Published in final edited form as:

Environ Sci Technol. 2012 June 5; 46(11): 6048–6055. doi:10.1021/es3003734.

COBRA: A Computational Brewing Application for Predicting the Molecular Composition of Organic Aerosols

David R. Fooshee^{1,†}, Tran B. Nguyen^{2,†}, Sergey A. Nizkorodov^{2,*}, Julia Laskin³, Alexander Laskin⁴, and Pierre Baldi^{1,*}

¹School of Information and Computer Sciences, University of California, Irvine, CA 92697-3435

²Department of Chemistry, University of California, Irvine, CA 92697-2025

³Chemical and Materials Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, USA

⁴Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington 99352, USA

Abstract

Atmospheric organic aerosols (OA) represent a significant fraction of airborne particulate matter and can impact climate, visibility, and human health. These mixtures are difficult to characterize experimentally due to their complex and dynamic chemical composition. We introduce a novel Computational Brewing Application (COBRA) and apply it to modeling oligomerization chemistry stemming from condensation and addition reactions in OA formed by photooxidation of isoprene. COBRA uses two lists as input: a list of chemical structures comprising the molecular starting pool, and a list of rules defining potential reactions between molecules. Reactions are performed iteratively, with products of all previous iterations serving as reactants for the next. The simulation generated thousands of structures in the mass range of 120–500 Da, and correctly predicted ~70% of the individual OA constituents observed by high-resolution mass spectrometry. Select predicted structures were confirmed with tandem mass spectrometry. Esterification was shown to play the most significant role in oligomer formation, with hemiacetal formation less important, and aldol condensation insignificant. COBRA is not limited to atmospheric aerosol chemistry; it should be applicable to the prediction of reaction products in other complex mixtures for which reasonable reaction mechanisms and seed molecules can be supplied by experimental or theoretical methods.

Keywords

Atmospheric aerosols; Chemoinformatics; High resolution mass spectrometry

*Corresponding authors: Pierre Baldi, pfbaldi@ics.uci.edu, phone: (949) 824-5809. Sergey Nizkorodov, nizkorod@uci.edu, phone: (949) 824-1262.

† These authors contributed equally to this project.

5. Supporting Information Available

The following supporting information is available for this article:

Supporting Table 1:

Isoprene high-NO_x SOA high-resolution negative ion mode ESI-MS data, compared with molecular formulas predicted by COBRA. This information is available free of charge via the Internet at <http://pubs.acs.org>

1. Introduction

Atmospheric organic aerosols (OA), airborne particles comprised primarily of organic material, are estimated to contribute up to 50% of the total particulate matter mass at continental mid-latitudes and up to 90% at forested areas (1). OA impact climate and visibility by interacting with solar radiation, atmospheric oxidants, and water vapor (2), and are associated with adverse effects on human health as discussed in Ref. (3), and references therein. The climate, visibility, and health effects of OA are dependent on their chemical composition (4–6). However, the diversity of OA formation and growth mechanisms makes detailed molecular chemical composition of OA difficult to predict by traditional modeling or characterize by experimental methods. During their residence time in the atmosphere, OA undergo chemical aging processes (7, 8) that further enhance molecular complexity through, for example, the formation of nitrogen-containing organic compounds (NOC) (9–11).

Recent advances in high-resolution mass spectrometry (HR-MS) have enabled simultaneous detection of hundreds of individual molecules in OA (9, 12–15). HR-MS tools are useful in providing the molecular formulas for OA constituents. Ion fragmentation patterns observed in tandem mass spectrometry (MS^n) experiments provide additional information about the structures. However, interpretation of MS^n data is complicated by the presence of multiple structural isomers and lack of sufficiently diagnostic fragmentation patterns. In previous work, reliable structural information could only be extracted for the low molecular weight (MW) species (16, 17) and homologous oligomers (18–20) present in OA, leaving the structures of the majority of complex oligomers uncharacterized.

HR-MS experiments suggest that in many cases, oligomeric compounds in atmospheric OA are produced from repetitive reactions between the OA constituents (17, 20–22). There is strong evidence that condensation reactions such as esterification and aldol condensation (18, 20, 23–28), and addition reactions such as hemiacetal formation (29–31) are quite common in both organic aerosols and in aqueous solutions of OA. Advanced computational approaches are needed for predicting the OA composition using a bottom-up approach, in which the starting low-molecular weight compounds are known from experiments, and reaction rules for combining these compounds into oligomers can be unambiguously defined.

This work describes the first application of a Computational Brewing Application (COBRA) to modeling oligomerization products observed in the detailed composition of OA. Figure 1 shows a diagrammatic representation of COBRA, which is a customizable simulation engine for chemical reactions. Prior work from our group related to predicting the mechanisms of organic chemical reactions has focused on a rule-based approach with broad knowledge of general organic chemistry (32), or a machine learning approach to reaction prediction based on ranking mechanistic steps by productivity (33). COBRA differs from these approaches in that it is optimized to simulate the evolution of a complex mixture by combinatorial computation of many thousands of chemical reactions between mixture constituents, based on a chosen pool of starting (seed) compounds and specific reaction mechanisms. For the case study presented in this work, the formation of a large set of experimentally-observed high-MW oligomeric products in OA derived from the photooxidation of isoprene (20, 22) is modeled by considering chemical transformations of a basic set of monomers through the following oligomerization reactions: esterification, aldol condensation, and hemiacetal formation. We demonstrate that COBRA succeeds at modeling relevant oligomerization chemistry, predicting and visualizing high-MW compound structures, and predicting unique reaction products in OA. More broadly, the COBRA approach is applicable to studying the evolution of a wide range of organic mixtures as long as the starting components and reaction mechanisms can be suggested.

