Using a reconstructed flavonoid subnetwork to study anthocyanin biosynthesis

Clark, S. and W. Verwoerd

Centre for Advanced Computational Solutions, Cell Biology Group,
Agriculture and Life Sciences Division, Lincoln University, Christchurch
Email: clarks3@lincoln.ac.nz

Keywords: *anthocyanins, elementary modes, reaction participation, minimal cut sets, fragility coefficients*

EXTENDED ABSTRACT

Flavonoids are ubiquitous secondary plant metabolites that play a variety of roles in the reproduction and protection of plants. Anthocyanin is a major subgroup of flavonoids that assist plants in attracting pollinators and seed dispersers by providing red to blue pigmentation in flowers and fruits. The compounds are water soluble and occur mostly as glucosylated pigments in fruits, leaves and flowers.

A self-contained flavonoid subnetwork, consisting of 137 metabolites and 117 reactions, is extracted from the AraCyc database which contains biochemical pathway information of the model plant *Arabidopsis thaliana* (*Arabidopsis*). *Arabidopsis* is used in this study because of the vast metabolic information available for it as well as the fact that it is very similar to most other plants so should make a good representation of flowering plants.

Using the stoichiometric matrix to mathematically represent the connections between the reactions, and convex analysis to identify all possible and feasible metabolic routes at a steady state, 199 elementary modes (EMs) or pathways are derived from the subnetwork. Eighty of these lead to the formation of flavonoid compounds. The rest lead to other compounds such as lignin and phenylpropanoid esters, whose association and interplay with flavonoid production is an interesting result to be further investigated.

The study uses the subnetwork to investigate the structural functionality of the anthocyanin biosynthetic pathway (ABP). Two anthocyanin compounds, pelargonidin and cyanidin glucosides, are present in *Arabidopsis* and six EMs lead to their formation. By identifying the enzymes related to the reactions involved in the six EMs, the structural functionality of the enzyme related genes in the ABP are studied. Genes that play important roles in multiple phenotypic traits are identified and their relationship with other EMs in the flavonoid subnetwork investigated. Analysis is also done to identify and study those genes which, when deleted, would result in the non-production of anthocyanin compounds. Emergent properties of the anthocyanin biosynthetic pathway, such as reaction participation and minimal cut sets are used for the investigations.

The reaction participation looks at the multiple phenotypic trait of the anthocyanin biosynthetic pathway genes by determining the number of times the enzyme product of each gene appears in the set of 80 elementary modes responsible for the formation of flavonoid compounds. The results show that, in terms of their sequence in the pathway, genes that occur early in the pathway are involved in more EMs, and thus the formation of other flavonoids, than those found later in the ABP, which are more specific to the anthocyanin compounds.

Minimal cut sets (MCSs) are used to identify the sets of reactions which, when removed from the network, ensure that no feasible balanced flux distribution (or EM) involves reactions that form the anthocyanin compounds. In effect, these MCSs provide the full set of candidates for genetic changes that are needed to eliminate anthocyanin production.

The minimal cut sets are also analysed to investigate the effect each has on the remaining non-target set of 74 flavonoid modes, as well as to determine the fragility of the anthocyanin biosynthetic pathway. Results suggest that the ABP is quite fragile which makes sense considering that there are only six EMs involved.

A general observation about the flavonoid subnetwork is that it has a remarkably small number of elementary modes compared to other metabolic networks of similar size in other organisms. This suggests a highly constrained network, which could be due to fact that it deals with secondary metabolites. The results are discussed in more detail in the main paper.

1. INTRODUCTION

The convergent evolution of phenotypes is believed to be a result of parallel selection pressures exerted by similar environmental conditions (Orr 2005). The example studied here is the common loss of the expression of particular genes when independent populations of flowering plants lose their pigmentation.

A study of the expression of genes active in the anthocyanin biosynthetic pathway (ABP) of *Aquilegia* was carried out by Whitall et al. (2006). Their investigation started with the experimental and observational study of genes involved in the convergent loss of floral anthocyanins in independent populations of *Aquilegia* and then used the results to make predictions about the metabolism (network) of the organism; for example, why the genes concerned were being targeted for mutations rather than others.

