HEINRICH HEINE
UNIVERSITÄT
DÜSSELDORF

A Metagenomic Approach towards Novel LOV domain containing Blue-Light Photoreceptors

Inaugural-Dissertation

zur

Erlangung des Doktorgrades der

Mathematisch-Naturwissenschaftlichen Fakultät

der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Gopal Prasad Pathak

aus Syangja, Nepal

May 2010

Angefertigt im Max Planck Institut für Bioinorganishce Chemie
(Performed at Max Planck Institute for Bioinorganic Chemistry)

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. Wolfgang Gärtner ***(first referee)***
Koreferent: Prof. Dr. Karl-Erich Jäger ***(second referee)***

Tag der mündlichen Prüfung:01.06.2010 ***(date of the oral examination)***

Acknowledgement

Firstly, I would like to thank Prof. Dr. Wolfgang Gärtner for providing me with an opportunity to conduct this challenging work. His trust, continuous support, guidance and productive discussions played key roles for the successful completion of this multidisciplinary work. Furthermore, I am grateful to him for the critical readings and corrections of the manuscript in spite of his busy schedules, which led this dissertation to come out in this form. I would like to thank Prof. Dr. Karl-Erich Jäger (Institute for Molecular Enzyme Technology, University of Düsseldorf, Jülich) for kindly agreeing to be the second referee of this dissertation.

My sincere thanks go to Prof. Dr. Wolfgang Streit (University of Hamburg, Department of Microbiology and Biotechnology) for valuable discussions, suggestions and providing the metagenomic samples. I would like to thank Dr. Aba Losi (University of Parma, Italy) for helpful discussions. Many thanks to Dr. Armin Ehrenreich (Technical University of Munich) for providing spotted microarray slides as per my requirement and fruitful discussions related to DNA microarray techniques. I am thankful to all of them for the great and successful cooperation.

Many thanks also go to Dr. Virginia Albaracin and Dr. M.E. Farias (Universidad Nacional de Tucumán/PROIMI-CONICET, Argentina) for providing the strains from HAALs and for a great cooperation. I extend my sincere thanks to Dr. Nicole Tandeau de Marsac (Pasteur Institute, France) for providing the DNA from some cyanobacterial strains. Similarly, I would like to thank Dr. Gabriela Schaule (IWW Water Centre) for giving me a chance to use the fluorescence plate reader in her lab. I would like to thank Dr. Susannah Green Tringe (Department of Energy, Joint Genome Institute, USA) for providing the selected clones from Whale fall and Farm soil metagenome.

I would like to thank Dr. Hideaki Ogata in our institute for his great support for crystallization of HT-Met1 LOV protein. I would like to thank Dr. Dagmar Krysciak, Patrick Bijtenhoorn and Ulrich Köhler in University of Hamburg (Department of Microbiology and Biotechnology) for their help in Hamburg for preparation of

metagenomic samples. Many thanks to Dr. Christian Elend (Technical University of Hamburg-Harburg) for his help and valuable suggestions related to microarray-system set up at the beginning.

A good atmosphere is always necessary for a fresh mind and hence a fruitful research output. Many thanks are given to all friends and colleagues in our laboratory for the friendly communications, advices and good working atmosphere. I would like to thank Helene Steffen for her great help by making laboratory material and basic support available whenever required. Gülümse Koc-Weier and Dr. Markus Knipp were always helpful and gave me access to using their laboratory facility whenever I required. Many thanks go to colleagues Alessandra Hoppe, Dr. Amrit Kaur, Björn Zorn, Jana Riethausen, Dr. Koji Nakayama, Dr. Madina Mansurova, Sarah Raffelberg, Sebastian Gandor, Shivani Sharda, Simone Ringsdorf, Yifen Tang and Zhen Cao. It was really a pleasant environment in the lab due to their cooperative and friendly attitude.

My especial thanks goes to my parents back home for their faith, support and hard work to make me able for this achievement. Finally, my most special thanks goes to my beloved son Garbit, who was born when I was on the way of this manuscript preparation and my wife Rashmi. Garbit always showed his patience, kept smiling and never became a reason for interruption of our regular work. Rashmi not only provided moral and usual support but also helped in some laboratory work when required as she was also working in our group.

Publications and manuscripts

Pathak, G.P., Ehrenreich, A., Losi, A., Streit, W., and Gärtner, W. (2009), Novel Blue Light-Sensitive Proteins from a Metagenomic Approach. *Env. Microbiol. 11* (9), 2388-2399

Pathak, G.P. and Gärtner, W., (2010), Detection and isolation of selected genes of interest from metagenomic libraries by a DNA microarray approach.In: Methods in Molecular Biology, Streit, W.R.; Daniel, R (ed.) (in press).

Pathak, G.P., Losi, A. and Gärtner, W., LOV is global: Metagenome-based screening reveals worldwide distribution of LOV-domain proteins (Manuscript submitted).

Summary

LOV domain-containing blue light (BL) receptor proteins, which were initially identified from *Arabidopsis thaliana*, have since then been reported from a significant and rapidly growing number of prokaryotes. This was made possible by a genome-level investigation of culturable microorganisms sequenced so far. Recently, a new field of molecular biology called “metagenomics” emerged after the realization that more than 99% of all microbes are inaccessible by the cultivation techniques available currently. This hidden prokaryotic diversity has been regarded as a potential reservoir for the novel genes of biotechnological importance.

