Drug Discovery Inspired by Mother Nature: Seeking Natural Biochemotypes and the Natural Assembly Rules of the Biochemome

Qiong Gu, Xin Yan, and Jun Xu*

Research Center for Drug Discovery, School of Pharmaceutical Sciences, and Institute of Human Virology, Sun Yat-Sen University, 132 East Circle at University City, Guangzhou, China.

Received, March 14, 2013; Revised, July 10, 2013; Accepted, July 28, 2013; Published, July 30, 2013.

ABSTRACT- Purpose. The Human Genome Project is producing a new biological 'periodic table', which defines all genes for making macromolecules (proteins, DNA, RNA, etc) and the relations between genes and their biological functions. We now need to consider whether to initiate a *biochemome project* aimed at discovering biochemistry's 'periodic table', which would define all molecular parts for making small molecules (natural products) and the relations between the parts and their functions to regulate genes. By understanding the Biochemome, we might be able to design biofunctional molecules based upon a set of molecular parts for drug innovation. **Methods**. A number of algorithms for processing chemical structures are used to systematically derive chemoyls (natural building blocks) from a database of compounds identified in Traditional Chinese Medicine (TCM). The rules to combine chemoyls for biological activities are then deduced by mining an annotated TCM structure-activity database (ATCMD). Based upon the rules and the basic chemoyls, a chemical library can be biochemically profiled, virtual synthetic routes can be planned, and lead compounds can be identified for a specific drug target. **Conclusions**. The Biochemome is the complete set of molecular components (chemoyls) in an organism and Biochemomics studies the rules governing their assembly and their evolution, together with the relations between the Biochemome and drug targets. This approach provides a new paradigm for drug discovery that is based on a comprehensive knowledge of the synthetic origins of biochemical diversity, and helps to direct biomimetic syntheses aimed at assembling quasi-natural product libraries for drug screening.

INTRODUCTION

Although modern synthetic chemistry has made great progress only one *de novo* compound, initiated by combinatorial chemistry, has been approved as a drug (the antitumor compound known as sorafenib from Bayer, approved by the FDA in 2005), in over 25 years from 1981 to 2006 (1). According to a report from the World Health Organization (WHO), between 70% and 95% of people in developing countries rely on traditional medicine as their primary source of medication, suggesting that new drugs might be discovered by paying more attention to learning from Nature (2).

There are two ways to learn from Nature. One way is to learn from natural products, the results of Nature's creation. People have been doing this for many years. For more than a century, natural products, including marine products and phytochemicals, have been extracted from natural organisms and screened against biological targets. However, there are major challenges to this approach: more and more natural organisms are becoming endangered and utilizing and modifying their natural products can cause ecological problems. Thus, when a natural product is officially approved as a drug or as a nutritional supplement, the organisms that produce the product are potentially at risk. Globalization and worldwide environmental changes exacerbate the situation. Many traditional Chinese medicines (TCM) are losing their medical functions due to the shortage of authentic traditional Chinese herbs. For example, TCM herbs usually depend on the locations where they are cropped; however, there is constant deterioration to the ecological systems of these locations. Moreover, chemically reproducing natural products is still difficult and costly. We have to seek new ways of producing natural products without provoking ecological crises.

Corresponding Author: Jun Xu, PhD; Professor; Research Center Centre for Drug Discovery, School of Pharmaceutical Sciences, and Institute of Human Virology, Sun Yat-Sen University, 132 East Circle at University City, Guangzhou, China; Email: junxu@biochemomes.com; xujun9@mail.sysu.edu.cn

Another way is to learn from the processes that Nature uses to create natural products. Most natural products are made in cells with non-extreme conditions (aqueous solution, neutral pH, room temperature, and normal atmospheric pressure). However, the cells can make natural products from a few simpler endogenous molecules efficiently. The secret is that the cells use enzymes as catalysts, and these enzymes can be regulated through physical conditions, other enzymes, ions, and small molecules. If these processes can be deciphered and mimicked, natural products or quasi-natural products can be made more efficiently without ecological consequences.

In this paper, we focus on deciphering how Nature reproduces and creates natural products. We started by identifying chemoyls(3) that Nature uses to make its bioactive compounds. Then we tried to identify the rules for combining these chemoyls to make organic compounds for bioactivities. Finally, we applied these rules to make a quasi-natural product library for the identification of drug leads. The rules of combining chemoyls are not so simple as to be represented in a "chemoyl periodic table". Instead, the rules have to be represented in networks.

METHODS

Defining Chemoyls and Biological Building Blocks

Drug design should be a process that assembles a number of molecular components to make a biofunctional molecule. Therefore, the first question for drug design is what is a molecular component? As shown in Figure 1, a natural product (A) can be disassembled into a number of simpler structural fragments, chemoyls (B). A chemoyl is not a molecular entity; it is a molecular fragment. An endogenous molecule that can contribute a chemoyl to a target molecule (natural product) in an enzyme-catalyzed chemical reaction is a natural product building block (Figure 1C).

Distinguishing Chemoyls and Building Blocks

The concept of molecular chemical parts has been proposed for many years. For example, the "isoprene rule" was first recognized by Wallach in the 19th century(4), and was extended into the "biogenetic isoprene rule" by Ruzicka sixty-six years later (5). Isoprene is a typical molecular part for many natural product molecules. Today, there are about ten natural product molecular components recognized by natural product chemists (6). To be concise, we term the molecular components as shown in Figure 1B as chemoyls, and term the molecule containing a chemoyl with an activation group or groups as a building block (Fig. 1C).

In chemistry, a building block means a compound that can be used to make other novel compounds. In biochemistry, a building block is a molecular entity that exists in a life system, and is an endogenous molecule (3). A biological building block can be used to make many different natural product molecules. Therefore, it is not obvious which building blocks have been used to build a natural product molecule. Moreover, there are many ways to disassemble a natural product molecule into chemoyls because there are many ways to make a natural product molecule in Nature.

Therefore, drug design requires the identification of molecular components to construct a natural product like molecular library, screening the library against drug target(s), seeking optimized synthetic routes and building blocks to make promising hits, converting the hits to leads, and improving the pharmacophore properties of the leads.

