Genetics and Molecular Biology, 28, 3 (suppl), 630-633 (2005)
Copyright by the Brazilian Society of Genetics. Printed in Brazil
www.sbg.org.br

Short Communication

Using the FORESTS and KEGG databases to investigate the metabolic network of *Eucalyptus*

José C.M. Mombach, Ney Lemke, Norma M. da Silva, Rejane A. Ferreira, Eduardo Isaia Filho, Cláudia K. Barcellos and Rodrigo J. Ormazabal

Universidade do Vale do Rio dos Sinos, Laboratório de Bioinformática e Biologia Computacional, São Leopoldo, RS, Brazil.

Abstract

In this work we apply a bioinformatics approach to determine the most important enzymes of the metabolic network of *Eucalyptus* to determine the coverage of the genome in the FORESTS library. We conclude that the library does not cover completely the metabolism of the organism. However, some important pathways could be analyzed, especially the lignin synthesis. We found that four of the most important enzymes predicted are involved in this pathway.

Key words: Eucalyptus, metabolism.

Received: June 3, 2004; Accepted: March 17, 2005.

The genome sequences of several organisms are available (Kanehisa and Goto, 2000) currently and the extraction of relevant physiological information from these data is a major challenge. The metabolic networks reconstructed from the annotated genomes provide an abundance of information for exploration with bioinformatics (Karp *et al.*, 1999). In this article we apply a recently developed graph analysis of metabolism that predicts important enzymes (Lemke *et al.*, 2004) of the metabolic network of Eucalyptus. Data were obtained from the non-public database of the Eucalyptus Genome Sequencing Project Consortium (FORESTS, https://forests.esalq.usp.br/) that contains approximately 100,000 expressed sequence tags (ESTs) of this plant.

Cellular metabolism is a complex network of biochemical reactions catalyzed by specialized proteins called enzymes. The reactions are organized in modules called metabolic maps with specific catabolic or anabolic functions. The complete set of metabolic maps forms the metabolic network.

An exponentially growing number of organisms have sequenced genomes (Devos and Valencia, 2001), and assuming that the annotated proteins are expressed, we can reconstruct the metabolic network of the organism (Karp *et al.*, 1999).

The investigation of the influence of enzymes on the network is a critical issue for the bioengineering and pharmaceutical industry since they can be targets for drugs (Karp *et al.*, 1999) or they can be genetically engineered to change the metabolic output of specific metabolites (Edwards and Palsson, 1997). Our approach investigates the static structure of the components of the network to infer causal and physiological relationships. In a previous work we applied our method to *Escherichia coli,* whose metabolism has been studied in depth, and found a strong correlation between the damage an enzyme causes to the network and its essentiality (Lemke *et al.*, 2004), thus showing the predictive power the method has for determining important enzymes. In this work we apply this method to the genome of *Eucalyptus* that is an important commercial source of wood and cellulose. In particular, there is a great deal of interest in the bioengineering of plants involving the metabolic pathways of cellulose and lignin to generate genetically modified organisms with enhanced production of cellulose and decreased production of lignin.

In our method we have introduced a new quantitative criterion for enzyme importance: the damage its removal causes to the metabolic network (Lemke *et al.*, 2004). In the absence of complete information about kinetic parameters and the influence of the regulatory network, we cannot predict all consequences of the deletion of a specific enzyme; however, we assume that the essentiality of a protein is not necessarily related to its level of expression as shown by Rocha and Danchin (2003) for *Escherichia coli* and *Bacillus subtilis*. Using only information about the reactions oc-

Send correspondence to José Carlos M. Mombach. Universidade do Vale do Rio dos Sinos, Laboratorio de Bioinformática e Biologia Computacional, Av. Unisinos 950, 93022-000 São Leopoldo, RS, Brazil. E-mail: mombach@unisinos.br.

curring in an organism, we can determine the number of metabolites whose production is prevented by the absence of the enzyme , which we define as the damage *d* to the network. *d* is a measure of the number of disrupted pathways generated by the removal of the enzyme from the network. To build the network of metabolic reactions for an organism, we collect all enzyme codes (ECs) and reactions involved in the small molecule metabolism (Edwards *et al.*, 2001; Masanori, 2004) of *Eucalyptus* using the KEGG and FORESTS databases. The small molecule metabolism is a subset of the complete metabolism that excludes DNA replication and protein synthesis reactions (Masanori, 2004).

