SCIENCE CHINA
Chemistry

• FEATURE ARTICLES •

January 2013 Vol.56 No.1: 71–85
doi: 10.1007/s11426-012-4761-0

Chemomics and drug innovation

XU Jun[1*], GU Qiong[1], LIU HaiBo[2], ZHOU JiaJu[1], BU XianZhang[1], HUANG ZhiShu[1], LU Gui[1], LI Ding[1], WEI DongQing[3], WANG Ling[1] & GU LianQuan[1]

[1]*Research Center for Drug Discovery (RCDD), School of Pharmaceutical Sciences and Institute of Human Virology, Sun Yat-Sen University, Guangzhou 510006, China*
[2]*Institute of Medicinal Plant Development, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing 100194, China*
[3]*Biostatistics College of Life Sciences and Biotechnology, Shanghai Jiaotong University, Shanghai 200240, China*

Received August 6, 2012; accepted August 14, 2012; published online October 15, 2012

Chemomics is an interdisciplinary study using approaches from chemoinformatics, bioinformatics, synthetic chemistry, and other related disciplines. Biological systems make natural products from endogenous small molecules (natural product building blocks) through a sequence of enzyme catalytic reactions. For each reaction, the natural product building blocks may contribute a group of atoms to the target natural product. We describe this group of atoms as a chemoyl. A chemome is the complete set of chemoyls in an organism. Chemomics studies chemomes and the principles of natural product syntheses and evolutions. Driven by survival and reproductive demands, biological systems have developed effective protocols to synthesize natural products in order to respond to environmental changes; this results in biological and chemical diversity. In recent years, it has been realized that one of the bottlenecks in drug discovery is the lack of chemical resources for drug screening. Chemomics may solve this problem by revealing the rules governing the creation of chemical diversity in biological systems, and by developing biomimetic synthesis approaches to make quasi natural product libraries for drug screening. This treatise introduces chemomics and outlines its contents and potential applications in the fields of drug innovation.

chemomics, chemoinformatics, bioinformatics, drug innovation, biomimetic synthesis

1 Introduction

The Human Genome Project was completed in 2003. The achievement is a great leap in life sciences. It was believed that the rest of the work was to discover genes and related functions [1]. Genomics involves acquiring and processing large amounts of data, which leads to the flourishing of bioinformatics. Bioinformatics mainly studies macromolecular structures, gene annotations, and genome experimental data mining [2]. Bioinformatics predicts biological functions at the macromolecular level and considers the gene as a fundamental unit of genetic material. Proteins are the functional molecules that are responsible for executing biological functions; their building blocks are amino acids. The properties of a gene are determined by the chemical properties of nucleotides; the properties of a protein are determined by amino acids *per se*.

Chemoinformatics mainly studies relations between the chemical structures and properties of molecules, the interactions between small molecules and macromolecules, and the construction of chemical libraries.

Bioinformatics and chemoinformatics overlap more and more. In chemistry, atoms are the building blocks of a molecule. The rules of combining atoms to make a molecule have been manifested in the Periodic Table of the Elements. In biology, a genome has complete instructions for making cells. However, a cell has to provide a set of endogenous molecules as building blocks to synthesize more complicated molecules (small or macro molecules) for its survival and proliferation. Each building block can contribute a group of atoms to a target molecule that has specific bio-

*Corresponding author (email: junxu@biochemomes.com)

logical functions. We describe this group of atoms as a chemoyl (chemical unit). The complete set of the chemoyls is a chemome. Chemomics is the study of the chemomes of organisms. Chemomics mainly determines the chemoyls and their corresponding building blocks in organisms, and reveals the common rules in biochemical synthetic pathways. The goal of chemomics is to come out with a "biological periodic table", which manifests the fundamental rules of combining chemoyls [3]. However, a rectangular table may be no longer able to describe the relationship among these chemoyls [4], because the relationships between them may be unprecedentedly sophisticated. How do we derive the chemoyl from biomolecules or natural products? What are nature's synthetic principles for making biomolecules from chemoyls? These questions should be answered by the joint efforts of chemoinformatics, bioinformatics, and other related disciplines; chemomics exists for these purposes.

Since the triplet genetic codons were deciphered in the 1960s, a "biological periodic table" has been pursued. For example, Biro designed a rectangular "biological periodic table" using amino acids as biological elements in order to find the relationship between the pattern of nucleotide codon combinations and the physicochemical properties of amino acids. In this table, nucleotides control the pharmacophore properties of amino acids. If a central nucleotide codon is purine, the hydrophilicity of an amino acid side chain will be determined; if the central nucleic acid codon is a pyrimidine, the pH of the amino acid side chain (positively or negatively charged) will be determined [5].

2 Building blocks for natural products

It is easy to recognize building blocks for biological macromolecules. For example, a protein is made by linking a group of amino acids with peptide bonds. The recipes for building biological macromolecules are encoded in a cell in order to maintain the hereditability of the organism's trait. However, it is difficult to recognize building blocks for natural products (Generally speaking, biological macromolecules are also natural products. In this article, natural products refer to small molecules made in living things). The recipes for making natural products are not explicitly preserved in a cell. However, natural products, including amino acids and nucleic acids, are still made from a set of simpler building blocks (chemoyls) and a group of enzyme-catalytic reactions. What should be done is to identify the chemomes and their related biochemical reactions. In the past decades, a number of models have been proposed to interpret natural products building blocks.

2.1 Pharmacophore model

The pharmacophore concept was suggested by medicinal chemists. A pharmacophore represents a key feature resulting in a molecule possessing biological or therapeutic activity. If this feature is missing, the efficacy is lost. Along with the development of computer-aided drug design, the pharmacophore concept has evolved into a three-dimensional spatial configuration (Figure 1) [6].

Therefore, the pharmacophore model has nothing to do with natural product construction.

2.2 Privileged structure model

The concept of a privileged structure was proposed by Evans and co-workers [7] in 1988. A privileged structure is a chemical scaffold with good drug-like properties and versatile binding properties, and provides potent and selective ligands for a range of different biological targets through the modification of functional groups. A privileged structure leads to more drug-like compound libraries and leads [8]. Four examples are listed in Table 1 [9].

There is no precise definition of a privileged structure. Therefore, the concept of a privileged structure is not associated with natural product construction either. Medicinal chemists use the concept to describe some common skeleton features for drug leads against a set of targets. This concept has not been investigated systematically.

2.3 Synthon model

The synthon model was initiated by Nobel Laureate Elias J. Corey in 1967. Creating a rigorous logical reasoning system, Corey developed a protocol to reveal all possible precursors (synthons) for a given molecule based on recursively applying known reaction rules. The reaction rules are operations (forming bonds, breaking bonds, changing bond types, etc) on molecules if they satisfy the prerequisite structural patterns of the rules. If a molecular structure is operated on by a rule, it will result in synthons. If the synthon is an unavailable molecule, reaction rules will be recursively applied until no rule can be applied or further synthons are found. The reasoning process results in synthon pathways called retrosynthetic trees (Figure 2) [10].

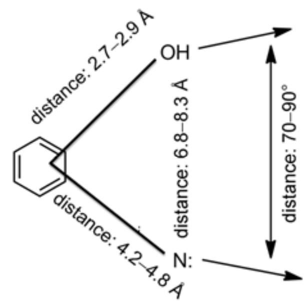
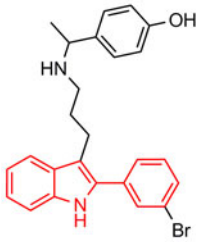
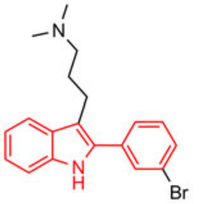
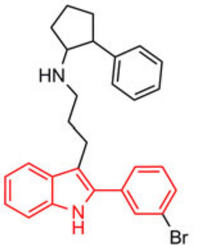
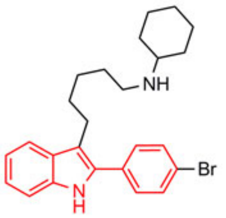


Figure 1 The three-dimensional pharmacophore model.

