Finding Precursor Compounds in Secondary Metabolism

Masanori Arita [1] arita@etl.go.jp
Kiyoshi Asai [1] asai@etl.go.jp
Takaaki Nishioka [2] nishioka@scl.kyoto-u.ac.jp

[1] Electrotechnical Lab., Tsukuba-shi Umezono 1-1-4, Ibaraki 305-8568, Japan
[2] Department of Agriculture, Kyoto University, Kyoto-shi Sakyo-ku Kitashirakawa, Kyoto 606-8562, Japan

Abstract

A precursor is a compound which is transformed to a class of functional molecules within short steps. It is an important process in the production of natural drugs to decide whether a given compound is a precursor or not. We present two strategies to select precursor compounds in the secondary metabolism of terpenoids: one is to find the packing of basic molecules in the given cyclic structure, and the other is to find the synthetic map of the given set of compounds. Both strategies play important roles in reproducing tracer experiments on a computer.

1 Introduction

Finding the biosynthetic pathway of a hormone or a natural drug is the key issue for its industrial production. Even a limited increase in its production rate may lead to a drastic change in the synthetic scheme at a commercial level, because of its very low yield from raw materials. For example, the average yield of paclitaxel (a recently approved anticancer drug) from the yew bark is in the range of 0.014–0.017%. About 7.5kg of bark is required to produce 1g paclitaxel [11].

When compared with the metabolism of nucleotides and amino acids, the metabolism of such a compound, called the secondary metabolism, is much less known. Since most drugs are synthesized in a species-specific manner, their synthetic pathway is likely to consist of yet unidentified enzymes. The function of these enzymes may resemble that of already known ones, but there remains a possibility that it has unique function and ligand specificity. For this reason, the prediction of the pathways in the secondary metabolism is prohibitively difficult.

However, there is a solution to a strategic drug discovery. Even without knowledge of the entire synthetic pathway, the yield of a specific compound can be increased by locating its key precursor, i.e. a gateway compound which can be transformed to a class of functional molecules within short steps. Note that a precursor is a structural, not functional, template of its downstream molecules. The enzymatic reaction synthesizing the precursor often forms a bottleneck in its succeeding metabolism. Since metabolism is considered an equilibrium process, over-feeding of a precursor to cultured cells usually produce the increased amount of compounds of its downstream.

Thus, the decision whether a given compound is a precursor or not is important in the production of natural drugs. In this paper, we restrict the target compounds to terpenoids (meaning terpene-like compounds), and show how to decide a given terpenoid is a precursor or not.

Let us briefly introduce terpenoids. Not only a vast class of natural products with medical use, but many fragrant compounds in plants—the essential oils in more daily terms—are formally dissected into C_5 units called isoprene. For example, vitamin A can be synthesized from four isoprenes structures, although isoprene itself can not be the functional unit used by nature (Fig. 1). Such compounds are called terpenes and are classified by the number of isoprene units (e.g. monoterpenes, diterpenes, triterpenes, and so on). Physiologically important compounds such as adrenal hormones, sex hormones, and vitamins are also among the terpenes.

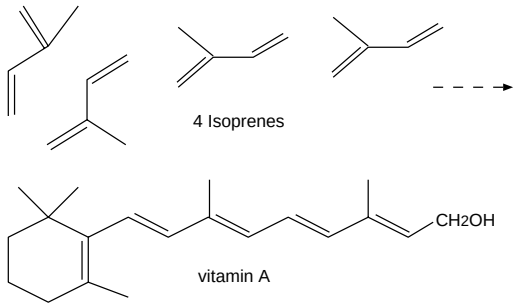

Figure 1: Synthesis of Vitamin A.

The formation of terpenoids is rationalized by a few basic reaction types [15]: (1) generation of a carbonium ion by de-phosphorylation or by opening a double bond, (2) cyclization, or aliphatic shift of a carbonium ion, (3) proton elimination or addition of water. An example is shown in Fig. 2. The important observation is that the connected part of isoprene units is singly bonded. With these observations, we introduce two strategies to find precursors in terpenoid biosynthesis.

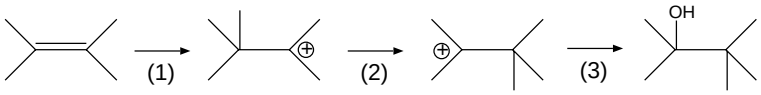

Figure 2: Basic Reactions.

In Section 2, we introduce a strategy to decide whether a given compound G is a polymer of isoprenes. More specifically, the following two conditions are checked for this judgment.

