An Algorithm to Assemble Gene-Protein-Reaction Associations for Genome-Scale Metabolic Model Reconstruction

João Cardoso[1,2], Paulo Vilaça[1,2], Simão Soares[1], and Miguel Rocha[2]

[1] SilicoLife Lda., Spinpark, Avepark, Apart. 4152, 4806-909 Guimarães, Portugal
{jcardoso,pvilaca,ssoares}@silicolife.com
[2] CCTC, School of Engineering, University of Minho
mrocha@di.uminho.pt

Abstract. The considerable growth in the number of sequenced genomes and recent advances in Bioinformatics and Systems Biology fields have provided several genome-scale metabolic models (GSMs) that have been used to provide phenotype simulation methods. Given their importance in biomedical research and biotechnology applications (e.g. in Metabolic Engineering efforts), several workflows and computational platforms have been proposed for GSM reconstruction. One of the challenges of these methods is related to the assignment of gene-protein-reaction (GPR) associations that allow to add transcriptional/ translational information to GSMs, a task typically addressed through manual literature curation. This work proposes a novel algorithm to create a set of GPR rules, based on the integration of the information provided by the genome annotation with information on protein composition and function (protein complexes, sub-units, iso-enzymes, etc.) provided by the UniProt database. The methods are validated by using two state-of-the-art models for *E. coli* and *S. cerevisiae*, with competitive results.

Keywords: Metabolic models, gene-protein-reaction rules, genome annotation.

1 Introduction

Genome-scale metabolic models (GSMs) are being increasingly used tools for the understanding of the metabolic behaviour of micro-organisms, allowing the simulation of their phenotypes in distinct environmental and genetic conditions. They have been used to find genetic modifications able to synthesize desired compounds within the realm of Metabolic Engineering (ME) [8] (e.g. *E. coli* strains have been designed *in silico* to overproduce lactate, ethanol, succinate and aminoacids), but also used to guide biological discovery by comparing predicted and experimental data, to analyse global network properties and to study evolution [2]. So, GSMs have become a core element of biological systems analysis and a common denominator for computational and experimental studies.

T. Shibuya et al. (Eds.): PRIB 2012, LNBI 7632, pp. 118–128, 2012.

GSMs gather information regarding different cellular entities. All models have basic information on the portfolio of metabolic reactions and the metabolites involved (including stoichiometry and reversibility information), and in many cases the compartment where reactions occur. Most GSMs also include information on the transcriptional/ translational level, including the enzymes that catalyse the reactions, information on the peptides making the protein complexes and, finally, the genes encoding those peptides [4]. The relationship between genes, proteins and reactions is usually represented using logical rules, commonly called Gene-Protein-Reaction (GPR) rules. These rules represent these relationships using the logical operators AND and OR at two levels: the former states how proteins are encoded by their genes and the latter how the reactions depend on the enzymes.

The inclusion of GPRs within GSMs is essential to allow the phenotype prediction of the cell under different genetic conditions, e.g. gene knockouts or over/underexpression. The capability of performing these predictions is fundamental, for instance in determining gene essentiality and in strain optimization efforts, where the best set of genetic modifications to impose over the the wild type strain is sought, for a given industrial application related to the overproduction of a given compound [13]. In this last case, it has been shown in previous work that the ability to perform simulations of gene knockouts, instead of reaction deletions used in earlier approaches, is essential to obtain more robust and biologically meaningful solutions [10].

The reconstruction of GSMs is being increasingly automated by structured pipelines [5,4] and making use of several Bioinformatics tools, related to genome annotation and re-annotation, homology searches, database integration, protein localization, among others [12,1]. However, in spite of the growing availability of such tools, some of the steps in GSM reconstruction are still done by semi-automated processes with need for manual curation by experts. The determination of the GPRs associated to each metabolic reaction is one of these steps, where the lack of computational tools for the automation of their generation is particularly felt being this task typically conducted by a laborious and time consuming literature search [12].

Therefore, the main aim of this work consists in developing an algorithm that allows to fully automate the process of adding GPR rules to GSMs in the context of their reconstruction process. Thus, the objective is to discover the best GPR rule for each reaction in a GSM, taking as input the information connecting genes and metabolic activities resulting from the genome annotation. The result of this work will be a computational tool to address this task that makes use of existing information in Bioinformatics databases, mainly UniProt [6].

