Formal TCA cycle description based on elementary actions

Pierre Mazière[1,4], Nicolas Parisey[2,4], Marie Beurton-Aimar[2,3,*] and Franck Molina[1,**]

[1]*Center for Pharmacology and Health Biotechnology, UMR 5160, Faculté de Pharmacie, 15 Av. Charles Flahault, BP 14491, 34093 Montpellier cedex 5, France*

[2]*Laboratoire de Physiopathologie mitochondriale, Université Bordeaux 2 – 146 rue Léo Saignat, 33076 Bordeaux, France*

[3]*LaBRI UMR 5800, Université Bordeaux 1 – 350 Avenue de la libération, 33405 Bordeaux, France*

[4]*Both authors worked equally on this paper*

[*,**]*Corresponding authors (Email, *marie.aimar@phys-mito.u-bordeaux2.fr; **franck.molina@cpbs.univ-montp1.fr)*

Many databases propose their own structure and format to provide data describing biological processes. This heterogeneity contributes to the difficulty of large systematic and automatic functional comparisons. To overcome these problems, we have used the BioΨ formal description scheme which allows multi-level representations of biological process information. Applied to the description of the tricarboxylic acid cycle (TCA), we show that BioΨ allows the formal integration of functional information existing in current databases and make them available for further automated analysis. In addition such a formal TCA cycle process description leads to a more accurate biological process annotation which takes in account the biological context. This enables us to perform an automated comparison of the TCA cycles for seven different species based on processes rather than protein sequences. From current databases, BioΨ is able to unravel information that are already known by the biologists but are not available for automated analysis tools and simulation software, because of the lack of formal process descriptions. This use of the BioΨ description scheme to describe the TCA cycle was a key step of the MitoScop project that aims to describe and simulate mitochondrial metabolism *in silico*.

[Mazière P, Parisey N, Beurton-Aimar M and Molina F 2006 Formal TCA cycle description based on elementary actions; *J. Biosci.* **32** 145–155]

1. Introduction

Biological data and controlled vocabularies used to describe biological functions of molecular entities are as numerous as they are heterogeneous. This diversity contributes to the lack of a unified method accurate enough to represent this complex concept in biology (Rison *et al* 2000). Currently, the function descriptions of genes or gene products are available from several sources. From scientific literature to annotation in curated databases, these descriptions use different formats and standards that allow various levels of precision and completeness (Ouzounis *et al* 2003). Molecular processes and cellular processes are the two main types of processes described. The former are directly performed by molecular entities (e.g. kinase, peptidase or simple binding) whereas the latter are the integration of several molecular processes acting at the cellular level (e.g. apoptosis, transcription or glucose biosynthesis). The use of classifications and/or ontologies to achieve these descriptions is not sufficient to provide a description compatible with sophisticated process comparison, modelling and simulation. Consequently, functional pathway comparison and pathway alignment are difficult to assess, and require the extra complication of biochemical and genomic data integration (Dandekar *et al* 1999).

BioΨ is a four-level description scheme that strongly structures functional data by encapsulating and embedding

Keywords. BioPsi; formal description; MitoScop; process comparison; TCA cycle

Abbreviations used: α-KDH, α-ketoglutarate dehydrogenase; BAs, biological activities; BEAs, basic elements of action; BFs, biological functionalities; BRs, biological roles; FADH2, reduced flavin adenine dinucleotide; GO, Gene Ontology; NADH, reduced nicotinamide adenine dinucleotide; PDH, pyruvate dehydrogenase; SMILES, simplified molecular input line entry system; SMIRKS, SMIles ReaKtion Specification.

information from bottom to top (Maziere *et al* 2004). Compared to the two main types of processes cited above, BioΨ introduces two supplementary precision levels to describe biological processes. The first is composed of a limited set of elementary actions which formalizes pseudo-chemical transformations involved in molecular processes. The second describes how molecular processes are related to each other given the structural constraints imposed by molecular entities that perform them. Within the context of the mitochondrial metabolism description and simulation project MitoScop, BioΨ has been used to annotate the well-known TCA cycle from seven different species, demonstrating the advantages of such a formalized description scheme.

1.1 *Existing tools to describe biochemical pathways*

Natural language has been the main method used to describe biological processes. However, the lack of standardization of these descriptions led to the introduction of new kinds of annotation. These new description methods can be reduced to two denominations: classifications and ontologies [enzymes (Fleischmann *et al* 2004), protein kinases (Hanks and Quinn, 1991; Krupa *et al* 2004), transporters (Busch and Saier 2003), EcoCyc (Karp *et al* 2000), MIPS (Mewes *et al* 2004), MPW (Selkov *et al* 1998)]. Classifications are designed using tree structures that describe processes hierarchically based on their properties. They are mainly used to describe molecular processes. Ontologies use a different approach to organize information: they are composed of a set of terms rigorously defined, and build relationships between these terms (Soldatova and King 2005). Contrary to classifications where, usually, only leaves of a tree are used to annotate biological functions, all the terms of ontologies can be used for annotation. Both ontologies and classifications often cover specific species, or particular processes. Therefore, the description of biological processes occurring in a different cell types and the comparison of processes between species require the use of several functional schemes that may be incompatible (Rison *et al* 2000). GeneOntology (GO) (Harris *et al* 2004) is becoming the *de facto* ontology standard in the field of biological process description. It is distinguishable from other ontologies on three aspects: (i) process descriptions independent from the molecular entities that perform them; (ii) clear distinction between molecular processes (i.e. molecular functions) and cellular processes (i.e. biological processes); (iii) terms are structured in a directed acyclic graph (DAG) which allows them to have several parents using two kinds of link: "is a", and "is part of ".

