Photosynth Res
DOI 10.1007/s11120-013-9910-6

EMERGING TECHNIQUE

SHARP: genome-scale identification of gene–protein–reaction associations in cyanobacteria

S. Krishnakumar · Dilip A. Durai · Pramod P. Wangikar · Ganesh A. Viswanathan

Received: 25 March 2013 / Accepted: 7 August 2013

Abstract Genome scale metabolic model provides an overview of an organism's metabolic capability. These genome-specific metabolic reconstructions are based on identification of gene to protein to reaction (GPR) associations and, in turn, on homology with annotated genes from other organisms. Cyanobacteria are photosynthetic prokaryotes which have diverged appreciably from their nonphotosynthetic counterparts. They also show significant evolutionary divergence from plants, which are well studied for their photosynthetic apparatus. We argue that context-specific sequence and domain similarity can add to the repertoire of the GPR associations and significantly expand our view of the metabolic capability of cyanobacteria. We took an approach that combines the results of context-specific sequence-to-sequence similarity search with those of sequence-to-profile searches. We employ PSI-BLAST for the former, and CDD, Pfam, and COG for the latter. An optimization algorithm was devised to arrive at a weighting scheme to combine the different evidences with KEGG-annotated GPRs as training data. We present the algorithm in the form of software "Systematic, Homology-based Automated Re-annotation for Prokaryotes (SHARP)." We predicted 3,781 new GPR associations for the 10 prokaryotes considered of which eight are cyanobacteria species. These new GPR associations fall in several metabolic pathways and were used to annotate 7,718 gaps in the metabolic network. These new annotations led to discovery of several pathways that may be active and thereby providing new directions for metabolic engineering of these species for production of useful products. Metabolic model developed on such a reconstructed network is likely to give better phenotypic predictions.

Keywords Cyanobacteria · SHARP · Gene–protein–reaction (GPR) association · Genome scale · Metabolic network reconstruction · PSI-BLAST

Abbreviations

BLAST	Basic Local Alignment Search Tool
BRENDA	BRaunschweig ENzyme DAtabase
CDD	Conserved Domain Database
CINPER	CSBL INteractive Pathway BuildER
COG	Cluster of Orthologous Groups
EFICAz	Enzyme Function Inference by Combined Approach
GPR	Gene to Protein to Reaction
LMI	Library of Metabolic Information
KAAS	Kegg Automatic Annotation Server
KEGG	Kyoto Encyclopedia of Genes and Genomes
metaSHARK	metabolic SearcH And Reconstruction Kit
MNR	Metabolic Network Reconstruction
Pfam	Protein families
PROFAT	PROtein Functional Annotation Tool
PSI-BLAST	Position-Specific Iterated-BLAST
RAST	Rapid Annotation using Subsystem Technology

S. Krishnakumar and Dilip A. Durai have contributed equally.

Electronic supplementary material The online version of this article (doi:10.1007/s11120-013-9910-6) contains supplementary material, which is available to authorized users.

S. Krishnakumar · D. A. Durai · P. P. Wangikar (✉) · G. A. Viswanathan (✉)
Department of Chemical Engineering, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India
e-mail: wangikar@iitb.ac.in

G. A. Viswanathan
e-mail: ganeshav@iitb.ac.in

Published online: 24 August 2013

RPS-BLAST	Reverse Position-Specific BLAST
SHARP	Systematic, Homology-Based Automated Re-annotation for Prokaryotes
UniProt	Universal Protein Resource

Introduction

Metabolic network orchestrates the interconversion of metabolites catalyzed by specific enzymes. These enzymes control the network's response (Syed and Yona 2009 and references there in, Bauer-Mehren et al. 2009, Reed et al. 2006). Metabolic network of different cyanobacterial species has been implicated in engineering (Zhou and Li 2010; Ducat et al. 2011; Wang et al. 2012) the organism for production of various biofuel precursors (Peralta-Yahya et al. 2012; Quintana et al. 2011) such as ethanol (Deng and Coleman 1999; Dexter and Fu 2009), iso-butanol (Varman et al. 2013), butanol (Lan and Liao 2011; Lan and Liao 2012). Moreover, this paradigm (Wang et al. 2012) is being attempted for production of chemicals such as ethylene (Takahama et al. 2003), iso-butyraldehyde (Atsumi et al. 2009), poly-3-hydroxybutyrate (Tyo et al. 2006), isoprene (Lindberg et al. 2010) etc. Efficient engineering of a species and analyses of metabolic network hinges on the availability of accurate gene–protein–reaction (GPR) association of all the enzymes in the network (Klimke et al. 2011; Kyrpides 2009; Ouzounis and Karp 2002).

Functional annotation tools for identifying GPR associations are paramount to metabolic network reconstruction (MNR) (Ogata et al. 1999; Francke et al. 2005; Feist and Palsson 2008; Feist et al. 2009). Based on curated and experimental data, a number of tools have been developed to identify unannotated enzymes (Kumar et al. 2012; Copeland et al. 2012 and references therein, Faust et al. 2011; Kumar et al. 2007). GPRs of metabolic network can be identified using (a) manual curation method, which requires extensive primary experimental information (Furnham et al. 2009) (e.g. UniProtDB (UniProt Consortium 2010) and BRENDA (Scheer et al. 2011)); or (b) automated methods, which is a knowledge-based approach (Watson et al. 2005) and is primarily based on sequence similarity (Adriaens et al. 2008; Viswanathan et al. 2008). Due to limited availability of experimental information, development of efficient automated methods to decipher correct annotations is required (Radivojac et al. 2013). In addition to the local sequence similarity (Altshul et al. 1990), most of these methods use several evidences to improve confidence levels of annotation predictions. Some recently developed methods include KAAS (Moriya et al. 2007), RAST (Aziz et al. 2008), CINPER (Mao et al. 2012). Irrespective of integration of any number of sequence similarity-based evidences, certain level of uncertainty is expected. The Gibbs sampling-based global probabilistic annotation (Plata et al. 2012) which incorporates these uncertainties requires experimentally validated information and is not scalable.