Table 1 Examples of privileged structures and their targets

Structure	Target
	NPY5
	5HT-6
	CCR3, CCR5
	MCR4

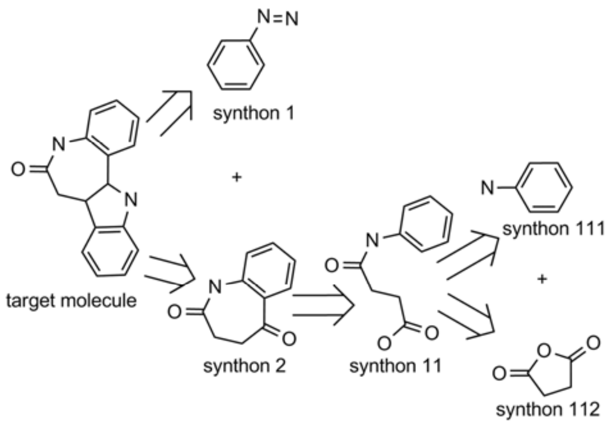


Figure 2 Corey's synthons and retrosynthetic tree.

Organic synthesis is the art of making chemicals. It was Corey who made the art a rigorous reasoning process. The reasoning system consists of i) a set of synthons that are like chess pieces on a chessboard and, ii) a set of chemical reaction rules that are like the rules of chess. Corey's protocol was implemented via computer program; these were successful examples of early artificial intelligence applications.

Building on Corey's work, Schreiber and co-workers proposed target oriented synthesis (TOS) and diversity oriented synthesis (DOS) [11] in order to generate more diverse and natural product-like compounds for drug screening or chemical biology.

2.4 Fragment-based model

Two decades ago, the poor hit rate of high throughput screening (HTS) was attributed to the fact that bigger molecules (MW~500 Daltons) may not fit in smaller protein binding pockets. Thus, compounds were divided into smaller fragments (MW < 250 Daltons) for HTS campaigns. This strategy is called the fragment-based approach [12]. The strategy resulted in smaller hits with weak binding affinities. The "fragment hits" were connected by chemical syntheses to enhance activity. Two examples are demonstrated in Table 2 [13].

What is interesting about this approach is that it finds a way to reveal nature-selected fragments through bioassays instead of subjective definition. However, the approach is not concerned with how a natural product is made.

2.5 From synthon to chemoyl

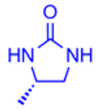
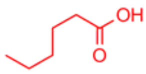
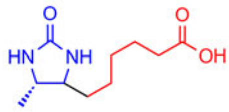
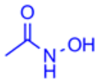
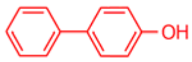
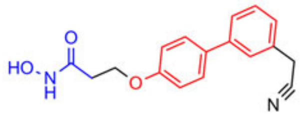
Most of the above-mentioned models focus on finding the building blocks of natural products or bioactive molecules. Nature's secret of making natural products was not systematically investigated.

The pharmacophore model reflects the complementary geometric and physical properties (such as charge, hydrophilicity, etc.) within arrangements involving a ligand and receptor. A pharmacophore is not related to the structural building blocks of natural products. The privileged structure model manifests the relations between small molecular scaffolds and target selectivity. The structural fragment model emphasizes fishing for smaller molecular hits (fragments) through bioassays and connecting them by chemical modification for better efficacies.

The synthon approach tells us how to derive building blocks (synthons) for a complicated natural product, and how to make it from the synthons—simpler molecules. The synthon protocol is recursive with the termination conditions of finding available synthons and trying all qualified reaction rules. This results in the combinatorial explosion of the retrosynthetic tree. However, the synthons are not necessarily the building blocks from which a cell makes natural products.

Inspired by the synthon approach, chemomics applies the rules from enzyme catalytic reactions on natural products to derive chemoyls and related building blocks. The building blocks are associated with biological building blocks, which are endogenous molecules. Thus, we can make natural product-like compounds (quasi natural products) by mim-

Table 2 Examples for fragment-based drug discovery

Target	Fragment 1	Fragment 2	Combined molecule
Avidin	 K_i = 34 μM	 K_i = 260 μM	 K_i = 0.0004 nM
Stromelysin	 K_d = 17 mM	 K_d = 280 mM	 K_d = 15 nM

icking a synthetic process occurring in a cell. To generate larger structural diversity space and reduce synthetic costs, we can also introduce chemical building blocks.

Nature makes amino acids from inorganic materials (water, nitrogen gas, oxygen gas, and carbon dioxide) by coupling the carbon fixation and nitrogen fixation reaction networks [14] (Figure 3). Through photosynthesis, plant cells make saccharides, and microorganisms make ammonia. Glycometabolic pathways coupled with the carbon fixation reaction produce glutamic acid, through which other amino acids are synthesized.

Endogenous molecules are constructed in an atomic group termed a building block in natural product chemistry [15]. In synthetic chemistry, small molecules that deliver a group of atoms to a target molecule are also called building blocks. To avoid confusion, we adopt the definition of synthetic chemistry, and the atomic group to be delivered to a target compound is termed a chemoyl. A number of chemoyls and their building block examples are listed in Table 3.

Therefore, a chemoyl is a group of atoms that can be delivered by an endogenous molecule (building block) to a target natural product. A chemoyl can have many building blocks. One building block can deliver a different chemoyl

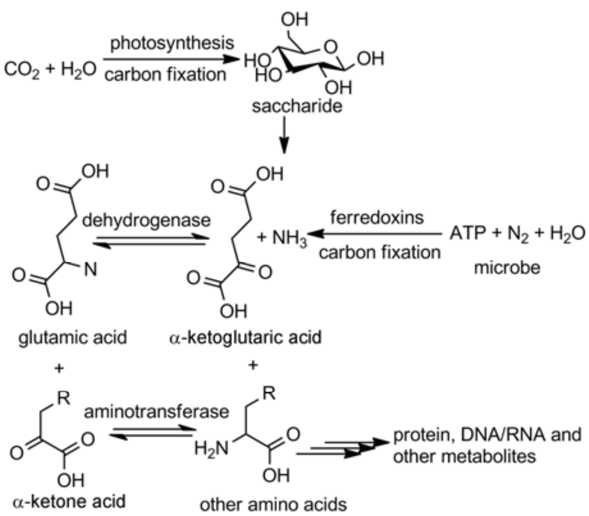
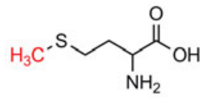
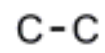
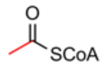
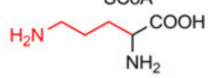
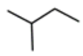
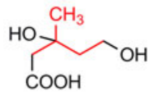
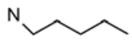
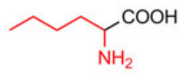
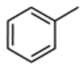
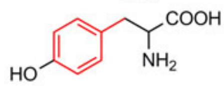
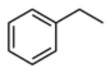
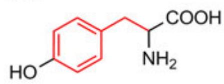
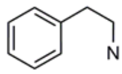
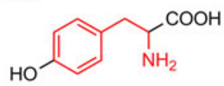
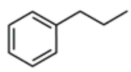
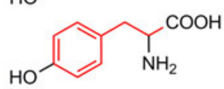
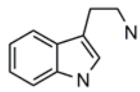
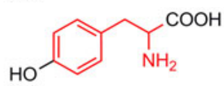


Figure 3 Making organic compounds from inorganic compounds by coupling the carbon fixation and nitrogen fixation reaction networks.

Table 3 Natural product chemoyls and their building blocks

Chemoyl*	Notatison	Example building blocks
	C1	
	C2	
	C4N	
	C5	
	C5N	
	C6C1	
	C6C2	
	C6C2N	
	C6C3	
	indole.C2N	

*A chemoyl is the building block's contribution to a target natural product. For example, X–C is a general form for a one-carbon chemoyl. Its corresponding building blocks will deliver one carbon (saturated or unsaturated) atom to a target natural product molecule.

to a target compound, depending on which enzyme catalytic reaction is coupled. Therefore, chemoyls and building blocks have a many-to-many relationship. The relationship is enzyme catalytic reaction *per se*. Take tyrosine as an example (Table 3): depending on which enzyme is coupled, tyrosine can contribute five different types of chemoyls (C6C1, C6C2, C6C2N, C6C3, and Indole.C2N) to a target natural product.

The missions of chemomics are i) to reveal chemoyls, corresponding building blocks (endogenous or exogenous compounds), and related reactions (enzyme catalytic or non-enzyme catalytic reactions) to generate natural products or quasi natural products; ii) to develop protocols for creating quasi natural product libraries or target-specific libraries; iii) to discover chemoyl evolution rules and the laws of chemoyl combinations. These studies will provide new technologies for sustainably acquiring, enhancing, and using natural products.

