A PATHWAY APPROACH TO ALIGN REGULATORY-METABOLIC NETWORKS

Y. LI[1,2,3,*], J.J. BOT[1,2], M.J.T. REINDERS[1,2,3] and D. DE RIDDER[1,2,3]

[1] *Information & Communication Theory Group, Faculty of Electrical Eng., Mathematics & Computer Science, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands*

[2] *Netherlands Bioinformatics Centre, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands*

[3] *Kluyver Centre for Genomics of Industrial Fermentation, Julianalaan 67, 2628 BC Delft, The Netherlands*

[*] *Email: y.li@tudelft.nl*

Integrating different types of biological networks and aligning networks across species are two useful but challenging comparative methods in systems biology nowadays. By combining these in one framework, we can expect to generate more reliable information and hypotheses. In this study, we systematically integrate the transcriptional regulation network of enzyme-coding genes and the corresponding metabolic network, and align these integrated networks between two species. By applying a comprehensive yet flexible scoring function to measure the alignment similarity, our method can be used to identify conserved elements (allowing for small variations) of evolution at both the regulatory and metabolic level, to reveal the interrelation and divergence between species and to use information at one level to predict missing information at the other level.

1. INTRODUCTION

Most metabolic reactions in cells are catalyzed by enzymes, and the genes which code for these enzymes are regulated by transcription factors (TFs). That is, TFs can bind to the promoter sequence of genes and subsequently activate or repress the transcription of these genes. This information flow from the regulatory level to the metabolic level is illustrated in Figure 1a. At each level, these interactions form a network, i.e. a transcriptional regulatory network and a metabolic network, respectively.

Comparing networks between species at each level individually can help to filter noise, and produce insights into the principles governing evolution. For example, Gasch *et al.*[1] found that many of the known cis-regulatory systems in *Saccharomyces cerevisiae* (yeast) have been conserved in 13 ancient fungi species. Tanay *et al.*[2] studied the promoter evolution of co-regulated genes in 17 yeast species, and suggested an intermediate redundant regulatory program underling the evolvability and increased redundancy of transcriptional regulation in higher organisms. Alkema *et al.*[3] improved the prediction of co-regulated genes based on the conservation of protein sequences and regulatory mechanisms. At the metabolic level, Jeong *et al.*[4] and Ravasz *et al.*[5] studied the global topological properties of the metabolic networks in 43 species. Heymans *et al.*[6] derived phylogenetic trees based on metabolic pathway comparison.

Comparing networks at different levels simultaneously can be even more informative. Since different types of network present different perspectives on the biological system, integrating them may offer a more comprehensive picture. Particularly when elements are conserved at multiple levels, we can be more confident about the reliability of the observed conservation. This allows us to make predictions, using information at one level to infer information at another level, or using information of one species to infer information for another species.

Although integrating different types of network within one species has received quite some attention[7–10], little advances have been made on the alignment of regulatory and metabolic networks across species. Here we present a method that searches for network elements that are conserved in evolution at both the regulatory and metabolic level, and measures the extent of this conservation. A schematic overview of our goal is given in Figure 1a.

Previously we developed M-PAS[11], a framework for metabolic pathway alignment and scoring based on the notion of building blocks (see Figure 2), to align the metabolic networks of *Saccharomyces cere-*

[*]Corresponding author.

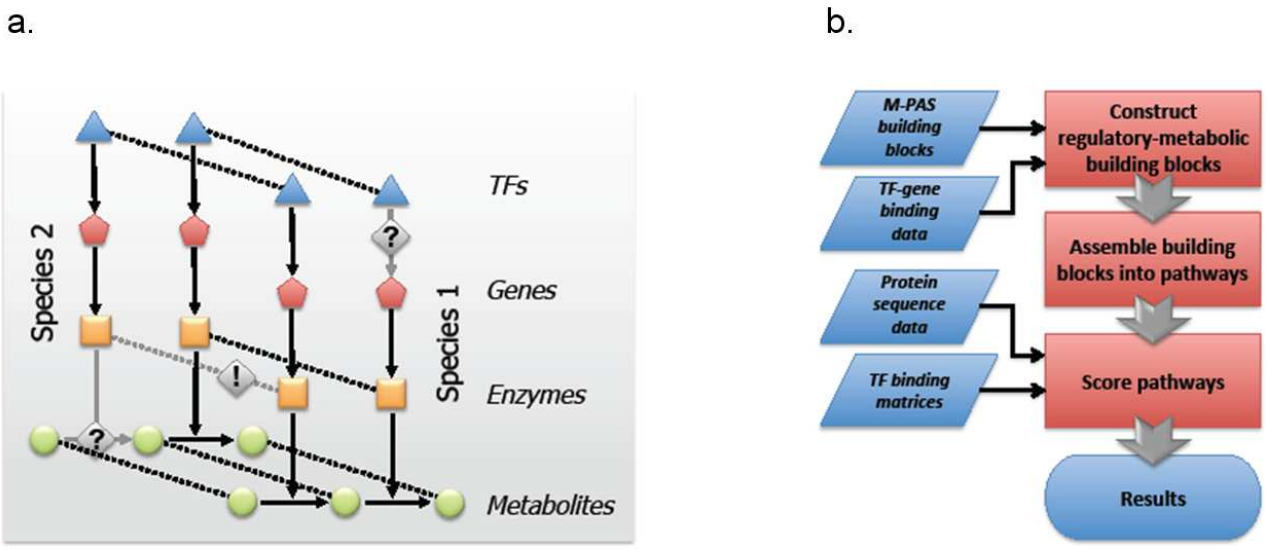

Fig. 1. Method overview. **a)** The goal of our method is to align metabolic pathways and their regulation between two species, using suitably defined similarity measures between compounds, enzymes and transcription factors (illustrated by the dotted lines), in order to find conserved elements and learn about differences between species (illustrated by the exclamation mark). From missing links in an otherwise conserved context, we can infer missing reactions or regulation within species (illustrated by the question marks). **b)** The RM-PAS flowchart.

visiae and *Escherichia coli*. In the current work, we integrate TF-gene interaction and TF binding site (TFBS) information into M-PAS, and form a more comprehensive method, RM-PAS. We applied RM-PAS to *S. cerevisiae* and *E. coli*, two of the best-annotated model organisms, with relatively much TF binding and TFBS data available. Since these species are not closely related, many differences are expected, and the resulting conservation is expected to be quite informative.

