tag{15.4}$$

The following two scaling procedures are often used. The first makes each column in the range $[0, 1]$: each column of the matrix A is modified using eq. (15.4) by setting the factors as

$$\begin{aligned} \beta_k &= \min_{1 \leq i \leq n} Xi_k \, , \\ \alpha_k &= 1 / \left(\max_{1 \leq i \leq n} Xi_k - \beta_k \right) . \end{aligned} \tag{15.5}$$

This scaling procedure is also termed “standardization of descriptors”.

The second scaling produces a new matrix A where each column has a mean of zero and a standard deviation of one. It does so by setting the factors (for $k = 1, 2, \ldots, m$) as

$$\begin{aligned} \beta_k &= \frac{1}{n} \sum_{i=1}^{n} Xi_k \, , \\ \alpha_k &= 1 / \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Xi_k - \beta_k)^2} \, . \end{aligned} \tag{15.6}$$

Both scaling procedures defined by eqs. (15.5) and (15.6) are based on the assumption that no one descriptor dominates the overall distance measures.

15.3.6 The Similarity and Diversity Problems

The Euclidean distance function $f(Xi, Xj) = \delta_{ij}$ based on the chemical descriptors can be used in performing similarity searches among the database compounds and between these compounds and a particular target. This involves optimization of the distance function over $i = 1, \ldots, n$, for a fixed j:

$$\text{Minimize } _{Xi \in \mathcal{S}} \, \{f(\delta_{ij})\} \, . \tag{15.7}$$

More difficult and computationally-demanding is the diversity problem. Namely, we seek to reduce the database of the n compounds by selecting a "representative subset" of the compounds contained in $\mathcal{S}$, that is one that is "the most diverse" in terms of potential chemical activity. We can formulate the diversity problem as follows:

$$\text{Maximize} \sum_{Xi, Xj \in \mathcal{S}_0} \{f(\delta_{ij})\} \tag{15.8}$$

for a given subset $\mathcal{S}_0$ of size n_0.

The molecular diversity problem naturally arises since pharmaceutical companies must scan huge databases each time they search for a specific pharmacological activity. Thus reducing the set of n compounds to n_0 representative elements of the set $\mathcal{S}_0$ is likely to accelerate such searches. 'Combinatorial library design' corresponds to this attempt to choose the best set of substituents for combinatorial synthetic schemes so as to maximize the likelihood of identifying lead compounds.

The molecular diversity problem involves maximizing the volume spanned by the elements of $\mathcal{S}_0$ as well as the separation between those elements. Geometrically, we seek a well separated, uniform-like distribution of points in the high-dimensional compound space in which each chemical cluster has a 'representative'.

A simple, heuristic formulation of this problem might be based on the similarity problem above: successively minimize $f(\delta_{ij})$ over all i, for a fixed (target) j, so as to eliminate a subset $\{Xi\}$ of compounds that are similar to Xj. This approach thus identifies groupings that *maximize intracluster similarity* as well as *maximize intercluster diversity*.

The *combinatorial optimization* problem, an example of a very difficult computational task, has *non-polynomial computational complexity* ('NP-complete') (see footnote in Chapter 11, Section 11.2). This is because an exhaustive calculation of the above distance-sum function over a *fixed set* $\mathcal{S}_0$ of n_0 elements requires a total of $\mathcal{O}(n_0^2 m)$ operations. However, there are many possible subsets of $\mathcal{S}$ of size n_0, namely $C_n^{n_0}$ of them, where

$$\begin{aligned} C_n^{n_0} &= \frac{n!}{n_0! \, (n - n_0!)} \\ &= \frac{n(n-1)(n-2)\cdots(n-n_0+1)}{n_0!} \, . \end{aligned} \tag{15.9}$$

As a simple example, for $n = 4$, we have $C_4^1 = 4/1 = 4$ subsets of one element; $C_4^2 = (4 \times 3)/2 = 6$ different subsets of two elements, $C_4^3 = (4 \times 3 \times 2)/(3!) = 4$ subsets of three elements, and $C_4^4 = (4 \times 3 \times 2)/(4!) =$ one subset of four elements.

Typically, these combinatorial optimization problems are solved by stochastic and heuristic approaches. These include genetic algorithms, simulated annealing, and tabu-search variants. (See Agrafiotis [5], for example, for a review).

As in other applications, the efficiency of simulated annealing depends strongly on the choice of cooling schedule and other parameters. Several potentially valuable annealing algorithms such as deterministic annealing, multiscale annealing, and adaptive simulated annealing, as well as other variants, have been extensively studied.

