Voor akkoord verklaring

Dit eindwerk is een examen; eventuele fouten die worden vastgesteld tijdens de eindwerkverdediging of erna werden niet gecorrigeerd. Het gebruik als referentie in publicaties is toegelaten na goedkeuring van de promotor en eindwerkbegeleider (van de stageplaats).

Dr. Paul Janssen

Stagementor

Dr. Is. Natalie Leys

Stagegever

Raphael Kiekens

Promotor

Woord vooraf

In dit woord vooraf wil ik iedereen bedanken die geholpen heeft om dit bachelorproef mogelijk te maken. Allereerst mijn stagebegeleider, Dr. Paul Janssen die mij begeleidde tijdens mijn stage. Ik heb erg veel bijgeleerd tijdens mijn stage, zowel over mijn stageonderwerp als over vele andere onderwerpen. Ook wil ik mijn ouders bedanken omdat die het mogelijk maakten voor mij om de opleiding bioinformatica te volgen en ook om deze stage te volgen. Verder wil ik ook mijn vriendin, Stéphanie Meurisse, bedanken voor de steun en aanmoediging tijdens het schrijven van mijn bachelorproef. Ten slotte wil ik ook mijn promotor, Dr. Raphael Kiekens bedanken voor de begeleiding tijdens mijn stage.

Samenvatting

De cyanobacterie *Arthrospira* werd door de European Space Agency geselecteerd voor de productie van zuurstof en als eetbaar eindproduct in het gesloten levensondersteunend systeem MELiSSA. Het is de bedoeling dat dit systeem, dat voorziet in de volledige herwinning van menselijk afval (faeces, urine, plantenresten, papier, etc.), kan worden gebruikt voor verre ruimtereizen of voor een permanente ruimtebasis. De keuze voor *Arthrospira* is niet zo vreemd omdat deze bacterie fotosynthetisch is, dus haar energie haalt uit zonlicht, CO_2 benut als koolstofbron, en water kan oxideren tot zuurstof. Bovendien werd deze bacterie reeds in de oudheid gebruikt als voedingssupplement en is *Arthrospira* vandaag een belangrijk commercieel product, met een geschatte wereldproductie van 50,000 ton per jaar.

De onderzoekseenheid Microbiologie van het SCK•CEN onderzoekt de genetische stabiliteit van de bacteriën die gebruikt worden in de MELiSSA reactor. Daarom werd in 2009 besloten om het volledige genoom van de *Arthrospira* stam PCC 8005 te sequeneren. Deze stam is sinds 1994 in gebruik door de European Space Agency. In de afgelopen twee jaar werden door andere onderzoeksgroepen, gesitueerd in de Verenigde Staten, Japan, en Thailand, de genomen van drie andere *Arthrospira* species of stammen gesequeneerd. Deze data zijn vrij beschikbaar via publieke databanken. Omdat de vier genoomprojecten dezelfde problemen ondervinden ten gevolge van het hoge aantal repetitieve sequenties werd getracht de vier genomen met elkaar te vergelijken. Dit bleek maar tot op zekere hoogte mogelijk mede door de hoge repetitiviteit maar ook wegens de veelvuldige DNA herschikkingen in de vier genomen.

Deze studie bestond uit drie delen: vergelijkende genoom analyse op DNA niveau, sequentie similariteitsonderzoek op eiwit niveau, en een fylogenetische analyse van twee enzymes van de citroenzuurcyclus.

Lijst met afkortingen en symbolen

SCK•CEN	StudieCentrum voor Kernenergie – Centre d'Étude de l'Énergie Nucléaire
BR	Belgian Reactor
LHMA	Laboratorium voor Hoge en Middelmatige Activiteit
ESA	European Space Agency
MELiSSA	Micro-Ecological Life Support System Alternative
CDS	Coding DNA Sequence
MEGA	Molecular Evolutionary Genetics Analysis
BRIG	BLAST Ring Image Generator
NCBI	National Center for Biotechnology Information
MYRRHA	Multi-purpose hybrid research reactor for high-tech applications
VITO	Vlaamse Instelling voor Technologisch onderzoek
ATP	adenosine trifosfaat
OGDC	2-oxoglutaraat decarboxylase
SSADH	succinic semialdehyde dehydrogenase

Verklarende woordenlijst

agar	bindmiddel dat gebruikt wordt om voedingsbodems te vormen voor bacteriële culturen
expect-value	verwacht aantal hits die bij kans voorkomen wanneer men een database van een bepaalde grootte doorzoekt
faeces	ontlasting
front-end	gebruikersomgeving van software
fylogenetisch	indeling volgens genetische eigenschappen
genoom annotatie	biologische relevantie relateren aan genensequenties
genoom assemblage	genoomstructuur afleiden door het overlappen van korte DNA-sequenties
motiel	de mogelijkheid om zelfstandig voort te bewegen
optical genome mapping	methode om restrictiemappen te maken op genoomniveau met één DNA-streng
phyto-chemicaliën	biologisch actieve bestanddelen die uit planten gehaald worden
proteoom	proteïne sequenties van een genoom
sequeneren	bepalen van de nucleïnezuur sequentie van een DNA-streng
similariteitsonderzoek	onderzoek naar de mate van overeenkomst tussen twee factoren
Smith-Waterman alignering	algoritme dat similariteit bepaalt tussen twee nucleotide of proteïne sequenties
transitiviteit	wanneer element A gerelateerd is aan elementen B en C dan moeten elementen B en C ook gerelateerd zijn.

Lijst van figuren

Lijst van tabellen

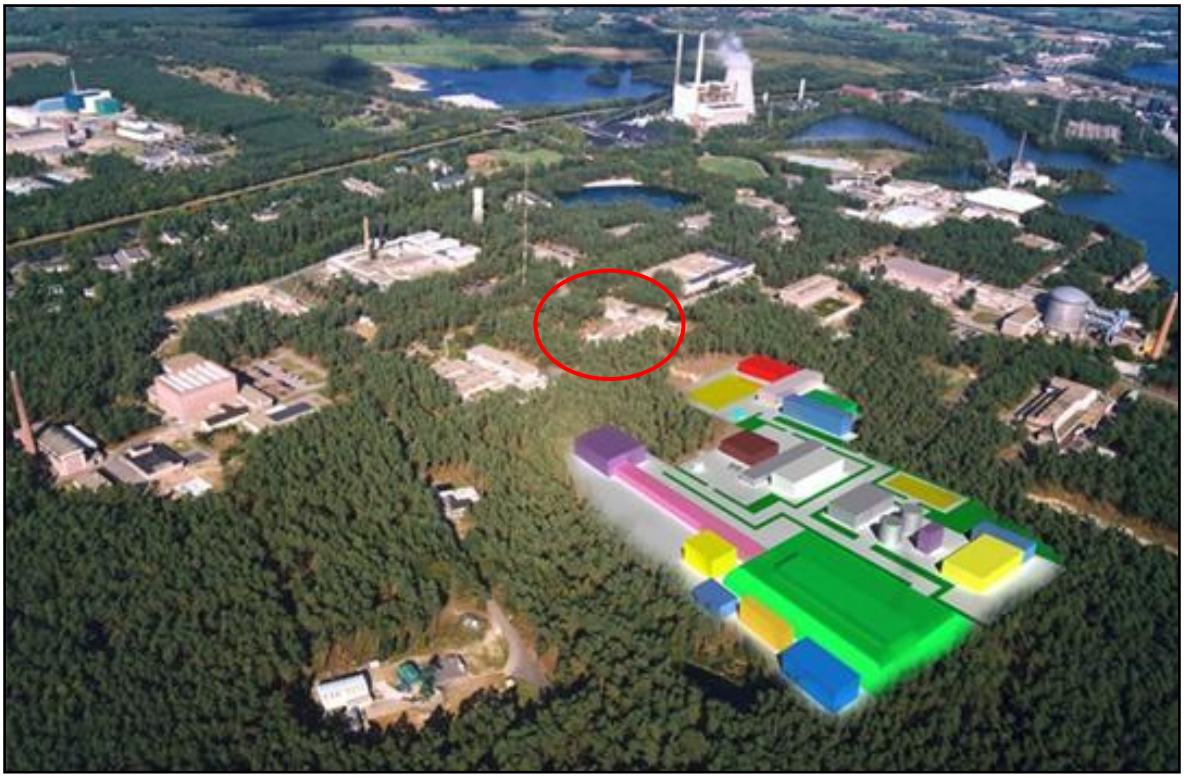
Inhoudsopgave

1 Inleiding, probleemstelling en situatieschets

1.1 SCK•CEN

1.1.1 Algemeen

Het StudieCentrum voor Kernenergie (SCK•CEN) te Mol werd opgericht in 1952 met als doel: het in stand houden van een excellentiecentrum voor onderzoek en vreedzame toepassingen van nucleaire wetenschap. Momenteel is het één van de grootste onderzoekscentra van België met ca. 680 personeel. Het SCK•CEN zet zich in voor de bevordering van de vreedzame industriële en medische toepassingen van ioniserende straling. Het is ook de toekomstige site voor de ontwikkeling van het Multi-purpose hybrid research reactor for high-tech applications project (in onderstaande figuur ingekleurd). De MYRRHA reactor moet op termijn de oude Belgian Reactor 2 (uit 1962) vervangen.



1.1 **Het domein van SCK•CEN, met de toekomstige MYRRHA reactor ingetekend. Het gebouw waarin het biologisch onderzoek is gesitueerd en waar de stage doorging is aangegeven door een rode cirkel.**

1.1.2 Onderzoeksdomeinen

Het SCK•CEN bestaat uit drie onderzoeksafdelingen die zich toespitsen op verschillende taken en een vierde afdeling die diensten en administratie omvat.

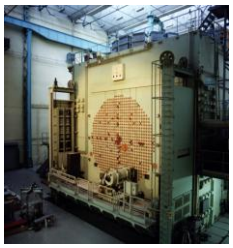
- instituut voor Nucleaire materiaalwetenschappen (NMS)
- instituut Geavanceerde Nucleaire Systemen (ANS)
- instituut voor Milieu, Gezondheid en Veiligheid (EHS)
- instituut voor Algemene Diensten en Administratie (CSA)

De afdeling Microbiologie is, samen met de onderzoekseenheden Radiobiologie en Radio-ecologie, onderdeel van het instituut EHS (milieu, gezondheid en veiligheid). In de eenheid Microbiologie wordt o.a. genetisch onderzoek verricht op de grondbacterie *Cupriavidus metallidurans* die resistent is voor een hele waaier aan zware metalen .Verder wordt onderzoek gedaan naar de bacteriële adaptatie aan extreme condities (UV, ioniserende straling), en wordt nagegaan welke de mogelijke invloed is van bacteriën op de verwerking en stockage van radioactief afval. Daarnaast doet heeft de Microbiologie eenheid ook enkele ESA projecten (zie verder).

1.1.3 Onderzoeksinstallaties

<u>BR1</u>

Belgian Reactor 1 is operationeel en is de oudste onderzoekreactor in België. De reactor kan gebruikt worden voor opleidingen en door onderzoekscentra, universiteiten en industrie.



1.2 BR1

<u>BR2</u>

1.3 BR2

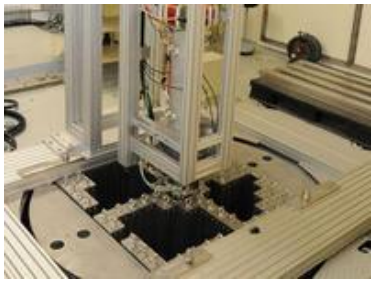
Belgian Reactor 2 is een onderzoeksreactor die gebruikt wordt voor testen met splijtstoffen en diverse materialen. De reactor is ook van groot belang voor de productie van medische en industriële radio-isotopen en silicium-dopering voor de elektronica-industrie

BR3

Belgian Reactor 3 is een prototype van drukwaterreactoren. Het was het Europese pilootproject voor reactorontmanteling en de daarbij voorkomende veiligheidsmaatregelen

1.4 BR3

VENUS-GUINEVERE



Deze reactor wordt gebruikt om reactorberekeningen te testen en zo de efficiëntie te verhogen van reactoren. Sinds 2008 wordt de reactor omgebouwd om te worden gebruikt bij het ontwikkelen van de volgende (vierde) generatie reactoren

1.5 BR4

LHMA

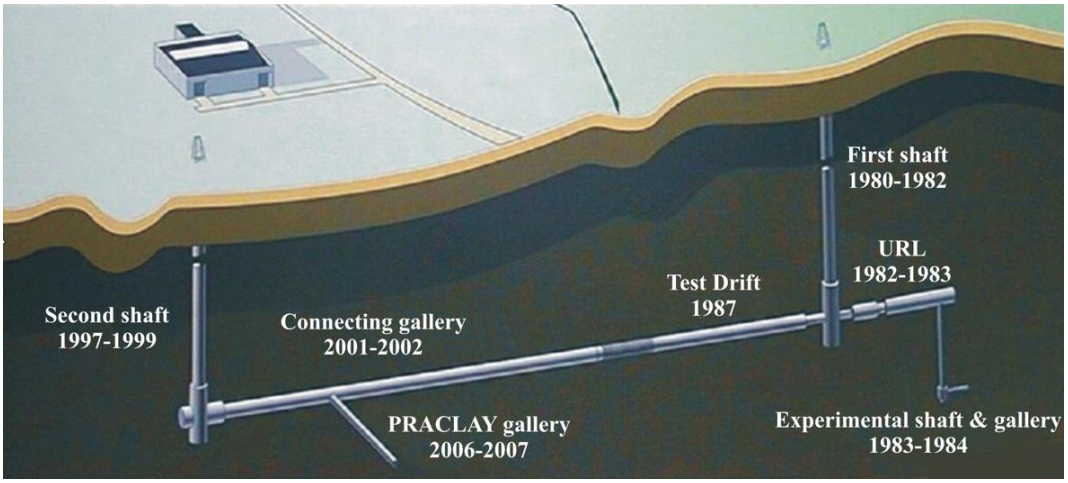
Het Laboratorium voor Hoge en Middelmatige Activiteit (LHMA) evalueert de gevolgen van straling op verschillende materialen die worden gebruikt in of bij reactoren. Met de resultaten zo verkregen kan men dan beter de veiligheid garanderen van het werken met bepaalde materialen.