2. Methods

The building block method used in M-PAS has shown to be an appropriate approach to align metabolic pathways[11, 12]. It is described briefly in Figure 2 and Appendix A. First, it is able to explore topological arrangement possibilities of reactions both between species (by building block construction) and within species (by pathway assembly). Second, by defining building blocks, we can focus on conserved pathways while allowing small variations. Third, the method is adaptable and can easily be extended to include more information.

Here, we extend the building block construction and the scoring function to include transcriptional regulation information. That is, for every enzyme in a reaction, we add the transcription factors that regulate the enzyme-coding genes. In the end, we consider the building blocks be the *conserved elements* that we are interested in. The flowchart is given in Figure 1b and will be explained in the remaining of this section. Note that given curated databases (see section 3) and user-defined parameters as input, each step in the flowchart is automated.

2.1. Regulatory-metabolic building blocks

We add transcriptional regulation to the metabolic building blocks in M-PAS, to construct regulatory-metabolic (RM) building blocks. That is, we add a link between a transcription factor and the enzyme in the reaction. This is only done when there is experimental evidence showing that the transcription factor indeed regulates the gene coding for the enzyme.

Like in the metabolic building block approach, we also categorize the RM building blocks with different TF regulation scenarios in the two species, as well as different TF similarity scenarios, i.e. (1) whether the TFs which bind to the enzyme-coding genes are

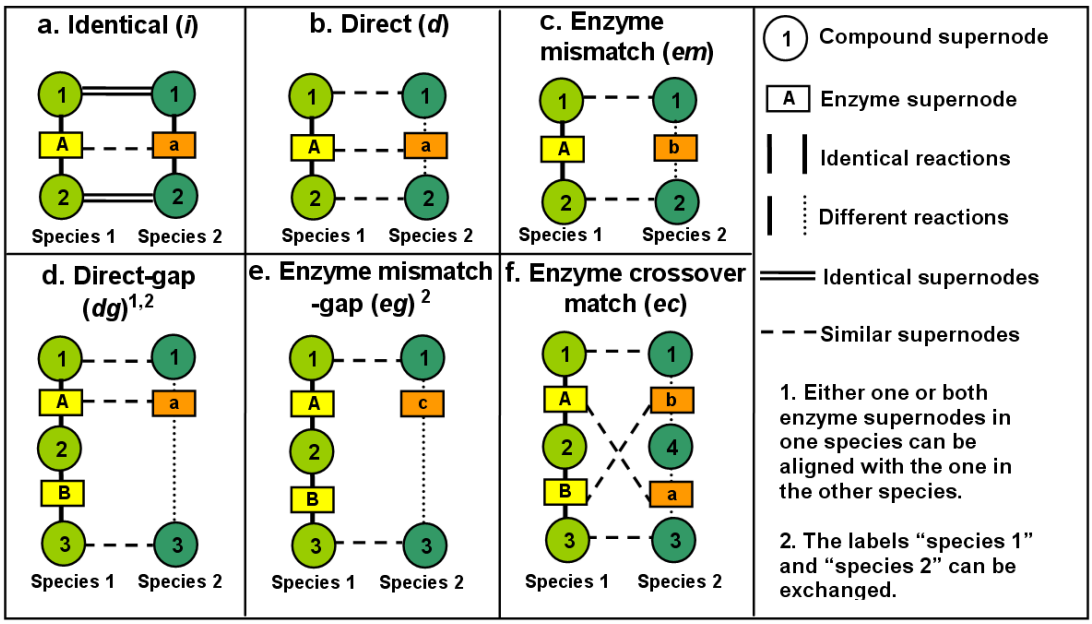

Fig. 2. Illustration of the six types of metabolic building blocks. A metabolic building block is formed if two reactions from two species transform the same substrate into the same product, by **a)** the same reaction which is present in both species, or **b-c)** different reactions with similar or dissimilar enzymes, or **d-e)** different number of reactions in two species, or **f)** different sequential order of the transformation. Note the reaction directions are omitted for simplicity. A *compound supernode* is the set of all substrates or products in a reaction. *Enzyme supernodes* are defined similarly. Two compound supernodes are considered similar if they share at least one common compound. Two enzyme supernodes are considered similar if there exists a pair of enzymes sharing the same first two digits in their EC numbers.

similar ("direct TF") or dissimilar ("mismatch TF"), and (2) whether there exist additional TFs ("alternative TF") in one species which are similar to the TFs in another species, but are not found to bind to the genes in that reaction. When one species has neither bound TFs nor alternative TFs, we call the RM building block has "absent TF" in that species. The seven possible cases where TFs are added to the metabolic building blocks are shown in Figure 3, cases 1-7.

In addition to the reactions present in the database, we also look for possible reactions which are currently missing in one of the species ("missing"). In this scenario, one reaction is present in only one species, but the other species does contain the reaction's compounds and enzymes with identical function in terms of EC number. An RM building block is then constructed when there is evidence from the transcriptional regulation control indicating that the missing reaction might be present. That is, when there exist "direct" and/or "alternative" TFs, we hypothesize the reaction might exist in both species. These three cases are shown in Figure 3, cases 8-10.

2.2. Pathway assembly

After building blocks are constructed, they are concatenated into pathways, if the product of the upstream building block is the substrate of the immediate downstream building block. Since we are interested in small differences (as illustrated in Figure 2 and Figure 3), instead of generating a few highly conserved longer pathways, we generate a ranked list of short pathways with the same length. Because the amount of overlap between pathways increases substantially when pathway length increases, we limit each pathway to contain four building blocks.

