RESEARCH ARTICLE



Open Access

BKM-react, an integrated biochemical reaction database

Maren Lang[†], Michael Stelzer[†] and Dietmar Schomburg^{*}

Abstract

Background: The systematic, complete and correct reconstruction of genome-scale metabolic networks or metabolic pathways is one of the most challenging tasks in systems biology research. An essential requirement is the access to the complete biochemical knowledge - especially on the biochemical reactions. This knowledge is extracted from the scientific literature and collected in biological databases. Since the available databases differ in the number of biochemical reactions and the annotation of the reactions, an integrated knowledge resource would be of great value.

Results: We developed a comprehensive non-redundant reaction database containing known enzyme-catalyzed and spontaneous reactions. Currently, it comprises 18,172 unique biochemical reactions. As source databases the biochemical databases *BRENDA*, *KEGG*, and *MetaCyc* were used. Reactions of these databases were matched and integrated by aligning substrates and products. For the latter a two-step comparison using their structures (*via InChIs*) and names was performed. Each biochemical reaction given as a reaction equation occurring in at least one of the databases was included.

Conclusions: An integrated non-redundant reaction database has been developed and is made available to users. The database can significantly facilitate and accelerate the construction of accurate biochemical models.

Background

For the construction of cellular models, the development of organism-specific reaction networks is essential. A number of sources for biochemical reactions exist, as the databases *BRENDA* [1], *KEGG* [2], and *MetaCyc* [3]. In general, the integration of biological databases is not trivial [4]. Due to the fact that the completeness of reaction data differs between the databases, it becomes important to combine the available reaction information of the used source databases in form of an integrated reaction database.

So a combination will lead to more complete and reliable metabolic networks. Therefore it is necessary to find identical reactions between the recognized databases. As different compound names and compound IDs, as well as reaction IDs, are in use within the described biochemical reactions a comparison is far from straightforward. A major obstacle results from the use of generic compound names, *e.g. 'an aldehyde'* or *'an alcohol'*. Furthermore some reactions even occur in the same database twice with different reaction IDs.

Integrated databases exist for diverse biological topics. The *TRANSPATH*[®] database for example is an integrated database which deals with signal transduction information [5]. As an example for an integrated metabolic database system the database BioSilico can be mentioned here [6]. For creation of this database, information of the metabolic databases *KEGG*, *ENZYME* [7], *EcoCyc* [8], and *MetaCyc* was combined, the latter two building parts of BioCyc [9]. The database BioSilico includes information on enzymes, biochemical compounds, and reactions. Radrich *et al.* [10] provide a semi-automated tool for the process of genome-scale network reconstruction demonstrated on the basis of data for Arabidopsis thaliana. Their integrated data set is built on the two sources KEGG and AraCyc [11]. Furthermore a reaction database on human biological pathways and processes named Reactome [12] exists as well as an annotated reaction database called *Rhea* [13], basically a modified version of



© 2011 Lang et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

^{*} Correspondence: D.Schomburg@tu-bs.de

⁺ Contributed equally

Department of Bioinformatics and Biochemistry, Institute for Biochemistry and Biotechnology, Technische Universität Braunschweig, Langer Kamp 19 B, 38106 Braunschweig, Germany

the reactions defined in the *IUBMB* enzyme list [14]. A collection of biochemical reactions and pathways in printed form contains the book *Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology* [15].

Methods

In this work information from the biological databases BRENDA [1], KEGG [2], and MetaCyc [3] was used (May 2011). Reaction comparisons were done by an *in silico* approach in which two steps, first a comparison of reactant structures using InChIs (linearized chemical structure descriptors [16]) and, second, a compound name comparison (incl. synonyms), were combined. An InChI structure coding was generated based on an original Mol*file* (contains molecular structure information [17]) by using a special converting tool (InChI version 1 (software version 1.03) for Standard and Non-Standard InChI/ InChIKey [18]). By using only relevant layers of an selfgenerated InChI, a higher matching rate was achieved. For this purpose we dropped the InChI layers dependent on the ionisation state so that e.g. acetic acid and the acetate ion were considered to be the same compound. Reactions without EC numbers were included as well as those reactions with incomplete EC numbers. Spontaneous reactions without EC number were labelled SPON-TANEOUS. Before the comparison, the compounds water (H_2O) and proton (H^+) were removed from the reactions. Additionally, a stoichiometry check was executed. This information was added as attribute to the reactions in the database as a quality measure. Stoichiometrically imbalanced reactions were marked as incom*plete* in the column *Stoichiometry*, except in cases where only a proton or water is missing. In two supplemental columns the incomplete cases are differentiated into Missing Substrate and Missing Product.