Our approach (although based on *Arabidopsis*) complements the observational approach of Whittall et al. (2006) in several ways. Rather than observing genes and inferring network properties, the approach starts from a known self-contained network containing all reactions necessary for the synthesis of the anthocyanin compounds. Without needing knowledge about the genes involved, metabolic pathway analysis (Schuster et al. 1999; Schuster et al. 2000), a constraints-based mathematical modelling approach, is used to determine elementary modes that constitute the anthocyanin biosynthetic pathway (ABP). Details of the process is covered in Section 2.

Elementary modes (EMs) are the minimal sets of enzyme reactions that allow the network to operate at a steady state (Schuster et al. 2002; Stelling et al. 2002). They represent non-decomposable pathways or routes that lead from an external substrate to an external product with the network maintained at a steady state, so they are based on the structure of the network and thus defined in the context of whole-cell metabolism. Using the EMs, emergent properties of the network are studied to gain a deeper understanding of the structural functionality of the ABP and deduce the roles of the related core genes and how their mutations would affect other flavonoid compounds. Section 3 discusses these aspects.

The fact that this EM-based study performs a comprehensive mathematical analysis of the network structure, means that it predicts all possible modes or pathways compatible with this structure. Such exhaustive characterisation is very hard to achieve experimentally. However many examples of a phenomenon one has observed, there might always still be others not yet observed. It is this aspect of completeness, subject only to a complete knowledge of the network itself, that makes it possible to make quantitative assessments e.g. of the relative importance of genes or reactions.

2. THE ANTHOCYANIN BIOSYNTHETIC PATHWAY

The subnetwork needed to study the anthocyanin biosynthesis is reconstructed for *Arabidopsis* using the biochemical pathway information in the AraCyc database (Rhee et al. 2003). Applying the method described in the previous paper (Verwoerd 2007) to AraCyc version 4.0, the smallest self-contained subnetwork containing anthocyanin biosynthesis in *Arabidopsis*, consists of 137 metabolites and 117 reactions.

This is also the flavonoid subnetwork because all flavonoids, including anthocyanins, originate from the combination of coumaryl-CoA and malonyl-CoA to yield chalcones that undergo a series of enzymatic modifications resulting in their production (Winkel-Shirley 2001). The set of EMs that contain both coumaryl-CoA and malonyl-CoA constitutes all pathways that lead to the formation of flavonoid compounds in *Arabidopsis*, including the anthocyanins. Figure 1 shows the full subnetwork and it is seen to be quite a complex structure. A detailed version of the subnetwork that is relevant to anthocyanin production is shown in Figure 2. Analysis work is carried out using the CellNetAnalyzer software program (Klamt 2006).

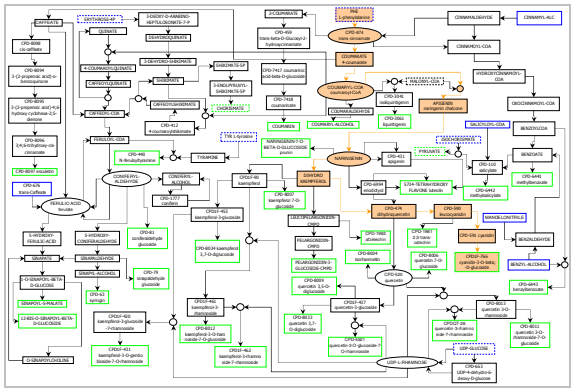

Figure 1. The *Arabidopsis* flavonoid subnetwork. Highlighted is one of the four EMs for the anthocyanin compound - cyanidin-3-O-beta-D-glucoside. A close-up of the highlighted EM is shown in Figure 2.

Two unusual properties of our reconstructed flavonoid subnetwork are that there are no

reversible reactions, and that the number of metabolites is greater than the number of reactions which suggests a non-redundancy of the network. In most cases the number of reactions exceeds the number of metabolites (Schilling and Palsson 2000; Stelling et al. 2002).