To implement an exhaustive search for all length-four pathways, we start a backtracking search from each substrate. During the search, all building blocks in a pathway should be different, and one reaction cannot appear more than once in one species. Note that twenty-six currency metabolites (ATP, ADP, UTP, UDP, GTP, GDP, AMP, UMP, GMP, NAD, NADH, NADP, NADPH, acetyl-CoA,

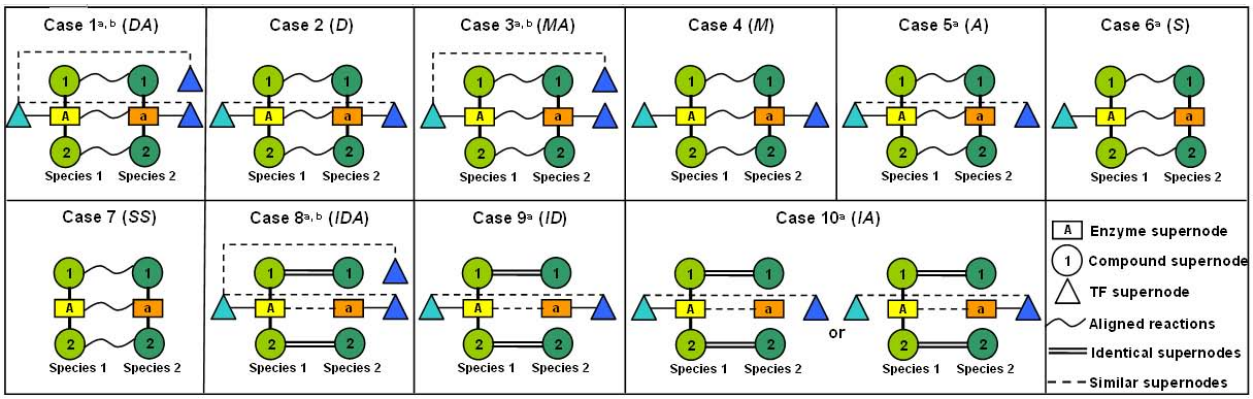

Fig. 3. Illustration of the ten cases of RM building blocks. ***D***: direct TF. ***A***: alternative TF. ***M***: mismatch TF. ***S***: absent TF. ***I***: missing reaction. [a] The labels "species 1" and "species 2" can be exchanged. [b] The alternative TF can be present in one or both species. Aligned reactions denote any of the six types of metabolic building blocks (see Figure 2). A *TF supernode* is the set of TFs which bind to the enzyme-coding genes in a reaction. Two TF supernodes are considered similar if their TFBS are more similar than average, i.e. for "direct TF": $Z_{TB}(B) > 0$ (Eq. 6), and for "alternative TF": $Z_{TU}(B) > 0$ (Eq. 8). In cases 8-10, the two enzyme supernodes have the same EC number.

CoA, propanoyl-CoA, L-glutamine, L-glutamate, 2-oxoglutarate, CTP, CDP, CMP, H_2O, CO_2, NH_2, and phosphate) are excluded from consideration during pathway assembly to avoid finding large numbers of pathway shortcuts[5, 13]. That is, we do not match or connect two reactions if they *only* share the same currency metabolites.

2.3. Scoring function

We rank the aligned pathways according to the extent of conservation, in order to prioritize the interesting pathways for further investigation. The M-PAS scoring function[11] integrates multiple similarity scores of all reaction components. It has a generic form and is capable of measuring pathway similarity given different biological emphases. This allows user to specifically look for certain characteristic differences between species in otherwise highly conserved pathways: by setting the appropriate parameters, differences will be allowed between enzymes, compounds and/or TFs. Due to its hierarchical integration structure, it is readily extensible to include other relevant similarity measures. In this study, the M-PAS scoring function[11] is adapted such that transcriptional regulation similarities are included.

2.3.1. *Total score*

Our goal is to reflect all aspects of an aligned pathway in the total similarity score. These include similarities at the regulatory level and the metabolic level, i.e. similarities between transcription factors, substrate sets, product sets, enzyme functions, enzyme sequences and alignment topology, respectively.

To account for their diverse distributions of similarities, we first compute similarity scores independently for each aspect, and then convert the raw scores into z-scores before integration. The integration of multiple z-scores is done hierarchically using Liptak-Stouffer's method[14]. In this way, we obtain a decomposable score for a pathway:

$$Z(P) = \frac{1}{\sqrt{N}} \sum_{\forall B \in P} Z(B) \qquad (1)$$
$$= \frac{1}{\sqrt{N}} \sum_{\forall B \in P} \frac{1}{\sqrt{3}} [Z_0(B) + Z_R(B) + Z_T(B)]$$

where $Z(P)$ denotes the total z-score of an aligned pathway P, which contains N building blocks B. $Z_0(B)$ is the user-specified bias for the building block alignment type. For example, if the user is interested in building blocks with gaps, then the building blocks with gaps, i.e. "*direct-gap*" and "*enzyme mismatch-gap*" in Figures 2d-e, can be assigned a large positive bias. $Z_R(B)$ and $Z_T(B)$ denote the reaction and transcription factor similarity z-scores in B. $Z_R(B)$

is discussed in detail in ref. 11, and will be briefly described below and in Appendix A. Here we mainly focus on the TF similarity score.

2.3.2. *Reaction score*

The reaction similarity score $Z_R(B)$ is a weighted sum of its compound score $Z(C_B)$ and enzyme score $Z(E_B)$ (Appendix A):

$$Z_R(B) = \frac{1}{\sqrt{\omega_c^2 + \omega_e^2}} [\omega_c Z(C_B) + \omega_e Z(E_B)] \quad (2)$$

Compound weight ω_c and enzyme weight ω_e can be used to assign different relative importance to compound similarity and enzyme similarity. The compound score $Z(C_B)$ combines the similarities between the substrate sets and between the product sets in a building block B, considering the amount and specificity of the overlapping compounds. The enzyme score $Z(E_B)$ is a weighted sum of a functional similarity score (with weight ω_f) and a sequence similarity score (with weight ω_q).