For the compound name based comparison step all found synonyms were used as well as generated 'DAY-LIGHT names' (Chemical Information Systems, Inc. [19]). We applied a special conversion that removed most of the common sources of differences in equivalent compound names like hyphens, parentheses, etc. Most of the special characters, except '+' and apostrophe ('), were deleted. For identifying common reactions, all available synonyms and 'DAYLIGHT names' (see above) of the compounds are included in form of a link table containing assigned compound IDs. Where possible, KEGG glycan IDs (G number) were exchanged by their corresponding compound IDs (C number). Reactions with NAD(P)/H (BRENDA) and NADP/H_OR_NO_P (MetaCyc) were split into two reactions, one with NADH, the other with NADPH. The reaction ID of the form without phosphate was labelled as the original but with _*WOP* (= <u>WithOut Phosphate</u>) at the end.

Data download, storage, and comparison was realized by *C*++ as well as *Python* scripts and embedded *MySQL* statements. By executing a cron-job in regular time points, the information about metabolites, enzymes, reactions, *Molfiles*, and *InChIs* was downloaded from the source databases and so kept up to date automatically.

The access to the integrated database is possible *via* the link to *BKM-react* [20], Figure 1A, or *via* the *BRENDA* website, making use of the *BRENDA* query engine. Figure 1B illustrates the access to the integrated non-redundant reaction database [21] \rightarrow *Reaction & Specificity* \rightarrow *Biochemicals Reactions Aligned* (see arrow). Parameters for doing queries are presented in Figure 2A for the reaction table. Figure 2B shows an example for a query result. The downloadable content of the database consists of three tables, containing the compared reactions, the according compounds as well as a link table connecting both with each other.

Results and discussion

The combined database contains a unique list of reactions that occur in any of the compared databases *BRENDA* [1], *KEGG* [2], and *MetaCyc* [3] and the associations between equivalent reactions. Additionally these reactions are assigned to *KEGG* and *MetaCyc* pathways. Table 1 lists the data used for the comparison. The largest number of reactions originates from the *BRENDA* database, followed by *MetaCyc*, and *KEGG*.

A significantly improved matching of reactions was achieved by removing the compounds H^+ and water (H₂O) from the reactions before comparing them because the reactions in the databases are not always stoichiometrically balanced. The order of executing first the *InChI* comparison followed by the name comparison was chosen because identical synonyms may occur for different compounds. To rely on synonyms could therefore result in incorrect links. By using the reverse order more false positive matchings would appear.

One of the difficulties in the comparison consists in the - sometimes implied - stereochemistry not given in the compound name. Whereas cases like "alanine" being used for "*L*-alanine" are obviously to be expected, sometimes things become more complicated. For example, in *BRENDA* and *MetaCyc beta*-stereochemistry is implied for C5 of *D*-fructose-1,6-bisphosphate, being the major stereoisomer (see Figure 3A and 3B), the *KEGG* database includes in fact two different reactions, one with *beta*stereochemistry at C5, the other with undefined stereochemistry (see Figure 4A and 4B) where pathway information is only assigned to the reaction with the full stereochemistry. In general metabolites with complete stereochemistry are favored in *BKM-react*.

