Automated Atom Mapping of Biochemical Reactions

Heinz Ekker

Institute for Theoretical Chemistry
University of Vienna

mailto:hekker@tbi.univie.ac.at

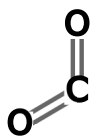
February 17, 2010

tbi

Outline

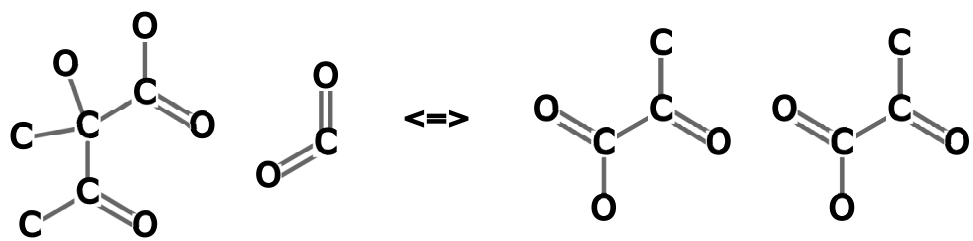
Reaction Mapping

1. Molecule graphs
2. Reactions as set of molecule graphs
3. Reaction as *graph rewrite rule*
4. Node *mapping* from substrates to products
5. Many different possibilities: We are looking for the atom mapping that requires the minimal editing distance!



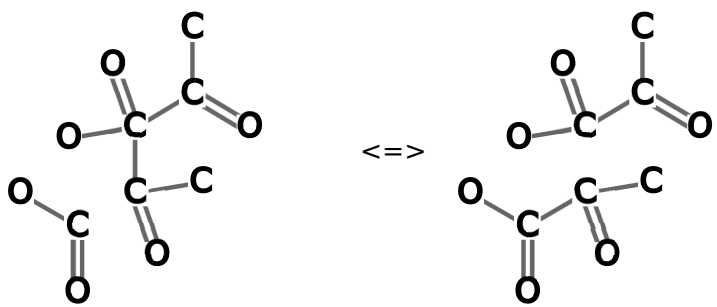
Reaction Mapping

1. Molecule graphs
2. Reactions as set of molecule graphs
3. Reaction as *graph rewrite rule*
4. Node *mapping* from substrates to products
5. Many different possibilities: We are looking for the atom mapping that requires the minimal editing distance!



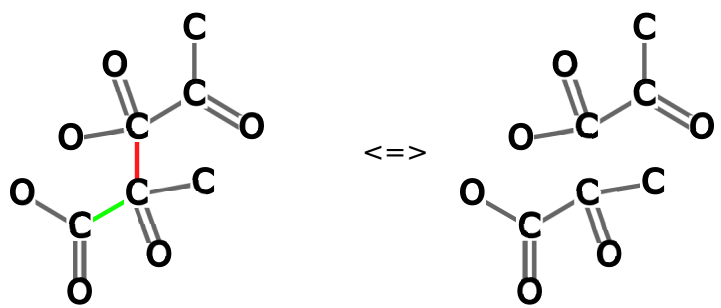
Reaction Mapping

1. Molecule graphs
2. Reactions as set of molecule graphs
3. Reaction as *graph rewrite rule*
4. Node *mapping* from substrates to products
5. Many different possibilities: We are looking for the atom mapping that requires the minimal editing distance!



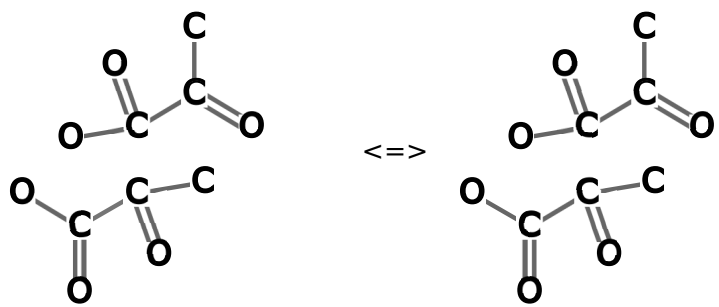
Reaction Mapping

1. Molecule graphs
2. Reactions as set of molecule graphs
3. Reaction as *graph rewrite rule*
4. Node *mapping* from substrates to products
5. Many different possibilities: We are looking for the atom mapping that requires the minimal editing distance!



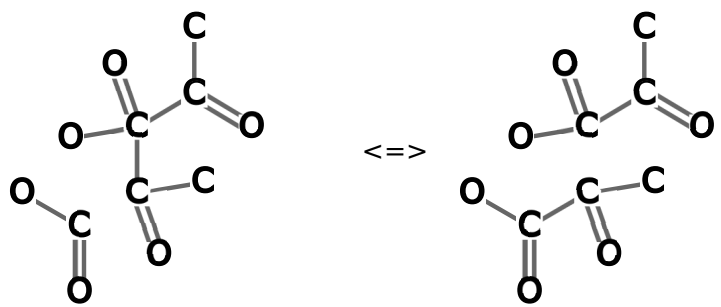
Reaction Mapping

1. Molecule graphs
2. Reactions as set of molecule graphs
3. Reaction as *graph rewrite rule*
4. Node *mapping* from substrates to products
5. Many different possibilities: We are looking for the atom mapping that requires the minimal editing distance!



