Published online 10 April 2013

Nucleic Acids Research, 2013, Vol. 41, No. 11 ***e113***
doi:10.1093/nar/gkt234

Metabolic tinker: an online tool for guiding the design of synthetic metabolic pathways

Kent McClymont[1] and Orkun S. Soyer[2,*]

[1]Computer Science, College of Engineering, Mathematics, and Physical Sciences, University of Exeter, Exeter EX4 4QF, UK and [2]Systems Biology Program, College of Engineering, Mathematics, and Physical Sciences, University of Exeter, Exeter EX4 4QF, UK

Received October 23, 2012; Revised March 14, 2013; Accepted March 15, 2013

ABSTRACT

One of the primary aims of synthetic biology is to (re)design metabolic pathways towards the production of desired chemicals. The fast pace of developments in molecular biology increasingly makes it possible to experimentally redesign existing pathways and implement *de novo* ones in microbes or using *in vitro* platforms. For such experimental studies, the bottleneck is shifting from implementation of pathways towards their initial design. Here, we present an online tool called 'Metabolic Tinker', which aims to guide the design of synthetic metabolic pathways between any two desired compounds. Given two user-defined 'target' and 'source' compounds, Metabolic Tinker searches for thermodynamically feasible paths in the entire known metabolic universe using a tailored heuristic search strategy. Compared with similar graph-based search tools, Metabolic Tinker returns a larger number of possible paths owing to its broad search base and fast heuristic, and provides for the first time thermodynamic feasibility information for the discovered paths. Metabolic Tinker is available as a web service at http://osslab.ex.ac.uk/tinker.aspx. The same website also provides the source code for Metabolic Tinker, allowing it to be developed further or run on personal machines for specific applications.

INTRODUCTION

Synthetic biology aims to apply engineering principles to (re)design natural biological systems or engineer *de novo* ones from reliable and pre-developed 'parts'. Within this paradigm, synthetic (re)engineering of metabolism is potentially the most successful and promising research area. In particular, the synthetic implementation and redesign of natural pathways in microbes has allowed novel production of, or increased yields in, several medically and industrially relevant compounds including artemisinin (the precursor of a malaria drug), alkanes and taxol (a cancer drug) (1–3). Characteristically, these studies first identify natural pathways in specific species and then alter their dynamics (e.g. over expression of bottleneck enzymes) or redesign them (e.g. implementation of some or all enzymes in a different organism). This reliance on *a priori* knowledge of pathways, however, is not in line with a traditional engineering pipeline, where one would start from parts (enzymes in this case) to draw up a completely *de novo* pathway. While the complexity of biological systems might never permit this kind of engineering, it is desirable to attempt developing appropriate 'system design' tools to achieve the true potential of synthetic biology (4).

Through developments in sequencing technologies and systems biology, the knowledge of biochemical reactions that are present in different organisms is increasing at a fast pace. Currently, databases such as KEGG (5) compile metabolic reactions identified in the literature from a large number of organisms. The use of this data, however, is mostly confined to analyses that aim to catalogue metabolic pathways in different organisms, or developing genome-level metabolic models (6). In the former category, for example, tools such as PathComp [available under the KEGG database (5)] and Rahnuma (7) concentrate on searching for existing metabolic paths in the context of specific species. This species-centred view is not in line with the aims of synthetic biology, where emerging molecular techniques increasingly allow 'implementation' of desired enzymes in any cellular context and creating 'patched' pathways [e.g. the re-implementation of the artemisinin pathway in yeast (2)]. Within this context, any approach for the design of *de novo* synthetic pathways should consider the entire known universe of biochemical reactions.

At present, the CHEBI and RHEA databases offer one of the most comprehensive collection of known metabolic reactions and reaction directionality (8,9). At the time of writing, these databases list ~20 000 compounds and

*To whom correspondence should be addressed. Tel: +44 1392 723615; Fax: +44 1392 217965; Email: O.S.Soyer@exeter.ac.uk

30 000 reactions. This data can be represented as a hyper-graph, where nodes and edges correspond to compounds and biochemical reactions, respectively. We can refer to this hyper-graph arising from the known universe of biochemical reactions as the 'Universal Reaction Network' (URN). Within the URN, possible biochemical paths that can result in the conversion of one compound into another can be found by using standard search algorithms developed in computer science and graph theory. These algorithms, in their generalized form, however, are not the most efficient for searching specialized large highly connected networks such as the URN, as they do not in all cases account for associated context-specific knowledge such as thermodynamics information on reactions. Tools like From Metabolite to Metabolite (FMM) (10) and Metabolic Route Search and Design (MRSD) (11) that approach the 'pathway design' through searching known reactions might consider the whole URN, but do not exploit the full context of available information such as reaction thermodynamics.

An alternative approach (or possible extension) to mining the URN and these reaction databases with search algorithms is to encapsulate and generalize the stored information by distilling a set of rules, which describe core biochemical reactions and conversions. Using these rules, it is then possible to reconstruct the database and even predict new reactions (based on the rules) and construct novel pathways. This approach has been developed successfully in the Biochemical Network Integrated Computational Explorer (BNICE) framework (12–14) and extensions of BNICE (15) as well as related tools such as RDM (16), META (17) and Pathway Prediction System (PPS) (18). The resulting paths can feature known reactions or completely novel ones (based on the reaction rules used). The use of reaction rules gives this approach the potential for the discovery of paths composed of novel reactions; however, at the same time, it creates a limitation for its broad application. Even for BNICE, which is the most developed tool applying this approach, the method requires some manual curation, which naturally limits the number of compounds and reactions available for searching to just select subsets. For example, in a recent study (19), the 86 reaction rules generated by BNICE cover only ~50% of the reactions contained within the KEGG database. Thus, BNICE could be applied only in the discovery of pathways between a select set of compounds. The limited search space resulted in the application of this approach only to study-specific metabolic conversions [see (12–14)].