1.6 LHMA

HADES

Het HADES-laboratorium bevindt zich 225 meter onder de grond en wordt gebruikt om onderzoek te verrichtten naar het opslaan van hoogradioactief afval in de Boomse kleilaag. Ook worden er testen uitgevoerd met warmte stimulatoren om na te gaan wat de mogelijke verstoring is van de warmteontwikkeling op de autochtone microbiële gemeenschappen. Immers, het hoogradioactief afval dat uiteindelijk in de Boomse klei zal worden opgeslagen zal nog een temperatuur van ca. 80 graden celsius hebben. Hierdoor kunnen bepaalde bacteriën de overhand

krijgen met als mogelijk gevolg gas- en zuurvorming, wat nadelig zou kunnen zijn voor de langdurige opslag van het radioactieve afval.



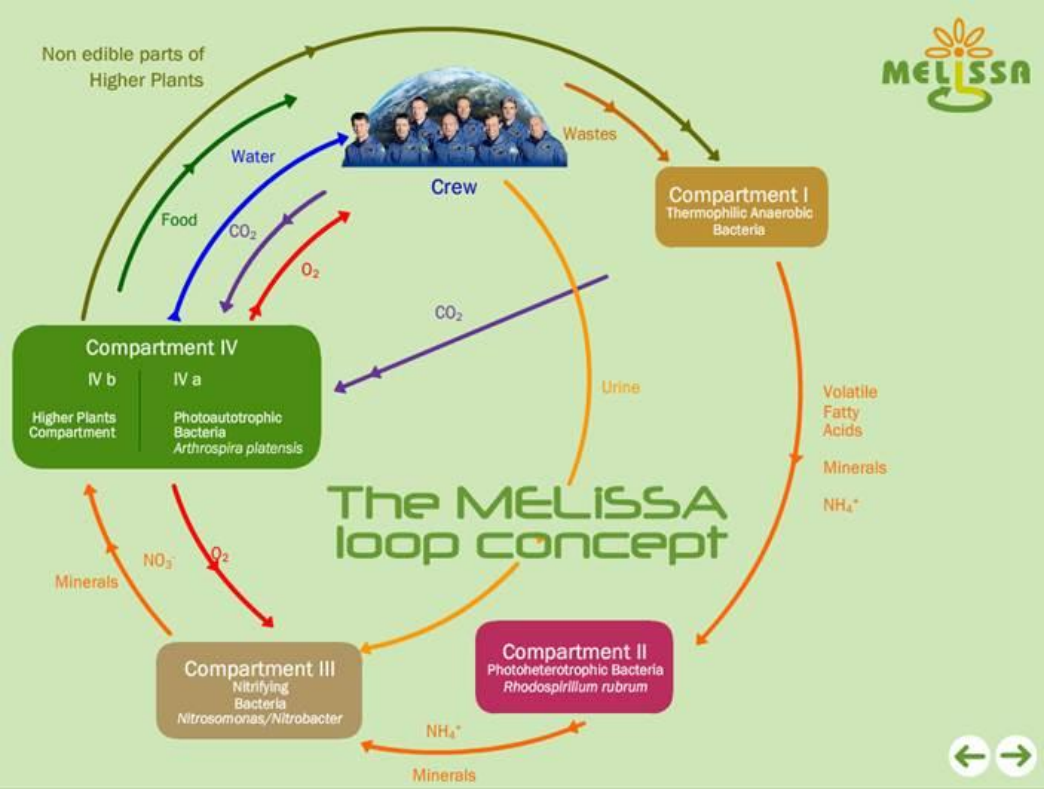
1.7 Ondergrondse galerijen en laboratorium (HADES)

(SCK•CEN Website), http://www.sckcen.be/

1.2 ESA en MELiSSA

Het Micro-Ecological Life Support System Alternative (MELiSSA) is een project van het ESA. Het is een gesloten recyclage systeem voor de recuperatie van voedsel uit menselijk afval. Het systeem bestaat uit vier compartimenten. In het 1ste compartiment wordt menselijk afval (faeces) en plantaardig afval, papier, etc., verwerkt waarbij lange koolstofketens worden afgebroken. In het 2^{e} compartiment worden belangrijke mineralen, vluchtbare vetzuren, en ammonium (NH_4^+) geëxtraheerd. Het ammonium wordt verder omgezet naar nitriet en nitraat in het 3^{e} compartiment en het bekomen nitraat dient tenslotte als stikstofbron voor het 4^{e} compartiment waarin planten en cyanobacteriën worden opgegroeid. Deze verbruiken het CO_2 en genereren zuurstof. Het is de bedoeling dat dit systeem gebruikt zal worden bij langdurige projecten in de ruimte zoals een maanbasis of een reis naar mars. Delen van de reactor o.a. de waterzuivering vinden ook hun toepassing op aarde, vooral dan in ver afgelegen, moeilijk bereikbare gebieden. Proefopstellingen van de diverse compartimenten zijn momenteel aanwezig in Clermont-Ferrand, Barcelona, en Gent. Het VITO (Vlaamse Instelling voor Technologisch

onderzoek – gelegen op dezelfde campus als het SCK•CEN – huisvest een pilootproject (Belissima) waarin de vier onderdelen worden geïntegreerd in één systeem.



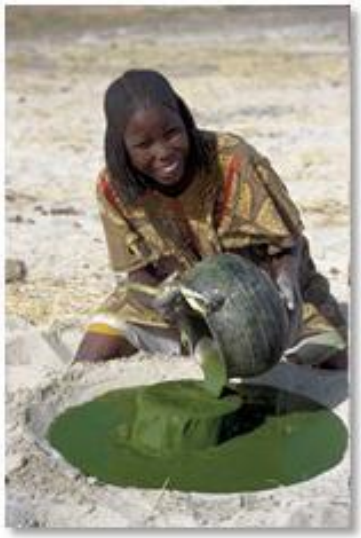
1.8 De MELiSSA cyclus schematisch voorgesteld. Planten en cyanobacteriën (*Arthrospira*) nemen CO_2 op, benutten gevormde nitraten, en produceren O_2. *Arthrospira* zijn rijk aan vitaminen, mineralen, eiwitten, en essentiële vetzuren.

1.3 Arthrospira

1.3.1 Algemeen

De cyanobacterie *Arthrospira* werd gekozen door de ESA voor de productie van zuurstof en als nutritioneel eindproduct in de MELiSSA loop. Deze eetbare bacterie heeft tal van nuttige eigenschappen. Zo is *Arthrospira* goed verteerbaar omdat het verhoudingsgewijs veel meer eiwitten bevat dan DNA. Het is ook rijk aan vitaminen (B12, C, E), mineralen, spoorelementen (micro-mineralen), omega-3 vetzuren, bètacaroteen, chlorofyl, en een breed spectrum van phyto-chemicaliën. In

feite bevatten zij elk nutriënt dat een menselijk lichaam nodig heeft. *Arthrospira* werden reeds als voedsel gebruikt door de Azteken in Midden-Amerika (waar het gekend was als 'tecuitlatl') en door inwoners van centraal Afrika ten tijde van het Kanem Keizerrijk (waar het gekend was als 'dihé'). De natuurlijke habitat van de bacterie is in warme meren in het evenaargebied. Deze meren zijn erg rijk aan carbonaten, de primaire C-bron voor deze soort van cyanobacterien. Momenteel wordt *Arthrospira* biomassa op industriële schaal aangemaakt, vooral in Azië met name in Thailand, Indië en Mongolië. Dit gebeurt in kunstmatige vijvers of in glasbuizen. Omdat de bacterie in meercellige filamenten groeit is het gemakkelijk oogstbaar door eenvoudige filtratie. Het commerciële product is ook gekend onder de naam "Spirulina" en is vrij verkrijgbaar in grote winkelketens (bv. Colruyt) of in bioshops. Een belangrijk gegeven is dat *Arthrospira* zich zeer goed heeft weten aan te passen en bestand is tegen hoge zout concentraties (tot 0.5 M NaCl), hoge alkaliwaarden (tot pH 11) en intensief licht (UV). Belangrijk is ook het feit dat *Arthrospira* goed bestand is tegen ioniserende straling. Onderzoek aan het SCK•CEN heeft aangetoond dat *Arthrospira* kan groeien onder acute dosissen van 5,000 Gy gammastraling. Dit is uiteraard van belang voor de toepassing van *Arthrospira* in de ruimte waar kosmische straling een belangrijke factor is. De reden voor deze resistentie is onbekend en is het onderwerp voor wetenschappelijk onderzoek in de eenheid Microbiologie van het SCK•CEN.



1.9 Traditionele winning van *Arthrospira*

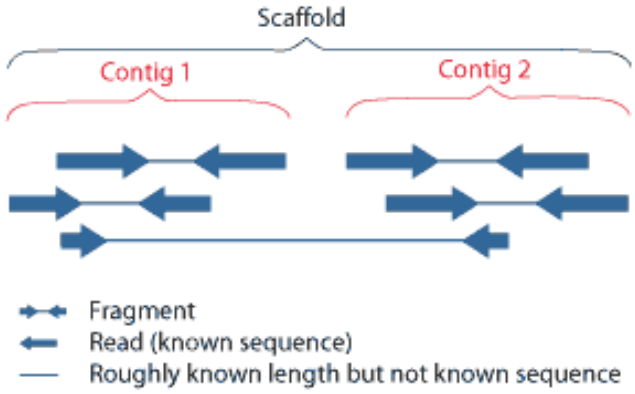
1.10 *Arthrospira* in commerciële vorm ("Sprulina")

1.11 *Arthrospira* groeit in meercellige filamenten

1.3.2 Genoomannotatie

Om *Arthrospira* te optimaliseren voor zijn functie in het MELiSSA project en om het organisme verder te bestuderen werd, onder leiding van de SCK•CEN eenheid Microbiologie, het genoom van *Arthrospira* stam PCC 8005 door GénoScope (Evry, Frankrijk) gesequeneerd. Dit is de stam die de ESA in 1994 heeft gekozen voor de MELiSSA reactor en die in de SCK•CEN eenheid Microbiologie wordt bestudeerd. Omdat deze stam nog niet tot op species niveau geïdentificeerd is wordt de notatie *Arthrospira* sp. PCC 8005 gebruikt. Andere onderzoeksgroepen startten ook een genoomproject, op andere species of stammen. Hiertoe behoren de genoomprojecten *Arthrospira maxima* CS-328 (Department of Energy, USA), *Arthrospira platensis* NIES-39 *(*National Institute of Technology and Evaluation, Japan), *Arthrospira platensis* C1 (King Mongkut's University of Technology, Thailand), en van *Arthrospira platensis* sp. Paraca (Fachhochschüle Westschweiz, Zwitserland). Deze laatste werd niet opgenomen in het stageonderzoek omdat de data van dit genoom te fragmentarisch is. Momenteel is voor geen enkele van deze stammen het genoom volledig in kaart gebracht (Tabel 1 op volgende blz). Wel is voor de NIES-39 en C1 genomen de oriëntatie en volgorde van de scaffolds gekend.

De assemblage en annotatie van een genoom is net als het oplossen van een grote puzzel. De korte reads, dit zijn telkens 40-100 nucleotiden die "gelezen" worden door de sequentie technologie, worden aan de hand van speciale software met elkaar verbonden (er wordt naar een overlap gezocht) om grotere fragmenten op te bouwen. Deze worden contigs genoemd. Van deze contigs worden dan grotere fragmenten samengesteld ge-



1.12 De "assemblage" van een genoom

naamd scaffolds en van deze scaffolds wordt dan uiteindelijk het volledige genoom samengesteld.