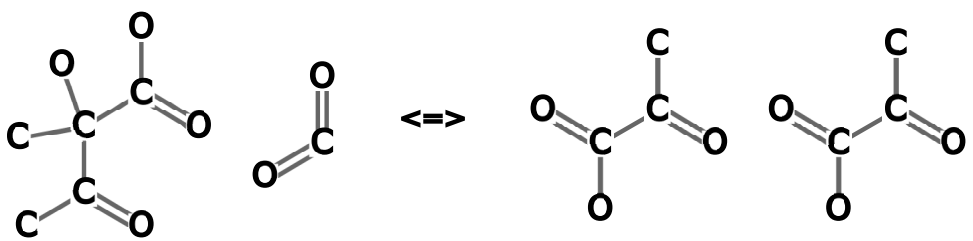
Reaction Mapping

1. Molecule graphs
2. Reactions as set of molecule graphs
3. Reaction as *graph rewrite rule*
4. Node *mapping* from substrates to products
5. Many different possibilities: We are looking for the atom mapping that requires the minimal editing distance!



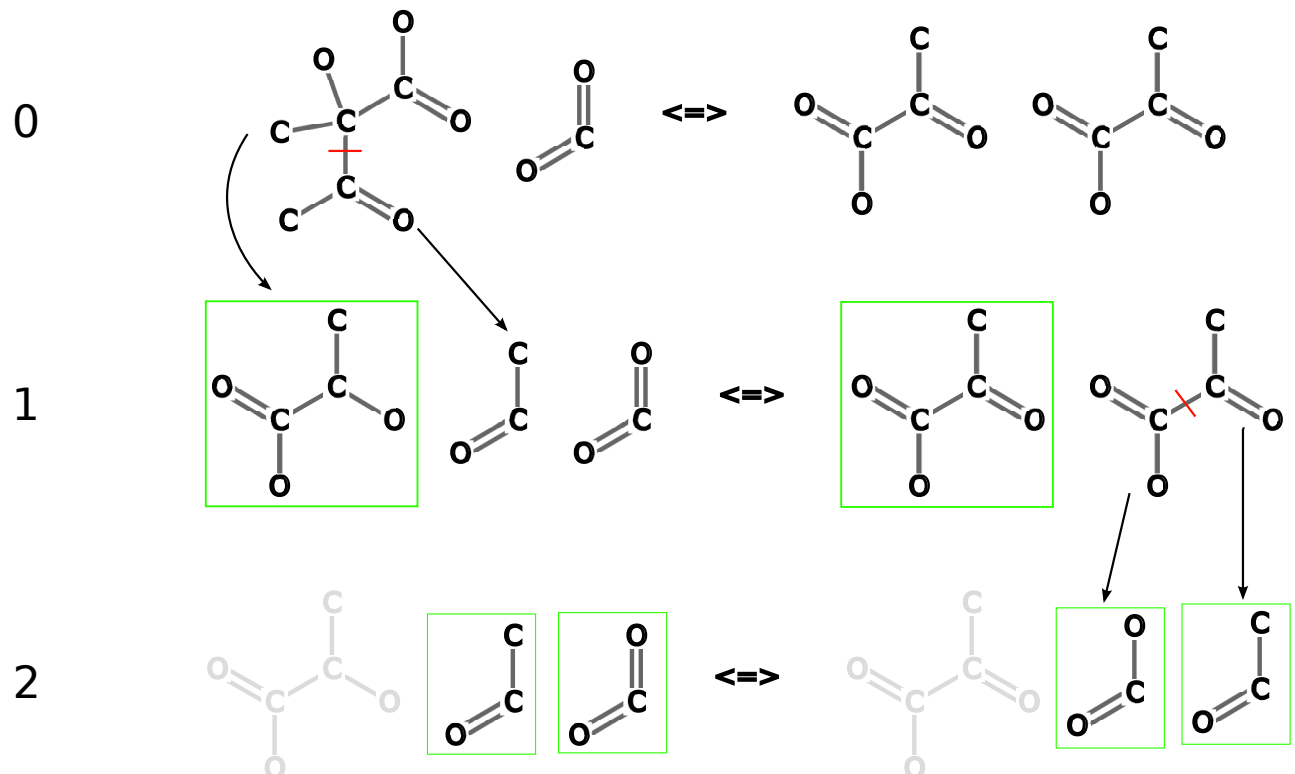
Reaction Mapping

1. Molecule graphs
2. Reactions as set of molecule graphs
3. Reaction as *graph rewrite rule*
4. Node *mapping* from substrates to products
5. Many different possibilities: We are looking for the atom mapping that requires the minimal editing distance!