GPR associations based on automated methods, and other related information for various species are stored in many public databases which include KEGG (Kanehisa and Goto 2000), MetaCyc (Caspi et al. 2006), SEED (Overbeek et al. 2004). Despite the availability of so many methods, GPR association for many enzymes is missing for several organisms. It is possible that these enzymes may have originated from phylogenetically distant species; particularly in those organisms (such as cyanobacteria) that diverged appreciably from their counterparts during evolution. Therefore their functional annotation prediction will require consideration of remote context-specific homologs (Ouzounis and Karp 2002). PSI-BLAST is a tool that can be used for this purpose (Altschul et al. 1997). Based on PSI-BLAST, we develop and demonstrate a novel methodology "Systematic, Homology-based Automated Re-annotation for Prokaryotes (SHARP)" to predict GPR associations. In addition to distant context-specific sequence similarity, this method uses as evidences functional region similarities, and bidirectional hits. An optimization, probabilistic-based approach using preexisting annotations (from KEGG) as training set is used for annotation prediction and for estimation of the species-specific and enzyme-specific significances of different evidences. We report GPR association for several enzymes unannotated by KEGG in 10 different bacterial species, predominantly from cyanobacteria family.

Results

Systematic, Homology-based Automated Re-annotation for Prokaryotes (SHARP)

Flow chart depicting the steps involved in SHARP is presented in Fig. 1. Consider a species $o_i \in O$, metabolic enzymes of which need to be annotated. As a first step, we create an offline Library of Metabolic Information (LMI) by distilling various metabolism-related information about the organism o_i available in a variety of public databases (see Materials and methods). Next, for a query enzyme $e_j \in E$ in the query organism o_i, using PSI-BLAST (Altschul et al. 1997), we find the candidate sequence set $C_{i,j} = \{C_{i,j,k}\}$ (see Materials and methods). Hereafter, subscripts i, j and k will, respectively, refer to query species o_i, query enzyme e_j, and the kth candidate in $C_{i,j}$.

Candidates obtained in the first iteration of PSI-BLAST consist of close hits. However, in subsequent iterations, the

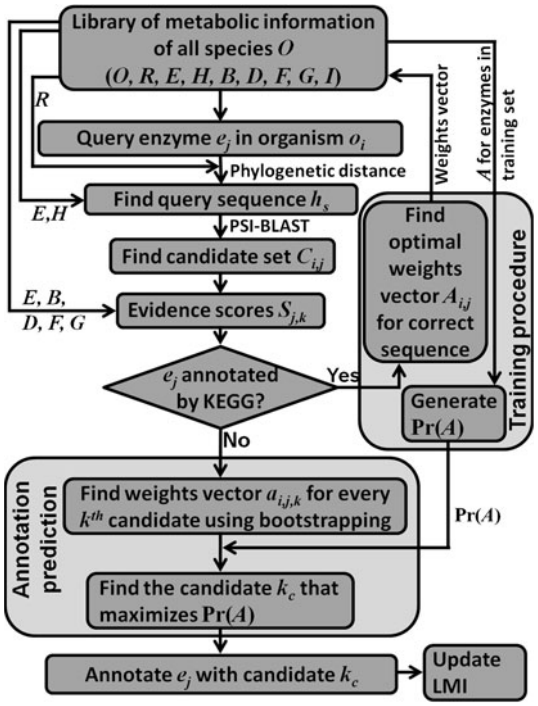

Fig. 1 Flow chart depicting the methodology of SHARP. *O*, *R*, *E*, *H*, *B*, *D*, *F*, *G*, *I* and *A*, respectively, refer to list of bacterial species, list of 16s rRNA sequence of all species in *O*, metabolic enzyme superset, amino acid sequences of all enzymes in *E*, name of the pathways from KEGG, domain names, family names, orthologous groups, gene IDs and weights. o_i, e_j, h_s, and k_c, respectively, are query organism, query enzyme, query sequence and SHARP predicted candidate for annotation of enzyme e_j. $C_{i,j}$ and $S_{j,k}$ are the vectors of candidates obtained from PSI-BLAST and evidence score matrix, respectively. $A_{i,j}$ and $a_{i,j}$, respectively, are the optimal weight vectors for enzymes in training set and enzymes for SHARP annotation. While the training procedure block consists of finding optimal weight vectors corresponding to the correct gene ID and generation of the joint probability distribution Pr(A), the annotation prediction uses a combination of bootstrapping method for optimal weight vector estimation and the training information for inferring the GPR association for a query enzyme. Subscripts i, j, and k, respectively, refer to the query organism o_i, query enzyme e_j, and kth candidate sequence in $C_{i,j}$

sequences picked are enriched with those remote homologs functions of which may deviate significantly from those obtained in the first iteration. Therefore, additional function-based evidences can be used to identify the best candidate in $C_{i,j}$ for GPR association of enzyme e_j. Therefore, we use as evidences the following:

(A) Reverse BLAST-based pathway similarity: Evolutionarily conserved enzymes across different species are expected to participate in similar metabolic pathways. In order to capture the extent of participation of the candidate sequence in relevant pathways, we define a score $S_{j,k,b}$ (see Materials and methods) that quantifies the average pathway similarity that the candidate sequence has with its homologs.

(B) Domain conservation: Proteins have a modular structure consisting of domains (Marchler-Bauer et al. 2009, 2010, 2011) and is based on the function performed by these domains. Therefore, comparison of the families into which these domains fall can serve as good evidences (Overbeek et al. 2007). We consider three different domain qualifiers to identify their extent of conservation:

(i) Domain conservation (CDD): For every candidate $C_{i,j,k}$, a score $S_{j,k,d}$ (see Materials and methods) measures the fraction of the conserved domains in query enzyme e_j present in the candidate.

(ii) Clusters of orthologous groups of proteins (COG) conservation: As each group corresponds to an ancient conserved domain (Tatusov et al. 1997, 2003), candidates are expected to share these with e_j. For a candidate $C_{i,j,k}$, we capture in score $S_{j,k,c}$ (see Materials and methods) the fraction of the expected COGs of the query enzyme e_j that is conserved in the candidates.