The resulting subnetwork generates a total of 199 EMs which is remarkably low when compared with other studied networks of similar numbers of compounds, e.g., the 2.4 million EMs from 112 reactions and 89 compounds in *Escherichia coli* (Gagneur and Klamt 2004). It indicates that the flavonoid network is highly constrained. That makes sense, bearing in mind that the number of reactions determines the dimension of the EM subspace while the number of internal compounds represent the number of constraints.

A biological interpretation of this observation might be the fact that it comprises mainly of reactions necessary for the formation of flavonoids which are secondary metabolites, so the reactions are more specific and not as many as there would be for the more essential primary compounds.

Of the total 199 EMs, 80 form flavonoid compounds and the rest produce other compounds such as lignin and phenylpropanoid esters. The 80 flavonoid EMs can be further classified according to the flavonoid subclass with which they are associated, as shown in Table 1.

Table 1. EM distribution: flavonoid subclasses

Flavonoids	EMs	Flavonoids	EMs
Flavone	8	Anthocyanin	6
Pro-anthocyanidin	6	Flavanone	4
Flavonol	56	**Total EMs**	**80**

The rest of this paper is mainly concerned with the 6 anthocyanin modes, of which an example was highlighted in Figure 1. Figure 2 shows the subset of nodes that are involved in these 6 EMs, and can be considered as an extended version of the simple linear ABP shown in (Whittall et al. 2006). Note that this is not a coherent self-contained subnetwork in the sense of (Verwoerd 2007) but merely an extract for display purposes.

In addition to the six ABP core genes specified by Whitall et al (2006), there are additional genes that play a role in the pathway. Two of these in *Aquilegia*, F3'H and F3'5'H, affect the class of anthocyanins produced, cyanidins (red) and delphinidins (blue) respectively, with pelargonidins being produced if neither of the genes are active (Whittall et al. 2006). As shown in Figure 2, the F3'H gene is also present in *Arabidopsis* and is important in the formation of the anthocyanin compound cyanidin-3-O-β-glucoside, while F3'5'H is not present.

Additional genes such as LDOX and 3GT found in *Arabidopsis* are shown on the extracted ABP and represented by their equivalent gene (catalysing the same reaction in the network) in the analyses results, e.g., LDOX and ANS are represented as ANS in the analyses; UF3GT and 3GT as UF3GT.

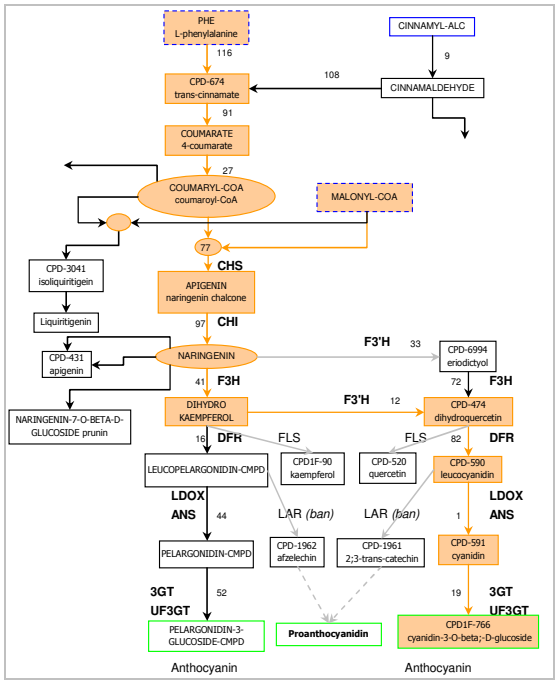

Figure 2. The ABP extracted from the reconstructed flavonoid subnetwork. The shaded part relates to the shaded anthocyanin EM in Figure 1 and the numbers relate to the *Rxn nos* in Table 2 and *R* in Tables 4 - 7

Genes such as FLS and LAR are related to the production of other flavonoid compounds that are not related to floral functions but responsible for other functions such as UV-protection, herbivore and pathogen resistance and so on. They are, therefore, not included in the analyses.

