Building the KOMODO media database

Matthew Oberhardt (mattoby@gmail.com)
Ruppin Lab, University of Maryland
Uri Gophna (urigo@post.tau.ac.il)
Tel Aviv University
Eytan Ruppin (eyruppin@gmail.com)
University of Maryland
Raphy Zarecki
Tel Aviv University

Method Article

Keywords: systems biology, bacterial culturing, microbial database, microbial media

Posted Date: September 3rd, 2015

DOI: https://doi.org/10.1038/protex.2015.075

Abstract

This protocol describes the process used to convert >1300 microbial media recipes, listed on the website of the Leibniz Institute DSMZ \(see: https://www.dsmz.de/?id=441), into a machine readable format, and then to read them into a SQL database. In the process, we combine compounds with degenerate names and include concentrations of all compounds in standard units. The protocol includes steps for error-checking, as well as a process to recursively add media when they cross-reference each other. The protocol might be useful in other, similar database building tasks, and especially might help in efforts to incorporate other microbial media recipes from literature into this or another database.

Introduction

A large collection of media recipes for microbial strains is available through the German Leibniz Institute DSMZ strain and media collection \(accessible here: https://www.dsmz.de/?id=441). These recipes are publicly available, but they are contained as instructions in portable document format \(PDF) files that must be searched on an organism-by-organism basis. Putting these recipes into a usable database form required extensive and non-trivial work parsing, merging, and organizing, as well as handling cross-references between media and sub-media compound mixtures such as 'trace element solutions,' which could be detailed and referenced from any DSMZ medium. Dealing with such cross references involves handling multiplication of volumes, masses, and concentrations, even in cases when the same media component is included both in a sub-medium mixture and in the main medium description, often with non-matching names and/or units \(e.g., once in g/L and once in moles/L). Medium and sub-medium volumes also are often not listed in media, but are assumed by microbiologists to be 1 liter per the number of grams \(or moles, or milliliters) of compounds listed for inclusion. However, there is no general rule for this, as some media do list specific volumes, some of which do not sum up to 1 liter. Often, the volumes are left to be deciphered through common sense. We tackled these challenges with a pipeline that is part manual and part automated. We used this pipeline to read in more than 1500 PDF media descriptions and to create the KOMODO database, containing media compositions with standardized units. The pipeline is depicted as a 15-step process in Figure 1 of this protocol. Each step is explained in detail below.

Procedure

Step 1. Copy media PDFs into text file PDFs of all the media in the DSMZ database were copied verbatim into a text file. **Step 2. Manually reformat PDFs for machine reading** The resulting ~27,000 lines of text were manually reformatted in a way that could be machine read, using tags such as /ph/ \(set the pH tag of the medium), /replace/ \(replace one compound with another), /conc/ \(change the concentration of a compound), and /rm/ \(remove a compound from the medium) to denote media features and instructions. These tags were embedded in a specialized syntax that was similar to natural language media instructions, and thus required minimal alterations from the instructions listed in the original pdfs, but that followed a defined syntactical structure that could be interpreted by a computer

