JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

CLCA: Maximum Common Molecular Substructure Queries within the MetRxn Database

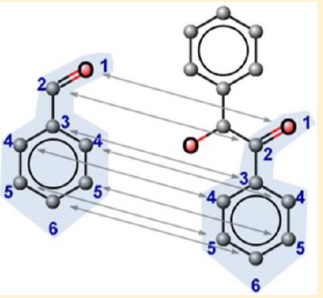
Akhil Kumar[†] and Costas D. Maranas*[,‡]

[†]The Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, Pennsylvania 16802, United States

[‡]Department of Chemical Engineering, Pennsylvania State University, University Park, Pennsylvania 16802, United States

Supporting Information

ABSTRACT: The challenge of automatically identifying the preserved molecular moieties in a chemical reaction is referred to as the atom mapping problem. Reaction atom maps provide the ability to locate the fate of individual atoms across an entire metabolic network. Atom maps are used to track atoms in isotope labeling experiments for metabolic flux elucidation, trace novel biosynthetic routes to a target compound, and contrast entire pathways for structural homology. However, rapid computation of the reaction atom mappings remains elusive despite significant research. We present a novel substructure search algorithm, canonical labeling for clique approximation (CLCA), with polynomial run-time complexity to quickly generate atom maps for all the reactions present in MetRxn. CLCA uses number theory (i.e., prime factorization) to generate canonical labels or unique IDs and identify a bijection between the vertices (atoms) of two distinct molecular graphs. CLCA utilizes molecular graphs generated by combining atomistic information on reactions and metabolites from 112 metabolic models and 8 metabolic databases. CLCA offers improvements in run time, accuracy, and memory utilization over existing heuristic and combinatorial maximum common substructure (MCS) search algorithms. We provide detailed examples on the various advantages as well as failure modes of CLCA over existing algorithms.



1. INTRODUCTION

MetRxn[1] is a standardized nonredundant searchable collection of published metabolic models and databases from a wide variety of organisms. MetRxn[1] primarily focuses on organizing metabolic information such as metabolites, reactions, and pathways using atomistic details.[2,3] Atomistic details enable standardization algorithms to scrutinize and remove duplications and inaccuracies from the metabolic information producing a curated data set. The curated data set is leveraged to automatically annotate and add missing information important for various metabolic modeling projects. Our annotations involve generating atom/bond transition maps, EC number classifications for all reactions, and semantic ontologies for all metabolites. The aforementioned annotations are generated using our novel polynomial time maximum common substructure search algorithm, which we we refer to as canonical labeling for clique approximation (CLCA).

The reaction atom mapping problem involves the task of matching atoms and bonds between reactants and products. This goal is computationally abstracted by identifying the mapping that conserves the maximum common substructure (MCS) between reactants and products.[4] By denoting atoms as vertices and bonds as edges, this task can be posed as a graph matching problem. There exist significant prior efforts devoted to the computational identification of biochemically correct mappings between reactants and products through arbitrary reactions. Jochum and Gasteiger[5] introduced the heuristic known as the principle of minimum chemical distance (PMCD) to computationally solve reaction atom mapping problems as graph matching problems. PMCD is based on the assumption that in most cases reactants (R) are converted into products (P) with the minimal number of bond additions and deletions. MCS search algorithms can thus be used to match and infer a one-to-one mapping between the vertices of R and P. PMCD forms the basis for the methodology described here.

Maximum common substructure (MCS) searches on graphs of general class[6] are computationally challenging,[7] being classified under the larger class of maximum subgraph isomorphism problems with nondeterministic polynomial time complexity (NP-hard). A large number of efforts have addressed this challenge with varying degrees of success and scope.[4,8−16] These algorithms can identify a maximum substructure between most molecular graphs but they have distinctive failure modes. For example, the algorithm by Lynch and Willet,[13] an adaptation of the extended connectivity (EC) algorithm[17] fails to identify a matching when the compared molecular graphs are too small[4] (<7 atoms). It also fails to identify the bonds changes in isomerization reactions involving functional group shifts.[4] The procedures by Vleduts,[18] McGregor and Willett,[12] and Funatsu et al.[19] also make use of the EC algorithm to identify a MCS. All these methods have similar failure modes as the error from an incorrect starting MCS solution provided by the EC algorithm propagates into the final mapping solution.[4] Nevertheless, due to computational tractability imperatives, commercial reaction atom mapping applications largely employ the EC algorithm to find MCS solutions.[4] Non-EC based methods traditionally use

Received: July 3, 2014

   dx.doi.org/10.1021/ci5003922 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

combinatorial search strategies such as branch and bound to identify a MCS. Combinatorial searches require $m!n!/k!(m-k)!(n-k)!$ comparisons to identify a substructure containing k atoms that is common to two structures containing m and n atoms, respectively.[12] Various algorithms attempt to speed up the MCS detection by using heuristics to reduce the combinatorial search space. For example, in the algorithm by Barker et al.,[20] functional groups are reduced to aggregate vertices with the connectivity between the aggregate vertices reflecting the connectivity of atoms in the original chemical graph. Such a representation greatly reduces the search space for the branch and bound algorithm but might fail to detect a MCS between a cyclic and a linear structure.[20,21] The branch and bound algorithm presented by Caboche et al.[22] can detect a MCS between a cyclic and linear graph; however, the algorithm requires converting the original graphs into a data structure referred to as a compatibility graph (CG). In CG, vertices represent a possible matching between a pair of atoms, and edges represent the bonds in the two structures. The largest clique in a CG represents a MCS solution between the original graphs. A branch and bound algorithm is then used to exhaustively search the CG for the largest clique. However, the size of a CG scales exponentially to the number of vertices and edges in the original graph.[21] Therefore, for larger graphs, such a representation is extremely dense, and MCS searches do not scale.[21] Similar to branch and bound-based algorithms for MCS searches between pairs of molecular graphs, a number of algorithms for MCS detection between multiple molecules in reaction graphs exist. The MCS of a reaction graph is used to detect a one-to-one mapping between the vertices/atoms and bonds/edges[12] of the reactant (R) and product (P) graph. First et al.[15] recently described an efficient mixed integer linear optimization based technique to map reactions as well as identify multiple reaction mechanisms by minimizing the number of bond or edge changes between R and P. The MWED algorithm by Latendresse et al.[9] improves upon the mathematical formulation by First et al.[15] by incorporating chemical knowledge as edge weights or bond costs based on atom species. In addition, to speed up the matching process, ring detection and matching is done prior to the branch and bound search. However, such mixed integer linear optimization techniques cannot identify reaction maps for unbalanced reactions.[9,15]

Unlike the previously mentioned algorithms, CLCA uses a local search strategy to canonically label (uniquely order) vertices/atoms and detect a MCS between molecular graphs. CLCA uses the property of prime factorization to uniquely label vertices and is an adaptation of the labeling technique used in the SMILES algorithm by Weinengier et al.[2] Compared to the labels generated by the EC algorithm, the labels generated by CLCA uniquely represent a vertex in the entire topological space and not just within a fixed neighborhood. Vertex labels between each reactant and product graph is compared, and unmatchable vertices are iteratively removed from the MCS search calculation. For reactions with multiple reactant or product molecular graphs, many combinations of MCS matches exist. We therefore follow the heuristic known as the principle of minimum chemical distance[5] (PMCD) to choose a matching that reduces the overall number of bond changes between reactants and products. When compared to MetaCyc,[23] KEGG,[24] and ReactionMap,[25] CLCA has 97.8%, 99.3%, and 97.9% agreement, respectively, with their reported atom mapping solutions. We have also manually verified the atom mapping results for 1293 reactions from the *E. coli* *i*AF1260[26] for accuracy and found an error rate of only 0.5%. In addition, we tested and accurately mapped over 95.2% of the reactions in the ICMap[27] test suite. Failure modes of CLCA, MetaCyc, MWED,[9] and ReactionMap[25] for different classes of reactions[4,27] are analyzed and discussed in detail. The atom mapping solution for approximately 27,000 reactions in the MetRxn database is available online at http://metrxn.che.psu.edu/

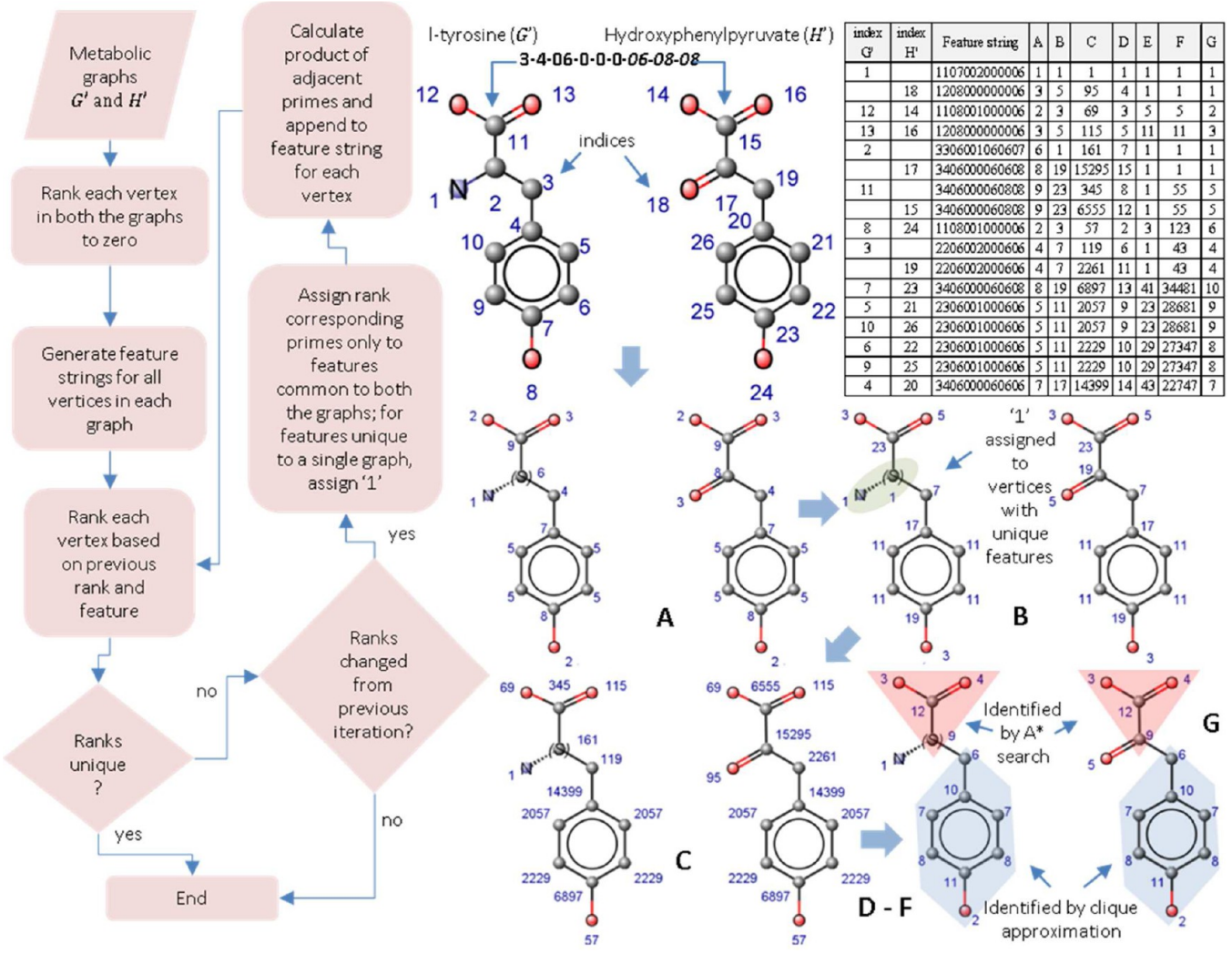
2. METHODS

A molecular graph represents a chemical compound wherein the vertices denote the atoms, and the edges represent the bonds. MetRxn[1] stores information about the atoms, bonds, and connectivity as SMILES.[3] This information is leveraged to construct molecular graphs for the algorithm we present in this article. Every graph G is an ordered pair of $\{V,E\}$, where $v \in V$ is a set of vertices or nodes and $e \in E$ is the set of edges or connections. The induced graph of G, represented by G', is homomorphic to G (i.e., has the same topology of G). The induced graph G' by definition preserves all the structural properties of G and contains the same number of vertices and edges. Let, for two graphs G' and H', the graph $g = \{v',e'\}$ be the (sub)graph that is isomorphic, i.e., after a finite set of edge and vertex deletions on G' and H', and both G' and H' would contain subgraphs that are isomorphic to g. A clique is a fully connected graph that is represented as $q = \{v'',e''\}$ where v'' represents a mapping between the elements of G' and H' and e'' represents a mapping between the edges of G' and H'. CLCA identifies the largest g between G' and H' by iteratively identifying the mappings represented by cliques q. Vertices in G' and H' are compared for common labels, and unmatchable ones are successively deleted. Multiple cliques q, identified by common labels, are then combined into the graph g, which is (sub)graph isomorphic to both G and H.

In order to identify vertex labels, a set of reordering operations are performed on the vertices of G' and H'. The labels generated after reordering the vertices of G' and H' are representative of the vertex locations in the entire topological space (i.e., vertices in the center of the graph *generally* receive larger integer labels than the vertices in the periphery). Such a reordering also follows the popular graph conjecture[28] that states that if two induced graphs are identical when reordered, then the parent graphs are identical. Therefore, to identify a mapping between G and H, we identify a mapping between the reordered graphs G' and H'. The reordering, labeling, and mapping of two induced molecular graphs G' and H' by CLCA is as follows: (i) Identify features for each vertex/atom in G' and H' and initialize label on each vertex to "0". (ii) Rank order with prime numbers only for vertices with features common to at least two compared molecules. Vertices with unique features are assigned the value "1" (i.e., deletion from MCS search space) (iii) (Re)assign labels and features based on product with neighboring primes. (iv) Iterate through steps (ii) and (iii) until atom labels do not change anymore. (v) Identify all cliques q by common labels. (vi) Expand the isomorphic subgraph g by appending cliques q. (vii) Identify and keep the largest connected subgraph.[21] (viii) Extend the largest connected subgraph to maximum common subgraph (MCS) using the A* search algorithm.[11] The subgraph identified after the A* search procedure g is (sub)graph isomorphic to G' as well as G.

Before we iterate, we extract distinguishing features of each atom and convert it into a string representation. The first seven features we extract are as follows: (i) number of non-hydrogen connections, (ii) number of non-hydrogen bonds, (iii) atomic numbers, (iv) sign of charge, (v) absolute charge, (vi) number of connected hydrogens, and (vii) atomic numbers of neighboring

dx.doi.org/10.1021/ci5003922 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX



index G′	index H′	Feature string	A	B	C	D	E	F	G
1		1107002000006	1	1	1	1	1	1	1
	18	1208000000006	3	5	95	4	1	1	1
12	14	1108001000006	2	3	69	3	5	5	2
13	16	1208000000006	3	5	115	5	11	11	3
2		3306001060607	6	1	161	7	1	1	1
	17	3406000060608	8	19	15295	15	1	1	1
11		3406000060808	9	23	345	8	1	55	5
	15	3406000060808	9	23	6555	12	1	55	5
8	24	1108001000006	2	3	57	2	3	123	6
3		2206002000606	4	7	119	6	1	43	4
	19	2206002000606	4	7	2261	11	1	43	4
7	23	3406000060608	8	19	6897	13	41	34481	10
5	21	2306001000606	5	11	2057	9	23	28681	9
10	26	2306001000606	5	11	2057	9	23	28681	9
6	22	2306001000606	5	11	2229	10	29	27347	8
9	25	2306001000606	5	11	2229	10	29	27347	8
4	20	3406000060606	7	17	14399	14	43	22747	7

Figure 1. CLCA workflow. The numbers shown in the table in columns A, D, and G represent the labels generated by CLCA. Atom features such as the number of non-hydrogen connections, number of non-hydrogen bonds, atomic numbers, sign of charge, absolute charge, number of connected hydrogens, and atomic numbers of neighboring atoms are integer encoded and concatenated into strings in the "Feature string" column. The columns G' and H' represent indices for lookup. Panel B shows the prime number assignment after ordering step A. Panel C shows the product with adjacent prime for each vertex. After the termination of the canonical labeling step (i.e., steps A through G), we extend the subgraph size using the A* search methodology. A* traversal and subgraph extension always starts from the largest fragment (i.e., subgraph highlighted in blue) toward the unmapped vertices in the graph (i.e., subgraph in red). The two nonequivalent atoms, identified after the termination of A* search, are stamped with different numbers (i.e., 1 and 5).