2. Methods

2.1 COBRA

Chemical structures are input into COBRA using the widely-used chemical string representation SMILES (34). Reaction transforms are defined using the reaction transform language SMIRKS (35). The SMIRKS language is a superset of SMILES, and can be used to define reaction transforms to an arbitrary degree of specificity. We leverage the programming language Python in conjunction with OpenEye's OEChem library (36) to process the input SMILES and SMIRKS into predicted products. Standard valence rules are explicitly programmed into the reaction transforms in order to predict chemically meaningful structures. Specifically, R-groups in the reactions shown in Fig. 1 were allowed to be alkyl or acyl groups, but not hydrogen alone. Filters can be imposed to prevent compounds with certain properties from participating in reactions, or from being included in the pool of final products.

The simulation includes the following iterative steps. 1) For each molecule and each pair of molecules in the reactant pool, all unimolecular reaction transforms and all bimolecular transforms, respectively, are applied. Identical reactions already performed in the previous iterations are recognized and skipped. 2) The product molecules are filtered. For this simulation, we automatically filter out any molecule with greater than 40 heavy (C, O, N) atoms. This filter restricts the molecular weights of the products below approximately 500 Da, the limit at which molecular assignments made with 0.1 mDa accuracy can be uniquely determined for $C_cH_hO_oN_nS_s$ species (37). This restriction significantly decreased the model run time while still capturing the chemistry of the most abundant compounds in isoprene high-NO_x SOA (22). 3) Molecules that were not filtered out during the previous step are added back into the reactant pool, and the system returns to step one. The simulation is complete either after the requested number of iterations has passed, or after a specified number of unique reactions has been simulated (30,000 reactions for the full simulation and the exclusion simulations, in this study). Results can then be conveniently visualized and searched for specific chemical structures matching any specified structural criteria. Additionally, we can compute the sequence of reactions that led to a given product's formation.

2.2 Experimental methods

Isoprene OA was photochemically generated in a 5 m³ Teflon chamber, as described previously (20). Samples were generated in dry air, under high-NO_x (VOC:NO_x <1) conditions, and in the absence of inorganic seed particles. H₂O₂ was used as an OH precursor. The initial mixing ratio of isoprene was 250 ppbv (parts per billion by volume). Blank samples were produced in an identical manner as OA samples, without UV photooxidation. Samples were collected on Teflon filters (0.2 μm pore size, Millipore), vacuum sealed, and frozen prior to analysis. Negative ion mode direct injection ESI (tip voltage 4kV) was used as an ionization method for solvent-extracted OA samples. The solvents used for analysis were water and acetonitrile (both HPLC grade, Fluka) at a 1:1 volume ratio. Mass analysis was done with a high-resolution linear ion trap (LTQ-) OrbitrapTM (Thermo Corp.) at Pacific Northwest National Lab (PNNL) Environmental Molecular Science Laboratory facility (EMSL), with a mass resolution of 60,000 m/Δm at *m/z* 400. MSⁿ studies were performed in the LTQ, with mass selection in the 0.5 *m/z* range and collision-induced-dissociation energies of 20–40 energy units. Product ions of MSⁿ were analyzed in the Orbitrap.

3. Results and Discussion

Table 1 lists the seed molecules for the simulation. These low-MW species have been identified as important building blocks in the formation of isoprene photooxidation OA (22). The majority of these molecules have been detected in isoprene OA (Ref. (38), and references therein), and are primarily multifunctional carbonyl, alcohol, and carboxyl compounds derived from the oxidation of isoprene with the hydroxyl radical (OH) (18, 22, 39–46). Some of these molecules have sufficiently high vapor pressure to exist primarily in the gas phase. For example, glycolaldehyde has a room temperature vapor pressure of 0.028 Torr (47), and may participate in aerosol mass-growth reactions initiated in the gas- or heterogeneous phase. Note that glyoxal is not included among the seed molecules used in these simulations because it was not found among the monomer units inferred from mass spectrometry of isoprene oxidation products as described in Ref. (22). The 3-nitrate ester of 2-methylglyceric acid (2MGA) (18, 20, 27, 45) represents the sole NOC in the list of seed molecules. Available information about NOC monomers is limited due to the small number of studies of NOC composition in the literature (Ref. (22) and references therein).

COBRA was used to simulate the evolution of a virtual OA mixture by applying the four reaction transforms shown in Figure 1 to the set of 27 starting molecules in Table 1. The resulting product pool contained 135,107 predicted structures after performing 30,000 unique reactions, representing a total of 758 unique elemental formulas of the type $C_cH_hO_oN_n$ with $MW < 500$ Da. The two orders of magnitude difference between the number of structures and elemental formulas reflects a large number of predicted structural isomers. The experimental HR-MS data from Ref. (22) (Table S1 in the Supporting Information section) contains 464 neutral elemental formulas, and 323 of these formulas (70%) were predicted by COBRA. Figure 2 shows the experimental HR-MS spectrum separated into two panels. The first panel shows all HR-MS peaks in the experimental spectrum that were predicted by COBRA (Fig. 2A), and the second one shows peaks that did not appear in the COBRA output (Fig. 2B). As Figure 2 demonstrates, the 70% fraction of the experimentally observed peaks predicted by COBRA represents the most abundant compounds with $MW < 500$ Da. The remaining 30% fraction has average peak intensities that are more than an order of magnitude smaller compared to the average peak intensities associated with the successfully predicted compounds, suggesting that COBRA captures the essential chemistry producing oligomers in isoprene high-NO_x OA. The high degree of overlap between the predicted and observed formulas confirms that the oxygenated hydrocarbon seed molecules proposed in our previous work are the relevant oligomer building blocks in this type of OA (22). The fact that a relatively small set of modeled reactions was sufficient to account for roughly 70% of the HR-MS peaks suggests that the oligomerization chemistry in isoprene OA is well constrained.