This last point conveys more biological meaning to GO annotations than can be found in other schemes. However, as also observed in these other schemes, the use of GO biological process descriptions is limited to a convenient way of tagging molecules with standardized terms. Although carrying a useful semantic value, the two types of connection are limited to the linking of terms describing the same kind of processes: there are no possible cross-link between molecular functions and biological processes. For example, the annotation of a molecule using the GO terms "kinase activity" (GO:0016301) and "regulation of transcription" (GO:0045449) does not indicate how the kinase activity regulates the transcription. Consequently, biologists are still missing a scheme for biological process description that allows detailed and formalized annotation at different level of abstraction.

1.2 *BioΨ: a biological process description scheme based on elements of action*

Biological molecules are involved in different kinds of biological functions. Their environment modifications related to these functions are the consequence of enzymatic and non-enzymatic biological processes. To develop a standardized annotation scheme compatible with the different ways to describe function in biology, it was necessary to dissociate three key properties: the physical molecular entity, the actions it performs (hereafter called biological processes), and the context in which these actions are performed (Maziere *et al* 2004). Structural and functional organization of biological molecules can then be described independently of all known possible biological processes. Despite their number and their heterogeneity, these biological processes can be viewed as a combination of simpler biological processes which can in turn be described as a combination of identifiable elementary actions. On the basis of this observation, the BioΨ description scheme (Maziere *et al* 2004) defines four abstraction levels: basic elements of action (BEAs), biological activities (BAs), biological functionalities (BFs) and biological roles (BRs). The BA and BR levels can be compared to the usual molecular processes and cellular processes respectively: they describe processes performed by molecule subparts (e.g. protein functional domains) for one, and processes occuring at the level of biochemical pathways for the other. Therefore, they can be mapped to molecular functions and biological processes as defined by GO. BEAs and BFs constitute two original levels of biological process description. The former are related to the key sets of atom of a molecule that are involved in the described process, whereas the latter are associated with full molecular entities. Associated with each of these levels, formalized biological constraints help to define the biological context required for accurate biological process descriptions.

Compared to existing description schemes such as the enzyme commission classification or GO, BioΨ does not rely on natural language to define biological

processes. Hence, it is not limited to a simple designation of processes by using tags. By using a formalized method for representing molecules and reactions, BioΨ allows a full description of molecular transformations occuring during the described process. The SMILES (simplified molecular input line entry system) formalism and its SMIRKS (SMIles ReaKtion Specification) extension were designed to describe molecule chemical structures and molecular transformations respectively, through a linear textual representation (Weininger 1998). Simple writing rules allow generic molecular transformations to be treated by computer analysis software, based on freely available toolkits and specifications: given the SMILES representation of substrates and the SMIRKS representation of a generic transformation based on chemical motifs, these software are able to compute the possible output products. The first level of BioΨ, BEAs, uses this SMIRKS representation to describe a hundred of elementary chemical transformations, sufficient to rebuild every enzymatic and non-enzymatic processes currently known. These processes are described at the BA level by combining BEAs. The processes described at the BF level combine BAs that are performed by a molecule, or a molecular complex, given specific constraints. Finally, processes described at the BR level are a combination of BFs in a cellular context. BioΨ embeds each of its levels within the immediately superior level, i.e. BEAs are embedded in BAs, BAs in BFs, and BFs in BRs. Consequently, the relations between biochemical reactions and a given cellular process in which they are involved can easily be extracted and interpreted within the context associated with this process.

2. Results

2.1 *Gathering the current annotation of the TCA cycle*

The TCA cycle was chosen as an entry point for the more global objective set by the MitoScop project: namely, to annotate precisely the mitochondrial metabolism for further system simulations. In most eukaryotes, the main source of ATP used to maintain homeostasis is produced by the oxidation of pyruvate in the TCA cycle. During this oxidation process, reduced nicotinamide adenine dinucleotide (NADH) and reduced flavin adenine dinucleotide (FADH2) are generated. NADH and FADH2 are mainly used to drive the process of oxidative phosphorylation, which is responsible for converting the reducing potential of NADH and FADH2 into the high energy phosphate of ATP (Saraste 1999).

The TCA cycle is a very well-studied metabolic network occuring specifically in mitochondria. It can be represented by a chain of 14 enzymatic reactions. In figure 1, partly inspired from the reference pathway found in KEGG (Ogata *et al* 1999), the TCA cycle enzymatic processes are represented by their EC number and surrounded by their substrates and products. Several steps, such as the transformation of oxaloacetate into citrate, or succinyl-coA into succinate, can be performed by following different paths depending on species and implicated compartments. They are represented with the different enzymatic activities corresponding to these paths.

One can note, that using this KEGG-inspired representation, mitochondrial and cytosolic enzymes are mixed. Mitochondrial transporters that could support the necessary metabolite exchanges exist, but are ignored and not annotated in the KEGG representation. Compared to the KEGG reference pathway, figure 1 describes a more detailled representation of the enzymatic activities involved in the pyruvate dehydrogenase (PDH) complex.

Besides this reference pathway, the KEGG database also contains descriptions of the TCA cycle pathway for several species. The species specific enzyme descriptions are often incomplete. Hence, data from several databases had to be gathered and integrated using a formal description independent of the original database formats.