Various formulations of the diversity problem have been used in practice. Examples include the maximin function — to maximize the minimum intermolecular similarity:

$$\text{Maximize}_{i,\; Xi \in \mathcal{S}_0} \;\{ \min_{\substack{j \neq i \\ Xj \in \mathcal{S}_0}} \;(\delta_{ij}) \;\} \tag{15.10}$$

or its variant — maximizing the sum of these distances:

$$\text{Maximize}_{Xi, Xj \in \mathcal{S}_0} \sum_i \{\min_{j \neq i} (\delta_{ij}) \;\} \,. \tag{15.11}$$

The maximization problem above can be formulated as a minimization problem by standard techniques if $f(x)$ is normalized so it is monotonic with range $[0, 1]$, since we can often write

$$\max[f(x)] \Leftrightarrow \min[-f(x)] \;\text{ or }\; \min[1 - f(x)] \,.$$

In special cases, combinatorial optimization problems can be formulated as integer programming and mixed-integer programming problems. In this approach, linear programming techniques such as interior methods can be applied to the solution of combinatorial optimization problems, leading to branch and bound algorithms, cutting plane algorithms, and dynamic programming algorithms. Parallel implementation of combinatorial optimization algorithms is also important in practice to improve the performance.

Other important research areas in combinatorial optimization include the study of various algebraic structures (such as matroids and greedoids) within which some combinatorial optimization problems can more easily be solved [263].

Currently, practical algorithms for addressing the diversity problem in drug design are relatively simple heuristic schemes that have computational complexity of at most $\mathcal{O}(n^2)$, already a huge number for large n.

15.4 Data Compression and Cluster Analysis

Dimensionality reduction and data visualization are important aids in handling the similarity and diversity problems outlined above. Principal component analysis (PCA) is a classic technique for data compression (or dimensionality reduction). It has already shown to be useful in analyzing microarray data (e.g., [1009]), as discussed in Chapter 1. The singular value decomposition (SVD) is another closely related approach. Data visualization for cluster analysis requires dimensionality reduction in the form of a projection from a high-dimensional space to 2D or 3D so that the dataset can be easily visualized. Cluster analysis is heuristic in nature.

In this section we outline the PCA and SVD approaches for dimensionality reduction in turn, continue with the distance refinement that can follow such analyses, and illustrate projection and clustering results with some examples.

15.4.1 Data Compression Based on Principal Component Analysis (PCA)

PCA transforms the input system (our database matrix A) into a smaller matrix described by a few uncorrelated variables called the **principal components** (PCs). These PCs are related to the eigenvectors of the covariance matrix defined by the component variables. The basic idea is to choose the orthogonal components so that the original data variance is well approximated. That is, the relations of similarity/dissimilarity among the compounds can be well approximated in the reduced description. This is done by performing eigenvalue analysis on the covariance matrix that describes the statistical relations among the descriptor variables.

Covariance Matrix and PCs

Let a_{ij} be an element of our $n \times m$ database matrix A. The covariance matrix $C_{m\times m}$ is formed by elements $c_{jj'}$ where each entry is obtained from the sum

$$c_{jj'} = \frac{1}{n-1}\sum_{i=1}^{n}(a_{ij} - \mu_j)\,(a_{ij'} - \mu_{j'})\,. \tag{15.12}$$

Here μ_j is the mean of the column associated with descriptor j:

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} a_{ij}\,. \tag{15.13}$$

C is a symmetric semi-definite matrix and thus has the spectral decomposition

$$C = V\Sigma V^T\,, \tag{15.14}$$

where the superscript T denotes the matrix transpose, and the matrix V $(m\times m)$ is the orthogonal eigenvector matrix satisfying $VV^T = I_{m\times m}$ with m component vectors $\{v_i\}$. The diagonal matrix Σ of dimension m contains the m ordered eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0\,.$$

We then define the m PCs Yj for $j = 1, 2, \cdots, m$ as the product of the original matrix A and the eigenvectors v_j:

$$Yj = Av_j\,, \qquad j = 1, 2, \cdots, m\,. \tag{15.15}$$

We also define the $m \times m$ matrix Y corresponding to eq. (15.15), related to V, as the matrix that holds the m PCs $Y1, Y2, \cdots, Ym$; this allows us to write eq. (15.15) in the matrix form $Y = AV$. Since $VV^T = I$, we then obtain an expression for the dataset matrix A in terms of the PCs:

$$A = YV^T\,. \tag{15.16}$$

Dimensionality Reduction

The problem dimensionality can be reduced based on eq. (15.16). First note that eq. (15.16) can be written as:

$$A = \sum_{j=1}^{m} Yj \cdot v_j^T\,. \tag{15.17}$$

Second, note that Xi, the vector of compound i, is the transpose of the ith row vector of A:

$$Xi = A^T\, e_i\,, \tag{15.18}$$

where e_i is an $n \times 1$ unit vector with 1 in the ith component and 0 elsewhere. Thus, compound Xi is expressed as the linear combination of the orthonormal set of eigenvectors $\{v_j\}$ of the covariance matrix C derived from A:

$$Xi = \sum_{j=1}^{m} (Yj_i)\, v_j\,, \quad i = 1, 2, \cdots, n\,, \tag{15.19}$$

where Yj_i is the ith component of the column vector Yj.