3. EMERGENT NETWORK PROPERTIES

The following network analyses involve quantitatively characterising the importance of enzymes and reactions at an increasing level of detail.

3.1. Genes and Reaction Participation

First, we start off by studying the most global system of classification for genes, their participation (Papin et al. 2002) in the flavonoid subnetwork. That is shown in Figure 3 as the

number of times the enzyme product of each ABP gene appears in the set of 80 elementary modes. The number of modes that would remain unaffected by blocking each gene is simply obtained from the figure by subtracting the participation number from 80.

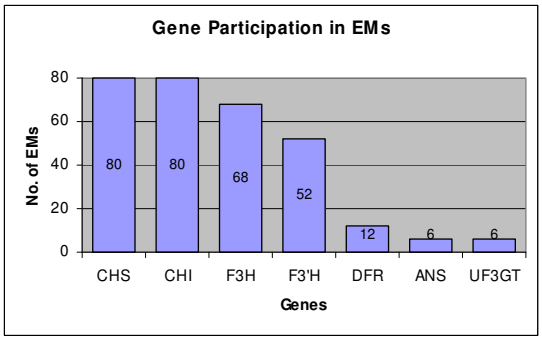

Figure 3. Participation of reactions related to APB genes/enzymes in the whole subnetwork.

The genes on the graph correspond to their position in the ABP (Figure 2) with the left hand side gene (CHS) on the graph being the earliest in the pathway and that on the right hand side (UF3GT) being the latest.

Somewhat more detail is added to this picture by noting in Table 2 the participation numbers of individual reactions in the set of pathways.

Table 2. Participation counts of reactions in the subnetwork.

Rxn No in Fig.2	Genes/ Enz	Partic pwys	Rxn No in Fig.2	Genes/ Enz	Partic pwys
1	ANS	4	52	UF3GT	2
12	F3'H	24	57	CHS	2
16	DFR	4	72	F3H	24
19	UF3GT	4	77	CHS	78
33	F3'H	28	81	CHI	2
41	F3H	44	82	DFR	8
44	ANS	2	97	CHI	78

As indicated in Figure 3, CHS and CHI at the start of the ABP have reactions participating in all 80 flavonoid pathways (EMs) with a decrease in participation when moving along the pathway. The genes ANS and UF3GT furthest down the pathway participate in only six EMs. This means that blocking genes further down the ABP does not affect as many EMs as blocking those earlier in the pathway. That is the argument used in general terms by (Whittall et al. 2006) but here we can quantify the difference.

In particular, some ambiguity was expressed by (Whittall et al. 2006) on whether, in terms of their sequence in the pathway, gene F3H should be classified as an 'early' or 'late' gene; here, its participation in 68 pathways, puts it closer to the 'early' genes (80 EMs) than the 'late' genes participating in 12 or fewer EMs.

Enzyme subsets is another useful tool in elementary mode analysis. These subsets contain reactions showing the same values in each flux distribution obeying the steady state condition (Pfeiffer et al. 1999), so they always have to operate together and structurally need each other. For the ABP we find two pairs of genes that form enzyme subsets. The subsets are the CHS and CHI pair early in the ABP, and the ANS and UF3GT pair lower down the pathway (refer to Figure2).

Such subsets have previously been found to indicate co-regulation (Klamt and Stelling 2003). However, Whittall et al (2006) observed co-regulation of ANS & DFR rather than the subset ANS & UF3GT we find. The reason that DFR and ANS do not belong to the same enzyme subsets becomes clear from Table 3.

Table 3. EMs that rely on DFR

EM	Product	Compound
1, 2, 3, 4	cyanidin glucoside	anthocyanin
5, 6, 7, 8	trans-catechin	
9, 10	pelargonidin glucoside	anthocyanin
11, 12	afzelechin	

DFR occurs in twelve EMs, six of which are related to the production of anthocyanin and six to the monomers, afzelechin and trans-catechin, responsible for the formation of proanthocyanidin. Thus if the DFR-related gene was removed, not only the formation of the anthocyanin compounds is affected but proanthocyanidin production as well. The discrepancy could be because proanthocyanidin is not produced in *Aquilegia* and DFR is only related to anthocyanin formation in that species.