From the FORESTS database, we obtained HTML files with the set of ECs and corresponding metabolic maps. Using this information, we searched KEGG for corresponding reactions to build the metabolic network. This initial metabolic set of reactions can present inconsistencies since the annotation of biological sequences is error prone (Devos and Valencia, 2001). Among the most common problems we find on annotated genomes is the incorrect attribution of function to some genes and the lack of function of others. When dealing with annotations based on ESTs libraries, we also have to deal with large errors in the nucleotide sequence and the fact that we do not have the complete genome since low-expressed genes are missed. A key advantage of this technique is that we can be confident that all the genes we have are really expressed. Absent genes imply the generation of inconsistent sets of chemical reactions with missing or false reactions. In our analysis we verify the consistency of this set, impossible reactions are removed and highly probable missing reactions are introduced. This analysis of consistency is helped by the fact that reactions produce or use metabolites from other reactions. For example, the reaction corresponding to an annotated EC may require a metabolite that is not produced by any other reaction or is not available from an external source. Then the reaction is not included in the input reaction set for the simulation. In addition, some reactions require metabolites that are not produced inside the cell. We determine all external metabolites required and use them as an input for the simulation. Consequently, our generated set of chemical reactions has higher confidence than the one obtained solely from the annotation process, but is dependent on its accuracy.

The metabolic analysis uses a graphical representation of metabolism. The graph is directed and has two types of nodes (formally we classify this structure as a bipartite graph). One node type represents chemical reactions and the other metabolites. A link between a reaction and a metabolite points towards the metabolite, if the metabolite is a product, and in the opposite direction, if the metabolite is a reactant. We treat reversible reactions as two separate reactions. To calculate the damage, we select an enzyme and delete all reactions it catalyzes; the number of deleted metabolites is defined as *d*. For more details see reference Lemke *et al.* (2004). The method was motivated by experiments of systematic mutagenesis where the importance of each gene is tested for the survival of an organism. The algorithm can be thought of as an *in silico* knock out of the ORF that encodes a given enzyme.

The analysis of the EST sequences of *Eucalyptus* in the FORESTS database was restricted to ESTs coding enzymes which involved the small molecule metabolism. Some enzymes were excluded from the analysis (32 ESTs); these proteins catalyzed reactions whose substrates cannot be produced by any other reaction and their products are never used. The reactions we used in our simulation are present in the following metabolic pathways:

- Alanine and aspartate metabolism;
- D-arginine and D-ornithine metabolism;
- Arginine and proline metabolism;
- D-glutamine and D-glutamate metabolism;
- Electron transport and oxidative phosphorylation;
- Galactose metabolism;
- Glutamate metabolism;
- Glyoxylate and dicarboxylate metabolism;
- Inositol phosphate metabolism;
- Inositol metabolism;
- Lysine biosynthesis;
- Lysine degradation;
- Pyrimidine metabolism;
- Purine metabolism;
- Reductive carboxylate cycle;
- Riboflavin metabolism;
- B-alanine metabolism;
- Cysteine metabolism;
- Stilbene, coumarine, and lignin biosynthesis;
- Starch and sucrose metabolism;
- Sterol, vitamin K, vitamin E, carotenoids biosynthesis;
- Tyrosine metabolism;
- Phenylalanine metabolism;
- Tryptophan metabolism;
- Urea cycle and metabolism of amino groups;
- Phenylalanine, tyrosine and tryptophan biosynthesis.

We found incomplete metabolic pathways with missing enzymes; however, the pathway of lignin production was possible to analyze. In Table 1 we list, in decreasing order of damage, the enzymes associated with high damage (*d*) to the network. These enzymes have a higher probability than others of being essential for metabolism (Lemke *et al.*, 2004). The list includes only enzymes with damage equal or higher than 6. We found that 204 (58.1% of 351) enzymes cause damage 0 to the network implying that they are redundant for the metabolism according to our simulation.

Here we give a brief description of the function of some of the enzymes in Table 1: phenylalanine ammonia lyase is involved in the elimination of CH3 and may also act

Table 1 - List of enzymes of *Eucalyptus* predicted to be associated with highest damage. First column: Enzyme name and EC. Second column: Damage *d*. Third column: metabolic maps in which enzymes participate.