Based on eq. (15.19), the problem dimensionality m can be reduced by constructing a k-dimensional approximation to Xi, Xi^k, in terms of the first k PCs:

$$Xi^k = \sum_{j=1}^{k} (Yj_i)\, v_j\,, \quad i = 1, 2, \cdots, n\,. \tag{15.20}$$

The index k of the approximation can be chosen according a criterion involving the threshold variance γ, where

$$\left(\sum_{i=1}^{k} \lambda_i\right) / \left(\sum_{i=1}^{m} \lambda_i\right) \geq \gamma\,. \tag{15.21}$$

The eigenvalues of C represent the variances of the PCs. Thus, the measure $\gamma = 1$ for $k = m$ reflects a 100% variance representation. In practice, good approximations to the overall variance (e.g., $\gamma > 0.7$) can be obtained for $k \ll m$ for large databases.

For such a suitably chosen k, the smaller database represented by components $\{Xi^k\}$ for $i = 1, 2, \cdots, n$ approximates the variance of the original database A reasonably, making it valuable for cluster analysis.

As we show below, the singular value decomposition can be used to compute the factorization of the covariance matrix C when the 'natural scaling' of eq. (15.6) is used.

15.4.2 Data Compression Based on the Singular Value Decomposition (SVD)

SVD is a procedure for data compression used in many practical applications like image processing and cryptanalysis (code deciphering) [296, for example]. Essentially, it is a factorization for rectangular matrices that is a generalization of the eigenvalue decomposition for square matrices. Image processing techniques are common tools for managing large datasets, such as digital encyclopedias, or images transmitted to earth from space shuttles on limited-speed modems.

SVD defines two appropriate *orthogonal coordinate systems* for the domain and range of the mapping defined by a rectangular $n \times m$ matrix A. This matrix maps a vector $x \in \mathcal{R}^n$ to a vector $y = Ax \in \mathcal{R}^m$. The SVD determines the orthonormal coordinate system of $\mathcal{R}^n$ (the columns of an $n \times n$ matrix U) and the orthonormal coordinate system of $\mathcal{R}^m$ (the columns of an $m \times m$ matrix V) so that A is diagonal.

The SVD is used routinely for storing computer-generated images. If, a photograph is stored as a matrix where each entry corresponds to a pixel in the photo, fine resolution requires storage of a huge matrix. The SVD can factor this matrix and determine its *best rank-k approximation*. This approximation is computed not as an explicit matrix but rather as a sum of k outer products, each term of which requires the storage of two vectors, one of dimension of n and another of dimension m ($m+n$ storage for the pair). Hence, the total storage required for the image is reduced from nm to $(m + n)k$.

The SVD also provides the *rank of A* (the number of independent columns), thus specifying how the data may be stored more compactly via the best rank-k approximation. This reformulation can reduce the computational work required for evaluation of the distance function used for similarity or diversity sampling.

SVD Factorization

The SVD decomposes the real matrix A as:

$$A = U\Sigma V^T, \tag{15.22}$$

where the matrices U $(n \times n)$ and V $(m \times m)$ are orthogonal, i.e., $UU^T = I_{n\times n}$ and $VV^T = I_{m\times m}$. The matrix Σ $(n \times m)$ contains at most m nonzero entries $(\sigma_i, i = 1, \cdots, m)$, known as the *singular values*, in the first m diagonal elements:

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \sigma_2 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & \sigma_r & \cdots & \cdots & \cdots & 0 \\ \vdots & & & & & & & \\ 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \sigma_m \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & & & & & & & \\ 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix} \tag{15.23}$$

where

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \ldots \geq \sigma_m \geq 0 \,.$$

The columns of U, namely $u_1, \ldots, u_n$, are the *left singular vectors*; the columns of V, namely $v_1, \ldots, v_m$, are the *right singular vectors*. In addition, $r =$ rank of $A =$ number of nonzero singular values. Thus if $r \ll m$, a rank-r approximation of A is natural. Otherwise, we can set k to be smaller than r by neglecting the singular values beyond a certain threshold.

Low-Rank Approximation

The rank-k approximation to A can be obtained by noting that A can be written as the sum of rank-1 matrices:

$$A = \sum_{j=1}^{r} \sigma_j \, u_j \, v_j^T \,. \tag{15.24}$$

The rank-k approximation, A_k, is simply formed by extending the summation in eq. (15.24) from 1 to k instead of 1 to r. In practice, this means storing k left singular vectors and k right singular vectors. This matrix A_k can also be written as

$$A_k = \sum_{j=1}^{k} \sigma_j \, u_j \, v_j^T \; = \; U \Sigma_k V^T \tag{15.25}$$

where

$$\Sigma_k = \text{diag} \left(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0 \right).$$

This matrix is closest to A in the sense that

$$\| A - A_k \| = \sigma_{k+1}$$

for the standard Euclidean norm.