Genome Name	*A. platensis C1*	*A. platensis NIES-39*	*A. maxima* CS-328	*A. platensis Paraca*	*A. platensis PCC8005*
Genome size (bp)	6,089,210	6,788,435	6,003,314	4,997,563	6,145,553
Total genes	6,153	6,676	5,730	5,401	5,718
Protein coding genes	6,108	6,630	5,690	5,370	5,675
Protein with function prediction	3,757	2,542	3,315	3,023	3,023
RNA genes	45	46	40	31	43
Enzymes	952	905	889	816	882
% Enzymes	15.47%	13.56%	15.51%	15.11%	15.42%
Transporter Classification	345	NA	NA	NA	NA
%Transporter Classification	5.61%	NA	NA	NA	NA
KEGG pathways	1,012	993	931	904	954
% KEGG pathways	16.45%	14.87%	16.25%	16.74%	16.68%
KEGG Orthology (KO)	1,837	1,702	1,623	1,492	1,658
% KEGG Orthology (KO)	29.86%	25.49%	28.32%	27.62%	29.00%
COGs	3,459	3,570	3,306	3,050	3,234
% COGs	56.22%	53.48%	57.70%	56.47%	56.56%
Pfam	3,529	3,598	3,564	3,431	3,526
% Pfam	57.35%	53.89%	62.20%	63.53%	61.66%
TIGRfam	1,180	1,213	1,160	1,185	1,203
% TIGRfam	19.18%	18.17%	20.24%	21.94%	21.04%
InterPro	4,244	3,969	3,938	4,207	4,294
% InterPro	68.97%	59.45%	68.73%	77.89%	75.10%
signal peptides	570	1,319	545	559	1,150
% signal peptides	9.26%	19.76%	9.51%	10.35%	20.11%
Transmembrane proteins	1,094	1,123	1,053	1,094	1,057
% Transmembrane proteins	17.78%	16.82%	18.38%	20.26%	18.49%
COG clusters	1,569	1,566	1,491	1,489	1,472
KOG clusters	729	723	722	737	717
Pfam clusters	1,740	1,732	1,702	1,729	1,730
TIGRfam clusters	932	936	902	924	931

Tabel 1: Genoom statistieken van vijf Arthrospira genomen (ontleend uit: Cheevadhanarak et al. (2012). Stand. Genomic Sci. 2012 6:1)

De genoom assemblage en annotatie voor stam PCC 8005 is zeer problematisch gebleken omdat het een groot aantal repetitieve elementen bevat, wat het correct aan elkaar passen van de contigs en scaffolds erg bemoeilijkt. De huidige versie van het PCC 8005 genoom bestaat uit vier grote scaffolds en 530 kleine "scaffolds" (Tabel 2). De kleine scaffolds, die eerder als contigs kunnen beschouwd worden omdat ze gemiddeld slechts 1 kb lang zijn en slechts 1-3 genen bevatten met bovendien onbetrouwbare sequenties, zijn een gevolg van de onmogelijkheid om deze vaak voorkomende sequenties te lokaliseren of te assembleren aan de hand van sequentie similariteiten. Dit betekent concreet dat de sequenties in de 530 kleine scaffolds (in feite contigs) ofwel technische duplicaten zijn, veroorzaakt door technologische tekortkomingen, ofwel natuurlijke duplicaten zijn, dat wil zeggen paraloge sequenties die ontstaan door gen duplicatie. Dit alles maakt dat de data van de 530 kleine scaffolds, of ca. 530 kb of 8% van het genoom voor vergelijkende studies onbruikbaar is. (P.J.Janssen et. al, 2010)

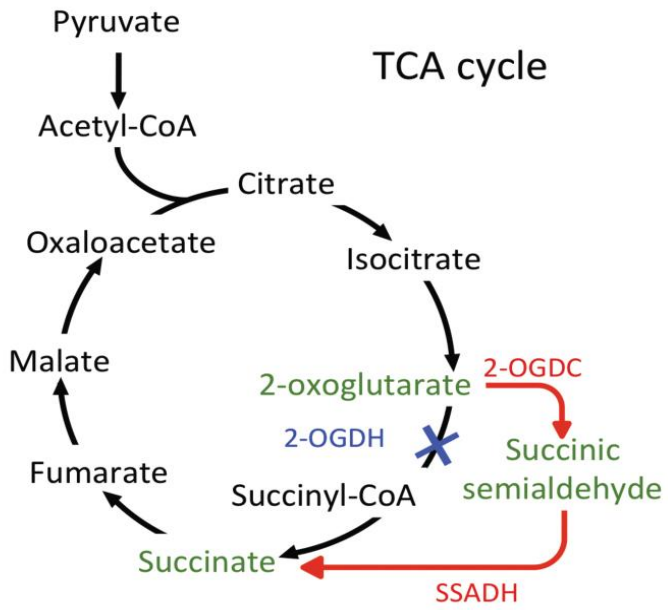
Een vierde versie van het PCC 8005 genoom wordt verwacht einde 2012 en zal bestaan uit 1 tot 4 scaffolds, zonder extra contigs.

	Version 1	Version 2	Version 3
Sequence length (bp)	6,279,260	5,397,084	6,715,674
GC%	44.72	44.65	44.63
Repeated Regions (%)	29.15	15.08	42.96
Number of genes (CDS,RNA,tRNA,…)	6,032	5,454	6,497
Predicted CDS	5,844	5,365	6,372
Average CDS length (bp)	905	900	850
Average intergenic length (bp)	227	185	273
5' overlapping CDS	213	NA	330
Protein coding density (%)	76	81	80
Number of contigs	119	170	692
Number of scaffolds	16	8	534

Tabel 2: De drie versies van PCC 8005 genoom data. Version 1: Janssen *et al.* (2010) J. Bacteriol. 192(9):2465-2466. Version 2: tussentijdse versie met aangepaste assemblage en sequenering. Version 3: laatste versie met nieuwe assembly deels gebaseerd op nieuwe paired-end sequencing.

1.3.3 Citroenzuurcyclus (TCA-cyclus)

Tot voor kort werd algemeen aanvaard dat de citroenzuurcyclus ("tricarboxylic acid" of TCA cyclus; ook gekend als de Krebs-cyclus) onvolledig was in cyanobacteriën doordat het gen dat voor het 2-oxoglutaraat dehydrogenase codeert in hun genomen ontbreekt. Hierdoor zou 2-oxoglutaraat niet omgezet kunnen worden naar succinaat, althans niet volgens de klassieke route via succinyl-coenzyme A (CoA) (Fig. 1.13). Uit een recent artikel van Shuyi Zhang en Donald A. Bryant blijkt echter dat de meeste cyanobacteriën twee tot dusver onbekende genen tot hun beschikking hebben coderend voor alternatieve enzymen die de cyclus kunnen vervolledigen. Dit zijn 2-oxoglutaraat decarboxylase (OGDC) en succinic semialdehyde dehydrogenase (SSADH) (Fig. 1.13). Omdat de TCA cyclus een zeer centrale rol speelt in de energievoorziening van de cel via de generatie van het energierijke adenosine trifosfaat (ATP), en in cyanobacteria het gevormde 2-oxoglutaraat een belangrijk signaalmolecule is in stikstof- en koolstof metabolisme, is het interessant na te gaan wat de situatie is voor *Arthrospira* en aanverwante species. (Zhang & Bryant, 2011)



1.13 Alternatieve TCA-cyclus in cyanobacterien

2 Materiaal en methoden

2.1 Data

De data werden gedownload van diverse databanken, enerzijds als DNA sequentie in FASTA formaat (als .fna bestanden), anderzijds als eiwit sequenties van de coderende genen (als .faa bestanden).

Stam nummer	bron	Grootte (bp)	# scaffolds	# contigs	complete	ordered
PCC 8005	MaGe	6.715.674	4 (+530)	692	no	no
CS-328	NCBI	6.131.314	-	129	no	no
NIES-39	NCBI	6.788.435	19	-	no	yes
C1 (&)	IMG	6.089.210	1 (*)	-	no	yes

(&) data pas beschikbaar in mei 2012; (*) 63 onbekende regio's ("gaps")

Tabel 3: Bron en eigenschappen van de vier *Arthrospira* genomen

2.2 BRIG

BLAST Ring Image Generator (BRIG) is een programma dat het standaard BLAST programma gebruikt. De BLAST-parameters, de input en de weergave van de output kunnen met een front-end interface worden ingesteld en worden dan meegegeven aan het BLAST programma. De output van de BLAST wordt uitgezet op concentrische cirkels in % identiteit met de referentie.

De BLAST berekening werd uitgevoerd voor een standaard expect-waarde (e) van $1 * 10^{-10}$ (Alikhan et al., 2011)

2.3 CIRCOLETTO

Circoletto is een programma dat BLAST samenbrengt met Circos. Circos is een veelzijdig visualisatie programma dat zich richt op circulaire voorstellingen. Circoletto zet beide sequenties uit op een cirkel en verbindt dan de gebieden die met elkaar overeenkomen. Op deze manier kan men gemakkelijk de posities van genen tussen verschillende sequenties bestuderen.
Circoletto werd uitgevoerd voor een expect-waarde (e) van $1 * 10^{-40}$

(Darzentas, 2010)

2.4 NCBI blast / fylogenie

De website van het Nationale Center for Biotechnology Information biedt een uitgebreide set van tools aan gaande van aligneringen tot fylogenetische bomen opmaken. Er werd gebruikt gemaakt van BLASTp om te bepalen of de genen coderend voor de twee enzymen van de alternatieve TCA-pathway aanwezig zijn in de *Arthrospira* genomen. Twee fylogenetische bomen werden opgemaakt met telkens de honderd beste BLAST resultaten van de twee enzymen tegen de NCBI protein databank met behulp van het UPGMA algoritme. (Camacho, et al., 2009)

2.5 MEGA

MEGA (Molecular Evolution Genetics Analysis) brengt verschillende software voor alignering en fylogenetisch onderzoek samen in één programma. Aligneringen werden uitgevoerd met ClustalW en fylogenetische bomen werden geconstrueerd met het UPGMA algoritme. Om de bomen te controleren werd de bootstrap methode gebruikt. Hierbij werden 500 (aantal kan zelf bepaald worden) bomen gemaakt startend van een willekeurige sequentie en wordt dan een “gemiddelde” of consensus boom samengesteld. (Tamura K, 2011)

2.6 Generage

Generage gebruikt een all-versus-all BLASTp pipeline om de similariteit tussen grote datasets van eiwit sequenties (proteomen) te bepalen. Hierdoor wordt een similariteitsmatrix opgebouwd die erg groot kan worden, bv. bij de vergelijking van de proteomen van NIES-39 en PCC 8005 (Tabel 1) wordt dat een matrix van 6.630 x 5.370. Voor gans de matrix wordt de transitiviteit gecheckt. Een voorbeeld van foute transitiviteit is als eiwit A een match vertoont met eiwit B, maar eiwit B heeft geen match met eiwit A. Dit is een 1-0 situatie in de similariteitsmatrix en dient uitgeklaard te worden met behulp van Smith-Waterman alignering (deze stap wordt ook de 'symmetrificatie' genoemd). Dit zorgt voor een zeer betrouwbaar resultaat maar vertraagt ook het proces, vooral voor grote families van eiwitten of als

bepaalde eiwit domeinen veelvuldig voorkomen. Omdat de volledige matrix wordt ingelezen en alle data en de gecorrigeerde resultaten worden opgeslagen in een tijdelijk geheugen, dient men over de nodige hardware te beschikken met voldoende RAM geheugen en rekenkracht.

Generage werd uitgevoerd met standaardwaarden voor BLASTp; enkel resultaten met een expect-waarde $1 * 10^{-10}$ of beter werden behouden. Voor transitiviteitsmetingen en clustering werden eveneens standaardwaarden gebruikt, met een *Z*-score voor de symmetrificatie and multi-domain detection van 10 and 3, respectievelijk.

(Enright & Ouzounis, 2000)

2.7 Ortho-MCL/Perl

Ortho-MCL maakt ook gebruik van all-versus-all BLASTp maar slaat alles onmiddellijk op in een MySQL databank. Bovendien is de clustering van de eiwit sequenties uitgevoerd met een Markov clustering algoritme. Ortho-MCL bestaat uit een reeks Perl-scipten die stap voor stap het proces doorlopen.