program. We were able to extract and reformat the majority of media from the DSMZ database in this way. A few examples of final formatted media descriptions \(ready for machine reading) are shown here: Example 1: #nondefined# #medium# 233. METHANOLOBUS-I-MEDIUM medium 141 /rm/ Yeast extract /rm/ Trypticase /rm/ Na-acetate /conc/ NaHCO3 @ 1 g/l /anaerobic/ /atm/ N2, CO2 /notag/ methanol @ 0.5 %% /ph/ 6.5 Example 2: #nondefined# #medium# 237. ENB medium ##Nutrient broth \(Difco) @ 100 %% /notag/ Peptone @ 5 g /notag/ KH2PO4 @ 1.5 g /notag/ K2HPO4 @ 3.5 g /notag/ NaCl @ 5 g /notag/ Glucose @ 1 g /notag/ Agar @ 15 g /notag/ Distilled water @ 1000 ml /ph/ 6.8 - 7.0 Example 3: #defined# #medium# 194. DESULFOBULBUS MEDIUM medium 193 /conc/ NaCl @ 1 g/l /conc/ MgCl2 x 6 H2O @ 0.4 g/l /rm/ Na-acetate x 3 H2O /notag/ Sodium propionate @ 1.5 g/l #194.1 For DSM 1744 /conc/ NaCl @ 10 g/l /rm/ Sodium propionate /notag/ sodium lactate @ 2.5 g/l /notag/ Yeast extract @ 1 g/l #194.2 For DSM 14880 /rm/ Sodium propionate /notag/ Yeast extract @ 0.5 g/l /notag/ Sodium pyruvate @ 2.2 g/l #194.3 For DSM 21556 /rm/ Sodium propionate /notag/ Sodium butyrate @ 2.2 g/l Ordinary form for a component: /notag/ NaCl @ 5 g this indicates that there are 5 grams of NaCl added to the medium. The line has the form: \[/tag/, compound name, @, concentration value, concentration unit]. If the concentration unit is a percent, then the molar amount is calculated based on a volume conversion. All of the tags we use are listed here: # = indicates the title of a new medium which is a strain-specific variant of the base medium being described. For example, in example 3 above, media 194.1, 194.2, and 194.3 are strain specific variants of medium 194, with some added instructions. ## = indicates that all components of a medium or submedium need to be added in some amount. // = a comment \(to be functionally ignored) /also/ = same as /or/ /anaerobic/ = no oxygen in the medium. /atm/ = this will be followed by all of the components in the atmosphere of the medium \(e.g., H2, CO2). /atm+/ = add the listed components to the atmosphere. /conc/ = alter the concentration of the component to the given quantity. /editnotes/ = a comment /notag/ = no tag \(this is just a placeholder). /or/ = this component can replace the one on the previous line \(can rack up multiple /or/'s this way) /ph/ = the pH of the medium is listed after this tag, along with any components used to adjust it \(e.g.: /ph/ 6 – 7, KOH) /rm/ = remove this component from the medium. /s/ = defines the substrate of a medium. Medium = add all components of the referenced medium to the current medium. ## = reference to another medium **Step 3. Add special instructions for specific organisms** We noticed that a large number of organisms had specialized growth instructions listed either within the media descriptions, or in the organism-medium mapping file provided to us by DSMZ. We considered these instructions critical to building an accurate database. To incorporate them, we copied the components of the base media and then implemented the stated changes to create medium definitions for each media variant. In all, this process resulted in nearly a doubling of the number of media in the database, from 1946 to 3672. In the DSMZ listing \(http://www.dsmz.de/?id=441), each medium is referenced by an ID number. We generated unique new media IDs for these media variants by following the base media IDs with a period \(.) or an underscore \(_), and then a unique numerical or text string. Additionally, many media included in their compositions submedia, which were to be mixed independently and then combined. To ease the formation of the database, we treated each sub-medium as an independent medium with a new medium ID of 2000 or above. This then allowed us to calculate cross-references between media and submedia using a standardized methodology. **Step 4. Map all media components to unique component names**