The fraction of the experimentally observed compounds that are correctly predicted by the simulation can be a misleading metric of success in some cases. For example, if the simulation were significantly underdefined, and generated every possible $C_cH_hO_oN_n$ formula ($\sim 10^5$) allowed by the valence rules, this fraction would become 100%. Therefore the reverse comparison of the fraction of predicted peaks that show up in the experiment is just as important. In the present case, 43% of predicted formulas correspond to compounds detected by HR-MS, and the remaining 57% are not observed. This level of agreement can be viewed as good, considering that the simulation does not use any kinetics restrictions and predicts the formation of compounds that may be below the limit of detection of HR-MS.

It is remarkable that a total of 62 NOC molecular formulas were predicted stemming from 2MGA-3-nitrate alone, representing 41% of the total NOC observed in the isoprene OA sample in the m/z 120–500 range (Fig. 2A). Note that this particular simulation's ability to

predict NOC compounds is limited because only a single NOC seed molecule is included. Nonetheless, the data confirm that 2MGA-3-nitrate is a prolific oligomer building block in isoprene OA that produces a variety of products through oligomerization in the condensed phase (18, 20, 27, 45).

Since structural complexity increases with molecular size, the number of structural and stereo isomers that can be produced by oligomerization reactions also grows exponentially with increasing mass of the individual products (Fig. 2C). For the 323 experimentally observed molecular formulas, COBRA predicts 102,650 unique structures, with number of isomers ranging from 1 to nearly 3,000. The apparent decrease in the number of hits as MW approaches 500 Da is a result of filtering; we artificially limit structures in the product pool to those with less than 40 heavy atoms, and do not consider structures with molecular weights in excess of 500 Da. Despite the large number of predicted isomers, the simulation results help constrain possible structures, especially for lower-MW compounds. For example, over 1,000 unique molecular structures associated with a molecular mass of 142.063 Da, assigned to $C_7H_{10}O_3$, can be retrieved from internet-based chemical inventories (e.g., SciFinder). In contrast, COBRA predicts only two structures of $C_7H_{10}O_3$, thereby dramatically reducing the complexity (Table 2). This represents more than a one hundred-fold increase in the level of confidence in assigning molecular structures to a given formula obtained by HR-MS. Even at high masses, COBRA predictions remain useful because the structural information for high-MW compounds is not readily available from chemical databases, and this type of modeling is a practical step toward understanding the possible structures of large oligomers.

A few published studies used MS^n fragmentation data to determine structures of the oligomers in isoprene high- NO_x SOA (18, 19, 22). All molecular structures described in the previous studies match structures predicted by COBRA. To further test the validity of the COBRA predictions, we made a comparison to MS^n studies previously performed by our group of NOC oligomers observed in SOA generated from isoprene photooxidation (22). Figure 3 shows the predicted structures of two NOC oligomers, $C_8H_{13}NO_9$ (267.059 Da) and $C_{14}H_{20}NO_{13}$ (411.101 Da), that are most consistent with MS^n experiments. The MS^n experiments were performed using the negative ion mode; therefore, the molecular masses of the detected compounds are those of deprotonated molecules. Oligomer fragmentation is often characteristic of the functional groups and monomer units present within the compound. For example, HNO_3 and CH_3NO_3 are characteristic MS^n losses from organic nitrates, and $C_4H_6O_3$ is a characteristic loss from the carbonyl ester unit of 2MGA (18, 22).

For a given molecular formula, the proposed structures based on MS^n experiments described in Ref. (22) match at least one of the structures predicted by COBRA. We note that structural isomers may not have distinct fragmentation patterns, and therefore it is possible that the experimental MS^n data represents several structural isomers predicted by COBRA. COBRA predicted four isomeric structures for $C_8H_{13}NO_9$, and at least one predicted structure is consistent with the fragmentation pattern produced by MS^n studies. For $C_{14}H_{21}NO_{13}$, at least one out of 44 predicted structures is consistent with the MS^n fragmentation pattern. Therefore, it is reasonable to hypothesize that other structures predicted by COBRA may be present in the OA, perhaps but not necessarily, in lower quantity than the most abundant isomer.

To elucidate which seed molecules were the most important as oligomer building blocks for isoprene OA, we performed 27 “exclusion” simulations, each with a different seed molecule omitted. Each exclusion simulation was limited to the same number of reactions (30,000). The total number of experimental molecular formulas recovered in each case was compared against a simulation in which no seed molecules were omitted. The “exclusion percent”

column in Table 1, calculated as a percent reduction in the number of predicted formulas matching experimental formulas, can be viewed as a measure of a compound's relative importance in generating structures with experimentally observed molecular formulas. The highest exclusion percent (19%) was for the 3-nitrate ester of 2-methylglyceric acid, consistent with its special role of being the sole NOC precursor. Specifically, the removal of 2MGA-3-nitrate eliminated all NOC oligomers from the product pool. Removal of other seed molecules resulted in smaller variation in the number of predicted formulas, ranging from -6 to 6%. Note that because of the fixed number of reactions in these simulations, an exclusion of a less significant contributor to chemistry can actually lead to a slight increase in the total number of predicted formulas, yielding a negative exclusion percent, as happened with 2-hydroxy-2-Me-4-oxobutanoic acid. None of the nitrogen-free compounds appeared to stand out from the rest, suggesting some redundancy in the pool of seed molecules. The redundancy in seed molecules lowered the percent overlap between simulation and experiment. These results highlight the need for future work in determining missing precursors and reaction mechanisms. In general, this type of exclusion analysis is useful for assessing the contribution of a single molecule to the total product pool, e.g. the importance of glyoxal in heterogeneous reaction with various OA constituents (48–51), or the contribution of first vs. second generation VOC oxidation products to producing condensable organic compounds.