Using this generic pathway (figure 1), we gathered annotation from different sources to describe the TCA cycle from: *Homo sapiens*, *Drosophila melanogaster*, *Cænorhabditis elegans*, *Saccharomyces cerevisiæ*, *Schizosaccharomyces pombe*, *Mus musculus* and *Rattus norvegicus*. Figure 2 shows a comparative table of these metabolic pathways. For each column, the metabolite is the substrate of the enzyme which produces the next column's metabolite. For instance, for the first species (*Homo sapiens*), acetyl CoA and oxaloacetate are the substrates for citrate synthase, which produces citrate. Some enzymes are missing in the KEGG descriptions. By searching for the missing information in BRENDA (Schomburg *et al* 2004) and Swiss-Prot (Bairoch *et al* 2004), we were able to annotate 13 reactions performed by the two dehydrogenase complexes, the aconitase, the succinate thiokinase, the succinate dehydrogenase and the fumarase. Some species are still missing annotations for several reactions. Although the TCA cycle is a well-known metabolic pathway, dispersion and heterogeneity of annotation make difficult the species comparison on the basis of their processes.

2.2 *BioΨ formal multi-scale description of TCA cycle*

The MitoScop project aims at building a mitochondria metabolism model to compare *in silico* different instances of the model for different species. BioΨ has been used to integrate the TCA cycle-related knowledge of the seven species previously listed. The full BioΨ descriptions given in supplementary materials (*http://dept-info.labri-fr/~parisey/mitoscop*) use the annotations gathered as

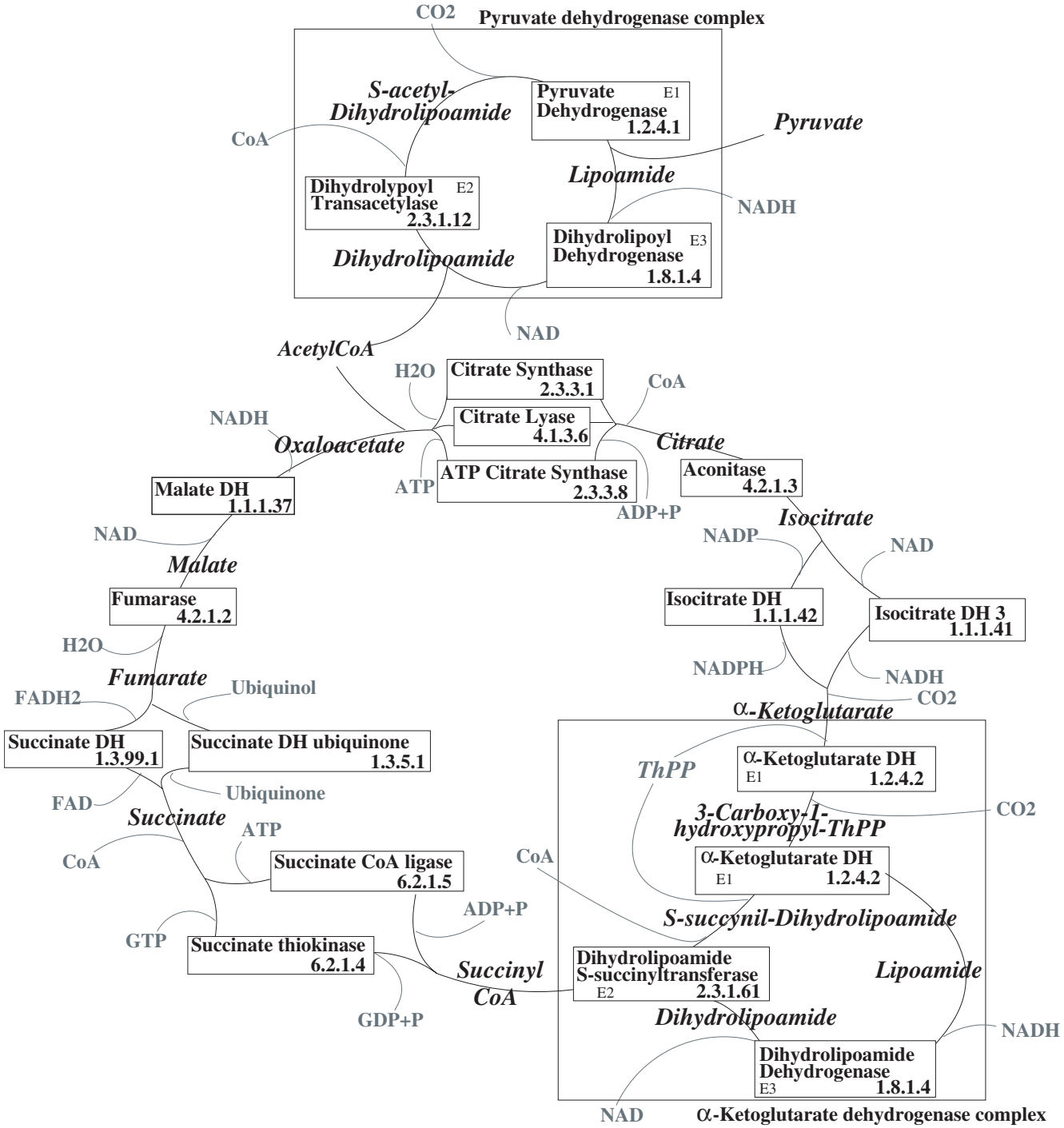

Figure 1. Reference TCA cycle.

described above and formalized through several steps that define the composition of processes from BAs to BR.