Natural products are materials which support organisms' survival and proliferation. In life's evolutionary history, there must have been many structurally diversified natural products produced by a species. Some of the natural products are helpful for survival and are favored; others were adverse or useless and disappeared. However, these extinct compounds can be useful for drug innovation by regulating other life systems. If the secrets of biosynthesis were deciphered, extinct natural products could be re-discovered and re-produced as new resources for drug discovery today.

On the other hand, if we can modify the natural synthesis protocols, for example, by changing the enzyme catalytic sequence, replacing enzymes with chemical catalysts, replacing endogenous reactants with chemicals, or hybridizing natural building blocks with unnatural building blocks, we may produce significant quasi natural product diversity for drug screening campaigns. This can be termed biomimetic synthesis (Figure 4).

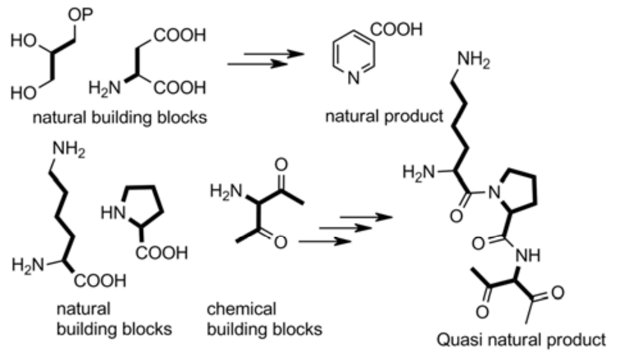


Figure 4 By hybridizing natural and unnatural building blocks, biomimetic syntheses can produce natural or quasi natural products. Bold substructures are chemoyls.

Over 60 years of drug innovation data show that natural products and quasi natural products are drug-like. Natural products and quasi natural products are therefore strategic resources for drug innovation and the life sciences [16].

3 Principles and technologies of chemomics

Natural products are produced through enzyme catalytic reactions in organisms from natural building blocks. Natural macromolecules (such as proteins) are made through enzyme catalytic reactions supervised by other biomolecules, such as mRNA or proteins. The goal of the supervised biosyntheses is to avoid synthetic errors. Moreover, synthetic protocols have to be encoded in genes for heredity. Most natural products are made through enzyme catalytic reactions without a supervisor. After a compound is produced, its synthetic pathway is "forgotten".

As shown in Figure 5, acetyl groups cannot be recognized as the building blocks of fatty acids. By examining the structure of the fatty acid, it is difficult to figure out what is the synthetic pathway associated with the activation, elongation, and termination.

3.1 Deciphering chemomes from natural products

Natural products seemed complicated to synthesize. In order to make them, we have to understand how nature assembles them from natural building blocks *in vivo*. Retro-synthetic analysis is a useful approach for deciphering chemoyl building blocks from natural products. Epothilone, for example, can be simplified into a number of chemoyls by applying the synthon of polyketide synthase (PKS) (Figure 6).

Deciphering chemomes from natural products relies on the following: i) A set of synthons derived from enzymatic reactions for making natural products; ii) An endogenous compound database, in which a compound is a natural product building block, associated with an enzyme (or enzymes) and a

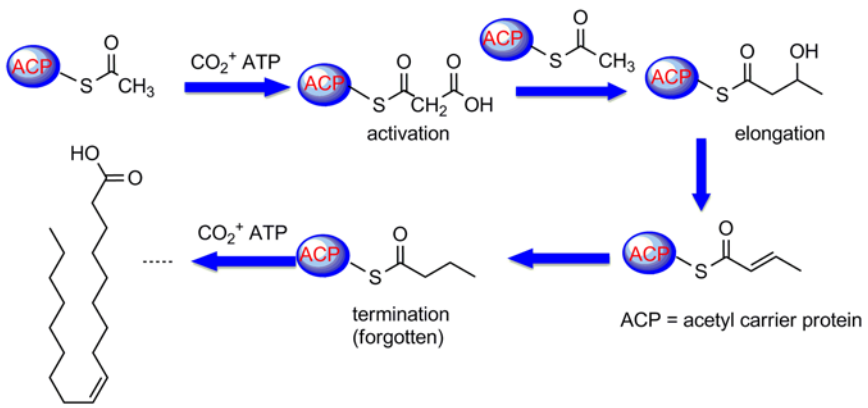


Figure 5 The fatty acid synthetic pathway and its chemoyls are "forgotten".

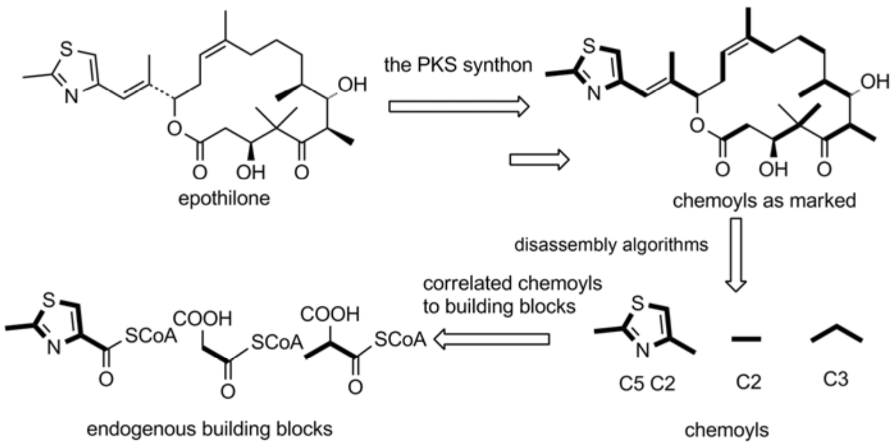


Figure 6 Deciphering chemoyls from epothilone.

chemoyl (or chemoyls); iii) The relationships between building blocks and chemoyls can be multiple-to-multiple; iv) A set of graph theoretic algorithms, with which the chemical structures of natural products are divided into chemoyls that are further mapped onto a set of endogenous (or non-endogenous) building blocks.

For a natural product, there can be more than one way to derive chemoyls, and more than one way to map the chemoyls to building blocks. Natural products, chemoyls, and building blocks form complicated networks.

Deciphering chemomes requires a large amount of metabolic chemical reaction data, which have been compiled in many resources, such as, "the biochemical pathways: biochemistry and molecular biology navigation diagram" [17], the KEGG database [18], and the BioPath database [19]. In the BioPath database, the reaction centers have been annotated and are searchable. Therefore, BioPath is more suitable for chemomics studies.

3.2 The principles governing assembly of natural products from endogenous building blocks

Both biopolymers and natural products are assembled from endogenous building blocks in cells. They use the same set of reagents and solution, the same set of synthetic devices (enzymes), and the same set of enzyme-catalyzed reactions (mainly hydrolysis, lysis, transference, isomerization, oxidation, reduction, and linking). The difference is that a biopolymer is assembled in a precisely supervised platform (e.g., mRNA), whereas a natural product is assembled more flexibly.

Assembly of a biological macromolecule involves programmed synthesis, which is strictly guided by a specific gene with an extremely low error rate. The error rate is between one and ten in a billion [20]. The low error rate ensures the species' genetic integrity and warrants flexibility in response to environmental stresses. This explains bacterial drug resistance as originating from the accumulated biological macromolecule assembly errors. Antibiotic abuse may only play a role in accelerating drug resistance. If a bacterial cell reproduction error rate is 1% (mutation rate), and the cell reproduces twice per hour, then it only takes 10 hours for a bacterial cell to produce millions (2^{20} = 1,048,576) of cells. Therefore, at least one mutant will be created from ten thousand bacterial cells. When the flora with mutants grows further, drug resistant mutants will increase significantly in number [21]. This explains why bacteria have resistance to yet-to-be-discovered drugs [22].

The general processes of assembling biological macromolecules (for example, proteins) can be summarized as follows:

Assembly platform A (ribosomal guided synthesis). Assembly process: activating amino acids→tRNA taking activated amino acids (aaa) to peptide binding sites→connecting aaa to the (n+1)th codon at an mRNA with a hydrogen bond →forming a peptide bond with the current amino acid, and releasing tRNA→shifting the mRNA to the (n + 2)th codon →repeating the process until the last amino acid is assembled.