2.3.3. *Transcription factor score*

We measure TF similarity to see whether regulation is conserved in the two species, and whether we can find possible alternative TFs. Therefore, the TF score contains two parts: (1) the similarity between the bound TFs in two species (Z_{TB}), and (2) the similarity between the bound TFs in one species and TFs that are not found to bind in the other species (Z_{TU}). Weights are given to these two parts for finding different cases in Figure 3. Thus the TF score can be written as an integrated z-score:

$$Z_T(B) = \frac{1}{\sqrt{\omega_{tb}^2 + \omega_{tu}^2}} [\omega_{tb} Z_{TB}(B) + \omega_{tu} Z_{TU}(B)] \quad (3)$$

First, we need to compute the raw similarity scores between TFs. A TF is characterized by its corresponding transcription factor binding site (TFBS), which can be quantitatively described by position weight matrices (PWM) or position frequency matrices (PFM)[15]. We take the standard approach of comparing PWM/PFM profiles[16, 17] to measure the similarities between different TFs in an RM building block. More specifically, we applied MatCompare[17] to calculate the Kullback-Leibler divergence[18] between the PWM/PFM matrices. This measures the information divergence between the matrix entries. If matrices m and m' have w columns, indicating the length of the TFBS sequence, the divergence between them is:

$$D(m, m') = \sum_{i=1}^{w} \sum_{j=A}^{T} (m_{ij} - m'_{ij}) \log(m_{ij}/m'_{ij}) \quad (4)$$

If one of the two matrices has fewer columns, that matrix is compared to all possible starting columns in the other matrix to find the best match.

For a building block B, there might be multiple TFs, each of which might have multiple PWM/PFM matrices. Let M_{B1} and M_{B2} denote the complete set of PWM/PFM matrices of all bound TFs involved in B in the two species, respectively. Then the raw TF similarity between bound TFs is the best match in all pairs of bound TF PWM/PFM matrices:

$$S_{TB}(B) = \max_{m \in M_{B1}, m' \in M_{B2}} -D(m, m') \quad (5)$$

This similarity is further transformed into a z-score:

$$Z_{TB}(B) = \frac{S_{TB}(B) - \mu_{TB}}{\sigma_{TB}} \quad (6)$$

where μ_{TB} and σ_{TB} are the average and standard-deviation of S_{TB} over all possible permuted pairs of M_{B1} and M_{B2}.

Similarly, we compute the raw similarity score between bound TFs in one species and the alternative TFs in the other species, which is the best match in all pairs between bound TF PWM/PFM matrices in one species and the alternative TF PWM/PFM matrices in the other species:

$$S_{TU}(B) = \max\{ \max_{m \in M_{B1}, m' \notin M_{B2}} -D(m, m'), \max_{m \notin M_{B1}, m' \in M_{B2}} -D(m, m') \} \quad (7)$$

$$Z_{TU}(B) = \frac{S_{TU}(B) - \mu_{TU}}{\sigma_{TU}} \quad (8)$$

where μ_{TU} and σ_{TU} are the average and standard-deviation of S_{TU} over all possible permuted pairs of M_{B1} and M_{B2}.

3. Data

Reaction definitions were obtained from Release 42.0 of the KEGG LIGAND composite database[19], updated on Aug. 18, 2008. The

species-specific reactions and enzyme lists were retrieved from KEGG/XML and KEGG/PATHWAY. Protein sequences were downloaded from UniProtKB/SwissProt[20] Release 56.0, updated on July 22, 2008.

For *S. cerevisiae*, the experimentally verified TF-gene binding data is collected from TRANSFAC[21] Release 11.4 and Yeastract[22] version 2008515. The PWM or PFM matrices are obtained from TRANSFAC, Yeastract, SwissRegulon[23], IMD[24], and ooTFD[25].

For *E. coli*, the experimentally verified TF-gene binding data is collected from EcoCyc[26] Release 11.6 and RegulonDB[27] Release 6.0. The TFBS matrices are obtained from RegulonDB and SwissRegulon.

4. Experiment and results

Based on 957 enzymatic reactions in yeast and 1175 enzymatic reactions in *E. coli*, we constructed 697 RM building blocks, including 5 of cases 8-10 in Figure 3. They are assembled into 8397 length-four pathways, starting from 259 substrates.

Here we demonstrate our method using three example queries, to find fully conserved pathways, missing TF-gene bindings, and differences between the regulatory and metabolic level. Each query uses a different parameter setting, including the building block type bias Z_0, four reaction score weights (i.e. ω_c, ω_e, ω_f and ω_q), and two TF score weights (i.e. ω_{tb} and ω_{tu}). In each query, the similarity scores of all pathways found are computed using Eq. 1, and the highest-scoring pathway(s) of a certain substrate is referred as the *best pathway* for that substrate.

Table 1 lists the parameter settings in the queries. The motivations for, and results of the queries are discussed in the following.

4.1. Identifying conserved regulatory-metabolic network elements

In Query 1, all aspects of known information at both the regulatory and the metabolic level are considered. Therefore, the resulting pathways represent elements fully conserved at both levels. Figure 4a gives an example, which is involved in the citrate cycle (TCA cycle) and the biosynthesis of several essential amino acids, i.e. valine, leucine and isoleucine.

The addition of TF similarity helps to refine the results of Query 1 in M-PAS, which only uses reaction similarity. Consequently, the ranks of found length-four pathways in RM-PAS might be different than those in M-PAS, revealing that regulatory mechanisms are not uniformly conserved in metabolic pathways.

For the 2427 pathways common in the results of RM-PAS and M-PAS, we calculated the rank of each pathway among the group of pathways which share the same starting substrate, using both scoring methods. This rank was then normalized by dividing by the size of the group to obtain a normalized rank in the range of [0,1], i.e. the most conserved pathway in a group ranks 1. In the end, 52% of pathways have normalized ranks higher in RM-PAS than in M-PAS, while 28% have lower ranks. Note that only 16% of the changes in the ranking is caused solely by changes in the group size.

In-depth analysis shows the TFs are indeed different in the pathways whose ranks are lower in RM-PAS. For instance, the pathway in Figure 4b has the highest score in M-PAS, but its RM-PAS score is the 30^{th} highest. This is because the TFs in the first building block are quite different; not only in TFBS matrices, but also in their functional annotations, binding domains, and protein sequences. In fact, the binding domain of the *E. coli* TF fruR is only present in bacteria.

4.2. Using one level to infer missing information at another level

Inferring missing reactions Here we use conservation at the regulatory level to infer missing reactions at the metabolic level. Based on the data collected, we constructed five building blocks corresponding to cases 8-10 (see Figure 3), which are shown in Figure 5. In particular, Figure 5d is found in six length-four pathways. In each example, although the reaction is not found in the database for one of the species, we hypothesize that it is actually present. The evidence comes from both metabolic and regulatory levels: all involved compounds and enzymes with the required function are present in the species, and they are also regulated similarly.