Recall that we can express each Xi as:

$$\text{Row } \; i \; \text{ of } (A) = (A^T \, e_i)^T \,,$$

where e_i is an $n \times 1$ unit vector with 1 in the ith component and 0 elsewhere. Using the decomposition of eq. (15.24), we have:

$$A^T\, e_i = \sum_{j=1}^{r} \sigma_j\, v_j\, u_j^T\, e_i = \sum_{j=1}^{r} (\sigma_j\, u_{j_i})\, v_j .$$

The SVD transforms this row vector to $[(A_k)^T\, e_i]^T$, where:

$$(A_k)^T\, e_i = \sum_{j=1}^{k} (\sigma_j\, u_{j_i})\, v_j . \tag{15.26}$$

Projection

This transformation can be used to project a vector onto the first k principal components. That is, the projection matrix $P_k = \sum_{j=1}^{k} [v_j v_j^T]$ maps a vector from m to k dimensions. For example, for $k = 2$, we have:

$$\begin{aligned} P_2 A^T\, e_i &= \sum_{j=1}^{r} (v_1\, v_1^T + v_2\, v_2^T)(\sigma_j\, u_{j_i})\, v_j \\ &= (\sigma_1\, u_{1_i})\, v_1 + (\sigma_2\, u_{2_i})\, v_2 \, . \end{aligned} \tag{15.27}$$

Thus, this projection maps the m-dimensional row vector Xi onto the two-dimensional (2D) vector Yi with components $\sigma_1 u_{1_i}$ and $\sigma_2 u_{2_i}$. This mapping generalizes to a projection onto the k-dimensional space where $k \ll m$:

$$Yi^k = (\sigma_1\, u_{1_i}\, , \sigma_2\, u_{2_i}\, , \cdots , \sigma_k\, u_{k_i})\, . \tag{15.28}$$

15.4.3 Relation Between PCA and SVD

It can be shown that the eigenvectors $\{v_i\}$ of the covariance matrix (eq. (15.14)) coincide with the right eigenvectors $\{v_i\}$ defined above when the second scaling (eq. (15.6)) is applied to the database matrix. Recall that this scaling makes all columns have zero means and a variance of unity.

Moreover, the left SVD vectors $\{u_i\}$ can be related to the singular values $\{\sigma_i\}$ and PC vectors $\{Yi\}$ of eq. (15.15) by

$$u_i = A\, v_i / \sigma_i = Yi / \sigma_i \, . \tag{15.29}$$

Therefore, we can use the SVD factorization as defined above (eq. (15.22)) to compute the PCs $\{Yi\}$ of the covariance matrix C. The SVD approach is more efficient since formulation of the covariance matrix is not required.

The algorithm ARPACK [728] can compute the first k PCs, saving significant storage. It requires an order $\mathcal{O}(nk)$ memory and $\mathcal{O}(nm^2)$ floating point operations.

15.4.4 Data Analysis via PCA or SVD and Distance Refinement

The SVD or the PCA projection is a first step in database visualization. The second step refines this projection so that the original Euclidean distances $\{\delta_{ij}\}$ in the m-dimensional space are closely related to the corresponding distances $\{d_{ij}\}$ in the reduced, k-D space. Here,

$$\delta_{ij} \equiv ||Xi - Xj||$$

and

$$d_{ij} \equiv ||Yi - Yj||$$

for all i, j, where the vectors $\{Y_i\}$ are the k-D vectors produced by SVD defined by eq. (15.28).

Projection Refinement

This distance refinement is a common task in distance geometry refinement of NMR models. In the NMR context, a set of interatomic distances is given and the objective is to find the 3D coordinate vector (the molecular structure) that best fits the data. Since such a problem is typically overdetermined — there are $\mathcal{O}(n^2)$ distances but only $\mathcal{O}(n)$ Cartesian coordinates for a system of n atoms — an optimal *approximate solution* is sought.

For example, optimization work on evolutionary trees [1001] solved an identical mathematical problem in an unusual context that is closely related to the molecular similarity problem here. Specifically, the experimental distance-data in evolutionary studies reflect complex factors rather than simple spatial distances (e.g., interspecies data arise from immunological studies which compare the genetic material among taxa and assign similarity scores). Finding a 3D evolutionary tree by the distance-geometry approach, rather than the conventional 2D tree which conveys evolutionary linkages, helps identify subgroup similarities.

Distance Geometry

The distance-geometry problem in our evolutionary context can be formulated as follows. We are given a set of pairwise distances with associated lower and upper bounds:

$$\{l_{ij} \leq \delta_{ij} \leq u_{ij}\}, \quad \text{for} \quad i, j = 1, 2, \ldots, n,$$

where each δ_{ij} is a target interspecies distance with associated lower and upper bounds l_{ij} and u_{ij}, respectively, and n is the number of species. Our goal is to compute a 3D "tree" for those species based on the measured distance/similarity data.