3.2. Minimal Cut Sets

In the next level of detail, we move on to studying combinations of reactions in the form of minimal cut sets (MCSs) (Klamt and Gilles 2004). These are the sets of reactions which (in relation to a defined target reaction) when removed from the network, ensure that no feasible balanced flux distribution (or EM) involves the target reaction.

This concept is highly relevant to loss-of-function mutations such as the loss of pigmentation in *Aquilegia*. Each minimal cut set can be seen as a candidate modification that will lead to the loss of the function associated with the target reaction. Because the mathematics guarantees that the

collection of MCS is complete, quantitative comparisons can be made to identify the cut set that will achieve the function loss most effectively.

There are two reactions responsible for forming anthocyanins in our network- reactions R 19 that forms cyanidin-3-O-beta;-D-glucoside and R 52 which forms pelargonidin-3-glucoside-cmpd (refer to R 19 and 52 in Figure 2). These reactions (highlighted in grey in Tables 4 - 7) are chosen as the target set. The MCSs are determined such that they will remove both *R* 19 & 52 from the feasible steady state, or equivalently block the six anthocyanin EMs. Table 4 shows the composition of the six anthocyanin EMs. Note that the reactions *R* 9, 27, 91, 108 and 116 in Tables, 4, 5, 6 & 7 are the reactions that occur prior to the combination of coumaryl-CoA and malonyl-CoA compounds at the beginning of the ABP and thus do not involve ABP genes.

The 26 MCSs in Table 5 provide the full set of candidates, for genetic changes that are needed to eliminate anthocyanin production. Figure 4 demonstrates the effect that each of these cut sets has on the remaining set of 74 flavonoid modes that are not the anthocyanin EMs being blocked. It provides a means of prioritising MCS candidates by their effect on other processes.

At the top of the list, MCS5, 13, 14 & 15 have seventy four unaffected modes and thus do not affect any other EMs, so are the best MCS to delete for eliminating anthocyanin production. Comparing Tables 5 and 2 shows that they respectively require UF3GT, ANS, ANS&UF3GT and ANS&UF3GT to be blocked. All of these would be equivalent according to the "minimum disturbance" principle invoked previously, but a second consideration is that blocking an 'early' gene avoids wasting resources on producing intermediate enzymes or compounds unnecessarily. This argues for MCS14, and/or MCS15 and MCS16 as the optimal cut sets, with ANS as the key enzyme to be suppressed.

The next best cut sets are MCS8 and MCS11, which both need DFR to be blocked in addition to either UF3GT or ANS. While this eliminates two additional modes, it may compensate for this by moving the cut higher up the pathway. Similarly, the next two levels represented by MCS19 and 20, then MCS17 involve DFR on its own or with ANS/UF3GT. After that, there is quite a big step in the number of affected modes to reach MCS9, 12 and 18; all of which involve F3H in addition. So this more detailed analysis confirms our earlier conclusion that ANS and DFR are the prime targets for anthocyanin suppression, with F3H as a runner up; in agreement also with the *Aquilegia* observations.

Table 4. Anthocyanin elementary modes
R - the reaction nos. in Figure 2;
EMs (1, 2, 5) ≡ EM Nos in Table 3;
x – indicates presence of reaction *R* in the EM.

R	1	9	12	16	19	27	33	41	44	52	72	77	82	91	97	108	116
EM																	
1	x		x		x	x		x				x	x	x	x		x
2	x				x	x	x				x	x	x	x	x		x
3	x	x	x		x	x		x				x	x	x	x	x	
4	x	x			x	x	x				x	x	x	x	x	x	
9				x		x		x	x	x		x		x	x		x
10		x		x		x		x	x	x		x		x	x	x	

Table 5. Minimal cut sets (MCSs) of the three anthocyanin EMs in Table 4.
x – essential reaction (*R*) for the MCS.