This task is not absent from important hurdles, being the first the inherent complexity of these GPRs. Indeed, two main factors contribute to this complexity: on one hand, different enzymes can have the same metabolic activity (iso-enzymes) and, on the other hand, an enzyme can be a protein complex formed by different sub-units encoded by different genes. Figure 1 illustrates the distinct cases and the corresponding representation in terms of Boolean rules, using examples from the iJR904 model for *Escherichia coli* [9].

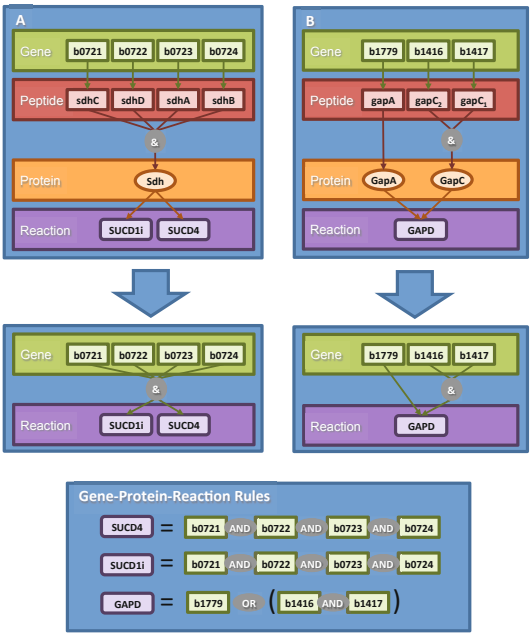

Fig. 1. Illustration of the different cases of GPRs: a) the Sdh enzyme is built from 4 sub-units and catalyses two reactions SUCD4 and SUCD1i; b) GAPD reaction is catalysed by two iso-enzymes (GapA and GapC); GapC is composed of two sub-units encoded by distinct genes.

The most recently published models include, as expected, GPR rules. This is the case with the iAF1260 model for *Escherichia coli* [3] and iMM904 for *Saccharomyces cerevisiae* [7] that will be used in this work to validate our approach.

The remaining of the paper is organized as follows: first, a detailed description of the proposed algorithm is given; next, the results obtained in the two case studies are provided and analysed; finally, conclusions and directions for further work are outlined.

2 Algorithm

An outline of the approach followed in this work is provided in Figure 2. The basic steps of this approach will be explained next with a high-level view. Specific details of each step will follow, organized in sub-sections.

The input for this process is an annotated genome of an organism, assumed in this work as a table containing a gene identifier, one or more Enzyme Commission (EC) numbers with (a list of) assigned metabolic functions and a textual definition of the function of the gene. EC numbers are a recommendation created in order to ensure a systematic organization to define the known metabolic conversions [14].