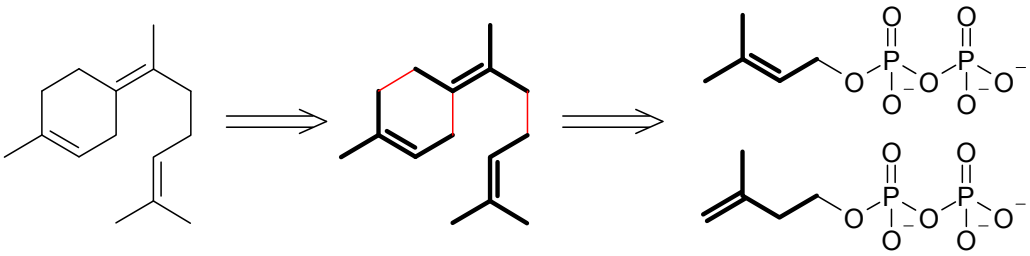


Figure 1. Relations of natural product, chemoyls and building blocks. A: Bisabolene, B: three chemoyls indicate how the molecule has been built, and C: two building blocks that consist of C_5-chemoyls activated by bisphosphate groups.

Seeking Chemoyls and Corresponding Building Blocks

Although, a chemoyl cannot be easily recognized by simply inspecting a natural product molecular structure, it can be derived from natural product molecular structures by systematic analyses. The "isoprene rule" is a good example.

To construct a quasi-natural product library, we have to discover rules to combine chemoyls. These rules can be revealed by systematically analyzing the partnerships of chemoyls in a natural product database, such as the TCM chemical structure database (7).

Based upon the discussion above, rational drug design requires the following elements: (1) drug targets; (2) interaction models of the target and drug; (3) chemoyls and their combinatorial rules; and (4) the biological or chemical building blocks related to the chemoyls. We consider both biological and chemical building blocks, because we want to employ both biological and chemical techniques to synthesize drug or drug like molecules from chemoyls.

Biological building blocks are organic compounds or metabolites that are present in many organisms or cells; they are utilized in other natural products as subunits. In short, building blocks are precursors of other natural products.

It is easy to recognize that proteinogenic amino acids, or the five nucleoside bases (including the purines A and G and the pyrimidines C, T and U), ribose, and deoxyribose are biological building blocks, because they are precursors to proteins or nucleic acids (DNA and RNA) respectively.

Other biological building blocks, such as short half-life endogenous compounds (*e.g.*, ornithine, homoserine, isoprene derivatives, shikimic acid, pyruvic acid, *etc.*) are important as precursors to the above-mentioned building blocks (*e.g.*, proteinogenic amino acids). However, they are not synthesized with well-known peptide polymerases. They are assembled through non-ribosomal peptide-synthetase (NRPS) enzymes. These building blocks are structurally more diverse (*i.e.*, they are not necessarily amino acids), and have to be derived from natural product databases and other references by data mining approaches.

The annotated TCM chemical structure database (ATCMD)(7) is a natural product database that contains TCM usage information, structural data and bioassay results discussed in the Encyclopedia of TCM. It is an integrated achievement of a long-term TCM research project by authors at the Chinese Academy of Sciences seeking to derive chemoyls from ATCMD.

Natural products (excluding proteins, DNA or RNA) are typically divided into primary metabolites (molecules that are directly involved in the regulation of life systems, *e.g.*, growth, development, and reproduction) and secondary metabolites (molecules that are not directly involved in the regulation of life systems, and often play roles in defense against environmental stresses). Some secondary metabolites are precursors to the primary metabolites.

RESULTS

Currently 129 chemoyls have been derived from ATCMD and other references (6, 8). These chemoyls are divided into 15 classes. Each chemoyl class associates with a number of bioactivities related to TCM herbs (as shown in Figure 2). The terpenene, alkaloid, and flavone classes are associated with the highest number of bioactivities; the saccharide, amino acid, steroid, lignin, quinone, coumarin, and tannin classes are associated with a medium number of bioactivities; and the benzofuran, stilbene, chromene, nucleoside, and polyene classes are associated with the lowest number of bioactivities. This is understandable because terpenenes, alkaloids, and flavones/ flavonols are the largest and most diverse groups of plant secondary compounds. Terpenes are common in most plants and fungi, but they rarely accumulate in bacteria. More than 15,000 terpenoids are discovered. They generally occur free, or derivatized as esters and glycosides, or attached to proteins. Steroids in mammals are products of terpenoid metabolism (9). More than 12,000 alkaloids have been discovered (10). Flavonols and flavones are produced widely in plants and are mainly concentrated in the outer tissues. Flavonols act as antioxidants and protect ascorbic acid from auto-oxidation (11). Benzofurans, stilbenes, chromenes, nucleosides, and polyenes without annotated bioactivities in the current version of the database may have lower occurrences in plants, or exhibit stability problems. Further investigations are required.

In nature, these chemoyls are covalently combined to form new molecular moieties, which occur for specific stress resistance purposes. The moieties, however, may be used for other biological purposes. For example, a natural product was produced to be a plant antibiotic; however, it can also be used as an anti-aging agent. Therefore, a chemoyl group can be associated with many bioactivities, as shown in Figure 3.