atoms. These features are sufficient for mapping molecules without chiral centers. All the features are concatenated as integers and then rank ordered using the natural ordering for strings. For example, in Figure 1, we show the feature string generation and ranking of each atom in L-tyrosine and hydroxyphenylpyruvate for identifying each atom uniquely. After we identify features on each atom, we iterate to rank order (Figure 1A), assign rank corresponding primes (Figure 1B), and calculate the product of adjacent primes (Figure 1C). If at any iteration the value for the product with adjacent primes is unique only to a single vertex in G' or H', we assign a value of "1" to it (Figure 1B). Assigning a value of "1" to a vertex can be considered as removing it from the MCS search. Prime numbers are assigned only to vertices with features and product values common to both G' and H'. The procedure is repeated until each atom is uniquely ranked or until the rank ordering does not change anymore (Figure 1G). The reason for using products of adjacent primes is to assign a unique rank order[6] to atoms by the properties of its adjacent neighbors. At each iteration for a given atom, the radii of influence on its rank by adjacent atoms increases. As illustrated in the inset table in Figure 1, the rank at step A for atom with index = 5 is only affected by the ranks of its surrounding atoms 4 and 6, but the rank at step D is affected by the ranks of atoms 3, 10, and 7. The procedure terminates when in two consecutive iterations the same rank ordering remains for all vertices. CLCA iteratively removes vertices with distinct topological features and labels only those vertices with common topological features. The canonical labels are then used to recognize the mappings between the vertices of G' and H'. We choose only the mappings for vertices in the largest connected subgraph and discard solutions to all other vertices. The mappings retained are stored as the graph g, which is (sub)graph isomorphic to both G' and H'. The mappings identifiable by g are then used as the starting solution for the combinatorial A* search.[11] A* search is a widely used graph traversal algorithm to efficiently traverse a weighted path between vertices. A* uses best first search to find a path with the least cost

dx.doi.org/10.1021/ci5003922 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

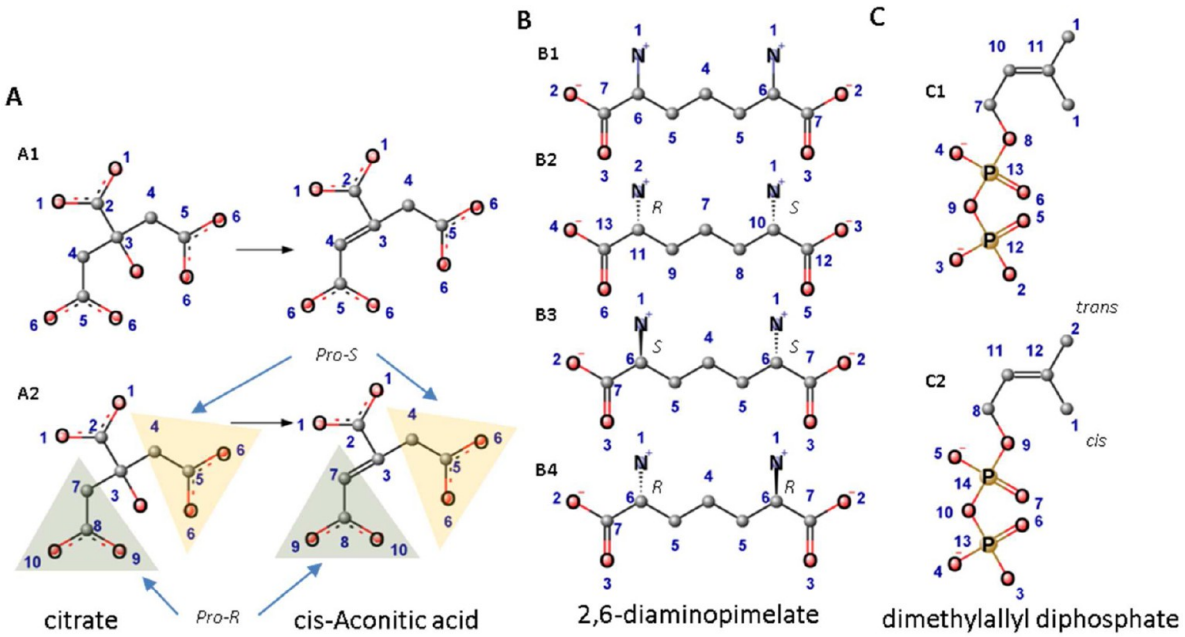


Figure 2. Canonical labeling with stereodescriptors. Labeling with the three additional stereodescriptors are applied on three molecules, citrate, 2,6-diaminopimelate, and dimethylallyl diphosphate, identified by graphs A, B, and C, respectively. Graphs A1, B1, and C1 identify chemically equivalent atoms when stereodescriptors are ignored, while A2, B2, and C2 differentiate stereochemically distinct atoms. Stereoisomers L,L-2,6-diaminopimelate and (2*R*,6*R*)-2,6-diammonioheptanedioate represented by graphs B3 and B4 have chemically and stereochemically equivalent atoms due to the presence of a rotation axis.

from source to goal vertex. In graphs G' and H', the source vertices are the mapped atoms/vertices, and the goal vertices are the unmapped atoms/vertices. The cost on each atom/vertex is the corresponding atomic number. Two goal vertices are mapped only if the least cost paths as well as the mapping for their initial vertices are equal in both G' and H'. As shown in Figure 1G, the source vertices in G' and H' contain the labels 2, 6, 7, 8, 10, and 11, respectively. The goal vertices mapped by the A* search are identified by the labels 3, 4, 9, and 12. Like many combinatorial search algorithms, to identify a path containing k atoms present in two structures with m and n atoms requires $m!n!/k!(m-k)!(n-k)!$ comparisons.[12] This generally enormous number of combinatorial comparisons needed prior by the A* is therefore avoided by identifying the source and goal vertices using vertex labels. We also note that CLCA always generates equivalent labels for homotopic groups. For example, in Figure 1G, vertices with label 7 represent carbons (indices = 5, 10, 21, and 26) that are symmetric due to an inversion center.

The mappings identified with only the seven aforementioned atomic properties can identify symmetry in a molecular graph if we ignore stereochemical distinctions. This is inadequate because stereospecificity is paramount in enzymatic reactions.[29] For example, there is a large number of NADH-dependent dehydrogenases that transfer a hydride ion either from/to the *re*-side or the *si*-side of the nicotinamide group in NADH/NAD+ implying that we need to differentiate between the *pro*-R and *pro*-S hydride ions.[30] Enzymes isocitrate dehydrogenase, malate dehydrogenase, lactate dehydrogenase, and alcohol dehydrogenase show stereospecificity to the *pro*-R hydride ion, while α-ketoglutarate dehydrogenase, glucose-6-phosphate dehydrogenase, glutamate dehydrogenase, and glyceraldehyde 3-phosphate dehydrogenase show stereospecificity to the *pro*-S side.[30] We therefore add additional characterizations related to stereochemistry in the feature string so as to stereochemically discern between symmetric atoms. Therefore, we characterize atoms based on an additional three stereodescriptors:[31] (viii) R or S descriptor for chiral atoms, (ix) pro-R or pro-S for prochiral arms, and (x) cis and trans descriptors.[31] For stereochemically characterizing all carbons, the characterization procedure traverses the molecular graph to identify all chiral atoms, prochiral atoms, and atoms with double bonds. To identify the prochiral arms/atoms, we calculate the dihedral angle between the heterotopic groups and the plane defined by the homotopic groups. For example, in the prochiral molecule citrate, for the plane defined by the prochiral carbon and the two carboxymethyl groups, we compare the angle of the hydroxyl group and the plane to the angle between the carboxyl group and the plane. For visual clarity, we only use hydrogen-suppressed molecular graphs in all the figures to explain CLCA. Figure 2A shows an example for mappings with and without stereodescriptors for the conversion from citrate to cis-aconitic acid, catalyzed by the enzyme aconitase. Ignoring stereodescriptors, CLCA identifies chemically equivalent, though biochemically distinct, symmetric groups. Citrate has a plane of symmetry and a prochiral carbon center. Aconitase differentiates between the pro-R and pro-S arms of citrate thus distinguishing between the carboxymethyl group and proton of the pro-R arm from the carboxymethyl group and proton of the pro-S arm and when forming the cis-aconitate intermediate.[32] Upon appending the stereodescriptors for the two prochiral substituents, we correctly distinguish between the pro-R and pro-S arms of citrate. Other metabolic graphs shown in Figure 2 have distinct stereo isomers. 2,6-Diammonioheptanedioate in Figure 2B has two chiral carbons and three distinct stereoisomers (i.e., meso-2,6-diaminopimelate, L,D-2,6-diaminopimelate, and (2*R*,6*R*)-2,6-diammonioheptanedioate). meso-2,6-Diaminopimelate has an inversion center but lacks a rotation axis, thereby rendering the two symmetric groups stereochemically distinct. L,L-2,6-Diaminopimelate and (2*R*,6*R*)-2,6-diammonioheptanedioate contain an axis of rotation, and therefore, the two symmetric groups are also biochemically equivalent. In *E. coli*, the

dx.doi.org/10.1021/ci5003922 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

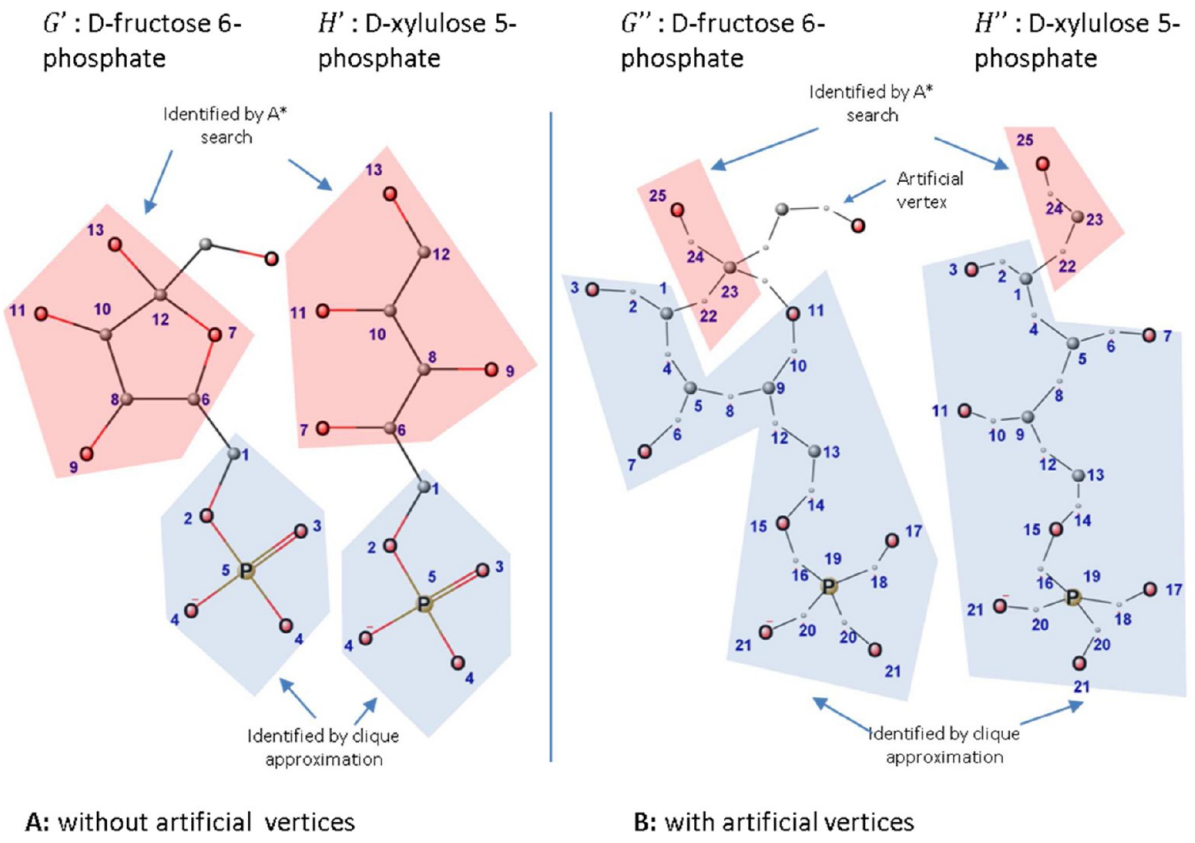


Figure 3. Addition of artificial vertices. Molecular graphs of D-fructose 6-phosphate (G') and D-xylulose 5-phosphate (H') are converted into their auxiliary graphs G'' and H'', respectively. The smaller vertices are the artificial vertices. The auxiliary graphs are created by replacing each bond with an artificial vertex. The region in blue is the mapping identified by CLCA. The region in red is the mapping identified using A* search. Notice the increase in the mapping regions between panels A and B, identified by CLCA after the introduction of artificial vertex. An artificial vertex allows for the removal of a bond in place of an atom from the MCS search (i.e., assignment of "1" instead of a prime number), thereby allowing CLCA to label a larger subgraph.

gene *lysA* encodes the enzyme diaminopimelate decarboxylase,[33] which stereospecifically catalyzes the decarboxylation of meso-2,6-diaminopimelate to L-lysine and is inactive for the RR- or SS-isomers of diaminopimelate. Similarly, as shown in Figure 2C, CLCA distinguishes between the symmetric groups that lie on the reside or si-side for a sp2 hybridized atom. Isopentenyl pyrophosphate isomerase stereospecifically catalyzes the reversible isomerization reaction between isopentenyl pyrophosphate and dimethylallyl pyrophosphate. As part of the standardization procedure, we characterize all atoms in each molecule in the MetRxn[1] database with chiral, prochiral, and cis−trans flags. This allows us to identify and present to the user biochemically relevant symmetry information.

Reduction of A* Search Space. As a test case, we compared 2000 pairs of molecular graphs for (sub)graph isomorphism using CLCA, and we noticed that the algorithm still took considerable time to identify accurate solutions. For each comparison, the procedure took an average time of 36 s to produce a solution. The canonical labeling step identifies a subgraph close to 61%, the size of the maximum common subgraph, and the A* search step identifies the remaining 39%. Over 99% of the total run time per comparison was used up by the A* extension step alone. For the examples shown in Figures 1 and 2, the run time was close to 1 s; however, for comparison with larger molecular graphs (i.e., atom count >100) and polycyclic molecules, the run time increased substantially. We reduced the reliance on the A* search by augmenting the input molecular graphs G' and H' with additional information. First, we introduce an artificial vertex between adjacent vertices in the induced graph G' by converting each bond into an artificial vertex. The new graph with the additional set of artificial vertices to graph G' is auxiliary graph G''. If there is a bond between v_1 and v_2, the bond is converted into an artificial vertex a_{12}, and new connections are formed between v_1 and a_{12} and v_2 and a_{12}.

In the example shown in Figure 3B, the vertex with index 24 is the artificial node in place of the edge between vertices 23 and 25. Similar to the previously mentioned vertex characterization step for induced graphs G' and H', we characterize each vertex in auxiliary graphs G'' and H'' by the properties of their corresponding atoms. These features are limited to (i) number of non-hydrogen connections, (ii) number of non-hydrogen bonds, (iii) atomic numbers, (iv) sign of charge, (v) absolute charge, (vi) number of connected hydrogens, and stereo descriptors such as (vii) R or S descriptor for chiral atoms, (viii) pro-R or pro-S for prochiral arms, and (ix) cis and trans.[31] Because the artificial vertices represent the bonds in the parent molecular graph, they are characterized by the bond order. The column with header FHV in Table 1 shows the new features for each vertex in G'' and H''. The iterative labeling routine is then run on the auxiliary graphs. Figure 3B shows the improvement in mapping due to the introduction of artificial vertices identified by the labeling step alone. The introduction of the artificial vertex mitigates the removal of atoms and allows for the removal of bonds, thereby allowing the labeling of a larger subgraph without invoking the costly A* procedure as shown in Figure 3B for G'' and H''.