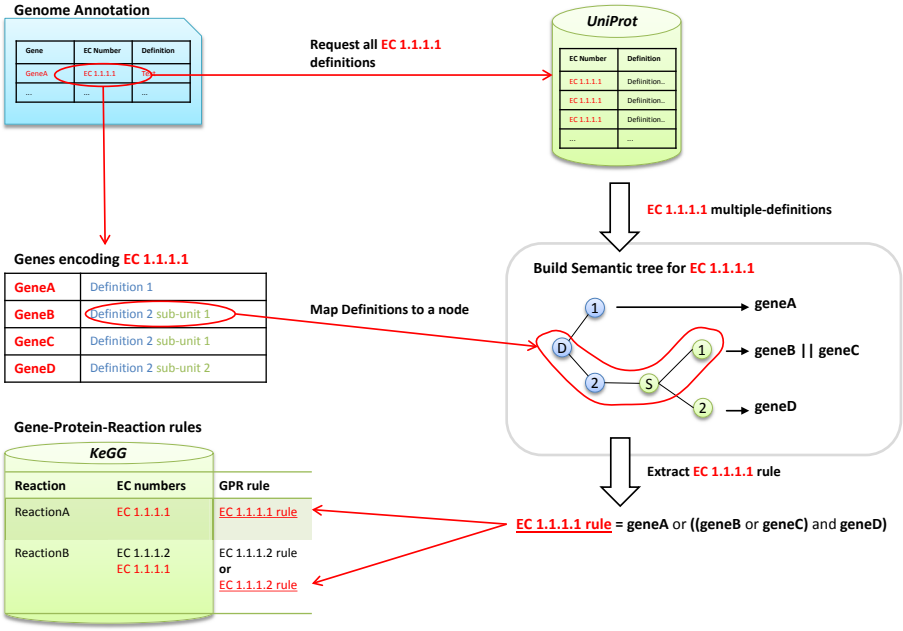

Fig. 2. Overall scheme of the approach followed in this work

For each EC number collected from the genome annotation, a search is conducted in SwissProt, the manually curated database from UniProtKB collection [6]. In each case, a list of matching entries in SwissProt is collected and a semantic tree is created from the definitions included in this list. The aim of this tree is to semantically represent the structure of proteins and their sub-units associated to the respective metabolic function.

In the next step of this process, the aim will be to associate the list of genes associated to that specific EC number to nodes in the semantic tree previously built. This will be done by matching definitions of specific genes to the definitions associated to the tree nodes. Once this association is complete, it is possible to infer a GPR rule by traversing the tree and gathering the linked genes, outputting a rule in the form of a Boolean function. Based on the biological meaning of each tree node, the algorithm can infer the biological association of the genes using the AND or OR logical operators and, thus, create a GPR rule for each EC number.

The final step of the algorithm is to create GPR rules for each reaction. From the GSM reconstruction process, a table is provided containing the list of reactions and their associated EC numbers (e.g. this information can be obtained from databases such as KEGG `http://www.genome.jp/kegg`). The GPR rule for a given reaction is obtained by the rules from the assigned EC numbers. If more than one EC number is assigned to a reaction, the respective rule will be created by joining the rules from the EC numbers using the operator OR.

2.1 Building the Semantic Tree

One of the most important steps of this algorithm is the creation of a semantic tree for each metabolic function (EC number). This tree is a n-ary tree structure, similar to a suffix tree, where the values are textual expressions representing biological definitions for functional roles. The input for this step will be a set of textual definitions, in this case from the list of entries as a result of a search for a specific EC number in the SwissProt database.

The first step is to create a matrix from the list of definitions, where each row is a definition and each column is obtained by splitting the expressions using white spaces and parenthesis as separators. The matrix is composed of all possible definitions available in the database. Figure 3 describes how the matrix is built from a definition set.

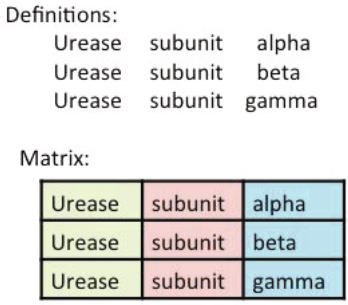

Fig. 3. Description of the matrix assembly process. The expressions are split into the terms and placed in the matrix resulting in 3x3 matrix structure.

To avoid problems with mismatches caused by synonyms of protein names or functional definitions, a dictionary is created for each EC number. This dictionary is filled with information from UniProt regarding synonyms or alternative names. The strategy is to keep in the matrix only one recommended name in each case and this strategy is applied to all definition rows.

Also, to prevent mismatches caused by typos or other small differences in terms, a global dictionary is used with common terms. A Levenshtein Automaton [11] is applied to every word, finding the closest word in the dictionary. If the distance is equal to 1, the word is replaced by the dictionary word. This allows to correct misspelled words, such as "putativ" or "cmponent", instead of "putative" or "component" respectively.

In order to reduce the information noise, some expressions were defined as useless to the definition match process and these terms are removed from the expressions. In this list the following are included: cellular localization terms, as the definition is the same; organs or organism structure, such as "leaf", "liver", etc; synonyms of homology or same function, such as "like" or "isoenzyme" are also not required. Those words are removed from the expressions before building the matrix.

The semantic tree is created by traversing the matrix row by row. The algorithm used to build the tree follows the ones used to build suffix trees, i.e. when a new row is considered the algorithm will match its words with the nodes in the tree, starting by the root and following the respective branches. When, at a certain level, the branch for that term does not exist, a new branch is created. The tree is composed by two types of nodes: terms, that represent each unique word available in the definition; and the genes that are associated to the last word in the expression.