(Enright et al, 2002)

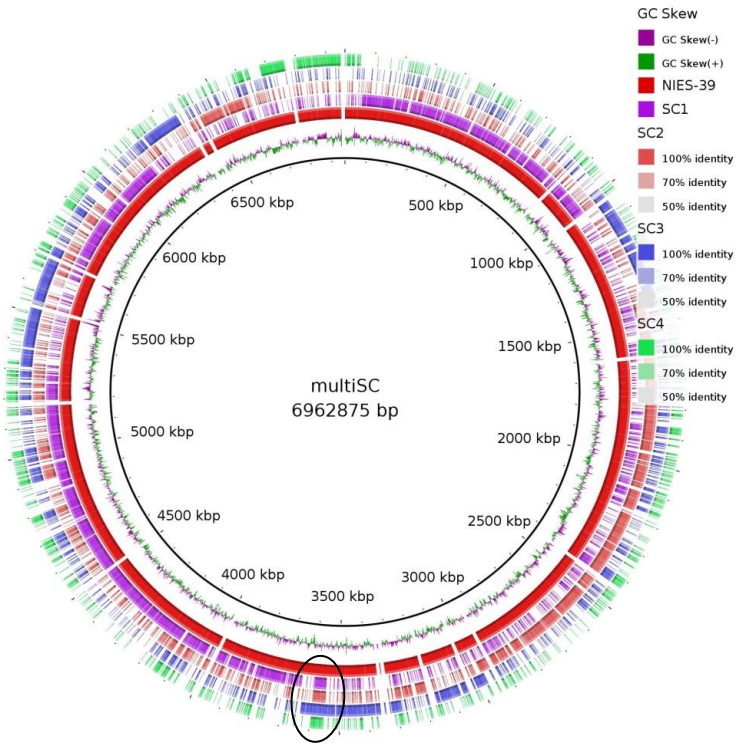
3 Resultaten

3.1 DNA-onderzoek (alignering en visualisatie)

3.1.1 BRIG

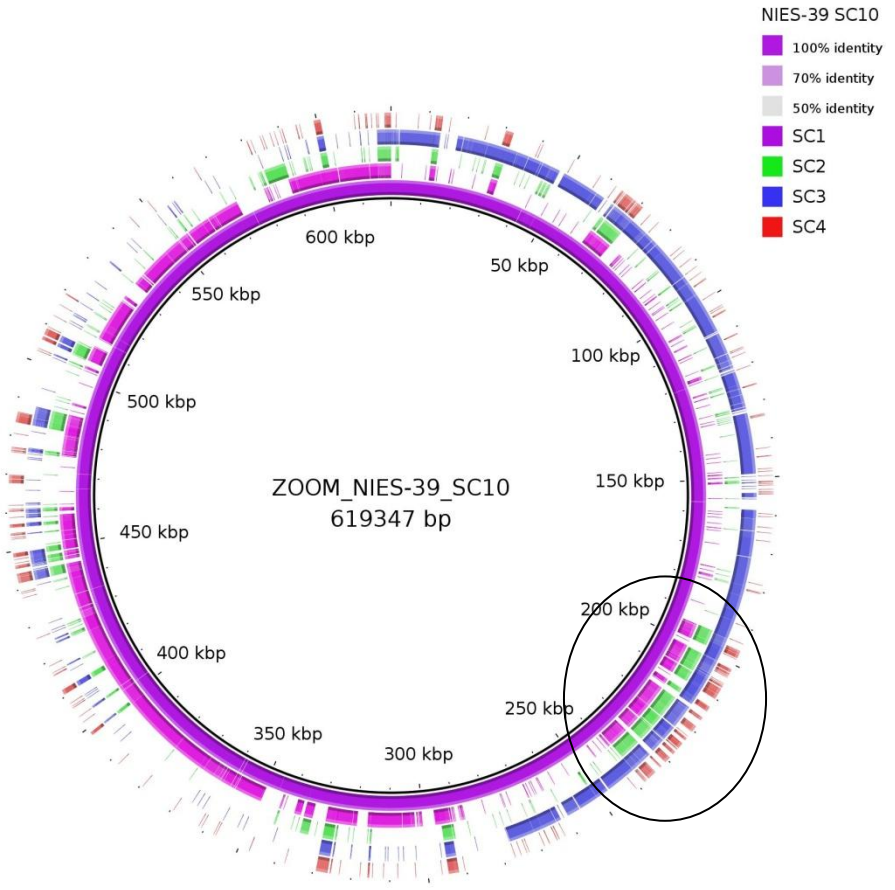
Om meer te leren over het PCC8005 genoom werd een vergelijkend onderzoek gevoerd tussen *Arthrospira* PCC8005 (vier scaffolds) en *Arthrospira* NIES-39 (negentien scaffolds). Belangrijk hierbij is dat bij NIES-39 de volgorde en oriëntatie van de scaffolds is vastgesteld met behulp van optical genome mapping. Om deze twee genomen te vergelijken werd gebruikt gemaakt van BRIG (Blast Ring Image Generator) (Alikhan, Petty, Zakour, & Beatson, 2011). BRIG maakt gebruikt een lokaal geïnstalleerde BLAST en verwerkt de resultaten tot een circulaire visualisatie. BLAST opties kunnen worden ingesteld in de BRIG software en de output is volledig aanpasbaar.

Bij de eerste alineëring werden de vier scaffolds van PCC8005 ge-BLAST tegenover de 19 scaffolds van NIES-39 (fig. 3.1)



3.1 Blast resultaten van PCC8005 scaffolds vs NIES-39 scaffolds uitgezet op concentrische ringen. De binnenste ring is de referentie en de ringen doorrond representeren elk één scaffold van PCC8005.

De binnenste rode ring bestaat uit de negentien scaffolds van NIES-39 met telkens een gap ertussen. De vier scaffolds van PCC8005 komen elk overeen met verschillende delen van het NIES-39 genoom. Dit kan er eventueel op wijzen dat de verdeling in vier scaffolds van het PCC8005 nog niet volledig correct is, anderzijds is het ook mogelijk dat dit veroorzaakt wordt door DNA uitwisseling ("shuffling") in het PCC8005 genoom. Op één positie, rond 3600 kbp op de figuur (omcirkeld), is er wel een regio die voorkomt op alle scaffolds van PCC8005 alsook op het tiende scaffold bij NIES-39. Deze regio werd verder onderzocht met BRIG met enkel het tiende scaffold van NIES-39 en verder ook met Circoletto.



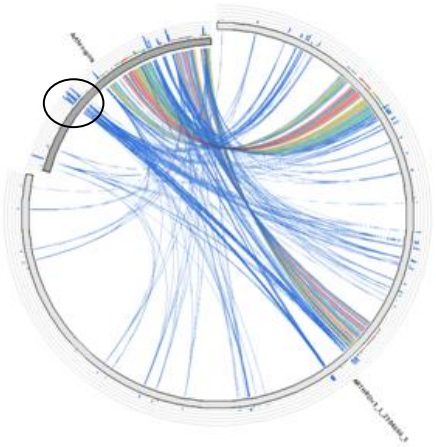
3.2 BRIG analyse van de vier scaffolds van PCC8005 t.o.v. het tiende scaffold van NIES-39

Op figuur 3.2 is duidelijk een regio van 45 kb zichtbaar die op alle scaffolds voorkomt. Deze regio bevat een 50-tal genen die blijkbaar op elk van de vier grootste scaffolds van PCC8005 gegroepeerd voorkomen (één groep genen per scaffold). Een dergelijk situatie kan zich voordoen als deze regio in feite een mobiel element is bvb. een genomisch eiland of een geïntegreerde fage. De BRIG analyse duidt echter enkel op de aanwezigheid van deze genen maar zegt niets over hoe de genen gelokaliseerd zijn. Daarvoor werd Circoletto gebruikt.

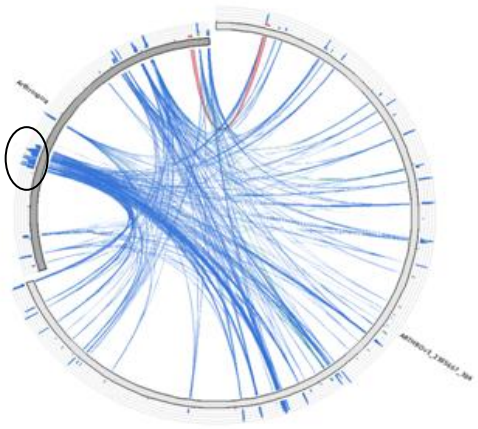
3.1.2 Circoletto

Bij Circoletto worden de genen met een similariteit boven een vastgestelde threshold met elkaar verbonden door middel van een gekleurde lijn. Op deze manier kunnen we op figuur 3.3 zien waar de genen die we willen onderzoeken gelegen zijn op de scaffolds van PCC8005.

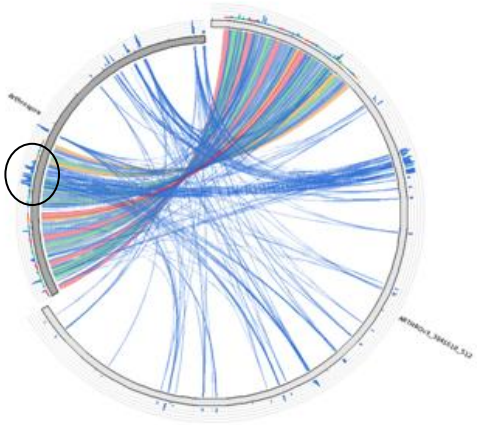
Bij drie van de vier scaffolds (1, 2, en 4) zijn de genen verspreid terwijl bij scaffold 3 de meeste genen samen zijn gegroepeerd. Dit wijst er op dat binnen de hoog geconserveerde regio tal van verplaatsingen hebben plaatsgevonden. Zulke interne verschuivingen binnen een genomic island (of geïntegreerde fage) zullen veelvuldiger voorkomen indien meerdere kopijen van het mobiele element aanwezig is aangezien de kans voor interne recombinatie toeneemt. Homologe DNA recombinatie kan gemakkelijker optreden tussen DNA gebieden in het genoom die door de evolutie heen gedupliceerd zijn en waarvan de sequentie vrij goed is bewaard.



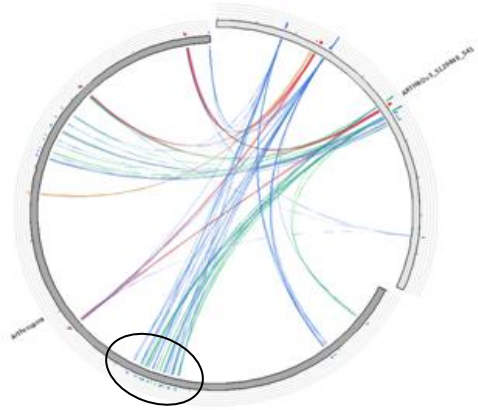
Scaffold 1



Scaffold 2



Scaffold 3

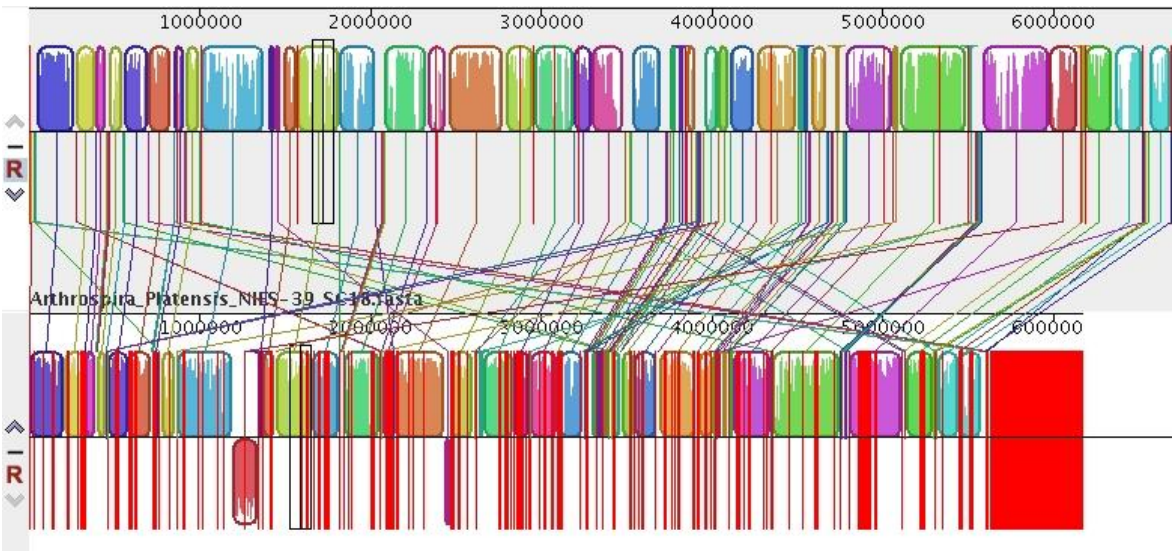


Scaffold 4

3.3 De vier scaffolds van PCC8005 (wit) tegenover het tiende scaffold van NIES-39 (grijs) uitgezet. De kleuren wijzen op de mate van simulariteit. (blauw < groen < geel < rood)

3.1.3 Mauve

Op figuur 3.4 is een alignering met het programma Mauve (Alikhan, Petty, Zakour, & Beatson, 2011) te zien. De verschillende gekleurde blokken geven regio's weer die een goede similariteit vertonen tussen de twee genomen. De pieken in deze blokken zijn een maat voor de hoge similariteit in deze stukken. Interessant hier is dat één regio (4929333 - 5128447) , bestaande uit twee opeenvolgende contigs (153 en 154) omgekeerd voorkomt (voorgesteld door de positie ten opzichtte van de middenas; links onderaan de figuur). Dit kan verklaard worden door het voorkomen van één of meerdere transposons in deze regio op het genoom. Tijdens één van de herschikkingen door de transposon genen werden deze contigs in omgekeerde volgorde terug in het genoom gebracht.