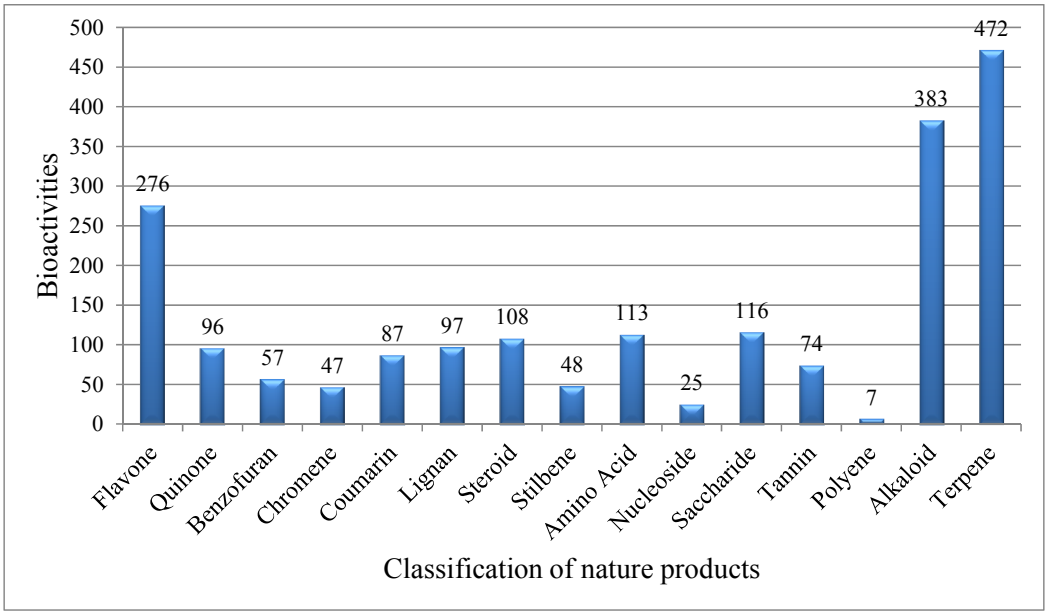


Figure 2. Fifteen chemoyl groups and their bioactivities in TCM.

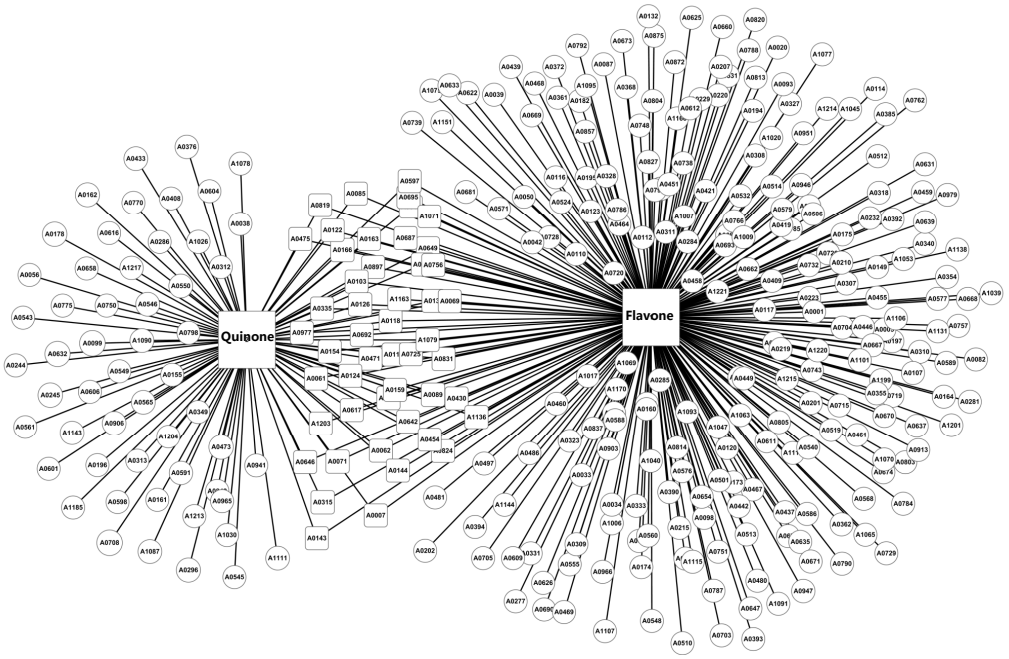
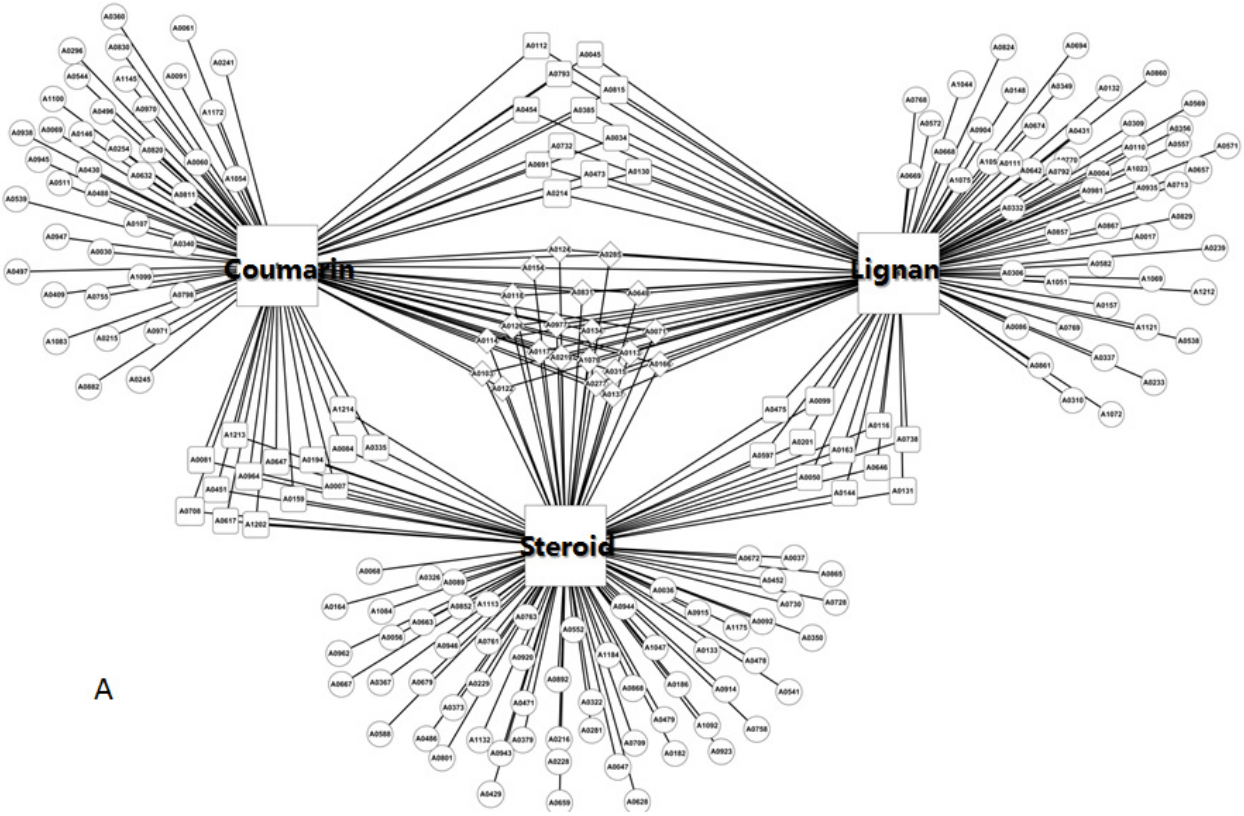