(iii) Pfam family conservation: Based on the different combination of functional domains in a protein, they are classified into various families (Punta et al. 2012). Candidates are therefore expected to belong to same families as e_j. For a candidate $C_{i,j,k}$, we capture this conservation in score $S_{j,k,f}$ (see Materials and methods).

Training procedure for weight parameter estimation

Evidences are species and enzyme specific. Confidence of an automated annotation method can be improved by weighting an evidence score according to its relative relevance. For this purpose, we introduce a query species and query enzyme-specific weight parameter vector $A_{i,j} = \{a_l\} \forall l = b, d, c, f$ where a's are the individual weights. This vector must be considered for identification of the candidate for annotation for a query enzyme e_j. Using these weights, we combine various evidence scores to define an overall score.

$$S^0_{j,k} = A_{i,j} \cdot S^t_{j,k,:} \quad ([1])$$

where $S_{j,k,:}$ is the raw score vector of kth candidate. Estimation of overall score requires priori estimates of $A_{i,j}$. As the weights are unavailable a priori, we introduce a systematic training procedure (Fig. 1) using known GPR association (reported by KEGG) for various enzymes across several species to find the optimal weights that will identify the correct candidate (see Materials and methods). The optimal weights are estimated for as many KEGG-

annotated enzymes for all the species considered. Using these optimal weights, we next generate a joint probability distribution Pr(A) which is the probability of finding a correct sequence whose weights vector A maximizes the overall score (across all candidates for the query enzyme e_j). This distribution will be used to identify the correct candidate sequence, as detailed below.

Annotation of unannotated enzymes

We employ a bootstrapping procedure to identify the best candidate $C_{i,j,k}$ for GPR association of query enzyme e_j as the one that maximizes the probability Pr(A) (see Materials and methods). The hallmark of this bootstrap method is that it enables identification of only those candidate(s) which are likely to have weights that can lead to a maximum overall score. As a result, this sequence is likely to be the correct sequence for the query enzyme e_j.

Completely automated software written using Perl programming language (version 5.8.8) of the proposed methodology SHARP is available, upon request. (Details of the software, installation, a list of libraries used, and execution protocol are in Suppl Mat. 2.)

Annotation of metabolic enzymes

We considered genome-wide annotation of 10 bacterial species (listed in Table 1) using SHARP. As different cyanobacterial species are being considered as model organisms for biofuel precursor production, we considered eight species from three clades of cyanobacteria species for re-annotation (Memon et al. 2013). For example, from Clade B, we consider *Cyanothece* sp. strain ATCC 51142 which exhibits temporal separation of oxygenic photosynthesis and oxygen-sensitive nitrogen fixation processes and therefore has recently been implicated for hydrogen production (Krishnakumar et al. 2013; Feng et al. 2010). Moreover, as cyanobacteria are photosynthetic prokaryotes which have diverged appreciably from their nonphotosynthetic counterparts, these species are amenable for demonstration of identification of GPR association using context-specific, remote homology. In addition, we also chose *Escherichia coli* K12 MG1655 and *Corynebacterium glutamicum* ATCC 13032 which are well annotated and considered as standards.

SHARP was able to find the optimal weights for ~86 % (4062/4734) of all the KEGG-annotated enzymes across these 10 species (Table 1). This suggests that information gathered by SHARP for training purposes is reliable. Key reasons for the inability to predict the optimal weights and therefore annotations for the remaining 14 % are (a) remote homologs belonging to the query species were not found for the query sequence, or (b) failure to predict optimal weights. Using a total of 4,062 enzymes across 10 species as training set, we predict annotation for 3,781 enzymes which is used for filling 7,718 gaps (Table 2) in the metabolic network. (A list of the EC numbers along with the gene IDs for all new annotations is provided in Suppl. Mat. 3.) Note that only enzymes that participate in active pathways reported by KEGG are considered. For most of the annotations predicted by SHARP, a reverse PSI-BLAST also resulted in the same GPR association. This suggests that the GPR associations predicted are one-to-one.

Individual weights for every enzyme reflect the relative importance of the corresponding evidence in identifying the correct annotations. The distribution of the nonzero weights normalized with the total number of corresponding species-specific evidence score available is presented in Fig. 2 for *Synechococcus* sp. PCC 7002. Distribution for the remaining nine species is presented in Fig. S1 (Suppl. Mat. 4). Note that the weights predicted to be zero by optimization suggests that individual evidence score does

Table 1 Number of KEGG-annotated enzymes for the 10 species considered and correspondingly those that were used for training SHARP

KEGG ID	Species	Total	Training set from KEGG
ana	*Anabaena* sp. PCC 7120	472	449
ava	*Anabaena variabilis ATCC 29413*	503	450
cgl	*Corynebacterium glutamicum* ATCC 13032	448	375
cyc	*Cyanothece* sp. PCC 7424	485	400
cyt	*Cyanothece* sp. ATCC 51142	462	415
eco	*Escherichia coli* K12 MG 1655	636	520
gvi	*Gloeobacter violaceus* PCC 7421	442	361
syc	*Synechococcus elongatus* PCC 6301	429	343
syn	*Synechocystis* sp. PCC 6803	445	420
syp	*Synechococcus* sp. PCC 7002	412	329
Total		4,734	4,062

Table 2 Total number of enzymes and gaps newly annotated by SHARP in the 10 species considered