In the work presented here, a metagenomic approach was applied to characterize LOV domain-containing BL receptor proteins. For this approach, a DNA microarray was developed that carried 149 54-mer oligonucleotides designed on the basis of consensus sequences of BL-photoreceptor encoding genes. Elaborate calibration experiments were performed with genomic, plasmid, and cosmid DNA, and even with entire gene libraries to determine the efficiency, selectivity and sensitivity of the array. The array was then used to screen the cosmid-cloned DNA from a library containing about 2000 clones which was constructed from the thermophilic fraction of a soil sample. A novel gene (*ht-met*1) could be isolated from a 40-kb cosmid clone insert, which encoded a protein consisting of several PAS domains, a LOV domain associated to a histidine kinase and a response regulator domain. From all reported LOV-domain proteins, the novel gene showed highest similarity to a known sequence from *Kineococcus radiotolerans* SRS30216 (58 % identity for the LOV domain only) and to a gene from *Methylibium petroleiphilum* PM1 (57 % identity for the LOV domain only). The LOV-domain region of *ht-met1* was subcloned and expressed yielding a fully functional, flavin containing LOV domain that showed typical LOV-photochemistry with a fast dark recovery. Aside from HT-Met1 LOV, further two partial LOV domains were isolated from an Elbe River metagenome using the newly developed array. This approach was extended for the screening of environmental strains from high altitude Andean lakes in Argentina and cyanobacterial strains from a desert environment. Also in this material we could detect LOV domains in several of the strains.

Summary

A further sequence-based investigation of LOV domain proteins from various environmental sample datasets was performed. A total of 231 LOV domains could be assigned via the highly conserved core region of this protein domain from already available metagenomic datasets. The majority of the putative LOV domains were originated from the Global Ocean Sampling expedition, but LOV domains were also found in metagenomes from extreme environments such as hypersaline ponds, acidic mine drainage (pH 0.8, 42 ^{o}C) or wastewater treatment facilities. Positively identified samples were found in material originating from 1.46 ^{o}C to 42 ^{o}C (excluding the simulated thermophilic samples). A proportion of ca. 73% of these identified gene fragments (169 out of 231) showed a sequence divergence of more than 20% from data base-deposited LOV-domain sequences originated from previously known species, despite the fact that they could clearly be identified as LOV domains. Sequence novelty was also observed in a subfamily approach determined by a maximum likelihood phylogenetic tree. More interestingly, a recently released phage metagenome contains LOV domain sequences, probably being acquired as an event of horizontal gene transfer. The virtually ubiquitous presence of LOV domains in metagenomes obtained worldwide from a wide range of environmental conditions emphasizes the important physiological role of these photoreceptors in light-induced signal transduction, stress adaptation and related survival mechanisms.

Zusammenfassung

LOV-Domänen enthaltende, Blaulicht- (BL-) sensitive Proteine waren erstmals in *Arabidopsis thaliana* identifiziert worden (Phototropine). Seitdem wurden diese Proteine in steigender Menge in einer Vielzahl prokaryotischer Genome entdeckt. Dies gelang bisher auf der Basis Genom-sequenzierter, kultivierbarer Mikroorganismen. Seit einiger Zeit findet ein neuer Ansatz, genannt „Metagenomics“ Anwendung, nachdem akzeptiert wurde, dass mehr als 99% aller Mikroorganismen auf der Basis gegenwärtiger Kultivationstechniken nicht zugänglich sein würden. Diese bisher verborgene prokaryotische Diversität wird zunehmend als Reservoir für neuartige Gene mit biotechnologischer Bedeutung betrachtet.

In der hier präsentierten Arbeit wird ein derartiger metagenomischer Ansatz für die Identifizierung LOV-Domänen enthaltender Proteine angewendet. Für diesen Zugang wurde ein micro-array entwickelt, der 149 Oligonukleotide mit einer Länge von jeweils 54 Basen enthielt. Die Sequenzen ergaben sich aus Consensus-Motiven von BL-Photorezeptoren. Ausgiebige und detaillierte Kalibrierungen des Array wurden unter Verwendung von genomischer, plasmid- und cosmid-basierter DNA und gene-libraries im Vorfeld durchgeführt, um die Effizienz, Selektivität und Sensitivität zu bestimmen. Der micro-array wurde dann verwendet, um cosmid-klonierte DNA aus einer library von ca. 2000 Klonen zu „screenen“, die aus einer thermophilen Fraktion einer Bodenprobe erhalten worden war. Es gelang, ein nauartiges Gen aus einem 40 kb-Insert zu identifizieren und zu isolieren, das für ein Multi-Domänen-Protein mit mehreren PAS- und einer LOV-Domäne, einer Histidin-Kinase und einem anfusionierten „response regulator“ bestand. Im Vergleich aller bisher bekannten LOV Domänen wies diese neuartige Domäne die größte Sequenzähnlichkeit zu zwei Proteinen auf: Zum einen zu einer LOV-Domäne aus *Kineococcus radiotolerans* SRS30216 (58 %) und zum anderen zu einer entsprechenden Domäne (57 %) aus *Methylibium petroleiphilum* PM1. Die LOV Domäne aus HT-Met1 wurde subkloniert und heterolog exprimiert. Wir fanden eine LOV-typische Absorption und Photochemie mit einer für prokaryotische LOV-Domänen kurzen Rückkehr-Kinetik. Zusätzlich

konnten in weiteren Ansätzen zwei partielle LOV-Domänen Sequenzen identifiziert werden. Dieser Zugang wurde ausserdem erweitert auf Material erhalten aus hochgelegenen Seen in den Argentinischen Anden und auf Cyanobakterien, die aus einer Wüstenregion isoliert wurden. Auch in diesem Material konnten LOV Domänen-kodierende Sequenzen isoliert werden.