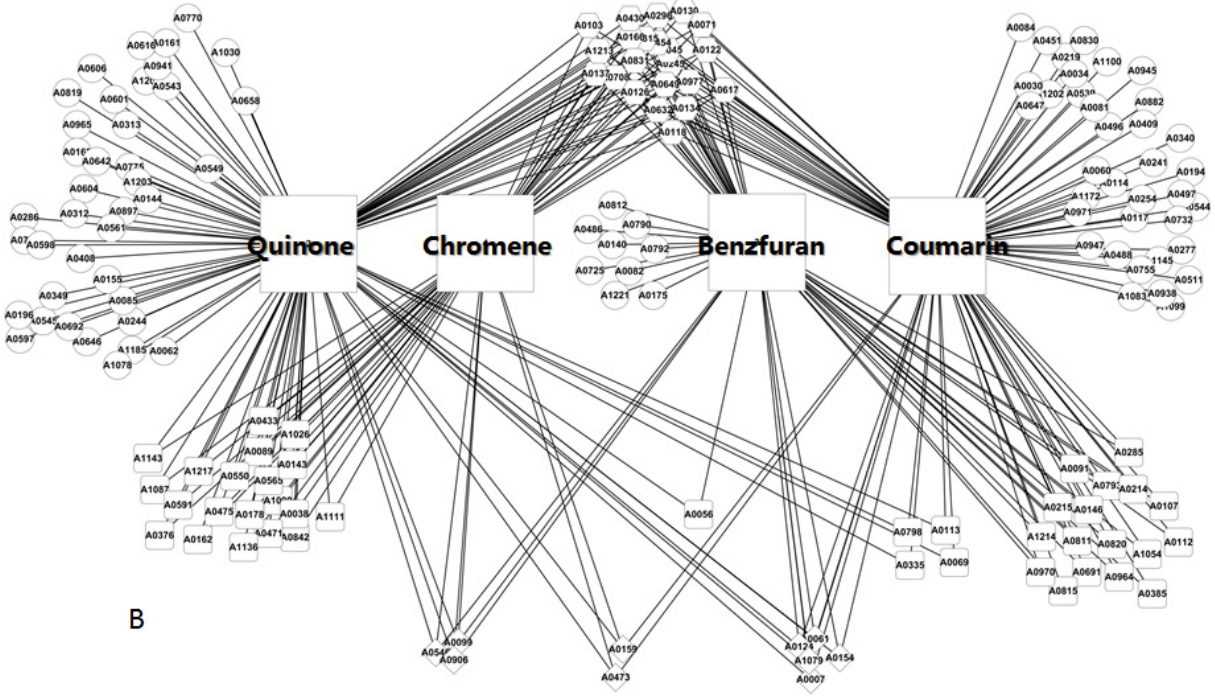
Figure 3. Quinone and flavone (labeled in the square boxes) and associated bioactivities in TCM. Codes in circles represent activities associated with quinone or flavone. Codes in smaller rectangles boxes represent activities associated with both quinone and flavone. The definitions for the activity codes in this figure can be found in reference (7).

In Figure 3, larger square nodes represent chemoyl groups; smaller circles represent associated bioactivities (as discovered in TCM); smaller rectangles represent combined chemoyls, composed of simpler chemoyls, among larger square nodes. The codes in the smaller circles represent the individual bioactivities described in reference (7). This figure demonstrates that smaller chemoyls tend to associate with a larger number of bioactivities. When two chemoyls are combined, the number of associated bioactivities is significantly reduced. This is because specificity increases when a molecule is made from more than one chemoyl.

The chemoyl combination networks can be formed based upon the combinations of 3, 4, and 5 chemoyls, which generate maps with enhanced complexity as shown in Figures 4 A to C.