2.2a *Biological activities composed of BEAs:* BEAs are divided into 4 main classes (Maziere *et al* 2004). To describe the TCA cycle, we have used 21 BEAs among the set of 97 defined by BioΨ (see supplementary materials). In figure 3A, the BEA Ba:CS.2 describes a chemical group transfer acting on a carbon-sulphur bond. Associated with chemical specificity constraints, this BEA participates to the BA_thioester_synthase process description. The next step consists in combining several BEAs into a biological activity (cf. figure 3B). The BEAs *Ba:PO.1*, *Ba:lab.2*,

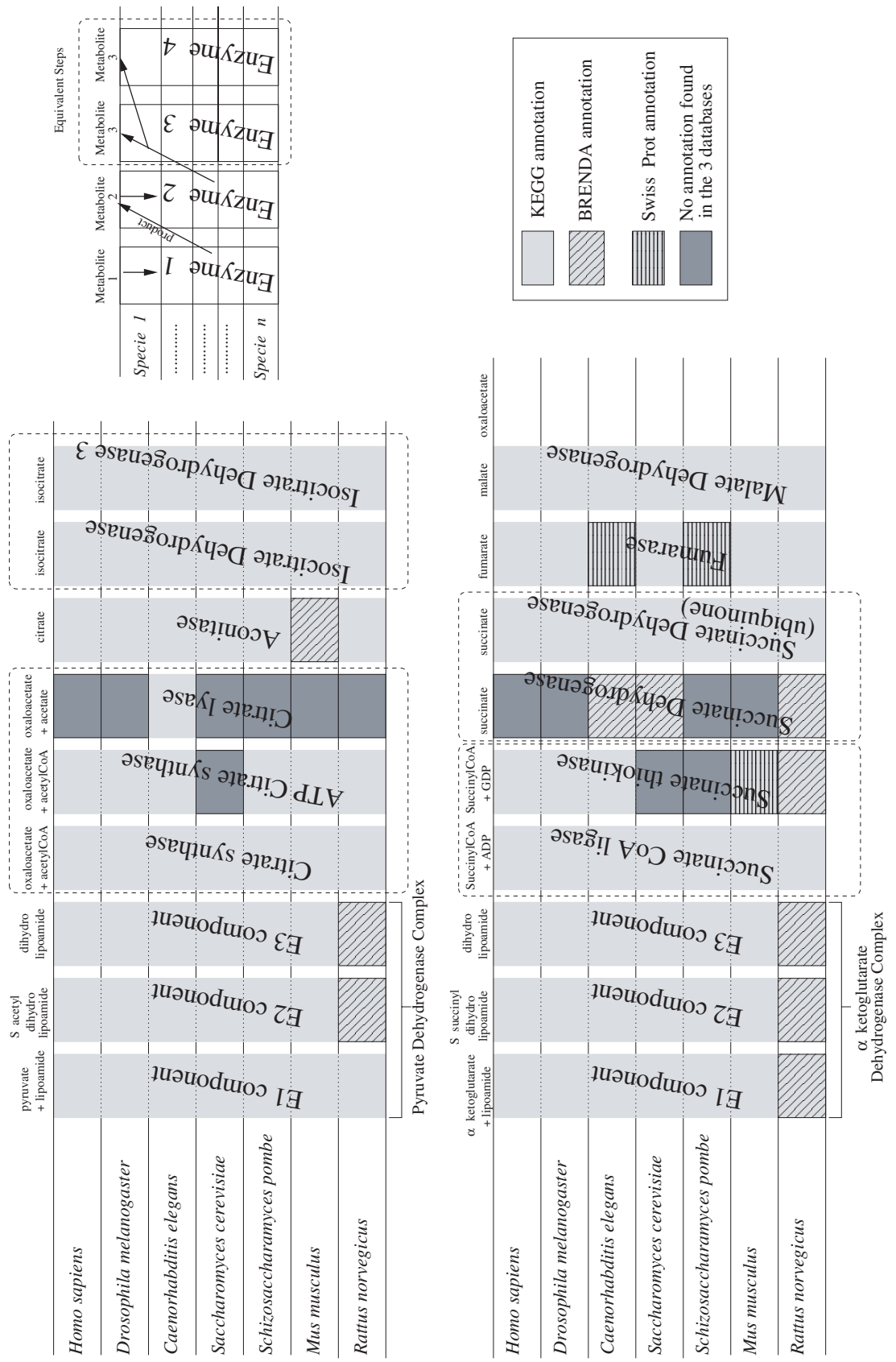

Figure 2. Comparison of the TCA cycle enzymes for seven species. Functional annotation for the enzymes involved in the TCA cycle is reported for 7 species. Three different resources were required to get all the necessary information.