This distance geometry problem can be reduced to finding a coordinate vector Y that minimizes the objective function

$$E(Y) = \sum_{i<j} \omega_{ij} \left(d_{ij}^2(Y) - \delta_{ij}^2 \right)^2 , \tag{15.30}$$

where $d_{ij}(Y)$ is Euclidean distance between points i and j in the vector Y, and the $\{\omega_{ij}\}$ are appropriately-chosen weights.

In the combinatorial chemistry context, we use the same function $E(Y)$ where Y is the vector of $2n$ components, listing the 2D projections of each compound in turn. Details of this data clustering approach are described in [1399, 1402]. Minimization can be performed so that the high-dimensional distance relationships are approximated.

Besides the value of the objective function (eq. (15.30)), a useful measure of the distance approximation in the low-dimensional space is the percentage of intercompound distances $\{i, j\}$ (out of $n(n-1)/2$) that are within a certain threshold of the original distances. We first define the deviations from the targets by a percentage η so that

$$\begin{aligned} |d(Yi, Yj) - \delta_{ij}| \leq \eta\, \delta_{ij} \quad & \text{when} \quad \delta_{ij} > d_{\min}\,, \\ d(Yi, Yj) \leq \tilde{\epsilon} \quad & \text{when} \quad \delta_{ij} \leq d_{\min}\,, \end{aligned} \tag{15.31}$$

where η, $\tilde{\epsilon}$, and $d_{\min}$ are given small positive numbers less than one. For example, $\eta = 0.1$ specifies a 10% accuracy; the other values may be set to small positive numbers such as $d_{\min} = 10^{-12}$ and $\tilde{\epsilon} = 10^{-8}$. The second case above (very small original distance) may occur when two compounds in the datasets are highly similar.

With this definition, the total number T_d of the distance segments $d(Yi, Yj)$ satisfying eq. (15.31) can be used to assess the degree of distance preservation of our mapping. We define the percentage ρ of the distance segments satisfying eq. (15.31) as

$$\rho = \frac{T_d}{n(n-1)/2} \times 100\,. \tag{15.32}$$

The greater the ρ value (the maximum is 100), the better the mapping and the more information that can be inferred from the projected views of the database compounds.

This minimization procedure (projection refinement) is quite difficult for scaled datasets. Experiments with several chemical datasets of size 58 to 27255 compounds show that the percentage of distances satisfying a threshold deviation ρ of 10% (eq. (15.31)) is in the range of 40% [1399, 1402]. Nonetheless, these low values can be made close to 100% with projections onto 10-dimensional space. This is illustrated in Figure 15.4, which shows the percentage of distances satisfying eq. (15.31) for $\eta = 0.1$ as a function of the projection dimension for a database ARTF.

A similar improvement can be achieved with larger tolerances η (e.g., distances that are within 25% of the original values rather than 10%) [1399, 1402].

15.4.5 Projection, Refinement, and Clustering Example

As an illustration, consider the model database ARTF of 402 compounds and $m = 312$ descriptors containing eight chemical subgroups. We have analyzed this database by performing 2D and 3D projections based on the SVD factorization followed by minimization refinement by TNPACK [1121, 1122, 1397] for performance assessment in terms of accuracy as well as visual analysis of the compound interrelationships.

From Figure 15.4 we note that the refinement stage that follows the SVD projection is important for increasing the accuracy in every dimension. Namely, the accuracy is increased by 25–40% in this example.

The 2D and 3D projection patterns obtained for ARTF in Figure 15.5 show the utility of such a projection approach. The resemblance between the 2D and 3D views is evident, and the various 3D views offer different perspectives of the intercompound relationships.

We note that clusters corresponding to individual pharmacological subsets appear very close to each other, though partial overlap of clusters is evident. The *ecdysteroid* group forms a diverse but separate set of points. The *estrogen* class is also clustered and somewhat separate from the others. The strong overlap of the three clusters corresponding to *D1 agonists*, *D1 antagonists*, and *H1 receptor ligands* is reasonable given the relative chemical similarity of these compounds: all act at receptors of the same pharmacological class (i.e., G-protein coupled receptors). Thus, such data compression and visualization techniques can be used as a quick analysis tool of the database structure.

The chemical structures in Figure 15.6 reveal that compounds that are nearer in the projection are more closely related than those that are distant; this is seen when compounds are compared both within the same subgroup and within different subgroup. For example, the two labeled estrogen representatives that are

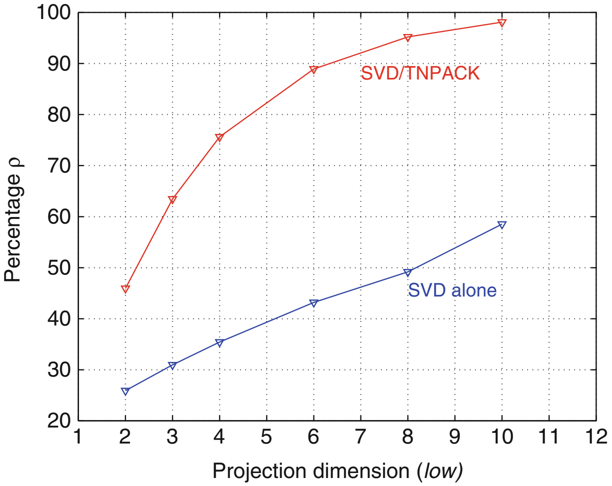


Figure 15.4. Performance of the SVD and SVD/minimization protocols for the ARTF chemical database in terms of the percentage of distances satisfying eq. (15.31) for $\eta = 0.1$ (reflecting 10% distance deviations) as a function of the projection dimension [1399, 1402].