Here we present an online tool called Metabolic Tinker, which develops the approach of searching the known set of reactions for guiding the design of synthetic metabolic pathways. It differs from previous tools using this approach (e.g. Rahnuma, FMM and MRSD) by extending the search space to the URN and implementing a specialized heuristic search algorithm that uses the embedded thermodynamics and compound similarity information to search the URN for thermodynamically feasible metabolic paths among two user-defined 'source' and 'target' compounds. To build the URN, Tinker periodically compiles an internal database of metabolic compounds and reactions using the CHEBI (9) and Rhea (8) databases. These recently developed databases have a wider coverage compared with the commonly used KEGG database (5), and are maintained as free and fully curated resources, which are under continual development. We show that Tinker is capable of finding both naturally existing paths among given compounds and many more alternative and thermodynamically feasible paths. In many cases, these paths contain ones that are not discovered by the currently available tools that are based on the same approach of searching the known reactions (see Table 2). We also show that Tinker's performance is comparable and complementary to that of tools using the reaction rule–based approach (e.g. BNICE).

MATERIALS AND METHODS

Creating the URN

The URN that Tinker searches is produced from a periodically updated internal database of metabolic compounds and reactions available from the CHEBI (9) and Rhea (8) databases. At the time of writing, this internal database contains ~20 000 reactions and 29 000 compounds, which have been manually qualified against references to the literature. The Rhea database provides a large set of curated metabolic reactions, which in many cases are annotated with the known direction of the reaction. Using this information a partially directed network (i.e. the URN) is produced, where reactions with missing direction information is completed using a prediction of the reaction thermodynamics, described in the next section below. The URN is shown in Figure 1. Similar to networks created from metabolic data of single species (20), the URN has a connectivity distribution that could be described by a power law with an average degree of 3.2 and peak degree of 4.9. Thus, the URN contains certain hubs, i.e. highly connected compounds such as cofactors ADP and NADH+. For the purposes of Tinker, the inclusion of these compounds in the pathway search is not feasible, as their usage as intermediary compounds in possible synthetic paths is not realistic and their presence exponentially increases the search time. With these considerations, we exclude the most highly connected compounds (those with a degree >650) from the search process, and also excluded a number of well-known cofactors (such as dATP) from being included as primary compounds in the pathways. It should be noted that these compounds can still be given as source and target compounds. The full list of excluded compounds is given in Table 1 and provided on the Tinker website and can be altered in the source code of the program.

Thermodynamic information on reactions

Where possible, the free energy change of reactions is calculated from the free energy of formation (ΔG^0) of the participating compounds. The latter is predicted using the group contribution method (21), which uses the known measured energy for a set of common small chemical structures, from hydrogen to functional groups, as a basis for predicting free energies of complex

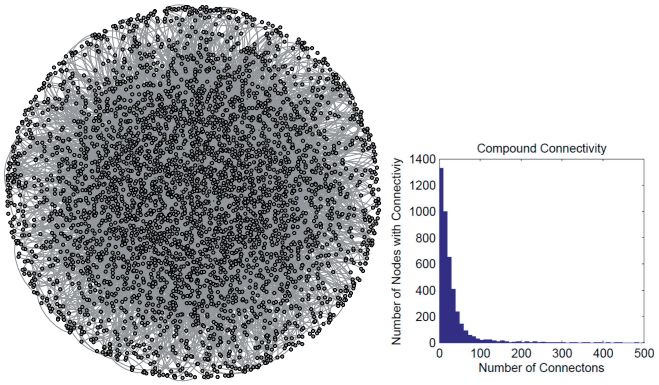

Figure 1. A graph representation of the URN to illustrate the scale and density of this directional graph. Nodes represent compounds, and edges represent reactions between these compounds. The more connected compounds are clustered in the centre of the graph and the lesser connected nodes at the extremities. The connectivity distribution of the nodes is shown in a histogram in the inset, which can be fitted to a power law distribution with an exponent of 3.4.

compounds (22). Effectively, the method predicts a chemical compound's free energies by adding together the known energy values for the constituent parts, which make up the more complex compound structure. The group contribution method does not allow calculating free energies for all the compounds. For cases where the free energy cannot be calculated for a compound in a reaction, the reaction is taken as a bidirectional link.

The energy change in reactions was then calculated by computing the sum of the free energies of compounds on either side of the reaction and determining the difference. If the absolute value of this difference is >10 and the reaction is not annotated in Rhea with a direction, then the reaction is assigned a direction going from the higher ΔG^0 to the lower ΔG^0. In the original article developing the group contribution method (21), the largest difference between the predicted and measured compounds was $\sim \pm 7.5$ kcal/mol. The additional calculations for reaction ΔG^0 took into account an analysis of environmental conditions, such as substrate concentrations, which were shown to affect the accuracy of the predictions (21). Therefore, a threshold of ± 10 kcal/mol was introduced to account for this variance in reliability, while still providing a good means for predicting reaction directionality. The thermodynamics of all compounds (including cofactors) in each reaction is included in the Tinker search algorithms and analyses.

It should be noted that the group contribution method cannot predict free energy of formation of certain compounds (21), limiting the calculation of thermodynamics for the reactions contained in URN. For cases where the ΔG^0 cannot be calculated for a compound in a reaction, the reaction is taken as a bidirectional link (unless directional information is given in the Rhea database). Reactions that contain compounds lacking thermodynamics information that are assigned bidirectionality in this way constitute 18% of all reactions in the URN. The thermodynamics information is used to provide a prediction of the direction of reactions that have not been annotated in Rhea and ensures only feasible reactions occur in the pathways produced by the search. In addition to completing the directionality of edges in URN, the predicted thermodynamics are an essential part of the heuristic search (see below) and provide a means of ranking the results.