1. G is formally dissected into isoprene units.
2. Adjacent isoprenes in G are connected with single bonds.

In Section 3, we introduce a strategy to find a biosynthetic map (pathway) of a set of structurally similar terpenoids $S = \{G_1, G_2, \ldots, G_s\}$. Given the set of terpenoids S, we compute the structural similarity for all the pair of compounds in S, and find its spanning tree. More specifically, the following conditions are checked.

1. The spanning tree is minimum in the sense of metabolic energy required.
2. The spanning tree is computed with reference to the amount of compounds in a cell.

The result of this paper and the discussion on the future works are given in Section 4.

2 Finding Isoprene Units

The problem of finding basic units in a graph is formulated as the 'generalized matching problem': given a graph G and a fixed pattern graph H, one wants to determine the maximum number of vertex-disjoint copies of H in G. The problem is known to be NP-complete if H contains more than three vertices [10], but for some restricted class of graph, it can be computed in linear-time [14]. It is known that such efficiency can be achieved by decomposing the graph G into a tree, followed by a dynamic programming [4].

Chemical structures do not confine themselves in a well-behaving class of graphs, but the number of cycles is limited, and the degree of nodes is bounded. Moreover, we only need to consider a case

in which H is a tree, because the basic metabolic unit is almost always a tree. [1] Therefore, a good strategy is to open cycles and to apply dynamic programming to the resulting tree structure. That is, each cycle is opened by deleting one of its edges, and the matching procedure is applied to each resulting tree. The matching problem between trees can be solved in linear time [2], and the total computational time becomes polynomial, if the number of cycles and the degree are bounded. This strategy is proven to be a polynomial-time algorithm by Akutsu [1].

Pattern-Packing algorithm

1. Locate small cycles in the structure. Let the number of cycles be k.
2. Delete k edges, one for each cycle, in a combinatorial manner. Apply the dynamic programming between the resulting tree and the pattern tree to find the maximum number of pattern trees, disjointly packed in the structure.

The first step, the detection of cycles, is performed by Horton's algorithm [9]. Its original purpose is to find a minimum cycle basis, and the final Gaussian elimination is required to remove such cycles that can be represented as a conjugation of smaller cycles. This step dominates the computational time of the algorithm, and its order is $O(VE^3)$.

Horton's algorithm

1. Compute the shortest path $\hat{p}_{uv}$ for each pair of nodes u and v in G.
2. For each node w and edge (u, v), create the cycle $C(w, u, v) = \hat{p}_{wu} + \hat{p}_{wv} + (u, v)$ and calculate its length. Degenerate cases in which $\hat{p}_{wu}$ and $\hat{p}_{wv}$ share nodes other than w are omitted.
3. Sort the cycles by length.
4. Consider the cycles as the rows of a binary matrix, whose columns and rows correspond to the edges and the incidence vectors of the cycles, respectively. Perform Gaussian elimination in the order of the length. When enough independent cycles have been found, the process stops.

The second step in the algorithm is the dynamic programming between trees. Assume we have a tree T representing the given molecular structure and the pattern tree H. The dynamic programming carries out a computation in a postorder traversal of the nodes of T. With each node of T, the algorithm associates an array of size $|H|$, each of whose cell represents a node of H, and registers the set of nodes in H which can be matched to the associated node of T. At the leaves of T, only the atomic type is considered in the matching. (That is, carbon is matched with carbon, nitrogen with nitrogen and so on.) Note that the values of each array can be computed in $O(1)$ time, since the size of the pattern tree is fixed. Moreover, if the nodes of a tree are given integer weights, the algorithm can find the matching with maximum total weight. The only modification is that the maximum weight is registered at each array cell together with the matching positions in H. To recover the optimal matching of pattern trees, we can use the usual dynamic programming technique of maintaining back pointers. Since the size of T is linear in the size of G, each iteration of the dynamic programming takes a linear time in the size of G.

2.1 Performance

The algorithm is implemented in C^{++} using LEDA library package [12]. Several sterols are tested for the possibility of isoprene packing. To our surprise, very few of them can be completely dissected into

[1] There are very few exceptions such as lignin. Lignin is formed by oxidative, radical polymerization of coniferyl alcohol, which contains a benzene ring.

[2] We assume that the pattern tree is of fixed size.

isoprenes, i.e. without leaving any carbons. Even lanosterol, the precursor of cholesterol family, cannot be dissected into five isoprenes. This question was rationalized by the introduction of hypothetical protosterol [15]. Its introduction clarifies not only the folding pattern of a carbon chain in cholesterol biosynthesis, but the synthesis of other triterpenoids such as cucurbitanes. Note that the conversion from lanosterol to cucurbitane is harder to occur than that from protosterol to cucurbitane, because of the shifted position of double bonds (Fig. 3). Consequently, protosterol is considered the true precursor of triterpenoids, although it seems to exist only as a reaction intermediate.