We performed additional exclusion simulations, in which one of the four reactions shown in Fig. 1 was disabled. The results of the simulations based on the three remaining reactions were compared to the results of the full simulation. Table 3 lists the percent reduction in the number of experimentally observed molecular formulas recovered by the simulation, the percent reduction in the number of theoretically predicted molecular formulas, and the percent reduction in the number of theoretically predicted structures generated by the simulation. The esterification reaction (Reaction 1) played the most significant role in recovering experimentally observed formulas, with an exclusion percent of 40.9%, followed by the hemiacetal formation reaction (Reaction 4) with an exclusion percent of 10.2%. The esterification and hemiacetal formation reactions also made the largest contributions to the total number of theoretically predicted structures. Exclusion of esterification from the simulation reduced the number of structures in the COBRA output by 37.8%, while exclusion of hemiacetal formation resulted in a 46.7% decrease. Although hemiacetal formation is prolific at generating theoretical structures, the results demonstrate that esterification is the most important reaction in promoting oligomerization in isoprene OA, while hemiacetal formation is of secondary importance. These results are consistent with a previous study, wherein reducing aerosol liquid water content changed the composition of isoprene high-NO_x SOA most drastically by hindering the esterification reaction (20). They are also consistent with the fast and efficient reactions of OA compounds containing carboxylic acid groups and carbonyls with alcohols observed in Ref. (52). In contrast, aldol condensation (Reactions 2 & 3) does not appear to play a significant role in forming the oligomers. Removal of either Reaction 2 or Reaction 3 resulted in a small reduction in the number of theoretically produced structures and recovered HR-MS formulas.

Once a pool of COBRA-generated simulation products is available, it can easily be mined for desired products or classes of products. For example, we are interested in the potential for isoprene photooxidation to generate α,β -unsaturated aldehydes, a class of genotoxic compounds found in cigarette smoke and air pollution (53–58). The most common atmospheric α,β -unsaturated aldehydes with adverse health effects are gas-phase species such as acrolein (C₃H₄O) (59). However, adverse health effects are also strongly linked to inhalation of particulate matter (3, 60–64), which contains more complex and poorly characterized organics. COBRA predicted 6,131 structures corresponding to α,β -unsaturated aldehydes in the isoprene photooxidation SOA data set. This large set of structures can be

filtered further, by mass, O/C ratio, or any other property required to yield a focused set of predicted structures. Applying a computational tool like COBRA along with the molecular-level experimental characterization of products is a powerful first step to identifying condensed-phase atmospheric toxins.

One limitation of COBRA is that it does not currently include kinetics information, and therefore cannot model the relative abundances of individual products. This limitation may be important in cases where the product branching ratios from a particular reaction are dissimilar or dependent on atmospheric conditions, in which case the results from COBRA will artificially enhance the importance of non-competitive products. For example, in the absence of nitrogen oxides (NO_x) the same OH-initiated oxidation chemistry produces more hydroperoxides and fewer carbonyl products (65). However, a scaling factor can be incorporated in the transformation rules if the branching ratios are known or can be estimated with *ab initio* methods. For the comparison with the ESI-based MS data, this is not a major limitation because there are no simple relationships between the detection efficiency in ESI and the concentration of the analyte species (66).

Applying COBRA to the simulation of oligomerization in isoprene OA is a good example of the utility of computational tools for understanding complex natural mixtures and verifying experimental data. An important strength of COBRA is its ability to handle a complex simulation and generate a large number of predicted compounds. While the combinatorial explosion of generated structures means computation time is currently on the order of several days, this time can be significantly reduced in future work by using parallel computing algorithms. If pertinent experimental data are available, direct comparison between the observed and predicted compounds can be made to better constrain chemistry and chemical structures. Furthermore, as COBRA reduces the number of possible structures available for a given chemical formula, the predicted structures may be useful for discriminating isobaric species in a mass spectrum. Applying a small set of transformation rules to predict an experimentally determined high-resolution mass spectrum may become a powerful tool for better understanding the chemistry of complex systems comprised of hundreds or thousands of individual compounds.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The initial support for this project came from a seed grant from the UCI Environmental Institute. TBN and SAN acknowledge additional support by the NSF grants ATM-0831518 and CHE-0909227. The work of DF and PB was supported in part by NIH grants LM010235-01A1 and 5T15LM007743 to PB, and NSF grant MRI EIA-0321390 to PB. JL and AL are supported by the intramural research and development program of the W. R. Wiley Environmental Molecular Sciences Laboratory (EMSL), a national scientific user facility sponsored by the Office of Biological and Environmental Research and located at PNNL. PNNL is operated for the U.S. Department of Energy by Battelle Memorial Institute under contract no. DE-AC06-76RL0 1830.

References

1. Kanakidou M, Seinfeld JH, Pandis SN, Barnes I, Dentener FJ, Facchini MC, Van Dingenen R, Ervens B, Nenes A, Nielsen CJ, Swietlicki E, Putaud JP, Balkanski Y, Fuzzi S, Horth J, Moortgat GK, Winterhalter R, Myhre CEL, Tsigaridis K, Vignati E, Stephanou EG, Wilson J. Organic aerosol and global climate modelling: a review. *Atmos Chem Phys*. 2005; 5:1053–1123.
2. Seinfeld JH, Pankow JF. Organic atmospheric particulate material. *Annu Rev Phys Chem*. 2003; 54:121–140. [PubMed: 12524426]