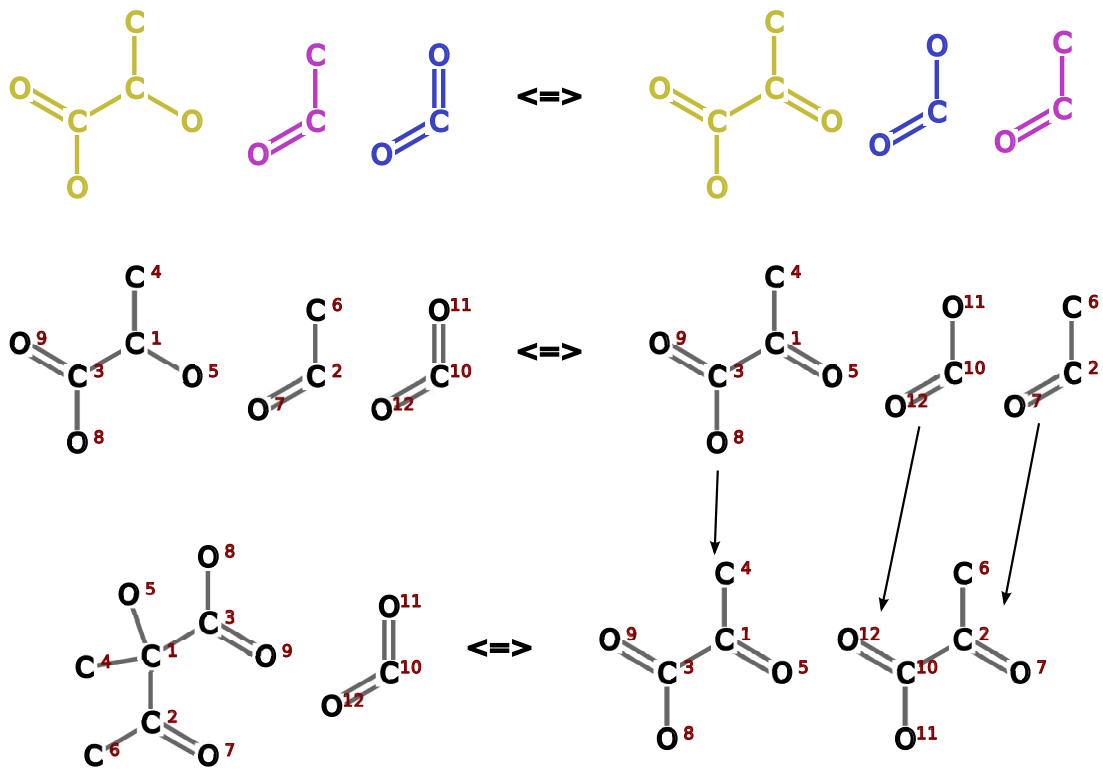
Cut Successive Largest Algorithm - Phase I

1. Take largest molecule of reaction
2. Break each bond in turn
3. Compare fragments to other molecules
4. If isomorphism, replace largest molecule with fragments
5. Repeat with 1



Cut Successive Largest Algorithm - Phase II

1. Label substrates
2. Map labeling to fragments
3. Relabel product fragments with isomorphic counterparts based on best mapping of bond orders
4. Map labeling to products



Bond Order, Aromatic Systems

Bond configuration in a multigraph

- single bond: one edge, double bond: two edges...
- Change in bond configuration just another edge removal/addition
- Problem size increases, could lead to invalid results

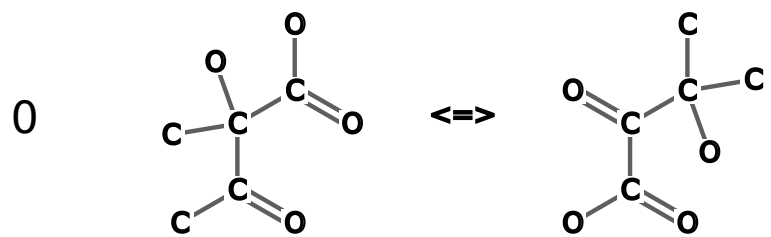
Bond configuration as edge labels

- Disregard during splitting
- Phase II: Choose best mapping from multiple possibilities based on bond order change

Isomerase Problem

What if two graphs have the same number and types of nodes, but are not isomorphic?
Intramolecular transfer of atoms or atom groups, structural rearrangements
A $\longrightarrow$ B, where A and B have the same sum formula

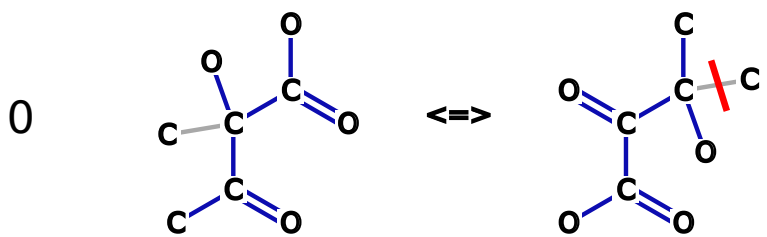
Use subgraph isomorphism, score fragments, keep best matches.