3.4 Mauve alignering. NIES-39 boven en PCC8005 onderaan

3.2 Vergelijking op Eiwit-niveau

3.2.1 GeneRAGE

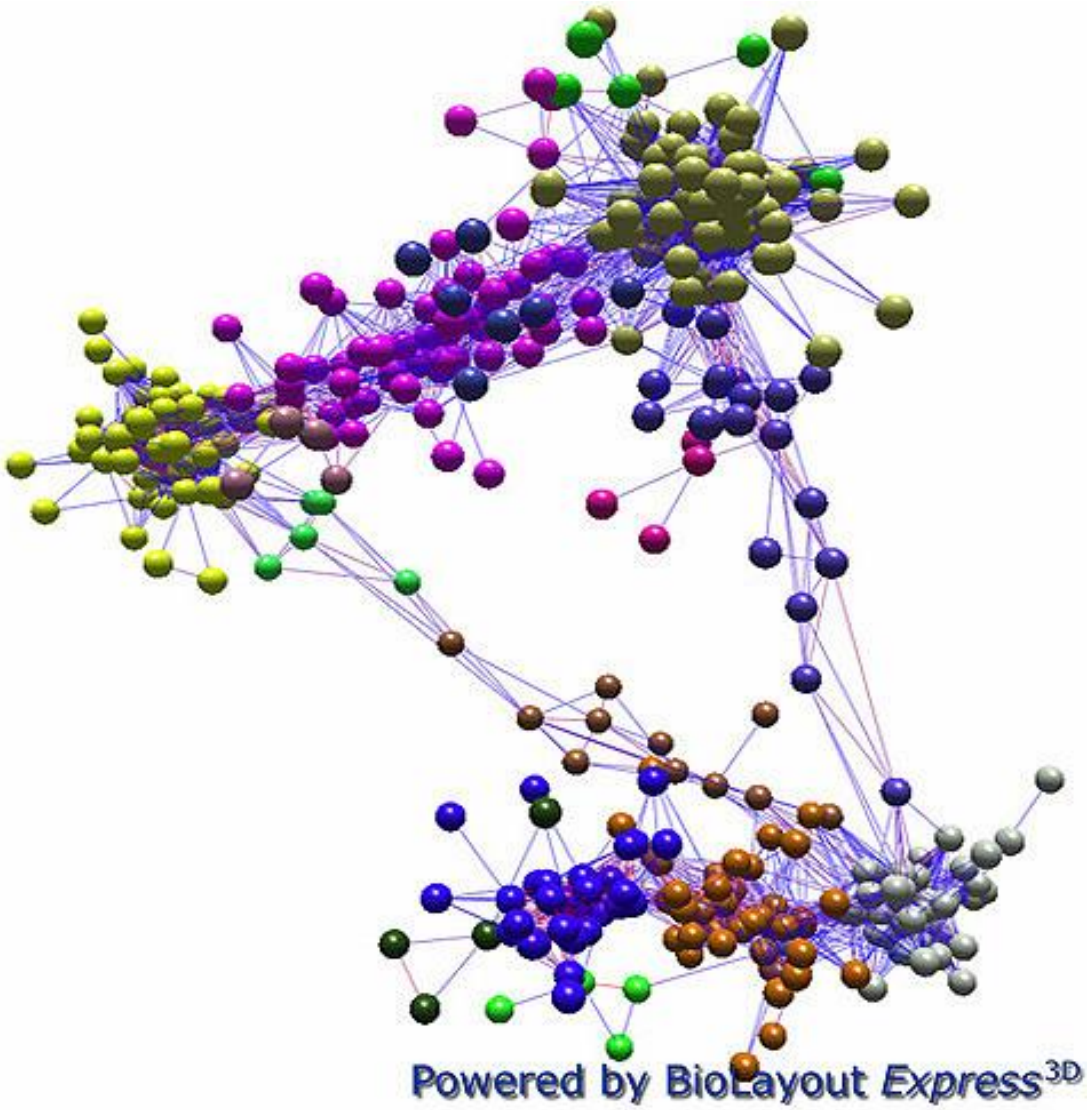
GeneRAGE werd uitgetest voor de vergelijking van *Arthrospira* proteomen (bv. NIES-39 en PCC 8005) en ook voor de vergelijking van eiwit sequenties van bv. scaffold 1 van PCC 8005 met het ganse proteoom van NIES-39. Hiervoor werd een LINUX machine gebruikt met 2 CPU en 4 Gb RAM. Helaas heeft het programma drie dagen lang gelopen en werd daarna stopgezet: in de logfiles werden telkens vele foutmeldingen genoteerd in verband met een tekort aan tijdelijk geheugen voor de transitiviteitscontrole (symmetrificatie) uitgevoerd met de Smith-Waterman alignerings methode. Dit werd veroorzaakt door de hoge interne redundantie van de genomen en de hoge verwantschap tussen de genomen, in combinatie met het rekenintensieve en dus langzame Smith-Waterman algoritme. Het GeneRage programma werd wel met succes uitgetest op kleinere datasets met minder redundantie. Het is dus geen softwarematig probleem maar een falen van de hardware.

3.2.2 Ortho-MCL/Perl

Ortho-MCL heeft twee verschillende Perl-modules nodig om naar behoren te kunnen werken: DBI en DBD. DBI staat voor DataBase Interface en DBD voor DataBase Driver. Deze twee modules zorgen voor de connectie met de MySQL databank via Perlsscripts. Bij het aanroepen van de DBD-module was er echter een probleem dat ondanks verwoede pogingen en veel zoekwerk niet kon worden opgelost. Het opzetten van de MySQL databank en de installatie van de basis software en de verschillende pipelines waren probleemloos maar we konden Ortho-MCL niet uittesten omdat resultaten niet konden worden weggeschreven in SQL. Een mogelijk oorzaak is een compatibiliteitprobleem tussen de versies van de Perl-modules en de MySQL server ,maar hierover kon niets over worden teruggevonden in de documentatie. We hebben verschillende versie van DBI en DBD uitgeprobeerd, echter zonder het verhoopte resultaat.

3.2.3 Alternatieven

Cytoscape en het visualisatieprogramma BioLayout hebben beiden een MCL plug-in. Echter, door de grote hoeveelheid data konden deze programma's de bewerkingen niet uitvoeren (de programma's kwamen tot stilstand of "crashten").



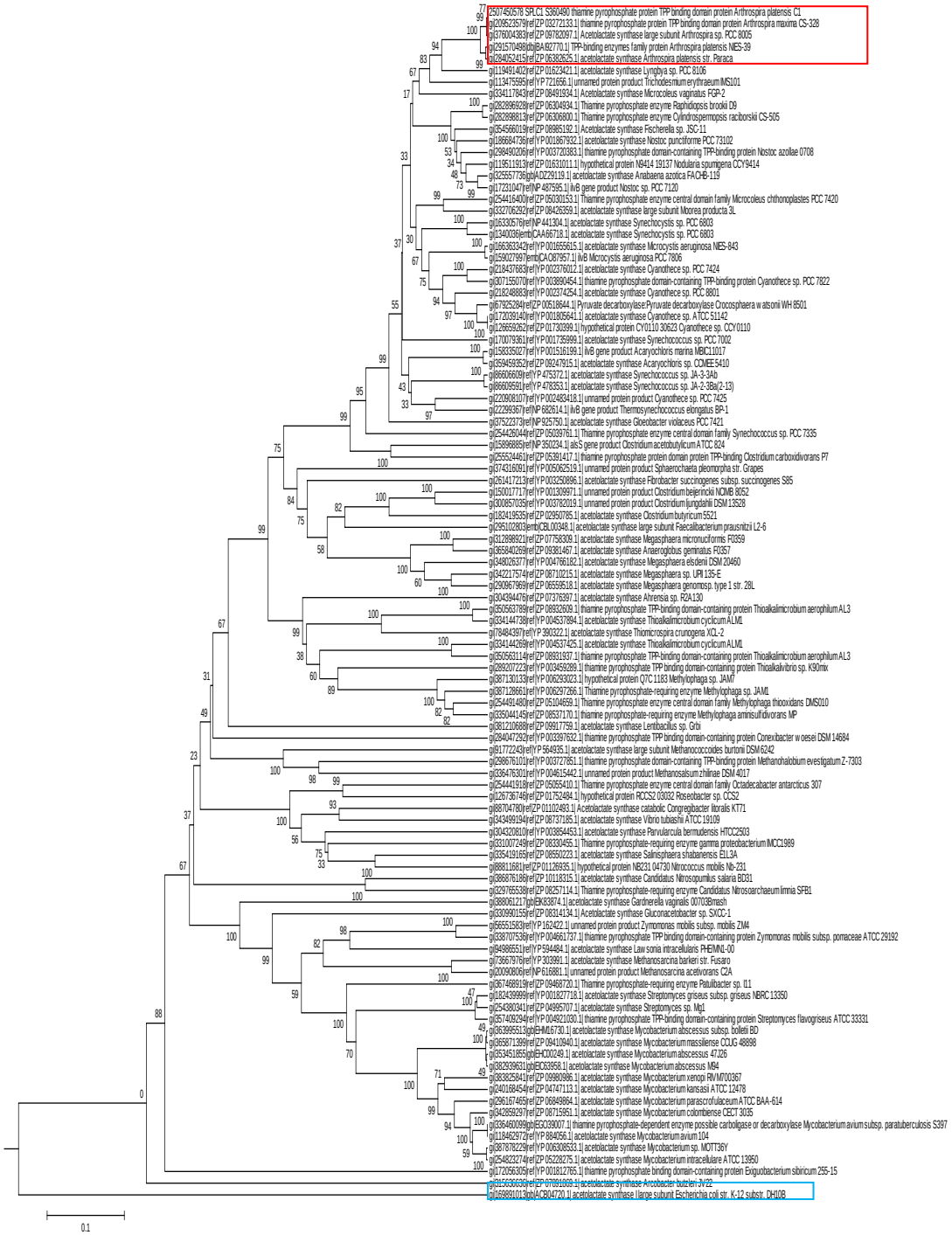
3.5 Voorbeeld van BioLayout clustering

3.3 Onderzoek naar een alternatieve TCA-cyclus pathway

De eiwitsequenties van Acetolactate synthase enSucciante-semialdehyde dehydrogenase die de alternatieve TCA-cyclus vormen werden verkregen via de NCBI databank. Om na te gaan of deze twee enzymen voorkomen in het genoom van *Arthrospira* gebruikten we BLASTp. Uit deze BLAST-resultaten kan worden geconcludeerd dat zowel *Arthrospira* sp. PCC8005, *A. maxima* CS-328*, A. platensis* C1, als *A. platensis* NIES-39 deze genen bevatten (bijlage 1). Dit wijst er op dat de TCA-cyclus kan doorgaan dankzij de alternatieve pathway.

Om dit verder te onderzoeken werd van elk enzym een fylogenetische boom opgesteld met de 100 beste resultaten van een BLASTp tegenover de algemene proteïnedatabank van NCBI. In deze 100 "best-hits" werden meer dan 30 cyanobacteriën gevonden en verder verscheidene andere soorten bacteriën. De fylogenetische bomen werden opgesteld met het UPGMA algoritme en met bootstrapping. Bij bootstrapping wordt met de sequenties meerdere malen in willekeurige volgorde een boom opgesteld en gecontroleerd of de onderlinge relaties veranderen. Als de bootstrapping aantoont dat de onderlinge relaties niet veranderen kan aangenomen worden dat de methode waarmee de boom werd opgesteld goed werd uitgevoerd. Als outlier werd *Escherichia coli* K12 gebruikt. (figuren 3.6 en 3.7).

Alle *Arthrospira* sequenties komen duidelijk samen voor in dezelfde vertakking ("clade") zodat men mag verwachten dat de alternatieve TCA-cyclus in de onderzochte *Arthrospira* stammen op een vrij identische manier verloopt.