Enzyme name (EC)	Damage	Metabolic maps in which enzyme participates
1. phenylalanine ammonia lyase (4.3.1.5)	22	tyrosine; phenylalanine; nitrogen; alkaloid (II) biosynthesis
2. 4-coumarate-Coa ligase (6.2.1.12)	16	flavonoids, stilbene, and lignin biosynthesis
3. aldehyde dehydrogenase (1.2.1.5)	12	glycolisis/gluconeogenesis; histidine; tyrosine; phenylalanine
4. mevalonate diphosphate decarboxylase (4.1.1.33)	12	sterol biosynthesis
5. monophenol monooxygenase (1.14.18.1)	12	tyrosine; riboflavin; flavonoids, stilbene and lignin; alkaloid (I) biosynthesis
6. alpha-galactosidase (3.2.1.22)	11	galactose; glycerolipid; sphingoglycolipid; globoside
7. isopentenyl-diphosphate delta-isomerase (5.3.3.2)	11	sterol; terpenoid biosynthesis
8. cinnamoyl-Coa reductase (1.2.1.44)	9	flavonoids, stilbene, and lignin biosynthesis
9. galactinol-sucrose galactosyltransferase (2.4.1.82)	8	galactose
10. cathecol O-methyltransf. (2.1.1.6)	8	tyrosine
11. L-aminoadipate-semialdehyde dehydrogenase (1.2.1.31)	8	lysine biosynthesis and degradation
12. UDP-glucose 6-dehydrogenase (1.1.1.22)	7	pentose and glucuronate interconversions; starch and sucrose; nucleotide sugars
13. homoserine dehydrogenase (1.1.1.3)	7	glycine, serine, and threonine; lysine biosynthesis
14. caffeate O-methyltransferase (2.1.1.68)	6	flavonoids, stilbene, and lignin biosynthesis
15. succinate-Coa ligase (6.2.1.5)	6	TCA cycle; propanoate; C5-branched dibasic acid; CO2 fixation
16. trimethyllysine dioxygenase (1.14.11.8)	6	lysine degradation
17. dihydrodipicolinate reductase (1.3.1.26)	6	lysine biosynthesis
18. glutathione dehydrogenase (1.8.5.1)	6	ascorbate and aldarate; glutamate; glutathione

on L-tyrosine; 4-coumarate-Coa ligase acts on acid-thiol ligation; aldehyde dehydrogenase is an oxidoreductase acting on the aldehyde or oxo group of donors, with NAD$^+$ or NADP$^+$ as acceptors; mevalonate diphosphate decarboxylase is a lyase involved in sterol biosynthesis; monophenol monooxygenase is an oxidoreductase acting on paired donors with incorporation of molecular oxygen, with another compound as one donor, and incorporation of one atom of oxygen; alpha-galactosidase is hydrolase involved in the hydrolysis of terminal, non-reducing alpha-D-galactose residues in alpha-D-galactosides, including galactose oligosaccharides, galactomannans and galactohydrolase; isopentenyl-diphosphate delta-isomerase acts on sterol and terpenoid biosynthesis cinnamoyl-Coa reductase is an oxireductase that acts on a number of substituted cinnamoyl esters of coenzyme A; galactinol-sucrose galactosyltransferase is involved in the hexosyl group transfer, the first step in biosynthesis of raffinose sugars; L-aminoadipate-semialdehyde dehydrogenase is an oxidoreductase acting on the aldehyde or oxo group of donors, with NAD$^+$ or NADP$^+$ as acceptors; UDP-glucose 6-dehydrogenase is an oxireductase and acts on the CH-OH group of donors, with NAD$^+$ or NADP$^+$ as acceptor. Also acts on UDP-2-deoxyglucose; succinate-Coa ligase is involved in acid-thiol ligation and formation of carbon sulfur bonds; dihydrodipicolinate reductase is an oxireductase that acts on paired donors with incorporation of molecular oxygen, with 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors. Requires Fe^{+2} and ascorbate.

We concluded that the current FORESTS library provides an incomplete coverage of the metabolic network of *Eucalyptus* since key metabolic compounds cannot be produced with the available set of enzymes. For example, cellulose, a key structural compound cannot be produced. In order to produce this substance from the compounds UDP-glucose and GDP-glucose, the enzymes cellulose synthase (UDP-forming, EC 2.4.1.12) and cellulose synthase (GDP-forming, EC 2.4.1.29) are required according to KEGG map's Starch and Sucrose Metabolism; however, they are missing in the FORESTS database. With our analysis we cannot determine whether this is due to an incomplete library or due to errors on the annotated enzymes or perhaps because this metabolic pathway is based on ORFs that have no homologs in any other sequenced organism. The clarification of this point demands experimental investigation. Some metabolic pathways seem incomplete, for example, in the Arginine and Proline Metabolism, only 2 enzymes out of 71 are present in the database and in the cellulose synthesis pathway, only 11 out of 71 are present. In contrast, some pathways, such as the Riboflavin Metabolism seem complete with 11 out of 13 enzymes present and the Alanine and Aspartate Metabolism with 33 out of 38. We must point out that KEGG's metabolic maps are general for all organisms and not necessarily all enzymes represented in these maps exist for a given organism. Additionally, metabolism is a complex network of reactions particular to different groups of organisms.

We also found that 32 enzymes catalyze isolated reactions, indicating that the library is not complete. The list of excluded ECs is presented in Table 2. However, the lignin synthesis could be investigated in our analysis. For instance, we determined that its production is prevented if the ORFs coding enzymes with ECs 6.2.1.12, 4.3.1.5, 1.2.1.44, 1.11.1.7, and 1.1.1.195 are deleted. This information can be also used to decrease the production of this substance. Among the most important enzymes, four are related to lignin biosynthesis (see Table 1).