Isomerase Problem

What if two graphs have the same number and types of nodes, but are not isomorphic?
Intramolecular transfer of atoms or atom groups, structural rearrangements
A $\longrightarrow$ B, where A and B have the same sum formula

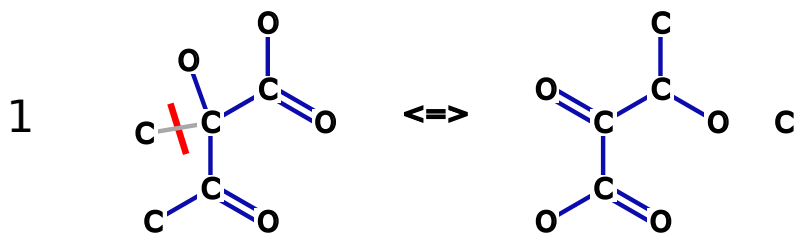
Use subgraph isomorphism, score fragments, keep best matches.



Isomerase Problem

What if two graphs have the same number and types of nodes, but are not isomorphic?
Intramolecular transfer of atoms or atom groups, structural rearrangements
A $\longrightarrow$ B, where A and B have the same sum formula

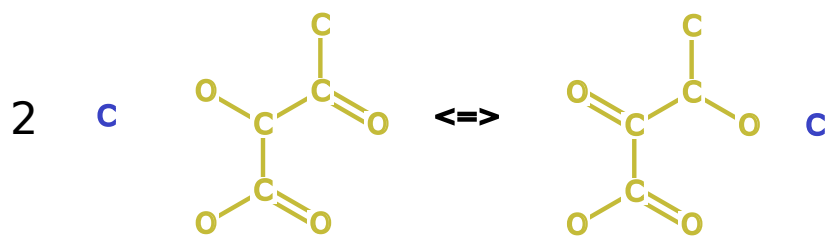
Use subgraph isomorphism, score fragments, keep best matches.



Isomerase Problem

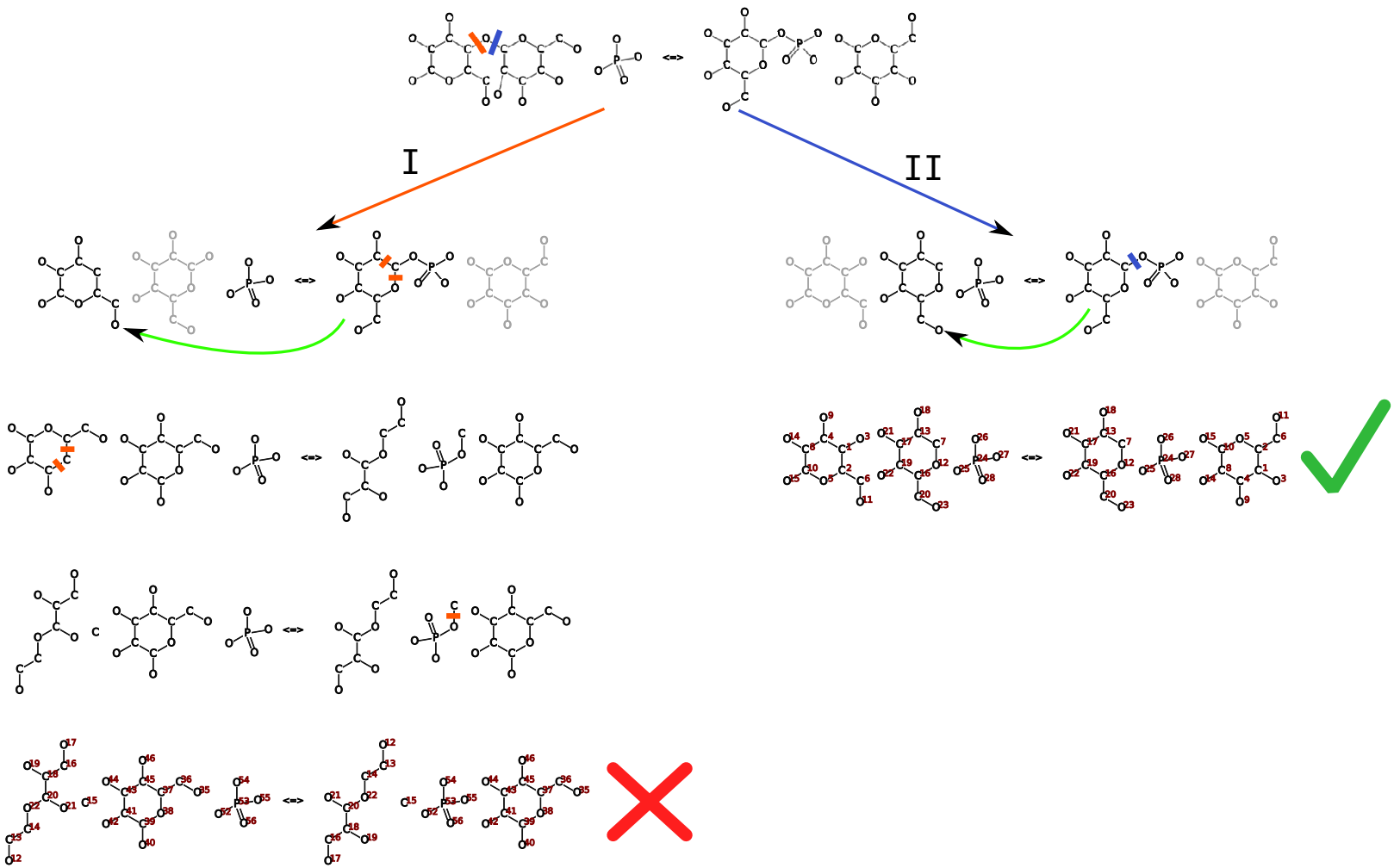
What if two graphs have the same number and types of nodes, but are not isomorphic?
Intramolecular transfer of atoms or atom groups, structural rearrangements
A $\longrightarrow$ B, where A and B have the same sum formula