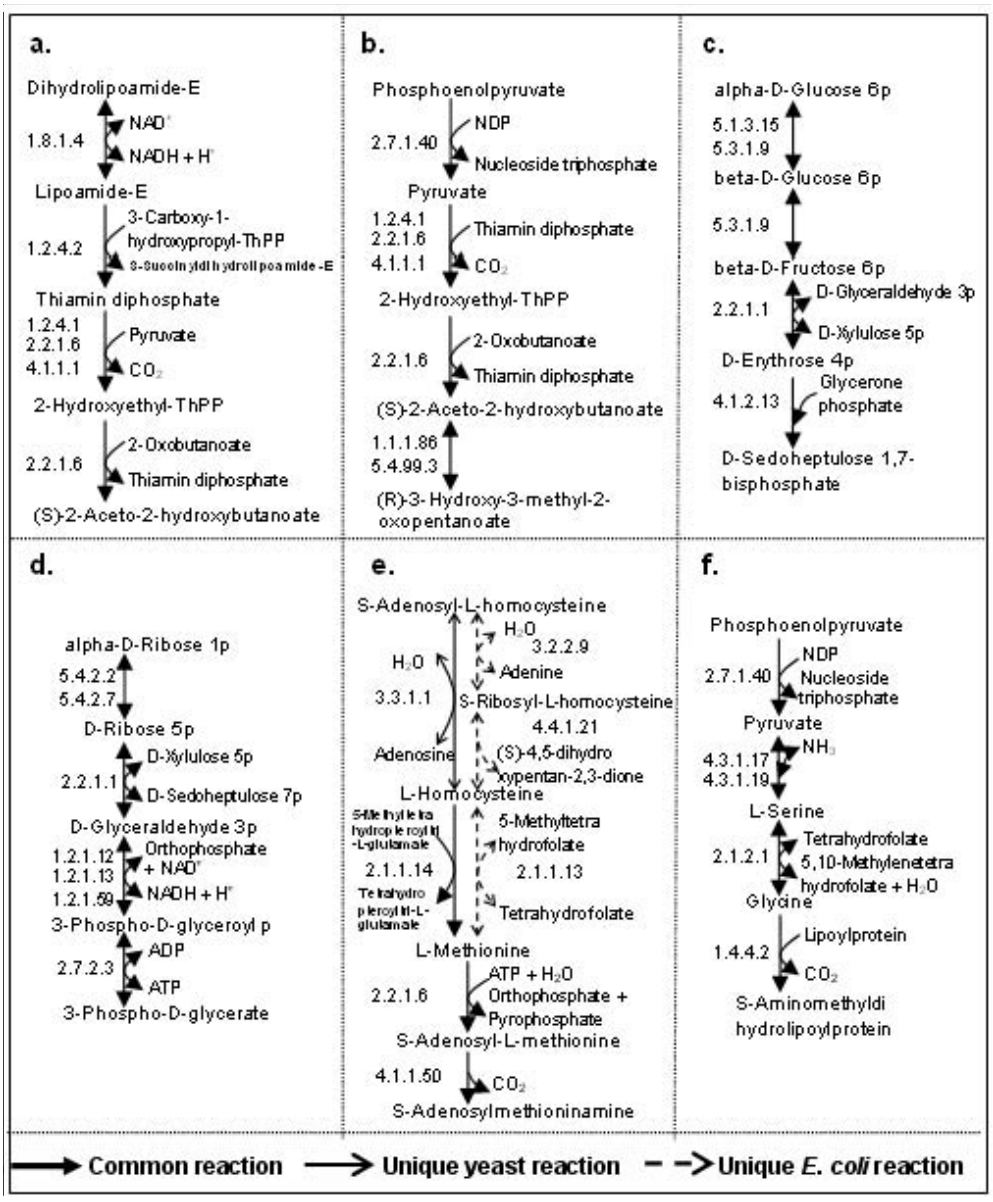

Fig. 4. Examples in the three queries. For conciseness, a common reaction in two species is drawn only once in each building block, indicated by a solid-headed arrow. **a)** An example best pathway in Query 1. **b)** One pathway which ranks differently in Query 1 of RM-PAS and M-PAS. **c-d)** Example best pathways in Query 2. **e-f)** Example best pathways in Query 3. See text for details.

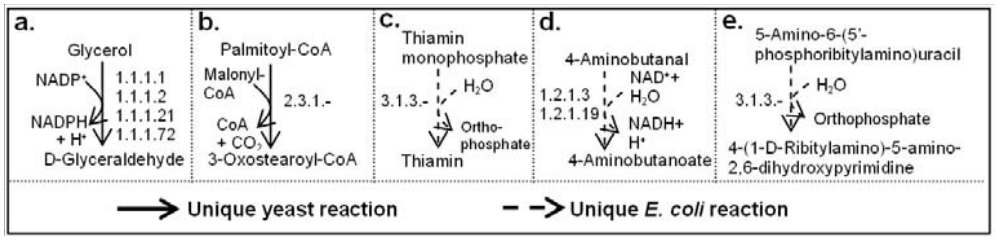

Fig. 5. Five building blocks belonging to cases 8-10 in Figure 3.

Table 1. The parameter settings in the three queries. "i" refers to the identical metabolic building block type in Figure 2. The cases refer to those in Figure 3, illustrating the scenarios for each query.

Query	ω_c	ω_e	ω_f	ω_q	ω_{tb}	ω_{tu}	Z_0	Target	Case
1	1	1	1	1	1	0	0 for all	Full conservation	1,2,8,9
2	1	1	1	1	0	1	0 for all	Missing TF-gene bindings	1,3,5,8,10
3a	1	1	1	1	-1	1	100 for non-"i"	Differences between two levels	3
3b	1	1	1	1	-1	1	100 for "i"	Differences between two levels	3

Inferring missing TF-gene bindings In Query 2, we try to use conservation at the metabolic level to prioritize a list of hypothetical TF-gene bindings with higher confidence. Overall, the predictions on yeast TF-gene bindings by RM-PAS are significantly better than random predictions. This is validated by a permutation test (see Appendix B), which shows that the TFs predicted by RM-PAS are more likely to bind to the respective genes than random predictions for 50% of the genes.