BKM-react online		BKM-read	ct online		TU Braunschweig Department of Bioinformatics & Biochemistry		
Home	1						
BKM-react online Query			for <u>B</u> RENDA- <u>K</u> EGG- <u>M</u> etaCyc- <u>rea</u>				
Download Database			e containing known enzyme-cata weig <u>EN</u> zyme <u>DA</u> tabase) [<u>1</u>], <u>K</u>				
teadme Documentation Table Architecture	integrated by alig	ning substrates and products.					
eam Profiles icense & Copyright eedback & Contact	BKM-react reaction comparisons were done by an <i>in silico</i> approach in which two steps, first a comparison of reactant structures using InChIs (linearized chemical structure descriptors) [4] and, second, a compound name comparison (incl. synonyms), were combined. After submitting an EC number or another attribute as substrate(s), product(s), or reaction ID of						
Release 2011.5 (May 2011)		ases, BKM-react online will re	trieve all results that match you				
А		of May 2011. The sources and ded the 11 th of May 2011.	e data from <i>BRENDA</i> release 201	11.1, <i>KEGG</i> (continually upd	ated), and <i>MetaCyc</i>		
~	Nucleic Acids R 2. Kanehisa M, Gott <u>Research 2009</u> 3. Caspi R, Altman Pujar A, Shearer <u>Acids Research</u> 4. Stein SE, Heller S	esearch 2011, 39:D670-D676. 5, Furumichi M, Tanabe M, Hirakawa M, 38:D355-D360. T, Dale JM, Dreher K, Fulcher CA, Gilham AG, Zhang P, Karp PD: The MetaCyc dat. 2009, 38:D473-D479. SR, Tchekhovski D: An open standard for	Rother M, Söhngen C, Stelzer M, Thiele J, Sc : <i>KEGG</i> for representation and analysis of me F, Kaipa P, Karthikeyan AS, Kothari A, Krum abase of metabolic pathways and enzymes a chemical structure representation - <i>The IUF</i>	olecular networks involving diseases a menacker M, Latendresse M, Mueller nd the <i>BioCyc</i> collection of pathway/g	ind drugs. <u>Nucleic Acids</u> LA, Paley S, Popescu L, enome databases. <u>Nucle</u>		
BRENDA home login history All enzymes	<u>Information Co</u>	nference Proceedings, Nimes, Franc The The	BRENDA		TU Braunschu Dept. of Bioinform		
Q Quick Search Fulltext Search Advanced Search	В	EC-Number Enzyme	Name Organism Protein Full tex Search Display 10 = entries New publications on BRENDA	Advanced Search			
TaxTree Explorer		Nomenclature	Reaction & Specificity	Functional Parameters			
EC Explorer Sequence Search Contology Explorer U Ontology Explorer Farameters SBML Output D Ownload		Enzyme Names EC Number Common/ Recommended Name Systematic Name Synonyms CAS Registry Number	Pathway Catalysed Reaction Reaction Type Natural Substrates and Products Substrates and Products Substrates Natural Substrate Products Natural Product	Km Value kcat/Km Value NEW Ki Value IC50 Value pl Value Turnover Number Specific Activity pH Optimum pH Range			
torial/Training		Isolation & Preparation	Inhibitors Cofactors	Temperature Optimum Temperature Range			
RENDA input opose new enzyme		Purification	Metals/lons Activating Compounds	Organism-related information			
roduction/References ntact and Impressum ws pyright		Cloned Expression NEW Renatured Crystallization	Ligands Ligand Views Biochemicals Reactions Aligned NEW	Organism Source Tissue Localization Protein-Specific Search			
elated Links Ip		Stability	Enzyme Structure	Disease & References			
knowledgements RENDA Professional		pH Stability Temperature Stability General Stability	Sequence/ SwissProt link	Disease/ Diagnostics References			
ommercial Version Release 2011.1 (January 2011)			3D-Structure/ PDB link Molecular Weight	Application & Engineering			
		Organic Solvent Stability Oxidation Stability	Subunits Posttranslational Modification	Engineering			
		Storage Stability	Webmaster: Maurice Scheer	Application			
			n,scheer@tu-bs,de		t i i i i i i i i i i i i i i i i i i i		

If no structural information is available, reactions are allowed to match by name comparison.

This example shows a general problem in biochemical compound name comparison. The large majority of biochemists refer to *S*-alanine just by the name alanine although the name in principle is ambiguous or should be used for the racemate. In most cases we assume that for the standard amino acids the name without stereo-descriptor implicitly means S- (or L-, respectively). This holds true also for some other compound names where the stereo-descriptor is implicitly given. A related problem occurs at positions where the