Assembly platform B (non-ribosomal guided synthesis). Assembly process: there are three enzyme chains associated with three operations (activation, elongation, and termination); selecting amino acids→activating amino acids→covalently binding amino acids onto phosphopantetheines→forming peptide bonds→post-assembly modifications [23].

General processes of assembling natural products can be summarized as follows: starting with endogenous primary or secondary metabolites and co-factors (for examples, ATP, CoA)→activating a building block→elongating the activated building block with a new building block→repeating the elongations till the environment disallows→terminating the process (see Figure 5). The process contains a group of enzymes, which form a biochemical reaction network.

In order to maintain the stability of biological systems, the building blocks are usually stable in cells, and will be activated when elongations are demanded. The activation is conducted through a carrier. As shown in Figure 7, Chemoyl C_2 is carried by SCoA in a form of acetyl, which is activated at the carbon atom of the methyl group by adding carbon dioxide. The activated carbon atom is ready to covalently bind to another chemoyl. The adenosine of SCoA supplies energy for the coming chemical reaction. SCoA can be compared to a missile, which contains three compartments: warhead (red), recognition (blue), and engine (black).

Although the assembly of biological macromolecules is precisely programmed, the assembly of natural products is not well programmed. This flexibility allows organisms to have more choices to response to environmental stresses. Figure 8 demonstrate an example of flexibility in the assembly of building blocks. Combining three building blocks with two chemoyls (C_2 and C_3) and three enzyme-catalyzed reactions can produce totally different natural products. The results can be arrived at simply by regulating/adding different combinations of enzymes, building blocks, and concentrations.

Building block A can be assembled with other building blocks at one end. It is used to assemble linear natural products. Building blocks B and C, however, can be assembled with other building blocks at multiple sites. They are used to assemble dendric or cyclic molecules. A fatty acid, a linear molecule, is generated by recursively assembling block A. The size of a natural product is regulated by enzymes.

Genes or a number of biological macromolecules synergistically make proteins from amino acids through aminations. It is more flexible to assemble a natural product in response to environmental stresses (Figure 9). Usually, cysteine and valine can be connected by a peptide bond. However, the two amino acids are fused to form a β-lactam

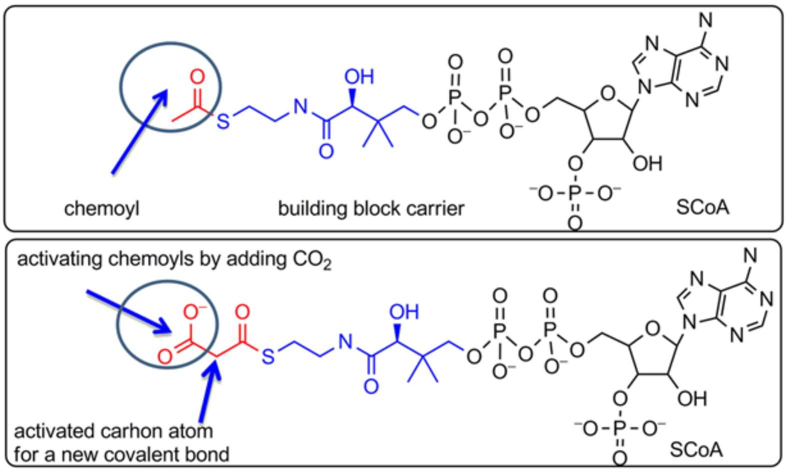


Figure 7 Activating a building block at its chemoyl.

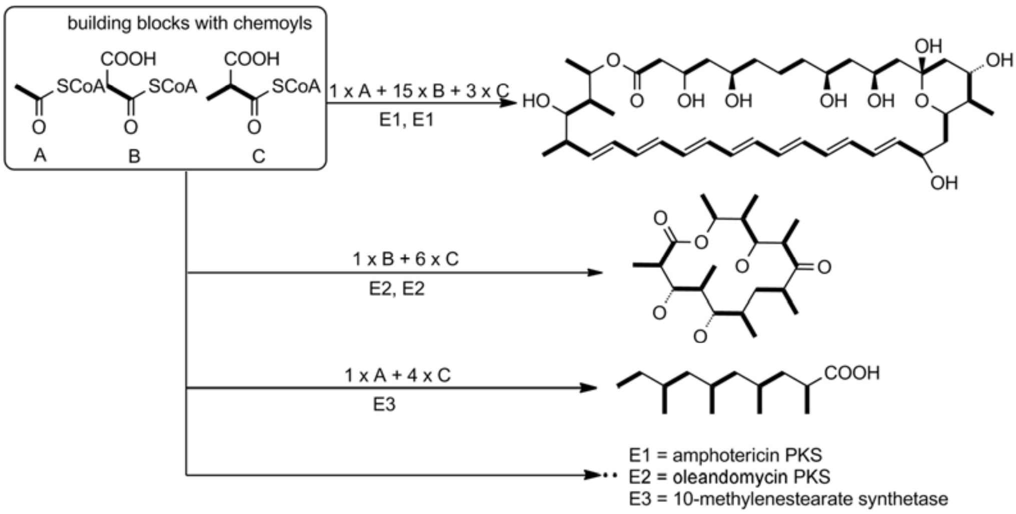


Figure 8 The assemblies of natural products from building blocks.

(penicillin) as an inhibitor to stop the condensation of two D-alanines crucial to other bacteria. The assembly of penicillin is an example of targeted synthesis in nature, and inspires us to find new ways of creating chemical diversity.

3.3 The evolution of chemical diversity in nature

The major stages in the life cycle are birth, growth, maturity, decline, and death. The types of compounds and their concentrations change at each stage. This knowledge is a guide for when, where, and how Chinese herbs should be harvested. A Chinese proverb says, Herba Artemisiae Capillaris is a herb in March, a weed in April, and firewood after May. For example, shikimic acid, a precursor of Tamiflu, can be found in many plants and microbes. However, it is only harvested from a few resources, such as the seeds of a small number of winteraceae plants. Metabolic pathways (Figure 10) show that shikimic acid is an intermediate for a number of other primary metabolites (such as amino acids). It is therefore not surprising that the amount of shikimic acid declines when a plant becomes mature. Therefore, it is critical to understand shikimic acid metabolic pathways in order to effectively harvest shikimic acid from plants.

In the 1980s, it was noticed that natural product production is related to the developmental history of an individual organism. For example, the alkaloids quinolizidine, coniine, and solanine are found in significant amounts in the buds of lupin, hemlock and potato [24]. These findings should be the scientific foundation of Traditional Chinese Mediane (TCM) herb harvest and processing. Unfortunately, conventional TCM harvesting and processing rely on empirical rules or traditions. In order to efficiently and sustainably utilize TCM resources, the harvest season or time, planting approach, extraction technology, and biodiversity protection strategy should be determined from the relationships between natural product building blocks, chemoyls, ontogene-

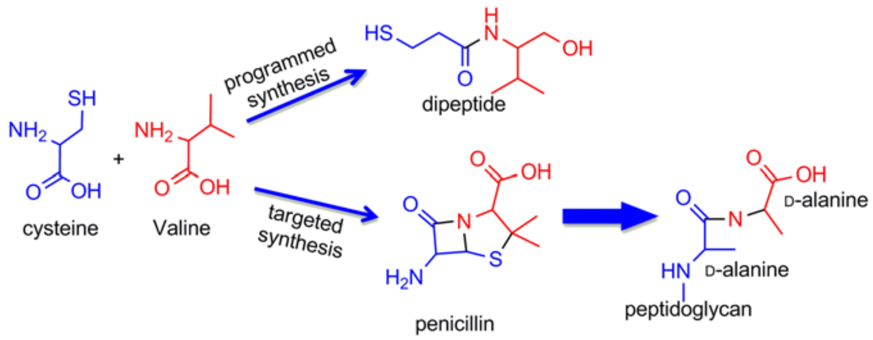


Figure 9 Programmed synthesis and targeted synthesis.