Media components as listed in literature are highly redundant and degenerate. For example, the compound Sodium sulfide is listed in the database in at least 9 different ways \(sodium sulfide, sodium sulphide, Na2S x 9 H2O, Na2S x 9H2O, etc.). To convert the database to the most versatile form, we manually mapped compound names to 'semi-unique names' as an intermediate layer, and then finally to 'unique names' that contained only the precise metabolites contributed to a medium by a metabolite. For example, the 'semi-unique' name mapped to all original forms of sodium sulfide \(including hydrated forms) from media descriptions is 'sodium sulfide', and the 'unique name' is 'SEED-cpd00239#cpd00971#,' which precisely depicts the two SEED compounds \(cpd00239 = Sulfide ion, and cpd00971 = Sodium ion). **Step 5. Map unique components to SEED compounds or 'complex' categories** We defined three classes of unique names, to which all components are mapped: \(1) SEED compounds, which are denoted with a "SEED-" tag and then up to three SEED metabolites contained within them \(e.g. "SEED-cpd00239#cpd00971#"). \(2) Complex components, which are denoted with a "rich-" tag \(e.g., "rich-peptone"). \(Note, this 'richness' is not to be confused with media richness; rather, it denotes complexity \(media richness is treated differently in the work). In the main text, complex components are presented with a complex- tag instead of a rich- tag. The two are interchangeable, and both denote complexity, not media richness). \(3) Other compounds, which are chemically defined but are not in SEED. These are simply written out in full \(e.g., "1,4-Naphthaquinone"). **Step 6. Calculate total volume of each medium** A rule of thumb in microbiology media recipes is that the quantities of compounds listed are those needed to produce 1 liter of final medium. Because of this, media volumes are often omitted \(and assumed to be 1 L), or are explicitly accounted for by mixing of media compounds with 1 liter of water. However, there are many exceptions to this rule, such as media or submedia compositions that include some volume of water that is not 1 liter, or that contain very small volumes of liquid \(from, e.g., addition of some volume of ethanol), which should not be considered the 'final volume' of the medium by any means. It was critical to determine the exact volume of media in order to properly convert compound units into concentrations \(see Steps 9-11). To deal with this, we classed media and submedia into categories called 'fill' and 'scale.' The 'fill' tag means that whatever volume a medium has should be 'filled' to 1 liter, i.e., that the volume listed should simply be ignored; the 'scale' tag means that the concentrations of compounds listed in a medium description should be scaled up with the listed volume until that volume comes out to 1 liter. Media were classed as 'fill' and 'scale' using general rules, which were overridden in ambiguous cases by manual curation \(filling and scaling pseudocode is listed below). **Step 7. Calculate unit multiplier for 1L of medium** Finally, we adjusted final volumes of 'fill' media and then determined a multiplier for each 'scale' medium and submedium composition in order to convert compound units from Moles to Moles per Liter \(see Steps 9-11). **Step 8. Unpack cross-media references** Large proportions of DSMZ media contain cross-references either to other media or to complex submedia \(~60% and >25%, respectively). Many of these references also contain references, so sometimes multiple layers of references must be unpacked in order to build a given medium. Faithfully unpacking these cross-references requires \(1) determining the molar concentrations of all compounds in the cross-referenced submedium/medium, \(2) determining the volume of the submedium/medium per liter of final medium, \(3) multiplying these two factors correctly to get the concentration of each submedium compound, and \(4) accounting for the volume of the cross-referenced

submedium/medium in determining the final medium volumes. This process was fully automated. **Step 9. Calculate component amounts per medium** A goal of this project was to include every compound if possible with standardized units, as this would ease analyses between media and between compounds. Compounds in the original media files were listed with over 30 distinct units. As a first step, we built a mapping with multipliers to convert all of these units into five standard ones: g/L, L/L, M/L, trace, and 'gas substrate'. **Step 10. Convert all compound units to Moles** The next step was to convert all of these units \(except for the 'trace' and 'gas substrate' ones, which were treated separately) into Moles. To do this, we needed to obtain the molecular weights of all defined media components, as well as the molar ratios of each component forming each semi-unique compound name. When available, molecular weights of SEED compounds were taken from the SEED database. For SEED compounds without molecular weights listed, as well as for compounds falling into the "Other" category \(i.e., defined but not listed in SEED), we curated molecular weights manually based on internet searches. Finally, we manually curated molar ratios of compounds in the original compound names, as well as the number of waters linked to each compound. With all of this information, we were able to calculate from, for example, the compound name "CoCl2 x 6 H2O," the exact molar amounts of cobalt and chloride in a final medium composition, even if the original compound was listed in grams and not Moles. For the subset of compounds listed with units of volume rather than grams or Moles, we universally assumed that the densities of the fluids were equal to the density of water \(1 gram per ml), in order to ease the conversion of units. This rule was not used for volumes of submedia or media, but only for units of individual compounds. **Step 11. Use media volumes to convert Moles into Moles/Liter** Finally, we needed to convert the units for each compound from a molar amount \(Moles, M) into a molar concentration \(Moles per Liter, M/L). This was done by multiplying the Molar amount of each compound by the medium volume multipliers as determined in steps 6-7. **\(For steps 12-14)** Many complicated bookkeeping calculations are automated in steps 6-11 of this workflow, and there are many potential sources for mistakes or errors. Therefore, it was important to validate several key results as a sanity check in order to ensure that the database was faithfully converted. To do this, we manually produced three 'gold standard' files for validation. Steps 12-14 each use one of these files as a manual check. In practice, validation files were built manually for steps 12-14, and then were used over multiple rounds of validation. These files were used for extensive troubleshooting and debugging of the conversion code and of the syntax in the files for conversion, until there were no mismatches left between the manual files and the automated results. **Step 12. Validate media volumes** Manually calculated media volumes for 149 media and "fill" or "scale" statuses for 138 media, to check against the results of step 6. **Step 13. Validate concentrations of unique compound names** Manually calculated quantities \(including units) of 973 compounds referenced across media, to validate the results of steps 7-9. **Step 14. Validate molar concentrations of SEED compounds** Manually calculated molar concentrations of 965 SEED compounds in media, to validate the results of steps 10-11. **Step 15. Add media & compounds to KOMODO database** The work in steps 1-14 ultimately produces a high confidence matrix of media versus the concentrations of compounds within them. This information was next integrated into a database format, along with information provided by DSMZ of which organisms grow on which media, and linkages of DSMZ organism IDs to NCBI IDs and SEED organism IDs, when available. This was done