3. Mauderly JL, Chow JC. Health effects of organic aerosols. *Inhal Toxicol*. 2008; 20(3):257–288. [PubMed: 18300047]
4. Poschl U. Atmospheric aerosols: Composition, transformation, climate and health effects. *Angew Chem Int Ed Engl*. 2005; 44(46):7520–7540. [PubMed: 16302183]
5. Docherty KS, Stone EA, Ulbrich IM, DeCarlo PF, Snyder DC, Schauer JJ, Peltier RE, Weber RJ, Murphy SM, Seinfeld JH, Grover BD, Eatough DJ, Jimenez JL. Apportionment of Primary and Secondary Organic Aerosols in Southern California during the 2005 Study of Organic Aerosols in Riverside (SOAR-1). *Environ Sci Technol*. 2008; 42(20):7655–7662. [PubMed: 18983089]
6. Andreae MO, Rosenfeld D. Aerosol-cloud-precipitation interactions. Part 1. The nature and sources of cloud-active aerosols. *Earth-Sci Rev*. 2008; 89(1–2):13–41.
7. Rudich Y, Donahue NM, Mentel TF. Aging of organic aerosol: Bridging the gap between laboratory and field studies. *Annu Rev Phys Chem*. 2007; 58:321–352. [PubMed: 17090227]
8. Jimenez JL, Canagaratna MR, Donahue NM, Prevot ASH, Zhang Q, Kroll JH, DeCarlo PF, Allan JD, Coe H, Ng NL, Aiken AC, Docherty KS, Ulbrich IM, Grieshop AP, Robinson AL, Duplissy J, Smith JD, Wilson KR, Lanz VA, Hueglin C, Sun YL, Tian J, Laaksonen A, Raatikainen T, Rautiainen J, Vaattovaara P, Ehn M, Kulmala M, Tomlinson JM, Collins DR, Cubison MJ, Dunlea EJ, Huffman JA, Onasch TB, Alfarra MR, Williams PI, Bower K, Kondo Y, Schneider J, Drewnick F, Borrmann S, Weimer S, Demerjian K, Salcedo D, Cottrell L, Griffin R, Takami A, Miyoshi T, Hatakeyama S, Shimono A, Sun JY, Zhang YM, Dzepina K, Kimmel JR, Sueper D, Jayne JT, Herndon SC, Trimborn AM, Williams LR, Wood EC, Middlebrook AM, Kolb CE, Baltensperger U, Worsnop DR. Evolution of Organic Aerosols in the Atmosphere. *Science*. 2009; 326(5959):1525–1529. [PubMed: 20007897]
9. Laskin J, Laskin A, Roach PJ, Slys GW, Anderson GA, Nizkorodov SA, Bones DL, Nguyen LQ. High-Resolution Desorption Electrospray Ionization Mass Spectrometry for Chemical Characterization of Organic Aerosols. *Anal Chem*. 2010; 82(5):2048–2058. [PubMed: 20146449]
10. Bones DL, Henriksen DK, Mang SA, Gonsior M, Bateman AP, Nguyen TB, Cooper WJ, Nizkorodov SA. Appearance of strong absorbers and fluorophores in limonene-O₃ secondary organic aerosol due to NH₄⁺-mediated chemical aging over long time scales. *J Geophys Res*. 2010; 115(D5):D05203.10.1029/2009jd012864
11. De Haan DO, Tolbert MA, Jimenez JL. Atmospheric condensed-phase reactions of glyoxal with methylamine. *Geophys Res Lett*. 2009; 36:5.
12. Nizkorodov SA, Laskin J, Laskin A. Molecular chemistry of organic aerosols through the application of high resolution mass spectrometry. *Phys Chem Chem Phys*. 2011; 13(9):3612–3629. [PubMed: 21206953]
13. Laskin A, Smith JS, Laskin J. Molecular Characterization of Nitrogen-Containing Organic Compounds in Biomass Burning Aerosols Using High-Resolution Mass Spectrometry. *Environ Sci Technol*. 2009; 43(10):3764–3771. [PubMed: 19544885]
14. Schmitt-Kopplin P, Gelencsér A, Dabek-Zlotorzynska E, Kiss G, Hertkorn N, Harir M, Hong Y, Gebefügi I. Analysis of the Unresolved Organic Fraction in Atmospheric Aerosols with Ultrahigh-Resolution Mass Spectrometry and Nuclear Magnetic Resonance Spectroscopy: Organosulfates As Photochemical Smog Constituents. *Anal Chem*. 2010; 82(19):8017–8026. [PubMed: 20879800]
15. Baltensperger U, Chirico R, DeCarlo PF, Dommen J, Gaeggeler K, Heringa MF, Li M, Prévôt AS, Alfarra MR, Gross DS, Kalberer M. Recent developments in the mass spectrometry of atmospheric aerosols. *Eur J Mass Spectrom (Chichester, Eng)*. 2010; 16(3):389–395.
16. Kautzman KE, Surratt JD, Chan MN, Chan AWH, Hersey SP, Chhabra PS, Dalleska NF, Wennberg PO, Flagan RC, Seinfeld JH. Chemical Composition of Gas- and Aerosol-Phase Products from the Photooxidation of Naphthalene. *J Phys Chem A*. 2009; 114(2):913–934. [PubMed: 19904975]
17. Altieri KE, Seitzinger SP, Carlton AG, Turpin BJ, Klein GC, Marshall AG. Oligomers formed through in-cloud methylglyoxal reactions: Chemical composition, properties, and mechanisms investigated by ultra-high resolution FT-ICR mass spectrometry. *Atmos Environ*. 2008; 42(7):1476–1490.
18. Surratt JD, Murphy SM, Kroll JH, Ng NL, Hildebrandt L, Sorooshian A, Szmigielski R, Vermeylen R, Maenhaut W, Claeys M, Flagan RC, Seinfeld JH. Chemical composition of