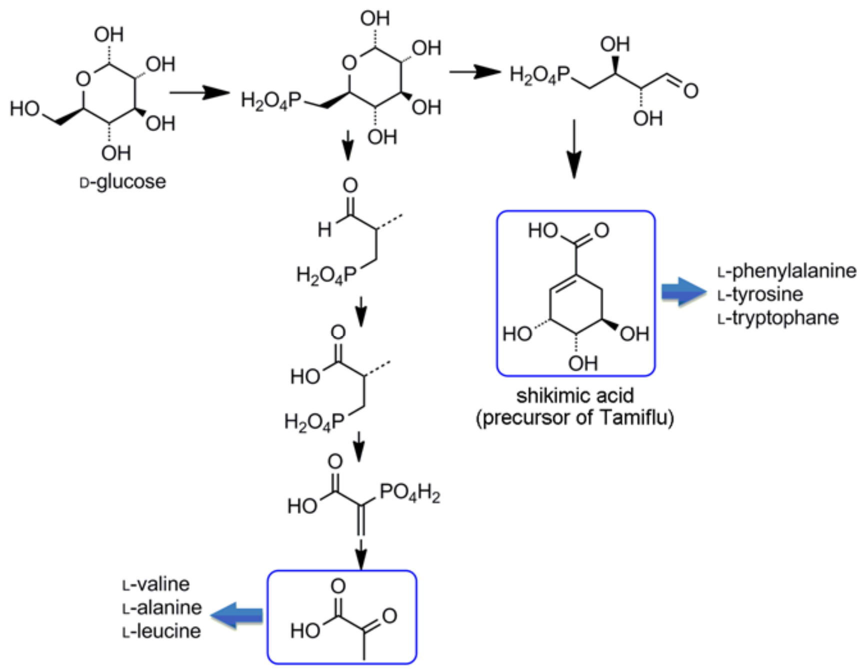


Figure 10 D-glucose and shikimic acid related metabolic pathways.

sis, and metabolic reaction pathways.

On the other hand, there is a relation between the diversity development of biochemical building blocks and the evolution of the species. The relation is the scientific foundation of TCM herb composition rules, such as, a herb's nature, taste, action, compatibility, and replacement. It is also the rationale for plans for the protection of endangered TCM species. Caffeic acid, for example, is widely found in plants. As the species group develops from lower to more advanced, the number of caffeic acid derivatives grows through caffeic acid autopolymerization and heteropolymerization, and demonstrates chemical diversity (Figure 11). Complicated caffeic acid metabolites demand specialized enzymes for production; for this, the organisms are regulated more precisely. If the evolutionary diagram of caffeic acid metabolites is clear, new ways of generating novel chemical diversity for drug screening can be discovered.

4 Chemomics and drug innovation

Drug discovery begins with active compound screening. Success relies on the screening technology and the druggability and structural diversity of compound libraries. Today, screening technology is highly advanced. Ultra-high throughput screening (uHTS), high content screening (HCS), and various types of omics are routinely used in the pharmaceutical industry. However, the number of new drug applications has not dramatically increased. One of the reasons is the chemical diversity for drug screening is still limited.

Most compound libraries for lead identification come from chemical syntheses (or combinatorial chemistry), which are limited by existing reactants and chemical synthetic protocols. Natural products are better resources for lead identification. However, the content of a natural product is usually low, and the cost for natural product extraction is high. Overconsumption leaves medical plants endangered. Strategies for making natural product-like compounds have been proposed [25]. One of the missions of chemomics is the development of another strategy for making natural product-like compounds (quasi natural product) via biomimetic syntheses (i.e., by mimicking natural endogenous synthetic rules) in order to efficiently make quasi natural product libraries as sustainable resources for drug screening.

4.1 Quasi natural product syntheses

In nature, synthesis of natural products occurs under mild conditions (ambient temperature and pressure with water and oxygen) beginning with simple endogenous molecules, such as, amino acids, creatine, and carbon dioxide, and enzyme-catalyzed reactions and reaction networks that couple to guarantee efficiency and chemical diversity. The structural diversity of natural products is sometimes unexpected sometimes, and often difficult to mimic by chemical synthesis. Although chemical synthesis seems incomparable with natural synthesis, it still has advantages. Chemical synthesis can be carried out under special conditions (e.g., no water, no oxygen, high temperature, high pressure, or microwave-heating). Also, chemical synthesis can be done in an isolated system to simplify the situation and produce a highly pure product.

Biomimetic synthesis should combine the pros of chemical syntheses and natural syntheses in order to make quasi natural products efficiently for drug screening (Table 4). As shown in Figure 12, acetyl carrier protein (ACP) is a platform for natural synthesis. A polymer bead is a platform for combinatorial chemical synthesis. Biomimetic synthesis, however, can adopt the protocol of the natural synthesis, but use a polymer bead as the synthetic platform. Replacing a protein with a polymer bead not only reduces synthesis costs, but also increases the synthetic capacity of the protocol. By means of chemical modifications, the bead can be a chemoyl carrier, and activate the chemoyl.

Natural synthetic protocols inspire chemical syntheses. If the mechanism of action of a natural synthetic protocol is understood, the biomimetic synthesis can be simplified, and the new synthetic protocol can be significantly more efficient. As shown in Figure 13, gramicidin S is synthesized at a protein carrier catalyzed by gramicidin synthase. After understanding the mechanism of action of the synthetic reaction, Bu and colleagues devised a simple and extremely efficient protocol to make gramicidin S [26]. Furthermore, with molecular dynamic simulations, Bu discovered that the peptide is easy to cyclize if it fits the following general formula:

$$(\text{D-X})\text{–L-P–}(\text{L-X})_n,$$

where D and L are configurations, X stands for any amino acid, P is proline, and $n \geqslant 1$.

4.2 Constructing quasi natural product libraries

To construct a quasi natural product library, we need a natural product scaffold. The scaffold can be a chemoyl or a combination of chemoyls, and has clear relations with various types of bioactivity. A network of chemical scaffolds and their bioactivities are demonstrated in Figure 14.

Having analyzed chemical structures collected by Zhou and colleagues [27], we conclude that some natural chemoyls or scaffolds are related to many bioactivity, while others are not. Excluding peptides, most natural products do not take α-amino acids as a scaffold. However, phenylalaninyl (chemoyl C_6C_3) is found in many natural products, and is associated with many types of bioactivity. Actually, a single carbon-substituted phenyl moiety plays important roles in natural product evolution. As shown in Figure 15, single carbon-substituted phenyl chemoyls appear in many TCM herb plants, and are related to many types of bioactivity.