Here, we give two examples. Figure 4c shows the highest-scoring pathway, involved in glycolysis/gluconeogenesis, pentose phosphate pathway, and carbon fixation. In the fourth building block, we find the bound yeast TF GCR1 is similar to an alternative *E. coli* TF cueR, with MatCompare score = 0.3 (the original paper defines two TFs are similar when this score is <= 1). It suggests cueR might bind to the *E. coli* enzyme fbaA.

We applied Regulatory Sequence Analysis Tools (RSAT[28]) to see whether the upstream region of fbaA contains the TFBS of cueR. RSAT scans the upstream coding sequence of fbaA for the TFBS matrices of cueR. It outputs a segment score for each sequence segment, which is calculated as the log-ratio between the probability to generate the sequence segment given the TFBS matrix, and the probability to generate the sequence segment given the first-order Markov chain-based background model. The result shows not only that there exists one matching site at -141bp to -120bp, but also that it has a higher segment score than all TFBS of the bound TFs (i.e. fruR and crp) with site-wise p-value = 0.0005.

Another example is shown in Figure 4d. In the first building block, we find the bound *E. coli* TF Fis is similar with an alternative yeast TF WAR1, with MatCompare score = 0.5. It suggests WAR1 might bind to the yeast enzyme PGM2. RSAT shows that the TFBS matrix of WAR1 has a higher segment score than 20 (83%) bound TFs, with site-wise p-value = 0.00002. In addition, WAR1 shares the same domain "Zn clus" with six bound TFs, according to Pfam[29].

We applied co-expression analysis to investigate the likelihood of this latter TF-gene binding. Our reasoning is that if a particular gene g is regulated by a particular TF T, then g should be more similar than random genes r to other genes g' also regulated by T, in terms of correlation of mRNA expression. This means the average co-expression coefficient between g and g' should be significantly larger than that between r and g'. We used an mRNA microarray dataset described earlier[30]. The result shows that the average co-expression coefficient between PGM2 and the set of genes known to be regulated by WAR1 is significantly higher than the co-expression between a randomly drawn gene and the same gene set ($p = 0.001$).

4.3. Revealing the differences between two levels

The target pathways in Query 3 are conserved at the metabolic level, yet differ at the regulatory level. As depicted in case 3 in Figure 3, the bound TFs are a "mismatch", even though there exist "alternative" TFs. We further refine our investigation by looking at two types of conservation at metabolic level.

Query 3a looks into the diverse regulation in non-"identical" metabolic building blocks, which contain unique reactions with different cofactors in two species. Therefore, the query actually is designed to find cofactor-specific TFs. Since the enzymes catalyze different reactions in two species, we hypothesize that the different cofactors might have induced different TFs to bind the enzyme-coding genes. These enzyme products in turn enable the same transformation of a particular substrate to

a particular product, when different cofactors are available.

Another possible explanation is that different species have evolved separately to produce different cofactors, e.g. ATP, which are actually the main products in some pathways. Several studies show that mutations in active-site residues produce new catalytic properties for enzymes, which enable the formation of new pathways[31]. In our results, we find examples of different TF binding domains that have evolved in different species. For instance, the first building block in Figure 4e contains unique reactions in both species, and the yeast TFs have a bHLH domain present in eukaryotes, and a Zn(2)-C6 fungal-type domain only present in fungal TFs. The second building block contains a unique reaction in *E. coli*, and its enzyme metR has a HTH lysR-type DNA-binding domain unique to bacteria.

Query 3b finds divergent TFBS in the most conserved pathways at metabolic level, with identical reactions in both species. This might indicate the evolution of TFBS[32], and the mutational robustness during the evolution.

Although binding sites are subject to random mutations, evolution has naturally driven TFBS to be unspecific so that the functional phenotype is somewhat insensitive to mutations[33]. Previous research also shows that orthologous transcription factors may regulate orthologous genes through divergent TFBS in distantly related species[3]. This is reflected in our results. For example, the TFBS in the first building block in Figure 4f are very dissimilar in two species with MatCompare score = 2.1, although the enzymes share similar sequences with BLAST E-value $= 4 \times 10^{-68}$.

5. Conclusions

RM-PAS combines biological knowledge across species, and across levels of cellular organization. By setting different weight parameters in the scoring function, we showed how RM-PAS can be applied to identify conserved regulatory-metabolic network elements, infer missing reactions, prioritize and corroborate TF-gene binding hypotheses, and reveal diverse regulation in pathways that are conserved at metabolic level.

Our findings may be further exploited to analyze the integrated and aligned network properties, study evolutionary processes in multiple species, seek metabolic engineering targets, predict operons, and provide more possibilities to construct such a multi-level network for a new genome.

Acknowledgments

The authors would like to thank Marc Hulsman, Marco de Groot, and Christof Franke for their help and constructive discussions. This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

Appendix A. M-PAS

A.1. Metabolic building blocks

To allow for some variation, we introduce six types of metabolic building blocks - (1) Identical (i): the two reactions are identical, i.e. the reaction is present in both species. (2) Direct (d): the two reactions are different, but the first two digits of the EC numbers of their enzymes are the same. (3) Enzyme mismatch (em): the first two digits of the EC numbers of their enzymes are not the same. (4) Direct-gap (dg): a direct building block with a gap. A gap occurs when one species uses one reaction for a certain substrate-product transformation, while the other species uses two reactions connected in tandem to complete the same transformation. (5) Enzyme mismatch-gap (eg): an enzyme mismatch building block with a gap. (6) Enzyme crossover match (ec): both species uses two reactions to transform a common substrate into a common end product, and the first two EC number digits of the first and second reaction in one species are the same as those of the second and first reaction in the other species, respectively.

Note that for the five types of non-identical building blocks, we enforce the constraint that they must contain at least one unique reaction in one of the species, in order to avoid redundant building blocks. For instance, if two reactions A and B which convert the same substrate into the same product

are present in both species, two "*identical*" building blocks A_1-A_2 and B_1-B_2 are constructed already. Therefore, any other combinations of these reactions (i.e. A_1-B_2 and B_1-A_2) are just worse matches.