Coumarin
Lignan
Steroid

A


Quinone
Chromene
Benzfuran
Coumarin

B

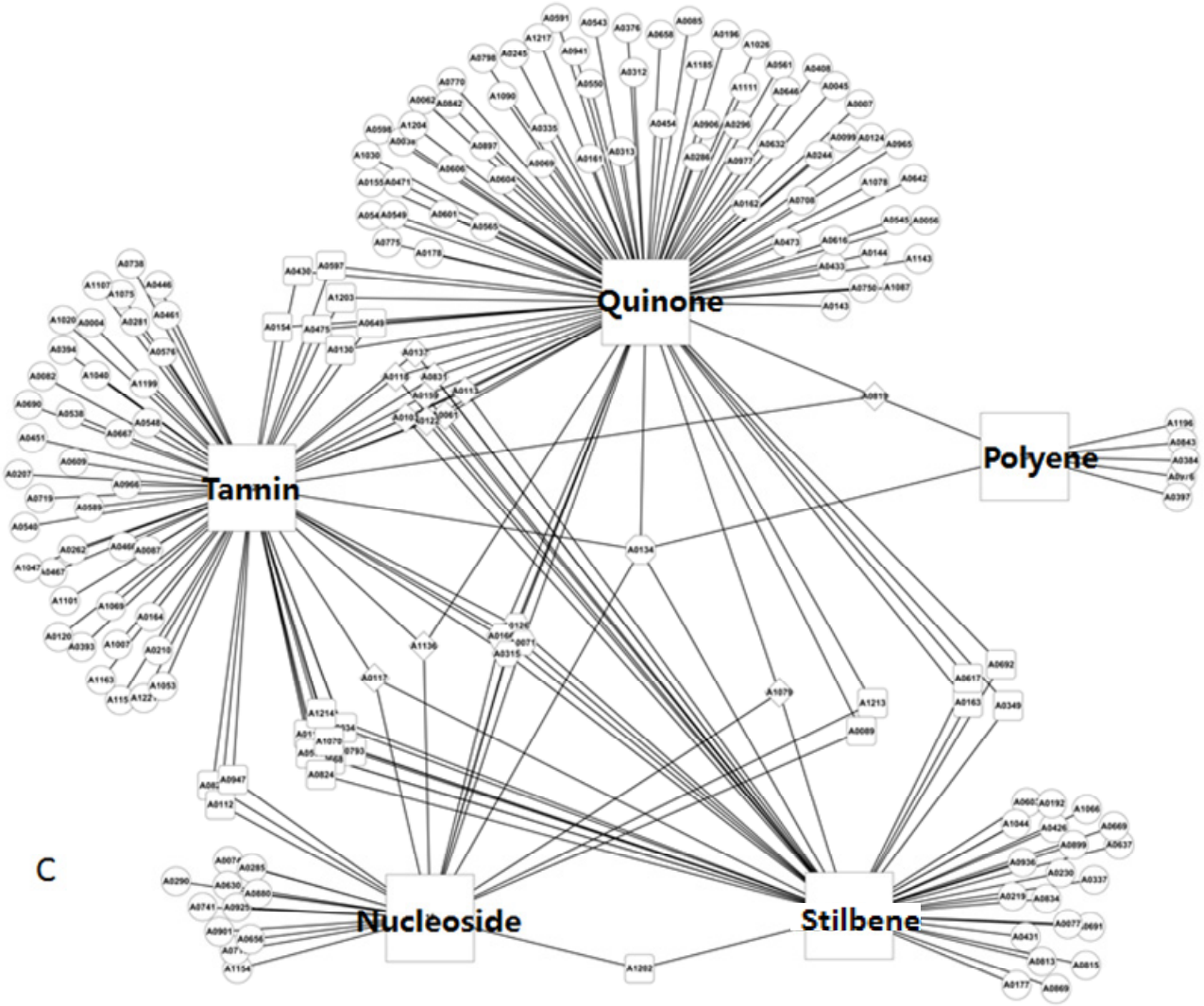


Figure 4. A: 3-chemoyl combination network. B: 4-chemoyl combination network. C: 5-chemoyl combination network.

These networks indicate that molecules generated by higher numbers of chemoyls will have fewer activities. For example, based on mining ATCMD data, combining quinone, tannin, and polyene can generate only one activity (A00819); similarly, combining quinone, tannin, nucleoside, and stilbene can generate only one activity (A0134).

When chemoyls are combined, the new molecule is larger in size, more rigid in shape, and has enhanced target specificity. Therefore, the new molecule has a smaller number of associated activities but individual activities are frequently enhanced. This outcome is frequently observed in the results from mining ATCMD structure data (as shown in Figure 5).

In Figure 5, the size of a pie-chart represents the number of bioactivities for the associated chemoyl or combination of chemoyls. Here, three chemoyls, benzofuran, flavone, and stilbene, are associated with more bioactivities than their combinations. The colors in the pie-chart stand for distinct bioactivities. For example, the major bioactivities for stilbene derivatives are anti-oxidization, anti-cancer, and anti-bacterial. Combining stilbene chemoyls and benzofurans produce new molecules that have fewer bioactivity types. Furthermore, combining the three chemoyls in Figure 5 produces new molecules with reduced spectra of bioactivity.

The 15 chemoyl classes form a chemoyl-type activity network which demonstrates the relations among fifteen chemoyl-types.

Figure 6 indicates that steroids (7) and terpenes (15) are more frequently covalently combined to produce the natural products that have the most diverse bioactivities in TCM. Polyene (13) is isolated from the network; it is not commonly combined with other chemoyls. Flavones (1), steroids (7), alkaloids (14) and terpenes (15) are most frequently (indicated by darker lines) present in natural products simultaneously, so as to exhibit more bioactivities.

In case of triple chemoyl-type combinations, we have created tri-chemoyl-type activity networks as shown in Figure 7. In this network, an edge stands for three chemoyl-types covalently combined together with amino acid scaffolds.

Chemoyl-type activity networks are both statistic and dynamic. The topology of a network depends on the threshold number of new bioactivities for new chemoyls produced from parent chemoyls. The networks indicate that chemoyl types can form more or less new chemical entities that exhibit a number of bioactivities. Consequently, the rules of chemoyl combination are determined by biological targets *per se*.

A note on the screening of the TCM database for lead compounds using computational methods

In addition to standard cell and/or animal-based drug target analysis, virtual screening provides an alternative strategy for the identification of lead compounds and would appear to be suitable for the analyses of the TCM biochemome project. Successful examples include the report of Giacomini and colleagues at UCSF who identified four large-neutral amino acid transporter 1 (LAT-1) ligands by virtual screening (12) and Leung and colleagues who identified inhibitors of tumor necrosis factor-α from a library of marketed drugs using virtual screening methods (13). For a review of successful cases of drug repositioning by virtual screening see Ref. (14).

CONCLUSIONS

By mining ATCMD data, we have generated three types of chemoyl-bioactivity networks. These networks represent the rules of chemoyl combination. Our studies reveal that biochemomes do indeed exist in the form of chemical building blocks. The rules governing chemoyl assemblies are in the form of networks, not in the form of a rectangular table. The networks are dynamic, and change with the evolution of life systems.

Based upon the rules of chemoyl combination and related bioactivities, quasi-natural product libraries can be constructed from a set of biochemoyls for specified bioactivities and virtual screening can be performed against drug targets using a number of computational tools (15-19). A consequent challenge is to develop feasible synthetic technologies to make the libraries. A number of technologies have been developed for chemical syntheses (20), biological syntheses (21-23), and biomimetic syntheses (24-29).