Braunschweig BKM-react online				BKI	M-read	ct onli	ne		TU Braunschweig Department of Bioinformatics Biochemistry
Home	Search EC numbe	er in the	integrated reactior	collection					
KM-react online Query			ontains v Sear	ch show 10	v results				
adme	Reaction ID BF			contains	v				
cumentation ble Architecture	Reaction ID			contains	~				
am Profiles ense & Copyright	Reaction ID M	letaCyc		contains	v			-	
edback & Contact Release 2011.5 (May 2011)		bstrates		contains	~			4	
	KEGG Pathwa	Products		contains	v				
	MetaCyc Pathwa			contains contains	v v				
		niometry		contains	v				
	Missing Su	ubstrate		contains					
	Missing			contains	v				
	Commentary	_		contains	v				
	Commentary M	letaCyc		contains	V				
Technische Universität Branschweig				BKI	M-react	online			TU Braunschweig Department of Bioinformatics & Biochemistry
BKM-react online	arch EC number in the integ	rated react	on collection						-
Freact online Query	contain	IS V	Search show 10 🔻	results					
niced Database 🔹 🐨	Reaction ID BRENDA		contains V						
Architecture	Reaction ID KEGG		contains V						
Profiles se & Copyright back & Contact	Reaction ID MetaCyc		contains V						
			contains 🛛 🔻						
Delawar, 2011 5 (May 2011)		•						В	
Rolease 2011.5 (May 2011)	Products	;	contains V contains V				I	В	
Release 2011.5 (May 2011)	Products KEGG Pathway Name MetaCyc Pathway Name	3					I	8	
Rolease 2011.5 (May 2011)	Products KEGG Pathway Name MetaCyc Pathway Name Stoichiometry	; 						В	
Release 2011.5 (May 2011)	Products KEGG Pathway Name MetaCyc Pathway Name Stoichiometry	2	contains V contains V contains V contains V				I	3	
Release 2011.5 (May 2011)	Products KEGG Pathway Name MetaCyc Pathway Name Stoichiometry Missing Substrate Missing Product Commentary KEGG	2	contains v				I	3	
Reference 2011.5 (May 2011)	Products KEGG Pathway Name KEGG Pathway Name Stolchiometry Missing Substrate Kissing Product Commentary KEGG Commentary MetaCyc		contains V				I	8	
Resser 2011 5 (Mry 2011)	Products KEGG Pathway Name KEGG Pathway Name Stoichiometry Missing Substrate Missing Product Commentary KEGG Commentary KEGG amino acid sequences lar-computer 1-0 off		contains V	alyzed reaction				В	
Release 2011.5 (May 2011)	Products KEGG Pathway Name MetaCyc Pathway Name Stoichiometry Missing Substrate Missing Product Commentary KEGG Commentary KeGG amino acid sequences D= com		contains V	alyzed reaction				В	
Releases 2011.5 (May 2011)	Products KEGG Pathway Name MetaCyc Pathway Name Siculionenty Missing Substrate Missing Product Commentary KEGG Commentary KEGG and Commentary MetaCyc antino acid aceuences D = com uta 1 - 10 of 0	prehensive or	contains V	alyzed reaction	Products	KEGC Pathway Name			Commentary KetaCye
Reference 2011 5 (May 2011) ♀ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓	Products KEGG Pathway Name MetaCyc Pathway Name Stoichionetry Missing Stokstate Missing Product Commentary KEGG Commentary KEGG amino acid sequences L= computed 1:0 of elementary and the computed of elementary	prehensive or Reaction ID KEGG	contains ▼		Products 2 CO2 + H2O2	Name Glyoxylate and dicarboxylate metabolism		, Missing Missing	Contains Mn2+ as a cofactor. The enzyme not a flavoprotein as had been thought [Cl [10493928]].
Resser 2011 5 (May 2011)	Products KEGG Pathway Name KEGG Pathway Name MetaCyc Pathway Name Sibichiometry Missing Substrate Missing Product Commentary KEGG Commentary MetaCyc annino acid sequences 1= com dist 1 = 0 of metaTota acid sequences 1= com Commentary MetaCyc Table Commentary MetaCyc Tabl	Prehensive or Reaction ID KEGG 1 R00273	contains V contains V	Substrates	2 CO2 + H2O2 formate + CO2	Name Glyoxylate and dicarboxylate metabolism Glyoxylate and dicarboxylate metabolism	MetaCyc Pathway Stoichli Nane oxalate degradation		Contains Mn2+ as a cofactor. The enzyme not a flavoprotein as had been thought [Cf [10493928]]. The enzyme from Bacillus subbilis containt
Release 2011 5 (May 2011)	Products KEGG Pathway Name MetaCyc Pathway Name Solichionery Missing Stubistrate Missing Product Commentary KEGG Commentary KEGG commentary KEGG commentary KEGG mithio acid sequences 2= computed ist 1=10 of 0 mithio acid sequences 2= computed Registrate acids use acids	Prehensive or Reaction ID KEGG 1 R00273	Contains V Contains Contains V Contains Contains Co	Substrates oxalate + O2	2 CO2 + H2O2	Name Glyoxylate and dicarboxylate metabolism Glyoxylate and dicarboxylate metabolism Glyoxylate and dicarboxylate metabolism	MetaCyc Pathway Stoichi Name oxalate degradation V V		Contains Mn2+ as a cofactor. The enzyme not a flavoprotein as had been thought [Cl [10493928]]. The enzyme from Bacillus subfilis contain: manganese and requires 02 for activity.
Resson 2011 5 (May 2011)	Products KEGG Pathway Name MetaCyc Pathway Name Subichiometry Mussing Substrate Missing Product Commentary KEGG Commentary KEGG Commentary KEGG Commentary MetaCyc asta 1-0 of Resconse	Reaction ID REGG R00273 R00522	Contains V Contains Contains V Contains Contains Co	Substrates oxalate + O2 oxalate ATP + oxalate +	2 CO2 + H2O2 formate + CO2 AMP + diphosphate +	Name Glyoxylate and dicarboxylate metabolism Glyoxylate and dicarboxylate metabolism Glyoxylate and dicarboxylate	MetaCyc Pathway Stoichir Name oxalate degradation oxalate degradation		Contains Mu2+ as a colactor. The enzyme not if incomparise in as had been hough (C) UM4930201018 sub181 contains the enzyme from Boselius sub181 contains manganese and requires 02 for activity, e though there is no net redox change.
Release 2011 5 (May 2011)	Products KEGG Pathway Name MetaCy: Pathway Name Stoichiometry Missing Stoichitete Missing Product Commentary KEGG Commentary KEGG Commentary KEGG Commentary KEGG Missing Product Commentary KEGG Missing Product Ramino acid sequences 1= comp Bitth 1:10 of Generation Minit 1:10 of Generation Minit 1:10 of Generation Minit 2:3.4 BI22430 BS189324 MD1:2:3.4 BI20420 BS189324	Reaction ID KEGG R00273 R00522 R01558	Contains V Contains V Contai	Substrates oxalate + O2 oxalate ATP + oxalate + CoA succinyl-CoA +	2 CO2 + H2O2 formate + CO2 AMP + diphosphate + oxalyI-CoA succinate +	Name Glyoxylate and dicarboxylate metabolism Glyoxylate and dicarboxylate metabolism Glyoxylate and dicarboxylate metabolism	MetaCyc Pathway Name Stoichi V oxalate degradation V oxalate degradation V oxalate degradation		Contains Mn2+ as a cofactor. The enzyme not a flavoprotein as had been thought [CI [10493928]]. The enzyme from Bacillus subfilis contains manganese and requires O2 for activity, e