- secondary organic aerosol formed from the photooxidation of isoprene. *J Phys Chem A*. 2006; 110(31):9665–9690. [PubMed: 16884200]
19. Chan AWH, Chan MN, Surratt JD, Chhabra PS, Loza CL, Crounse JD, Yee LD, Flagan RC, Wennberg PO, Seinfeld JH. Role of aldehyde chemistry and NO_x concentrations in secondary organic aerosol formation. *Atmos Chem Phys*. 2010; 10:7169–7188.
 20. Nguyen TB, Roach PJ, Laskin J, Laskin A, Nizkorodov SA. Effect of humidity on the composition of isoprene photooxidation secondary organic aerosol. *Atmos Chem Phys*. 2011; 11:6931–6944.
 21. Reinhardt A, Emmenegger C, Gerrits B, Panse C, Dommen J, Baltensperger U, Zenobi R, Kalberer M. Ultrahigh mass resolution and accurate mass measurements as a tool to characterize oligomers in secondary organic aerosols. *Anal Chem*. 2007; 79(11):4074–4082. [PubMed: 17411016]
 22. Nguyen TB, Laskin J, Laskin A, Nizkorodov SA. Nitrogen containing organic compounds and oligomers in secondary organic aerosol formed by photooxidation of isoprene. *Environ Sci Technol*. 2011; 45:6908–6918. [PubMed: 21732631]
 23. Tolocka MP, Jang M, Ginter JM, Cox FJ, Kamens RM, Johnston MV. Formation of oligomers in secondary organic aerosol. *Environ Sci Technol*. 2004; 38(5):1428–1434. [PubMed: 15046344]
 24. Barsanti KC, Pankow JF. Thermodynamics of the formation of atmospheric organic particulate matter by accretion reactions - 2. Dialdehydes, methylglyoxal, and diketones. *Atmos Environ*. 2005; 39(35):6597–6607.
 25. Barsanti KC, Pankow JF. Thermodynamics of the formation of atmospheric organic particulate matter by accretion reactions - Part 3: Carboxylic and dicarboxylic acids. *Atmos Environ*. 2006; 40(34):6676–6686.
 26. Hamilton JF, Lewis AC, Carey TJ, Wenger JC. Characterization of polar compounds and oligomers in secondary organic aerosol using liquid chromatography coupled to mass spectrometry. *Anal Chem*. 2008; 80(2):474–480. [PubMed: 18081325]
 27. Chan MN, Surratt JD, Claeys M, Edgerton ES, Tanner RL, Shaw SL, Zheng M, Knipping EM, Eddingsaas NC, Wennberg PO, Seinfeld JH. Characterization and Quantification of Isoprene-Derived Epoxydiols in Ambient Aerosol in the Southeastern United States. *Environ Sci Technol*. 2010; 44(12):4590–4596. [PubMed: 20476767]
 28. Yasmeen FN, Sauret Gal J-F, Maria P-C, Massi L, Maenhaut W, Claeys M. Characterization of oligomers from methylglyoxal under dark conditions: a pathway to produce secondary organic aerosol through cloud processing during nighttime. *Atmos Chem Phys*. 2010; 10(8):3803–3812.
 29. Garland RM, Elrod MJ, Kincaid K, Beaver MR, Jimenez JL, Tolbert MA. Acid-catalyzed reactions of hexanal on sulfuric acid particles: identification of reaction products. *Atmos Environ*. 2006; 40(35):6863–6878.
 30. Jang M, Carroll B, Chandramouli B, Kamens RM. Particle Growth by Acid-Catalyzed Heterogeneous Reactions of Organic Carbonyls on Preexisting Aerosols. *Environ Sci Technol*. 2003; 37(17):3828–3837. [PubMed: 12967102]
 31. Jang MS, Czoschke NM, Northcross AL. Semiempirical model for organic aerosol growth by acid-catalyzed heterogeneous reactions of organic carbonyls. *Environ Sci Technol*. 2005; 39(1):164–174. [PubMed: 15667091]
 32. Chen JH, Baldi P. No Electron Left Behind: A Rule-Based Expert System To Predict Chemical Reactions and Reaction Mechanisms. *J Chem Inf Model*. 2009; 49(9):2034–2043. [PubMed: 19719121]
 33. Kayala MA, Azencott CA, Chen JH, Baldi P. Learning to predict chemical reactions. *J Chem Inf Model*. 2011; 51(9):2209–22. [PubMed: 21819139]
 34. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*. 1988; 28(1):31–36.
 35. James CA, Weininger D, Delany J. Daylight Theory Manual. Jan.
 36. OpenEye Scientific Software, I. OEChem, version 1.7.4. OpenEye Scientific Software; Santa Fe, NM, USA: 2010.
 37. Kim S, Rodgers RP, Marshall AG. Truly “exact” mass: Elemental composition can be determined uniquely from molecular mass measurement at <0.1 mDa accuracy for molecules up to <500 Da. *Int J Mass Spectrom*. 2006; 251(2–3):260–265.