In addition to the aforementioned features for characterizing each atom vertex, we use a connectivity criterion to improve the

dx.doi.org/10.1021/ci5003922 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Characterization and Prime Number Assignment for Figure 5[a]

INDEX	WEIGHT	CHARGE	GEODESIC	FHV	A	A′	B	B′	C	C′	D	D′	E	E′	F	F′	g
14	006	0	6*6*6*6	1000519069986	17	218569	43		43		43		43		43		43
36	006	0	6*6*6*6	1000519069986	17	218569	43		43		43		43		43		43
10	006	0	6*6*6*6	1000519069986	17	386699	47		47		47		47		47		47
44	006	0	6*6*6*6	1000519069986	17	386699	47		47		47		47		47		47
25	015	0	15	1001423846568	2		2		2		2		2		2		2
52	015	0	15	1001423846568	2		2		2		2		2		2		2
27	008	0	8*15	1001479949820	19	703	53	3127	71	5183	73	5767	83	7387	89	8633	89
29	008	0	8*15	1001479949820	19	703	53	3127	71	5183	73	5767	83	7387	89	8633	89
56	008	0	8*15	1001479949820	19	703	53	3127	71	5183	73	5767	83	7387	89	8633	89
58	008	0	8*15	1001479949820	19	703	53	3127	71	5183	73	5767	83	7387	89	8633	89
9	001	0	*6*6*6*6	1001782211495	23	16031	41		41		41		41		41		41
45	001	0	*6*6*6*6	1001782211495	23	16031	41		41		41		41		41		41
18	001	0	*6*6*6*6	1001782211495	23	11339	67	223597	67		67		67		67		67
47	001	0	*6*6*6*6	1001782211495	23	11339	67	223597	67		67		67		67		67
17	001	0	*6*6*6*6	1001782211495	23	11339	67	2881	1	43	1	43	79		79		79
35	001	0	*6*6*6*6	1001782211495	23	11339	67	204551	1	3397	1	43	79		79		79
19	006	0	6*6*6*6*6	1002042243476	29	20677	71	347261	79	322873	71		71		71		71
48	006	0	6*6*6*6*6	1002042243476	29	20677	71	347261	79	322873	71		71		71		71
5	006	0	6*6*6*6*6	1002042243476	29	640987	1	4891	1	1	1	1	1	5767	83		83
34	006	0	6*6*6*6*6	1002042243476	29	20677	71	347261	79	79	1	1	1	5767	83		83
4	001	0	*6*6*6*6*6	1100111807953	31	899	1	1	1	1	1	1	1	1	1	1	1
59	001	0	*6*6*6*6*6	1100111807953	31	1189	1	1	1	1	1	1	1	1	1	1	1
20	001	0	*6*6*6*6*6	1100111807953	31	42253	73	191771	61		61		61		61		61
49	001	0	*6*6*6*6*6	1100111807953	31	42253	73	191771	61		61		61		61		61
6	001	0	*6*6*6*6*6	1100111807953	31	42253	73	2117	1	29	1	29	73		73		73
33	001	0	*6*6*6*6*6	1100111807953	31	42253	73	150307	1	2291	1	29	73		73		73
31	008	1	8*15	1100327504643	3		3		3		3		3		3		3
54	008	1	8*15	1100327504643	3		3		3		3		3		3		3
12	006	0	6*6*6	1100513988396	5		5		5		5		5		5		5
40	006	0	6*6*6	1100513988396	5		5		5		5		5		5		5
24	001	0	*8*6*6*6*6*6	1100553066779	7		7		7		7		7		7		7
51	001	0	*8*6*6*6*6*6	1100553066779	7		7		7		7		7		7		7
2	001	0	*6*6*6*6*6*6	1100615106393	1	1	1	1	1	1	1	1	1	1	1	1	1
30	001	0	*15	1100740988597	37	222	17		17		17		17		17		17
53	001	0	*15	1100740988597	37	222	17		17		17		17		17		17
26	001	0	*15	1100740988597	37	1406	59	6254	73	10366	79	11534	89	14774	97	17266	97
28	001	0	*15	1100740988597	37	1406	59	6254	73	10366	79	11534	89	14774	97	17266	97
55	001	0	*15	1100740988597	37	1406	59	6254	73	10366	79	11534	89	14774	97	17266	97
57	001	0	*15	1100740988597	37	1406	59	6254	73	10366	79	11534	89	14774	97	17266	97
22	008	0	8*6*6*6	1101319646818	11		11		11		11		11		11		11
42	008	0	8*6*6*6	1101319646818	11		11		11		11		11		11		11
1	008	0	8*6*6*6*6*6*6	1101528804776	1	1	1	1	1	1	1	1	1	1	1	1	1
16	008	0	8*6*6*6*6	1101689473384	41	533	19		19		19		19		19		19
38	008	0	8*6*6*6*6	1101689473384	41	533	19		19		19		19		19		19
8	008	0	8*6*6*6*6	1101689473384	41	943	23		23		23		23		23		23
46	008	0	8*6*6*6*6	1101689473384	41	943	23		23		23		23		23		23
15	001	0	6*6*6*6	1101798782938	13		13		13		13		13		13		13
37	001	0	6*6*6*6	1101798782938	13		13		13		13		13		13		13
21	001	0	*6*6*6	1101857886417	43	2365	31		31		31		31		31		31
41	001	0	*6*6*6	1101857886417	43	2365	31		31		31		31		31		31
13	001	0	*6*6*6	1101857886417	43	3655	61	13115	53		53		53		53		53
39	001	0	*6*6*6	1101857886417	43	3655	61	13115	53		53		53		53		53
11	001	0	*6*6*6	1101857886417	43	3655	61	14335	59		59		59		59		59
43	001	0	*6*6*6	1101857886417	43	3655	61	14335	59		59		59		59		59
7	008	0	8*6*6*6*6*6	1102057185826	47	1457	29		29		29		29		29		29
32	008	0	8*6*6*6*6*6	1102057185826	47	1457	29		29		29		29		29		29
23	008	0	8*6*6*6*6*6	1102057185826	47	10199	37		37		37		37		37		37
50	008	0	8*6*6*6*6*6	1102057185826	47	10199	37		37		37		37		37		37
3	006	0	6*6*6*6*6*6	1102141857694	1	31	1	1	1	1	1	1	1	1	1	1	1

[a]The feature hash vector (FHV) is the hash value of the string that encodes the features and longest graph geodesic for each vertex. Prime numbers in column *A* are assigned based on the FHV, and products *A′* are only calculated for vertices with no mapping. Subsequent prime numbers in columns *B* through *F* are assigned based on the products in columns *A′* through *E′* at each iteration. Iterations when the prime values in subsequent iterations are the same (i.e., *F* and *g*). At each iteration, if a mapping (green) is identified, the prime value assigned to those vertices will remain fixed.

dx.doi.org/10.1021/ci5003922 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

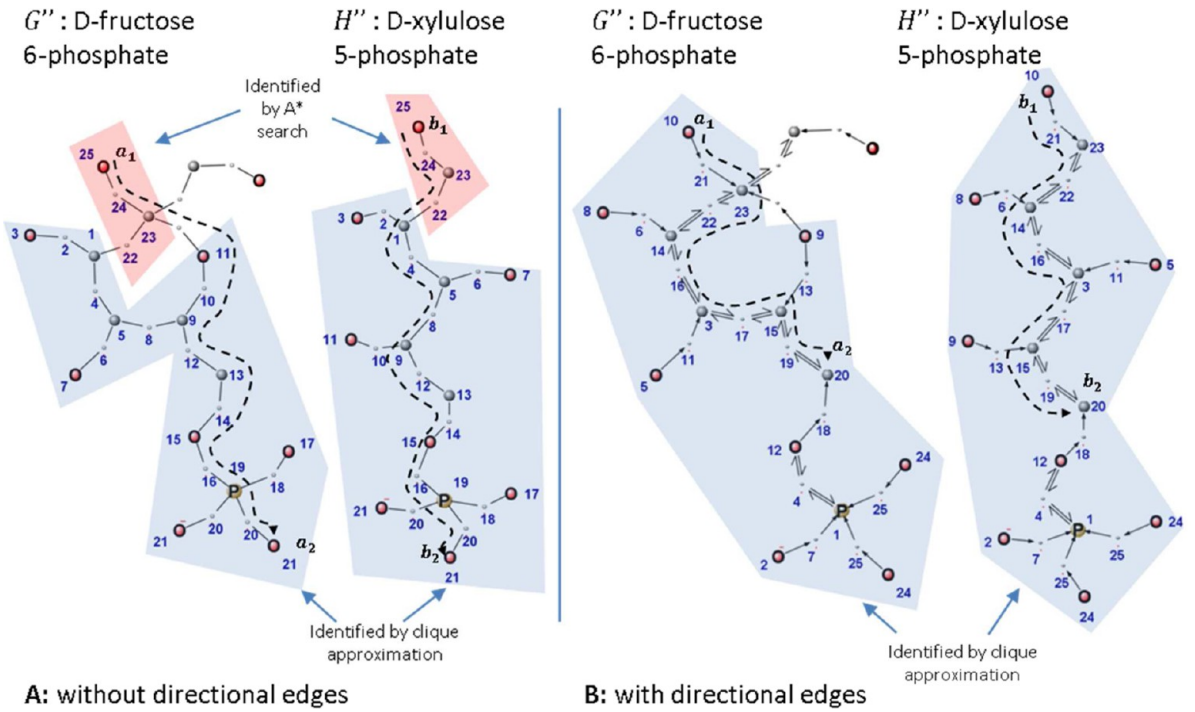


Figure 4. Addition of directional edges. Vertices representing carbon atoms have edges directed toward it, while vertices for noncarbon atoms have edges directed away from them. If two vertices containing incoming or outgoing edges are adjacent, then the edges between them are bidirectional. Artificial vertices are assigned directionality according to adjacent atom representing vertices. The dashed line represents the longest graph geodesic for vertices a_1 and b_1 in both panels A and B. The longest geodesic of a_1 and b_1 terminate at carbon vertices in panel B as noncarbon vertices are unreachable due to edge directions. Notice the increase in the mapping regions between panels A and B, identified by CLCA after the introduction of directional edges. By calculating the longest geodesic feature on a directed graph, a MCS without the A* search step is identified.

overall mapping accuracy of CLCA. We increase both the prediction fidelity[21] and size of the MCS by using path information. A vertex v_g in G'' is mapped to a vertex v_h in H'' if a path v_g to v_g' in G'' is that same as a path from v_h to v_h' in H'' (i.e., a path of same length, type of vertices (atom type), and sequence of vertices). We use the Floyd–Warshall algorithm to identify the shortest paths between all vertex/atom pairs. The set of shortest paths in an undirected graph with n vertices is of size n;[2] therefore, many vertices may share the same set of paths. In order to reduce the commonality in paths, edges are directed away from noncarbon vertices and toward carbon vertices. The set of paths originating from v_g and v_h is now smaller and more distinct. Because we require only one path as a feature for each vertex, we choose the longest of all the (common) paths from the smaller set. This path, called the longest geodesic, reduces the matching ambiguity between multiple atoms and at times also resolves matches between pairs of atoms prior to the *product with adjacent primes* step. This feature therefore also contributes toward reducing the total number of iterations needed to completely identify a complete MCS solution. The longest geodesic is illustrated in Figure 4B, where all the edges in G'' and H'' are directed toward the carbon atoms and away from the noncarbon atoms. The artificial vertex also preserves this directionality in the auxiliary graph. The rationale for the directionality is based on the type of molecules under consideration or the underlying data set we wish to map. We noticed that in MetRxn over 86% of reactions have reaction centers on non C–C bonds. Only 3761 reactions of 27,632 reactions in MetRxn involve breaking or formation of C–C bonds. Therefore, to ensure that a longest geodesic contains the least number of non C–C bonds, we direct the edges away from noncarbon atoms. The number of edges connecting two adjacent nodes will always be two. It is important to note the mapping accuracy involving C–C bonds was not affected by this assumption, as shown by the example in Figure 4B. The relationship between direction and bond type is not fixed and can be modified to best suit the data that has to be mapped. A similar strategy is also followed by Kraut et al.,[27] wherein reaction rules pritoritize breaking of bonds between heteroatoms over carbon–carbon atoms.

Figure 5 and Table 1 provide an example for MCS detection between the molecular graphs of D-fructose 6-phosphate and D-xylulose 5-phosphate. As shown previously, the features on each node are concatenated and integer hash-coded, as shown in column "FHV" in Table 1. The feature hash codes of all vertices in G'' and H'' are combined into a column vector FHV (feature hash code vector) and sorted for natural ordering. Equivalent hash codes in FHV receive equivalent rank order. We create another column vector, which is a prime number vector A of the same size of FHV. Rows of columns A and FHV link to each other by their indices. Each row in column A stores a prime number based on the rank order of a corresponding element in FHV. If the rank of an element is unique in FHV, the corresponding row in A stores "1". Each vertex in G'' and H'' can also be linked to a prime number or "1" in A through FHV. For each vertex, we calculate the product of primes with adjacent vertices. The adjacent product for each vertex for the first iteration is shown in column A' in Table 1. For example, in Figure 5, the vertex with index 10 has adjacent vertices with indices 9, 11, and 18. The prime numbers assigned to the adjacent vertices are 23, 23, and 43, respectively. The product for vertex 10 (prime = 17) is 386,699. The products are appended to the original feature and sorted again to check for uniqueness. If in the current iteration no change in rank ordering of pairs of vertices is noticed, then we identify them as a mapping.

dx.doi.org/10.1021/ci5003922 | J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

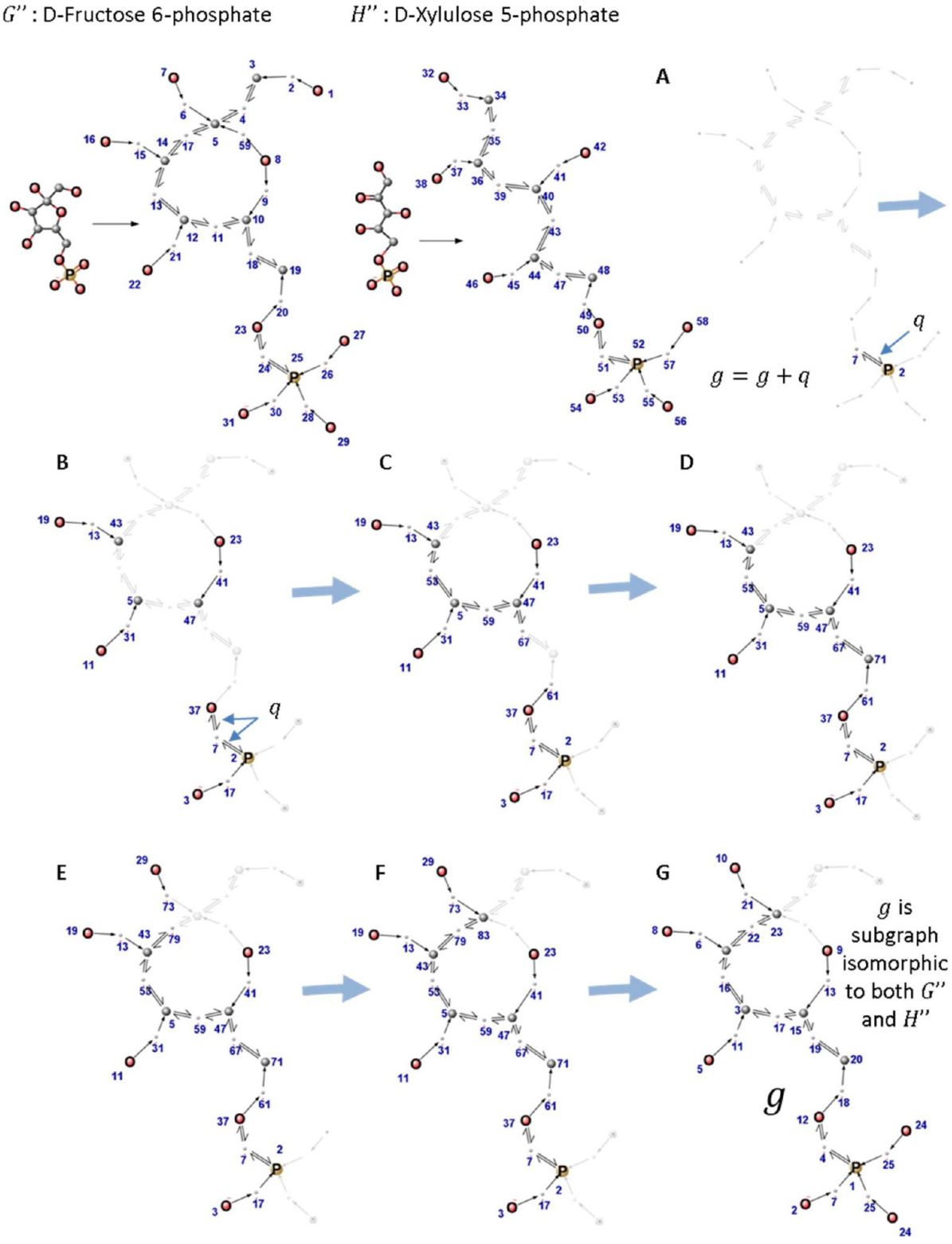


Figure 5. CLCA using the auxiliary graph data structure. Panels A through F show an increase in size of g due to the repeated addition of cliques q. Cliques q are identified if two adjacent labels are common to vertices in both G'' and H''. The final graph g after iterations terminate is a subgraph isomorphic to both G'' and H''.