Table 1. A table of restricted compounds, which are prevented from being used as primary compounds in the Tinker search tool at the time of writing

Compound	CHEBI Id	Number of reactions
H(+)	15378	11 269
H_2O	15377	7472
O_2	15379	2800
NADPH	57783	2520
NADP(+)	58349	2520
NAD(+)	57540	2144
NADH	57945	2128
ATP	30616	1830
Phosphate	43474	1618
$HP_2O_7^{(3-)}$	33019	1562
ADP	456216	1461
CO_2	16526	1350
CoA	57287	1136
NH4(+)	28938	961
S-adenosyl-L-methionine	59789	679
GTP	37565	140
GDP	58189	188
UDP	58223	568
dATP	61404	8
AMP	456215	457

These compounds have been excluded based on their connectivity and presence as well-known cofactors (such as dATP). Compounds with >650 incoming or outgoing reactions are automatically added to this list. This is also the current list used on the online implementation of Tinker (i.e. the Tinker website). Users can modify this list in the Tinker source code, which can be downloaded from the website.

Compound similarity

In addition to calculating thermodynamics, Tinker uses a compound similarity measure to calculate the similarity of substrates to the target compound in all reactions in a pathway. The similarity among each of the compounds in URN is calculated using a graph comparison technique similar to that presented in (23). As with the group contribution method (21), the compound similarity algorithm uses functional groups of atoms and bonds that make up the two compounds as a basis for calculating the number of differences in the structure of these functional groups. The algorithm calculates the ratio of functional groups that match information across the two compounds compared with those atoms and bonds that are not common in both structures.

This similarity score is implemented as part of the heuristic, to guide the search process (as detailed in the next section). Further, a similarity threshold can be set to prevent the inclusion of reactions that use substrates that are too similar to the target compound on the premise that if substrates are available in the system that are structurally similar to the target, then the path search would likely start from those compounds (see also below). This provides a filtering mechanism that 'sanitizes' results

Table 2. Comparison of Tinker's performance against that of Rahnuma (7), MRSD (11) and FMM (10)

Source	L-arginine (32682, C00062)	(R)-mevalonate (36464, C00418)	D-erythrose 4-phosphate (16897, C00279)	Pyruvate (15361, C00022)	IPP (128769, C00129)
Target	L-citrulline (57743, C00327)	amorpha-4,11-diene (52026, C16028)	3-amino-4-hydroxybenzoate (60005, C12107)	3HP (16510, C01013)	beta-Carotene (17579, C02094)
Max Length	3	7	7	6	7
Reference	26	2	27	19	28
Tinker	4 paths (100 unique paths)	4 paths (6291 unique paths)	20 paths (36 936 unique paths)	24 paths (5457 unique paths)	11 paths (6552 unique paths)
Rahnuma	34 paths	3 paths	0 paths	0 paths	0 paths
MRSD	2 paths	0 paths	0 paths	5 paths	0 paths
FMM	3 paths	0 paths	0 paths	1 path	0 paths

For a given pair of target and source compounds of interest (see related references for natural pathways among these compounds), a search is performed with the given pathway length cut-off. The CHEBI and KEGG IDs are given in parentheses with each compound. The number of unique paths returned by each search program is given in the final row. Some or all known pathways are found in all cases where pathways are returned by each respective search algorithm. Metabolic Tinker results are those obtained from running the heuristic search algorithm with the online server's time constrains (see Heuristic search algorithm, MATERIALS AND METHODS) and using a similarity threshold of 0.8. MRSD, Rahnuma and FMM are run using their default settings and as made available on their corresponding websites. For all the search results that are summarized in this Table, please see Supplementary Material.

and ensures only useful results are returned by the search. As discussed below, the similarity threshold is a user-defined setting and can be turned off (set to 1).

Heuristic search algorithm

Tinker implements a novel heuristic search algorithm designed specifically for searching the URN and which is capable of finding a large number of possible paths between two given compounds. Although a wide range of graph search heuristics and algorithms exist, these do not take advantage of contextual information, such as thermodynamics, to improve the search efficiency. As already stated, the URN is a highly connected hypergraph, which, assuming each compound is on average connected to 20 other compounds, through combinatorial expansion contains ~28 000 000 potential pathways of length 5 between two compounds. For longer pathways, this quickly becomes an infeasible number to enumerate.

To improve the likelihood of Tinker locating a good number of thermodynamically feasible pathways within a limited time, we devised a heuristic search algorithm that uses weighted edges and a distance measure to guide the heuristic, where the heavier weighted edges and compounds close to the target are searched first. In many cases, this results in the search locating the target more quickly as less favourable, and less likely pathways are searched last. The heuristic uses reaction thermodynamics to weight the edges and moves to nodes (i.e. compounds) that are more thermodynamically favourable and structurally similar to the target compound. To break ties between two or more links that have the same thermodynamic score, a secondary weighting is used, which ranks these available connections by the number of available reactions that enable them. In other words, the tie between equally feasible reactions is broken by moving to compounds that are accessible from the current compound via more reactions. The pseudocode for the heuristic is given in Figure 2 and given more visually in a schematic diagram in Figure 3. This heuristic algorithm significantly improves computation time and is more likely to return results when search time is limited, even when searching a significant fraction of all possible paths for larger pathway lengths set (see Table 3).

User-defined parameters

Tinker provides a set of user-settable search parameters, which allow the user to more effectively filter the results and control the extent of the search.

Pathway length

The 'pathway length' setting determines the range of graph 'exploration' that Tinker performs starting from the source compound (i.e. node). It dictates 'depth' of the search by setting a limit on the number of intermediate compounds allowed between the source and target in a path. This setting has a direct impact on the complexity and completion time of the search. Table 3 illustrates this explosion by showing the search results as the pathway length is increased. The default choice is a pathway length of 5 with a maximum of 20.