automatically using custom built code. **Pseudocode for steps 6-11 of database build** Here we provide pseudocode for steps 6-11 of the database building process, which are the automated portions for reading in the initial database information: \(1) Determine volumes of each of the media. a. All submedia are considered to have volumes. Therefore, convert ones with units of mass into units of volume with the 1 ml = 1 g conversion \(even though it's not always precise). b. For metabolites added with parentheses, add the volume if it exists in one of the parentheses. For example: in /notag/ NaCl @ \(100 ml)*\(5 g/l), the volume added is 100 ml. \(2) Adjust volumes based on the following formula: a. All rules about to be written are overridden by the tags put on specific media / submedia for determining the fill or scale status. The tags are: "fill" and "scale". "fill" means that the medium should have volume added to it so that the final volume is 1 liter, but without altering the amounts of compounds in the medium. "scale" means that concentrations in the medium should scale up along with the volume of the medium, until the volume is 1 liter. For example: i. Fill: if there's 1 g HCl in 700 ml medium, and the tag is "fill", then the final volume is 1 liter and the final concentration of HCl is 1 g/l. ii. Scale: if there's 1 g HCl in 700 ml medium, and the tag is "scale", then the final volume is 1 liter and the final concentration of HCl is \(1 g / 0.7 l) = 1.43 g/l. b. If a medium or submedium has a volume of 0, adjust the volume to 1 liter \(i.e., the rule is "fill"). c. If a medium or submedium has a volume of 1 liter, keep as it is. d. If a medium has a volume above 1 liter, the rule is "scale" e. All submedia with nonzero volumes should be "scaled" \(3) Determine the amount of each compound in each medium. For this, parenthesis are multiplied out \(e.g., \(100 ml)**\(5 g/l) □ 0.5 g/l), with the general principle that all compounds are in units of mass or moles \(i.e., g/l or M/l). A compound that has a volume should be converted to grams using the formula: 1 ml = 1 g \(even though this is not strictly accurate, it's a reasonable approximation for most compounds we're dealing with). Also, submedia are treated like more embedded parentheses. For example, if medium a contains 10 ml of medium b, and medium b contains 15 ml medium c, and medium c contains 5 ml of metabolite X, then medium a contains \(10 ml/l)**\(15 ml/l)**\(5 ml metabolite X)**\(1 g/ml conversion) = 0.00075 g metabolite X. Percentages are converted as shown in the conversion sheet. \(4) For all SEED compounds, convert grams into Moles. For this calculation, water molecules that are attached to the compound molecules should be accounted for. Water molecules that should be accounted for are always in the form "metabolite x N H2O". For example, the metabolite: /notag/ CaCl2 x 2 H2O @ 10 mg would be converted as such: a. Molecular weight of CaCl2 is 110 b. Molecular weight of H2O is 18 c. So 10 mg of CaCl2 x 2 H2O = \(10 mg) / \(\(110 + 2*18) mg/mmol) = 0.0684 mmol of CaCl2 **Coupling with SEED** An ultimate goal of this work is to combine the knowledge embedded in manually built media with modern sequencing and genomics databases, in a form that may be used for large-scale metabolic analysis. A natural choice for this linkage is the Model SEED, a project that utilizes the RAST genome annotation server to automatically build and store genome-scale metabolic models 1, 2. To this end, we converted all compounds that had SEED equivalents into SEED compound names and IDs, with each ingredient listed in a medium converted into between one and three SEED compounds \(see example in Figure 1). The quantities of these SEED compounds \(as well as compounds without SEED equivalents) were then combined in final media compositions and converted to molar units.

Timing

The database build took two researchers approximately 6 months to complete.