To summarize, these six types of building blocks emphasize the conservation between species, while taking alternative pathways, evolutionary diversity, annotation errors and possible variations in the order of the catalysis into consideration.

A.2. Compound score

$Z(C_B)$ is composed of the substrate similarity $Z(c_{sub})$ and the product similarity $Z(c_{pro})$ in building block B, where c_{sub} and c_{pro} denote the substrate supernode pair and the product supernode pair, respectively (see Figure 2). Each of these can be expressed by the *agreement* and *specificity* of the overlap between the paired compound supernodes in two species:

$$\begin{aligned} Z(C_B) &= \frac{1}{\sqrt{2}}[Z(c_{sub}) + Z(c_{pro})] \\ &= \frac{1}{\sqrt{2}}\{\frac{1}{\sqrt{2}}[Z_A(c_{sub}) + Z_S(c_{sub})] \\ &\quad + \frac{1}{\sqrt{2}}[Z_A(c_{pro}) + Z_S(c_{pro})]\} \end{aligned} \tag{A.1}$$

Let $c = \{c_{sub}, c_{pro}\}$. The agreement $Z_A(c)$ is the extent of the overlap in number of compounds between the compound supernodes. This is computed as the probability of observing the amount of overlap by chance, according to a hypergeometric distribution:

$$P_A(c) = \frac{\binom{|c_1|}{|c_1 \cap c_2|}\binom{|c_1 \cup c_2| - |c_1|}{|c_2| - |c_1 \cap c_2|}}{\binom{|c_1 \cup c_2|}{|c_2|}} = \frac{\binom{|c_1|}{|c_1 \cap c_2|}}{\binom{|c_1 \cup c_2|}{|c_2|}} \tag{A.2}$$

where c_1 and c_2 denote the compound supernodes in the two species which are paired to form c. $|x|$ denotes the number of compounds in x. To transform this probability to a z-score, the mean μ_{AC} and standard-deviation σ_{AC} of $P_A(c)$ over all possible compound supernode pairs in all reactions are needed:

$$Z_A(c) = \frac{P_A(c) - \mu_{AC}}{\sigma_{AC}} \tag{A.3}$$

The specificity $Z_S(c)$ of the overlapping compounds in c is considered in the scoring function, since some compounds appear more often than the others in the background. Therefore, we consider two compound supernodes to be more similar if the overlap is more specific, i.e. not observed frequently by chance. $Z_S(c)$ is calculated as follows:

$$P_S(c) = 1 - \frac{\#\text{observed } (c_1 \cap c_2) \text{ in the intersection}}{\#\text{all possible compound supernode pairs}} \tag{A.4}$$

$$Z_S(c) = \frac{P_S(c) - \mu_{SC}}{\sigma_{SC}} \tag{A.5}$$

The numerator in (A.4) is the number of times the specific overlap in compound node in c, i.e. $(c_1 \cap c_2)$, is observed in the intersections of all possible compound supernode pairs. μ_{SC} and σ_{SC} are the mean and standard-deviation of $P_S(c)$ computed over all possible compound supernode pairs.

A.3. Enzyme score

$Z(E_B)$ is a weighted sum of a functional similarity score $Z_F(e)$ and a sequence similarity score $Z_Q(e)$ for the enzyme supernode pair e, which is formed by the enzyme supernodes e_1 and e_2:

$$Z(E_B) = \frac{1}{\sqrt{\omega_f^2 + \omega_q^2}}[\omega_f Z_F(e) + \omega_q Z_Q(e)] \tag{A.6}$$

Like other weights, $\omega_f, \omega_q \in [-1, 1]$ indicate the the relative importance between functional and sequence similarity scores. $Z_F(e)$ is computed similar to Eqs. (A.1)-(A.5), containing the agreement and specificity of the EC number overlap, i.e. the common subclasses between the EC numbers of e_1 and e_2. For instance, for e_1 = 1.2.3.4 and e_2 = 1.2.4.4, the set of all subclasses $\mathcal{T} = \{1, 1.2, 1.2.3, 1.2.4, 1.2.3.4, 1.2.4.4\}$, and the common subclasses $\mathcal{M} = \{1, 1.2\}$. Then the enzyme functional similarity is calculated as follows:

$$Z_F(e) = \frac{1}{\sqrt{2}}[Z_A(e) + Z_S(e)] \tag{A.7}$$

$$P_A(e) = \frac{\binom{4}{|\mathcal{M}|}\binom{|\mathcal{T}| - 4}{4 - |\mathcal{M}|}}{\binom{|\mathcal{T}|}{4}} = \frac{\binom{4}{|\mathcal{M}|}}{\binom{|\mathcal{T}|}{4}} \tag{A.8}$$

$$Z_A(e) = \frac{P_A(e) - \mu_{AE}}{\sigma_{AE}} \tag{A.9}$$

$$P_S(e) = 1 - \frac{\#\text{observed } \mathcal{M} \text{ in the overlapping subclasses}}{\#\text{all possible enzyme supernode pairs}} \quad \text{(A.10)}$$

$$Z_S(e) = \frac{P_S(e) - \mu_{SE}}{\sigma_{SE}} \quad \text{(A.11)}$$

where $\{\mu_{AE},\sigma_{AE}\}$ and $\{\mu_{SE},\sigma_{SE}\}$ are computed from $P_A(e)$ and $P_S(e)$ over all possible enzyme supernode pairs, respectively. Finally, the sequence similarity score $Z_Q(e)$ is derived from the BLAST E-value $L(e)$:

$$Q(e) = -log_{10}L(e), \qquad Z_Q(e) = \frac{Q(e) - \mu_q}{\sigma_q} \quad \text{(A.12)}$$

where μ_q and σ_q are the mean and standard-deviation of $Q(e)$ over all possible enzyme supernode pairs.

Now we know how to calculate the enzyme similarity score for a pair of enzyme supernodes. If there are two such pairs e^a and e^b in a building block, as in "*enzyme crossover match*", we integrate the scores of the two supernode pairs as $Z(E_B) = [Z(e^a) + Z(e^b)]/\sqrt{2}$.