Similarity threshold

Despite including thermodynamical information, the search and enumeration of paths between two given compounds can result in pathways that do not make sense from biochemical and design perspective. In particular, a graphically and thermodynamically feasible path might involve substrates that are structurally highly similar to the desired target compound. To limit the appearance of such paths, Tinker allows evaluating for each path, the similarity between substrates involved in each of the reactions in that path and the target compound. This filtering mechanism is used to reduce the number of illogical pathways being returned to the user and is achieved by checking, for each reaction in a pathway, whether any of the reactions use cofactors that are chemically similar to the desired target compound. If a reaction uses a cofactor that is more similar to the target than a given user-set threshold, then the reaction and associated pathways are excluded from the search and results. This prevents the

```
function EXPANDNODE(Compound c, Path p)
    linkingReactions ← c.reactions
    heuristicSort(linkingReactions)
    for all r ∈ linkingReactions do
        if r links to target then
            storePath(p ∪ target)
        else if r is a valid reaction then
            for all n ∈ r.compounds do
                expandNode(n, p ∪ n)
            end for
        end if
    end for
end function
```

Figure 2. Psuedo code listing of the Tinker heuristic search. The algorithm describes how Tinker searches the graph from the source compound c by retrieving and sorting all the reactions that link out from the current compound c. The search algorithm then checks each of the sorted reactions in order and if it links to the target compound adds the current path to the list of valid paths. Otherwise, the algorithm checks if the reaction is valid for searching (i.e. all the substrates are acceptable and not too similar to the target) and then recursively calls itself for each of the products produced in the current reaction.

search from traversing links between compounds that are created by reactions that require compounds as substrates that are almost identical to the target compound. The threshold varies between 0 and 1, where a setting of 1 allows even identical compounds to the target to be incorporated into the pathway (i.e. would result in Tinker returning all possible paths). The default similarity threshold is 0.5.

The reasoning behind this filtering mechanism was to allow Tinker to return paths that are more logical from a design perspective. Consider a theoretical example, where Tinker is used without such filtering to search for paths between a source S and a target T. If there was a path from S to a compound X, and a reaction such as $Z + X \rightarrow T$, where Z is highly similar to T, then Tinker would return a path from S to T, through X, as a plausible solution. In practice, this path would not be useful because availability of Z would defy the logic of building such a path. The implemented filtering option (based on compound similarity) allows ruling out the possibility of Tinker returning such 'trivial' and illogical paths. Furthermore, the compound similarity is used to guide the heuristic search (as explained in the main text) and thereby reducing computation times.

Outputs

Tinker returns the found paths in html, comma separated, tabulated and graph formats. The results are all ranked (ascending) ΔG^0 for the overall pathway. Given the large number of results produced by Tinker, the ranking by thermodynamics is an important step in making the results more accessible to the user. Indeed, similar search methods such as (24) identify the importance of ranking and filtering results in pathway-prediction tools like Tinker.

For the first three output formats, each of the listed paths start from the source compound, listing the intermediary compounds and reactions at each step and ending

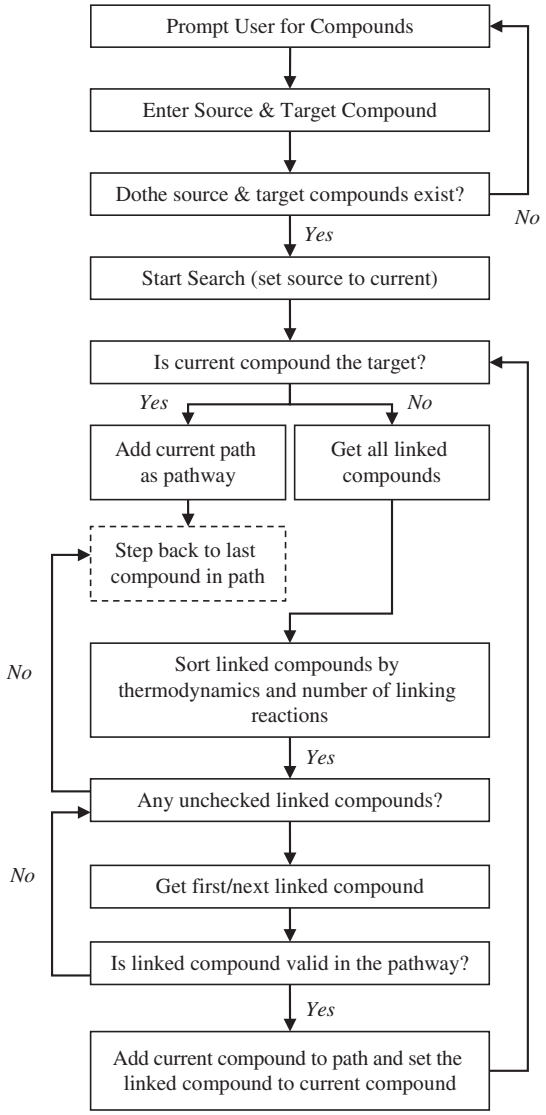

Figure 3. Schematic representation of the search algorithm. This figure shows the recursive process flow of the algorithm, where the linked compounds from each compound are searched using a depth-first approach. The depth-first search is guided using a heuristic sorting algorithm, which considers the number of reactions linking the two compounds and the thermodynamic feasibility of the connecting reactions.

with the target compound. This list distinguishes between a 'path' and corresponding 'unique paths'; the former corresponds to all metabolic paths that consist of the same set of intermediary compounds and all possible reactions linking these, while the latter corresponds to paths with specific reactions linking these intermediate compounds. The paths, along with the number of unique paths they entail, constitute the main html output. An additional tabulated list can be downloaded that lists 10 unique paths for each path, selected by their ATP and NADP usage. The graph output format provides a visual representation of found paths in the GEXF format, which can be read with several open-source network visualization and analysis tools including Gephi (25). Within this graphical output, the source and target compounds are