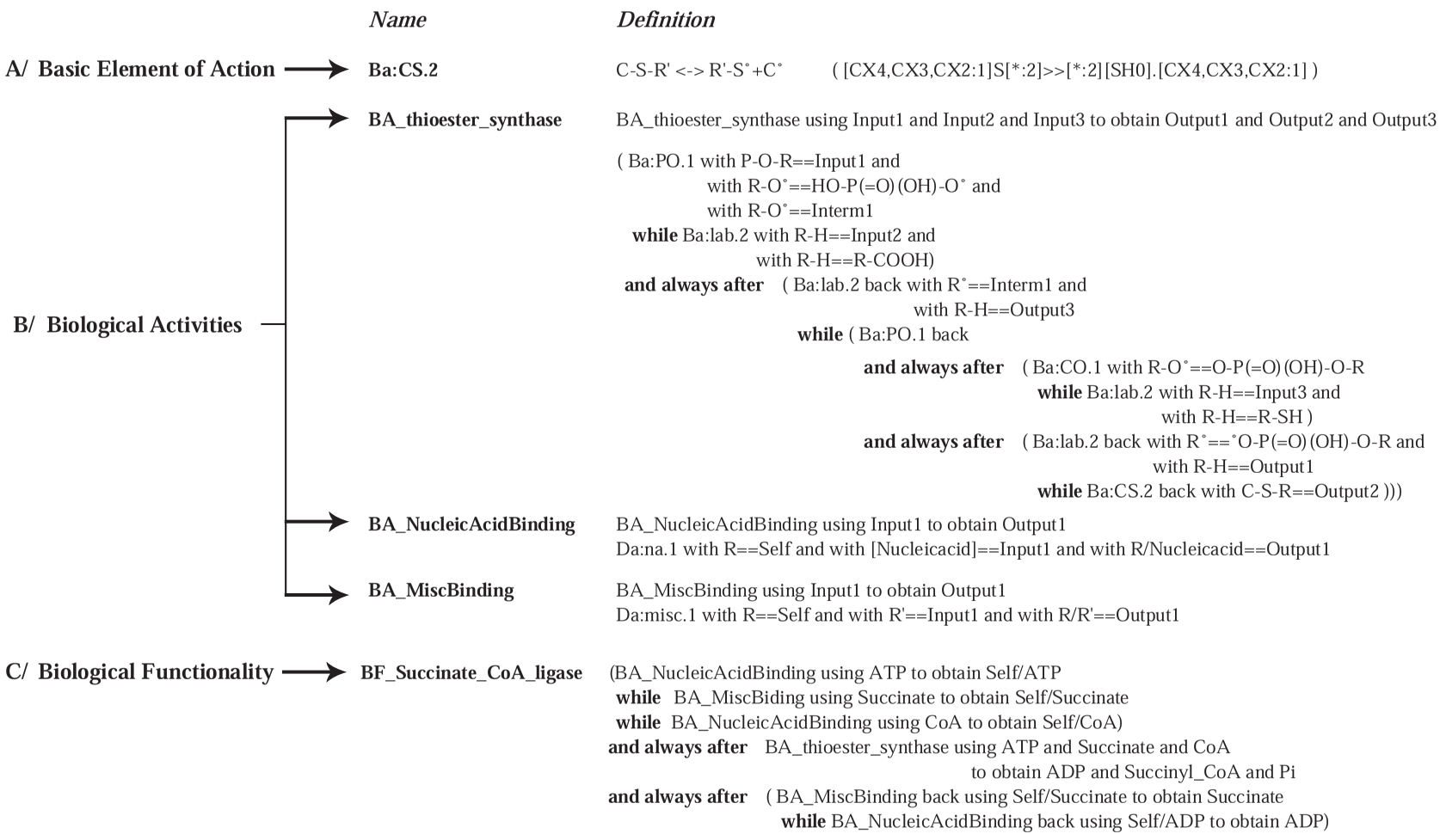

Figure 3. Example of BioΨ descriptions used for the formalization of the TCA cycle process. (A) Representation of the BEA Ba:CS.2 that cuts a bond between a carbon and a sulfur atom; The SMIRKS representation of this transformation is given between parenthesis. (B) Representation of 3 BA level processes. The BA_thioester_synthase requires 3 molecules to build a thioester and makes use of the BEA Ba:CS.2 associated with scheduling and specificity constraints. (C) Representation of a BF level process. The BF_Succinate_CoA_ligase is composed of the previously described BAs associated with scheduling constraints and more stringent specifity constraints.

Ba:CO.1 and *Ba:CS.2* are combined to describe the BA_ thioester_synthase process. The scheduling constraints are attached to the BEAs by operators that allow one to define boolean tests, sequential events or to specify the orientation of the processes ('back' operator). Other kinds of constraints could be included in this BioΨ description as described elsewhere (Maziere *et al* 2004). The same method is used to build BA_NucleicAcidBinding and BA_MiscBinding processes.

2.2b *Biological functionalities composed of biological activities:* The next step involves the design of the biological functionality (cf. figure 3C). In our example, the biological functionality BF_Succinate_CoA_ligase is composed of biological activities BA_thioester_synthase, BA_NucleicAcidBinding and BA_MiscBinding processes associated with biochemical specificity constraints. At this level, the different inputs and outputs are identified as the metabolites involved in the reaction.

2.2c *TCA cycle biological role composed of biological functionalities:* The biological role level of BioΨ combines the biological functionalities described above and associates them with biological constraints. As stated above, this biological role level description can be mapped to the GO term "tricarboxylic acid cycle" (GO:0006099). Whereas the GO annotation is limited to a process designation tag, BioΨ embeds the details of the process down to the description of pseudo-chemical transformations.

2.3 *Functional comparison of TCA cycles*

The comparison of BioΨ descriptions across the TCA cycles of these seven species leads to the construction of a BioΨ generic TCA cycle (cf. figure 4). The seven fully annotated TCA cycles are described by 21 different BEAs combined in a total of 15 BAs. For *Homo sapiens*, 14 BAs are used to compose 13 BFs (cf. figure 4), which is not the case for all species (cf. supplementary material). In figure 4, the generic TCA cycle (shown in figure 1) is represented using its process composition as defined by BioΨ, neglecting biological constraints (generic BioΨ TCA cycle). Each enzyme of the cycle is represented by the biological process it performs at the BF level (the name of the corresponding BF) and at the BA level (symbols representing BAs combined in this BF). Using the same data, this generic BioΨ description of the reference TCA cycle pathway is already able to emphasize information that is not available with the traditional representation of figure 1. This can be achieved by a simple comparison of the lists of formalized processes of each TCA cycle.

First, the two dehydrogenase complexes that are described by different EC numbers in figure 1 are described by exactly the same BAs in the generic BioΨ TCA cycle . In other words, PDH complex and α-ketoglutarate dehydrogenase (αKDH) complex perform the same kind of molecular transformations (with different kinetic, specificity etc. parameters), although constituted of different proteins (cf. figure 5). This information is not new to biologists, but it is either not included in current databases or requires curation assistance from a human expert of the pathway.