Species		ana	ava	cgl	cyc	cyt	eco	gvi	syc	syn	syp	Total
Total number of enzymes newly annotated		433	397	273	315	389	392	456	485	236	405	3,781
Number of gaps filled in different pathway categories												
KEGG-based pathway category	Code											
Carbohydrate metabolism	CM	223	191	177	153	212	178	225	236	103	203	1,901
Lipid metabolism	LM	105	100	88	74	43	121	135	130	57	139	981
Amino acid metabolism	AM	191	170	115	154	197	172	219	244	145	221	1,828
Xenobiotics biodegradation and metabolism	XM	97	135	49	80	111	149	99	108	58	65	951
Metabolism of cofactors and vitamins	VM	48	45	35	37	47	41	58	66	21	55	453
Biosynthesis of other secondary metabolites	BM	10	9	5	8	4	5	10	6	5	9	71
Energy metabolism	EM	63	64	30	64	51	40	73	72	21	72	550
Metabolism of terpenoids and polyketides	TM	14	8	5	5	11	10	13	14	11	9	100
Metabolism of other amino acids	AA	33	26	24	18	25	43	35	40	21	31	296
Nucleotide metabolism	NM	59	51	45	43	56	26	58	65	31	55	489
Glycan biosynthesis and metabolism	GM	3	2	1	4	2	3	10	9	4	7	45
Translation	TL	2	5	5	3	7	6	9	6	4	6	53
Total gaps filled		848	806	579	643	766	794	944	996	481	872	7,718

The gaps are presented for 13 KEGG identified pathway categories. Detailed list of new annotations by SHARP and corresponding scores, optimal weights is in Suppl. Mat 3 and the corresponding pathway-wise compilation is in Suppl. Mat 6

not contribute to the overall score for that enzyme. Weights distribution suggests that the order of significance of various evidences in terms of its contribution to overall score for GPR association prediction is (1) pathway similarity, (2) COG conservation, (3) PFAM conservation, and (4) CDD conservation. The distribution of the actual (available) scores supports this (Fig. S2 in Suppl. Mat. 4).

The predictive ability of the proposed method depends on the extent of training information captured by the joint probability distribution Pr(A). We find that the variation in Shannon entropy (Shannon 1948) associated with Pr(A) due to the number of species included sequentially in the training set is insignificant (Fig. S3, Suppl. Mat. 5). This suggests that information content in Pr(A) is, relatively, independent of the number of species. Therefore, we believe that the method is scalable.

New annotations that we report fall in several important metabolic pathway categories such as carbohydrate metabolism, and amino acid metabolism. SHARP is able to fill (Table 2) significant number of gaps in several pathways. A consolidated list of newly annotated enzymes participating in each of these individual pathways for every species is presented in Suppl. Mat 6.

Discussion

In this study, we present a novel PSI-BLAST-based methodology SHARP for GPR association of the metabolic enzymes involved in prokaryotes. We believe that PSI-BLAST-based method for functional annotation can predict annotations of several enzymes for which local sequence similarity method fails. In Fig. 3 (for *Synechococcus* sp. PCC 7002) and in Fig. S4 (for the other nine species, Suppl. Mat 7), we compare the histograms of the phylogenetic distance between the query organism and the species from which the functional annotation for an enzyme is picked for those in training set and for those newly annotated by SHARP. This comparison clearly shows a shift in the histogram toward higher phylogenetic distance indicating that the new annotations have indeed been picked from remote homologs. For example, consider the butanoate metabolism pathway in *Synechococcus* sp. PCC 7002 KEGG pathway entry of which is syn00650 (Fig. S5, Suppl. Mat. 8). In this pathway, SHARP predicted annotations for 16 new enzymes with a very good query coverage for most of them (Table 3). Table 3 also shows that BLAST failed to predict annotations for these enzymes. This may be because they may have evolutionarily deviated significantly as substantiated by the phylogenetic distance of the species from which the annotation is picked. So, these enzymes may have a homology with species that is phylogenetically not so close and therefore, their annotations cannot be predicted by local sequence similarity.

For more than 83 % of all the enzymes considered by SHARP in each of the four species cyt, cyc, syn, and syp, the GPR association identification matched with that reported in Pathway Genome Database (Karp et al. 2010). (For a detailed comparison, see Table S4 in Suppl. Mat 9.)

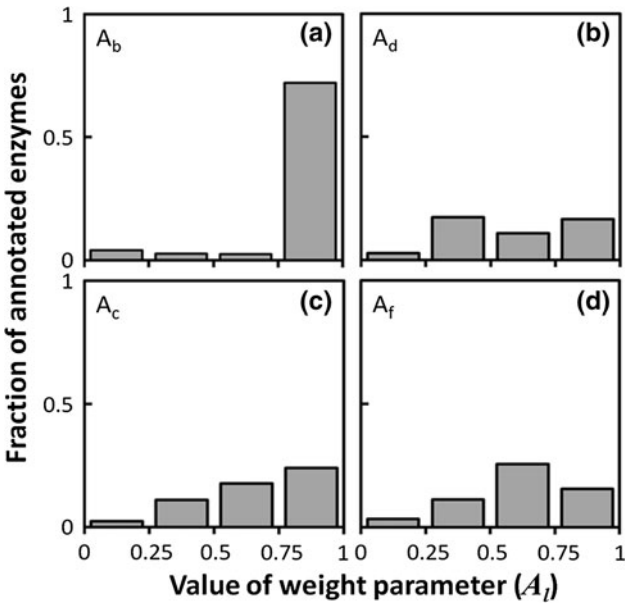

Fig. 2 Distribution of optimal weights for (**a**) Pathway similarity evidence, (**b**) Domain conservation evidence, (**c**) COG conservation evidence, and (**d**) Pfam conservation evidence for *Synechococcus* sp. PCC 7002. The total number of species-specific scores available for each of evidences was used for normalizing the corresponding distribution

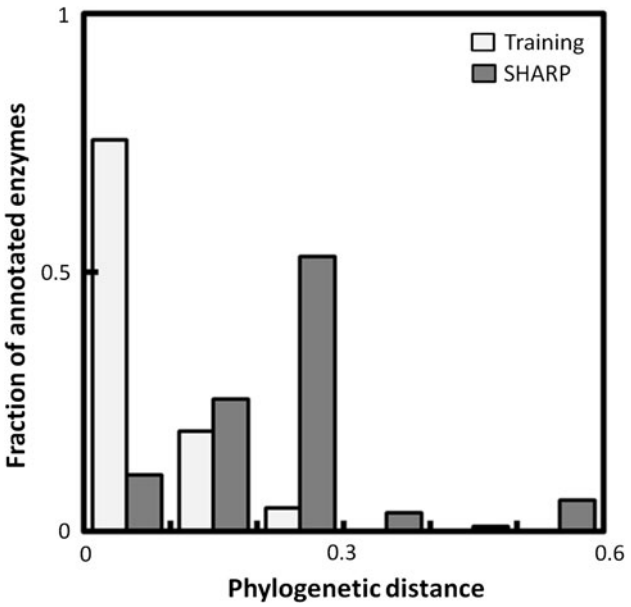

Fig. 3 Histograms of the phylogenetic distance between the query enzyme and the species from which the functional annotation is picked for enzymes in the training set (*white*) and for those GPR association of which is predicted by SHARP (*gray*) for *Synechococcus* sp. PCC 7002. Phylogenetic distance histograms for other species are given in Fig. S4 in Suppl. Mat 7