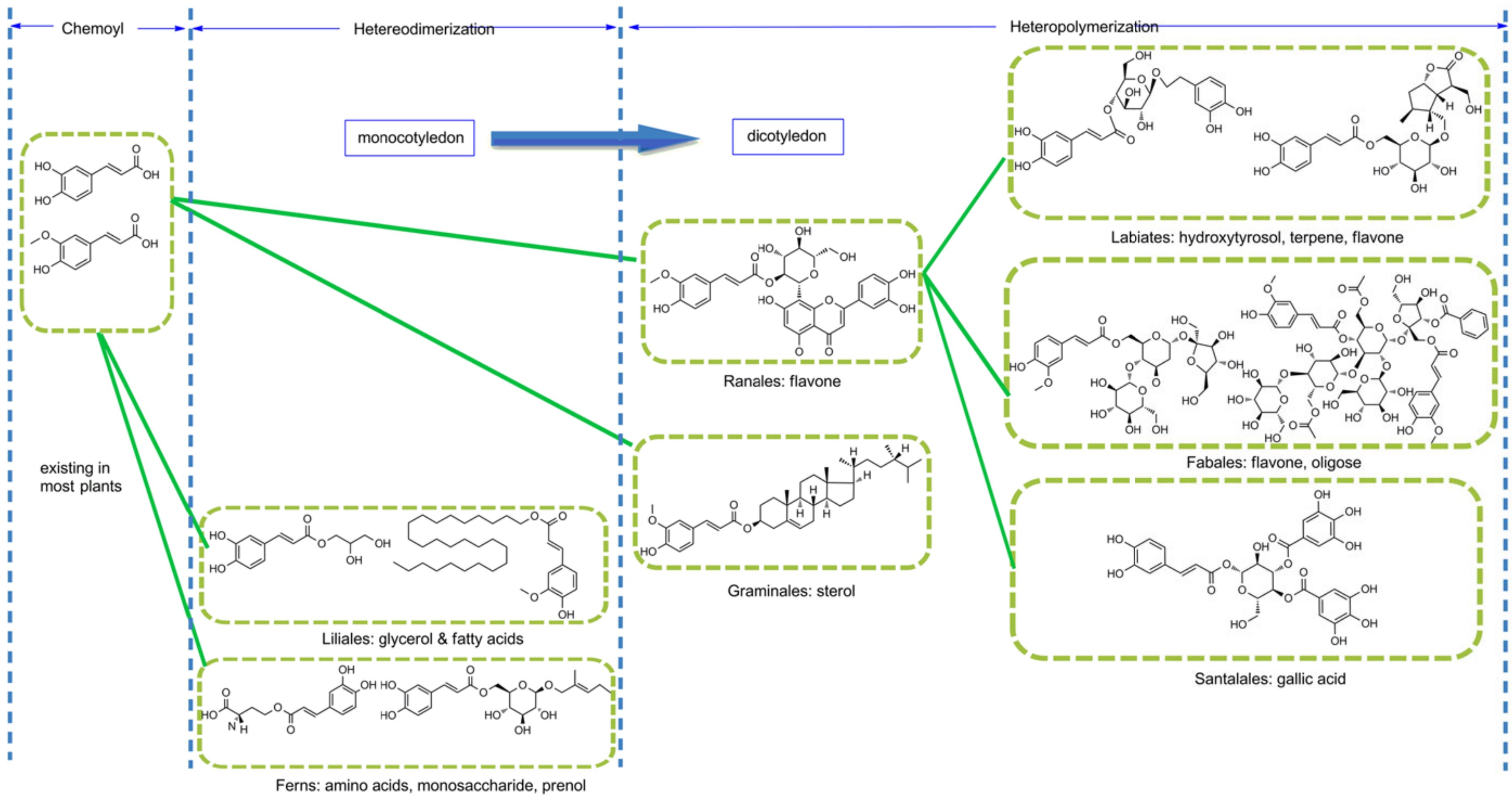


Figure 11 The derivatives and species evolution of caffeic acid.

Table 4 Natural synthesis, synthetic and biomimetic synthesis

	Building blocks	Synthetic condition	Synthetic protocol
Natural synthesis	endogenous molecules	natural condition complex system	loading & activating→elongating→terminating
Chemical synthesis	chemical reagents	controllable condition isolated system	finding synthon→activating precursor→ synthesizing (chiral control + elongation + release)
Biomimetic synthesis	endogenous molecules and chemical reagents	controllable condition simplified system	loading & activating→elongating→terminating

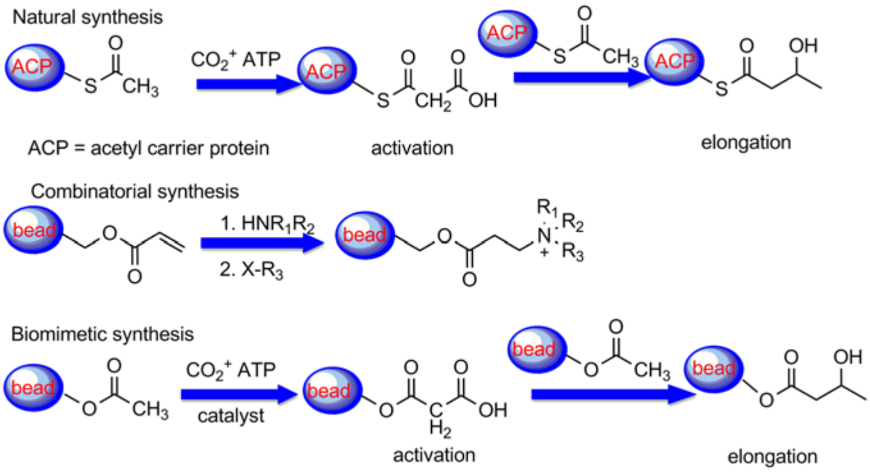


Figure 12 Examples of natural, combinatorial and biomimetic syntheses.

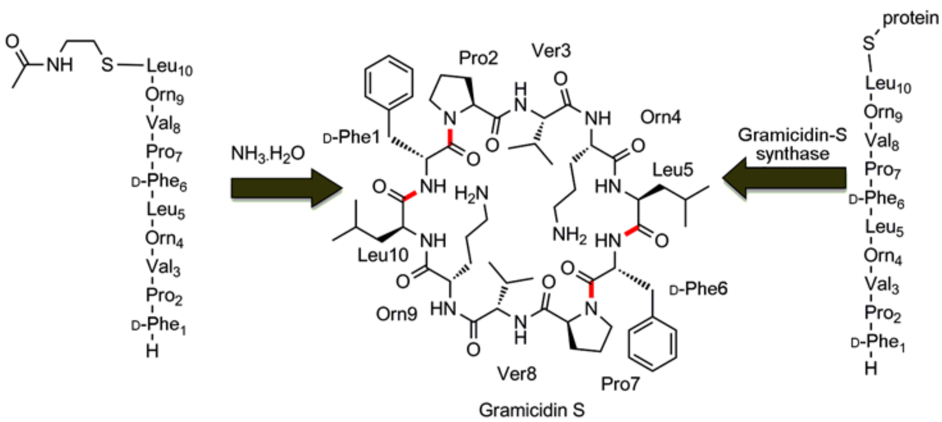


Figure 13 The biomimetic synthesis and natural synthesis of gramicidin S.

The bioactivities associated with these chemoyls are: cytotoxicity, antioxidant, antibacterial/anti-virus, anti-inflammatory, anti-metabolic diseases, and anti-tumor. For example, the benzopyran class of antioxidants mainly come from licorice, *epimedium*, skullcap, iris, *ginkgo*, *rotenone*, *artemisia annua*, orange peel, arrowroot, *morus alba*, *evodia*, *erythrina*, *millettia incense*, *matrine*, *capillaris*, shegan, *andrographis paniculata*, *salvia*, and *oriental arborvitae*.

This analysis can be extended. For example, the structural features of antioxidant benzopyran compounds can be summarized in the form of a Markush structure, as shown in Figure 16. The Markush structure is a quasi natural product library for anti-oxidation bioactivity. Through computer simulation techniques, a quasi natural product library can be enumerated from the Markush structure in Figure 16. The virtual compounds can be validated by virtual screening programs. Furthermore, biomimetic synthetic protocols can be designed, and lead compounds eventually made and assayed eventually.

4.3 Chemoyl-based pharmacological networks

The conventional drug discovery paradigm focuses on finding a compound that is highly selective against a single-target. However, more and more evidence demonstrates that most ligands act on multiple targets, and these targets may

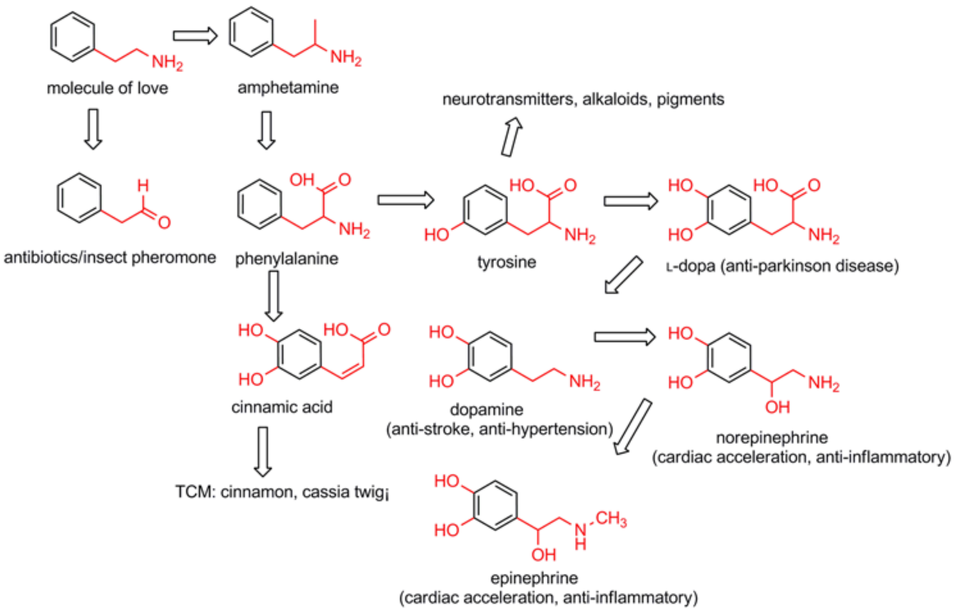


Figure 14 A network of chemical scaffolds and bioactivities.