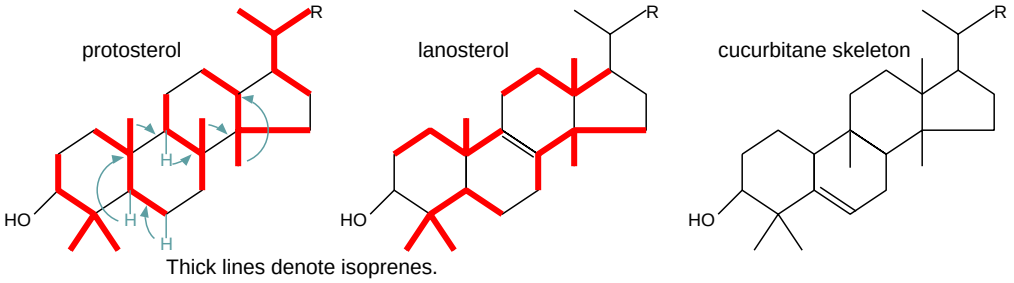

Figure 3: Protosterol can contain five isoprenes, whereas lanosterol or cholesterol can contain at most four (think lines). The arrows of Protosterol are used in the later section.

3 Finding a Synthetic Map

In general, a precursor is selected from two criteria: its amount and structure. First, a precursor is considered to exist more in amount than its derived forms. Second, a precursor is less oxidized, because oxidation generally occurs during the degradation process. These criteria are exemplified in the estimation of the synthetic pathway of paclitaxel [8].

We shall discuss how to apply the minimum spanning tree (MST) algorithm to find the metabolic pathway of a given set of compounds. MST algorithm is one of the classics introduced in many textbooks (See [5] for the algorithmic detail.). It is considered useful in finding biosynthetic pathways for the following reasons.

- It groups compounds in the order of similarity until all the compounds are grouped to a single set. This is the same process as what expert biologists intuitively do in reconstructing pathways.
- There are efficient algorithms for computing sub-optimal solutions [6].

Given a set of compounds, we compute the structural similarity of all the pair of compounds in it. From this all-to-all comparison, we can determine the difference or 'distance' between any two terpenoids. The result can be regarded as a dense graph whose set of nodes is the given compounds, and whose set of edges is the computed distances. In the following, this graph structure is denoted G.

3.1 Structural Distance

In contrast to the reactions in basic metabolism, the enzymatic modification of terpenoids preserves substrates' carbon structure. For example, the tetra-cyclic backbone in androgen and estrogen metabolism is modified only at some positions in a restricted manner (Fig. 4). There are two types of reactions: one that accompanies a conversion between a single and double bond on the backbone, and the other that does not. The latter reaction is basically reversible and low-energy, while the former one may be high-energy and thus irreversible in many cases. In graph G, these two types are distinguished simply by assigning different edge-distances as in Table 1.

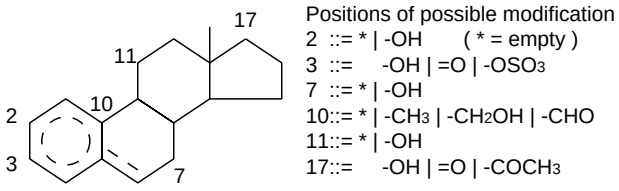

Figure 4: Modification pattern of the tetra-cyclic backbone. Dotted lines are the possible positions of double bonds.

Table 1: The notation A/B indicates a conversion between a chemical group A and a group B. An addition (or elimination) of a group B is denoted /B. In reaction number 5, X denotes a sulfate or carbonate.

No.	conversion	score	No.	conversion	score
1:	/ =O	2.3	3:	/ -CR	14.3
2:	=O / -OH	3.2	4:	-C=C- / -C-C-	6.6
			5:	-OH/ -OX	12.3

Distance Assignment

1. Given two structures, compute their maximum common substructure using a branch-and-bound method [2].

2. Identify the chemical groups modified between them, and sums up the penalty score according to Table 1. This summed score is the edge distance between the two structures.

The reactions occurring in steroid metabolism are covered by the few conversions in Table 1. The distance for each conversion is the reciprocal of its occurrences (frequency) in the 86 reactions in C_{21}-Steroid Hormone metabolism and in Androgen and Estrogen metabolism of KEGG database [13]. For example, the reaction "/ =O" appears in 38 reactions, 86 divided by 38 makes 2.3.