38. Carlton AG, Wiedinmyer C, Kroll JH. A review of secondary organic aerosol formation from isoprene. *Atmos Chem Phys*. 2009; 9(14):4987–5005.
39. Edney EO, Kleindienst TE, Jaoui M, Lewandowski M, Offenberg JH, Wang W, Claeys M. Formation of 2-methyl tetrols and 2-methylglyceric acid in secondary organic aerosol from laboratory irradiated isoprene/NO_x/SO₂/air mixtures and their detection in ambient PM_{2.5} samples collected in the eastern United States. *Atmos Environ*. 2005; 39(29):5281–5289.
40. Grosjean D, Williams EL, Grosjean E. Atmospheric chemistry of isoprene and of its carbonyl products. *Environ Sci Technol*. 1993; 27(5):830–840.
41. Spaulding RS, Schade GW, Goldstein AH, Charles MJ. Characterization of secondary atmospheric photooxidation products: Evidence for biogenic and anthropogenic sources. *J Geophys Res*. 2003; 108(D8):4247.10.1029/2002jd002478
42. Kleindienst TE, Lewandowski M, Offenberg JH, Jaoui M, Edney EO. The formation of secondary organic aerosol from the isoprene plus OH reaction in the absence of NO_x. *Atmos Chem Phys*. 2009; 9(17):6541–6558.
43. Claeys M, Graham B, Vas G, Wang W, Vermeylen R, Pashynska V, Cafmeyer J, Guyon P, Andreae MO, Artaxo P, Maenhaut W. Formation of secondary organic aerosols through photooxidation of isoprene. *Science*. 2004; 303(5661):1173–1176. [PubMed: 14976309]
44. Paulot F, Crounse JD, Kjaergaard HG, Kroll JH, Seinfeld JH, Wennberg PO. Isoprene photooxidation: new insights into the production of acids and organic nitrates. *Atmos Chem Phys*. 2009; 9:1479–1501.
45. Szmigielski R, Surratt JD, Vermeylen R, Szmigielska K, Kroll JH, Ng NL, Murphy SM, Sorooshian A, Seinfeld JH, Claeys M. Characterization of 2-methylglyceric acid oligomers in secondary organic aerosol formed from the photooxidation of isoprene using trimethylsilylation and gas chromatography/ion trap mass spectrometry. *J Mass Spectrom*. 2007; 42(1):101–116. [PubMed: 17154243]
46. Tan Y, Carlton AG, Seitzinger SP, Turpin BJ. SOA from methylglyoxal in clouds and wet aerosols: Measurement and prediction of key products. *Atmos Environ*. 2010; 44(39):5218–5226.
47. Petitjean M, Reyes-Perez E, Perez D, Mirabel P, Le Calve S. Vapor Pressure Measurements of Hydroxyacetaldehyde and Hydroxyacetone in the Temperature Range (273 to 356) K. *J Chem Eng Data*. 2009; 55(2):852–855.
48. Healy RM, Wenger JC, Metzger A, Duplissy J, Kalberer M, Dommen J. Gas/particle partitioning of carbonyls in the photooxidation of isoprene and 1,3,5-trimethylbenzene. *Atmos Chem Phys*. 2008; 8(12):3215–3230.
49. Liggio J, Li SM, McLaren R. Reactive uptake of glyoxal by particulate matter. *J Geophys Res-Atmos*. 2005; 110(D10):D10304.
50. Liggio J, Li S-M, McLaren R. Heterogeneous reactions of glyoxal on particulate matter: identification of acetals and sulfate esters. *Environ Sci Technol*. 2005; 39(6):1532–1541. [PubMed: 15819206]
51. Zhao J, Levitt NP, Zhang R, Chen J. Heterogeneous reactions of methylglyoxal in acidic media: implications for secondary organic aerosol formation. *Environ Sci Technol*. 2006; 40(24):7682–7687. [PubMed: 17256513]
52. Bateman AP, Walser ML, Desyaterik Y, Laskin J, Laskin A, Nizkorodov SA. The Effect of Solvent on the Analysis of Secondary Organic Aerosol Using Electrospray Ionization Mass Spectrometry. *Environ Sci Technol*. 2008; 42(19):7341–7346. [PubMed: 18939568]
53. Witz G. Biological interactions of alpha,beta-unsaturated aldehydes. *Free Radic Biol Med*. 1989; 7(3):333–349. [PubMed: 2673948]
54. Wynder EL, Hoffmann D. Tobacco and Health. *N Engl J Med*. 1979; 300(16):894–903. [PubMed: 370598]
55. Berhane K, Widersten M, Engstrom A, Kozarich JW, Mannervik B. Detoxication of base propenals and other alpha, beta-unsaturated aldehyde products of radical reactions and lipid peroxidation by human glutathione transferases. *Proc Natl Acad Sci USA*. 1994; 91(4):1480–1484. [PubMed: 8108434]

56. Trombetta D, Saija A, Bisignano G, Arena S, Caruso S, Mazzanti G, Uccella N, Castelli F. Study on the mechanisms of the antibacterial action of some plant alpha,beta-unsaturated aldehydes. *Lett Appl Microbiol.* 2002; 35(4):285–290. [PubMed: 12358689]
57. Lambert C, McCue J, Portas M, Ouyang Y, Li J, Rosano TG, Lazis A, Freed BM. Acrolein in cigarette smoke inhibits T-cell responses. *J Allergy Clin Immunol.* 2005; 116(4):916–922. [PubMed: 16210070]
58. Andre E, Campi B, Materazzi S, Trevisani M, Amadesi S, Massi D, Creminon C, Vaksman N, Nassini R, Civelli M, Baraldi PG, Poole DP, Bunnett NW, Geppetti P, Patacchini R. Cigarette smoke-induced neurogenic inflammation is mediated by alpha,beta-unsaturated aldehydes and the TRPA1 receptor in rodents. *J Clin Invest.* 2008; 118(7):2574–2582. [PubMed: 18568077]
59. Kehrer JP, Biswal SS. The molecular effects of acrolein. *Toxicol Sci.* 2000; 57(1):6–15. [PubMed: 10966506]
60. Rohr AC, Shore SA, Spengler JD. Repeated exposure to isoprene oxidation products causes enhanced respiratory tract effects in multiple murine strains. *Inhal Toxicol.* 2003; 15(12):1191–1207. [PubMed: 14515222]
61. Delfino R, Sioutas JC, Malik S. Potential role of ultrafine particles in associations between airborne particle mass and cardiovascular health. *Environ Health Perspect.* 2005; 113(8):934–46. [PubMed: 16079061]
62. Dockery DW, Pope CA 3rd, Xu X, Spengler JD, Ware JH, Fay ME, Ferris BG Jr, Speizer FE. An association between air pollution and mortality in six U.S. cities. *N Engl J Med.* 1993; 329(24):1753–9. [PubMed: 8179653]
63. Pope CA III, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, Thurston GD. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *J Am Med Assoc.* 2002; 287(9):1132–1141.
64. Samet JM, Dominici F, Currier FC, Coursac I, Zeger SL. Fine particulate air pollution and mortality in 20 U.S. cities, 1987–1994. *N Engl J Med.* 2000; 343(24):1742–1749. [PubMed: 11114312]
65. Atkinson R. Atmospheric chemistry of VOCs and NOx. *Atmos Environ.* 2000; 34(12–14):2063–2101.
66. Tang K, Page JS, Smith RD. Charge competition and the linear dynamic range of detection in electrospray ionization mass spectrometry. *J Am Soc Mass Spectrom.* 2004; 15(10):1416–1423. [PubMed: 15465354]

Figure 1.

A schematic representation of COBRA. COBRA converts a set of seed molecules into a set of predicted products using predefined chemical reaction rules. This process is repeated for several iterations, with predicted products going back into the pool of reacting molecules after each step. The four reactions used in this work are shown in the center panel, and the full list of 27 seed molecules is shown in Table 1.