The clique identified by the mapping between vertices with common labels between G'' and H'' is $q = \{v'',e''\}$, where v'' represents a mapping between G'' and E''. e'' is the edge between two vertices and represents a mapping between the edges of G'' and H''. As illustrated in Figure 5, the common cliques are appended to g at each iteration. For example, in Table 1, iteration A identifies 6 bijections. Only the bijections between the vertex pairs 24, 25 and 51, 52 form a clique (Figure 5A). Figure 5 illustrates the clique recognition and common subgraph expansion at each step. The graph $g = \{v'',e''\}$ obtained upon

dx.doi.org/10.1021/ci5003922 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

termination is subgraph isomorphic to both G'' and H''. We iterate to rank, identify corresponding primes, and the product of adjacent primes until we notice no change in the rank ordering of each vertex. The procedure terminates after the sixth iteration as vectors F and g are identical and improvement in the size of the common subgraph is not possible. For the example shown in Figure 5, we do not require the use of A* search method[11] to maximize the subgraph isomorph. If for any comparison the A* search is indeed needed, we limit the search depth to 3 as now CLCA without A* generates in >99% of the cases an optimal mapping solution. After the introduction of the auxiliary graph data structure, the average time per MCS solution reduced from 36 s to 140 ms.

Applying CLCA on the auxiliary graph provides multiple benefits. In addition to improved computational performance, CLCA can now deal with reactions involving ring formation or breaking. Without the auxiliary graph, CLCA would only identify common substructures close to the reaction center. Using the auxiliary graph data structure, we were able to extend the common subgraph to include reaction centers as well as to increase the size to the maximum subgraph in most cases. This dramatically reduces the reliance on the costly A* search for identifying the maximum subgraph. Strategies other than the A* search can also be utilized to extended the MCS identified by CLCA.[34] CLCA generates homotopic solutions for the atom mapping problem (i.e., same mapping solution to symmetric groups). Figure 5 shows the equivalent oxygen atoms, with mapping = 24, on the phosphate arm of D-fructose 6-phosphate and D-xylulose 5-phosphate. A simple post-processing step to find all possible permutations of mappings produces alternate solutions. All alternate solutions are encoded within a single solution; therefore, the mappings provided by CLCA greatly reduce the computational overhead needed to identify and iterate through the entire set of solutions due to symmetry.

Reaction Atom Mapping. A pairwise comparison for the maximum common substructure (MCS) between each product and reactant auxiliary graph is performed using CLCA for computing the atom mapping solution for every reaction. The subset of vertices and edges of the molecular graphs that undergo transformation are denoted as reaction centers. Figure 6 shows the atom mapping solution for the transketolase transformation between the substrates D-erythrose 4-phosphate and D-xylulose 5-phosphate. Illustrated in Figure 6C, CLCA is first applied to find common subgraphs between the molecular graphs of D-fructose 6-phosphate and D-xylulose 5-phosphate, D-glyceraldehyde 3-phosphate and D-xylulose 5-phosphate, D-fructose 6-phosphate and D-erythrose 4-phosphate, and D-glyceraldehyde 3-phosphate and D-erythrose 4-phosphate. The maximum subgraphs identified between the pairs of substrate products are combined to identify the maximum subgraph between all substrates and products. The identification of the optimal MCS solution for the reaction requires that the solutions from the pairwise matching are combined so as to reduce the overall bond changes.

Multiple mapping solutions exist not only due to the presence of equivalent or symmetric groups but also due to possibly alternate combinations of MCS solutions. Therefore, it is imperative to ensure that all the MCS solutions are correctly identified and assessed for biochemical feasibility. As in most methodologies,[8,11,35,36] we use the principle of minimal chemical distance (PMCD) heuristic proposed by Jochum et al.[5] PMCD states that computational procedures that predict reaction mechanisms based on the largest substructure overlap alone might not always be chemically acceptable. Instead, a MCS solution that minimizes the reordering of bonds is generally closer to the actual reaction mechanism. Figure 6 presents two optimal solutions for atom mappings of D-glyceraldehyde-3-phosphate glycolaldehyde transferase reaction. In both cases, a minimum of three bond breaks are needed for the transformation between the substrates and the products. Both the reaction atom mapping solutions are predicted using the PMCD "largest set of largest substructure" heuristic.[5] The mapping solutions have been identified by combining the maximum subgraph solutions between the substrate−product pairs. The order of this comparison dictates that the outcome of the atom mapping solution and hence each comparison has to be assigned a precedence order. A precedence order is needed to ensure that we correctly recognize the largest overlap with minimum bond breaks between a pairwise comparison as the first solution. For example in Figure 6C, only one bond is broken in the MCS between D-xylulose 5-phosphate and D-Glyceraldehyde 3-phosphate. The MCS solution indicated by ald receives the highest precedence. Also, the MCS between D-erythrose 4-phosphate and D-glyceraldehyde 3-phosphate involves a single broken bond and would get the highest precedence in an alternate solution scheme. The subgraphs that match with the least bond changes are chosen and successively removed from additional MCS comparisons. This greedy heuristic can be solved using a variety of algorithms. We chose to use the minimum spanning tree (MST) to identify a precedence order for the various MCS solutions. The following summarizes the procedure for building an MST precedence order:

(i) Identify the MCS between the pairs of molecular graphs of products and reactants and ascertain the number of bonds broken and formed.

ii) Generate a transition graph with weighted and directed edges. The vertices in this graph represent the task of comparing the two molecular graphs. The edges incident from these vertices are designated edge weights that represent the output for the comparison. If we consider the minimization of bond changes, each edge is given a positive weight. The positive value on the outgoing edge represents the number of bonds that need to be broken or formed to get a maximum common subgraph. For example, in Figure 6, vertex ald represents a comparison between D-xylulose 5-phosphate and D-glyceraldehyde 3-phosphate. The outgoing edge is always given a weight of "1" because one edge has to be broken to derive the maximum common subgraph.

(iii) Using the Kruskal minimum spanning tree (MST) algorithm, we identify the spanning tree with the least weight. Each vertex in the tree identifies a MCS comparison with the highest precedence given to the root.

(iv) The MCS solution from the root is fixed, and the edge weights are recalculated. The edge weights are modified to reflect the solution proposed by the root. For example, in Figure 6D, after we consider the MCS solution ald, comparison alc requires effectively no bond breaks, as no MCS between "a" and "c" can be found anymore. bld requires two bond breaks for a MCS between "b" and the remaining section of "d". Therefore, all edges connected to the vertices comparing "a" or "d" are updated with new weights in the transition graph.

(v) Upon completion of the ald operation, all edges outgoing from this vertex are removed.

dx.doi.org/10.1021/ci5003922 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

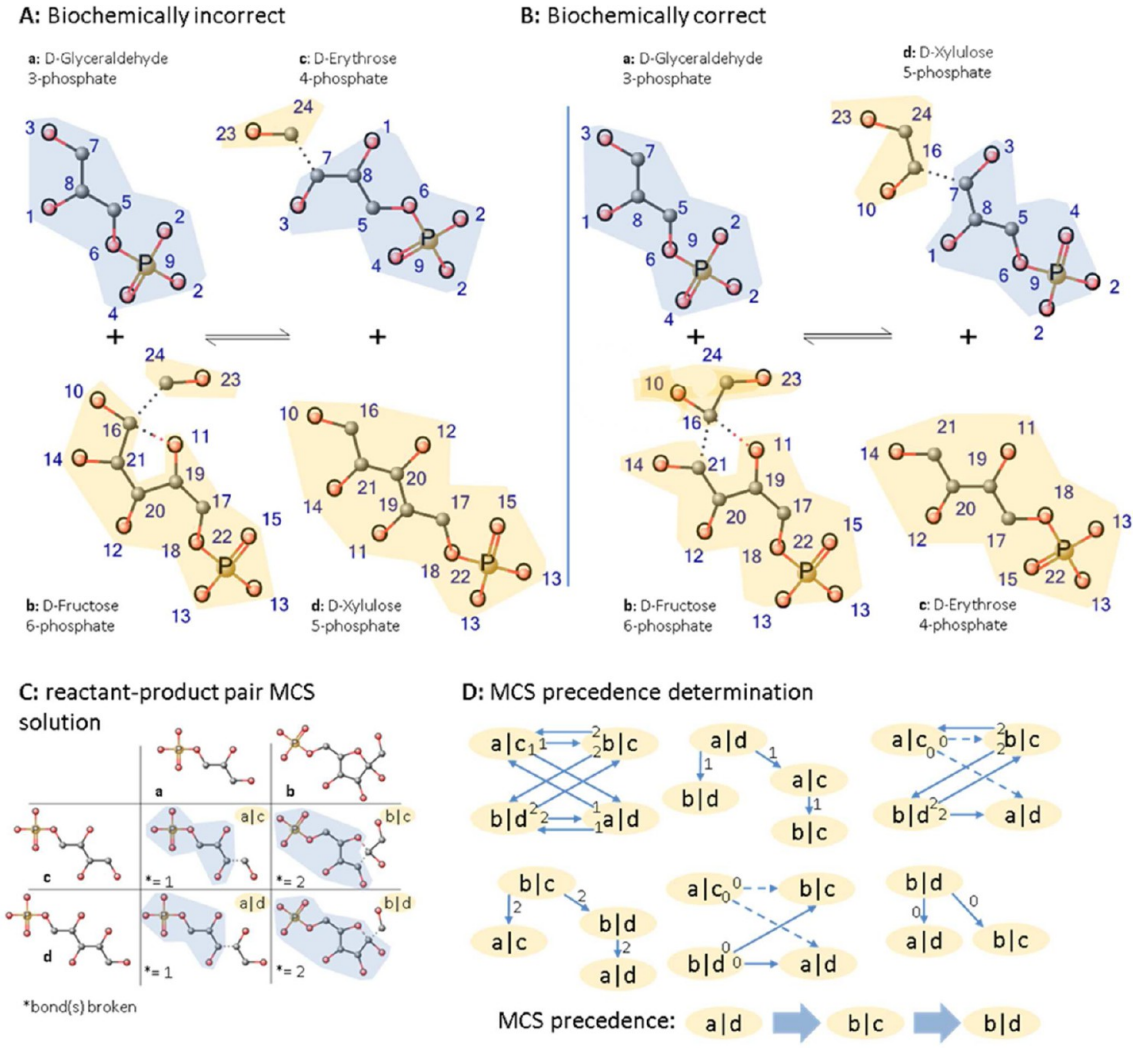


Figure 6. Alternate solutions. The reaction D-fructose 6-phosphate:D-glyceraldehyde-3-phosphate glycolaldehyde transferase has two atom mapping solutions. Both solutions shown in panels A and B follow the principle of minimum chemical distance[5] (PMCD). The highlighted regions in panels A and B represent the common substructures between reactant and product. Reactant−product pairwise MCS searches are performed (panel C), and MCS combinations that obey PMCD are chosen (panels A and B). To identify a MCS combination with the least number of bond breaks, a greedy procedure using minimum spanning tree (MST) is implemented (panel D). After each MST is identified, the subgraphs in the MCS solution suggested by the root are removed, and a reactant−product pair MCS search is repeated. Because four MCS solutions exist, at most four MSTs are required. The root for the first MST receives the highest precedence, while the root for the last MST receives the least precedence. The pairwise MCS solutions are then combined using the root precedence order to identify complete reaction atom mapping solutions. The MCS precedence depicted in panel D identifies the biochemically correct solution shown in panel B.

(vi) A MST is identified again, and the MCS solution suggested by the new root is appended to earlier MCS comparisons. We iterate to update the transition graph and identify a new root from the MST again every time we fix a MCS solution.

(vii) The iterations continue until all the vertices are fully disconnected or there is no change in the solutions proposed.

Moreover, we also identify all possible minimum spanning trees possible for a reaction. Each alternate spanning tree identifies a unique route to identify an optimal solution. In the example shown, for the minimum bond change condition, two minimum spanning trees are suggested with highest precedence with roots alc and ald. Each tree suggested by the minimum spanning tree indicates the order of comparison to be considered to obtain a complete atom mapping solution. For example, the two solutions suggested in Figure 6A by the Kruskal minimum spanning tree algorithm have the root as alc, while the example in Figure 6B has the first root as ald.

Using CLCA for MCS detection and the above stated greedy approach for the MCS solution combination, we generate atom mapping solutions for over 27,000 MetRxn reactions. Figure 7 shows examples of six different reactions each representing common transformations in metabolism with their reaction centers highlighted. The reactions are (A) methionine synthase, (B) glutamate decarboxylase, (C) L-threonine dehydrogenase, (D) glutamine synthetase, (E) D-alanine-D-alanine dipeptidase, and (F) glucose-6-phosphate isomerase. The reaction graph of D-alanine-D-alanine dipeptidase depicts the symmetry due to a stoichiometry of two for D-alanine due to the cleavage of D-alanyl-D-alanine. The reaction is catalyzed by the enzyme D-alanine-D-alanine dipeptidase, a Zn^{2+}-dependent enzyme. The enzyme encoded by the gene vanXB in *Enterococcus faecium* BM4147 has stereospecificity to only D-alanyl-D-alanine and does not accept

dx.doi.org/10.1021/ci5003922 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

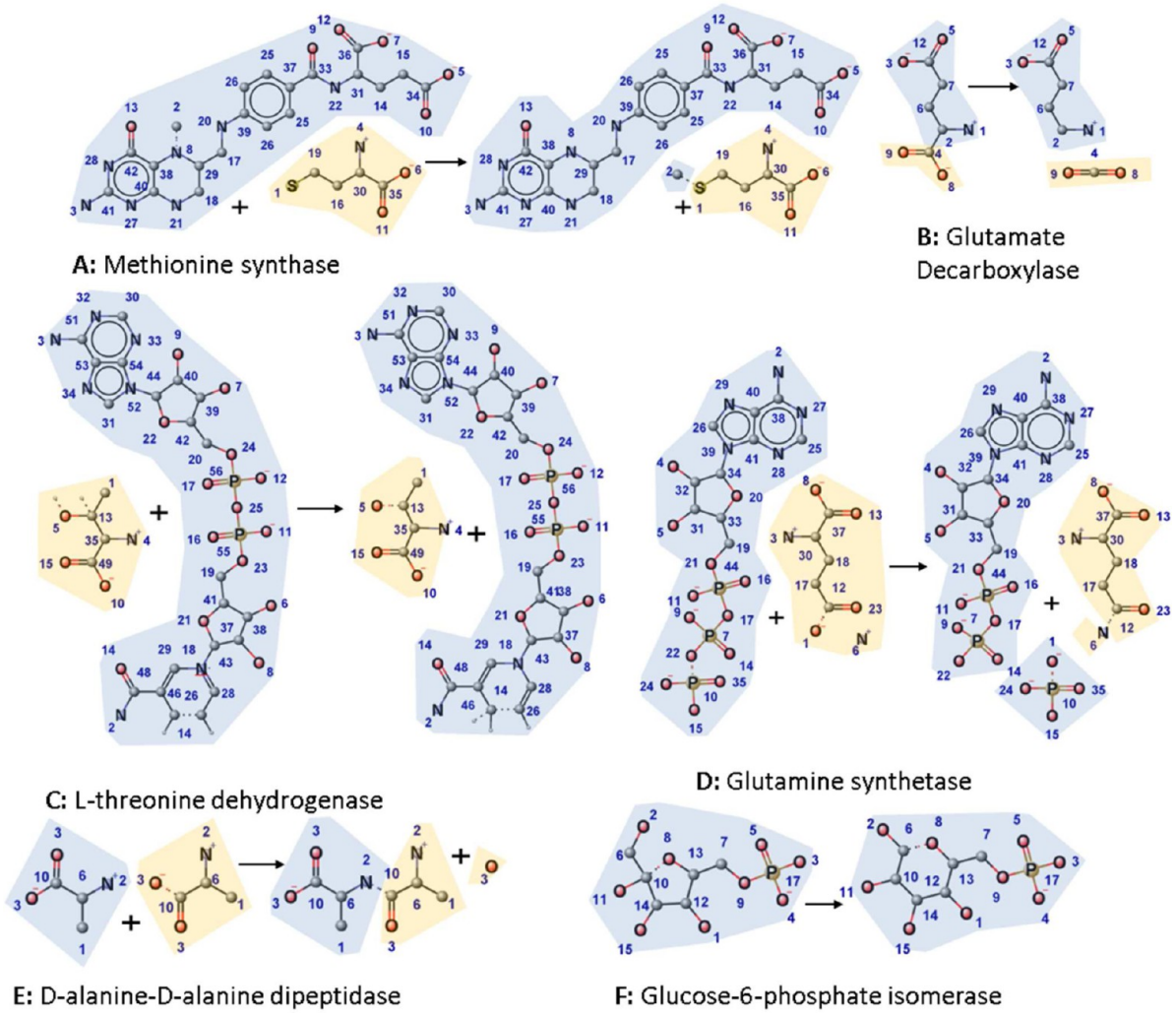


Figure 7. Reaction atom mapping. CLCA identifies correct reaction atom maps for reactions representing six common EC classifications. The highlighted regions identify the common fragments between reactant and product.

the other three RS, SS, and SR stereoisomers as substrates.[37] L-Threonine dehydrogenase is an oxidoreductase enzyme providing an example where only changes in bond order were identified without changes in the position of atoms. Methionine synthase, a transferase enzyme, transfers a methyl group from L-homocysteine to 5-methyltetrahydrofolate to form L-methionine. The changes involved are bond breaks, bond formation, and change in position of two vertices. Glutamate decarboxylase, a hydrolase, provides an example of a cleavage reaction with changes in atom and bond positions between the reactant and product graphs. Glucose-6-phosphate isomerase leads to changes in more than one bond, and glutamine synthetase results in bond and atom position changes in all the three molecular graphs. All six reactions combine cleavage, condensation, substitution, and unimolecular reaction operations.[38]

The utility of CLCA in MCS detection is not limited to metabolite graphs alone. We extend the application of CLCA to larger graphs and identify the conserved subsections between pairs of reaction or pathway graphs. The conserved subsections between pairs of reaction or pathway graphs elucidate the common transformation mechanism from one molecule into another. In MetRxn, we incorporate reaction and pathway information from multiple sources. In order to identify common transformations, we perform a comparison for similarity between two reaction graphs and two pathway graphs as shown in Figures 8 and 9, respectively.