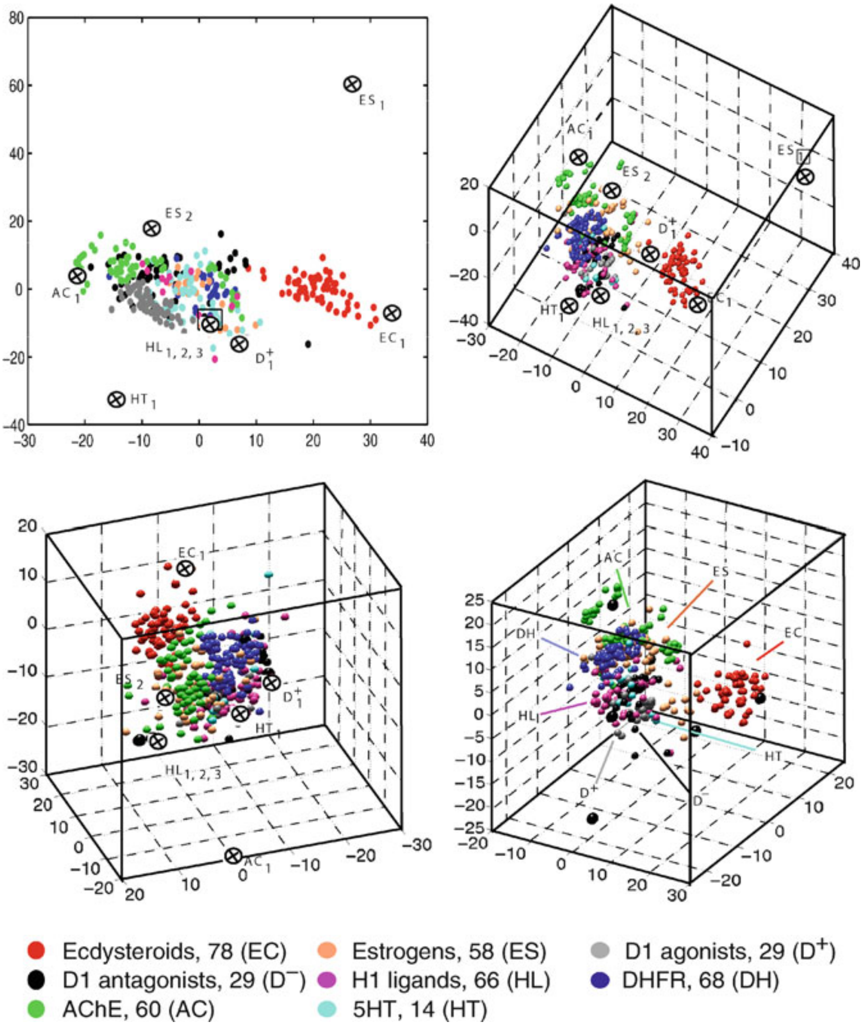


Figure 15.5. Two and three-dimensional projections of the chemical database ARTF of 402 compounds composed of the eight chemical subgroups ecdysteroids (EC), estrogens (ES), D1 agonists (D$^+$), D1 antagonists (D$^-$), H1 ligands (HL), DHFR inhibitors (DH), AchE inhibitors (AC), and 5HT ligands (HT) using the projection/refinement SVD/TNPACK approach [1399, 1402]. Three views are shown for the 3D projection. The accuracy of the 2D projection is about 46% and that of the 3D is 63% (with $\eta = 0.1$); see eq. (15.31). The 2D projection was obtained by refining the 3D projection. The nine chemical structures labeled in the projections are drawn in Figure 15.6.