Zusätzlich wurde eine weitere Sequenz-basierte Methode auf der Basis vorhandener Metagenomics-Datenbanken durchgeführt, die durch Datensammlung aus verschiedenen Habitaten entstanden sind. Durch diese Suche wurden unter Verwendung des konservierten Bereichs 231 neuartige LOV Domänen identifiziert. Die Mehrheit dieser Sequenzen entstammte Materials der „Global Ocean Sampling Expedition“, allerdings wurden LOV Domänen aus Metagenomen gefunden, die aus extremen Umgebungen stammten, z.B., aus hypersalinen Seen, sauren Abwässern von Minen (pH 0.8, 42 °C) oder Wasseraufbereitungs-Anlagen stammten. Eindeutig identifizierte Proben stammten aus Temperaturbereichen zwischen 1.46 und 42 °C, wobei rein thermophile Habiltate ausgeschlossen waren. Ein Anteil von 73 % (169 von 231) der gefundenen Sequenzen zeigte eine Sequenz-Abweichung von mehr als 20 %, unabhängig von der Tatsache, dass sie eindeutig als LOV Domänen identifiziert wurden. Die Neuartigkeit der gefundenen Sequenzen ergab sich auch aus eine „maximum lokelihood phylogenetic tree“. Sogar in einem kürzlich präsentierten Cyano-Phagen Metagenome wurden Sequenzen für LOV Domänen nachgewiesen, dies möglicherweise als Effekt eines horizontalen Gentransfers.

Die praktisch ubiquitäre Verbreitung der LOV Domänen in diesen Metagenomen, die weltweit aus einem breiten Bereich verschiedener Umweltproben stammen, unterstreicht die bedeutende physiologische Rolle dieser Photorezeptoren in der Licht-induzierten Signaltransduktion für eine Adaptierung an verschiedene Stress-Faktoren und vergleichbare Überlebensmechanismen.

.

Table of Contents

Abbreviations

The abbreviations repeatedly used throughout this dissertation are summarized below. SI units have been excluded from the lists. DNA bases were abbreviated using their one letter codes (A, C, G and T) and amino acids were abbreviated using the common one- or three letter codes.

aa	Amino acid(s)
ALOHA	A Long-Term Oligotrophic Habitat Assessment station
AMD	Acid Mine Drainage
Amp	Ampicillin
AP	Alkaline phosphatase
BAC	Bacterial artificial chromosome
BL	Blue light
BLUF	Sensor of Blue-Light Using FAD domain
bp	Base pairs
BSA	Bovine serum albumin
CAMERA	Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis
cDNA	Complementary DNA
cGMP	Cyclic guanosine mono phosphate
Cm	Chloramphenicol
CPD	Cyclobutane pyrimidine dimer
cry	Cryptochrome
CTAB	Hexadecyl-trimethyl-ammonium bromide
C-terminal	Carboxy-terminal
Cy3-dCTP	5-Amino-propargyl-2'-deoxycytidine 5'-triphosphate coupled to Cy3 fluorescent dye
Cy5-dCTP	5-Amino-propargyl-2'-deoxycytidine 5'-triphosphate coupled to Cy5 fluorescent dye
dATP	2'-deoxyadenosine 5'-triphosphate
dCTP	2'-deoxycytidine 5'-triphosphate
dGTP	2'-deoxyguanosine 5'-triphosphate
Dig or DIG	Digoxigenine
DIG-11-UTP	Digoxigenine-11-2'-deoxy-uridine-5'-triphosphate
DMF	Dimethyl formamide
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleoside triphosphate
DTT	Dithiothreitol
dTTP	2'-deoxythymidine 5'-triphosphate
EAL	Putative diguanylate phosphodiesterase

EBPR	Enhanced Biological Phosphorus Removal
E-cup	Eppendorf cups (1.5 or 2 ml plastic tubes)
EDTA	Ethylene diamine tetra-acetic acid
FAD	Flavin adenine dinucleotide
FGA	Functional gene array
FMN	Flavin mononucleotide
GAF	cGMP-specific phosphodiesterases/Adenylcyclase/FhlA
GGDEF	Diguanylate cyclase, domain with conserved GGDEF motif present in a variety of bacteria
GMP	Guanosine mono phosphate
GOS	Global Ocean Sampling
His tag	Histidine residues identification tag
HK	Histidine kinase domain
HTH	Helix turn helix
IDA	Imino diacetic acid
IMAC	Immobilized metal affinity chromatography
IMG	Integrated Microbial Genomes
IPTG	Isopropyl--D-thiogalactopyranoside
Km	Kanamycin
kb	Kilobase
kDa	Kilo dalton
LB	Luria Bertani broth
LOV	Light, Oxygen, Voltage domain
M	Molar (mol/l)
MCS	Multiple cloning site
mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information
NBT/BCIP	nitro blue tetrazolium/5-Bromo-4-chloro-3-indolyl phosphate, toluidine salt
NMR	Nuclear magnetic resonance
NPH1	Non phototrophic hypocotyl 1
NPSG	North Pacific Subtropical Gyre
N-terminal	Amino-terminal
OD	Optical density
ORF	Open reading frame
PAC	PAS associated /C-terminal PAS domain
PAGE	Polyacrylamide gel electrophoresis
PAS	Per-Arnt-Sim domain
PCC	Pasteur Culture Collection
PCR	Polymerase chain reaction
Pefabloc	4-(2-aminoethyl)-benzene-sulfonylfluoride hydrochloride
PEG	Polyethylene glycol
phot	Phototropin

Phy	Phytochrome
PVDF	Polyvinylidene difluoride
PVP	Polyvinylpyrrolidone
PYP	Photoactive yellow protein
RNA	Ribonucleic acid
RNase	Ribonuclease
ROS	Reactive oxygen species
rpm	Rotations per minute
RR	Response regulator
rRNA	Ribosomal ribonucleic acid
RT	Room temperature
SAP	Shrimp alkaline phosphatase
SDS	Sodium dodecylsulfate
SI	Signal intensity
SNR	Signal to noise ratio
SSC	Standard-saline-citrate
STAS	Sulfate transporter anti-sigma factor antagonist domain
TAE	Tris- acetate /EDTA
TBS	Tris buffer saline
TBY	Terrific broth/yeast extract excess
td	Tri-distilled
TE	Tris /EDTA
TIGR	The Institute for Genomic Research
Tris	2-amino-2-(hydroxymethyl)-1,3-propanediol
Triton X-100	t-octylphenoxypolyethoxyethanol
tween	Poly-oxyethylene-sorbitan
UV	Ultra violet
V	Volt
v/v	Volume per volume
VIS	Visible spectrum
w/v	Weight per volume
X-Gal	5-Brom-3-Chlor-3-Inoyl-β-D-galactopyranoside