Common Subgraphs between Two Reactions. As shown in the previous sections, a one-to-one MCS search between a pair of molecules is required for reaction atom mapping. The utility of the CLCA algorithm is however not limited to MCS searches between a pair of molecular graphs alone and can also be extended to calculate similarity between larger multiple molecular graphs. We can use similarity scores to cluster, organize, and annotate reactions with information such as EC numbers. For example, in a reaction similarity calculation, a MCS search between the entire set of molecular graphs representing a pair of reaction is performed. We first find the common subgraph between the two reactions and calculate the similarity in the number of atom/vertices present in the subgraph. As shown in Figure 8, we compare two amino acid oxidoreductase reactions. CLCA clearly identifies the conserved phenyl, carboxyl, and amino groups in the reactant side and the conserved phenyl, carboxyl, and oxo groups in the product side of both the reactions. These conserved components, including the reaction center (shown as dotted bonds), are part of the common subgraph between the two reaction graphs. The only difference arises in the hydroxyl groups present in L-tyrosine:NAD+ oxidoreductase. The common subgraph identifies how similar

dx.doi.org/10.1021/ci5003922 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

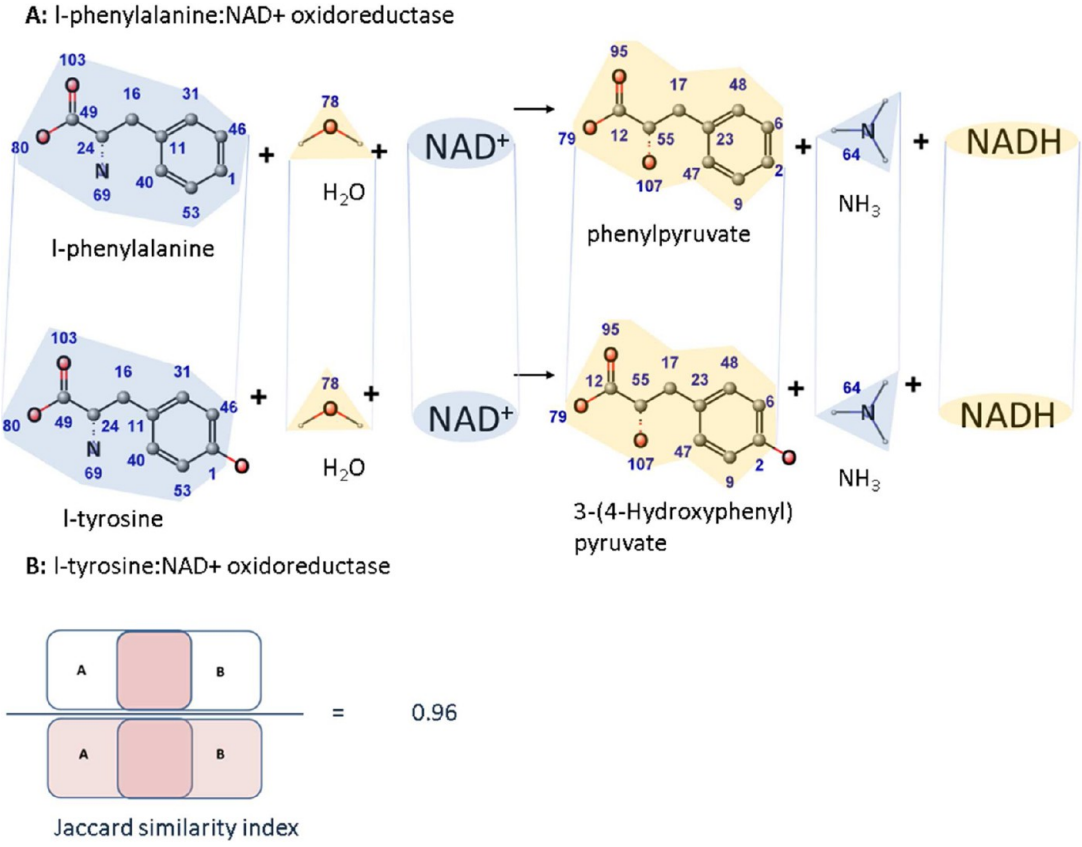


Figure 8. Reaction similarity. Panel A shows reaction L-phenylalanine:NAD+ oxidoreductase, and panel B shows reaction L-tyrosine:NAD+ oxidoreductase. Reaction centers are depicted using dotted bonds, and common substructures are highlighted with contrasting colors for visual clarity. A Jaccard similarity value of 0.96 is calculated by the number of atoms in the common substructures.

the two graphs are to each other. We calculate the Jaccard similarity score by finding the ratio of twice the number of vertices in the common subgraph to the total number of vertices in both the reaction graphs. Users of MetRxn[1] can search for reactions based on similarity scores for a query reaction or even sets of molecular graphs. Similarity scores in MetRxn can assist users to transfer ontological information such as enzyme commission (EC) number annotations to EC unannotated reactions from EC annotated reactions. Historically, all EC information for reactions are manually annotated and take into account the cofactors, reaction center, and the substrate/product structures. Recent efforts for automated EC classifications are not powerful enough to classify a reaction fully and still depend upon expert knowledge to annotate the fourth number correctly. To identify the best EC class, these automated methods calculate similarity scores, (e.g., Tanimoto, Jaccard, Sorensen-Dice, etc.) using chemical fingerprints[39] or computationally intensive MCS calculations[35,40−44] for reaction center detection. CLCA has advantages over existing methods because it does not require chemical fingerprints and prior reaction center information for assisting EC annotation of reactions.

Common Subgraphs between Two Metabolic Pathways. Similar to the reaction similarity detection problem, CLCA can be used to identify similarity between metabolic pathways. The pairwise pathway alignment problem[45,46] is the task of calculating similarity between two pathways. A large number of pathway alignment algorithms align protein−protein interaction networks or gene−protein interaction networks. Few algorithms for metabolic pathway alignment align metabolic networks using EC numbers, reaction centers, or metabolite identity. Methods that align using EC classifiers, align pathways without considering the fourth classifier of the EC number.[47] Methods that align with metabolite identity cannot handle branched or cyclic pathways because the aligned pathways need the start substrate and end product to be common.[48] Similarity calculated on atomistic details provides higher resolution by identifying the conserved moieties as well as conserved reaction mechanisms. As mentioned previously, the pairwise pathway alignment problem is also a MCS search problem. Aligning pathways[49] using atomistic details of metabolites would be intractable because they cannot handle large graphs. CLCA therefore has clear advantages over existing techniques because it (i) improves resolution of overlap by comparing atoms and bonds, (ii) accommodates reactions without EC annotations, (iii) can handle branched pathways, and (iv) has polynomial time complexity. Figure 9 shows a comparison between two branched chain amino acid degradation pathways. We compare valine degradation to isoleucine degradation and correctly identify the common moieties and conserved reaction mechanism between the two. The first three reactions of transamination, decarboxylation, and dehydrogenation are catalyzed by three common enzymes. The next two reactions of hydration and dehydrogenation are catalyzed by distinct enzymes. We note that CLCA correctly identifies the reaction centers as well as the moiety transferred at each step.

3. RESULTS AND DISCUSSION

Ahead of the mapping calculations, all reactions were balanced and standardized using the MetRxn[1] workflow. All informa-

dx.doi.org/10.1021/ci5003922 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

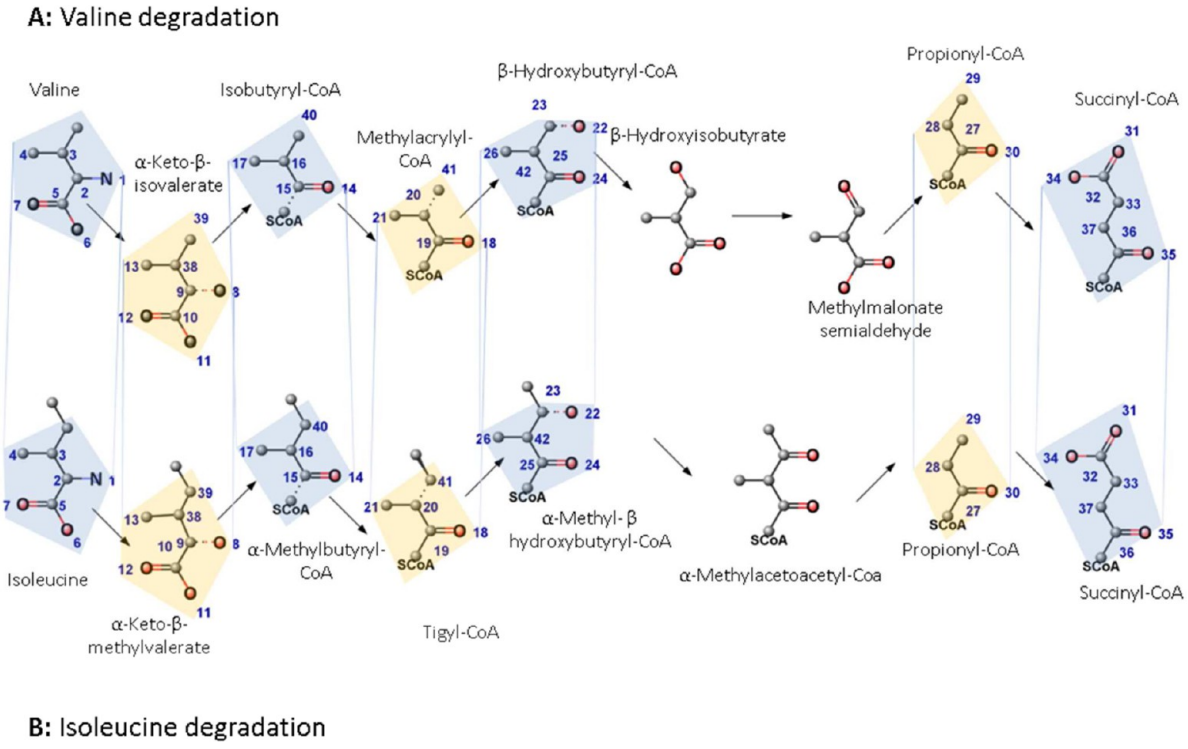


Figure 9. Comparison of the two branched chain amino acid degradation pathways for valine (A) and isoleucine (B). Common subgraphs between the two pathways are highlighted, and the reaction centers are identified by the dotted bonds for visual clarity.

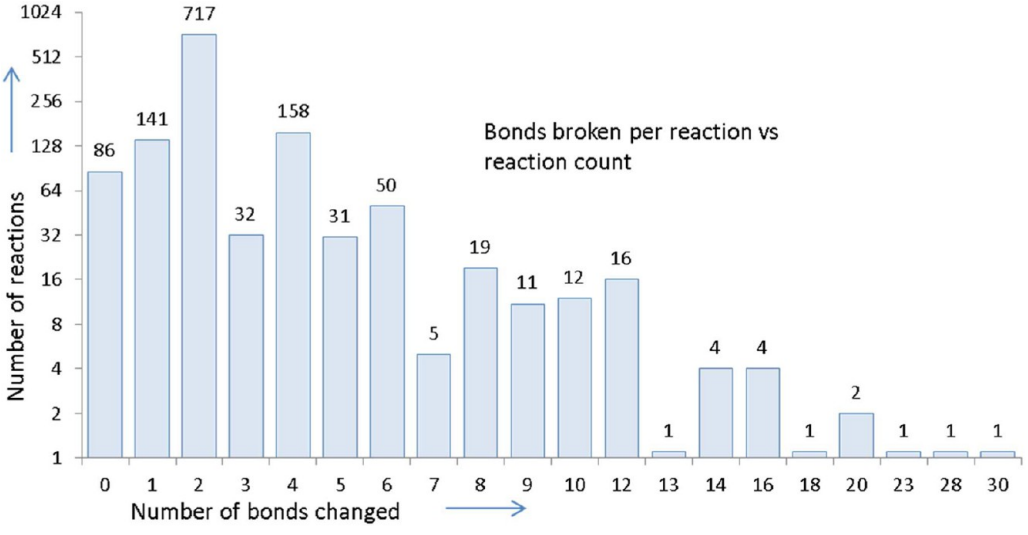


Figure 10. Bond changes per reaction statistics for the *E. coli* *i*AF1260 metabolic model. The *x*-axis depicts the number of bond changes all substrates and products in a reaction. This graph allows us to visually isolate reactions that may have been mapped incorrectly as denoted by too many bond changes. Note that 86 reactions were mapped without any bonds changes, as we do not track changes due to bond order change (i.e., single to double) or bonds with H atoms.

tion regarding protonation, stereochemistry, and bond order of atoms were removed or standardized to produce a unique canonical form. Each reaction is unique only in its connectivity. All examples and results presented henceforth consider this canonical form of molecular graphs, unless stated otherwise. Metabolic information from over 8 different databases and 112 different metabolic models is combined to produce a canonical data set of over 27,000 reactions. CLCA and A* search took an average of 140 ms per reaction or 1.3 ms per atom to calculate a single atom mapping solution. Overall, when only a single solution is requested, the run over the entire database was completed in 64 min using a standard desktop with 2.3 GHz CPU and 8 GB memory. When all possible alternative solutions were requested, the complete task took 160 ms on average to terminate. When compared with existing methodologies, CLCA performs better than other procedures in terms of literature reported run time and accuracy. Alternative mapping solutions arising due to group equivalence as well as alternative objectives were identified rapidly. The largest run time for any reaction was 7 s,

dx.doi.org/10.1021/ci5003922 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

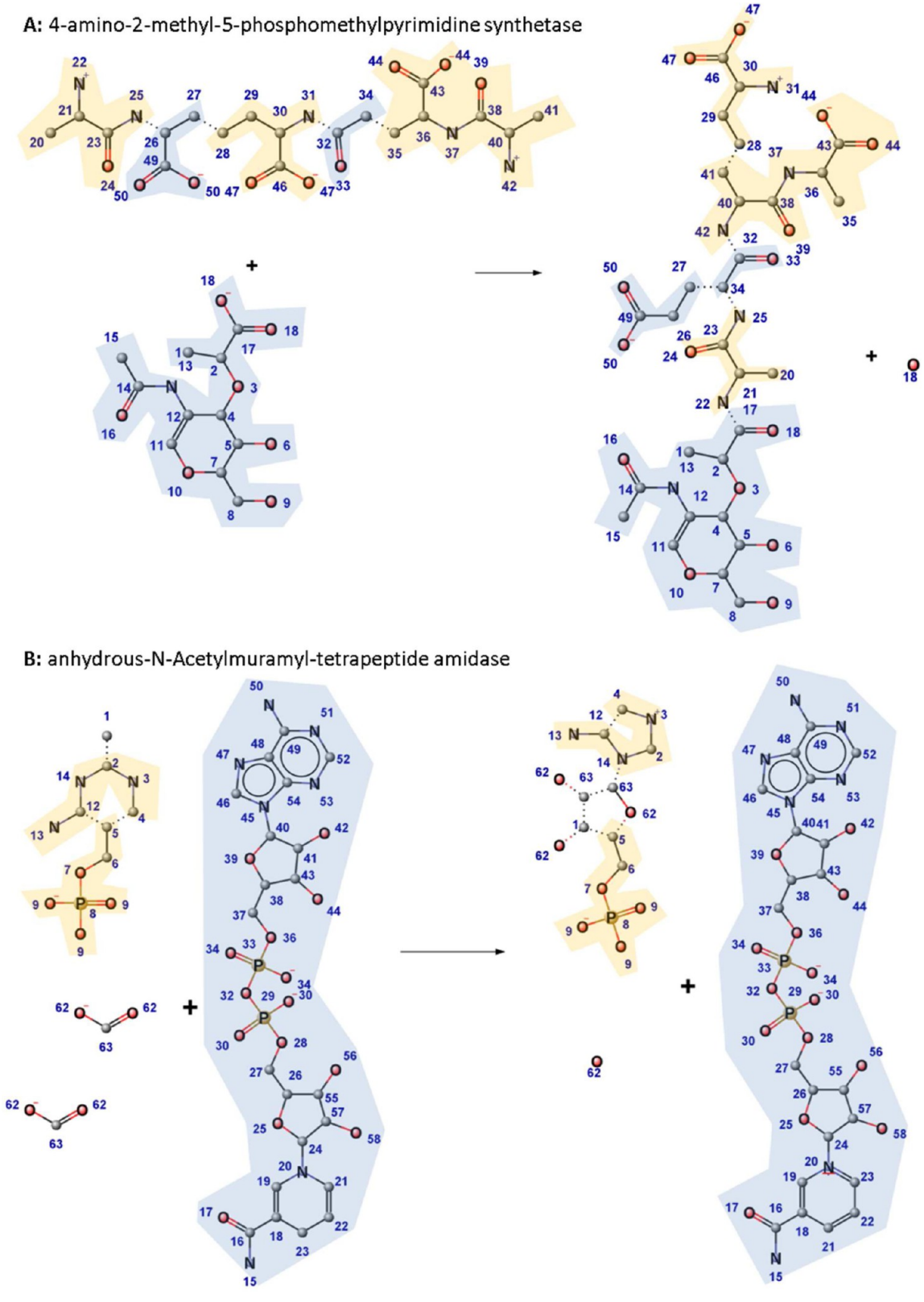


Figure 11. Example of possibly incorrect mapping from *i*AF1260. Reaction 4-amino-2-methyl-5-phosphomethylpyrimidine synthetase (panel A) and anhydrous-N-Acetylmuramyl-tetrapeptide amidase (panel B) as mapped by CLCA involved too many bond breaks and formations (>6). The connected common subgraphs between reactant and product are highlighted with contrasting colors for visual clarity.

for the reaction O16 antigen (×4) ligase (periplasm),[26] including more than 2000 atoms.