Use subgraph isomorphism, score fragments, keep best matches.



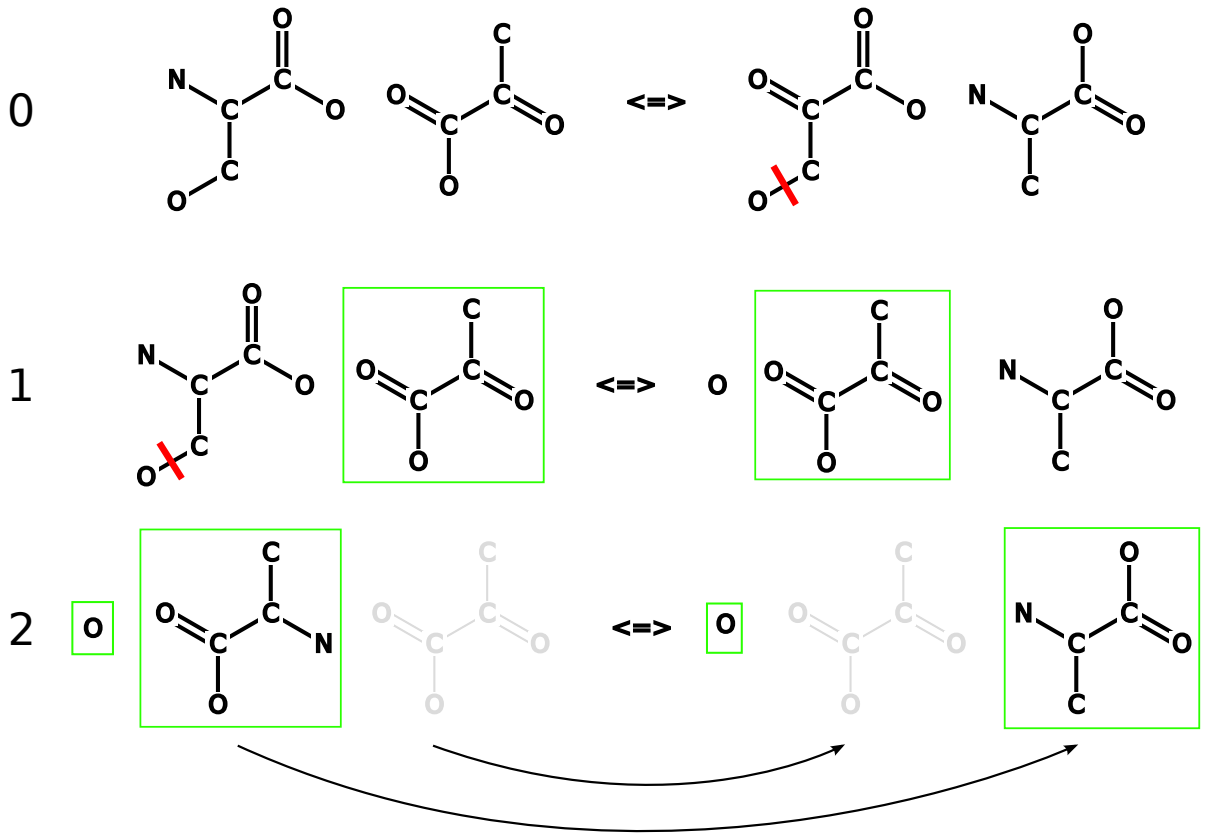
Split Selection

Evaluate all splits with same score: Some splits lead to dead ends or suboptimal mappings.