Second, the different paths used by the seven species to realize the steps of the cycle are more easily identified and compared. Both succinate-CoA ligase and succinate thiokinase are described by the same BAs whereas succinate dehydrogenase and succinate dehydrogenase ubiquinone processes result in the same outputs but use BF level processes differing by one BA. Citrate synthase, citrate lyase and ATP citrate synthase are clearly using different BF level processes to achieve their transformation of oxaloacetate into citrate. This process comparison can be done in the same manner after including biological constraints, allowing a comparison of the biological contexts required by the different processes.

3. Discussion

Based on the well-known TCA cycle, we conclude that functional annotation found in various databases is often incomplete. The gathering of information from several databases is then required to obtain complete set of data related to biological processes. To integrate such heterogeneous formats used by these databases, the BioΨ description scheme was chosen. This formal representation allows us to describe biological processes independently of the molecular entities by using four abstraction levels. BioΨ properties lead to a new kind of functional comparison independent of protein sequence alignments or clusterings, whose link to function has been shown to be often indirect and noisy (Thornton *et al* 2000). Since much of our existing functional knowledge comes from genomic approaches, widely based on sequence comparisons, BioΨ does not intend to be a predictive tool. Rather, it is a mean to uncover information which is, to some extent, currently present in databases, but not readily accessible to automated analysis tools.

The BioΨ-based TCA cycle provides a model of the cycle compatible with several species by using 21 BEAs combined in 15 BAs. BF level descriptions allow the instantiation of this model by including specificity. This type of biological process description hence provides a new way to deal with metabolic networks. We found that two distinct enzymatic complexes involved in the TCA cycle are, in fact, using the same building blocks for their action. One can notice that the use of only EC numbers in KEGG representation does not allow an accurate description of PDH and αKDH complex processes. The process which is described as EC 1.8.1.4 (transfer of protons from dihydrolipoamide to NAD) is actually composed of a first transfer of protons from dihydrolipoamide

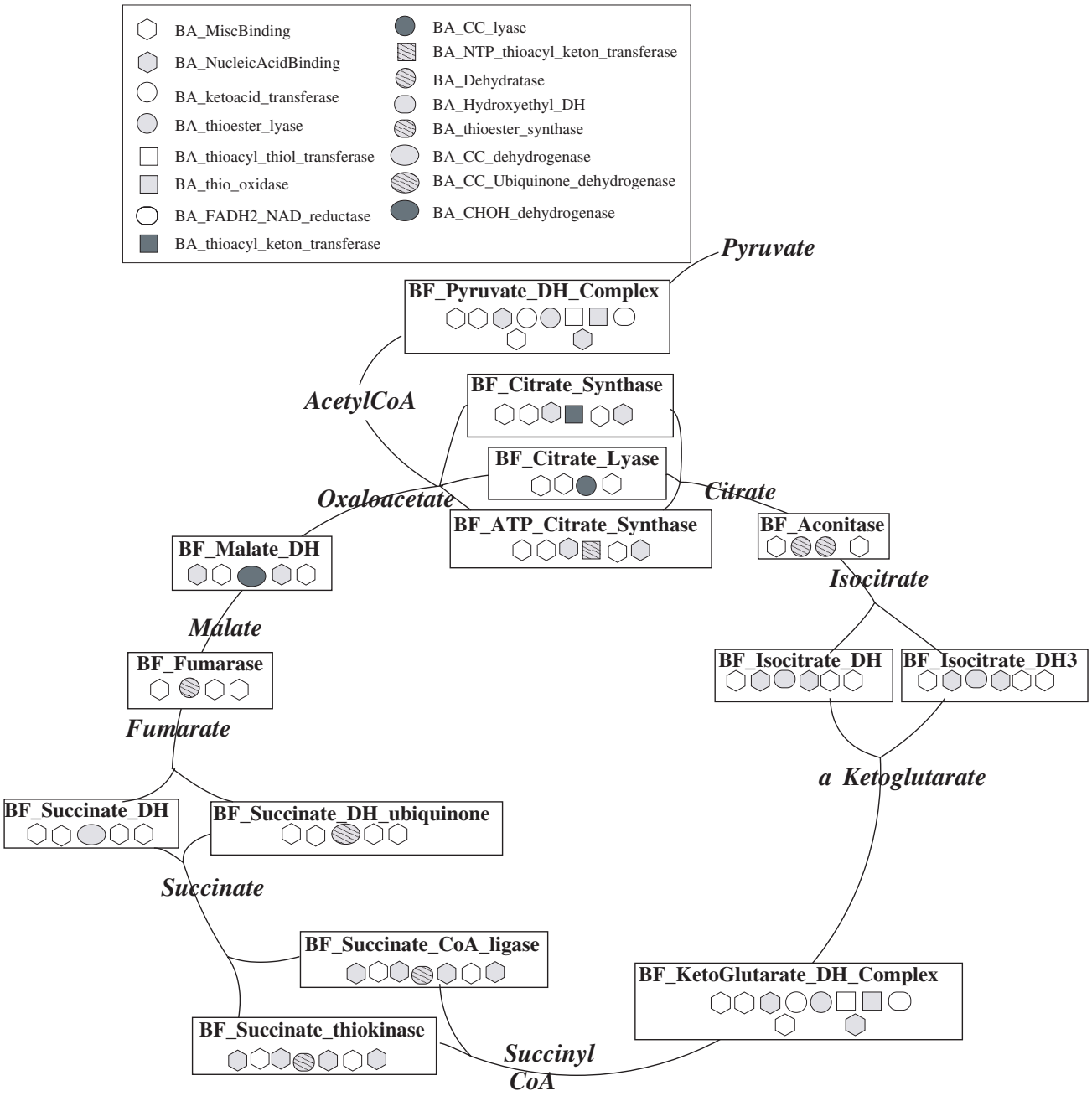

Figure 4. Generic TCA cycle as described by BioΨ. The TCA cycle is represented by using BioΨ: each biological functionalities involved in the biological role BR_TCA_Cycle is described through the biological activities it is composed of.

to FAD, followed by a second transfer of protons from FADH2 to NAD. This decomposition is impossible with EC classifications, because it does not define the FADH2 to NAD proton transfer. The numerous combinations of the hundred defined BEAs give more flexibility and accuracy in process description than what could be accomplished with a set of predefined processes such as EC or GO.