Application to *E. coli* *i*AF1260 Metabolic Model. We perform various comparisons, automated as well as manual, to validate and test the robustness of CLCA. We started with a manual inspection of all the reactions mapped in the *E. coli* *i*AF1260[26] metabolic model. We flagged reactions with an abnormally high number of bond changes for manual curation.

dx.doi.org/10.1021/ci5003922 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

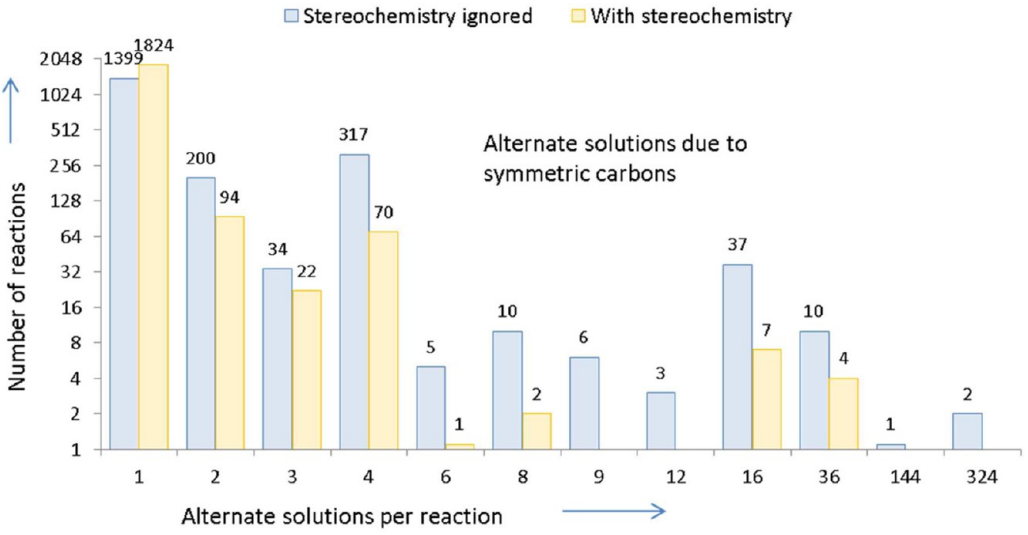


Figure 12. Alternate solutions due to equivalent or symmetric carbon groups in the *E. coli* *i*AF1260[26] metabolic model. The *x*-axis depicts the number of alternate mapping solutions a reaction might have when stereochemistry is ignored/considered. As many as 425 reactions that were producing alternative mappings were reduced to unique mappings when stereospecifity information was considered.

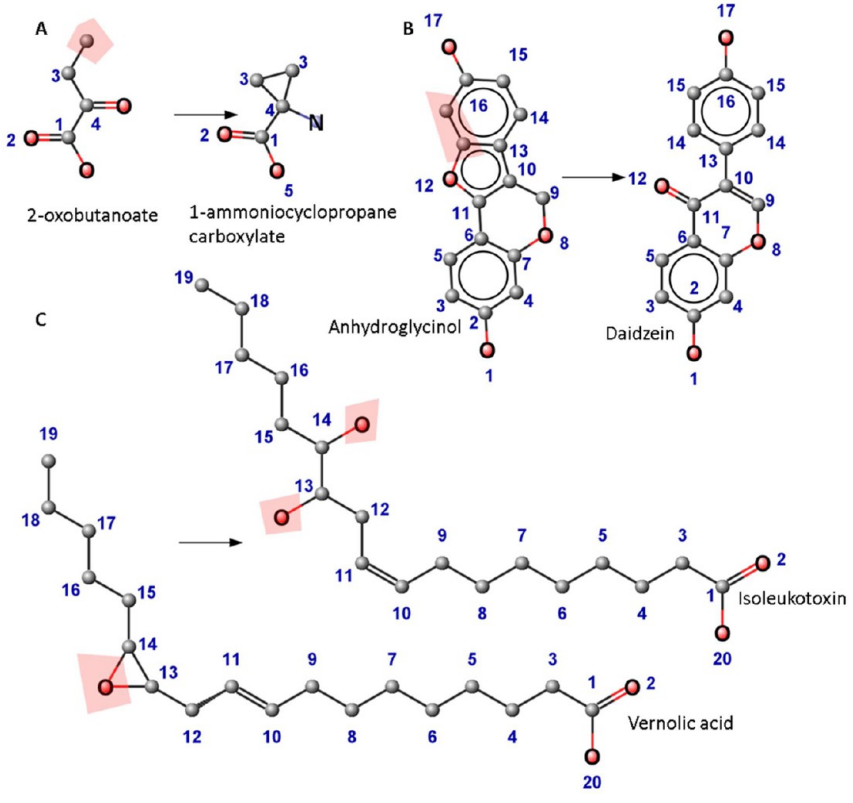


Figure 13. CLCA incomplete mapping. CLCA was unable to proceed with maximal mapping due to the presence of symmetric members in the reaction center of cyclic moieties. Such cases were rare, and only 33 such reactant pairs were found in the entire MetRxn database. A complete mapping solution for transformation in panel B was however realized by A* search, but the mapping solution for atoms highlighted in panels A and C require additional chemical information.

The bar graph shown in Figure 10 shows the number of reactions vs number of bond changes per atom mapping solution excluding transport and exchange reactions. Overall atom mapping statistics for the remaining 1293 reactions in the *E. coli* metabolic model are depicted in Figure 10; 84 reactions were mapped with zero bond changes. These reactions either undergo protonation changes (no hydrogen atoms are tracked) or bond order changes (e.g., single to double). There are also reactions with a large number of bond changes (~30). Such reactions have substrates with a large stoichiometry (>15) combining multiple elementary steps into a single one. We found seven reactions to be incorrectly mapped in the *E. coli* *i*AF1260[26] by CLCA. Figure 11 shows two such examples. We find that all the reactions failed to map aggregate multiple elementary transformations. The reaction

dx.doi.org/10.1021/ci5003922 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

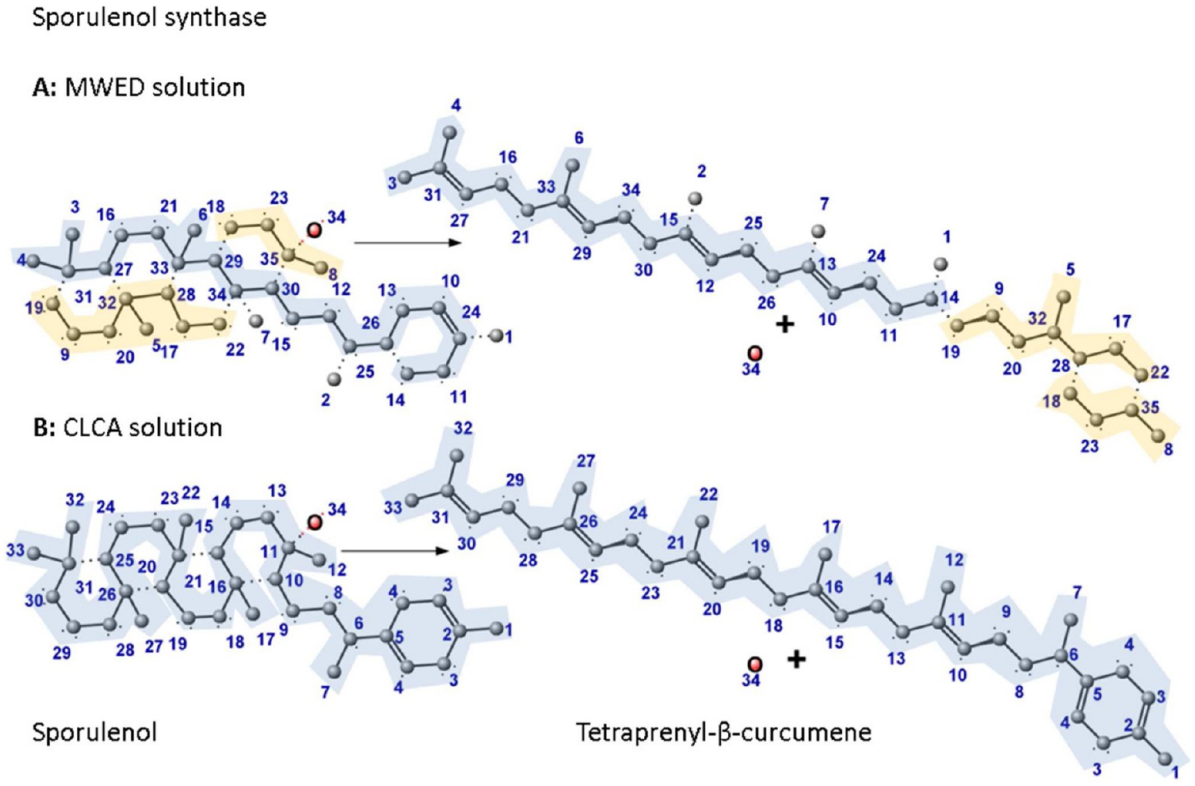


Figure 14. Comparison with MetaCyc MWED. For reactions involving polycyclic compounds, MWED suggested a large number of bond changes, often suggesting breaking up of the entire reactant molecule before its transformation into the product molecules. As many as 34 additional reactions where the polycyclic topology was not preserved during the transformation led to incorrect atom mapping solutions. Panel A shows the MWED generated solution, whereas panel B shows the CLCA correctly mapped solution as confirmed in KEGG. CLCA identifies the mapping solution with only five bond deletions on the reactants side. The highlighted regions in contrasting colors depict the connected common subgraphs between reactant and product. When compared to MWED, CLCA identifies a larger connected common subgraph with a lesser number of bond rearrangements (i.e., shown as dotted bonds).

anhydrous-N-acetylmuramyl-tetrapeptide amidase combines multiple reactions of N-acetylmuramoyl-ala amidohydrolase and tetrapeptide L,D-carboxypeptidase, part of the murein recycling pathway. N-Acetylmuramoyl-ala amidohydrolase at each step elongates the peptide chain on N-acetylmuramate by adding L-alanine. The reaction tetrapeptide L,D-carboxypeptidase cleaves the molecule L-alanine-D-glutamate-meso-2,6-diaminoheptanedioate-D-alanine one by one generating five L-alanine molecules. Reactions represented in highly lumped form may lead to incorrect mappings as the exact bond operations are concealed.

Alternate solutions in *E. coli* *i*AF1260 Metabolic Model Due to Equivalent Groups. We generate the atom mapping solution for two cases. In the first case, we generate homotopic solutions by ignoring the stereochemical information on the molecule. We consider stereochemistry in the second case. Figure 12 shows the difference in the number of unique solutions when stereochemistry is considered. We notice a sharp increase in the number of reactions with only one unique solution. This is representative of the fact that most enzymes are highly stereospecific, and the transitions identified have to consider stereochemical information in its solution. For example, the enterobactin transport reaction allows the transport of enterobactin from cytosol to the extracellular environment, involving nine alternate solutions as the number of symmetries in the reaction graph that can be calculated by $\Pi_i S_i! \times c_i^{S_i}$, where S_i is the stochiometric coefficient and c_i is the number of inherent symmetries of the ith molecule.[50]

Comparison with Existing Efforts. MetRxn aggregates information is primarily from three metabolic databases, BRENDA,[51] KEGG,[52] and MetaCyc,[23] of which KEGG and MetaCyc also provide atom mapping information. In addition, recent efforts from Fooshee et al.[25] and Kraut et al.[27] have made available online reaction mapping tools. We compare atom mapping data from KEGG,[52] MetaCyc,[9] ReactionMap,[25] and ICMAP[27] with the atom mapping data generated by CLCA. The KEGG RPAIR[24] database is a manually compiled list of reactant pair alignments. To generate reactant pair alignments, each reactant molecular graph is decomposed into 68 functional groups and atom microenvironments.[24] The functional group, atom microenvironment, and alignment information are available in a proprietary format called the KEGG chemical function (KCF) format. We downloaded KCF and MOL files for 2636 reactant pairs from http://www.kegg.jp/, version 71.0, and converted them to the SMILES[3] format before running CLCA. We found high agreement (99.3%) between the alignments suggested by CLCA and the manually identified alignments from KEGG. Only 16 of the alignments suggested by CLCA disagreed with KEGG RPAIR because they were nonmaximal or incorrect in the alignments. Few examples for the disagreements are shown in Figure 13. The first alignment between 2-oxobutanoate and 1-aminocyclopropane-1-carboxylate was incomplete because one of the two symmetric carbons in the cyclopropane ring was not mapped. CLCA was unable to determine the correct bond breakage because structurally both the C−C bonds are equivalent, and breakage of any one of them would give a

dx.doi.org/10.1021/ci5003922 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

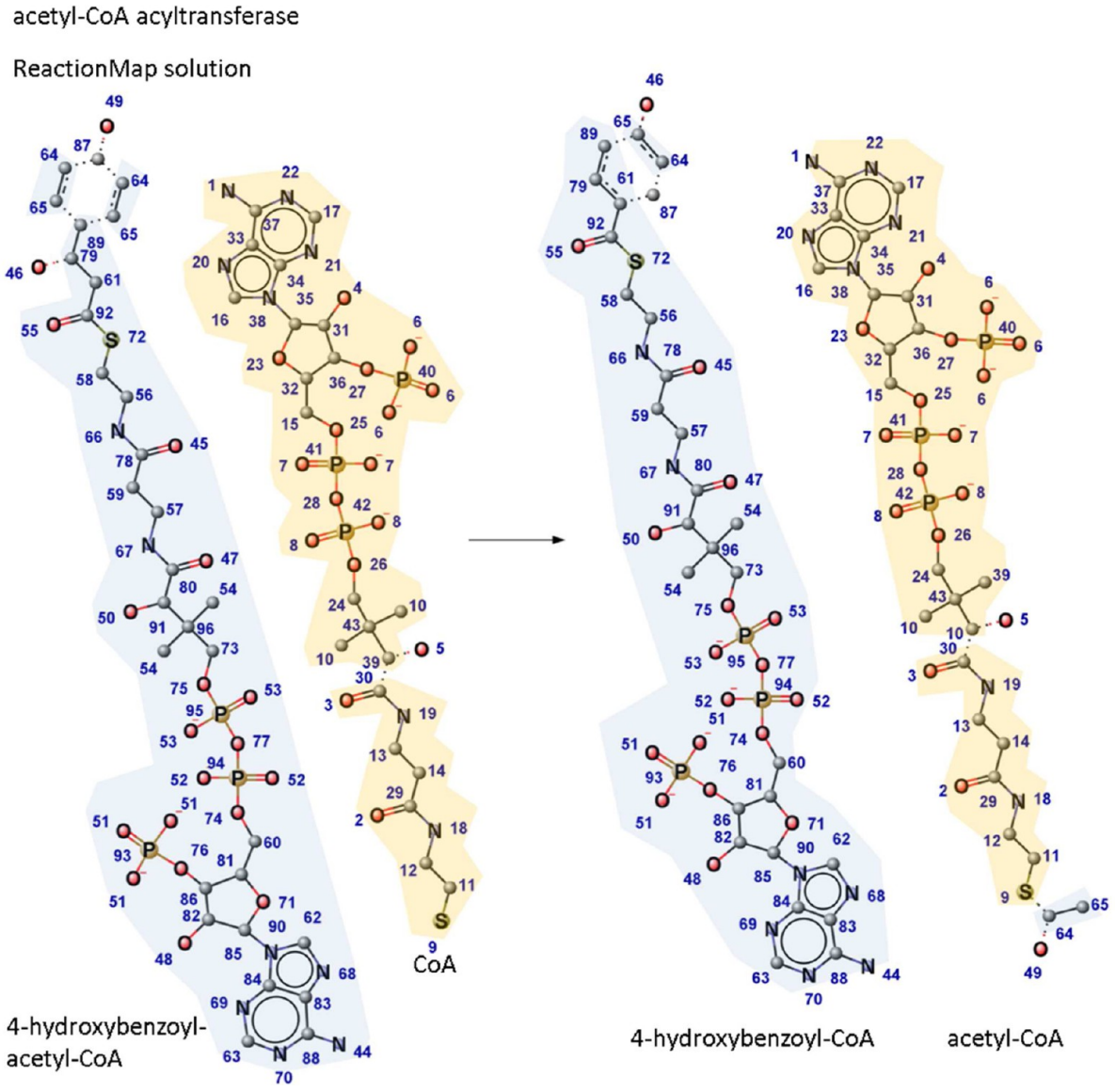


Figure 15. ReactionMap solution for acetyl-CoA acyltransferase. For the conversion of CoA (coenzyme A) to acetyl-CoA by acetyl transfer, ReactionMap suggests 16 bond rearrangements. The solution suggested by ReactionMap indicates the reaction center (shown as dotted bonds) exists on hydroxybenzoyl and neopentane moieties for the molecules 4-hydroxybenzoyl-acetyl-CoA and CoA, respectively.