Chapter 1

Introduction

In general, the term “light” denotes the wavelength range of the solar radiation that is visible to us, probably including spectral regions at both edges of the spectrum, e.g., the ultraviolet (UV) and the far-red regions that are not visible for us but for other organisms. Light is a nearly ubiquitous environmental signal which regulates developmental and behavioral responses in plant, fungal, bacterial, and animal cells. Be it for a unicellular or multicellular organism, light is important for various activities. From a daily schedule of a person to the flying of a bird, from localization of molecules inside the cells to trigger numerous biological responses up to its utilization as the universal energy source, light has direct or indirect influences. In other words, light is one of the most important elements in the living world that plays a crucial role for survival and maintenance of the organisms. According to Presti and Delbrück [1] living organisms utilize light as a source of energy and as a means of obtaining information about their environment.

Only a small part of the overall radiation emitted by the sun reaches the earth. Any wavelengths shorter than 295 nm are filtered out by the ozone layer in the stratosphere so the shortest wavelength of sunlight reaching the surface of the earth is about 295 nm. It is the visible radiation, with wavelengths from 380 nm to 780 nm, which is important for vision and many other photobiological responses (Figure 1.1).

Certain bacteria, algae and higher plants can utilize the light energy directly for the synthesis of essential components by photosynthesis. On the other hand, animals depend on these “synthesizers” as they can not utilize the light for the synthesis of biomolecules directly. Not all light radiations are acceptable to the organisms, e.g., ultraviolet radiation, including wavelengths shorter than 400 nm, controls some photobiological reactions and many important photochemical reactions, but also has detrimental effects. Non-ionizing radiation produces excited states in molecules due to

the absorption of one or more photons. Excited-state molecules are highly reactive and can react with adjacent molecules, but quite often they undergo photochemical and photophysical changes within their own molecular structure [2]. In order to avoid the detrimental effects caused by light radiations, certain organisms have developed mechanisms for sensing and eventually avoiding such radiations.

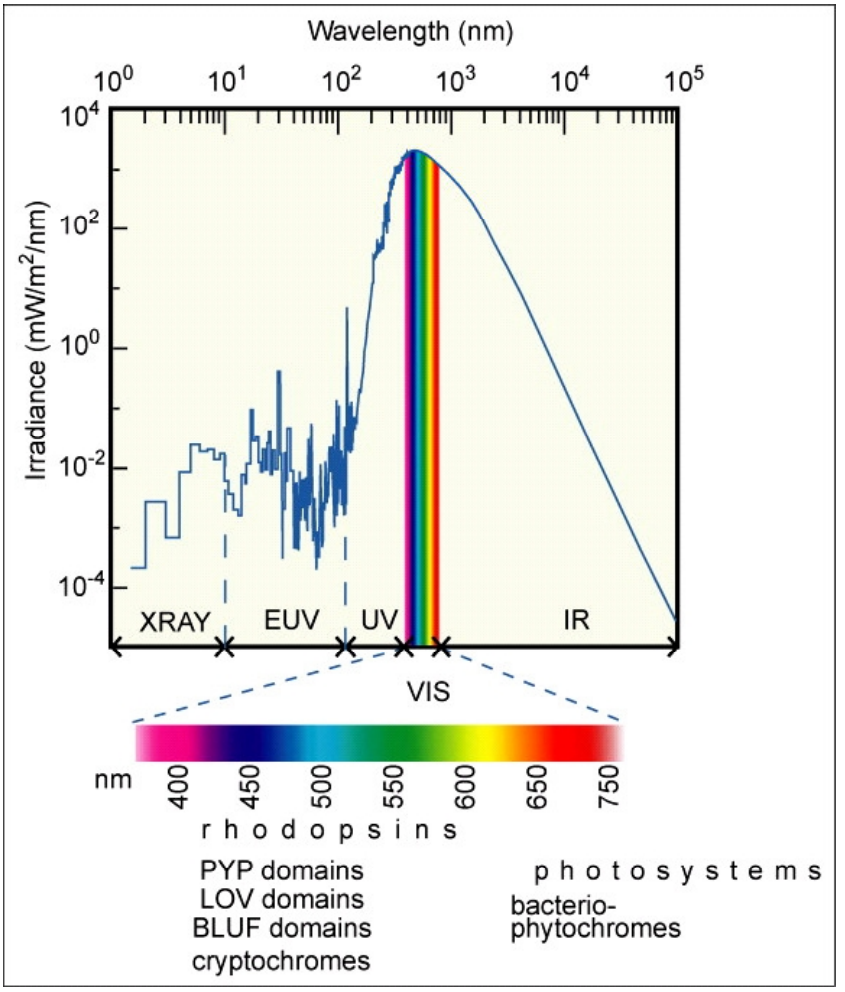


Figure 1.1 Solar radiation and the light visible to the human eye. Photoreceptor proteins from different organisms utilize only the small portion of the spectrum from ultraviolet to the far red [3]. EUV= extreme ultraviolet; UV= ultraviolet; VIS= visual; IR= infra red.

It depends on the environment what wavelengths range organisms are expected to deal with [4], since different environmental niches are inhabitated by various forms of life and light radiation is not equally distributed in all environments. It is important for the organisms to adjust their metabolism to perceive the useful radiation. On the other hand, it is equally important to avoid harmful radiations for optimal cellular functioning and survival. In either way, the organisms are dependent on receptor substances ("photoreceptors") to detect light. The role of these photoreceptors is directly associated with the survival and maintenance of the organisms in their particular habitat.