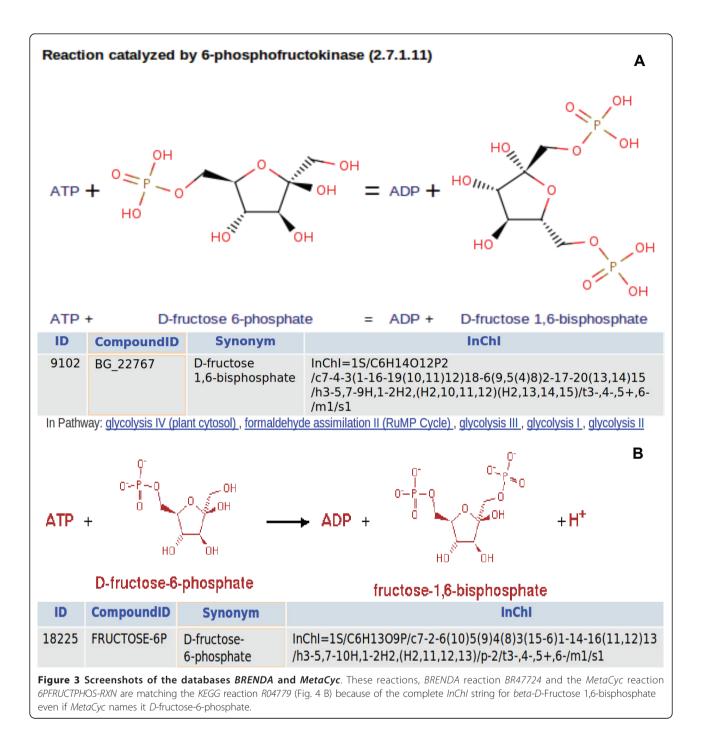
stereochemistry is ambiguous like in the case of C1 of D-glucose. In some cases the stereochemistry for this position is undefined in the *Molfiles* [17], in others the more stable form (*e.g. beta* in the case of glucose) is used and defined.

Although all three databases offer their own *InChIs*, they are not directly comparable because *KEGG* uses the non-standard form of an *InChI*, whereby *MetaCyc* and *BRENDA* use the standard *InChI* format. So for a standardized comparison it is necessary to use self-generated

Table 1 Overview reaction sources and data

	Different EC numbers	Incomplete EC numbers	Reaction IDs	Reaction IDs without EC number	Compound IDs	Synonyms	Molfiles	InChIs
KEGG	3,761	122	8,452	1,288	6,522	11,597	6,327	5,416
MetaCyc	4,159	138	9,343	2,236	6,095	19,707	6,035	5,782
BRENDA	4,425	207	10,109	55	9,750	20,922	9,750	5,242

The number of reactions in BRENDA is in fact close to 180,000. In this case only complete reactions with natural substrates were included.



InChIs based on *Molfiles*. For this purpose the official *IUPAC* converting tool was utilized [18]. A higher matching rate was achieved by using only essential layers (see *Methods* section) of an *InChI* string. A drawback is that not for each compound an *InChI* is available, *e.g.* for macromolecular reactants or for generic compounds.

A pairwise comparison of reactions revealed a high identity between *KEGG* &*MetaCyc*. About 50% reactions were equal, out of which most were also found in

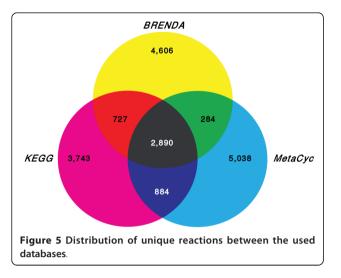
BRENDA (Figure 5). Between *MetaCyc* &*BRENDA* 3,174 reactions were identified to be equal. Comparing *KEGG* &*BRENDA*, even more reactions (3,617) could be assigned to each other.