We are presently engaged in identifying natural biochemoyls and the natural assembly rules of the biochemome by data-mining TCM and other natural product databases. Biochemomics is still young; there are many things to be done.

ACKNOWLEDGEMENTS

This work was supported by a grant from the National High Technology Research and Development Program of China (863 Program) (No. 2012AA020307), the Introduction of Innovative R&D Team Program of Guangdong Province (2009010058), the National Natural Science Foundation of China (No. 81001372, and 81173470), and the Special Funding Program for the National Supercomputer Center in Guangzhou (2012Y2-00048/2013Y2-00045, 201200000037), the International Collaboration & Exchange Project (2010DFB30830). Authors would like to thank Prof. Jiaju Zhou for his contributions to the ATCMD database and inspiring discussions.

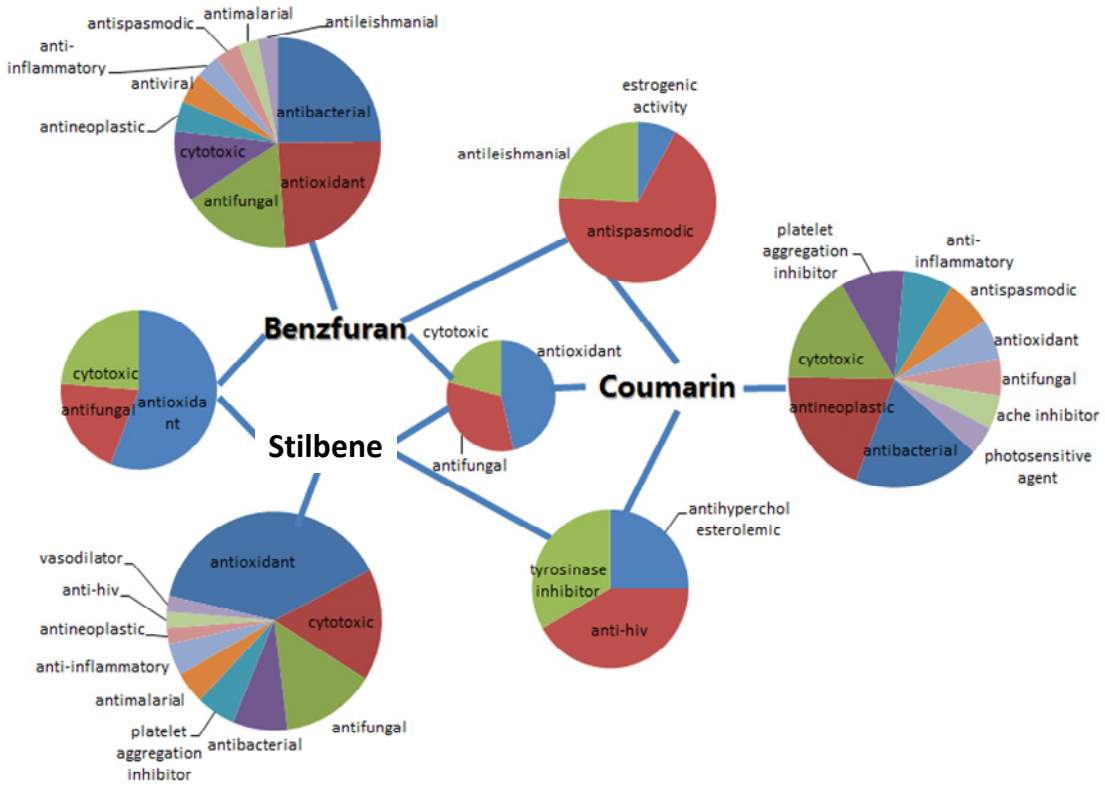


Figure 5. Chemoyl combinations and bioactivities in TCM.

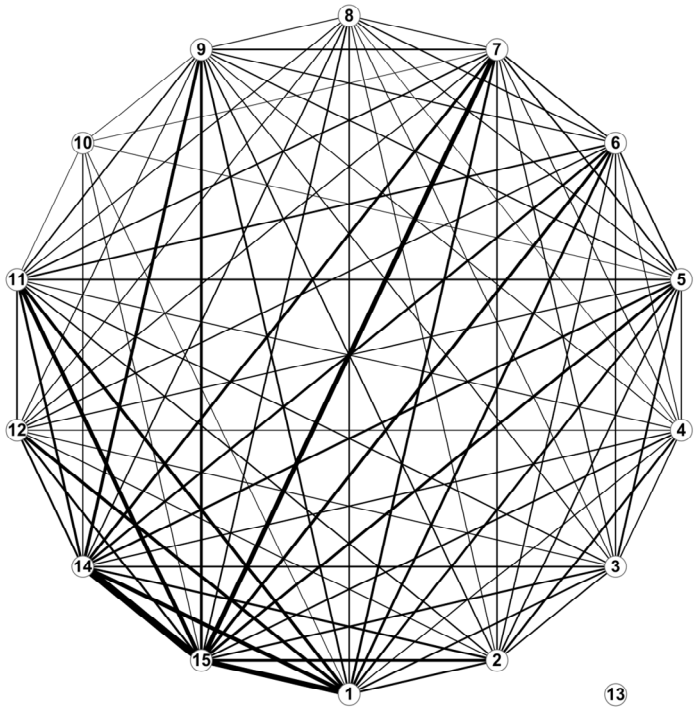


Figure 6. Chemoyl-type activity network derived from ATCMD. Nodes are labeled in numbers. Each number stands for a chemoyl-type. 1: Flavone; 2: Quinone; 3: Benzofuran; 4: Chromene; 5: Coumarin; 6: Lignan; 7: Steroid; 8: Stilbene; 9: Amino Acid; 10: Nucleoside; 11: Saccharide; 12: Tannin; 13: Polyene; 14: Alkaloid; 15: Terpene. An edge, for example ① — ②, means that the molecules are generated by combining chemoyl-types 1 and 2; and the molecules exhibit more than 10 activities.