Table 3. Comparison of the computing time and search performance of the heuristic search algorithm implemented in Tinker against that of a standard depth-first search

Target–source pair	Pathway length searched	Depth-first search	Heuristic search
Arginine–Citrulline (urea cycle)	3	58	58
Arginine–Citrulline (urea cycle)	5	113[a]	176[a]
Arginine–Citrulline (urea cycle)	10	104[a]	189[a]
Arginine–Citrulline (urea cycle)	20	124[a]	206[a]

Arginine and Citrulline, which constitute part of the natural urea cycle, were used as source–target pair as a sample search case. Computing time is approximated as number of pathways found (not unique pathways, which would be considerably higher) in searches limited to 10 s run time and using different pathways lengths (results averaged over 10 runs). For all searches, the similarity threshold was set to 1.
[a]Runs that timed out.

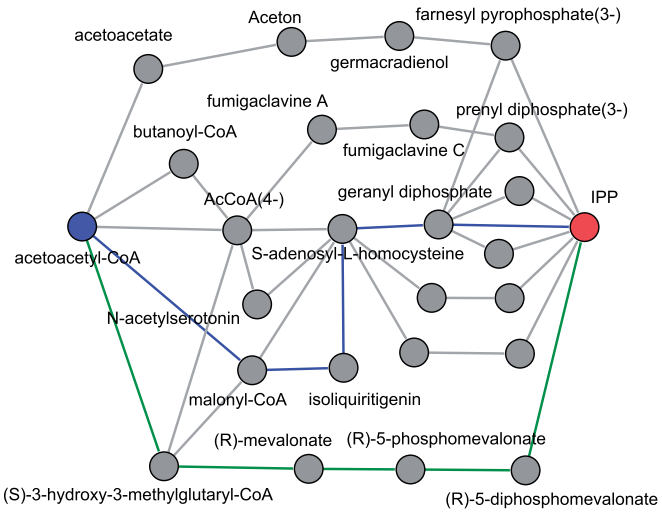

Figure 4. Graph of pathways from acetoacetyl-CoA to IPP. The naturally occurring pathway is shown at the bottom of the figure in green. An entirely novel pathway is also shown (in blue), which has good thermodynamic properties ($\Delta G^0 = -52.6$), second only to the natural pathway. This path is acetoacetyl-CoA → malonyl-CoA → isoliquiritigenin → S-adenosyl-L-homocysteine → geranyl diphosphate → isopentenyl diphosphate.

colour-coded, all compounds are labelled with chemical names and the reactions connecting them are included with their reaction identifiers given as edge labels. A maximum of three reactions are shown for any one compound-to-compound link to make the results more readable. The edges are weighted by number of linking reactions and using edge width as a visual indicator of this weighting (see Figure 4). An image of the results is also provided on the Tinker website, which gives a quick view of the GEXF graph.

The Tinker paths are summarized and returned in two forms: pathways and unique pathways. A unique pathway represents an individual pathway from one compound to another via some intermediate pathways using a specific unique combination of reactions. In many cases, there are multiple reactions that link two compounds in the URN and so, in such cases, there exist multiple unique pathways, which use the same intermediate compounds but different reactions. These sets of unique pathways can be grouped into a single more general pathway in which all share the same intermediate compounds. In other words, for each pathway, there may be multiple unique paths that represent different combinations of reactions but the same combination of compounds.

Availability and requirements

Metabolic Tinker is publicly available at http://osslab.ex.ac.uk/Tinker.aspx. The tool can be used as an online service via a web submission form, or its source code in Java can be downloaded and run on a personal computer or adapted for specific applications. The program uses indexed data to reduce computational burden; however, the computer should have at least 1 GB of memory for the best performance. The server version puts limits on search time (searches terminated after 10 s) and output size (output terminates after 30 s).

RESULTS AND DISCUSSION

Unlike other available metabolic pathway search tools, Tinker uses the entire known metabolic reaction set (the URN) to perform its searches among given compounds. Tinker implements a specific heuristic search, which is capable of finding large numbers of paths within a limited search time (see 'Materials and Methods' section). Its broader search scope and the novel search algorithm it implements allow Tinker to find the known paths among given compounds as well as many novel ones that are thermodynamically feasible (see Figure 4 for an example). In all the test cases we have analysed, Tinker returned a larger number of paths compared with the existing pathway search tools (Table 2). It must be noted that some of the paths found by Tinker might involve reactions that are not all present in a single species. This should only facilitate more applications of *de novo* metabolic system development, which will be increasingly possible with the development of experimental tools such as genetic switchboards in synthetic biology (29).

De novo synthetic paths

To illustrate how Tinker can facilitate synthetic design of pathways of interest, we discuss two simple example searches. In the first one of these, we look at the Mevalonate pathway, and particularly search for paths involving the conversion of the Acetoacetyl-CoA to 3-isopentenyl pyrophosphate (IPP) (Figure 4). Using the Tinker search tool, all possible pathways were found between this source and target pair, with a maximum of 5 intermediate compounds allowed and no similarity cut-off (i.e. similarity cut-off = 1). The results of this search are shown in Figure 4, where the naturally observed pathway is shown at the bottom of the graph. The search returns 22 pathways, involving a total of 26 compounds and representing 13 443 unique pathways.