Table 2 shows the assignment of diverse reactions between the databases which are equal. There are examples of reactions that have a 1:n relation because of redundant reactions occurring within the same database. In *KEGG* for example, metabolites are differentiated into

Entry	R00756 Reaction A					
Name	ATP:D-fructose-6-phosphate 1-phosphotransferase					
Definition	ATP + D-Fructose 6-phosphate <=> ADP + D-Fructose					
	1,6-bisphosphate					
Equation	C00002 + C00085 <=> C00008 + C00354					
	$\begin{array}{c} \begin{array}{c} 0\\ HO-P-O-P-O-P-O-P-O-P-O-P-O-P-O-P-O-P-O-P$					
RPair	RP00003 C00002_C00008 main RP00052 C00085_C00354 main RP06297 C00002 C00354 trans					
Enzyme	2.7.1.11					
Compoun	dID Synonym InChl					
C00354	D-Fructose InChI=1S/C6H14O12P2 1,6-bisphosphate /c7-4-3(1-16-19(10,11)12)18-6(9,5(4)8)2-17-20(13,14)15 /h3-5,7-9H,1-2H2,(H2,10,11,12)(H2,13,14,15) /t3-,4-,5+,6?/m1/s1					
Entry	R04779 Reaction B					
Name	ATP:D-fructose-6-phosphate 1-phosphotransferase					
Definition	ATP + beta-D-Fructose 6-phosphate <=> ADP + beta-D-Fructose 1,6-bisphosphate					
Equation	C00002 + C05345 <=> C00008 + C05378					
	$\begin{array}{c} 0 \\ HO - P - O \\ OH \\ COS346 \end{array} \xrightarrow{0} P \\ HO - P - O \\ HO $					
RPair	RP00003 C00002_C00008 main RP00191 C05345_C05378 main RP10335 C00002 C05378 trans					
Enzyme	2.7.1.11					
Pathway	<pre>rn00010 Glycolysis / Gluconeogenesis rn00030 Pentose phosphate pathway rn00051 Fructose and mannose metabolism rn00680 Methane metabolism rn01100 Metabolic pathways rn01110 Biosynthesis of secondary metabolites rn01120 Microbial metabolism in diverse environments</pre>					
Orthology	K00850 6-phosphofructokinase [EC:2.7.1.11]					
Compound	dID Synonym InChl					
C05378	C05378 beta-D-Fructose InChI=1S/C6H14O12P2 1,6-bisphosphate /c7-4-3(1-16-19(10,11)12)18-6(9,5(4)8)2-17-20(13,14)1 /h3-5,7-9H,1-2H2,(H2,10,11,12)(H2,13,14,15) /t3-,4-,5+,6-/m1/s1					
	ts of the database <i>KEGG</i> . <i>KEGG R00756</i> and <i>R04779</i> . The second reaction is the preferred one. <i>C04779</i> possess and is therefore matched with the more complete described metabolites of the other databases.					

glycans and compounds, respectively. This means that identical compounds may get two different IDs, starting with G and C. This results in reactions with different reaction IDs (no. 3 in Tab. 2). Sometimes there are

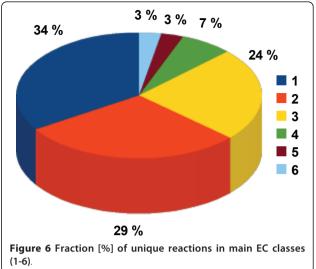
synonyms or keto-enol tautomers which describe one reaction in various forms (no. 1 in Tab. 2) or other alternative writing styles (no. 2 in Tab. 2). Further *KEGG* uses one reaction-ID for the same reaction being catalysed by



enzymes with different *EC* numbers, whereas *MetaCyc* often uses different reaction IDs in such cases (no. 1 in Tab. 2).

In Figure 5 the distribution of equal reactions occurring in any of the three databases is illustrated. 2,890 of all reactions are contained in all three databases, corresponding to 34% of all *KEGG* reactions, 31% of all *MetaCyc* reactions, and 29% of the included *BRENDA* reactions, respectively. In the present version of the data set, 3,743 *KEGG* reactions, 5,038 *MetaCyc* reactions, and 4,606 *BRENDA* reactions occur only in the respective database (Figure 5). Altogether the non-redundant reactions and 20,358 *EC*/reaction combinations as some reactions are catalyzed by a number of different enzymes.

In Figure 6 the fraction of all unique reactions belonging to the six main *EC* classes is shown. The largest fractions belong to *EC* classes 1 and 2, followed by class 3. Statistical data about the *EC* numbers occurring in the



non-redundant reaction database are given in Table 3. Additionally to all EC numbers, complete and incomplete, the latter ones are listed separately. Furthermore it is distinguished between EC numbers representing at least one single reaction or more than one. A detailed look on the EC numbers with the highest number of reactions is given in Table 4 together with the number of reactions.

The only database with a similar goal is *BioSilico* [6]. One important difference consists of the fact that the assignment of identical reactions in our database is done by an actual comparison of the compounds structure in combination with synonyms whereas in *BioSilico*, the matching is only a simple assignment of reactions having the same *EC* number without redundancy check.