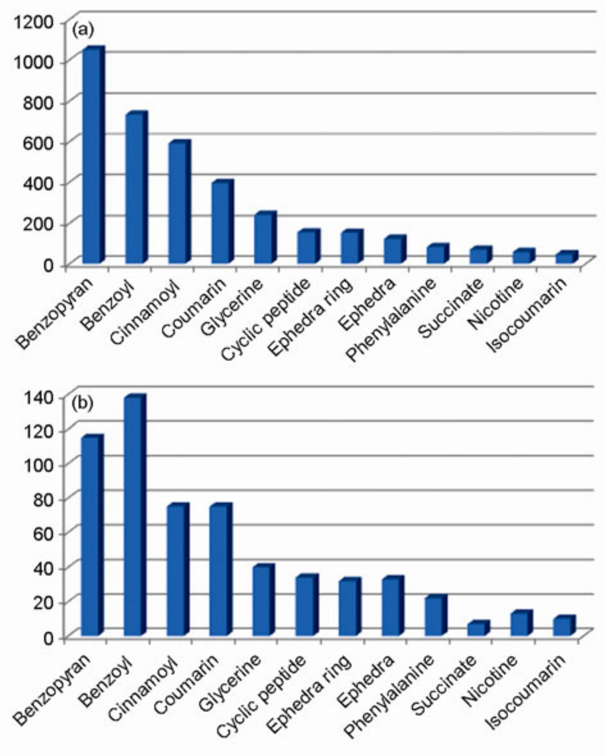


Figure 15 (a) Chemoyls versus TCM herbs; (b) chemoyls versus bioactivity.

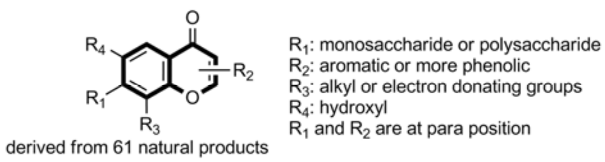


Figure 16 A quasi natural product library for antioxidants.

form complex signal transduction networks. It will be difficult to design or discover an effective drug without side-effects. In fact, there is no such a drug, which absolutely acts on a single-target. It is safer to announce that you have not yet found a drug that acts on more than one target. This is particularly true for diseases caused by complex mechanisms of actions [28]. For example, β-lactams kill viruses through more than two protein-induced cell apoptoses and Gleevec, a targeted anti-cancer drug kills cancer cells through multiple signal transduction pathways for apoptoses.

Recently, network pharmacology has attracted a lot of attention in the pharmaceutical industry. Network pharmacology adopts approaches from systems biology, creates pharmacological networks, and identifies the changes in gene polymorphism, connectivity, and redundancy due to pharmaceutical regulations. Some scientists believe network pharmacology is a new term for the systematical quantitative structure–activity relationships (QSAR) studies on drug effects and side effects done by Janssen years ago [29]. Others argue that it is actually systems pharmacology [30]. The networks studied by network pharmacology are as follows: i) protein-protein interaction networks (PPI), where the node is a protein molecule, and the edge represents two proteins' similarity or regulations; ii) drug-drug interaction networks (DDI), where the node is a drug molecule, and the edge represents common target(s) or medical indication(s) between two drugs. Figure 17 demonstrates examples of drug and target interactions.

Network pharmacology mainly tries to solve the following problems: i) Identifying drug targets at the system level. An ideal drug target satisfies these criteria: it is an enzyme that is able to be screened; it is the main cause and the driving force of the disease; it is a unique determining factor for

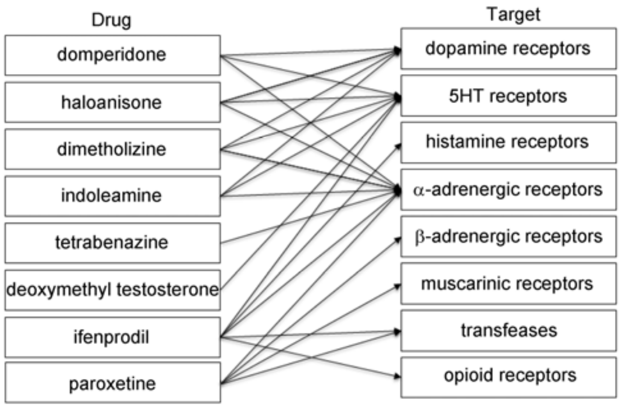


Figure 17 Examples of drug-target interactions.

disease development; it is the only target associated with the clinical observation; there is ligand affinity at the active site; it has low molecular weight and a three-dimensional structure or at least a structural model; preferably, it is not present in the human body; it does not disturb other proteins or genes unrelated to the disease process. ii) Identifying drug leads. An ideal drug lead mainly regulates on its target without disturbing anti-targets. Preferably, if the disease is caused by more than one target, the lead can synergistically regulate multiple drug targets (multiple targeted drugs). Alternatively, the leads can regulate multiple targets (compound drugs).

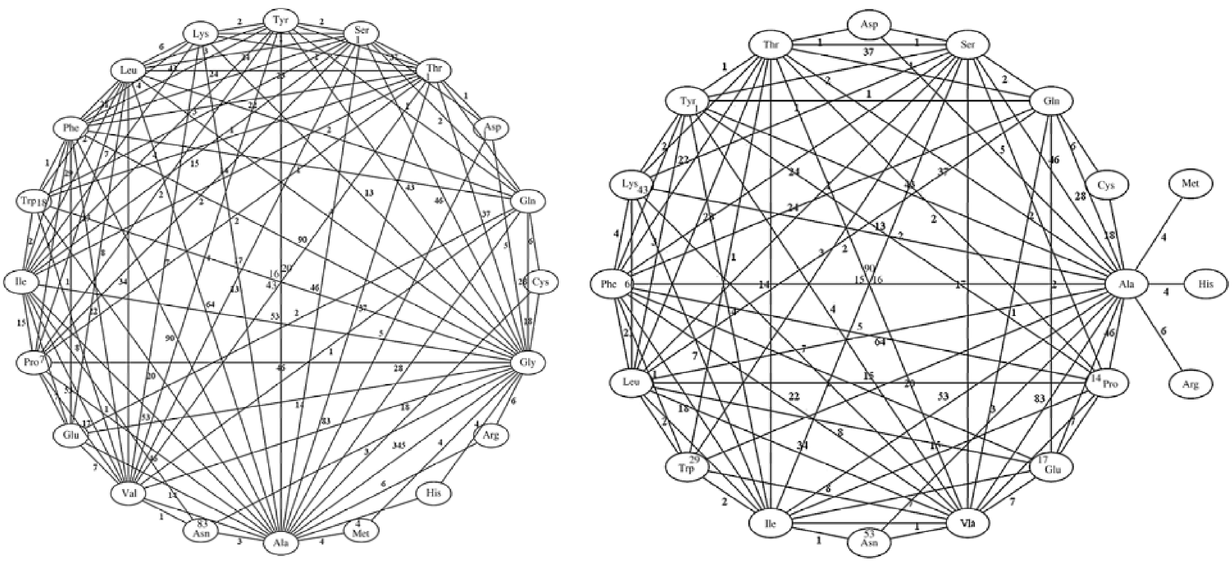
The drug-drug interaction network should have another version: chemoyl-based pharmacological networks (CPN). In CPNs, the node is a chemoyl, the edge represents a relation if two chemoyls can be combined to result in natural product(s) or bio-functional molecule(s). By investigating a natural products database, DNP (Dictionary of Natural Products, www.chemnetbase.com), we found interesting phenomena for CPNs. As shown in Figure 18(a), 20 natural amino acids (primary metabolic chemoyls) form a CPN: some amino acid pairs can productively generate many natural products; others cannot. For example, the tyrosine and phenylalanine pair results in more than 43 natural products; proline and phenylalanine, and tryptophan and phenylalanine, only result in one natural product each. Since glycine has no side chain, it can be removed from the network, resulting in a simplified CPN as shown in Figure 18(b).