While the two previous complexes were characterized several years ago, some crucial information are still not available in databases. Each of them is composed of three different enzymes with only the dihydrolipoamide dehydrogenase common to both complexes. In the KEGG database, the processes performed by the α-KDH complex are described more precisely than those performed by the

BF_Pyruvate_DH_compplex

(BA_MiscBinding using Pyruvate to obtain Self/Pyruvate
while
BA_MiscBinding using TPP to obtain Self/TPP
while
BA_NucleicAcidBinding using NAD to obtain Self/NADH)
and always after BA_ketoacid_transferase using Pyruvate and E1/TPP to obtain E1-hydroxyethylTPP
and always after BA_thioester_lyase using E1/hydroxyethylTPP and E2-Lypoyllysine
to obtain E2-S-Acetyldihydrolipoyllysine and E1/TPP
and always after BA_thioacyl_thiol_transferase back using CoA and E2-S-Acetyldihydrolipoyllysine
to obtain AcetylCoA and E2-Dihydrolipoyllysine
and always after ((BA_thio_oxidase using E2-Dihydrolipoyllysine and E3/FAD
to obtain E2-lipoyllysine and E3/FADH2
and always after BA_FADH2_NAD_reductase using E3/FADH2 and NAD
to obtain E3/FAD and NADH2
and always after BA_NucleicAcidBinding using NADH2)
while BA_MiscBinding back using AcetylCoA)

BF_KetoGlutarate_DH_Complex

(BA_MiscBinding using AlphaKetoGlutarate to obtain Self/AlphaKetoGlutarate
while
BA_MiscBinding using TPP to obtain Self/TPP
while
BA_NucleicAcidBinding using NAD to obtain Self/NADH)
and always after BA_ketoacid_transferase using AlphaKetoGlutarate and E1/TPP
to obtain E1-hydroxycarboxypropylTPP
and always after BA_thioester_lyase using E1/hydroxycarboxypropylTPP and E2-Lypoyllysine
to obtain E2-S-succinyldihydrolipoyllysine and E1/TPP
and always after BA_thioacyl_thiol_transferase back using CoA and E2-S-Succinyldihydrolipoyllysine
to obtain SuccinylCoA and E2-Dihydrolipoyllysine
and always after ((BA_thio_oxidase using E2-Dihydrolipoyllysine and E3/FAD
to obtain E2-lipoyllysine and E3/FADH2
and always after BA_FADH2_NAD_reductase using E3/FADH2 and NAD
to obtain E3/FAD and NADH2
and always after BA_NucleicAcidBinding using NADH2)
while BA_MiscBinding back using SuccinylCoA)

Figure 5. Pyruvate dehydrogenase complex and α-KDH complex process comparison. The BioΨ formal description of these two complexes composed of different molecules emphasize the similarity of the process they perform.

PDH complex. By using the BioΨ description of the TCA cycle, processes performed by these two different enzymatic complexes are described by the same set of BAs, confirming that the two complexes perform the same biological process. If these similarities are well known to biologists, they are not explicitly indicated in database annotations: it usually requires the knowledge of an expert of the pathway. The simple comparison of the BioΨ description levels can unveil this information in an automated manner. Despite their equivalent processes, these complexes have different specificities in terms of substrates and kinetics. These differences expressed by the biological constraints in the BioΨ description are also available by a simple comparison of each process formalized description: expert intervention is no longer required, allowing the access of this information to modelling and simulation software.

Within a given metabolic pathway, the same biological process can be performed by different molecular entities and at different steps of the metabolic network. As discussed above, such situations can easily be identified using BioΨ descriptions. Moreover, whereas KEGG representation does not take into account the cellular compartments spanned

by the TCA cycle, BioΨ description of this same cycle requires to include the mechanisms enabling the transport of molecules from one compartment to another. For example, (figure 4) despite the switch of symbol from EC numbers to BioΨ levels, the incomplete BioΨ description is still suffering from KEGG-like issues such as the lack of compartment information. This is clearly illustrated by the isocitrate to α-keto glutarate step. The isocitrate dehydrogenase and the isocitrate dehydrogenase 3 seem to perform exactly the same transformations, but they actually are localized respectively in the mitochondria and cytosol. This can not be inferred from the KEGG representation. The BioΨ description of those two processes includes biological constraints that precise the requirement for a specific localization of the molecule that will perform it. Since BioΨ descriptions are formalized, the analysis of the TCA cycle description by dedicated software would not take into account the transformation of isocitrate into α-keto glutarate described by process BF_Isocitrate_DH3: one or several other steps are required to export isocitrate from mitochondria to cytosol, and import α-keto glutarate from cytosol to mitochondria. The same is true for the ATP citrate synthase which is localized into the cytosol.