The number of reactions in the database described in this paper is far beyond that in *BioSilico*. Selecting three *EC* numbers by chance resulted in *e.g. EC* number $1.14.14.1 \rightarrow 4$ reactions in *BioSilico vs.* 116 reactions in

No.	KEGG	MetaCyc	BRENDA	Definition			
1	R04915			Quinoline-3,4-diol + Oxygen <=> Formylanthranilate + CO			
	R05719			3-Hydroxy-1H-quinolin-4-one + Oxygen <=> Formylanthranilate + CO			
		1.13.11.47-RXN		3-hydroxy-1H-quinolin-4-one + oxygen = carbon monoxide + N-formylanthranilate			
			BR22597	3-hydroxy-1H-quinolin- 4 -one + O2 = N- formylanthranilate + CO			
2	R00004			Diphosphate + H2O <=> 2 Orthophosphate			
		INORGPYROPHOSPHAT-RXN		diphosphate + $H_2O = 2$ phosphate + H^+			
			BR22749	diphosphate + $H2O = 2$ phosphate			
3	R00010			alpha, alpha-Trehalose + H2O <=> 2 D-Glucose (<i>C01083</i>)			
	R06103			Trehalose + H2O <=> 2 D-Glucose (G00293)			
		TREHALA-RXN		trehalose + H2O \rightarrow 2 β -D-glucose			
			BR15991	alpha, alpha-trehalose + $H2O = 2 D$ -glucose			
			BS370856	alpha, alpha-trehalose + H2O = beta-Dglucose			

Table 2 Some instructive cases of different forms for identical reactions

Table 3 Statistics about *EC* numbers occurring in the integrated non-redundant reaction database

EC numbers	Different EC numbers	Incomplete EC numbers
in total	4,288	365
with > 1 reaction	2,681	185
with > 5 reactions	561	73
with > 10 reactions	184	49

our reaction database, *EC* number $2.1.1.103 \rightarrow 1$ reaction in *BioSilico vs.* 4 reactions in our database, $3.1.1.47 \rightarrow 1$ reaction in *BioSilico vs.* 12 reactions in our database. The fact that in these examples not even all available *KEGG* reactions were found in *BioSilico* indicates that this database is not updated.

Additionally, our reaction database contains the information whether a reaction is stoichiometric incomplete or not. This test is performed before the removal of H⁺ and H₂O. Non-balanced reactions are labeled in a separate table column. 2,803 out of 18,172 reactions are at present in this category. The labeling allows modelers to select only balanced reactions for the reconstruction of organism-specific models and networks.

The tool of Radrich *et al.* [10] also includes a stoichiometric evaluation. Their method includes a name comparison where they compare the similarity of compound names. Further they use *SMILES* strings for a structural comparison. The tool was executed only for *Arabidopsis*

Conclusions

In this work we present an integrated and non-redundant reaction database implementing a combined approach of structure and name based comparison.

The tool, integrated into the *BRENDA* [1] query engine but not confined to *BRENDA* data is offering a novel valuable tool that can be used *e.g.* for the construction of biological models. The resulting models will be much more complete than if only one of the databases is used as the three biological databases *BRENDA*, *KEGG* [2], and *MetaCyc* [3] complement each other. Regular 6monthly updates of this database will make guarantee actuality.

Availability and requirements

The integrated and non-redundant reaction database is accessible *via BKM-react* [20] and the website of the *BRENDA* [1] database: *BRENDA* website [21] \rightarrow *Reaction & Specificity* \rightarrow *Biochemicals Reactions Aligned* (Figure 1). The complete dataset is additionally provided as a CSV-formatted download at the same site. Available is a reaction table, a table with all compounds occurring in the reactions, and an assignment table with the linkage between reactions and compounds.

EC number	Enzyme	Number of reactions		
1.14.14.1	unspecific monooxygenase	116		
2.4.1.17	1.17 glucuronosyltransferase			
3.2.1.21	beta-glucosidase	74		
1.1.1.100	3-oxoacyl-[acyl-carrier-protein] reductase	55		
3.5.1.4	amidase	46		
3.6.3.44	xenobiotic-transporting ATPase	46		
3.1.3.16	phosphoprotein phosphatase	44		
1.3.1.10	enoyl-[acyl-carrier-protein] reductase (NADPH, B-specific)	43		
3.2.1.1	alpha-amylase	43		
1.1.1.50	3alpha-hydroxysteroid dehydrogenase (B-specific)	42		
2.3.1.41	beta-ketoacyl-acyl-carrier-protein synthase l	41		
3.6.1.9	nucleotide diphosphatase	39		
1.1.1.1	alcohol dehydrogenase	37		
2.3.1.86	fatty-acyl-CoA synthase	37		
1.14.13.72	methylsterol monooxygenase	36		
1.2.1.3	aldehyde dehydrogenase (NAD ⁺)	34		
1.2.1.5	aldehyde dehydrogenase [NAD(P) ⁺]	33		
3.2.1.24	1.24 <i>alpha-</i> mannosidase			
3.2.1.51	alpha-L-fucosidase	33		
.14.13.8	flavin-containing monooxygenase	32		
1.4.3.3	D-amino-acid oxidase	32		

Recommended names of enzymes: source BRENDA database.