3.6 Acetolactate synthase fylogenetische boom

Rood : Arthrospira genomen; Blauw : Outgroup E. Coli K12

3.7 Succinic semialdehyde dehydrogenase fylogenetische boom

Rood : *Arthrospira* genomen; Blauw : Outgroup E. Coli K12

4 Discussie

4.1 DNA-niveau

Een vergelijkende studie van microbiële genomen beoogt gemeenschappelijke kenmerken vast te leggen alsook de verschillen aan te tonen, om dan eventueel deze verschillen verder te onderzoeken. Hierbij speelt de visualisatie een grote rol en werden enkele programma's gekozen (BRIG, Circoletto) die voor deze studie het meest interessant leek. Dankzij deze visualisatie en de vergelijking van *Arthrospira* genomen kwamen elementen voor verder onderzoek aan het licht. Zo werd een regio van 45 kb met veelvuldige recombinatie waargenomen en bleek uit de studie duidelijk dat het genoom van *Arthrospira* zeer dynamisch is en veel repetitieve elementen bevat. Door in te zoomen in dergelijke regio's bvb. met MaGe, het annotatie platform van GénoScope (www.cns.fr/agc/mage) kan men de mogelijke functies van de genen in deze regio's achterhalen, en kan men een idee krijgen over de manipuleerbaarheid van het *Arthrospira* genoom. Dit is van belang bij de ontwikkeling van een genetisch systeem, waarbij *Arthrospira* met extern DNA wordt getransformeerd om zodoende nieuwe genen te introduceren of bestaande genen door homologe recombinatie uit te wisselen met mutant genen. Bovendien is een hoog dynamisch genoom een eerste indicatie voor zgn. 'micro-evolutie' waarbij het genoom onder evolutionaire druk veranderingen ondergaat in het raam van een snelle cellulaire adaptatie t.g.v. van plotse veranderingen in het leefmilieu. De hoge sequentie redundantie (gen duplicatie, paralogie) en het potentieel van een frequente interne herschikking in het *Arthrospira* genoom zou kunnen verklaren waarom dit organisme in staat is extreme condities zoals ioniserende straling te weerstaan. Dit is namelijk ook het geval voor *Deinococcus radiodurans*, een grondbacterie die hoge acute dosissen van gamma straling weerstaat door meerdere kopijen van belangrijke genen in zijn genoom als "back-up" te gebruiken (de functie van een defect gen wordt dan onmiddellijk overgenomen door een paraloog) en dat in staat is zijn eigen genoom, indien vernietigd door de hoge straling, volledig opnieuw op te bouwen d.m.v. een complexe reeks recombinaties.

4.2 Eiwit-niveau

De *Arthrospira* genomen waar mee gewerkt werd hebben een zeer hoge gen paralogie, welke zich uit in sequentie redundantie, en zijn bovendien hoog aan elkaar verwant en coderen voor tal van eiwitten met eenzelfde functioneel domein. Zelfs de uitsluiting van de 530 kleinere PCC 8005 scaffolds, die al dan niet technische of natuurlijke duplicaties bevatten, bleek niet te helpen. De vergelijkende studie op proteoom niveau werd door deze hoge redundantie sterk bemoeilijkt. De gebruikte software had hierdoor zodanig veel berekeningen uit te voeren dat dit leidde tot problemen met tijdelijke computer bestanden die in het RAM geheugen weggeschreven worden, tot op het punt dat het programma vastliep of dezelfde berekeningen eindeloos bleef herhalen.

Clustering van de proteïnen zou veel informatie kunnen opleveren over de onderlinge relaties tussen de genomen en over elk genoom afzonderlijk. De identificatie van zowel de gemeenschappelijke als de unieke genen (en de eiwitten die zij coderen) voor elk van de genomen is hoogst interessant voor verder onderzoek omdat elk van de onderzochte stammen bepaalde eigenschappen heeft. Zo produceert *A. maxima* CS-328 veel meer waterstof dan de andere drie stammen, wat interessant is op vlak van bio-energie, en de *A. platensis* C1 stam is niet motiel (geen voortbeweging op agar, wat handig is voor de ontwikkeling van een genetisch systeem in respect tot cellulaire uitgroei in colonies) maar de andere stammen wel, en is PCC 8005 de enige *Arthrospira* stam met een volledig operon *nthRAB* voor de afbraak van nitriet, een organische verbinding met een -C≡N functionele groep en belangrijk in de aanmaak van acrylaten en rubbers.

4.3 TCA-cyclus

Dankzij het onderzoek van Zhang en Bryant (Zhang & Bryant, 2011) is aangetoond dat de TCA-cyclus wel kan doorgaan in cyanobacteriën dankzij een alternatief reactiepad. Door aan te tonen dat deze genen voorkomen in de door ons bestudeerde genomen is het duidelijk dat ook in *Arthrospira* deze "bypass" wordt gebruikt. Een fylogenetische studie toont ook aan dat de genen voor deze alternatieve enzymen bijna in alle cyanobacteriële genomen voorkomen. Dit kan interessant zijn in het bestuderen van de werking van *Arthrospira* omdat de TCA-cyclus een cruciale rol speelt in veel metabolische processen. De ontdekking van een alternatieve route voor de TCA cyclus heeft ook directe implicaties in de ontwikkeling van nieuwe reactiepaden bvb. voor de productie van biobrandstoffen en bioplastics.

5 Besluit

Deze stage werd opgedeeld in drie luiken. Eerst werd aan de hand van diverse softwares (BLASTp, BRIG, Circoletto) de genomen van vier *Arthrospira* stammen met elkaar vergeleken. De visualisatie hiervan is erg belangrijk in het raam van de incomplete *Arthrospira* genomen (alle *Arthrospira* genomen hebben klaarblijkelijk te kampen met dezelfde problemen van hoge sequentie redundantie) en de mogelijke voortzetting van de uiteindelijke assemblage van het PCC 8005 genoom en de andere genomen. Vervolgens werd op eiwit niveau getracht de gemeenschappelijke en unieke eiwitten op te sporen (GeneRage, Ortho-MCL) maar jammer genoeg faalde de hardware met name het werkgeheugen bleek onvoldoende en er waren ook problemen met de connectie tussen Ortho-MCL en de mySQL databank. Een derde en laatste deel bestond uit een fylogenetische studie van twee enzymes in een alternative TCA cyclus, kenmerkend voor cyanobacteriën en aanwezig in alle vier de onderzochte *Arthrospira* stammen. Deze studie was een boeiende eerste kennismaking met de organisatie van een genoom project, de moeilijkheden die voorkomen bvb. wat de assemblage van een genoom betreft, of de hardware matige tekortkomingen die aan het licht kwamen.

Literatuurlijst

A.J., Enright, & C.A., Ouzounis (2000). GeneRAGE : a robust algorithm for sequence clustering and domain detection. *BIOONFORMATICS* , 451-457.

Alikhan, N.-F., Petty, N. K., Zakour, N. L., & Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple. *BMC Genomics* , 12:402.

Camacho, C., G, C., V, A., N, M., J, P., K, B., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* , 10:421.

Darling, A., Mau, B., & Perna, N. (2010). progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS One* , 5(6):e11147.

Darzentas, N. (2010). Circoletto: visualizing sequence similarity with Circos. *Bioinformatics* , 2620–2621.

Enright, A., Van Dongen, S., & Ouzounis, C. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* , 30(7):1575-1584.

P.J.Janssen, N.Morin, & M.Mergeay. (2010). Genome Sequence of the Edible Cyanobacterium Arthrospira. *Journal of Bacteriology* , 2465-2466.

SCK•CEN Website. (sd). Opgeroepen op Maart - Mei 2012, van http://www.sckcen.be/

Tamura K, P. D. (2011). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution* , 28: 2731-2739.

Zhang, S., & Bryant, D. A. (2011). The Tricarboxylic Acid Cycle in Cyanobacteria. *Science* , 334:1551.

Bijlagen

Bijlage 1 – BLAST-resultaten Acetolactate synthase Succinic semialdehyde dehydrogenase

Query= SYNPCC7002_A2771 succinate-semialdehyde dehydrogenase
Length=454

> 2507449251 SPLC1_S203470 aldehyde Dehydrogenase [Arthrospira platensis C1]
Length=459

Score = 132 bits (333), Expect = 4e-32, Method: Compositional matrix adjust.
Identities = 92/340 (28%), Positives = 152/340 (45%), Gaps = 11/340 (3%)

```
Query  114  SYVCYQPLGILLAVMPWNFPFWQVFRFAAPALMAGNVAVLKHASNVPQCALAVEAILEAA  173
            SY+  +PLG++L + PWN+PF  V      A+ AGN A+LK +      +  V  ++
Sbjct  99   SYIYPEPLGVVLIIGPWNYPFQLVISPLVGAIAAGNCAILKPSELAVHTSQVVADLISKT  158

Query  174  GFPEGVFQTLLIGASQVEQVIKDPRVKAATLTGSEPAGASLASLAGQEIKPTLLELGGSD  233
              P   +   + G  +V Q + D        TG +  G  + + A + + P  LELGG
Sbjct  159  FSPN--YIATVQGGVEVSQQLLDEPFDHIFFTGGKRIGKIVMTAAAKHLTPVTLELGGKS  216

Query  234  PFVVFPSADLDEAVEVGTVARTMNNGQSCIAAKRFILHEAIAAEFLEKLHLKFASLKIGD  293
            P +V     LD   +     + +N GQ+CIA    ++   I ++ +  +     S    +
Sbjct  217  PCIVDADTQLDYTAKRIVWGKFINAGQTCIAPDYLLVDRRIKSDLITAMIGCIESFYGSE  276

Query  294  PMAPETDIGPLATEGILQDISRQVDQAVAAGAKILLGGRPLDRAGYFYPPTILTEIPPGA  353
            P     D G +        ++  +        KI+ GG+  D +  +  PT++ E+
Sbjct  277  PQQ-SPDYGRIINHYHFHRLTELIHDG-----KIVAGGK-FDESDRYISPTLIDEVSWED  329

Query  354  KILQEELFAPVAMVFTVKDLDQAIALANDIPFGLGASAWTNDPAEQQRFIQELDAGAVFI  413
             I+Q+E+F P+  + T  DL  AI+  N  P  L    ++ +   QQR +QE  +G +
Sbjct  330  PIMQDEIFGPILPILTYNDLGDAISQINARPKPLALYLFSRNKQSQQRVLQETSSGGICF  389

Query  414  NG--MVKSDPRLPFGGTKRSGYGRELGLAGIRTFVNAKTV  451
            N   M      LPFGG   SG G+  G A   TF + K++
Sbjct  390  NETIMQVGGQSLPFGGVGESGIGKYHGQATFDTFCHYKSL  429
```

>dbj|BAI92769.1| putative succinate-semialdehyde dehydrogenase [Arthrospira platensis NIES-39]
Length=455

Score = 615 bits (1586), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 301/455 (66%), Positives = 355/455 (78%), Gaps = 1/455 (0%)

```
Query  1    MAIATINPTTGEICQRFKALTPAEIDAKLAKAQEAFQAYRRTSFSQRRQWLENAAAILER  60
            M IATINP TGE+ + F++LT +EI+A L  A++AF  YR   F ++ +W+  AA ILER
Sbjct  1    MGIATINPATGEVVKTFESLTDSEIEACLESAEQAFSRYRYLPFEKKAEWMNRAAEILER  60

Query  61   DTSKFAEIMTTEMGKTHQSAIAEAEKSALVCRYYAEHGEQFLANEYTETQATESYVCYQP  120
            D  +FA+IMT EMGK  + AIAEA+KSALVCR+YAE+  QFLA+    + AT S+V YQP
Sbjct  61   DAQRFAQIMTLEMGKPIKDAIAEAKKSALVCRFYAENAPQFLADVPAASDATRSFVRYQP  120

Query  121  LGILLAVMPWNFPFWQVFRFAAPALMAGNVAVLKHASNVPQCALAVEAILEAAGFPEGVF  180
            LG +LAVMPWNFPFWQVFRFAAPALMAGNV +LKHASNVPQCALA+  I  AAGFP GVF
Sbjct  121  LGAILAVMPWNFPFWQVFRFAAPALMAGNVGLLKHASNVPQCALAIAEIFTAAGFPPGVF  180

Query  181  QTLLIGASQVEQVIKDPRVKAATLTGSEPAGASLASLAGQEIKPTLLELGGSDPFVVFPS  240
            QTLL+GA +V Q++ DPRVKAATLTGSEPAGASLAS AG+ +K T+LELGGSDPF+V  S
Sbjct  181  QTLLVGADKVAQIVADPRVKAATLTGSEPAGASLASTAGKNLKKTVLELGGSDPFIVLES  240

Query  241  ADLDEAVEVGTVARTMNNGQSCIAAKRFILHEAIAAEFLEKLHLKFASLKIGDPMAPETD  300
            AD++ AV+    AR +NNGQSCIAAKRFIL  +IA EF +KL  KF +LK+GDPM P+T+
Sbjct  241  ADIEAAVQTAVTARMLNNGQSCIAAKRFILASSIAGEFEKKLVAKFKALKVGDPMLPDTE  300

Query  301  IGPLATEGILQDISRQVDQAVAAGAKILLGGRPL-DRAGYFYPPTILTEIPPGAKILQEE  359
            +GPLAT GIL+DI  QV   +AAGAK L GG  L DR G FY PTIL  IPPG    QEE
Sbjct  301  VGPLATPGILKDIDEQVQTCLAAGAKALTGGHRLSDRPGNFYAPTILASIPPGTPADQEE  360

Query  360  LFAPVAMVFTVKDLDQAIALANDIPFGLGASAWTNDPAEQQRFIQELDAGAVFINGMVKS  419
             F PVA++F VK++D+AIALAN   FGLGASAWT D AEQ R   EL+AGAVFING+VKS
Sbjct  361  FFGPVALLFRVKNIDEAIALANSTSFGLGASAWTTDTAEQDRLTLELEAGAVFINGLVKS  420

Query  420  DPRLPFGGTKRSGYGRELGLAGIRTFVNAKTVWLK  454
            DPRLPFGG KRSGYGREL + GI  FVN KTVW+K
Sbjct  421  DPRLPFGGIKRSGYGRELSIQGIHEFVNVKTVWIK  455
```