Similar comparison for species cyt (Saha et al. 2012) and syn (Knoop et al. 2010) with those used in Flux balance analyses shows about 94 % match in the GPR association (Table S5 in Suppl. Mat. 9). GPR association identification in the training set and predictions by SHARP agree very well with those by existing methods such as pathway tools (Karp et al. 2010), flux balance analyses. (Knoop et al. 2010).

Enzymes annotated by SHARP have led to identification of several industrially important pathways. For example, in the network involved in butanoate metabolism (Fig. S5 in Suppl. Mat. 8) in *Synechococcus* sp. PCC 7002, we predict annotations for EC 4.1.1.15, 2.6.1.19, and 1.2.1.24 which are involved in the metabolic pathway transforming L-Glutamate to succinate with 4-aminobutanoate and succinate semialdehyde as intermediates. Moreover, in methane metabolism pathway, we associate geneID SYNPCC7002_A0853, syc0566_c, and PCC7424_5201 to EC 1.14.13.25 in *Synechococcus* sp. PCC 7002, *Synechococcus elongatus* PCC 6301 and *Cyanothece* PCC 7424, respectively (Suppl. Mat. 3). This enzyme catalyzes methane to form methanol. Besides, in *Anabaena variabilis* ATCC 29413, *Anabeana* sp. PCC 7120, and *S. elongates* PCC 6301, SHARP was able to annotate EC 1.1.1.244, 1.2.1.46, and 1.2.1.2, respectively (Suppl. Mat. 3). These enzymes are directly involved in pathways that convert methanol to CO_2. These GPR association predictions can now be used to design pertinent metabolic engineering strategies to improve the productivity of appropriate biofuels (Quintana et al. 2011). In addition, SHARP predicts new annotation(s) for several enzymes in amino acid metabolism (Table 2). These enzymes can now be considered targets for novel synthetic biology strategies (Purnick and Weiss 2009) to increase the ability of the species to process amino acids. Phosphorylation of D-glycerate to 3-phosphoglycerate in "glycoxylate and dicarboxylate metabolism" pathway is an important biochemical step that helps in carbon fixation in a few cyanobacterial species (Bartsch et al. 2008). In *Anabeana* sp. PCC 7120, SHARP predicted GPR association for 24 enzymes in this pathway—particularly, those of EC 1.1.1.26, 1.1.1.29, 1.1.1.60, and 1.1.1.79 which are directly involved in the production of D-glycerate and those of EC 4.1.3.1 involved in the production of glyoxylate.

Annotation predictions based on computational biology methods must be validated experimentally, which provide the highest confidence (Thiele and Palsson 2010). The annotation that we report for EC 1.2.1.24, which catalyzes conversion of succinate semialdehyde to succinate has been experimentally validated (Zhang and Bryant 2011). In the isoleucine synthesis pathway of *Cyanothece* ATCC 51142, geneID cce_0248 has been predicted as the correct annotation (see Fig. S6 in Suppl. Mat. 8) for EC 2.3.1.182, which has recently been established using experimental approaches (Wu et al. 2010). Such a validation is not possible for most of the annotations due to unavailability of

Table 3 Metabolic enzymes in butanoate metabolism network (KEGG pathway 00650) for *Synechococcus* sp. PCC 7002 for which GPR association was predicted by SHARP

Enzyme	Species (from which query sequence was taken)	Phylogenetic distance	PSI-BLAST e-value (from 4th iteration)	BLAST e-value	Query coverage	Identity
1.1.1.304	*Staphylococcus epidermidis* RP62A	0.247543	2.00E−12	NF	45.13618677	22.22
1.1.1.35	*Brevibacillus brevis* NBRC 100599	0.235522	1.00E−37	NF	38.77805486	14.86
1.1.1.36	*Bacillus cereus* ATCC 14579	0.237744	3.00E−26	4.00E−23	99.58506224	23.86
1.1.1.4	*Bacillus cereus* ATCC 14579	0.237744	2.00E−25	5.00E−21	94.79768786	20.86
1.1.1.61	*Clostridium difficile* 630	0.244814	4.00E−28	NF	97.03504043	17.13
1.1.1.83	*Acidaminococcus fermentans* DSM 20731	0.21706	2.00E−77	4.00E−34	99.15730337	30.68
1.2.1.24	*Bacillus amyloliquefaciens* FZB42	0.242012	2.00E−59	3.00E−54	97.4025974	32.17
1.2.7.1	*Halothermothrix orenii* H 168	0.231005	9.00E−66	NF	66.66666667	20.42
2.3.3.10	*Nostoc punctiforme* PCC 73102	0.120439	6.00E−66	NF	79.90196078	16.32
2.6.1.19	*Anabaena variabilis* ATCC 29413	0.112158	1.00E−56	2.00E−19	97.99554566	20.61
2.6.1.19	*Anabaena variabilis* ATCC 29413	0.112158	1.00E−56	2.00E−29	99.33184855	20.61
4.1.1.15	*Prochlorococcus marinus* MIT 9313	0.122498	2.00E−22	NF	83.8362069	16.45
4.1.1.15	*Prochlorococcus marinus* MIT 9313	0.122498	3.00E−18	NF	83.40517241	14.11
4.1.3.4	*Brevibacillus brevis* NBRC 100599	0.235522	3.00E−48	NF	97.99331104	20.08
4.2.1.17	*Anabaena variabilis* ATCC 29413	0.112158	3.00E−45	NF	51.45067698	20.5
4.2.1.55	*Brevibacillus brevis* NBRC 100599	0.235522	2.00E−38	2.00E−22	98.08429119	33.33
6.2.1.16	*Rhodospirillum rubrum* ATCC 11170	0.247109	7.00E−23	NF	53.74251497	16.39
6.2.1.2	*Pseudomonas fluorescens* Pf 5	0.26	1.00E−11	3.00E−05	30.98591549	18.54