To create the Boolean rules, it is required to identify how the components are assembled together. Thus, it is necessary to identify for each branch if it will associated to an AND or to an OR relationship. This process will take into account the semantics of the terms found in the annotations. The gene nodes associated connected to the same root are associated by an OR expression. Terms such as "subunit", "chain", "component", "peptide" or any synonym to these words identify the existence of a complex structure and therefore will be associated to an AND relationship. The remaining terms under the same node are also related with an OR relationship.

There are also distinct identifiers for different types of substructure: Greek alphabet characters, Roman numerals, digits and Latin alphabet characters. In some cases, the complex is made by a pair of a "small" and a "large" or "heavy" and "light" chain or units described by their molecular weight. Figure 4 exemplifies the generated semantic tree for an example.

2.2 Matching the Tree with the Genome Annotation

Given a table with the annotated genome, containing for each metabolic gene a set of EC numbers and the textual definition of its functional role, the next step will be to map the genes onto the trees created in the previous step.

For each EC number, a tree is created as explained in the previous section. Also, a list of genes related to that EC number is extracted from the genome annotation table. Each of these genes will then be mapped to the tree by matching its definition text with the one on the tree nodes. The matching algorithm is similar to the one using in the construction of the tree explained above. The gene will be linked to the deepest node in the tree where the matching process is possible.

When all genes for a given EC number are matched onto the tree, it is possible to create a rule for this EC number. The tree is traversed generating a string; each branch has a Boolean function, i.e. the nodes in that branch are connected by either "AND" or "OR". Sub-trees without genes are disregarded and nodes with genes will add the gene identifier to the string.

Figure 5 shows an example, based on the tree shown in the previous subsection. In this case, the generated GPR will be the following: *BCE_3662 AND BCE_3663 AND BCE_3664*.

The last step is to generate rules for the reactions in the target model. Assuming there is information available on the set of EC numbers for each rule, this step is achieved by joining together the rules for the set of EC numbers through an OR operator.

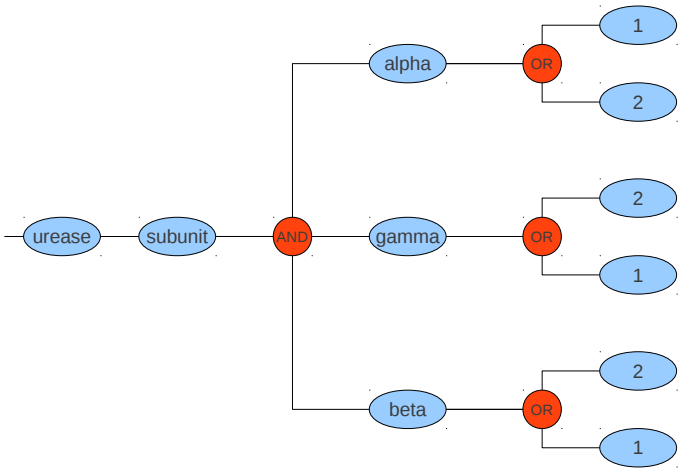

Fig. 4. Illustration of an example semantic tree

2.3 Implementation

The previous algorithm was implemented using the Java programming language, being the software available on demand to the authors. To collect all information from the SwissProt database, the UniProtJAPI provided by European Bioinformatics Institute (EBI) (`http://www.ebi.ac.uk/uniprot/remotingAPI`) has been used. The mappings of reactions to EC numbers can be taken from the KEGG database. In the experiments, this information is available from the models.

3 Results

In order to validate the proposed algorithm, two existing GSMs were used: the iAF1260 model for *Escherichia coli* [3] and iMM904 for *Saccharomyces cerevisiae* [7]. Since these methods have a set of GPR rules associated to most reactions, in both cases as a result of thorough literature curation process Table 1 shows basic statistics of both models, including the number of reactions with an assigned EC number, the ones with GPR rules available and the intersection of both sets. These last sets will be the ones of interest in the analysis of the results, to provide a fair comparison with the proposed method.