List of abbreviations used

BRENDA: BRaunschweig ENzyme DAtabase; EC: Enzyme Commission; InChI: IUPAC International Chemical Identifier; IUBMB: International Union of Biochemistry and Molecular Biology; IUPAC: International Union of Pure and Applied Chemistry, KEGG: Kyoto Encyclopedia of Genes and Genomes; SMILES: Simplified Molecular Input Line Entry System.

Acknowledgements and funding

The authors are grateful to Ron Caspi from *MetaCyc* for help with the implementation of the *MetaCyc* data, Adam Podstawka for technical support, Maurice Scheer for implementing the webinterface, and René Rex for providing a stoichiometry verification tool. Financial support: *European Union (FELICS* [22], *SLING* [23]).

Authors' contributions

ML and MS executed the data acquisition and implemented the reaction comparison. ML and MS were involved in the construction of the integrated reaction database and the scientific evaluation. DS had the idea to develop the reaction database and supervised the development. ML, MS, and DS wrote the manuscript. All authors read and approved the final manuscript.

Received: 10 January 2011 Accepted: 8 August 2011 Published: 8 August 2011

References

- 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 Research 2011, 39:D670-D676.
- Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Research 2009, 38:D355-D360.
- Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Paley S, Popescu L, Pujar A, Shearer AG, Zhang P, Karp PD: The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research* 2009, 38:D473-D479.
- Stein LD: Integrating biological databases. Nat Rev Genet 2003, 4:337-345.
 Krull M. Vars N. Chai, C. Distar S. Database, A. Wienender, F. TRANSPATING and S. S. Stein S. Ste
- Krull M, Voss N, Choi C, Pistor S, Potapov A, Wingender E: TRANSPATH[®]: an integrated database on signal transduction and a tool for array analysis. *Nucleic Acids Research* 2003, 31:97-100.
- Hou BK, Kim JS, Jun JH, Lee D-Y, Kim YW, Chae S, Roh M, In Y-H, Lee SY: BioSilico: an integrated metabolic database system. *Bioinformatics* 2004, 20:3270-3272.
- 7. ExPASy ENZYME. [http://www.expasy.org/enzyme/].
- EcoCyc: Encyclopedia of Escherichia coli K-12 Genes and Metabolism. [http://ecocyc.org/].
- 9. BioCyc Home. [http://biocyc.org/].
- Radrich K, Tsuruoka Y, Dobson P, Gevorgyan A, Swainston N, Baart G, Schwartz J-M: Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. *BMC Systems Biology* 2010, 4:114.
- 11. TAIR AraCyc. [http://www.arabidopsis.org/biocyc/index.jsp].
- Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, Kanapin A, Lewis S, Mahajan S, May B, Schmidt E, Vastrik I, Wu G, Birney E, Stein L, D'Eustachio P: Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Research* 2009, **37**:D619-D622.
- Rhea Annotated reactions database. [http://www.ebi.ac.uk/rhea//home. xhtml#].
- 14. IUBMB Nomenclature Home Page. [http://www.chem.qmul.ac.uk/iubmb/].
- 15. Michal G, Schomburg D: Biochemical Pathways: An Atlas of Biochemistry and Molecular. *Biology*. 2 edition. Wiley-Interscience; 2011.
- Stein SE, Heller SR, Tchekhovski D: An open standard for chemical structure representation-the IUPAC chemical identifier. Nimes International Chemical Information Conference Proceedings, Nimes, France 2003. 131-143.
- Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, Laufer J: Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *Journal of Chemical Information and Computer Sciences* 1992, 32:244-255.
- International Union of Pure and Applied Chemistry. [http://www.iupac. org/inchi/download/index.html].

- 19. Daylight. [http://www.daylight.com/].
- 20. BKM-react online. [http://bkm-react.tu-bs.de/].
- 21. Enzyme Database BRENDA. [http://www.brenda-enzymes.org/].
- 22. FELICS | FELICS. [http://www.felics.org].
- 23. SLING | sling-fp7.org. [http://www.sling-fp7.org/].

doi:10.1186/1471-2091-12-42

Cite this article as: Lang *et al.*: *BKM-react*, an integrated biochemical reaction database. *BMC Biochemistry* 2011 **12**:42.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

) BioMed Central

Submit your manuscript at www.biomedcentral.com/submit