Exploring all these possible unique paths experimentally may not be feasible; however, these paths provide clear

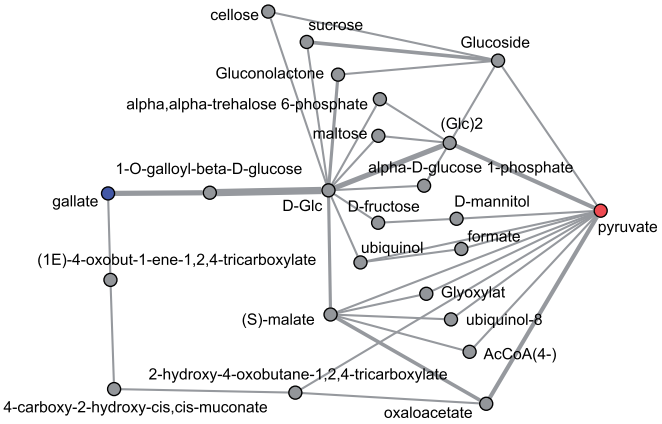

Figure 5. A graph showing the combined pathways from gallate (blue node) to pyruvate (red node). The edge weight depicts the number of reactions connecting the two compounds, where thicker lines indicate larger number of possible reactions and potentially a more feasible pathway.

ideas for both novel synthetic design of the Acetoacetyl-CoA to IPP conversion. In particular, we discover several alternative paths that use 'malonyl-CoA' and 'geranyl diphosphate' as intermediate compounds, rather than mevalonate. Interestingly, this pathway is the second most thermodynamically favourable pathway (after the natural pathway), which indicates that alternative and thermodynamically feasible pathways can be found that use a distinct set of compounds and reactions to transform Acetoacetyl-CoA into IPP.

A second example is shown in Figure 5. In this case, we performed a search between gallate (a polyphenolic compound) and pyruvate with a similarity cut-off 0.5 and path length 4. This returns 26 paths and 35 406 unique paths (or 31 paths and 45 450 unique paths when similarity cut-off is set to 1). We find that these 26 paths fall mostly into two classes. One of these (represented by ~2 paths and 468 unique paths) corresponds to the known naturally occurring gallate degradation path (30) and its slight variations in terms of reactions used. The two paths differ in their intermediate compounds in only one instance. The other class of paths (24 paths and ~35 000 unique paths) result from the assimilation of gallate into 1-O-galloyl-beta-D-glucose and subsequent production of glucose. Once glucose is produced, this can lead into pyruvate through a range of paths that make up this second class of results. While these paths do not directly represent a chemical conversion from gallate to pyruvate (i.e. the carbon atoms from the former does not end up in the latter), they present an interesting alternative to the natural gallate degradation, which involves using it as an enabling compound linking into diverse paths leading to pyruvate. Thus, this example illustrates both discovering primary natural paths and thermodynamically feasible alternative solutions that are not immediately apparent. We note that the required additional substrates for the gallate assimilation and subsequent reactions are compounds such as UDP-D-glucose, riboflavin and phosphoenolpyruvate, which are all plausible to include in a reaction system.

These simple examples demonstrate how Tinker can aid the adaptation of existing and/or design of entirely new metabolic pathways.

Searching URN in a feasible way

By extending the search scope to the entire known metabolic reaction space, Tinker is faced with much more complex search tasks compared with those set in species-specific reaction domains. This is exemplified in sample results shown in Table 3; as the allowed path length increases from 5 to 20 for a 'source'–'target' coupling from the urea cycle, the increasing search scope results in a standard depth-first search finding fewer pathways in a given search time. As simply increasing computing time might not allow overcoming this issue in a feasible way, an alternative is to reduce the search complexity. One approach is to facilitate standard search approaches by biologically motivated heuristics. Tinker implements one such heuristic that is based on thermodynamics and compound similarity (see 'Materials and Methods' section and Figure 2) which, in these results, produce a greater number of pathways compared with a standard depth-first search and is more robust to increasing the search space (path length), shown in Table 3. While the heuristic approach may in some cases limit the final search results to only a subset of all possible paths among the source and target, it provides a significant improvement in efficiency over off-the-shelf search methods (see Table 3).

Addressing limitations of graph-based search approaches

As with many other existing metabolic pathway search tools, Tinker uses a graph representation of metabolic reactions, where nodes correspond to compounds and edges correspond to reactions, which involve these compounds as either substrate or product. This approach does not include information on the availability and concentrations of appropriate substrates and enzymes, temperature and pH, and thus cannot consider the actual feasibility of reactions that make up the graph representation. Two main concerns are that the graph-based searches return paths that contain thermodynamically infeasible reactions or use substrates whose availability is not in line with the logic of the path search. Tinker addresses the former issue by using the annotated direction information from the Rhea database in combination with the inclusion of thermodynamic feasibility in the search, an approach that has only been used so far in the context of designing specific pathways (13). Tinker uses thermodynamics information both to direct its heuristic search and to rank resulting paths from all algorithmic searches (see 'Materials and Methods' section).

The latter issue of ensuring substrate availability to be in line with the logic of the path search is addressed in Tinker with a user-settable similarity threshold parameter (see 'Materials and Methods' section). A good example to illustrate the utility of this parameter can be found in a search for two-step paths from '*N*-carbamoyl-L-aspartate' to 'N2-succinyl-L-citrulline'. This search returns a path, where the second reaction combines carbamoyl phosphate with 'N2-succinyl-L-ornithine' to produce

'N2-succinyl-L-citrulline'. While this path is feasible, its implementation is not sensible from a design perspective because 'N2-succinyl-L-ornithine' is structurally highly similar to 'N2-succinyl-L-citrulline' and would not be expected to be freely available. The user-settable similarity threshold parameter allows Tinker to disregard such paths. For a given threshold value, Tinker returns only those paths where any of the substrates (excluding substrates that are produced within the path) are less similar to the target compound than this threshold (see 'Materials and Methods' section).