maximal alignment. Similarly in Figure 13, the second alignment between anhydroglycinol and daidzein involves a C−O bond breakage on the furan moiety with two possibilities. The phenol and naphthol moieties are connected by a rotatable bond, and CLCA could not ascertain which of the two carbons (in phenol) in the third position was previously connected to the oxygen atom. In the third example between (−)-leukotoxin B and isoleukotoxin, CLCA was unable to map the oxygen atom on the oxirane ring of (−)-leukotoxin B. In cases wherein symmetric members of ring moieties were involved in the breakage or formation of bonds, CLCA could not proceed to identify a complete match. As shown in Figure 13, maximal overlaps might not necessarily give a correct solution, and biochemical knowledge will always be needed to ascertain the correct bond breakage and formation. The remaining 13 cases of disagreement are shown in the Figure S2 of the Supporting Information. Within the MetRxn database, 33 reactant pairs were found with such characteristics, and we flag such alignments for manual validation.

MetaCyc[23] is a reference database consisting of metabolites, enzymes, reactions, and metabolic pathways database from more than 30,000 publications. As part of our goal is to align large metabolic databases, we have successfully cross-referenced all metabolite and reaction information from MetaCyc with KEGG,[52] BRENDA,[51] RHEA,[53] CHEBI,[54] HMDB,[55] and Reactome.[56] Atom transition information using the recently published MWED (minimum weighted edit distance) procedure for all reactions in MetaCyc was made available with the release of version 16. The MWED model is an MILP formulation with an objective to minimize the cost of breaking and forming bonds while matching reactants and products. MWED also uses an empirically derived bond breakage/formation propensity metric to get an optimal matching solution. Ring detection and matching is performed separately as a preprocessing step to the complete matching. We obtained the reaction atom mapping information as a SMILES output for over 10,000 reactions for version 18.1 directly from the author. We compared the 10,585 reaction atom mapping solutions of MWED[9] with the

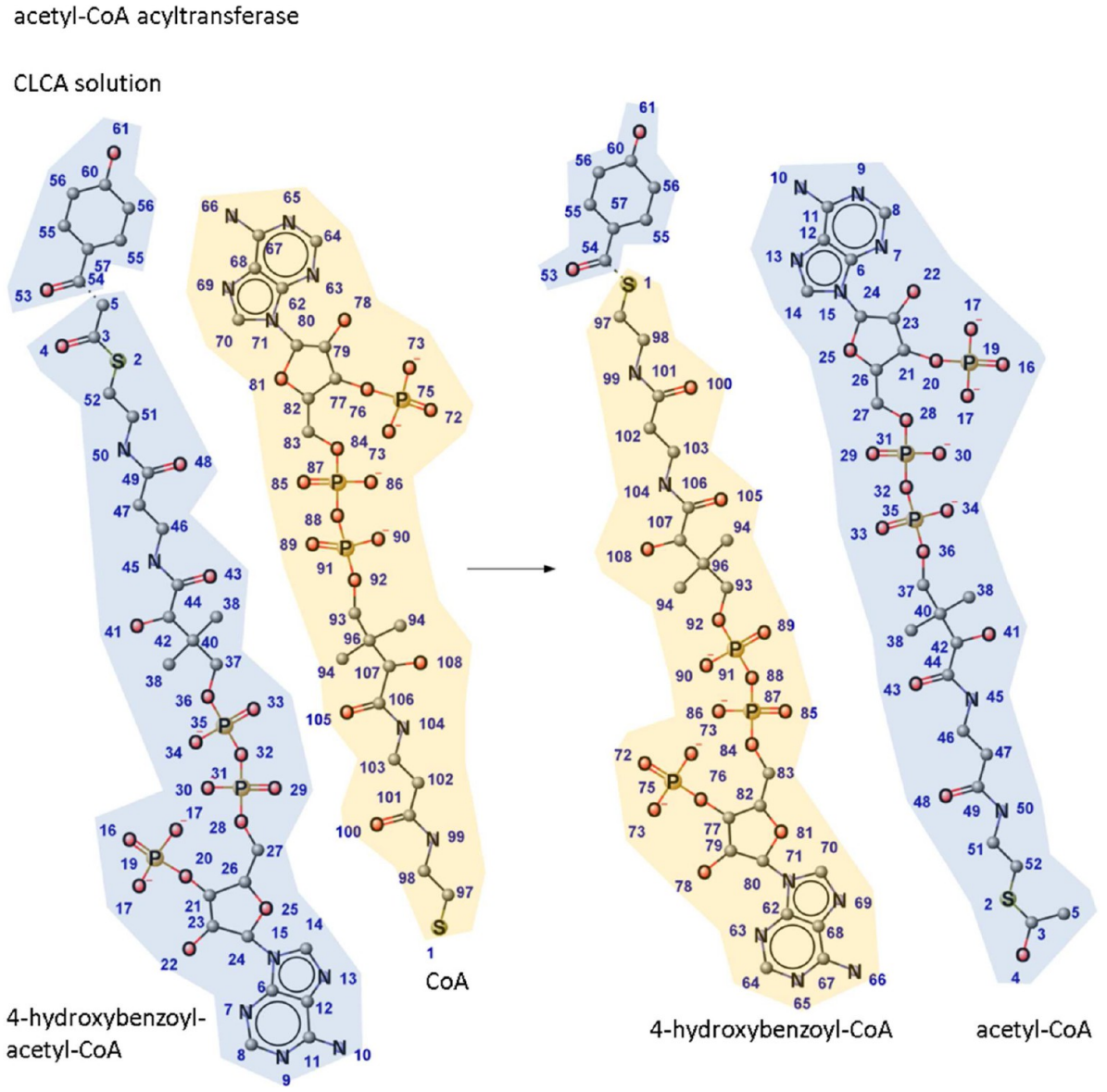


Figure 16. CLCA solution for acetyl-CoA acyltransferase. For the conversion of CoA to acetyl-CoA by acetyl transfer, CLCA suggests two bond rearrangements. The solution suggested by CLCA indicates the transfer of 3-hydroxybenzaldehyde from 4-hydroxybenzoyl-acetyl-CoA onto CoA to produce 4-hydroxybenzoyl-CoA and acetyl-CoA.

atom mapping solutions using CLCA. We noticed disagreement with only 232 mismatches. CLCA suggested fewer bond breaks for 198 reactions, while MWED suggested fewer bond breaks for 34 mapping solutions. Biochemical knowledge was used to manually verify the 232 CLCA solutions and 66 erroneous reaction atom mapping solutions were identified. Most of these cases were similar to the examples in Figure 13 or cases where elementary reactions were aggregated into a single reaction. Figure 14 provides an example of MWED failure mode where CLCA was able to correctly identify a solution with fewer number of bond rearrangements. In the reaction sporulenol synthase, for the conversion from sporulenol to tetraprenyl-β-curcumene, 11 bond deletions on sporulenol were suggested by MWED, whereas CLCA identified an optimal mapping solution with only five bond deletions as shown in Figure 14. We found 38 such reactions wherein linear molecules transformed into polycyclic molecules. Most of these reactions were ring forming/breaking reactions with multiple bonds being formed or broken in the transformation between products and reactants. The MWED ring detection and matching step identifies only the rings that are conserved between the reactant and product. Therefore, in some cases where the polycyclic topology is not conserved between the reactant and products, MWED fails to detect the correct mapping.

In addition to the comparisons presented in the previous section, we compare the CLCA atom mapping solutions of 1000 randomly chosen MetRxn reactions with the atom mapping solutions provided by the ReactionMap algorithm. The ReactionMap algorithm by Fooshee et al.[25] uses a combination of MCS search and bipartite matching steps to produce the atom mapping solution. In the first step of ReactionMap, a partial MCS mapping is identified using the OEChem toolkit.[57] Then, the bipartite matching step[25] extends the MCS solution by incorporating chemical knowledge in the form of SMILES encoded cost functions. For 979 reactions, at least one CLCA generated solution matched the atom mapping solution provided

dx.doi.org/10.1021/ci5003922 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

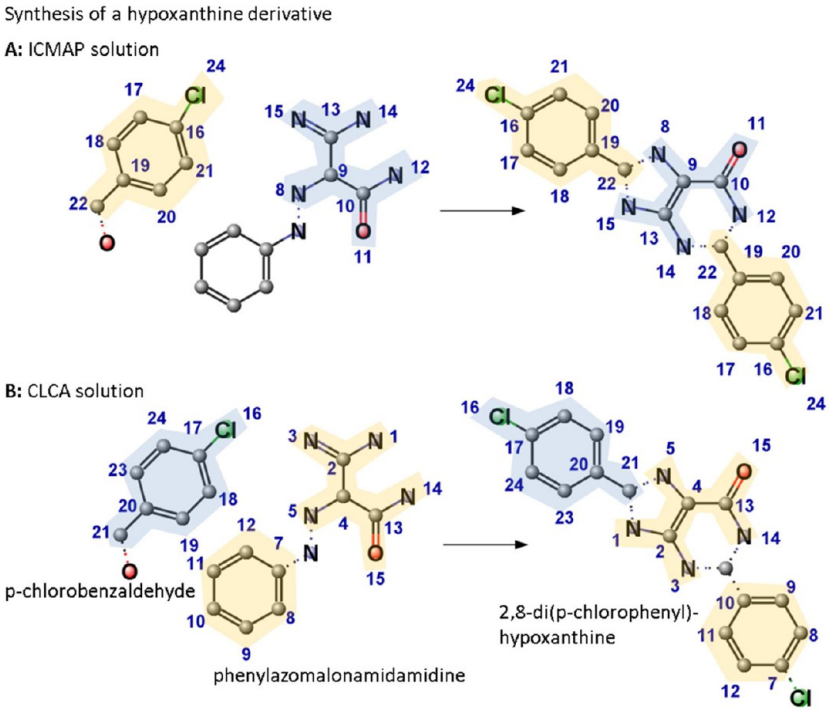


Figure 17. Incorrect CLCA solution for a CLASSIFY data set reaction. Panel A shows a correct solution from ICMAP wherein the phenyl moiety of phenylazomalonamidamidine remains unmapped. Panel B shows an incorrect mapping solution from CLCA wherein the phenyl moeity from phenylazomalonamidamidine is mapped to the chlorophenyl moiety 2,8-di(*p*-chlorophenyl)-hypoxanthine.

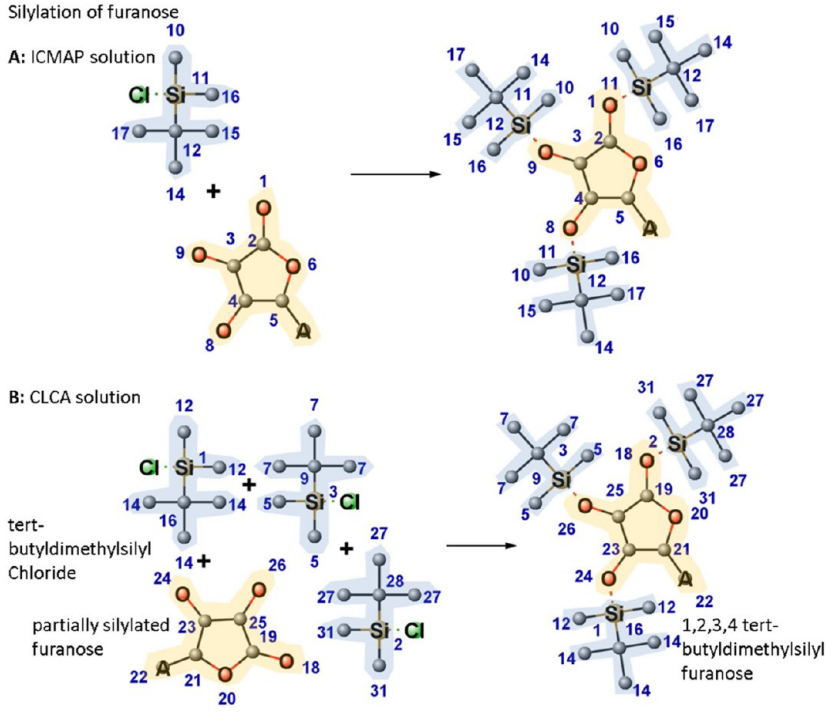


Figure 18. Correct CLCA solution for a CLASSIFY data set reaction. Panel A shows the mapping solution generated by ICMAP wherein the atoms from *tert*-butyldimethylsilyl chloride are mapped thrice to atoms in 1,2,3,4 *tert*-butyldimethylsilyl furanose. Panel B shows the same solution generated by CLCA after reactant copies of *tert*-butyldimethylsilyl were added.

by the ReactionMap Web interface (http://cdb.ics.uci.edu). For the remaining 21 reactions, CLCA suggested mappings with fewer bond changes than ReactionMap for 20 reactions and an incomplete mapping for phytoene synthetase. Figure 15 and Figure 16 show mapping solutions for acetyl-CoA acyltransferase by ReactionMap and CLCA, respectively. The reaction mechanism proposed by the CLCA mapping solution suggests a transfer of the hydroxybenzoyl group from hydroxybenzoyl-acetyl-CoA to CoA after the formation of a C−S bond to produce 4-hydroxybenzoyl-CoA and acetyl-CoA. Unlike MWED,

dx.doi.org/10.1021/ci5003922 | *J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX

wherein failure modes were specific to certain topological features, ReactionMap failures were primarily associated with reactions with greater than 200 atoms and reactions involving acyl-CoAs as the reactants or products. As shown in Figure 15, ReactionMap frequently suggested reaction centers on the acyl and neopentane moieties. The 21 reactions for which CLCA and ReactionMap provide differing atom mapping solutions are available as SMILES in Table S3 of the Supporting Information.

Applicability of the CLCA reaction mapping procedure is not limited to the MetRxn curated and mass balanced reaction database and can be extended to other large unbalanced reaction databases as well. Various efforts devoted to the reaction atom mapping of large reaction databases have been reviewed in recent articles by Ehrlich and Rarey[21] and Chen et al.[4] They compare various MCS and reaction mapping algorithms and identify specific shortcomings while mapping chemical structures with certain topological features. Following the conclusions presented in the aforementioned reviews, Kraut et al.[27] present a comprehensive data set of 104 reactions. The CLASSIFY test set, version 1.0, is categorized by difficulty into seven groups and is topologically representative of reactions in large commercial databases. We compare the solutions suggested by CLCA with the InfoChem ICMAP-generated atom mapping solutions. Prior to mapping, reactant and product copies were added to minimize the difference in the number of reactant and product atom in unbalanced reactions. There were disagreements for only five reactions with two in group G7, two in group G6, one in group G5. Figure 17 shows the two mapping solutions by ICMAP and CLCA for a hypoxanthine derivative synthesis reaction.[58] Figure 17A shows the correct mapping solution by ICMAP.[58] Figure 17B shows the incorrect mapping solution by CLCA, wherein the phenyl moiety of phenylazomalonamidamidine is mapped to the chlorophenyl moeity of 2,8-di(*p*-chlorophenyl)-hypoxanthine. This reaction was grouped under groups G6 by Kraut et al.[27] as the reaction information provided is incomplete and requires another copy of a reactant (i.e., a total of two *p*-chlorobenzaldehyde). An incomplete mapping was suggested by CLCA, as the reactant copy addition step did not suggest the addition of another *p*-chlorobenzaldehyde to the reaction. Figure 18B shows a correct mapping solution for a furanose syliation reaction[59] generated by CLCA after reactant copies were added. CLCA-suggested mappings for all the 104 test case reactions are available as reaction SMILES in Table S4 of the Supporting Information.