Figure 2.

Comparison between the HR-MS experiment and COBRA predictions. Panel (a) shows the HR-MS peaks that are predicted by COBRA (70%). Panel (b) shows the remaining 30% of HR-MS peaks that are not predicted by our simulation. In both panels, NOC peaks are colored green. Panel (c) indicates the number of isomeric structures predicted at each observed molecular mass.

Figure 3.

Experimental MS² fragmentation spectra for precursor (I) ions (a) C₈H₁₂NO₉⁻ and (b) C₁₄H₂₀NO₁₃⁻. Two structures predicted by COBRA that are consistent with the fragmentation patterns are overlaid on the spectra. Product ions are obtained from fragmentation at the dashed bonds (after losses of neutral molecules). Product ions for (a) are: II. *m/z* 203.056 (I – HNO₃); III. *m/z* 189.040 (I-CH₃NO₃); IV. *m/z* 164.020 (I-C₄H₆O₃); V. *m/z* 119.035 (I-C₄H₅NO₅). Product ions for (b) are: II. *m/z* 347.098 (I – HNO₃); III. *m/z* 333.082 (I-CH₃NO₃); IV. *m/z* 308.062 (I-C₄H₆O₃); V. *m/z* 307.082 (I-C₃H₅NO₃).

Table 1

Seed molecules used for the COBRA simulations described in this work. Exclusion percent is a measure of the molecule's contribution to the pool of correctly predicted products.

Formula	Name	Exclusion percent	Structure
$C_2H_4O_2$	glycolaldehyde	0
$C_2H_4O_2$	acetic acid	0
$C_2H_4O_3$	glycolic acid	-0.6
$C_3H_4O_2$	methylglyoxal	-0.3
$C_3H_4O_2$	acrylic acid	-0.9
$C_3H_6O_2$	hydroxyacetone	0
$C_3H_6O_2$	propanoic acid	2.2
$C_3H_4O_3$	pyruvic acid	2.8
$C_3H_6O_3$	lactic acid	-2.2
$C_3H_4O_4$	3-hydroxypyruvic acid	-5.6
$C_3H_4O_4$	2-hydroxy-3-oxopropanoic acid	-2.5

Formula	Name	Exclusion percent	Structure
$C_4H_6O_2$	methacrylic acid	1.2
$C_4H_6O_3$	2-methyloxopropanoic acid	-1.9
$C_4H_8O_3$	2-methylglyceraldehyde	-0.3
$C_4H_8O_4$	2-methylglyceric acid	-1.5
$C_4H_8O_4$	erythrose/threose	-0.9
$C_4H_6O_4$	methylmalonic acid	-0.6
$C_4H_6O_4$	2-hydroxy-2-methyl-3-oxopropanoic acid	-2.5
$C_4H_6O_5$	hydroxy(methyl) propanedioic acid	-0.6
$C_4H_7O_6N$	2-methylglyceric acid 3- nitrate	18.8
$C_4H_8O_5$	2,3,3-trihydroxy-2- methylpropanoic acid	-2.2
$C_5H_8O_3$	2-hydroxy-2- methylbutanedial	2.8

Formula	Name	Exclusion percent	Structure
$C_5H_8O_4$	methylsuccinic acid	3.4
$C_5H_8O_4$	2-hydroxy-2-Me-4-oxobutanoic acid	-4.6
$C_5H_8O_5$	2-hydroxy-2-methylbutanedioic acid	-1.9
$C_5H_{10}O_3$	2,4-dihydroxy-2-methylbutanal	6.2
$C_5H_{10}O_5$	2,3,4-trihydroxy-2-methylbutanoic acid	-0.6

Table 2

A small fraction of the full output from COBRA. For each observed product, several different isomers (SMILES structures) are predicted (number of hits). The full number of hits for each mass is shown in Figure 2C.

Product Mass (Da)	Product Formula	Hits (n)	SMILES # 1	SMILES # 2	... SMILES # n
128.047	C ₆ H ₈ O ₃ N ₀	4	CC(=C)C(=O)OCC=O	CC(=C)C(=O)OC(=O)C	CC(=O)COC(=O)C=C
130.027	C ₅ H ₆ O ₄ N ₀	4	CC(=O)C(=O)OCC=O	CC(=O)C(=O)OC(=O)C	C=CC(=O)OCC(=O)O
132.042	C ₅ H ₈ O ₄ N ₀	9	CCC(=O)OC(=O)CO	CCC(=O)OCC(=O)O	CC(C(=O)O)OC(=O)C
136.037	C ₄ H ₈ O ₅ N ₀	2	C(C(O)OC(=O)CO)O	C(C(O)OCC(=O)O)O	
142.063	C ₇ H ₁₀ O ₃ N ₀	2	CCC(=O)OC(=O)C(=C)C	CC(=C)C(=O)OCC(=O)C	
146.022	C ₅ H ₆ O ₅ N ₀	8	CC(=O)C(=O)OC(=O)CO	CC(=O)C(=O)OCC(=O)O	CC(=O)OC(C=O)C(=O)O

Table 3

Results from an exclusion analysis, in which one of the four reactions shown in Fig. 1 is disabled. The values listed in the last three columns represent: 1) the percent reduction in the number of experimentally observed formulas recovered by the simulation, 2) the percent reduction in the total number of theoretical formulas predicted by the simulation, and 3) the percent reduction in the total number of theoretical structures predicted by the simulation.

Reaction Transform	Reaction Type	Percent reduction in experimentally observed formulas	Percent reduction in theoretically predicted formulas	Percent reduction in theoretically predicted structures
1	esterification	40.9	12.8	37.8
2	aldol condensation	0.3	0.7	5.6
3	elimination of water from aldol condensates	0	1.5	0.01
4	hemiacetal formation	10.2	-0.1	46.7