1.1 Biological photoreceptors

Light absorption by photoreceptor proteins is the first condition for the photo-response by any living cell. The ability of photoreceptor proteins to absorb visible light is, at the molecular level, directly correlated to the type of chromophore (from the Greek word *chromos* = color) that is bound within the respective photoreceptor protein.

Though there exist several photoreceptor protein classes/types, they can be grouped into few families based on their absorption range and chromophore modules. The most rational approach is on the basis of the chemical structures of their chromophores and information based on sequence alignments [5]. Accordingly, the biological photoreceptors can be classified into six major families viz., rhodopsins, the phytochromes, the xanthopsins, the cryptochromes, the phototropins and the BLUF proteins. In the first three families, photon absorption by the chromophore causes a double bond cis-trans isomerization of the chromophores, while in the latter three families there is a transfer of electrons [3].

Rhodopsins contain seven membrane-embedded alpha helices that form an internal pocket in which the retinal chromophore is bound. Functionally, two types of rhodopsins exist: light driven energy pumps and sensory rhodopsins. Bacteriorhodopsin (BR) and halorhodopsin (HR) are light driven energy pumps. Sensory rhodopsins I (SRI), sensory rhodopsin II (SRII) and the visual rhodopsins perform sensory functions. The first four rhodopsins were initially reported from

Halobacterium salinarum, an archaeon while the visual rhodopsins are found in higher eukaryotes including human eyes [6;7]. Visual rhodopsins employ 11-cis isomer of their retinal chromophore whereas the rhodopsins of prokaryotes, unicellular algae and fungi all carry the all-trans isomer of retinal in their resting state irrespective of their function as sensors or energy convertors. Apart from archaea, new members of the rhodopsins were reported from marine proteobacteria [8], cyanobacteria [9], lower eukaryotes like *Chlamydomonas reinhardtii* [10] and *Neurospora* [11]. Other opsins (e.g., melanopsin, parapinopsin) are additional retinal-, extra retinal- and extra-occular photoreceptors that may play important roles in circardian clocks, camouflage, detection of ambient light conditions and seasonal variation in photoperiod [7].

Another important and widely distributed family of photoreceptor proteins is the group of red- and far red light-sensing **phytochromes**. The phytochromes bind an open-chain tetrapyrrole ligand as chromophore and have been reported from higher plants, cyanobacteria, non-photosynthetic eubacteria and also from some fungi [12-14]. In contrast to the rhodopsins, phytochromes in general are soluble proteins. The chromophore-binding part of phytochromes in general is composed of PAS-GAF-Phy domains. The proteobacterial and fungal phytochromes bind the tetrapyrrole biliverdin as chromophore via a thioether-bond to a conserved cysteine residue in a PAS domain. On the other hand, the plant and cyanobacterial phytochromes bind their chromophore (phytochromobilin or phycocyanobilin) via a conserved cysteine residue in the GAF domain [15]. In many flowering plants the phytochromes regulate the time of flowering based on the length of the day and night, they detect the quality of the light spectra (e.g., shade avoidance), and control cellular responses, tropism and circadian rhythms. A functional role of bacterial phytochromes in red/far-red light dependent photomorphogenesis is not yet established but they have been reported to control synthesis of the photosynthetic apparatus [16], the complementary chromatic adaptation [17], and were also reported to be necessary for the growth of *Synechocystis* sp PCC6803 under blue light [18].

The **xanthopsins** are a family of photoreceptors that have been coined according to their first identified representative “photoactive yellow proteins” (PYP).

PYPs use p-coumaric acid as a chromophore [19], which is linked to the sulfur of a cysteine residue by a thioester linkage. PYP proteins were first isolated from *Ectothiorhodospira halophila* [20], an archaeon. PYPs are water soluble proteins and exhibit a light-induced photocycle triggered by the photo-isomerization of its chromophore [19]. Based on their small size (about 125 aa, 14 kDa) and simple photocycle they are regarded as a structural prototype of the Per-Arnt-Sim (PAS) domain superfamily [21]. They are photochemically and spectroscopically well characterized proteins. This class of receptor proteins is thought to be responsible for the negative phototactic response [22]. It is also involved in the expression of chalcone synthase [13], which is a key enzyme for the synthesis of photo-protective pigments. PYP domains have been found to be followed by a bacteriophytochrome in *Rhodospirillum centenum* [13] and *Thermochromatium tepidum* [23].

1.1.1 Flavin containing photoreceptors

The blue region of the spectra (430-500 nm) is the range with highest energy in the visible radiation. Also, this spectral region shows highest capacity to penetrate into the oceanic water reaching up to a depth of 200 m [24]. Organisms inhabiting oceanic water are supposed to be able to “deal with” the blue light as to detect the light source for photosynthetic purposes on one hand and for the avoidance of harmful effect of the “high energy” light on the other hand. It is well documented that the UV radiation is the causative agent of DNA damage mostly by the formation of cyclobutane pyrimidine dimer (CPD) [25]. Similarly, the blue light generates with a high quantum yield the triplet state of porphyrins ubiquitous in all living organisms, in case they are present in their metal-free form, the porphyrin triplets, in turn, react with present oxygen and produce the highly reactive singlet form of oxygen and other reactive oxygen species (ROS) that can be harmful and even deleterious for cells [26]. Hence, sensing the light in the blue region is critical for microorganisms in habitats where blue light is profound. Three major types of flavin-containing photoreceptors, viz. cryptochromes, BLUF domains and LOV domains have been identified in the different forms of life.