4. SUMMARY

We present a robust algorithm capable of identifying common substructures even for large molecular graphs. CLCA is novel and differs from all other algorithms in conceptual design and overcomes the previously mentioned drawbacks of existing algorithms. We show that CLCA is accurate even with large molecular graphs with complex topologies (rings and symmetric groups) and outperforms other algorithms in terms of computational complexity. The key operations performed in CLCA are the feature string sorting and prime number assignment step, product of primes step, and the Floyd–Warshall all pair shortest path calculation step. The sorting, prime assignment, and product of primes step has a convergence complexity similar to the canonical labeling algorithm presented by Weinenger et al.[2] of $\omega(n)$, and the Floyd–Warshall algorithm has a run time complexity of n.[3] Although CLCA terminates in deterministic time, it is important to note that CLCA does not guarantee an optimal MCS search solution for all molecular graphs. Particularly for nonplanar molecular graphs of molecules such as dodecahedrane, adamantane, or twistane, no known MCS search algorithm guarantees an optimal solution in deterministic polynomial time. Fortunately, most molecular graphs are planar (i.e., trees, outerplanar, etc.), and nonplanar molecular graphs seldom occur, enabling CLCA and many other MCS search algorithms to identify an optimal isomorph for most molecules in polynomial time. All the steps in CLCA are highly parallelizable, making it highly suitable for vectorization. CLCA has advantages over existing algorithms in performance and accuracy with the additional capability to handle unbalanced reactions, find multiple optimal mappings handle stereochemistry, and handle large complex structures with polynomial time computational complexity. CLCA is integrated within the MetRxn[1] database allowing rapid searches for similar molecules and biochemical transformations using Jaccard and Tanimoto similarity indices. Atom maps for all reactions are available online on MetRxn[1] for download. The java executable and usage directions will be made available online for download at www.metrxn.che.psu.edu. Marvin 6.3.1[60] by ChemAxon (http://www.chemaxon.com) was used for drawing, displaying, and characterizing chemical structures, substructures, and reactions.

■ ASSOCIATED CONTENT

Supporting Information

Figure S1 provides an example for a correct mapping solution when maximum atom/bond overlap is considered. Figure S2 provides 16 examples of failure modes of CLCA when compared with KEGG. Table S1 is a list of MetaCyc reaction IDs where the atom mapping solutions from CLCA disagree. Table S2 contains values of the longest graph geodesics for each vertex mentioned in Figure 5. Table S3 contains the disagreements between ReactionMap and CLCA solutions for 21 of the 1000 random chosen MetRxn reactions. Table S4 contains the CLCA solution for the 104 reactions in the CLASSIFY test set. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: costas@psu.edu.

Funding

This work was supported by the U.S. Department of Energy (DOE) at the Pennsylvania State University, University Park, under Grant DE-SC10822882. The Java API for minimum spanning tree can be accessed at http://jgrapht.org/. The Java API for A* search can be accessed at http://graphstream-project.org/.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge Saratram Gopalakrishnan for contributions during the manual verification of the atom mapping results for over 2000 reactions. We thank Anupam Chowdhury, Rajib Saha, and Ali Khodayari for their valuable inputs during the preparation of the manuscript. We thank Mario Latendresse from SRI International for sharing the MetaCyc reaction atom mappings for comparison and evaluation with CLCA atom mapping data. We also thank the reviewers for their invaluable comments and suggestions for improving the manuscript.

ABBREVIATIONS

CLCA, canonical labeling for clique approximation; MCS, maximum common substructure; MST, minimum spanning tree; EC, enzyme commission or extended connectivity algorithm; MWED, minimum weighted edit distance; GED, graph edit distance; FHV, feature hash vector

REFERENCES

(1) Kumar, A.; Suthers, P. F.; Maranas, C. D. MetRxn: A knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC Bioinf.* **2012**, *13*, 6.

(2) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Model.* **1989**, *29*, 97−101.

(3) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.

(4) Chen, W. L.; Chen, D. Z.; Taylor, K. T. Automatic reaction mapping and reaction center detection. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3*, 560−593.

(5) Jochum, C.; Gasteiger, J.; Ugi, I. The principle of minimum chemical distance (PMCD). *Angew. Chem., Int. Ed. Engl.* **1980**, *19*, 495−505.

(6) Faulon, J. J. Isomorphism, automorphism partitioning, and canonical labeling can be solved in polynomial-time for molecular graphs. *J. Chem. Inf. Comput. Sci.* **1998**, *2338*, 432−444.

(7) Garey, M. R.; Johnson, D. S. *Computers and Intractability: A Guide to the Theory of NP-Completeness* ; Series of Books in the Mathematical Sciences; W. H. Freeman: San Francisco, 1979; p 340.

(8) Apostolakis, J.; Sacher, O.; Körner, R.; Gasteiger, J. Automatic determination of reaction mappings and reaction center information. 2. Validation on a biochemical reaction database. *J. Chem. Inf. Model.* **2008**, *48*, 1190−1198.

(9) Latendresse, M.; Malerich, J. Accurate atom-mapping computation for biochemical reactions. *J. Chem. Inf. Model.* **2012**, *52*, 2970−2982.

(10) Kouri, T.; Mehta, D. *Experimental Algorithms*; Lecture Notes in Computer Science; Pardalos, P. M., Rebennack, S., Eds.; Springer: Berlin, Heidelberg, 2011; Vol. *6630*, pp 157−168.

(11) Heinonen, M.; Lappalainen, S.; Mielikäinen, T.; Rousu, J. Computing atom mappings for biochemical reactions without subgraph isomorphism. *J. Comput. Biol.* **2011**, *18*, 43−58.

(12) McGregor, J. J.; Willett, P. Use of a maximum common subgraph algorithm in the automatic identification of ostensible bond changes occurring in chemical reactions. *J. Chem. Inf. Model.* **1981**, *21*, 137−140.

(13) Lynch, M. F.; Willett, P. The automatic detection of chemical reaction sites. *J. Chem. Inf. Model.* **1978**, *18*, 154−159.

(14) Mann, M.; Nahar, F.; Ekker, H. *Principles and Practice of Constraint Programming*; Lecture Notes in Computer Science; Schulte, C.., Ed.; Springer: Berlin, Heidelberg, 2013; Vol. *8124*.

(15) First, E. L.; Gounaris, C. E.; Floudas, C. a. Stereochemically consistent reaction mapping and identification of multiple reaction mechanisms through integer linear optimization. *J. Chem. Inf. Model.* **2012**, *52*, 84−92.

(16) Fontain, E. The problem of atom-to-atom mapping. An application of genetic algorithms. *Anal. Chim. Acta* **1992**, *265*, 227−232.

(17) Morgan, H. L. The generation of a unique machine description for chemical structures − A technique developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107−113.

(18) Vléduts, G. É. Concerning one system of classification and codification of organic reactions. *Inf. Storage Retr.* **1963**, *1*, 117−146.

(19) Funatsu, K.; Endo, T.; Kotera, N.; Sasaki, S.-I. Automatic recognition of reaction site in organic chemical reactions. *Tetrahedron Comput. Methodol.* **1988**, *1*, 53−69.

(20) Barker, E. J.; Buttar, D.; Cosgrove, D. a; Gardiner, E. J.; Kitts, P.; Willett, P.; Gillet, V. J. Scaffold hopping using clique detection applied to reduced graphs. *J. Chem. Inf. Model.* **2006**, *46*, 503−511.

(21) Ehrlich, H.-C.; Rarey, M. Maximum common subgraph isomorphism algorithms and their applications in molecular science: A review. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 68−79.

(22) Caboche, S.; Pupin, M.; Leclère, V.; Jacques, P.; Kucherov, G. Structural pattern matching of nonribosomal peptides. *BMC Struct. Biol.* **2009**, *9*, 15.

(23) Caspi, R.; Altman, T.; Dreher, K.; Fulcher, C. a; Subhraveti, P.; Keseler, I. M.; Kothari, A.; Krummenacker, M.; Latendresse, M.; Mueller, L. a; Ong, Q.; Paley, S.; Pujar, A.; Shearer, A. G.; Travers, M.; Weerasinghe, D.; Zhang, P.; Karp, P. D. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **2012**, *40*, D742−53.

(24) Shimizu, Y.; Hattori, M.; Goto, S.; Kanehisa, M. Generalized reaction patterns for prediction of unknown enzymatic reactions. *Genome Inf. Ser.* **2008**, *20*, 149−158.

(25) Fooshee, D.; Andronico, A.; Baldi, P. ReactionMap: An efficient atom-mapping algorithm for chemical reactions. *J. Chem. Inf. Model.* **2013**, *53*, 2812−2819.

(26) Feist, A. M.; Henry, C. S.; Reed, J. L.; Krummenacker, M.; Joyce, A. R.; Karp, P. D.; Broadbelt, L. J.; Hatzimanikatis, V.; Palsson, B. Ø. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **2007**, *3*, 121.

(27) Kraut, H.; Eiblmaier, J.; Grethe, G.; Löw, P.; Matuszczyk, H.; Saller, H. Algorithm for reaction classification. *J. Chem. Inf. Model.* **2013**, *53*, 2884−2895.

(28) Corneil, D. G.; Gotlieb, C. C. An efficient algorithm for graph isomorphism. *J. Assoc. Comput. Mach.* **1970**, *17*, 51−64.

(29) Voet, D.; Voet, J.; Pratt, C. *Fundamentals of Biochemistry: Life at the Molecular Level*; Wiley: New York, 2006.

(30) Glasfeld, A.; Leanz, G. F.; Benner, S. A. The stereospecificities of seven dehydrogenases from *Acholeplasma laidlawii*. The simplest historical model that explains dehydrogenase stereospecificity. *J. Biol. Chem.* **1990**, *265*, 11692−11699.

(31) Moss, G. P. Basic terminology of stereochemistry (IUPAC Recommendations 1996). *Pure Appl. Chem.* **1996**, *68*, 2193−2222.

(32) Rose, I. A.; O'Connell, E. L. Mechanism of aconitase action. I. The hydrogen transfer reaction. *J. Biol. Chem.* **1967**, *242*, 1870−1879.

(33) Momany, C.; Levdikov, V.; Blagova, L.; Crews, K. Crystallization of diaminopimelate decarboxylase from *Escherichia coli*, a stereospecific D-amino-acid decarboxylase. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 549−552.

(34) Raymond, J. W.; Gardiner, E. J.; Willett, P. Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 305−316.

(35) Rahman, S. A.; Bashton, M.; Holliday, G. L.; Schrader, R.; Thornton, J. M. Small molecule subgraph detector (SMSD) toolkit. *J. Cheminf.* **2009**, *1*, 12.

(36) Crabtree, J. D.; Mehta, D. P.; Kouri, T. M. An open-source Java platform for automated reaction mapping. *J. Chem. Inf. Model.* **2010**, *50*, 1751−1756.

(37) Lessard, I. A. D.; Pratt, S. D.; McCafferty, D. G.; Bussiere, D. E.; Hutchins, C.; Wanner, B. L.; Katz, L.; Walsh, C. T. Homologs of the vancomycin resistance D-Ala-D-Ala dipeptidase VanX in *Streptomyces toyocaensis*, *Escherichia coli* and *Synechocystis*: Attributes of catalytic efficiency, stereoselectivity and regulation with implications for function. *Chem. Biol. (Oxford, U. K.)* **1998**, *5*, 489−504.

(38) Antoniewicz, M. R.; Kelleher, J. K.; Stephanopoulos, G. Elementary metabolite units (EMU): A novel framework for modeling isotopic distributions. *Metab. Eng.* **2007**, *9*, 68−86.

(39) Latino, D. a R. S.; Zhang, Q.-Y.; Aires-de-Sousa, J. Genome-scale classification of metabolic reactions and assignment of EC numbers with self-organizing maps. *Bioinformatics* **2008**, *24*, 2236−2244.

(40) Rahman, S. A.; Cuesta, S. M.; Furnham, N.; Holliday, G. L.; Thornton, J. M. EC-BLAST: A tool to automatically search and compare enzyme reactions. *Nat. Methods* **2014**, *11*, 171−174.

(41) Egelhofer, V.; Schomburg, I.; Schomburg, D. Automatic assignment of EC numbers. *PLoS Comput. Biol.* **2010**, *6*, e1000661.

(42) Yamanishi, Y.; Hattori, M.; Kotera, M.; Goto, S.; Kanehisa, M. E-Zyme: Predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs. *Bioinformatics* **2009**, *25*, i179−86.
(43) Sacher, O.; Reitz, M.; Gasteiger, J. Investigations of enzyme-catalyzed reactions based on physicochemical descriptors applied to hydrolases. *J. Chem. Inf. Model.* **2009**, *49*, 1525−1534.
(44) O'Boyle, N. M.; Holliday, G. L.; Almonacid, D. E.; Mitchell, J. B. O. Using reaction mechanism to measure enzyme similarity. *J. Mol. Biol.* **2007**, *368*, 1484−1499.
(45) Pinter, R. Y.; Rokhlenko, O.; Yeger-Lotem, E.; Ziv-Ukelson, M. Alignment of metabolic pathways. *Bioinformatics* **2005**, *21*, 3401−3408.
(46) Ay, F.; Kellis, M.; Kahveci, T. SubMAP: Aligning metabolic pathways with subnetwork mappings. *J. Comput. Biol.* **2011**, *18*, 219−235.
(47) Tipton, K.; Boyce, S. History of the enzyme nomenclature system. *Bioinformatics* **2000**, *16*, 34−40.
(48) Li, Y.; de Ridder, D.; de Groot, M. J. L.; Reinders, M. J. T. Metabolic pathway alignment between species using a comprehensive and flexible similarity measure. *BMC Syst. Biol.* **2008**, *2*, 111.
(49) Ay, F.; Dang, M.; Kahveci, T. Metabolic network alignment in large scale by network compression. *BMC Bioinf.* **2012**, *13* (Suppl 3), S2.
(50) Zhou, W.; Nakhleh, L. Quantifying and assessing the effect of chemical symmetry in metabolic pathways. *J. Chem. Inf. Model.* **2012**, *52*, 2684−2696.
(51) Scheer, M.; Grote, A.; Chang, A.; Schomburg, I.; Munaretto, C.; Rother, M.; Söhngen, C.; Stelzer, M.; Thiele, J.; Schomburg, D. BRENDA, the enzyme information system in 2011. *Nucleic Acids Res.* **2011**, *39*, D670−6.
(52) Ogata, H.; Goto, S.; Sato, K.; Fujibuchi, W.; Bono, H.; Kanehisa, M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **1999**, *27*, 29−34.
(53) Alcántara, R.; Axelsen, K. B.; Morgat, A.; Belda, E.; Coudert, E.; Bridge, A.; Cao, H.; de Matos, P.; Ennis, M.; Turner, S.; Owen, G.; Bougueleret, L.; Xenarios, I.; Steinbeck, C. Rhea—A manually curated resource of biochemical reactions. *Nucleic Acids Res.* **2012**, *40*, D754−60.
(54) Degtyarenko, K.; de Matos, P.; Ennis, M.; Hastings, J.; Zbinden, M.; McNaught, A.; Alcántara, R.; Darsow, M.; Guedj, M.; Ashburner, M. ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **2008**, *36*, D344−50.
(55) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M.-A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G. E.; Macinnis, G. D.; Weljie, A. M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L. HMDB: The human metabolome database. *Nucleic Acids Res.* **2007**, *35*, D521−6.
(56) Vastrik, I.; D'Eustachio, P.; Schmidt, E.; Joshi-Tope, G.; Gopinath, G.; Croft, D.; de Bono, B.; Gillespie, M.; Jassal, B.; Lewis, S.; Matthews, L.; Wu, G.; Birney, E.; Stein, L. Reactome: A knowledge base of biologic pathways and processes. *Genome Biol.* **2007**, *8*, R39.
(57) *OEChem TK*, version 1.7. 4.3; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2010.
(58) Yoneda, F.; Koga, R.; Higuchi, M. A One-step synthesis of purine derivatives by the reaction of phenylazomalonamidamidine with aryl aldehydes. *Chem. Lett.* **1982**, *4*, 365−368.
(59) Dureau, R.; Legentil, L.; Daniellou, R.; Ferrières, V. Two-step synthesis of per-O-acetylfuranoses: Optimization and rationalization. *J. Org. Chem.* **2012**, *77*, 1301−1307.
(60) *Marvin*, version 6.3.1. ChemAxon. http://www.chemaxon.com (accessed 2014).