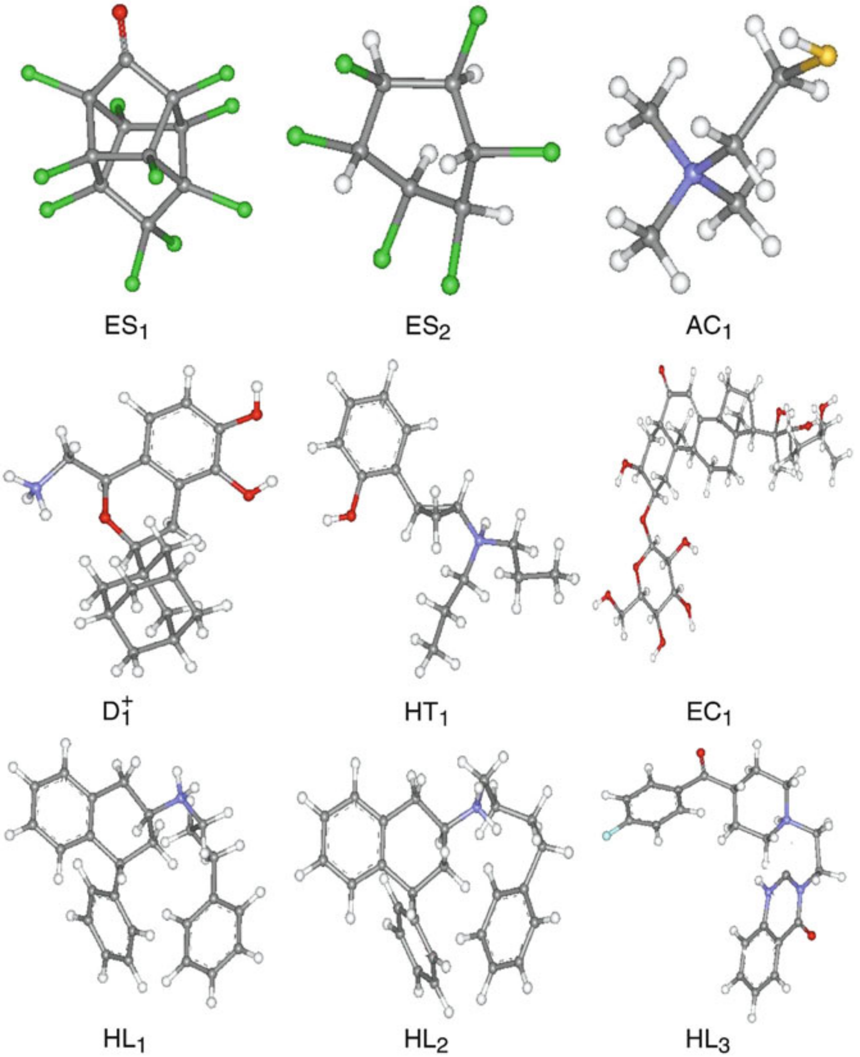


Figure 15.6. Selected chemical structures from the ARTF projection shown in Figure 15.5 reveal similarity of nearby structures and dissimilarity of distant compounds.

distant in the projection appear chemically quite different, while the three clustered H1 ligands appear similar to each other and perhaps to the nearby D1 agonist representative.

An example of a database projection in 2D by the alternative PCA approach followed by distance refinement is shown in Figures 15.7 and 15.8 for 832 compounds from the MDL Drug Data Report (MDDR) database using topological indices. (This work was performed in collaboration with Merck

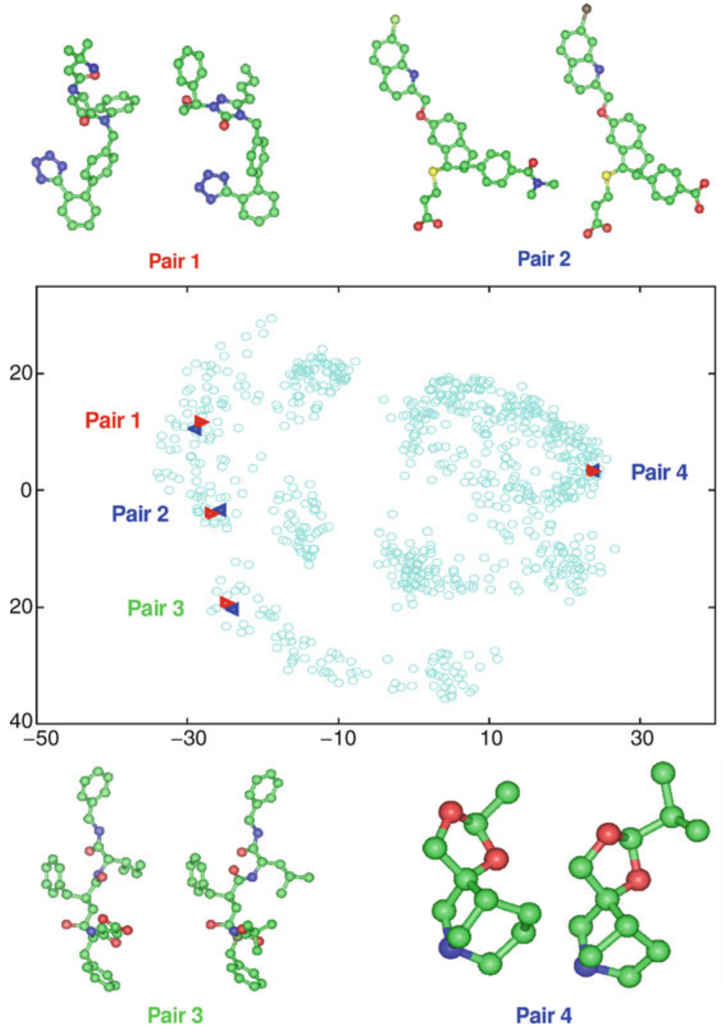


Figure 15.7. 2D projection using PCA for 832 compounds in the MDDR database showing the similarity of four compound pairs that are near in the projection.