The **cryptochromes** were initially identified in plants by Ahmad and Cashmore in 1993 [27]. Since then they have been reported from lower and higher eukaryotes

(mammals, insects, plants and algae) [28;29]. These proteins use FADH as chromophore. Cryptochromes share a high degree of sequential and structural similarity to DNA photolyases, but lack the activity of typical DNA photolyase enzymes; instead, they carry a C-terminal tail for signal transduction [30]. A new family of cryptochrome protein named as CRY-DASH (DASH: *Drosophila*, *Arabidopsis*, *Synechocystis*, *Homo*) was recently reported from some eubacteria (*Synechocystis* sp. PCC6803, *Vibrio cholerae*) [31;32]. CRY-DASH proteins do not possess the C-terminal tail found in typical chryptochromes and, interestingly, show DNA photolyase activity in single stranded DNA. The typical DNA repair activity on CPDs in double stranded DNA has not been demonstrated [30]. The cryptochromes (excluding CRY-DASH) are involved in the functioning of circardian clocks in animals and seed germination, hypocotyl elongation and pigment accumulation in plants.

A comparatively young class of flavin-containing photoreceptors is the **BLUF proteins**. They have been reported from bacteria and lower eukaryotes [29;33] . They were named BLUF according to their function and chromophore: “sensor for blue light sensing using FAD (Flavin Adenine Dinucleotide)”. Members of this family are known to be involved in photophobic responses in *Euglena gracilis* and *Synechocystis* [29] and transcriptional regulation in *Rhodobacter sphaeroides* [34].

Another important family of blue light receptor proteins is the **LOV domain protein** family, first identified as a genetic locus and designated as NPH1 (non phototrophic hypocotyl 1) [35] in *Arabidopsis*. The NPH1 protein was named phototropin 1 (phot1) after its functional role in phototropism [36]. A sequential and biochemical characterization showed that the initially identified phototropin consists of two stretches of PAS domains in its N-terminal part that show light detection capability due to their incorporated flavin chromophore. Such domains are regulated by light, oxygen and voltage, after which the name LOV domain was assigned to the blue light absorbing PAS domain of phot [37]. They bind a single molecule of flavin mononucleotide (FMN) (Figure 1.2) as chromophore.

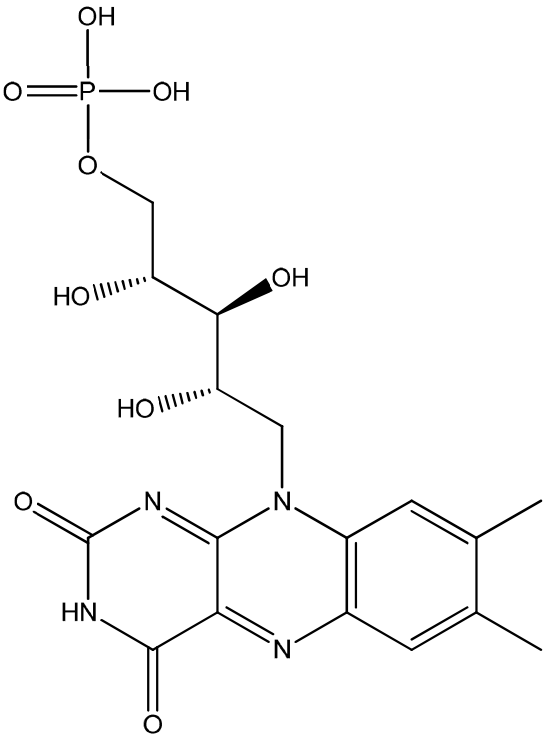


Figure 1.2 A flavin mononucleotide (FMN) molecule. The three-ring system (isoalloxazine) part of FMN plays a key role in protein-FMN interaction.

A highly conserved central motif (GxN**C**RFLQ) with a cysteine residue is responsible for chromophore binding and the photoreaction mechanisms of LOV proteins (Figure 1.3). Christie and colleagues [36] have first expressed the LOV domains from *Avena sativa* (oat) and *Arabidopsis* phototropin heterologously in *E. coli*. They also showed that the LOV domains of phy3, a putative photoreceptor from the fern, *Adiantum capillus*-veneris also binds FMN. Besides plants, the LOV-domain proteins have been reported from *Neurospora crassa,* a fungus [38], the alga *Chlamydomonas reinhardtii* [39] and many bacterial species [15].

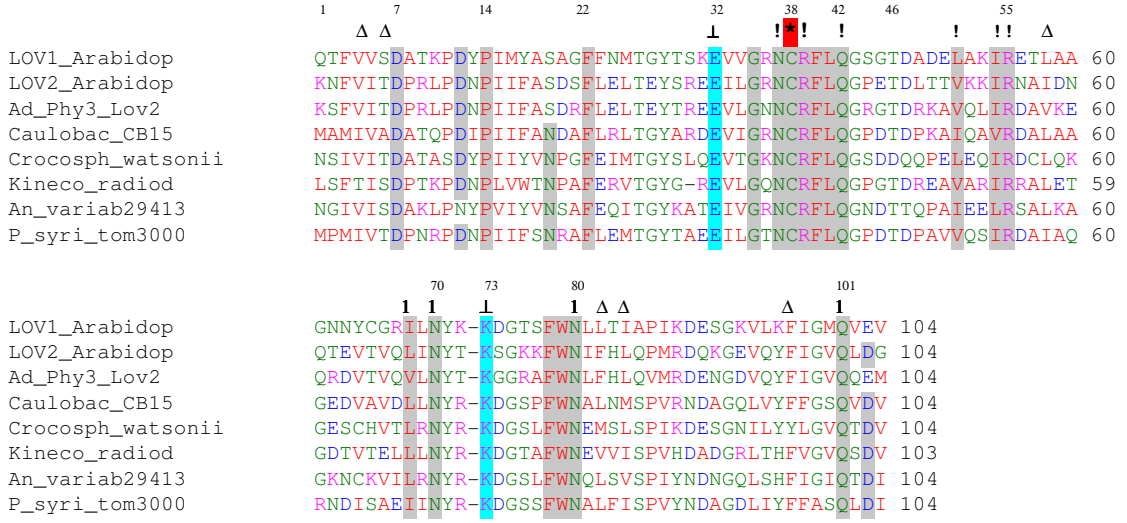


Figure 1.3 Sequence alignment of LOV domains from plants and bacteria displaying the conserved functional residues (shaded). The first two sequences are from Arabidopsis LOV1 and LOV2, the third one from Adiantum LOV2 while the others are from bacteria. LOV1_Arabidop: LOV1 from *Arabidopsis thaliana* phot1; LOV2_Arabidop: LOV2 from *Arabidopsis thaliana* phot1; Ad_Phy3_LOV2: LOV2 from *Adiantum capillus* phy; Caulobac_CB15: LOV from *Caulobacter crescentus*; Crocosph_watsonii: LOV from *Crocosphera watsonii*; Kineco_radiod: LOV from *Kineococcus radiotolerans*; An_variab29413: LOV from *Anabaena variabililis*; P_syri_tom3000: LOV from *Pseudomonas syringae* DC3000.