We searched EC numbers in each enzyme class and found that oxireductases are the most abundant followed by transferases, hydrolases, lyases, ligases, and isomerases. The comparison of the number of ECs with the number of reactions in each enzyme category suggests that oxireductases and transferases are less substrate specific (ratio of number of reactions to number of ECs) than any of the other enzyme classes.

Table 2 - List of enzymes excluded from our analysis. These enzymes catalyze reactions whose substrates are not produced by any other reaction and whose products are also not used.

Enzyme name	(EC)
nicotinate-nucleotide-dimethylbenzimidazole	(2.4.2.21)
beta-mercaptopyruvate sulfurtransferase	(2.8.1.2)
cystathione beta-lyase	(4.4.1.8)
hydroxymethylglutaryl-Coa reductase	(1.1.1.34)
lathosterol oxidase	(1.3.3.2)
menadione reductase	(1.6.99.2)
mevalonate kinase	(2.7.1.36)
2’, 3’ -cyclic-nucleotide 2’ -phosphodiesterase	(3.1.4.16)
adenosine-tetraphosphatase	(3.6.1.14)
aspartoacylase	(3.5.1.15)
6.3.2.15	(not available in KEGG)
tyrosine 3-monooxygenase	(1.14.16.2)
pantoate-beta-alanine ligase	(6.3.2.1)
pectinesterase	(3.1.1.11)
polygalacturonase	(3.2.1.15)
oxalate oxidase	(1.2.3.4)
formamidase	(3.5.1.49)
N-formylglutamate deformylase	(3.5.1.68)
glutamine-fructose-6-phosphate transaminase	(2.6.1.16)
GMP reductase	(1.7.1.7)
cysteine dioxygenase	(1.13.11.20)
homoaconitate hydratase	(4.2.1.36)
cistathionine gamma-lyase	(4.4.1.1)
guanidinoacetate kinase	(2.7.3.1)
formylmethionine deformilase	(3.5.1.31)
glyceraldehyde 3-phosphate dehydrogenase	(1.2.1.12)
glutathione-cystine transhydrogenase	(1.8.4.4)
aminobutyraldehyde dehydrogenase	(1.2.1.19)
alpha,alpha-trehalase	(3.2.1.28)
oxalate decarboxylase	(4.1.1.2)
phosphoric monoester hydrolase	(3.1.3.-)

Our results also indicate a high level of redundancy, this was also observed in *E. coli*, and even in Mycoplasmas (Mombach *et al.*, 2005) that have very small metabolic networks. The damage values we observed can be considered small (the maximum value is only 22 compounds); this is a consequence of the property of clusterization of the metabolic network observed by Ravasz *et al.* (2002).

Despite the fact that the metabolic network of *Eucalyptus* is incomplete, this work presents a first step towards the understanding of its overall metabolism. In future studies we intend to complement our methodology with other techniques using flux balance analysis that predict the matter flux through each reaction (Edwards *et al.*, 2001). Another possible extension of our work is to use the technique proposed by Masanori (2004) that considers explicitly the flux of the atoms participating in a given reaction.

Acknowledgements

We would like to thank Prof. Renata Vieira for her help and CNPq and FAPERGS for their support. We would also like to thank the Eucalyptus Genome Sequencing Project Consortium, the KEGG, and PEC databases for providing free access to their data.

References

Devos D and Valencia A (2001) Intrinsic errors in genome annotation. Trends Genet 17:429-431.

Edwards JS, Ibarra RU and Palsson BO (2001) *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. Nat Biotechnol 19:125-130.

Edwards JS and Palsson BO (1997) How will bioinformatics influence metabolic engineering. Biotechnol Bioeng 58:162-169.

Kanehisa M and Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28:27-30.

Karp PD, Krummenacker M, Paley S and Wagg J (1999) Integrated pathway-genome databases and their role in drug discovery. Trends Biotechnol 17:275-281.

Lemke N, Herédia F, Barcellos CK, Reis AN and Mombach JCM (2004) Essentiality and damage in metabolic networks. Bioinformatics 20:115-119.

Masanori A (2004) The metabolic world of *Escherichia coli* is not small. Proc Natl Acad Sci USA 101:1543-1547.

Mombach JCM, Lemke N, Machado da Silva N, Ferreira RA, Isaia EF, Barcellos CK and Ormazabal RJ (2004) Bioinformatics analysis of mycoplasma metabolism: Important enzymes, metabolic similarities, and redundancy. Accepted for publication in Comput Biol Med.

Ravasz E, Somera AL, Mongru DU, Oltvai ZN and Barabási A-L (2002) Hierarchical organization of modularity in metabolic networks. Science 297:1551-1554.

Rocha PCR and Danchin A (2003) Essentiality, not expressiveness drives gene-strand bias in bacteria. Nat Genet 34:377-378.

Associate Editor: Claudia Monteiro-Vitorello