Graph Edit Distance and the Real World

Transaminase Reactions



Graph Edit Distance and the Real World

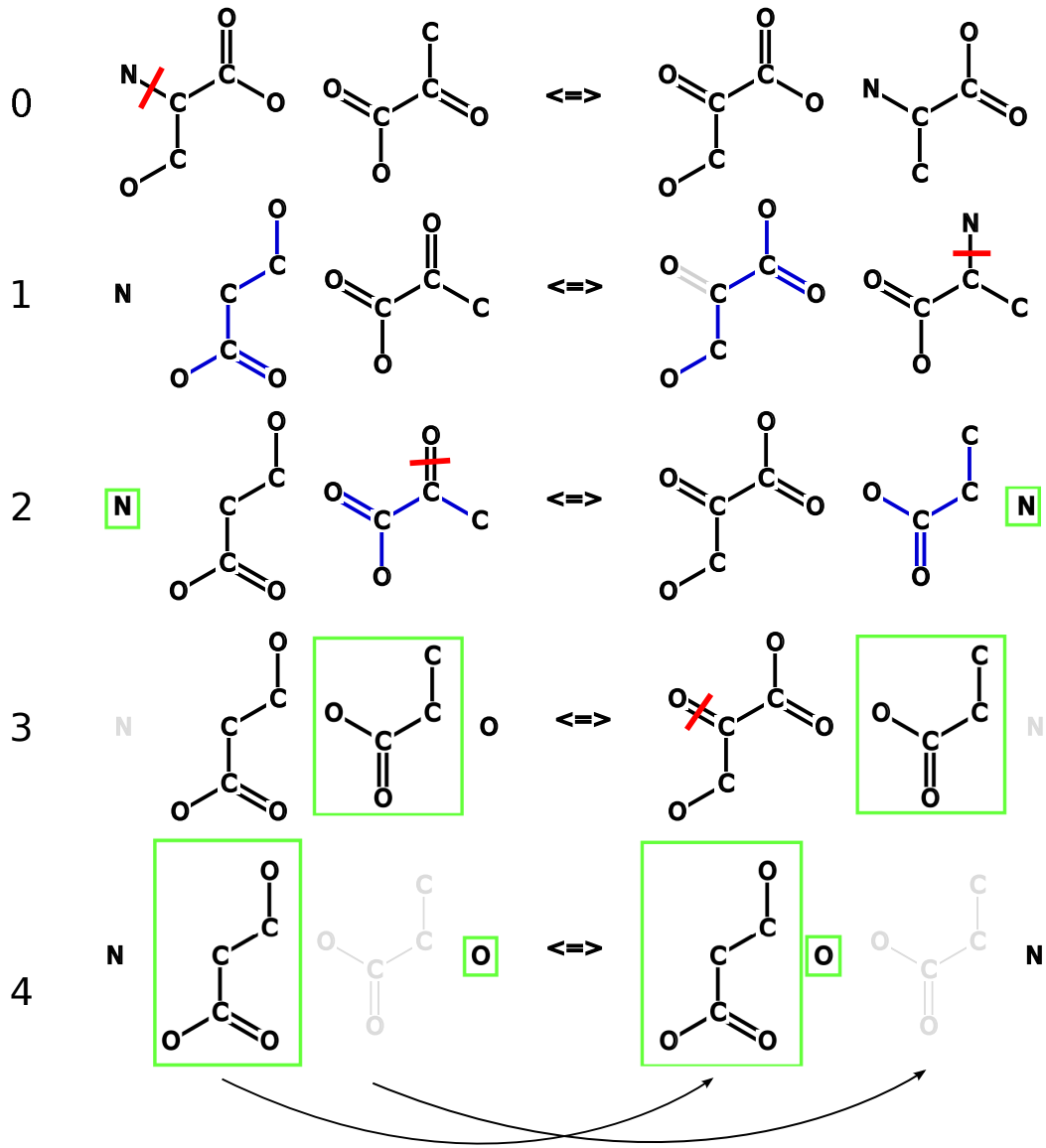
Transaminase Reactions

D-alanine transaminase (EC 2.6.1.21), taken from: MACiE Entry M0066

Step 1	The amine of the substrate L-glutamate attacks the PLP cofactor in a nucleophilic addition and the bound Lys145 deprotonates the newly attached amine.
Step 2	The secondary amine that results from the initial attack initiates an elimination of the covalently bound lysine, resulting in free PLP and lysine in a neutral state.
Step 3	Lys145 deprotonates the CH adjacent to the bound amine, resulting in double bond rearrangement as the PLP acts as an electron sink.
Step 4	The PLP feeds the electrons back, resulting in the C=C attached to the aromatic ring deprotonates Lys145.
Step 5	Lys145 deprotonates water, which initiates a nucleophilic attack on the carbon of the C=N group in an addition reaction.
Step 6	The secondary amine deprotonates the attached hydroxyl group, initiating an elimination which releases 2-oxoglutarate.
Step 7	The amine of PMP initiates a nucleophilic attack on the carbonyl carbon of pyruvate. The oxyanion deprotonates the newly formed secondary amine in the first step of a Schiff base formation.
Step 8	The secondary amine initiates an elimination, forming the Schiff base and releasing water with concomitant deprotonation of Lys145.
Step 9	Lys145 deprotonates the CH2 adjacent to the nitrogen, resulting in double bond rearrangement as the PLP acts as an electron sink.
Step 10	The PLP feeds the electrons back, the N+=C bond deprotonates Lys145.
Step 11	The amine of Lys145 attacks the PLP in a nucleophilic addition reaction, the secondary amine of the attached substrate reprotonates from the bound Lys145.
Step 12	The secondary amine that results from the initial attack initiates an elimination of the covalently bound product, resulting in alanine and the regenerated PLP cofactor.