References

1. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol 28, 977-982 \(2010). 2. Henry CS, et al. Connecting genotype to phenotype in the era of high-throughput sequencing. Biochim Biophys Acta, \(2011).

Figures

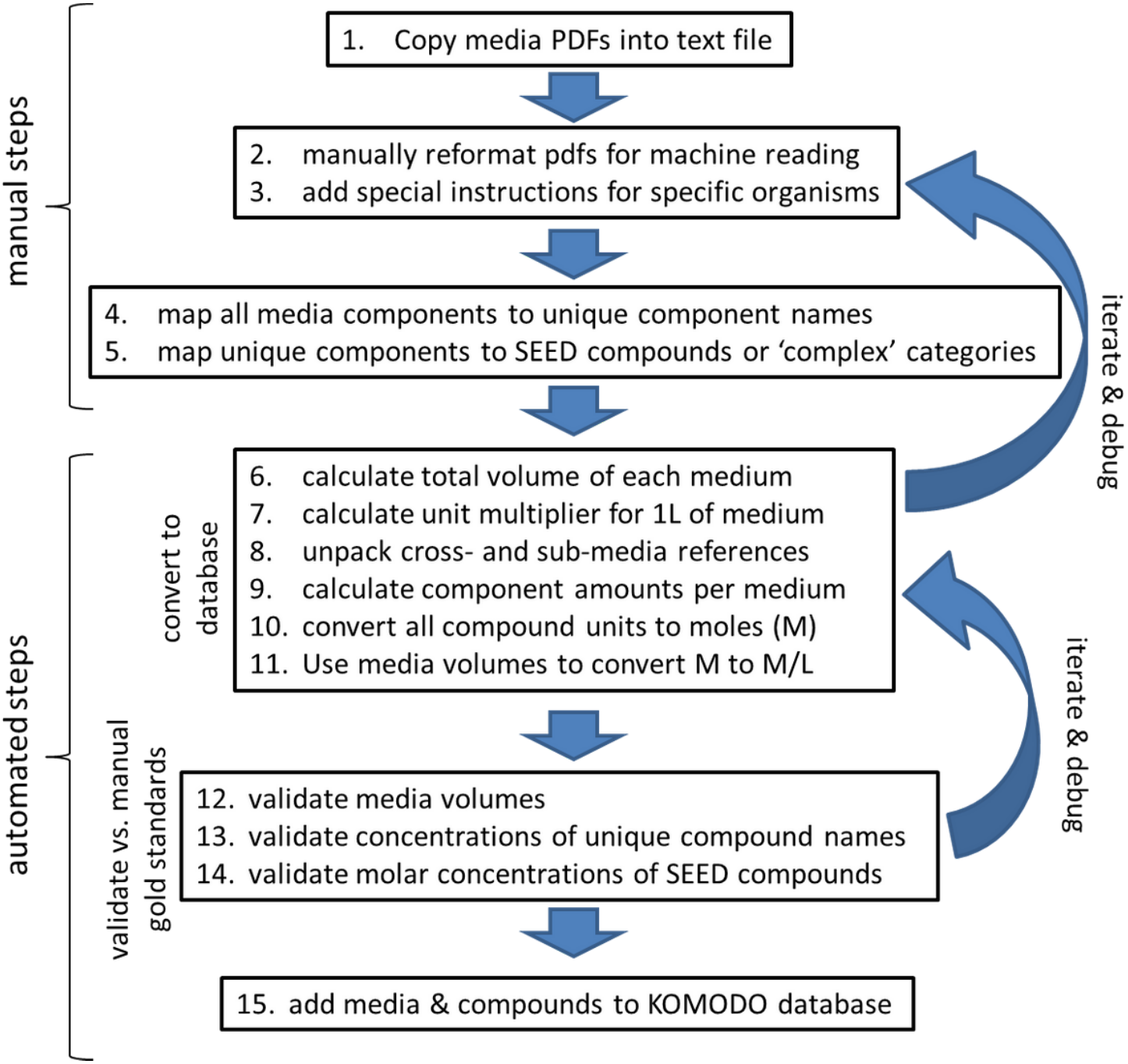

Figure 1

Workflow for building KOMODO, the Known Media Database This partially manual and partially automated workflow enabled the building of KOMODO, based on media recipes publically available on the Leibniz DSMZ website.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement0.xlsx