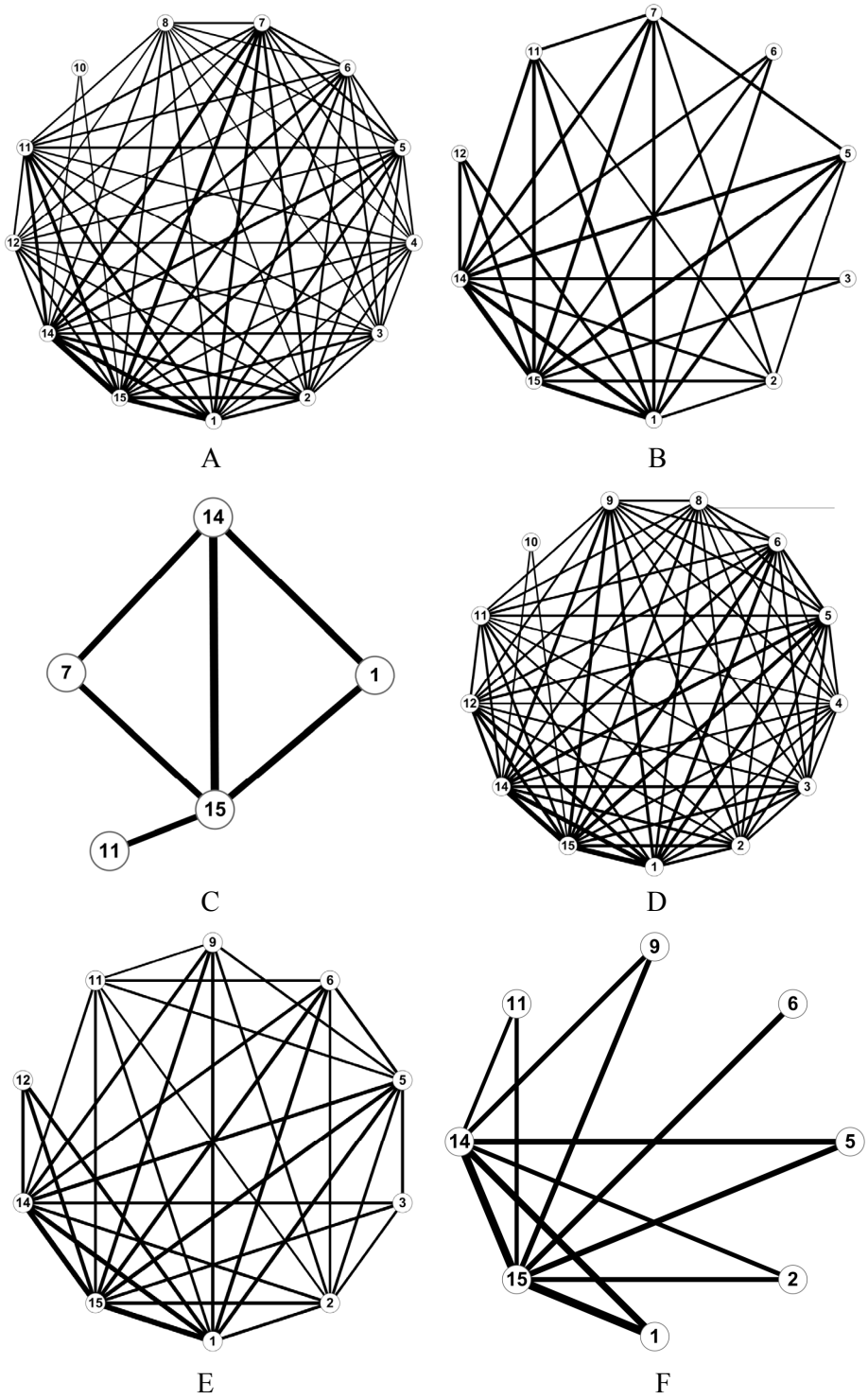


Figure 7. Tri-chemoyl-type activity networks. 1: Flavone; 2: Quinone; 3: Benzofuran; 4: Chromene; 5: Coumarin; 6: Lignan; 7: Steroid; 8: Stilbene; 9: Amino Acid; 10: Nucleoside; 11: Saccharide; 12: Tannin; 13: Polyene; 14: Alkaloid; 15: Terpene. A, B, and C: an edge stands for the new chemoyl types formed from amino acid scaffolds and two other chemoyl types, and which exhibit n (>10) activities. For example: edge ⑨ – ⑧ in Figure 7A means that an amino acid, a stilbene, and a saccharide can form new chemoyl types that exhibit more than 10 bioactivities; edge ①– ⑤ in Figure 7B means that an amino acid, a flavone, and a coumarin can form new chemoyl types that exhibit more than 20 bioactivities; and edge ⑨ – ⑮ in Figure 7C means that an amino acid, a saccharide, and a terpene can form new chemoyl types that exhibit more than 30 bioactivities. C, D, and E: an edge stands for the new chemoyl types formed from steroid scaffolds and two other chemoyl types, and which exhibit multiple (>10) activities.