Deeper investigations are required in order for CPN to answer the following questions: what are the targets associated with the edges (natural products) of the CPN? Why are some combinations of amino acid pairs not selected in nature? If these products unselected in nature are identified, what will be their targets?

5 Conclusions

Natural chemoyls are basic chemical units of biopolymers and natural products, and keys to understanding how these molecules are assembled from simple building blocks in nature. Each chemoyl is associated with a number of small molecular building blocks, and *vice versa*. Therefore, the relationships between the chemoyls and the building blocks are multiple-to-multiple. This relationships should be fully investigated in order to decipher the natural secrets of natural product syntheses.

Correctly predicting the druggability of a compound is crucial to a drug discovery project. It is well known that the druggability relies on the chemical structure of a compound,



(a) (b)

Figure 18 (a) CPN for 20 natural amino acids; (b) simplified CPN for natural amino acids (glycine has been removed).

but, we have not investigated if the druggability is associated with chemoyls in the compound. As time goes on, more rules of natural product assembly will be revealed, and we will invent more techniques to make quasi natural products for drug screening and to predict druggable molecules by recognizing druggable chemoyl combinations. To understand the natural chemoyl assembly process, one has to study the chemical reaction mechanism. The core of the mechanism is the conformation changes of a substrate molecule at transition states as it is catalyzed by an enzyme. Many existing drugs work because their active conformations are similar to those of enzyme substrate transition states, and have higher affinities to the active sites of enzymes. Hence, chemomics will be useful for design compound libraries that mimic enzyme substrate transition states. Maybe, enzyme substrate transition state conformationomics will appear in the future as a new virtual screening approach. The changes in chemoyls and related natural building blocks along with individual medical plant life-time development, the changes along with species evolution, and the relationships between the chemoyls and bioactivity will be important information for sustainably utilizing natural products and TCM modernization.

In recent years, synthetic biology has developed new technology to synthesize basic components in cells from nucleic acids [31]. Through external interventions, synthetic biology tries to force biological systems, for example bacteria, to make desired natural products. The common part of chemomics and synthetic biology is to make natural products (or quasi natural products) by learning from nature. However, chemomics mimics the chemical mechanism of biological syntheses, and synthetic biology mimics the biological process of the biological syntheses. The ultimate goal of synthetic biology is able to synthesize life.

Chemomics focuses on deciphering the building blocks of natural products, discovering natural synthetic laws, and developing new chemical technology to make quasi natural products. Conventional natural product chemistry studies metabolites and optimizing natural products by chemical modification for biological activity. The modification is limited by existing scaffolds and reagents. In contrast, chemomics starts with elucidating natural products, interpreting natural synthetic laws, and inventing new ways to make quasi natural products by combining chemical and biological synthetic approaches in order to expand chemical diversity space. Chemomics is interdisciplinary, and requires the joint efforts of scientists from chemistry, biology, informatics, and mathematics.

This work was supported by the National Science and Technology Major Project of China (2010ZX09102-305), the National High-tech R&D Program of China (863 Program, 2012AA020307), the Introduction of Innovative R&D Team Program of Guangdong Province (2009010058), and the National Natural Science Foundation of China (81173470).

1 Collins FS, Morgan M, Patrinos A. The human genome project: Lessons from large-scale biology. *Science*, 2003, 300: 286–290
2 Luscombe NM, Greenbaum D, Gerstein M. What is Bioinformatics? A proposed definition and overview of the field. *Methods Inf Med*, 2001, 40: 346–358
3 Dhar PK. The next step in biology: A periodic table. *J Biosci*, 2007, 32: 1005–1008
4 Lander ES. The new genomics: Global views of biology. *Science*, 1996, 274: 536–539
5 Biro JC, Benyó B, Sansom C, Szlávecz Á, Fördös G, Micsik T. A common periodic table of codons and amino acids. *Biochem Biophys Res Commun*, 2003, 306: 408–415
6 Van Drie JH. Pharmacophore discovery–Lessons learned. *Curr Pharm Design*, 2003, 9: 1649–1664
7 Evans BE, Rittle KE, Bock MG, Dipardo RM, Freidinger RM, Whitter WL. Methods for drug discovery: Development of potent, selective, orally effective cholecystokinin antagonists. *J Med Chem*, 1988, 31: 2235–2246
8 DeSimone RW, Currie KS, Mitchell SA, Darrow JW, Pippin DA. Privileged structures: Applications in drug discovery. *Comb Chem & High T Scr*, 2004, 7: 473–493
9 Lewell XQ, Judd DB, Watson SP, Hann MM. RECAP—retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J Chem Inf Comput Sci*, 1998, 38: 511–522
10 Corey EJ. General methods for the construction of complex molecules. *Pure Appl Chem*, 1967, 14: 19–38
11 Burke MD, Schreiber SL. A planning strategy for diversity-oriented synthesis. *Angew Chem Int Ed*, 2004, 43: 46–58
12 Böhm HJ. Site-directed structure generation by fragment-joining. *Perspect Drug Discovery Des*, 1995, 3: 21–33
13 Rees DC, Congreve M, Murray CW, Carr R. Fragment-based lead discovery. *Nat Rev Drug Discov*, 2004, 3: 660–672
14 Miflin BJ, Lea PJ. Amino acid metabolism. *Ann Rev Plant Physiol*, 1977, 28: 299–329
15 Dewick PM. *Medicinal Natural Products: A Biosynthetic Approach*. Wiley, 2009
16 Munos B. Lessons from 60 years of pharmaceutical innovation. *Nat Rev Drug Discov*, 2009, 8: 959–968
17 Michal G. *Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology*. Wiley, 1999
18 Kotera M, Hirakawa M, Tokimatsu T, Goto S, Kanehisa M. The KEGG databases and tools facilitating omics analysis: Latest developments involving human diseases and pharmaceuticals, next generation microarray bioinformatics. *Methods Mol Biol*, 2012, 802: 19–39
19 Reitz M, Sacher O, Tarkhov A, Trümbach D, Gasteiger J. Enabling the exploration of biochemical pathways. *Org Biomol Chem*, 2004, 2: 3226–3237
20 Johnson RE, Washington MT, Prakash S, Prakash L. Fidelity of human DNA polymerase η. *J Biol Chem*, 2000, 275: 7447–7450
21 Pray L A. DNA replication and causes of mutation. *Nat Educ*, 2008, 1: 1
22 Brown JR, Gentry D, Becker JA, Ingraham K, Holmes DJ, Stanhope MJ. Horizontal transfer of drug-resistant, aminoacyl-transfer-RNA synthetases of anthrax and Gram-positive pathogens. *EMBO reports*, 2003, 4: 692–698
23 Strieker M, Tanović A, Marahiel MA. Nonribosomal peptide synthetases: Structures and dynamics. *Curr Opin Struct Biol*, 2010, 20: 234– 240
24 Wiermann R. Secondary plant products and cell and tissue differentiation. In: Conn EE, Ed. *The Biochemistry of Plants*. New York: Academic Press, 1980. 85–137
25 Nicolaou KC, Pfefferkorn JA, Roecker AJ, Cao GQ, Barluenga S, Mitchell HJ. Natural product-like combinatorial libraries based on privileged structures. 1. General principles and solid-phase synthesis of benzopyrans. *J Am Chem Soc*, 2000, 122: 9939–9953

26 Bu XZ, Wu XM, Ng NLJ, Mak CK, Qin CG, Guo ZH. Synthesis of gramicidin S and its analogues via an on-resin macrolactamization assisted by a predisposed conformation of the linear precursors. *J Org Chem*, 2004, 69: 2681–2685
27 Zhou J, Xie G, Yan X. *Encyclopedia of Traditional Chinese Medicines—Molecular Structures, Pharmacological Activities, Natural Sources and Applications*. Berlin: Springer-Verlag, 2011
28 Janga SC, Tzakos A. Structure and organization of drug-target networks: Insights from genomic approaches for drug discovery. *Mol Bio Syst*, 2009, 5: 1536–1548
29 Hopkins AL. Network pharmacology: The next paradigm in drug discovery. *Nat Chem Biol*, 2008, 4: 682–690
30 Berger SI, Iyengar R. Network analyses in systems pharmacology. *Bioinformatics*, 2009, 25: 2466–2472
31 Purnick PEM, Weiss R. The second wave of synthetic biology: From modules to systems. *Nat Rev Mol Cell Biol*, 2009, 10: 410–422