R	1	9	12	16	19	27	33	41	44	52	72	77	82	91	97	108	116
MCS																	
1						x											
2														x			
3																x	x
4		x															x
5					x					x							
6												x					
7															x		
8	x			x													
9	x							x									
10							x	x									
11				x	x												
12					x			x									
13	x								x								
14					x				x								
15	x									x							
16								x			x						
17				x									x				
18								x					x				
19									x				x				
20										x			x				
21			x	x			x										
22			x				x		x								
23			x				x			x							
24			x	x							x						
25			x						x		x						
26			x							x	x						

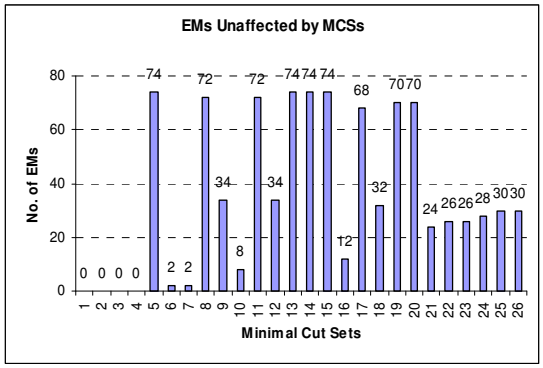

Figure 4. Effect of MCSs on remaining 74 flavonoid EMs.
MCS Nos. ≡ Nos. in Table 5.

3.3. Fragility Coefficients

The previous sections relied on counting of affected modes to prioritise genes; but it is conceivable, for example, that a gene involved in

only a small number of modes may nevertheless be crucial for those. To address that in terms of network structure, the final level of detail in the study introduces a measure of how essential the reactions are for a specific objective.

This is the fragility coefficient f_R of reaction R, defined as the average of the reciprocals of the number of reactions participating in a MCS (Klamt and Gilles 2004). If the reaction is part of a larger MCS, its malfunction will be less crucial for the objective reaction.

Table 6. The calculated fragility coefficient f_R for the ABP reactions R from Figure 2.

R	1	9	12	16	19	27	33	41	44	52	72	77	82	91	97	108	116
f_R	0.50	0.50	0.33	0.43	0.50	1.00	0.38	0.50	0.43	0.43	0.38	1.00	0.50	1.00	1.00	0.50	0.50

To generalise the concept to genes, we take the average for all reactions catalysed by the corresponding enzyme. This is shown as F in **Table 7** below.

Table 7. Fragility coefficients, F, of ABP genes showing related reactions, R, from Figure 2.

R	77	97	1, 44	16, 82	19, 52	41, 72	12, 33
Genes	CHS	CHI	ANS	DFR	UF3GT	F3H	F3'H
F	1	1	0.5	0.5	0.47	0.44	0.35

The gene entries in **Table 7** are sorted by descending fragility values. This shows CHS and CHI as crucial to the ABP, then, in decreasing order, ANS or DFR, UF3GT and F3H, while F3'H is noticeably less important for anthocyanin formation.

The overall network fragility coefficient F (averaged over all its reactions) is 0.60 so the ABP is not very robust. This makes sense considering that it only contains six EMs.

4. CONCLUSIONS

Isolation of the flavonoid subnetwork and its analysis based on the calculated elementary modes, gives quite a detailed picture of the roles played by the genes, enzymes and reactions involved in anthocyanin production.

Overall conclusions, based entirely on the known metabolic network structure of *Arabidopsis Thaliana*, are that (i) The early enzymes CHS and CHI are essential for flavonoid production; (ii) In contrast, the enzymes DFR, ANS and UF3GT furthest down the ABP only occur in EMs leading to the formation of anthocyanin compounds; (iii) The enzyme F3H is involved in many modes, similar to early enzymes, but in terms of the effects of its suppression appears to behave rather more like a late enzyme.

These conclusions are supported by the experimental findings of (Whittall et al. 2006) on independent losses of floral anthocyanins in *Aquilegia*. One minor difference is that the experimental results indicate coregulation of DFR and ANS, while these enzymes do not form an enzyme subset in *Arabidopsis* as was expected on the basis of previously found correspondence between enzyme subsets and coregulation.