Symbols:

!= Residues that bind with FMN, Δ = Residues that form hydrophobic pocket around the dimethyl phenyl ring of FMN, 1: Residues responsible for interaction with isoalloxazine ring ⊥= Conserved E and K residues forming a salt bridge at the surface of LOV domain, * = Reactive cysteine.

1.1.1.1 LOV domain modules

The LOV domains from different organisms regulate the activity of numerous output or effector domains such as kinases, phosphodiesterases, zinc fingers and stress sigma factors (Figure 1.4).

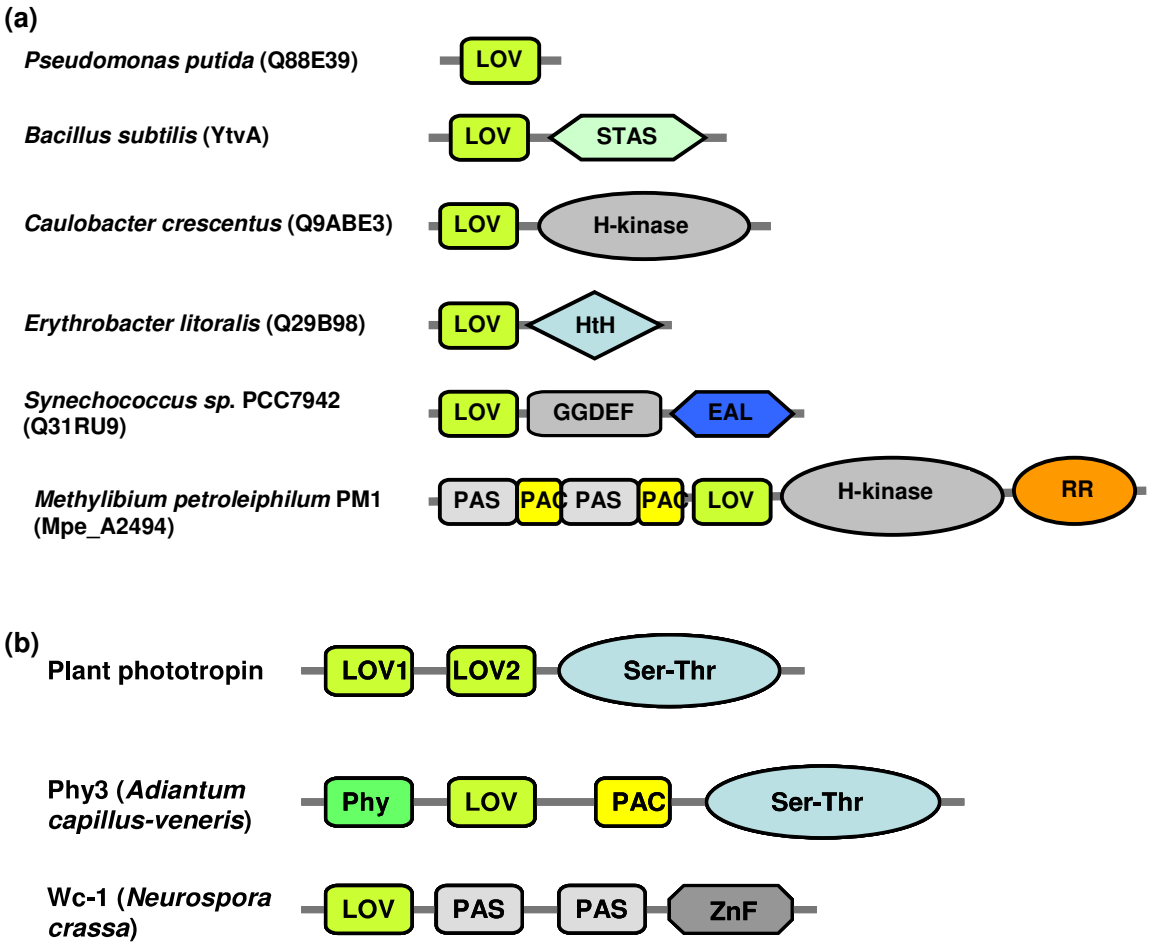


Figure 1.4 Domain architecture of few selected LOV domain-containing proteins. (a) Selected bacterial LOV domain-containing proteins with increasing domain complexity from single domain short LOV from *Pseudomonas putida* to multi domain proteins like that of *Methylibium petroleiphilum* PM1. (b) The eukaryotic LOV domain-containing proteins from plants and fungi. Phototropin proteins from plants contain a tandem array of two LOV domains (LOV1 and LOV2) in a single receptor protein with a C-terminal serine threonine kinase. The Phy3 protein from a fern contains an N-terminal phytochrome domain followed by a LOV domain and a C-terminal serine threonine kinase. (Abbreviations: STAS: sulphate transporter antisigma factor; H-kinase: histidine kinase; HtH: helix turn helix DNA binding motif; GGDEF: named after a conserved sequence motif; this domain is likely to catalyze the synthesis of cyclic diguanylate; EAL: named after the conserved sequence EAL, found in phosphodiesterases with a hydrolysis function of cyclic-di-GMP; PAS: Per, Arnt, Sim domain; PAC: PAS-associated C-terminal motif; RR: response regulator; Ser-Thr: serine threonine kinase; Phy: phytochrome; ZnF: zinc-finger motif)