4. Conclusion

While we primarily used KEGG examples to illustrate our point, new databases such as Reactome (Joshi-Tope *et al* 2005) seem to be more accurate. Such databases, however, still do not offer an easy way to compare biological processes. The integration of BioΨ as the main biological process annotation scheme in a database would permit direct access to existing knowledge for both modelling and simulation software. A drawback of this solution is the requirement for descriptions of every known biological process. It took two weeks to gather all the data and to describe the processes presented in this article. This could be eased by automatic methods which generate a skeleton of BioΨ description based on existing mappings between GO annotations and the BAs and BRs levels. In order to facilitate the use of BioΨ by existing bioinformatic tools, an XML (Hedley 2000) version has been designed, implementing each property of the original language (cf. supplementary material). Compared to CellML (Lloyd *et al* 2004) and SBML (Hucka *et al* 2003), two XML based languages tackling biological process description and model issues, BioΨ is capable of accurately representing the current state of knowledge, as could be described by an expert of a field, rather than a model based on the simplification of this knowledge. This approach should overcome the difficulties encountered by previous attempts to compare and align biological network based on their functionalities (Dandekar *et al* 1999). Because the BioΨ formal description can be represented as a graph, it is for instance compatible with network topology analysis, and elementary action comparison. Both SMILES and SMIRKS (Weininger 1998) are descriptions of chemical entities and chemical patterns that are in common use in databases related to chemical compounds. In BioΨ, BEA descriptions are expressed as SMIRKS and the substrates/products inputs as SMILES. Thus, when applied to biochemical processes, we are able to take advantage of all the theoretical and practical approaches that have been developed in virtual chemistry. Within the MitoScop project, we plan to use this advantage to describe other related metabolic networks in the mitochondria, such as respiratory chain and fatty acid β-oxidation.

Acknowledgments

We are indebted to Professor Jean-Pierre Mazat for useful scientific discussions. We thank Dr Anton Enright for checking the manuscript.

References

Bairoch A, Boeckmann B, Ferro S and Gasteiger E 2004 Swiss-prot: Juggling between evolution and stability; *Brief. Bioinform.* **5** 39–55.

Busch W and Saier M H Jr 2003 The IUBMB-endorsed transporter classification system; *Methods Mol. Biol.* **227** 21–36

Dandekar T, Schuster S, Snel B, Huynen M and Bork P 1999 Pathway alignment: application to the comparative analysis of glycolytic enzymes; *Biochem. J.* **343** Pt 1, 115–124

Fleischmann A, Darsow M, Degtyarenko K, Fleischmann W, Boyce S, Axelsen K, Bairoch A, Schomburg D, Tipton K and Apweiler R 2004 IntEnz, the integrated relational enzyme database; *Nucleic Acids Res.* **32** 434–437

Hanks S and Quinn A 1991 Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members; *Methods Enzymol.* **200** 38–62

Harris M A, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S *et al* 2004 The Gene Ontology (GO) database and informatics resource; *Nucleic Acids Res.* **32** D258–D261

Hedley W 2000 Xml languages for describingbiological models; *Proc. Physiol. Soc. New Zealand* 19

Hucka M, Finney A, Sauro H M, Bolouri H, Doyle J C, Kitano H, Arkin A P, Bornstein B J, *et al* 2003 The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models; *Bioinformatics* **19** 524–531

Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath G R, Wu G R, Matthews L, Lewis S, Birney E and Stein L 2005 Reactome: a knowledgebase of biological pathways; *Nucleic Acids Res.* **33** D428–D432

Karp P, Riley M, Saier M, Paulsen I, Paley S and Pellegrini-Toole A 2000 The ecocyc and metacyc databases; *Nucleic Acids Res.* **28** 56–59

Krupa A, Abhinandan K and Srinivasan N 2004 KinG: a database of protein kinases in genomes; *Nucleic Acids Res.* **32** D153–D155

Lloyd C M, Halstead M D B and Nielsen P F 2004 CellML: its future, present and past; *Prog. Biophys. Mol. Biol.* **85** 433–450

Maziere P, Granier C and Molina F 2004 A description scheme of biological processes based on elementary bricks of action; *J. Mol. Biol.* **339** 77–88

Mewes H, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, Munsterkotter M, Pagel P, Strac N, Stumpflen V, Warfsmann J and Ruepp A 2004 MIPS: analysis and annotation of proteins from whole genomes; *Nucleic Acids Res.* **32** D41–D44

Ogata H, Goto S, Sato K, Fujibuch W, Bono H. and Kanehisa M 1999 Kegg: Kyoto encyclopedia of genes and genomes; *Nucleic Acids Res.* **27** 29–34

Ouzounis C, Coulson R, Enright A, Kunin V and Pereira-Leal J 2003 Classification schemes for protein structure and function; *Nat. Rev. Genet.* **4** 508–519

Rison S, Hodgman T and Thornton J 2000 Comparison of functional annotation schemes for genomes; *Funct. Integr. Genomics* **1** 56–69

Saraste M 1999 Oxidative phosphorylation at the fin de siecle; *Science* **283** 1488–1493

Schomburg I, Chang A and Schomburg D 2004 Brenda, enzyme data and metabolic information; *Nucleic Acids Res.* **30** 47–49

Selkov E Jr, Grechkin Y, Mikhailova N and Selkov E 1998 Mpw: the metabolic pathways database; *Nucleic Acids Res.* **26** 43–45

Soldatova L N and King R D 2005 Are the current ontologies in biology good ontologies?; *Nat. Biotechnol.* **23** 1095–1098

Thornton J M, Todd A E, Milburn D, Borkakoti N and Orengo C A 2000 From structure to function: approaches and limitations; *Nat. Struct. Biol.* (*Suppl*.) **7** 991–994

Weininger D 1998 SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules; *J. Chem. Inf. Comput. Sci.* **28** 31–36

*e*Publication: 20 December 2006