For enzyme supernodes with multiple EC numbers and/or multiple sequences, we first compute all $Z(e)$ given all possible combinations of EC numbers and corresponding sequences in enzyme hypernode e, and take the highest $Z(e)$ to be the enzyme similarity score, which indicates the similarity of the most conserved part in this enzyme supernode pair. For the same reason, when gaps are present, we choose the higher $Z(e)$ of the two enzyme supernode pairs.

B. Permutation test

To validate whether the TFs predicted by RM-PAS are more likely to bind to a particular gene than random predictions, we employed the following procedure:

(1) Generate the RM-PAS prediction dataset. This dataset contains the TF-gene pairs predicted by RM-PAS in Query 2. In particular, the genes are the enzyme-coding genes in the best pathways of Query 2, with $Z_{TU} > 0$ (Eq. 8).
(2) Generate the permuted dataset. For each TF-gene pair in the prediction dataset, fix the gene and pair it with 10 random TFs that have matrices and that are not known/predicted to bind to this gene.
(3) Run RSAT on both the prediction dataset and the permuted dataset, to obtain a segment score for each TF-gene pair.
(4) For each gene, test whether the segment scores of predicted TFs are significantly higher than those of random TFs in permuted dataset. This is a one-tailed t-test, assuming that two sets of scores come from normal distributions with unknown and possibly unequal variances. If $p < 0.05$, RM-PAS "wins" this gene test.
(5) Perform (4) for all genes in the prediction dataset, and obtain the percentage of genes for which RM-PAS wins.

Results: Given 40 genes in total in the prediction dataset, RM-PAS is significantly better than random in predicting TFs for 20 genes. Out of the other 20 genes where $p > 0.05$, 19 genes only have 2 or 3 predicted TFs, indicating that small sample size is a major cause of lack of significance.

References

1. Gasch AP, Moses AM, Chiang DY, Fraser HB, Berardini M, Eisen MB. Conservation and evolution of cis-regulatory systems in ascomycete fungi. *PLoS Biol.* 2004; **2(12)**: e398.
2. Tanay A, Regev A, Shamir R. Conservation and evolvability in regulatory networks: The evolution of ribosomal regulation in yeast. *PNAS* 2005; **102(20)**: 7203-7208.
3. Alkema WB, Lenhard B, Wasserman WW. Regulog analysis: detection of conserved regulatory networks across bacteria: application on *Staphylococcus aureus*. *Genome Res.* 2004; **14**: 1362-1373.
4. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. *Nature* 2000; **406**: 651-654.
5. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. Hierarchical organization of modularity in metabolic networks. *Science* 2002; **297**: 1551-1555.
6. Heymans M, Singh AK. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics* 2003; **19**: i138-i146.
7. Sharan R, Ideker T. Modeling cellular machinery through biological network comparison. *Nature Biotechnology* 2006; **24(4)**: 427-433.
8. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BØ. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 2004; **429**: 92-96.

9. Ihmels J, Levy R, Barkai N. Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*. *Nature Biotech.* 2004; **22(1)**: 86-92.
10. Patil KR, Nielsen J. Uncovering transcriptional regulaton of metabolism by using metabolic network topology. *PNAS* 2005; **102(8)**: 2685-2689.
11. Li Y, de Ridder D, de Groot MJL, Reinders MJT. Metabolic pathway alignment between species using a comprehensive and flexible similarity measure. *BMC Systems Biology* 2008; **2**: 111.
12. Li Y, de Ridder D, de Groot MJL, Reinders MJT. Metabolic pathway alignment (M-Pal) reveals diversity and alternatives in conserved networks. In: Brazma A, Miyano S, Akutsu T (eds.), Advances in Bioinformatics & Computational Biology, Volume 6. Imperial College Press, London. 2008: 273-285.
13. Ma HW, Zeng AP. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* 2003; **19**: 270-277.
14. Hedges LV, Olkin I. *Statistical methods for meta-analysis*. Academic Press, Orlando. 1985.
15. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics* 2004; **5**: 276-287.
16. Kielbasa SM, Gonze D, Herzel H. Measuring similarities between transcription factor binding sites. *BMC Bioinformatics* 2005; **6**: 237.
17. Schones DE, Sumazin P, Zhang MQ. Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics* 2005; **21(3)**: 307-313.
18. Smith AD, Sumazin P, Zhang MQ. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *PNAS* 2005; **102(5)**: 1560-1565.
19. Goto S, Nishioka T, Kanehisa M. LIGAND: chemical database for enzyme reactions. *Bioinformatics* 1998; **14(7)**: 591-599.
20. The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 2008; **36**: D190-D195.
21. Wingender E, Chen X, Fricke E *et al.* The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.* 2001; **29(1)**: 281-283.
22. Teixeira MC, Monteiro P, Jain P *et al.* The YEASTRACT database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*. *Nucl. Acids Res.* 2006; **34**: D446-D451.
23. Pachkov M, Erb I, Molina N, van Nimwegen E. SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucl. Acids Res.* 2007; **35**: D127-D131.
24. Chen QK, Hertz GZ, Stormo GD. MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput. Appl. Biosci.* 1995; **11(5)**: 563-566.
25. Ghosh D. OOTFD (Object-Oriented Transcription Factors Database): an object-oriented successor to TFD. *Nucl. Acids Res.* 1998; **26(1)**: 360-361.
26. Keseler IM, Collado-Vides J, Gama-Castro S *et al.* EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucl. Acids Res.* 2005; **33**: D334-D337.
27. Gama-Castro S, Jacinto VJ, Peralta-Gil M *et al.* RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucl. Acids Res.* 2008; **36**: D120-D124.
28. http://rsat.ulb.ac.be/rsat/
29. Finn RD, Tate J, Mistry J *et al.* The Pfam protein families database. *Nucl. Acids Res.* 2008; **36**: D281-D288.
30. Li Y, de Ridder D, Duin RPW, Reinders MJT. Integration of prior knowledge of measurement noise in kernel density classification. *Pattern Recognition* 2008; **41**: 320-330.
31. Murzin AG. Can homologous proteins evolve different enzymatic activities? *Trends. Biochem. Sci.* 1993; **18**: 403-405.
32. Madan Babu M, Teichmann SA. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res.* 2003; **31**: 1234-1244.
33. van Nimwegen E, Crutchfield JP, Huynen M. Neutral evolution of mutational robustness. *PNAS* 1999; **96**: 9716-9720.