REFERENCES

1. Newman DJ, Cragg GM. Natural products as sources of new drugs over the last 25 years. Journal of natural products. 2007 Mar;70(3):461-77. PubMed PMID: WOS:000245118800027. English.
2. Robinson MR, Zhang X. The world medicines situation 2011 Traditional medicines: global situation, issues and challenges. Geneva,: World Health Organization; 2011.
3. Xu J, Gu Q, Liu HB, Zhou JJ, Bu XZ, Huang ZS, et al. Chemomics and drug innovation. Sci China Chem. 2013 Jan;56(1):71-85. PubMed PMID: WOS:000313456500009. English.
4. Wallach O. Zur Kenntnis der Terpene und der ätherischen Öele. Justus Liebigs Ann Chem 1887;239:1-54.
5. Ruzicka L. The isoprene rule and the biogenesis of terpenic compounds. Experientia. 1953 Oct 15;9(10):357-67. PubMed PMID: 13116962.
6. Dewick PM. Medicinal Natural Products: A Biosynthetic Approach. New York: Wiley; 2009.
7. Zhou J, Xie G, Yan X. Encyclopedia of Traditional Chinese Medicines. Zhou J, Xie G, Yan X, editors. New York: Springer-Verlag; 2011.
8. wermuth CG. The practice of Medicinal Chemistry. San Diego: Academic Press; 2003.
9. Seigler DS. Plant Secondary Metabolism. New York1998.
10. Koleva, II, van Beek TA, Soffers AE, Dusemund B, Rietjens IM. Alkaloids in the human food chain--natural occurrence and possible adverse effects. Molecular nutrition & food research. 2012 Jan;56(1):30-52. PubMed PMID: 21823220.
11. Herrmann K. Flavonols and flavones in food plants: a review. International Journal of Food Science & Technology. 2006;11(5):433-48.
12. Geier EG, Schlessinger A, Fan H, Gable JE, Irwin JJ, Sali A, et al. Structure-based ligand discovery for the Large-neutral Amino Acid Transporter 1, LAT-1. Proceedings of the National Academy of Sciences of the United States of America. 2013 Apr 2;110(14):5480-5. PubMed PMID: 23509259. Pubmed Central PMCID: 3619328.
13. Leung CH, Chan DS, Kwan MH, Cheng Z, Wong CY, Zhu GY, et al. Structure-based repurposing of FDA-approved drugs as TNF-alpha inhibitors. ChemMedChem. 2011 May 2;6(5):765-8. PubMed PMID: 21365767.
14. Ma DL, Chan DS, Leung CH. Drug repositioning by structure-based virtual screening. Chemical Society reviews. 2013 Mar 7;42(5):2130-41. PubMed PMID: 23288298.
15. Yan X, Gu Q, Lu F, Li J, Xu J. GSA: a GPU-accelerated structure similarity algorithm and its application in progressive virtual screening. Molecular diversity. 2012 Nov;16(4):759-69. PubMed PMID: 23081812.
16. Huang D, Gu Q, Ge H, Ye J, Salam NK, Hagler A, et al. On the value of homology models for virtual screening: discovering hCXCR3 antagonists by pharmacophore-based and structure-based approaches. Journal of chemical information and modeling. 2012 May 25;52(5):1356-66. PubMed PMID: 22545675.
17. Fang J, Huang D, Zhao W, Ge H, Luo HB, Xu J. A new protocol for predicting novel GSK-3beta ATP competitive inhibitors. Journal of chemical information and modeling. 2011 Jun 27;51(6):1431-8. PubMed PMID: 21615159.
18. Zhao W, Gu Q, Wang L, Ge H, Li J, Xu J. Three-dimensional pharmacophore modeling of liver-X receptor agonists. Journal of chemical information and modeling. 2011 Sep 26;51(9):2147-55. PubMed PMID: 21434646.
19. Xu J. GMA: A Generic Match Algorithm for structural Homorphism, Isomorphism, Maximal Common Substructure Match and Its Applications. J Chem Inf Comput Sci. 1996;36:25-34.
20. Nicolaou KC, Pfefferkorn JA, Barluenga S, Mitchell HJ, Roecker AJ, Cao GQ. Natural product-like combinatorial libraries based on privileged structures. 3. The "libraries from libraries" principle for diversity enhancement of benzopyran libraries. J Am Chem Soc. 2000 Oct 18;122(41):9968-76. PubMed PMID: WOS:000090107600012. English.
21. Lindgren D, Sjodahl G, Lauss M, Staaf J, Chebil G, Lovgren K, et al. Integrated Genomic and Gene Expression Profiling Identifies Two Major Genomic Circuits in Urothelial Carcinoma. Plos One. 2012 Jun 7;7(6). PubMed PMID: WOS:000305351700077. English.
22. Nandagopal N, Elowitz MB. Synthetic Biology: Integrated Gene Circuits. Science. 2011 Sep 2;333(6047):1244-8. PubMed PMID: WOS:000294406400040. English.
23. Tajbakhsh S, Cavalli G, Richet E. Integrated Gene Regulatory Circuits: Celebrating the 50(th) Anniversary of the Operon Model. Mol Cell. 2011 Aug 19;43(4):505-14. PubMed PMID: WOS:000294151000004. English.
24. Song LY, Yao HL, Zhu LY, Tong RB. Asymmetric Total Syntheses of (-)-Penicipyrone and (-)-Tenuipyrone via Biomimetic Cascade Intermolecular Michael Addition/Cycloketalization. Org Lett. 2013 Jan 4;15(1):6-9. PubMed PMID: WOS:000313156400003. English.
25. Li C, Dian LY, Zhang WD, Lei XG. Biomimetic Syntheses of (-)-Gochnatiolides A-C and (-)-Ainsliadimer B. J Am Chem Soc. 2012 Aug 1;134(30):12414-7. PubMed PMID: WOS:000306942600023. English.
26. Razzak M, De Brabander JK. Lessons and revelations from biomimetic syntheses. Nat Chem Biol. 2011 Dec;7(12):865-75. PubMed PMID: WOS:000297166200007. English.

27. Khupse RS, Sarver JG, Trendel JA, Bearss NR, Reese MD, Wiese TE, et al. Biomimetic Syntheses and Antiproliferative Activities of Racemic, Natural (-), and Unnnatural (+) Glyceollin I. J Med Chem. 2011 May 26;54(10):3506-23. PubMed PMID: WOS:000290651800005. English.
28. Dethe DH, Erande RD, Ranjan A. Biomimetic Total Syntheses of Flinderoles B and C. J Am Chem Soc. 2011 Mar 9;133(9):2864-7. PubMed PMID: WOS:000289455200019. English.
29. Boone MA, Tong RB, McDonald FE, Lense S, Cao R, Hardcastle KI. Biomimetic Syntheses from Squalene-Like Precursors: Synthesis of ent-Abudinol B and Reassessment of the Structure of Muzitone. J Am Chem Soc. 2010 Apr 14;132(14):5300-8. PubMed PMID: WOS:000276553700060. English.