What the network analysis adds to the picture is a more quantitative and therefore more differentiated characterisation of the enzyme roles. A key factor in this is the exhaustive nature of the mathematical analysis of network structure. For example, multiple cut sets (MCS) supply a complete listing of all options open to the plant for suppressing anthocyanin production, as well as characterisation of the impact each has on production of other flavonoids. This gives independent insight into the reasons for the observed molecular convergence in the evolution of floral pigmentation losses.

In their work, (Whittall et al. 2006) also address the issue of whether this loss of function takes place through mutation of primary genes or via changes in regulation. This matter cannot be clarified by network analysis as discussed here, as it is only the functional presence or absence of a reaction that determines if a particular elementary mode is active, irrespective of the mechanism responsible for its suppression.

Some general observations about the flavonoid subnetwork also follow from our study. In comparison to other metabolic networks of similar size in other organisms, the flavonoid subnetwork has a remarkably small number of elementary modes. This means that it is a highly constrained network; and this is also reflected in the high fragility number calculated for the anthocyanin part of the subnetwork. A plausible explanation of this phenomenon would be that, because this subnetwork is concerned with secondary metabolites, there is not such a strong biological need for robustness and redundancy as in more central parts of metabolism.

5. REFERENCES

Gagneur, J. and S. Klamt (2004), Computation of elementary modes: a unifying framework and the new binary approach, *BMC Bioinformatics* 5(175).

Klamt, S. (2006). CellNetAnalyzer.

Klamt, S. and E. Gilles (2004), Minimal cut sets in biochemical reaction networks, *Bioinformatics* 20, 226 - 234.

Klamt, S. and J. Stelling (2003), Two approaches for metabolic pathway analysis?, *Trends in Biotechnology* 21(2), 64-69.

Orr, H. A. (2005), The probability of parallel evolution, *Evolution* 59(1), 216-220.

Papin, J. A., N. D. Price and B. O. Palsson (2002), Extreme pathway lengths and reaction participation in genome-scale metabolic networks, *Genome Research* 12(12), 1889-1900.

Pfeiffer, T., I. Sanchez-Valdenebro, J. C. Nuno, F. Montero and S. Schuster (1999), METATOOL: for studying metabolic networks, *Bioinformatics* 15(3), 251-257.

Rhee, S. Y., W. Beavis, T. Z. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander, M. Montoya, N. Miller, L. A. Mueller, S. Mundodi, L. Reiser, J. Tacklind, D. C. Weems, Y. Wu, I. Xu, D. Yoo, J. Yoon and P. Zhang (2003), The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community, *Nucleic Acids Research* 31(1), 224-228.

Schilling, C. H. and B. O. Palsson (2000), Assessment of the metabolic capabilities of Haemophilus influenzae Rd through a genome-scale pathway analysis, *Journal of Theoretical Biology* 203(3), 249-283.

Schuster, S., T. Dandekar and D. A. Fell (1999), Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering, *Trends in Biotechnology* 17, 53-60.

Schuster, S., D. Fell and T. Dandekar (2000), A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks, *Nat Biotechnol* 18, 326 - 332.

Schuster, S., C. Hilgetag, J. H. Woods and D. A. Fell (2002), Reaction routes in biochemical reaction systems: Algebraic properties, validated calculation procedure and example from nucleotide metabolism, *Journal of Mathematical Biology* 45(2), 153-181.

Stelling, J., S. Klamt, K. Bettenbrock, S. Schuster and E. Gilles (2002), Metabolic network structure determines key aspects of functionality and regulation, *Nature* 420, 190 - 193.

Verwoerd, W. S. (2007), Identifying coherent subnetworks in genome scale metabolic networks. MODSIM07, Christchurch.

Whittall, J. B., C. Voelckel, D. J. Kliebenstein and S. A. Hodges (2006), Convergence, constraint and the role of gene expression during adaptive radiation: floral anthocyanins in Aquilegia, *Molecular Ecology* 15(14), 4645-4657.

Winkel-Shirley, B. (2001), Flavonoid Biosynthesis. A Colorful Model for Genetics, Biochemistry, Cell Biology, and Biotechnology, *Plant Physiol.* 126(2), 485-493.