3.2 MST Performance

We apply MST algorithm to the obtained graph G. Interestingly, the compounds appearing in about 90 reactions in steroid metabolism were well grouped by this simple scheme: by edges of short distances (number 1 and 2 in Table 1), compounds in the above two metabolic maps are largely grouped into seven types in Fig. 5 according to their backbone structures. Note that it is easy to find a representative node which is least oxidized within each group. Further application of MST algorithm using longer edges, however, produced several MSTs, different from the *in vivo* pathway.

We need to further consider the amount of compounds in a cell in addition to structures. Although actual experimental measurements are not available, we can hypothesize that the amount of precursor is maximum, and that the amount becomes less and less as the precursor is processed down the metabolic pathway.

Therefore, when the structural distance of an edge in MST algorithm ties, we define that an edge connecting a compound of larger amount has the priority in the MST generation. With this assumption, the MST of seven groups becomes the dotted lines in Fig. 5. There remains a discrepancy from the *in vivo* pathway (thick lines), and its reason is explained in the next subsection.

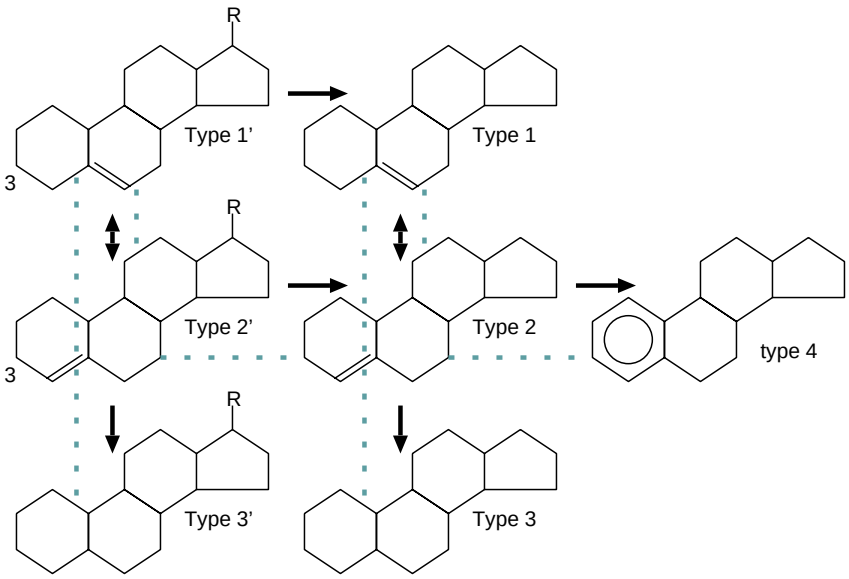

Figure 5: Seven basic backbones of steroids. Solid arrows show the biosynthetic order *in vivo*. Dotted arrows show the connection computed by MST. Type 1: pregnenolones and cholesterols. Type 1': epiandrosterones. Type 2: testosterones. Type 2': progesterones and cortisons. Type 3: androsterones. Type 3': urocortisons and pregnanes. Type 4: estrones.

3.3 Limitation

With the above distance, type 1 and 2 become more distant than type 1 and 3, because of the shift of a double bond. In reality, however, type 1 and 2 are inter-convertible. This discrepancy comes from the lack of consideration of the conjugated bonds and the concerted shift of chemical groups.

Double bonds are called *conjugated* when they are interspersed with single bonds. Conjugated bonds are more stable than distantly located double bonds. For example, the position 3 in type 2' (or 2) in Fig. 5 is likely to be oxidized in =O form, while the same position in type 1' (or 1) is not. The difference comes from the position of the conjugating double bond in the structure.

The other, and more complicated, discrepancy results from a concerted shift of chemical groups. With the above distance, the protosterol and cucurbitane are as different as four methyl groups. In reality, however, these conversions are concerted as the arrows in Fig. 3, and are facilitated with a much lower energy. In fact, the concerted shift occurs in relation with the conjugated bonds too. The conversion from -OH group to =O group at the position 3 in type 1' (or 1) is concerted with a shift of the double bond, resulting in type 2' (or 2).

4 Results and Discussion

4.1 Importance of Tracer Experiments

The powerful laboratory method of establishing a metabolic sequence is the use of isotopes. Radiotracing and mass spectrometry are the most sensitive methods. Radiotracing is used to obtain the quantitative data of isotope incorporation in compounds. Its shortcoming is that the information of the location of labels in each compound remains ambiguous. Fortunately, when a single compound is under focus, NMR spectroscopy provides data for the efficient structural determination. However, as the structure gets larger, it becomes more difficult to analyze its NMR data. If a computer simulation of metabolism can generate possible locations of incorporated isotopes, it would bridge the outputs of

these two laboratory schemes and boost the study in metabolism. This is the reason that we presented two strategies for analyzing chemical structures in terpenoid biosynthesis.