Pathway prediction tools

As summarized in Table 2, Tinker outperforms existing tools that use the same approach of searching the known reactions for paths among user-defined compounds. Benchmarking Tinker (and similar tools listed in Table 2) directly against tools using reaction rules (such as BNICE), however, is not possible. This is due to the fundamental differences between the two approaches used (see 'Introduction' section) and also due to the fact that currently none of the reaction rule–based tools are available as online search platforms. Despite these limitations, we have attempted a limited comparison between Tinker (and other search-based tools) and BNICE based on the most recent application of the latter (12). Using Tinker (and also Rahnuma, MRSD and FMM), we have searched for paths between pyruvate and 3-Hydroxypropanoate (3HP) as done in (21). Similar to the analysis in (21), we have focused on pathways of six metabolites or shorter. Using a similarity threshold of 0.8 and intermediate compound number cutoff at 4, Tinker found 24 paths corresponding to 5457 unique paths. The top result (in terms of ΔG^0 ranking) was one of the known paths (as shown in Figure 1 of (21)]; pyruvate → acetyl-CoA → malonyl-CoA → 3-oxopropanoate → 3HP ($\Delta G^0 = -186.4$). One of the novel paths reported in Figure 2 of (21) was pyruvate → oxaloacetate → 3-oxopropanoate → 3HP. Paths similar to this one, but using known reactions, were returned by Tinker. These were pyruvate → 2-oxoglutaramate → 2-oxoglutarate → 3-oxopropanoate → 3-hydroxypropionate ($\Delta G^0 = -47.3$) and pyruvate → Glyoxylat → 2-oxoglutarate → 3-oxopropanoate → 3-hydroxypropionate ($\Delta G^0 = -41.2$). Finally, Tinker returned also a 3 metabolite path that used known reactions: pyruvate → 3-oxopropanoate → 3-hydroxypropionate (dG0 = −30.2). This compares with the novel shortest path reported in Figure 2 of (21), which involves a conversion through lactate (by two novel reactions not found in databases).

These examples demonstrate that while BNICE provides a promising avenue for pathway discovery, approaches like Tinker, which are able to exploit the entire URN, are still capable of finding similar and, in some cases, shorter pathways consisting of known reactions (which are potentially more feasible for design purposes). It is important to note that any comparison between Tinker and BNICE results would need to be evaluated carefully, as these two tools aim to achieve similar ends but with completely different approaches. While BNICE can find completely novel paths (unlike Tinker), as suggested above, it cannot (currently) match the coverage available to Tinker. Hence, just considering the number of paths found by the two approaches as a sign of performance would be wrong. In practice, the best approach might be to use the complementary strengths of both tools. Future work towards developing tools that can combine the strength of the two approaches would be highly beneficial and we would aim to undertake such research.

DISCUSSION

In this study, we describe Metabolic Tinker, which is a user-friendly online tool that represents an effort towards developing a 'system design' tool for synthetic (re)design of metabolic pathways. The main novel features of Tinker are its search scope, which is the URN, and its integration of reaction thermodynamics and compound similarity into graph-based search algorithms. These features distinguish Tinker from existing tools, which do not incorporate compound similarity and thermodynamics and either specialize in searching natural paths defined in a specific species context [such as those presented in (5,7)] or attempt to search the URN by using only standard search algorithms [e.g. (11)].

Despite its novel features, Tinker does not provide a definite answer to all the challenges associated with *de novo* design of synthetic pathways. For example, the heuristic search in Tinker does not consider different reaction conditions such as pH and temperature when estimating thermodynamics and does not take into account substrate availability or toxicity. Rather, Tinker focuses on the search and prediction of compound-reaction pathways without reference to any specific organism or subcellular structure. The core aim is to provide a generalized simple tool to search the URN—encompassing all known reactions—for potentially useful pathways among two given compounds. The resulting pathways should be taken through more in-depth study and analysis before any experimental implementation might be attempted. Such implementations might involve creating the found reaction pathways *in vitro* or *in vivo* (i.e. by bringing together the indicated enzymes in a cell or in a test tube containing cell-free extract). In the latter case, subcellular structures might well provide issues hampering experimental implementation, but these would be part of a larger set of challenges inherent in any experimental design.

Overcoming these limitations would be difficult in a generally applicable tool like Tinker, but could be achieved within user-specified design problems. Metabolic Tinker is available both as an online resource and as open-source code, which will facilitate its utility and further development in such question-specific manner. A more interesting area of future development is the extension of heuristic search algorithms towards discovery of completely novel paths that are based on known enzymatic capabilities but not necessarily limited by the current known set of reactions. In particular, we can use

tools like Tinker to explore the effects of adding novel enzymatic reactions into the URN and assess its effect on the design space for a given chemical conversion problem.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1.

ACKNOWLEDGEMENTS

We are grateful to Cleo Kontoravdi, Ali Tavassoli, Aline Miller, Huabing Yin and Wei Huang for useful comments and discussions.

FUNDING

Project grant from Engineering, Physical Sciences Research Council (EPSRC). Funding for open access charge: EPSRC [EP/H04986X/1].

Conflict of interest statement. None declared.