The species from which the annotation was picked, its phylogenetic distance with *Synechococcus* sp. PCC 7002, PSI-BLAST e-value for the match, BLAST e-value for the match, if available, query coverage and identity

NF not found

appropriate experimental data. However, the predictions made by SHARP can provide pinpointed directions for experimentalists to validate annotations using techniques such as specific gene-knockout (Puchalka et al. 2008), NMR ligand screening (Chen et al. 2011), and gene expression studies (Moskal et al. 2007; Becker and Palsson 2008).

Use of species and enzyme-specific optimal weights, estimated by SHARP as a measure of the relative importance (Thiele and Palsson 2010) of various evidences, provides a novel approach for quantifying the confidence levels of the predicted GPR association. Such quantification can be incorporated appropriately in genome-scale metabolic models to identify the confidence levels of the model. For instance, a cut-off on the weights for profile search-based evidences reflecting the reliability of the existence of a particular biochemical reaction may be considered as additional constraint while estimating the metabolic flux distributions using appropriate optimization strategies (Shastri and Morgan 2005; Montagud et al. 2010). Ability of SHARP to predict annotations of metabolic enzymes and to estimate the associated confidence levels is only limited by the availability and veracity of the necessary information in LMI. Additional evidences such as operon information (Memon et al. 2013) which can be easily incorporated in SHARP will help in improving the confidence levels in the ability of automated methods to predict accurate GPR association.

Materials and methods

Library of metabolic information

The LMI was created with (a) list of bacterial species O containing $s_m = 1{,}332$ species along with a roster of active pathways in each of them, (b) 16s rRNA sequences R of all species in O (as of September 27, 2011) downloaded from NCBI (Sayers et al. 2011), (c) Metabolic enzyme superset (E) from KEGG (Kanehisa and Goto 2000) with corresponding Enzyme Commission (EC) number, (d) names of the pathways B in which the enzymes in E participate (available as of December 2011) from KEGG, (e) amino acid sequence H of all the enzymes in E from UNIPROT (Uniprot C, 2010), (f) highly conserved domains D (Marchler-Bauer et al. 2009, 2010), orthologous groups G (Tatusov et al. 1997, 2003) of all the enzymes in E using standalone tools (RPSBLAST) from NCBI, (g) names of family F (Finn et al. 2010) to which each of the enzyme in E belongs to using RPSBLAST, and (i) enzyme-specific gene IDs I.

Query sequence identification

Using multiple sequence alignment-based phylogenetic distance (CLUSTALW) as an evolutionary distance metric, all the 1,332 species were ranked in the increasing order of distance from the query organism $o_i \in O$, which is to be annotated. The closest organism cl_j which contains the sequence h_s corresponding to the enzyme of interest is identified. (Note that the sequences used here are those from UNIPROT.) This sequence (in the closest organism) is the query sequence.

Candidate set

With h_s as query sequence (described in the previous sub-section), candidate sets $C_{i,j} = \{C_{i,j,k}\} \forall k = 1, n$ were searched using PSI-BLAST (Altschul et al. 1997) against NCBI's non-redundant database. As parameters, e-value threshold was set to 1E−5, and the maximum number of alignments was set to 30,000. All sequences corresponding to the query organism o_i obtained in at most the fourth iteration of PSI-BLAST were considered as potential candidates.

Evidence score calculation

Four individual evidence scores are estimated as follows:

(i) Reverse BLAST-based pathway similarity score ($S_{j,k,b}$): $S_{j,k,b} = \frac{1}{h}\sum_r \left(\frac{P_{b,r}}{P_b}\right) \forall r = 1, h$ where h is number of homologs corresponding to a candidate sequence participating in any of the expected pathways, and where $P_{b,r}$ and P_b, respectively, are the number of the expected pathways in which the homolog r participates and total number of possible pathways in which the query enzyme e_j participates. A cut-off of 1E−5 was set for finding the homologs of the candidate sequence.

(ii) Domain conservation score ($S_{j,k,d}$): $S_{j,k,d} = {D_k}/{D_E}$ where D_k and D_E, respectively, are the number of highly conserved domains in the kth candidate sequence and the total number of highly conserved domains (see Suppl Mat. 1) in the LMI for the query enzyme e_j. (Note that only CDD information was included in this score.)

(iii) COG conservation score ($S_{j,k,c}$): $S_{j,k,c} = {C_k}/{C_E}$ where C_k and C_E, respectively, are the number of highly conserved orthologous groups in the kth candidate sequence and the total number of such groups (see Suppl Mat. 1) in the query enzyme e_j available in LMI.

(iv) PFAM conservation score ($S_{j,k,f}$): $S_{j,k,f} = {F_k}/{F_E}$ where F_k and F_E, respectively, are the number of conserved families in the kth candidate sequence k and the total number of such families (see Suppl Mat. 1) in the query enzyme e_j. Both these numbers are estimated based on the information available in LMI.

Weight vector estimation

Suppose a query enzyme e_j is mapped by KEGG to gene ID $I^A_{i,j}$ which corresponds to sequence $C_{i,j,k}$. Using constrained optimization (described in the next sub-section), weight vector $A_{i,j}$ corresponding to $C_{i,j,k}$ is estimated. Should multiple sequences be reported for a particular enzyme, weights are obtained for each of these by considering them independently. All enzymes e_j for which the optimization fails are excluded from further analyses. In these cases, the optimization failed because it was unable to maximize when two candidates had nearly same evidence scores and only one of them was correct.