Research Laboratories). The accuracy of this projection (the percentage of distances satisfying eq. (15.31) for $\eta = 0.1$) is only 0.2% after PCA and 24.8% after PCA/TNPACK. Figure 15.7 shows that compounds close in the projection appear similar, and Figure 15.8 shows that more distantly related compounds tend to be different. Without knowing the grouping of these compounds according to bioactivity, the clusters identified in Figure 15.8 suggest a ‘diversity subset’ consisting of a few members from each cluster.

The approach described here appears promising, but further work is required to make the technique viable for very large databases.

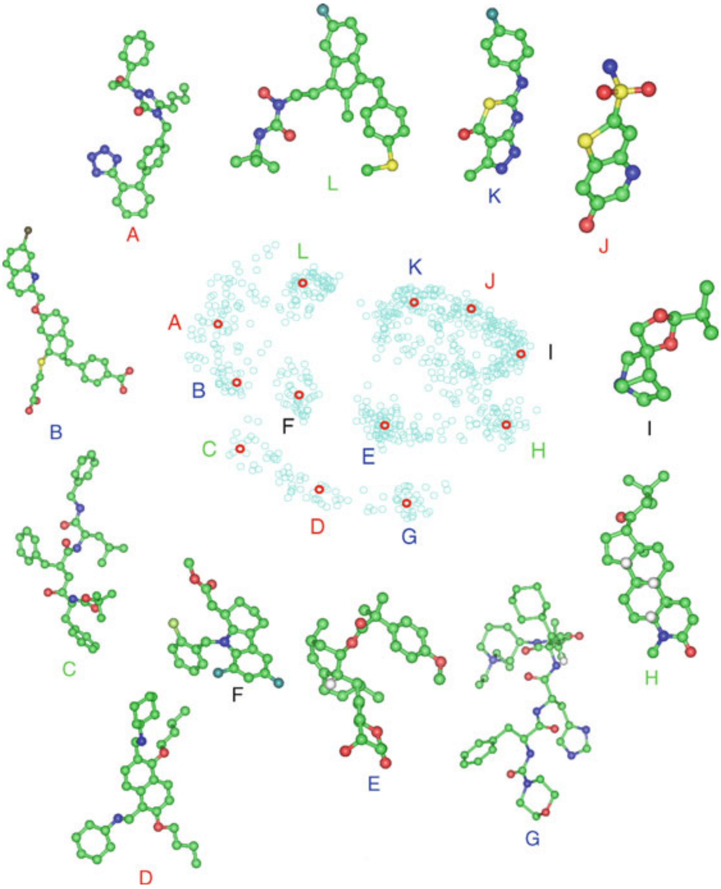


Figure 15.8. 2D projection using PCA for 832 compounds in the MDDR database showing the diversity of compounds that represent different clusters in the projection (distinguished by letters). A representative subset may thus consist of one or only a few members from each cluster.

15.5 Future Perspectives

Similarity and diversity sampling of combinatorial chemistry libraries is a field in its infancy. The choice of descriptors as well as metrics used to define similarity and diversity are empirical and perhaps application dependent. Thus, many challenges remain for future developments in the field, and the added involvement of mathematical scientists and new approaches borrowed from allied disciplines might be fruitful.

Developments are needed for formulation of descriptor sets, rigorous mathematical frameworks for their analysis, and efficient algorithms for very large-scale problems based on statistics, cluster analysis, and optimization. The algorithmic challenge of manipulating large datasets might also explain the tendency toward smaller and focused libraries [555]; still, as argued in [621], this assumed defeat is premature!

The central assumption of structure/activity relationships of course remains a challenge to validate, develop, and further apply.

More broadly, structure-based drug design is likely to increase in importance as many more protein targets are identified and synthesized [1301], and as modeling programs improve in their ability to predict binding affinities of certain ligands (e.g., peptide-like) that share chemical groups with macromolecules, the focus of many biomodeling packages. The difficulty in determining membrane protein structures continues to be a limitation since membrane receptors are important pharmacological targets.

While perhaps not the dominant technique, it is clear that structure-based drug design will be an important component of drug modification and optimization after available leads have been generated. The search for the needle in the haystack (i.e., a successful drug) will likely be guided by the steady light generated by computer modeling. And, with additional genetic and genomic screening, disease treatment is likely to move forward to a new phase of greater scientific precision and success.