Graph Edit Distance and the Real World

Transaminase Reactions



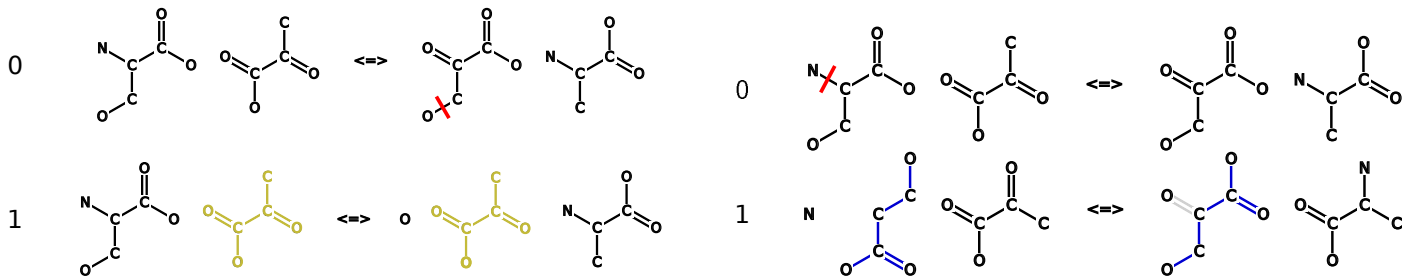
Bond Weighting

based on a crude estimation of bond energies

- Preprocessing: Assign weight to each edge
  Based on:
  1. Adjacent atoms from global table
  2. Expand neighborhood by one, search resulting fragment in predefined list
  3. Adjacent atoms from fragment-specific weight table
  4. Expand by one ...
- Score calculation:
  1. Score bond-bond mapping?
  2. Score one fragment?
  3. Score cost of bond breaking?
- Splitting: $score = \frac{score}{weight\ \ of\ \ broken\ \ bond}$
- Re-evaluate fragments?

Bond Weighting - Effects

- Push splitting into the right direction



- But: Impossible to find weighting scheme for *all* enzymes
- Dramatic overall performance improvement: Better bounding - fewer splits evaluated

Verification

- KEGG Ligand database: Composite database consisting of Compounds, Glycans, Reactions, Enzymes and RPAIRs.
- RPAIR: Chemical structure transformation patterns for substrate-product pairs

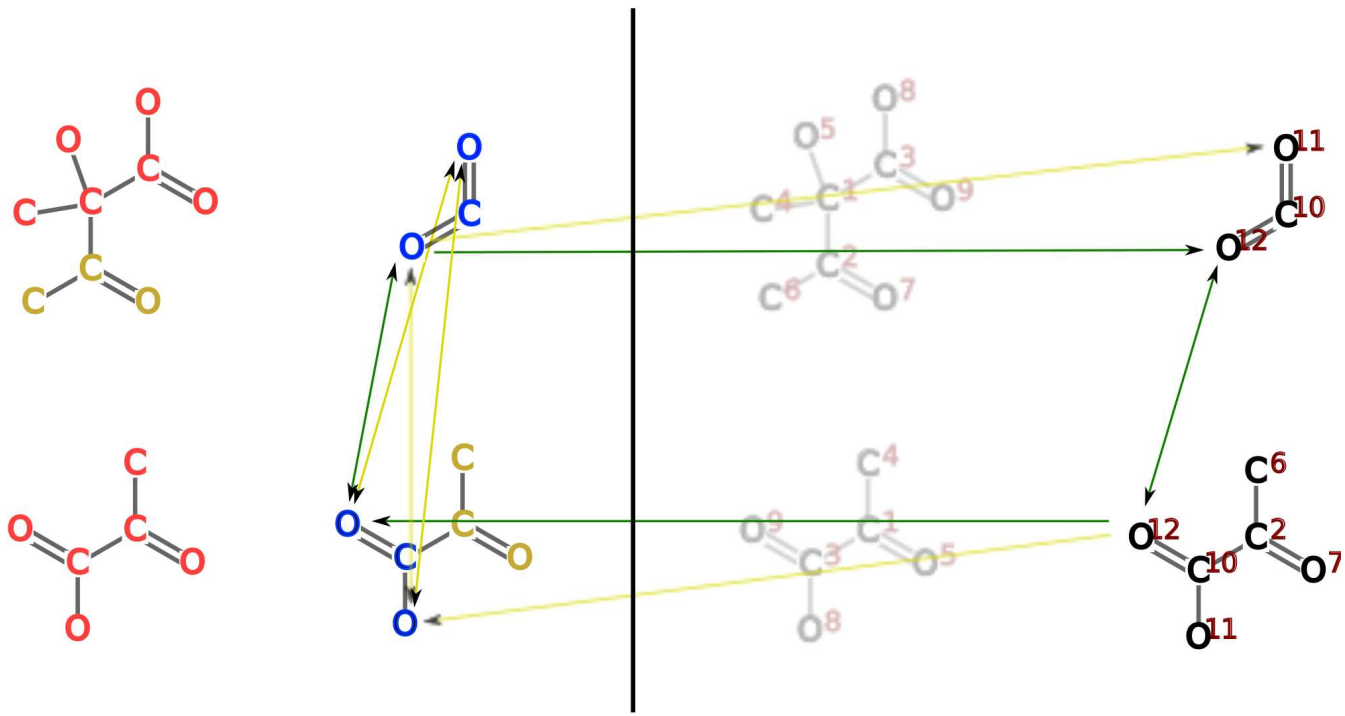
Verification steps:

1. Fetch reaction, parse components, fetch compounds, transform to graph
2. Map reaction, apply rule to substrates, check result
3. Fetch RPAIRs, map RPAIRs to compound graph representation
4. Compare our atom mapping with mapped RPAIRs

Verification - Example

2-acetolactate pyruvate-lyase

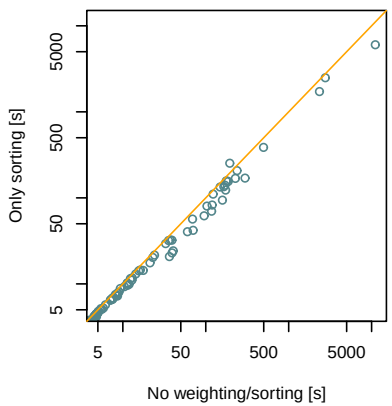
- Mapped by 3 RPAIRs
- Acetolactate mapped by 2 RPAIR entries
- Molarity implied, oxygens of carboxyl groups interchangeable, mapping not unique

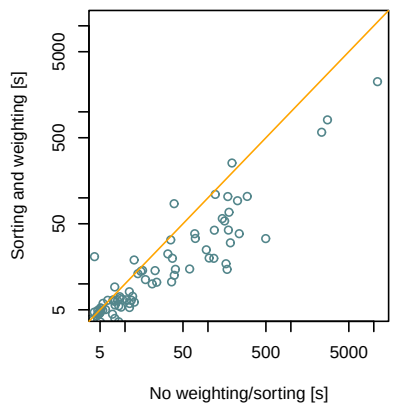


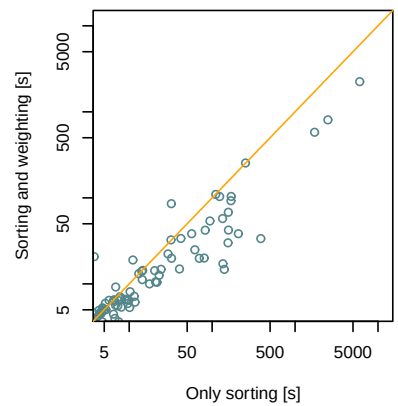
Preliminary Results

Preliminary Results			
	Reactions	Applicable	Bonds median/3rd quartile
KEGG Overall	7393	6604	86/124
Sample	600	447	98/130

	Mapped	Verified OK	total/mean/median/3rd qu [s]
w/o weighting and no bond sorting	447	383	28073/62.8/2.1/7.4
with weighting and bond sorting	447	390	9781/21.8/1.9/6.0







Program improvements

- Parallelization
  at bond or split level
- Subgraph isomorphism algorithm
  Ullmann: Problems with highly symmetric structures
  Subgraph check in C/C++-library
- Evaluate all atom mapping possibilities
- Perl bindings for toychem library
- Bells and Whistles
  More sophisticated web interface, better visualizations
- Integration into PerlMol (CPAN)
- Improve RPAIR check - cycle detection?

References

[1] G. Benkö, C. Flamm, and P. F. Stadler.
A Graph-Based Toy Model of Chemistry.
J. Chem. Inf. Comput. Sci., 43:1085–1093, October 2003.

[2] J. D. Crabtree and D. P. Mehta.
Automated Reaction Mapping.
Journal of Experimental Algorithmics, 13:1.15—-1.29, October 2009.

[3] G. Holliday, D. Almonacid, G. Bartlett, N. O'Boyle, J. Torrance, P. Murray-Rust, J. Mitchell, and J. Thornton.
MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms.
Nucleic Acids Research, 35(Database issue):D515, 2007.

[4] C. Jochum, J. Gasteiger, and I. Ugi.
The Principle of Minimum Chemical Distance (PMCD).
Angewandte Chemie International Edition in English, 19:495–505, January 1980.

[5] M. Kanehisa and S. Goto.
KEGG: Kyoto encyclopedia of genes and genomes.
Nucleic acids research, 28(1):27, 2000.

[6] M. Kotera, M. Hattori, M. Oh, R. Yamamoto, T. Komeno, J. Yabuzaki, K. Tonomura, S. Goto, and M. Kanehisa.
RPAIR: a reactant-pair database representing chemical changes in enzymatic reactions.
Genome Inform, 15:P062, 2004.

[7] F. Rossello and G. Valiente.
Chemical Graphs, Chemical Reaction Graphs, and Chemical Graph Transformation.
Electronic Notes in Theoretical Computer Science, 127(1):157–166, March 2005.

[8] J. R. Ullmann.
An Algorithm for Subgraph Isomorphism.
Journal of the Association for Computing Machinery, 23:31–42, April 1976.