REFERENCES

1. Ajikumar,P.K., Xiao,W.H., Tyo,K.E., Wang,Y., Simeon,F., Leonard,E., Mucha,O., Phon,T.H., Pfeifer,B. and Stephanopoulos,G. (2010) Isoprenoid pathway optimization for taxol precursor overproduction in *Escherichia coli*. *Science*, **330**, 70–74.
2. Ro,D.K., Paradise,E.M., Ouellet,M., Fisher,K.J., Newman,K.L., Ndungu,J.M., Ho,K.A., Eachus,R.A., Ham,T.S., Kirby,J. *et al.* (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, **440**, 940–943.
3. Schirmer,A., Rude,M.A., Li,S., Popova,E. and del Cardayre,S.B. (2010) Microbial biosynthesis of alkanes. *Science*, **329**, 559–562.
4. Andrianantoandro,E., Subhayu,B., Karig,D.K. and Weiss,R. (2006) Synthetic biology: new engineering rules for an emerging discipline. *Mol. Syst. Biol.*, **2**, 2006.0028.
5. Kanehisa,M. and Goto,S. (2000) Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
6. Feist,A.M., Thiele,I., HerrgrdM,J., Reed,R.L. and Palsson,B. (2009) Reconstruction of biochemical networks in microorganisms. *Microbiology*, **7**, 129–134.
7. Mithani,A., Preston,G.M. and Hein,J. (2009) Rahnuma: hypergraph-based tool for metabolic pathway prediction and network comparison. *Bioinformatics*, **25**, 1831–1832, doi:10.1093/bioinformatics/btp269.
8. Alcántara,R., Axelsen,A.B., Morgat,A., Belda,E., Coudert,E., Bridge,A., Cao,H., de Matos,P., Ennis,M., Turner,S. *et al.* (2012) Rhea–a manually curated resource of biochemical reactions. *Nucleic Acids Res.*, **40**, D754–D760.
9. Degtyarenko,K., de Matos,P., Ennis,M., Hastings,J., Zbinden,M., McNaught,A., Alcántara,R., Darsow,M., Guedj,M. and Ashburner,M. (2008) Chebi: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350.
10. Chou,C.H., Chang,W.C., Chiu,C.M., Huang,C.C. and Huang,H.D. (2009) Fmm: a web server for metabolic pathway reconstruction and comparative analysis. *Nucleic Acids Res.*, **37**, W129–W134.
11. Xia,D., Zheng,H., Liu,Z., Li,G., Li,J., Hong,J. and Zhao,K. (2011) Mrsd: a web server for metabolic route search and design. *Bioinformatics*, **27**, 1581–1582.
12. Finley,S.D., Broadbelt,L.J. and Hatzimanikatis,V. (2009) Computational framework for predictive biodegradation. *Biotechnol. Bioeng.*, **104**, 1086–1097.
13. Hatzimanikatis,V., Li,C., Ionita,J.A., Henry,C.S., Jankowski,M.D. and Broadbelt,L.J. (2005) Exploring the diversity of complex metabolic networks. *Bioinformatics*, **21**, 1603–1609.
14. Li,C., Henry,C.S., Jankowski,M.D., Ionita,J.A., Hatzimanikatis,V. and Broadbelt,L.J. (2004) Computational discovery of biochemical routes to specialty chemicals. *Chem. Eng. Sci.*, **59**, 5051–5060.
15. Wu,D., Wang,Q., Assary,R.S., Broadbelt,L.J. and Krilov,G. A computational approach to design and evaluate enzymatic reaction pathways: application to 1-butanol production from pyruvate. *J. Chem. Inf. Model.*, **51**, 1634–1647.
16. Oh,M., Yamada,T., Hattori,M., Goto,S. and Kanehisa,M. (2007) Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation path-ways. *J. Chem. Inf. Model.*, **47**, 1702–1712.
17. Klopman,G., Dimayuga,M. and Talafous,J. (1994) Meta. 1. a program for the evaluation of metabolic transformation of chemicals. *J. Chem. Inf. Comput. Sci.*, **34**, 1320–1325.
18. Hou,B.K., Ellis,L.B. and Wackett,L.P. (2004) Encoding microbial metabolic logic: Predicting biodegradation. *J. Ind. Microbiol. Biotechnol.*, **31**, 261–272.
19. Henry,C.S., Broadbelt,L.J. and Hatzimanikatis,V. (2010) Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate. *Biotechnol. Bioeng.*, **106**, 462–473.
20. Jeong,H., Tombor,B., Albert,R., Oltvai,Z.N. and Barabási,A.L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.
21. Henry,C.S., Broadbelt,L.J. and Hatzimanikatis,V. (2007) Thermodynamics-based metabolic flux analysis. *Biophys. J.*, **92**, 1792–1805.
22. Jankowski,M.D., Henry,C.S., Broadbelt,L.J. and Hatzimanikatis,V. (2008) Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys. J.*, **95**, 1487–1499.
23. Hattori,M., Okuno,Y., Goto,S. and Kanehisa,M. (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.*, **125**, 11853–11865.
24. Cho,A., Yun,H., Park,J., Lee,S. and Park,S. (2010) Prediction of novel synthetic pathways for the production of desired chemicals. *BMC Syst. Biol.*, **4**, 35.
25. Bastian,M., Heymann,S. and Jacomy,M. (2009) Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media.*
26. Voet,D. and Voet,J.G. (2010) *Biochemistry*, 4th edn. Wiley, ISBN-10: 0470570954.
27. Mao,Y., Varoglu,M. and Sherman,D.H. (1999) Molecular characterization and analysis of the biosynthetic gene cluster for the antitumor antibiotic mitomycin c from streptomyces lavendulae nrrl 2564. *Chem. Biol.*, **6**, 251–263.
28. Smolke,C.D., Martin,V.J. and Keasling,J.D. (2001) Controlling the metabolic flux through the carotenoid pathway using directed mrna processing and stabilization. *Metab. Eng.*, **3**, 313–321.
29. Callura,J.M., Cantor,C.R. and Collins,J.J. (2012) Genetic switchboard for synthetic biology applications. *Proc. Natl Acad. Sci. USA*, **109**, 5850–5855.
30. Kasai,D., Masai,E., Miyauchi,K., Katayama,Y. and Fukuda,M. (2005) Characterization of the gallate dioxygenase gene: three distinct ring cleavage dioxygenases are involved in syringate degradation by sphingomonas paucimobilis syk-6. *J. Bacteriol*, **187**, 5067–5074.