Optimization for estimation of weights

In order to obtain the optimal weights vector $A_{i,j}$ for every enzyme, a least square fitting problem is set up to maximize the overall score $S^0_{j,k} = A_{i,j} \cdot S^t_{j,k,:}$ where k represents correct sequence with the constraints

(i) $0 < A_{i,j,l} < 1 \quad \forall l = b, c, d, f$

(ii) $\sum_{l=b,c,d,f} A_{i,j,l} = 1$

If the score of a certain property is zero for all the candidates, then that property is omitted from the optimization problem. It is assumed that there is only one correct sequence. The optimization problem, implemented in Matlab®, is developed as a standalone module.

Bootstrap method for annotation prediction

The bootstrap method is described in the following steps:

(a) Assume that the kth candidate is the correct candidate, and all others as incorrect.

(b) Find the weight vector $a_{i,j,k,:} = \{a_{i,j,k,b}, a_{i,j,k,d}, a_{i,j,k,c}, a_{i,j,k,f}\}$ that maximizes the overall score $S^0_{j,k} = a_{i,j,k,:} \cdot S^t_{j,k,:}$ (Eq. [1]) if the candidate sequence k were to be the correct sequence.

(c) Repeat steps (a) and (b) for all candidate sequences $k = 1 \ldots n$ (bootstrap).

(d) Using the joint probability distribution Pr(A), find the sequence k_c that maximizes Pr($a_{i,j,k,:}$).

(e) Assign the gene ID ($I^A_{i,j}$) of this sequence C_{i,j,k_c}, which is the sequence corresponding to k_c to the query enzyme e_j. If multiple candidates satisfy (d) above, then all candidates are assigned to the query enzyme.

Acknowledgments This study was partially funded by the Department of Biotechnology, Ministry of Science and Technology, Government of India, under grant number: BT/Indo-Aus/04/04/2009 awarded to PPW. The authors thank Prof. Sharad Bhartiya for useful discussions and advice on the optimization problem. The optimization code was written by Satyajit Rao. The authors are grateful to CDAC, Pune for providing an access to the high-performance computing facilities.

References

Adriaens ME, Jaillard M, Waagmeester A, Coort SLM, Pico AR et al (2008) The public road to high-quality curated biological pathways. Drug Discov Today 13:856–862

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402

Altshul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Atsumi S, Higashide W, Liao JC (2009) Direct recycling of carbon dioxide to isobutyraldehyde using photosynthesis. Nat Biotechnol 27:1177–1180

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA et al (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75

Bartsch O, Hagemann M, Bauwe H (2008) Only plant-type (GLYK) glycerate kinases produce D-glycerate 3-phosphate. FEBS Lett 582:3025–3028

Bauer-Mehren A, Furlong LI, Sanz F (2009) Pathway databases and tools for their exploitation: benefits, current limitations and challenges. Mol Syst Biol 5:290

Becker SA, Palsson BO (2008) Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol 4:e1000082

Caspi R, Foerster H, Fulcher CA, Hopkinson R, Ingraham J, Kaipa P et al (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 34:D511–D516

Chen Y, Apolinario E, Brachova L, Kelman Z, Li Z, Nikolau BJ et al (2011) A nuclear magnetic resonance based approach to accurate functional annotation of putative enzymes in the methanogen *Methanosarcina acetivorans*. BMC Genomics 12(Suppl 1):S7

Consortium Uniprot (2010) Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res 39:D214–D219

Copeland WB, Bartley BA, Chandran D, Galdzicki M, Kim KH, Sleight SC et al (2012) Computational tools for metabolic engineering. Metab Eng 14:270–280

Deng MD, Coleman JR (1999) Ethanol synthesis by genetic engineering in cyanobacteria. Appl Environ Microbiol 65:523–528

Dexter J, Fu P (2009) Metabolic engineering of cyanobacteria for ethanol production. Energy Environ Sci 2:857–864

Ducat DC, Way JC, Silver PA (2011) Engineering cyanobacteria to generate high-value products. Trends Biotechnol 29:95

Faust K, Croes D, van Helden J (2011) Prediction of metabolic pathways from genome-scale metabolic networks. BioSystems 105:109–121

Feist AM, Palsson BØ (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. Nat Biotechnol 6:659–667

Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ (2009) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol 7:129–143

Feng X, Bandyopadhyay A, Berla B, Page L, Wu B, Pakrasi HB et al (2010) Mixotrophic and photoheterotrophic metabolism in *Cyanothece* sp. ATCC 51142 under continous light. Microbiology 156:2566–2574

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL et al (2010) The Pfam protein families database. Nucleic Acids Res 38(Database issue):D211–D222

Francke C, Siezen RJ, Teisink B (2005) Reconstructing the metabolic network of a bacterium from its genome. Trends Microbiol 13:550–558

Furnham N, Garavelli JS, Apweiler R, Thornton JM (2009) Missing in action: enzyme functional annotations in biological databases. Nat Chem Biol 5:521–525

Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30

Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee TJ, Kaipa P, Gilham F, Spaulding A, Popescu L et al (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief Bioinformatics 11:40–79

Klimke W, O'Donovan C, White O, Brister JR, Clark K, Fedorov B et al (2011) Solving the problem: genome annotation standards before the data deluge. Stand Genomic Sci 5:168–193

Knoop H, Zilliges Y, Lockau W, Steuer R (2010) The metabolic network of *Synechocystis* sp. PCC 6803: systemic properties of autotrophic growth. Plant Physiol 154:410–422

Krishnakumar S, Gaudana SB, Viswanathan GA, Pakrasi HB, Wangikar PP (2013) Rhythm of carbon and nitrogen fixation in unicellular cyanobacteria under turbulent and highly aerobic conditions. Biotechnol Bioeng 110:2371–2379

Kumar VS, Dasika MS, Maranas CD (2007) Optimization based automated curation of metabolic reconstructions. BMC Bioinformatics 8:212

Kumar A, Suthers PF, Maranas CD (2012) MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. BMC Bioinformatics 13:6

Kyrpides NC (2009) Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. Nat Biotechnol 27:627–632