By running the methods described in the previous section in the provided case studies, the following number of GPR rules were created (showing also the percentage over the total number of reactions with GPR and EC number):

- *Escherichia coli*: 674 (71%)
- *Saccharomyces cerevisiae*: 535 (70%)

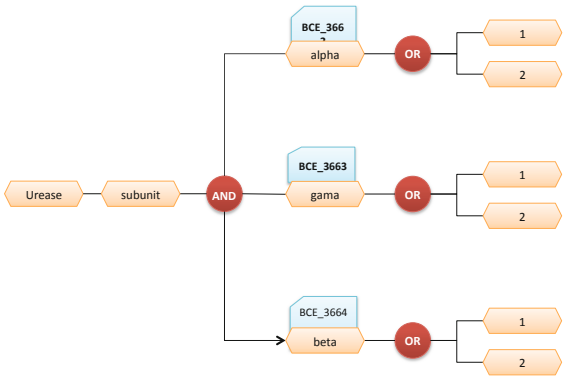

Fig. 5. Illustration of the process of mapping genes onto the semantic tree

Table 1. Model statistics

	E.coli	*S.cerevisiae*
Reactions with EC number	955	760
Reactions with GPR	1944	1043
Reactions with GPR and EC number	932	693

To provide an analysis of the results by comparing the rules obtained with the ones in the models, the Jaccard coefficient (J) will be defined to compare GPR rules for the same reaction. For each rule, the sets of genes used in the target rule (T) and the proposed rule (P) are taken and J is calculated as follows:

$$J(T, P) = \frac{|P \cap T|}{|P \cup T|} \tag{1}$$

Figure 6 shows the distribution of J values over all reactions for both case studies. The values are divided into four categories: $J = 1$ (perfect match), $J \geq 0.5$ (considered a good match), $J < 0.5$ (partial match) and $J = 0$ (no match). In both cases, the large majority of the rules obtain a good match with the rules in the model, with over 50% with a perfect match and more than half of the remaining with a match over 50%. It is important to notice that about half of the cases where there is no match are situations where the proposed method provides a rule and the model does not have one.

These results show the high correspondence between both data. However, since GPR rules are Boolean functions it is important not only to check the correspondence of the variables used, but also to compare the results of the function. This analysis was conducted for the cases where there was a full match of the sets of variables used. A truth table with all possible values for the genes was created in each case, where each row stands for a possible combination of the values of the genes involved. The output of the GPR rule was compared in

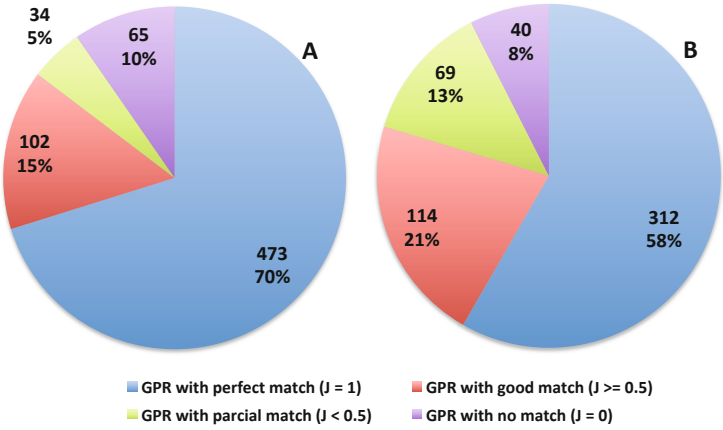

Fig. 6. Results of the proposed methods applied to the E. coli (a) and S. cerevisiae (b) models. Pie charts show the distribution of the Jaccard indexes (J) calculated over the GPR rules created using the proposed methods and obtained from the models

each case between the proposed rule and the existing one. It was verified that the results are 99.8% identical in *E. coli* and 100% in *S. cerevisiae*.

It is also important to notice that models are also composed by transport reactions that have a specific annotation - the Transport Commission numbers - and their semantic composition is more complex. For that reason, this algorithm is not suitable for assemble GPR for those conversions.

Other discrepancies between the existing rules have been found. For instance, the EC 1.2.1.3 (*aldehyde dehydrogenase*) is associated to *b1300* in the model, however the UniProt database describes it with EC number 1.2.1.5 . This mismatch can be either explained by two reasons: the annotation was reviewed and associated with a new function or the manual curation and literature mining process during the reconstruction determined that the gene is also related to the function. Another issue identified is the lack of EC function associated with the gene (e.g. *b3610* in the *E. coli* model). Although the model has an association with the metabolic function EC 1.8.4.2, there is no evidence at the UniProt database.