>ref|ZP_03272132.1| Aldehyde Dehydrogenase [Arthrospira maxima CS-328]
ref|ZP_09782098.1| putative Succinate-semialdehyde dehydrogenase (NAD(P)(+)), GabD-like [Arthrospira sp. PCC 8005]
gb|EDZ96284.1| Aldehyde Dehydrogenase [Arthrospira maxima CS-328]
emb|CCE17851.1| putative Succinate-semialdehyde dehydrogenase (NAD(P)(+)), GabD-like [Arthrospira sp. PCC 8005]
Length=455

Score = 609 bits (1570), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 300/455 (66%), Positives = 353/455 (78%), Gaps = 1/455 (0%)

```
Query  1    MAIATINPTTGEICQRFKALTPAEIDAKLAKAQEAFQAYRRTSFSQRRQWLENAAAILER  60
            M IATINP TGE+ + F++LT +EI+A L  A++AF  YR   F ++ +W+  AA ILER
Sbjct  1    MGIATINPATGEVVKTFESLTDSEIEACLESAEQAFTRYRYLPFDKKAEWMNRAAEILER  60

Query  61   DTSKFAEIMTTEMGKTHQSAIAEAEKSALVCRYYAEHGEQFLANEYTETQATESYVCYQP  120
            D  +FA+IMT EMGK  + AIAEA+KSALVCR+Y E+  QFLA+    + AT S+V YQP
Sbjct  61   DAQRFAQIMTLEMGKPIKDAIAEAKKSALVCRFYGENAPQFLADVPAASDATRSFVRYQP  120

Query  121  LGILLAVMPWNFPFWQVFRFAAPALMAGNVAVLKHASNVPQCALAVEAILEAAGFPEGVF  180
            LG +LAVMPWNFPFWQVFRFAAPALMAGNV +LKHASNVPQCALA+  I +AAGFP GVF
Sbjct  121  LGAILAVMPWNFPFWQVFRFAAPALMAGNVGLLKHASNVPQCALAIAEIFKAAGFPPGVF  180

Query  181  QTLLIGASQVEQVIKDPRVKAATLTGSEPAGASLASLAGQEIKPTLLELGGSDPFVVFPS  240
            QTLL+GA +V Q++ DPRVKAATLTGSEPAGASLAS AG+ +K T+LELGGSDPF+V  S
Sbjct  181  QTLLVGADKVAQIVADPRVKAATLTGSEPAGASLASEAGKNLKKTVLELGGSDPFIVLES  240

Query  241  ADLDEAVEVGTVARTMNNGQSCIAAKRFILHEAIAAEFLEKLHLKFASLKIGDPMAPETD  300
            ADL+ AV+    AR +NNGQSCIAAKRFIL + IA EF +KL  KF +LK+GDPM P+T+
Sbjct  241  ADLEAAVQTAVTARMLNNGQSCIAAKRFILADRIAGEFEKKLVAKFKALKVGDPMLPDTE  300

Query  301  IGPLATEGILQDISRQVDQAVAAGAKILLGGRPL-DRAGYFYPPTILTEIPPGAKILQEE  359
            +GPLAT GIL+DI  QV   +AAGA  L GG  L +R G FY PTIL  IPPG    QEE
Sbjct  301  VGPLATPGILKDIDEQVQTCLAAGAIALTGGHRLSERPGNFYAPTILASIPPGTPADQEE  360

Query  360  LFAPVAMVFTVKDLDQAIALANDIPFGLGASAWTNDPAEQQRFIQELDAGAVFINGMVKS  419
             F PVA++F VK LD+AIALAN   FGLGASAWT D AEQ R   EL+AGAVFING+VKS
Sbjct  361  FFGPVALLFRVKSLDEAIALANSTSFGLGASAWTTDTAEQDRLTLELEAGAVFINGLVKS  420

Query  420  DPRLPFGGTKRSGYGRELGLAGIRTFVNAKTVWLK  454
            DPRLPFGG KRSGYGREL + GI  FVN KTVW+K
Sbjct  421  DPRLPFGGIKRSGYGRELSIQGIHEFVNVKTVWIK  455
```

Score = 609 bits (1570), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 300/455 (66%), Positives = 353/455 (78%), Gaps = 1/455 (0%)

```
Query  1    MAIATINPTTGEICQRFKALTPAEIDAKLAKAQEAFQAYRRTSFSQRRQWLENAAAILER  60
            M IATINP TGE+ + F++LT +EI+A L  A++AF  YR   F ++ +W+  AA ILER
Sbjct  1    MGIATINPATGEVVKTFESLTDSEIEACLESAEQAFTRYRYLPFDKKAEWMNRAAEILER  60

Query  61   DTSKFAEIMTTEMGKTHQSAIAEAEKSALVCRYYAEHGEQFLANEYTETQATESYVCYQP  120
            D  +FA+IMT EMGK  + AIAEA+KSALVCR+Y E+  QFLA+    + AT S+V YQP
Sbjct  61   DAQRFAQIMTLEMGKPIKDAIAEAKKSALVCRFYGENAPQFLADVPAASDATRSFVRYQP  120

Query  121  LGILLAVMPWNFPFWQVFRFAAPALMAGNVAVLKHASNVPQCALAVEAILEAAGFPEGVF  180
            LG +LAVMPWNFPFWQVFRFAAPALMAGNV +LKHASNVPQCALA+  I +AAGFP GVF
Sbjct  121  LGAILAVMPWNFPFWQVFRFAAPALMAGNVGLLKHASNVPQCALAIAEIFKAAGFPPGVF  180

Query  181  QTLLIGASQVEQVIKDPRVKAATLTGSEPAGASLASLAGQEIKPTLLELGGSDPFVVFPS  240
            QTLL+GA +V Q++ DPRVKAATLTGSEPAGASLAS AG+ +K T+LELGGSDPF+V  S
Sbjct  181  QTLLVGADKVAQIVADPRVKAATLTGSEPAGASLASEAGKNLKKTVLELGGSDPFIVLES  240

Query  241  ADLDEAVEVGTVARTMNNGQSCIAAKRFILHEAIAAEFLEKLHLKFASLKIGDPMAPETD  300
            ADL+ AV+    AR +NNGQSCIAAKRFIL + IA EF +KL  KF +LK+GDPM P+T+
Sbjct  241  ADLEAAVQTAVTARMLNNGQSCIAAKRFILADRIAGEFEKKLVAKFKALKVGDPMLPDTE  300

Query  301  IGPLATEGILQDISRQVDQAVAAGAKILLGGRPL-DRAGYFYPPTILTEIPPGAKILQEE  359
            +GPLAT GIL+DI  QV   +AAGA  L GG  L +R G  FY PTIL  IPPG    QEE
Sbjct  301  VGPLATPGILKDIDEQVQTCLAAGAIALTGGHRLSERPGNFYAPTILASIPPGTPADQEE  360

Query  360  LFAPVAMVFTVKDLDQAIALANDIPFGLGASAWTNDPAEQQRFIQELDAGAVFINGMVKS  419
             F PVA++F VK LD+AIALAN   FGLGASAWT D AEQ R   EL+AGAVFING+VKS
Sbjct  361  FFGPVALLFRVKSLDEAIALANSTSFGLGASAWTTDTAEQDRLTLELEAGAVFINGLVKS  420

Query  420  DPRLPFGGTKRSGYGRELGLAGIRTFVNAKTVWLK  454
            DPRLPFGG KRSGYGREL + GI  FVN KTVW+K
Sbjct  421  DPRLPFGGIKRSGYGRELSIQGIHEFVNVKTVWIK  455
```

Query= SYNPCC7002_A2770 acetolactate synthase
Length=545
> 2507450578 SPLC1_S360490 thiamine pyrophosphate protein TPP binding
domain protein [Arthrospira platensis C1]
Length=545
Score = 929 bits (2401), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 429/543 (80%), Positives = 489/543 (91%), Gaps = 0/543 (0%)

```
Query  1    MNTAELLIRCLENEGVEYIFGLPGEENLHILEALKESPIRFITVRHEQGAAFMADVYGRL  60
            MNTAELL++CLENEGV+Y+FGLPGEEN+ +LE+LK+S I+FIT RHEQGAAFMADVYGRL
Sbjct  1    MNTAELLVKCLENEGVKYVFGLPGEENMEVLESLKDSSIQFITTRHEQGAAFMADVYGRL  60

Query  61   TGKAGVCLSTLGPGATNLMTGVADANLDGAPLIAITGQVGTDRMHIESHQYLDLVAMFAP  120
            TGKAGVCLSTLGPGATNLMTGVADANLDGAPL+AITGQVGTDRMHIESHQYLDLVAMF+P
Sbjct  61   TGKAGVCLSTLGPGATNLMTGVADANLDGAPLVAITGQVGTDRMHIESHQYLDLVAMFSP  120

Query  121  VTKWNKQIVRPNTTPEVVRRAFKIAQQEKPGAVHIDLPENIAAMPVEGQPLQRDGREKIY  180
            VTKWN QIVRP+ TPE+VR+AFK+AQ EKPGAVHIDLPENIA+M VEGQPL RD REKIY
Sbjct  121  VTKWNAQIVRPSITPEIVRKAFKLAQTEKPGAVHIDLPENIASMAVEGQPLNRDRREKIY  180

Query  181  ASSRSLNRAAEAIAHAKSPLILVGNGIIRADAAEALTDFATQLNIPVVNTFMGKGAIPYT  240
             + +S+N AA AI+ A +PLILVGNG +RA+A+EA+T+FATQLNIPV NTFMGKGAIPYT
Sbjct  181  CAYQSMNEAASAISKAVNPLILVGNGALRANASEAVTEFATQLNIPVANTFMGKGAIPYT  240

Query  241  HPLSLWTVGLQQRDFVTCAFEQSDLVIAVGYDLIEYSPKRWNPEGTTPIIHIGEVAAEID  300
            HPLSLWT GLQ RDF++CAF+++DLVIA+GYDLIEYSPK+WNP+GT PIIH+G  AAEID
Sbjct  241  HPLSLWTTGLQLRDFISCAFDKADLVIAIGYDLIEYSPKKWNPKGTIPIIHVGANAAEID  300

Query  301  SSYIPLTEVVGDIGDALNEIRKRTDREGKTAPKFLNVRAEIREDYERHGTDASFPVKPQK  360
            SSYIP+ E+VGDI D+L+EI +R+DR GK  P  + +R+EIR+DYER+  D  FP+KPQK
Sbjct  301  SSYIPIAEIVGDISDSLHEILRRSDRTGKPDPYGVKLRSEIRQDYERYANDDGFPIKPQK  360

Query  361  IIYDLRQVMAPEDIVISDVGAHKMWMARHYHCDRPNTCLISNGFAAMGIAIPGAVAAKLV  420
            IIYDLRQVM PEDIVISDVGAHKMWMARHYHCDRPNTCLISNGFAAMGIAIPGA+AAK V
Sbjct  361  IIYDLRQVMGPEDIVISDVGAHKMWMARHYHCDRPNTCLISNGFAAMGIAIPGAIAAKFV  420

Query  421  YPEKNVVAVTGDGGFMMNCQELETALRIGANFVTLIFNDGGYGLIGWKQINQFGAPAFVE  480
             PE  VVAVTGDGGFMMNCQELETALR+G  FVTLIFNDGGYGLI WKQ +QFG+ +F++
Sbjct  421  NPELKVVAVTGDGGFMMNCQELETALRVGTPFVTLIFNDGGYGLIEWKQEDQFGSSSFIK  480

Query  481  FGNPDFVQFAESMGLKGYRITAAADLVPTLKEALAQDVPAVIDCPVDYSENVKFSQKSGD  540
            FGNPDFV+FAESMGLKGYR+ AA DLVP LKEALA +VP VIDCPVDY EN +FSQ++G
Sbjct  481  FGNPDFVKFAESMGLKGYRVQAATDLVPILKEALASEVPVVIDCPVDYRENARFSQRAGG  540

Query  541  LIC  543
            L C
Sbjct  541  LCC  543
```