Lan EI, Liao JC (2011) Metabolic engineering of cyanobacteria for 1-butanol production from carbon dioxide. Metab Eng 13:353–363

Lan EI, Liao JC (2012) ATP drives direct photosynthetic production of 1-butanol in cyanobacteria. Proc Natl Acad Sci 109:6018–6023

Lindberg P, Park S, Melis A (2010) Engineering a platform for photosynthetic isoprene production in cyanobacteria, using *Synechocystis* as the model organism. Metab Eng 12:70–79

Mao X, Chen X, Zhang Y, Pangle S, Xu Y (2012) CINPER: an interactive web system for pathway prediction for prokaryotes. PLoS ONE 7:e51252

Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, Deweese-Scott C, Fong JH et al (2009) CDD: specific functional annotation with the conserved domain database. Nucleic Acids Res 37:D205–D210

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWesse-Scott C et al (2010) CDD: a conserved domain database for the functional annotation of proteins. Nucleic Acids Res 39:D225–D229

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, Deweese-Scott C et al (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 39(Database issue):D225–9

Memon D, Singh AK, Pakrasi HB, Wangikar PP (2013) A global analysis of adaptive evolution of operons in cyanobacteria. Antonie Van Leeuwenhoek 103:331–346

Montagud A, Navarro E, de Cordoba PF, Urchueguia JF, Patil KR (2010) Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium. BMC Syst Biol 4:156

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35:W182

Moskal WA Jr, Wu HC, Underwood BA, Wang W, Town CD, Xiao Y (2007) Experimental validation of novel genes predicted in the un-annotated regions of the *Arabidopsis* genome. BMC Genomics 8:18

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 27:9–34

Ouzounis CA, Karp PD (2002) The past, present and future of genome-wide re-annotation. Genome Biol 3:Comment2001.1–2001.6

Overbeek R, Dlsz T, Stevens R (2004) The SEED: a peer-to-peer environment for genome annotation. Commun ACM 47:46–50

Overbeek R, Bartels D, Vonstein V, Meyer F (2007) Annotation of bacterial and archeal genomes: improving accuracy and consistency. Chem Rev 107:3431–3447

Peralta-Yahya PP, Zhang F, del Cardayre SB, Keasling JD (2012) Microbial engineering for the production of advanced biofuels. Nature 488:320–328

Plata G, Fuhrer T, Hsiao T-L, Sauer U, Vitkup D (2012) Global probabilistic annotation of metabolic network enables enzyme discovery. Nat Chem Biol 8:848–854

Puchałka J, Oberhardt MA, Godinho M, Bielecka A, Regenhardt D, Timmis KN et al (2008) Genome-scale reconstruction and analysis of the *Pseudomonas putida* KT2440 metabolic network facilitates applications in biotechnology. PLoS Comput Biol 4:e1000210

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C et al (2012) The Pfam protein families database. Nucleic Acids Res 40:D290–D301

Purnick P, Weiss R (2009) The second wave of synthetic biology: from modules to systems. Nat Rev Mol Cell Biol 10:410–422

Quintana N, van der Kooy F, van de Rhee MD, Voshol GP, Verporte R (2011) Renewable energy from cyanobacteria: energy production optimization by metabolic pathway engineering. Appl Microbiol Biotechnol 91:471–490

Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A et al (2013) A large-scale evaluation of computational protein function prediction. Nat Methods 10:221–227

Reed JL, Famili I, Thiele I, Palsson BØ (2006) Towards multidimensional genome annotation. Nat Rev Genet 7:130–141

Saha R, Verseput AT, Berla BM, Mueller TJ, Pakrasi HB, Maranas CD (2012) Reconstruction and comparison of the metabolic potential of cyanobacteria *Cyanothece* sp. ATCC 51142 and *Synechocystis* sp. PCC 6803. PLoS ONE 7(10):e48285

Sayers EW, Barrett T, Benson DA, Bolton E, Bryan SH, Canese K et al (2011) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 39:D38–D51

Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, Rother M et al (2011) BRENDA, the enzyme information system in 2011. Nucleic Acids Res 39:D670–D676

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423

Shastri AA, Morgan JA (2005) Flux balance analysis of photoautotrophic metabolism. Biotechnol Prog 21:1617–1626

Syed U, Yona G (2009) Enzyme function prediction with interpretable models. Methods Mol Biol 541:373–420

Takahama K, Matsuoka M, Nagahama K, Ogawa T (2003) Construction and analysis of a recombinant cyanobacterium expressing a chromosomally inserted gene for an ethylene-forming enzyme at the psbAI locus. J Biosci Bioeng 95:302–305

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV et al (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41

Thiele I, Palsson BØ (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 5:93–121

Tyo KE, Zhou H, Stephanopoulos GN (2006) High-throughput screen for poly-3-hydroxybutyrate in *Escherichia coli* and *Synechocystis* sp. strain PCC6803. Appl Environ Microbiol 72:3412–3417

Varman AM, Xiao Y, Pakrasi HB, Tang YJ (2013) Metabolic engineering of *Synechocystis* sp. strain PCC 6803 for isobutanol production. Appl Environ Microbiol 79:908–914

Viswanathan GA, Seto J, Patil S, Nudelman G, Sealfon SC (2008) Getting started in biological pathway construction and analysis. PLoS Comput Biol 4:e16

Wang B, Wang J, Zhang W, Meldrum DR (2012) Applications of synthetic biology in cyanobacteria and algae. Front Microbiol 3:1–15

Watson JD, Laskowski RA, Thornton JM (2005) Predicting protein function from sequence and structural data. Current Opin Struct Biol 15:275–284

Wu B, Zhang B, Feng X, Rubens JR, Huang R, Hicks LM et al (2010) Alternative isoleucine synthesis pathway in cyanobacterial species. Microbiology 156:596–602

Zhang S, Bryant DA (2011) The tricarboxylic acid cycle in cyanobacteria. Science 334:1551

Zhou J, Li Y (2010) Engineering cyanobacteria for biofuels and chemicals production. Protein Cell 1:207