4 Conclusions and Further Work

In this work, a novel algorithm and computational tool has been proposed to address the task of gene-protein-reaction rule inference from the genome annotation. This is an important task within the larger effort of genome-scale metabolic model reconstruction that has been traditionally performed using laborious manual literature curation. Although the results are still preliminary and the methods can be improved, this contribution already shown interesting results when applied to well known and validated models from *E. coli* and *S. cerevisiae*.

Some issues are still preventing more accurate results. One one hand, the models used as case studies were built over the last decade in a process of iterative refinement involving huge resources and extensive manual curation. Also, in many cases, divergences on the EC number annotations between the models and the UniProt database are the reason for many mismatches. This should be further examined in posterior work, namely by considering the use of additional databases complementing UniProt.

Also, the approach proposed here is not able to encompass an important class of reactions that handle the transport of metabolites from the exterior of the cell and between cell compartments. Since these are mostly not covered by EC number nomenclature, a distinct approach needs to be developed, for instance based on TC numbers from the TCDB database (`http://www.tcdb.org/`). This will be a major task in future work, together with other possible improvements in the proposed methodology.

Acknowledgements. The work is partially funded by ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT (Portuguese Foundation for Science and Tech-nology) within projects ref. COMPETE FCOMP-01-0124-FEDER-015079 and PEst-OE/EEI/UI0752/2011.

References

1. Dias, O., Rocha, M., Ferreira, E., Rocha, I.: Merlin: Metabolic models reconstruction using genome-scale information. Computer Applications in Biotechnology 11, 120–125 (2010)
2. Feist, A.M., Palsson, B.: The growing scope of applications of genome-scale metabolic reconstructions using escherichia coli. Nature Biotechnology 26(6), 659–667 (2008)
3. Feist, A.M., Henry, C.S., Reed, J.L., Krummenacker, M., Joyce, A.R., Karp, P.D., Broadbelt, L.J., Hatzimanikatis, V., Palsson, B.Ø.: A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. Molecular Systems Biology 3 (2007)
4. Feist, A.M., Herrgård, M.J., Thiele, I., Reed, J.L., Ø Palsson, B.: Reconstruction of biochemical networks in microorganisms. Nature Reviews. Microbiology 7(2), 129–143 (2009)
5. Henry, C.S., DeJongh, M., Best, A.A., Frybarger, P.M., Linsay, B., Stevens, R.L.: High-throughput generation, optimization and analysis of genome-scale metabolic models. Nature Biotechnology 28(9), 977–982 (2010)
6. Magrane, M., Uniprot Consortium: Uniprot knowledgebase: a hub of integrated protein data. Database: the Journal of Biological Databases and Curation 2011:bar009 (January 2011)
7. Mo, M., Palsson, B.Ø., Herrgård, M.J.: Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC Systems Biology 3, 37 (2009)
8. Nielsen, J.: Metabolic engineering. Appl Microbiol Biotechnol. 55, 263–283 (2001)
9. Reed, J.L., Vo, T.D., Schilling, C.H., Palsson, B.Ø.: An expanded genome-scale model of escherichia coli k-12 (ijr904 gsm/gpr). Genome Biology 4, R54 (2006)

10. Rocha, M., Maia, P., Mendes, R., Pinto, J.P., Ferreira, E.C., Nielsen, J., Patil, K.R., Rocha, I.: Natural computation meta-heuristics for the in silico optimization of microbial strains. BMC Bioinformatics 9 (2008)
11. Schulz, K.U., Mihov, S.: Fast string correction with levenshtein automata. International Journal on Document Analysis and Recognition 5, 67–85 (2002)
12. Thiele, I., Palsson, B.Ø.: A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature Protocols 5, 93–121 (2010)
13. Vilaça, P., Maia, P., Rocha, I., Rocha, M.: Metaheuristics for Strain Optimization Using Transcriptional Information Enriched Metabolic Models. In: Pizzuti, C., Ritchie, M.D., Giacobini, M. (eds.) EvoBIO 2010. LNCS, vol. 6023, pp. 205–216. Springer, Heidelberg (2010)
14. Webb, E.C.: International Union of Biochemistry, and Molecular Biology. In: Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes, 6th edn., Academic Press (1992)