>TPP-binding enzymes family protein [Arthrospira platensis NIES-39]
Length=545

Score = 931 bits (2405), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 429/543 (79%), Positives = 491/543 (90%), Gaps = 0/543 (0%)

```
Query  1    MNTAELLIRCLENEGVEYIFGLPGEENLHILEALKESPIRFITVRHEQGAAFMADVYGRL  60
            MNTAELL++CLENEGV+Y+FGLPGEEN+ +LE+LK+S I+FIT RHEQGAAFMADVYGRL
Sbjct  1    MNTAELLVKCLENEGVKYVFGLPGEENMEVLESLKDSSIQFITTRHEQGAAFMADVYGRL  60

Query  61   TGKAGVCLSTLGPGATNLMTGVADANLDGAPLIAITGQVGTDRMHIESHQYLDLVAMFAP  120
            TGKAGVCLSTLGPGATNLMTGVADANLDGAPL+AITGQVGTDRMHIESHQYLDLVAMF+P
Sbjct  61   TGKAGVCLSTLGPGATNLMTGVADANLDGAPLVAITGQVGTDRMHIESHQYLDLVAMFSP  120

Query  121  VTKWNKQIVRPNTTPEVVRRAFKIAQQEKPGAVHIDLPENIAAMPVEGQPLQRDGREKIY  180
            VTKWN QIVRP+ TPE+VR+AFK+AQ EKPGAVHIDLPENIA+M VEGQPL RD REKIY
Sbjct  121  VTKWNAQIVRPSITPEIVRKAFKLAQTEKPGAVHIDLPENIASMAVEGQPLNRDRREKIY  180

Query  181  ASSRSLNRAAEAIAHAKSPLILVGNGIIRADAAEALTDFATQLNIPVVNTFMGKGAIPYT  240
            ++ +S+N AA AI+ A +PLILVGNG +RA+A+EA+T+FATQLNIPV NTFMGKGAIPYT
Sbjct  181  SAYQSMNEAASAISKAVNPLILVGNGALRANASEAVTEFATQLNIPVANTFMGKGAIPYT  240

Query  241  HPLSLWTVGLQQRDFVTCAFEQSDLVIAVGYDLIEYSPKRWNPEGTTPIIHIGEVAAEID  300
            HPLSLWT GLQQRDF++CAF+++DLVIA+GYDLIEYSPK+WNP+GT PIIH+G  AAEID
Sbjct  241  HPLSLWTTGLQQRDFISCAFDKADLVIAIGYDLIEYSPKKWNPKGTIPIIHVGANAAEID  300

Query  301  SSYIPLTEVVGDIGDALNEIRKRTDREGKTAPKFLNVRAEIREDYERHGTDASFPVKPQK  360
            SSYIP+ E+VGDI D+LNEI +R+DR GK  P  + +R++IR+DYER+  D  FP+KPQK
Sbjct  301  SSYIPIAEIVGDISDSLNEILRRSDRTGKLDPYGVKLRSDIRQDYERYANDDGFPIKPQK  360

Query  361  IIYDLRQVMAPEDIVISDVGAHKMWMARHYHCDRPNTCLISNGFAAMGIAIPGAVAAKLV  420
            IIYDLRQVM PEDIVISDVGAHKMWMARHYH DRPNTCLISNGFAAMGI+IPGA+AAKLV
Sbjct  361  IIYDLRQVMGPEDIVISDVGAHKMWMARHYHSDRPNTCLISNGFAAMGISIPGAIAAKLV  420

Query  421  YPEKNVVAVTGDGGFMMNCQELETALRIGANFVTLIFNDGGYGLIGWKQINQFGAPAFVE  480
             PE  VVAVTGDGGFMMNCQELETALR+G  FVTLIFNDGGYGLI WKQ +QFG+ +F++
Sbjct  421  NPELKVVAVTGDGGFMMNCQELETALRVGTPFVTLIFNDGGYGLIEWKQEDQFGSSSFIK  480

Query  481  FGNPDFVQFAESMGLKGYRITAAADLVPTLKEALAQDVPAVIDCPVDYSENVKFSQKSGD  540
            FGNPDFV+FAESMGLKGYR+ AA DLVP LKEALA +VP VIDCPVDY EN +FSQ++G
Sbjct  481  FGNPDFVKFAESMGLKGYRVEAATDLVPILKEALASEVPVVIDCPVDYRENARFSQRAGG  540

Query  541  LIC  543
            L C
Sbjct  541  LCC  543
```

> gb|EDZ96285.1| thiamine pyrophosphate protein TPP binding domain protein [Arthrospira maxima CS-328]

Length=545
Score = 929 bits (2401), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 429/543 (79%), Positives = 489/543 (90%), Gaps = 0/543 (0%)

```
Query  1    MNTAELLIRCLENEGVEYIFGLPGEENLHILEALKESPIRFITVRHEQGAAFMADVYGRL  60
            MNTAELL++CLENEGV+Y+FGLPGEEN+ +LE+LK+S I+FIT RHEQGAAFMADVYGRL
Sbjct  1    MNTAELLVKCLENEGVKYVFGLPGEENMEVLESLKDSSIQFITTRHEQGAAFMADVYGRL  60

Query  61   TGKAGVCLSTLGPGATNLMTGVADANLDGAPLIAITGQVGTDRMHIESHQYLDLVAMFAP  120
            TGKAGVCLSTLGPGATNLMTGVADANLDGAPL+AITGQVGTDRMHIESHQYLDLVAMF+P
Sbjct  61   TGKAGVCLSTLGPGATNLMTGVADANLDGAPLVAITGQVGTDRMHIESHQYLDLVAMFSP  120

Query  121  VTKWNKQIVRPNTTPEVVRRAFKIAQQEKPGAVHIDLPENIAAMPVEGQPLQRDGREKIY  180
            VTKWN QIVRP+ TPE+VR+AFK+AQ EKPGAVHIDLPENIA+M VEGQPL RD REKIY
Sbjct  121  VTKWNAQIVRPSITPEIVRKAFKLAQTEKPGAVHIDLPENIASMAVEGQPLNRDRREKIY  180

Query  181  ASSRSLNRAAEAIAHAKSPLILVGNGIIRADAAEALTDFATQLNIPVVNTFMGKGAIPYT  240
             + +S+N AA AI+ A +PLILVGNG +RA+A+EA+T+FATQLNIPV NTFMGKGAIPYT
Sbjct  181  CAYQSMNEAASAISKAVNPLILVGNGALRANASEAVTEFATQLNIPVANTFMGKGAIPYT  240

Query  241  HPLSLWTVGLQQRDFVTCAFEQSDLVIAVGYDLIEYSPKRWNPEGTTPIIHIGEVAAEID  300
            HPLSLWT GLQ RDF++CAF+++DLVIA+GYDLIEYSPK+WNP+GT PIIH+G  AAEID
Sbjct  241  HPLSLWTTGLQLRDFISCAFDKADLVIAIGYDLIEYSPKKWNPKGTIPIIHVGANAAEID  300

Query  301  SSYIPLTEVVGDIGDALNEIRKRTDREGKTAPKFLNVRAEIREDYERHGTDASFPVKPQK  360
            SSYIP+ E+VGDI D+L+EI +R+DR GK  P  + +R+EIR+DYER+  D  FP+KPQK
Sbjct  301  SSYIPIAEIVGDISDSLHEILRRSDRTGKPDPYGVKLRSEIRQDYERYANDDGFPIKPQK  360

Query  361  IIYDLRQVMAPEDIVISDVGAHKMWMARHYHCDRPNTCLISNGFAAMGIAIPGAVAAKLV  420
            IIYDLRQVM PEDIVISDVGAHKMWMARHYHCDRPNTCLISNGFAAMGIAIPGA+AAK V
Sbjct  361  IIYDLRQVMGPEDIVISDVGAHKMWMARHYHCDRPNTCLISNGFAAMGIAIPGAIAAKFV  420

Query  421  YPEKNVVAVTGDGGFMMNCQELETALRIGANFVTLIFNDGGYGLIGWKQINQFGAPAFVE  480
             PE  VVAVTGDGGFMMNCQELETALR+G  FVTLIFNDGGYGLI WKQ +QFG+ +F++
Sbjct  421  NPELKVVAVTGDGGFMMNCQELETALRVGTPFVTLIFNDGGYGLIEWKQEDQFGSSSFIK  480

Query  481  FGNPDFVQFAESMGLKGYRITAAADLVPTLKEALAQDVPAVIDCPVDYSENVKFSQKSGD  540
            FGNPDFV+FAESMGLKGYR+ AA DLVP LKEALA +VP VIDCPVDY EN +FSQ++G
Sbjct  481  FGNPDFVKFAESMGLKGYRVQAATDLVPILKEALASEVPVVIDCPVDYRENARFSQRAGG  540

Query  541  LIC  543
            L C
Sbjct  541  LCC  543
```

>Acetolactate synthase large subunit [Arthrospira sp. PCC 8005]
Length=545
Score = 932 bits (2409), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 430/543 (79%), Positives = 490/543 (90%), Gaps = 0/543 (0%)

```
Query  1    MNTAELLIRCLENEGVEYIFGLPGEENLHILEALKESPIRFITVRHEQGAAFMADVYGRL  60
            MNTAELL++CLENEGV+Y+FGLPGEEN+ +LE+LK+S I+FIT RHEQGAAFMADVYGRL
Sbjct  1    MNTAELLVKCLENEGVKYVFGLPGEENMEVLESLKDSSIQFITTRHEQGAAFMADVYGRL  60

Query  61   TGKAGVCLSTLGPGATNLMTGVADANLDGAPLIAITGQVGTDRMHIESHQYLDLVAMFAP  120
            TGKAGVCLSTLGPGATNLMTGVADANLDGAPL+AITGQVGTDRMHIESHQYLDLVAMF+P
Sbjct  61   TGKAGVCLSTLGPGATNLMTGVADANLDGAPLVAITGQVGTDRMHIESHQYLDLVAMFSP  120

Query  121  VTKWNKQIVRPNTTPEVVRRAFKIAQQEKPGAVHIDLPENIAAMPVEGQPLQRDGREKIY  180
            VTKWN QIVRP+ TPE+VR+AFK+AQ EKPGAVHIDLPENIA+M VEGQPL RD REKIY
Sbjct  121  VTKWNAQIVRPSITPEIVRKAFKLAQTEKPGAVHIDLPENIASMAVEGQPLNRDRREKIY  180

Query  181  ASSRSLNRAAEAIAHAKSPLILVGNGIIRADAAEALTDFATQLNIPVVNTFMGKGAIPYT  240
             + +S+N AA AI+ A +PLILVGNG +RA+A+EA+T+FATQLNIPV NTFMGKGAIPYT
Sbjct  181  CAYQSMNEAASAISKAVNPLILVGNGALRANASEAVTEFATQLNIPVANTFMGKGAIPYT  240

Query  241  HPLSLWTVGLQQRDFVTCAFEQSDLVIAVGYDLIEYSPKRWNPEGTTPIIHIGEVAAEID  300
            HPLSLWT GLQQRDF++CAF+++DLVIA+GYDLIEYSPK+WNP+GT PIIH+G  AAEID
Sbjct  241  HPLSLWTTGLQQRDFISCAFDKADLVIAIGYDLIEYSPKKWNPKGTIPIIHVGANAAEID  300

Query  301  SSYIPLTEVVGDIGDALNEIRKRTDREGKTAPKFLNVRAEIREDYERHGTDASFPVKPQK  360
            SSYIP+ E+VGDI D+L+EI +R+DR GK  P  + +R+EIR+DYER+  D  FP+KPQK
Sbjct  301  SSYIPIAEIVGDISDSLHEILRRSDRTGKPDPYGVKLRSEIRQDYERYANDDGFPIKPQK  360

Query  361  IIYDLRQVMAPEDIVISDVGAHKMWMARHYHCDRPNTCLISNGFAAMGIAIPGAVAAKLV  420
            IIYDLRQVM PEDIVISDVGAHKMWMARHYHCDRPNTCLISNGFAAMGIAIPGA+AAK V
Sbjct  361  IIYDLRQVMGPEDIVISDVGAHKMWMARHYHCDRPNTCLISNGFAAMGIAIPGAIAAKFV  420

Query  421  YPEKNVVAVTGDGGFMMNCQELETALRIGANFVTLIFNDGGYGLIGWKQINQFGAPAFVE  480
             PE  VVAVTGDGGFMMNCQELETALR+G  FVTLIFNDGGYGLI WKQ +QFG+ +F++
Sbjct  421  NPELKVVAVTGDGGFMMNCQELETALRVGTPFVTLIFNDGGYGLIEWKQEDQFGSSSFIK  480

Query  481  FGNPDFVQFAESMGLKGYRITAAADLVPTLKEALAQDVPAVIDCPVDYSENVKFSQKSGD  540
            FGNPDFV+FAESMGLKGYR+ AA DLVP LKEALA +VP VIDCPVDY EN +FSQ++G
Sbjct  481  FGNPDFVKFAESMGLKGYRVQAATDLVPILKEALASEVPVVIDCPVDYRENARFSQRAGG  540

Query  541  LIC  543
            L C
Sbjct  541  LCC  543
```