Our first strategy was to find the packing of basic molecules in the given cyclic structure. We showed that this procedure can be computed in polynomial time, and demonstrated the appropriateness of protosterol as a precursor compound.

The computational analysis of the packing pattern helps us to understand where the labeled atoms are incorporated in a structure in tracer experiments. The current algorithm, however, answers yes or no only. Ideally, when the answer is no (i.e. when a given compound cannot be dissected into isoprenes), the algorithm should *suggest* possible precursors of a similar structure. Therefore, an approximate dissection problem is our future work. We also note that this analysis is valuable also for alkaloids, another large class of cyclic molecules synthesized from amino acids through still unknown pathways.

Our second strategy was to find the synthetic map of the given compounds. Taking sterols as an example, we showed that their biosynthesis can be mostly reproduced with a simple MST algorithm. We also clarified the future work to be done: the consideration of the conjugate bonds, and the concerted shift. Also, the utilization of k best MSTs is another aspect of future work.

The biosynthetic map should be synchronized with the computational tracer experiment, by digitizing the atomic movement. We already achieved the digitization of each reaction by finding maximal common substructures between substrates and products. This is the first step of the fully automated simulation of tracer experiments, and the strategic drug discovery by locating key precursors.

4.2 AMR Project

Both algorithms are implemented in the Automated Metabolic Reconstruction (AMR) project, whose aim is the prediction of unknown or alternative metabolic pathways by computational simulation [3]. Its next major step is to formalize unknown enzymatic reactions. It seems impossible to predict a function of totally unique and novel enzymes, but most enzymes perform basic reactions such as oxidation or carbon transfer. We consider that the prediction of alternative pathways is possible by digitizing such basic reaction schemes.

Acknowledgment

The authors thank Dr. Shigeru Yamane and Dr. Nobuyuki Ohtsu, department heads at ETL, for supporting this collaborative work between ETL and Kyoto Univ. The work is partly supported by Grant-in-Aid for Scientific Research on Priority Area "Genome Informatics," from Ministry of Education, Science, Sports and Culture, Japan.

References

[1] Akutsu, T., A polynomial time algorithm for finding a largest common subgraph of almost trees of bounded degree, *IEICE Trans. Fundamentals*, E76-A(9):1488–1493, 1993.

[2] Arita, M., Graph modeling of biological mechanisms, *J. Japan. Soc. Artif. Intel.*, under submission.

[3] Arita, M., Automated metabolic reconstruction: theory and experiments, Ph.D Thesis, Unviersity of Tokyo, 1999.

[4] Bern, M.W., Lawler, E.L., and Wong, A.L., Linear-time computation of optimal subgraphs of decomposable graphs, *J. Algorithms*, 8:216–235, 1987.

[5] Cormen, T.H., Leiserson, C.E., and Rivest, R.L., *Introduction to Algorithms*, MIT Press, 1990.

[6] Eppstein, D., Finding the k smallest spanning trees, *BIT*, 32:237–248, 1992.

[7] Garey, M.R. and Johnson, D.S., *Computers and Intractability*, W.H. Freeman and Company, 1979.

[8] Hezari, M. and Croteau, R., Taxol biosynthesis: an update, *Planta Medica*, 63:291–295, 1997.

[9] Horton, J.D., A polynomial-time algorithm to find the shortest cycle basis of a graph, *SIAM J. Comput.*, 16(2):358–366, 1987.

[10] Kirkpatrick, D.G. and Hell, P., On the completeness of a generalized matching problem, *Proc. 10th ACM Symp. on Theor. on Comput.*, 240–245, 1978.

[11] Patel, R.N., Tour de Paclitaxel: biocatalysis for semisynthesis, *Annu. Rev. Microbiol.*, 98:361–395, 1998.

[12] Mehlhorn, K. and Näher, S., LEDA: a platform for combinatorial and geometric computing, *Comm. Assoc. Comput. Mach.*, 38(1):96–102, 1995.

[13] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M., KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, 27:29–34, 1999.

[14] Takamizawa, K., Nishizeki, T., and Saito, N., Linear-time computability of combinatorial problems on series-parallel graphs, *J. Assoc. Comput. Mach.*, 29(3):623–641, 1982.

[15] Torssell, K.B., *Natural Product Chemistry 2nd Ed.*, Apotekarsocieteten, 1997.