Among 176 LOV domain proteins from sequenced prokaryotic genomes a wide range of domain composition can be seen. Histidine kinases are found as the output domain in 89 prokaryotic LOV domain-containing proteins while further 35 of such proteins have a dicyclic-GMP-synthesis/degradation motif. Another 18 show a STAS domain and 23 do not have any fused output domain [15]. In the majority of the prokaryotes known so far, the directly interacting partner domains in the LOV photoreceptor proteins are fused on the same protein, while in few cases those receptor proteins may have only a short LOV motif without any further domain. Moreover, unlike in plant phototropins (e.g., *Adiantum* and *Arabidopsis* phototropins) where two LOV domains, LOV1 and LOV2, are present, the presence of more than one LOV domain in a single prokaryotic photoreceptor protein is not reported so far, though multiple PAS domains can be found in the same protein along with a LOV domain. *Adiantum*-phy3 is an unusual protein such that it has a phytochrome-like N-terminal domain and a phototropin like C-terminal domain (Figure 1.3b). Hence, phy3 displays the properties of both a red and a blue light photoreceptor [40].

1.1.1.2 Photo-excitation and structure dynamics

In the dark (or resting) state, LOV domains incorporate non-covalently a single molecule of FMN. In brief, the light absorption causes a transient, fully reversible bleaching of the optical absorption at 447 nm accompanied by an increase of the absorption at 390 nm (LOV_{390}), which represents the lit or signaling state (Figure 1.5). The signaling state (LOV_{390}) formation occurs via the decay of the triplet state (LOV_{660}), typically within 1-2 µs [41]. In molecular terms, the excitation by a blue light photon prompts the formation of a covalent bond between the conserved cysteine (SH group) and atom C (4a) of the FMN ring [42].

The dominant photoproduct constitutes a FMN-cysteinyl adduct [43]. This generated covalent bond of the “photoadduct” (LOV-390) decays and recovers spontaneously the parent state on a time scale of few seconds (ca 50 seconds) in *Avena sativa* [41] to several seconds (ca 200 seconds) in *Chlamydomonas reinhardtii* [44]. The thermal recovery to ground states takes hours in bacterial LOV domains from *B. subtilis* and *P. putida* SB2 [45;46]. Studies have shown that the replacement of the

highly conserved cysteine at the canonical motif by alanine or serine resulted in unbleachable LOV domains that do not form the intermediate LOV-390, but show a high, constitutive fluorescence [43;44].

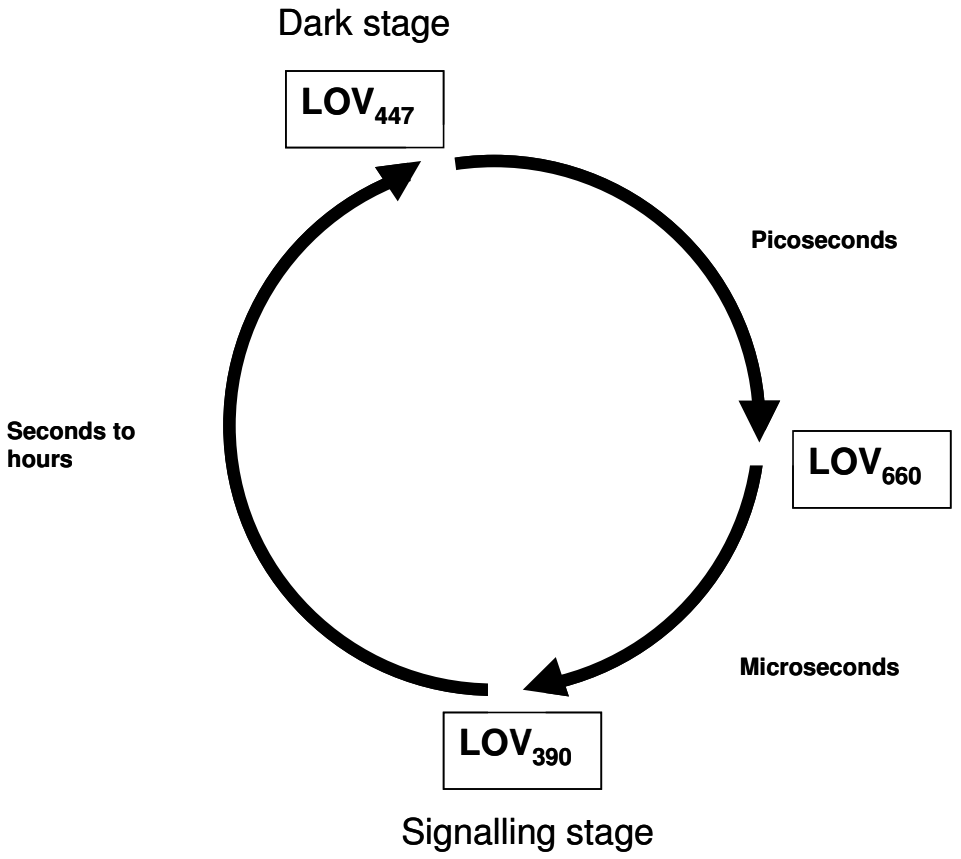


Figure 1.5 Schematic illustration of a LOV photocycle

The X-ray crystallographic studies of the LOV domains [47;48] exhibit a prototypical PAS fold that consists of five anti-parallel β–sheets and four α–helices (Figure 1.6). The conserved cysteine is located at a distance of 4.2 Å from the C-4a of flavin. Two N-terminal sheets βA and βB are linked by helical connectors (αA; αB; α'A and αC) to the βC, βD, and βE sheets. Another recently crystallized YtvA-LOV domain contains an additional connector helix (αJ) at the C-terminal end which was supposed to play an important role in signal transduction [49].

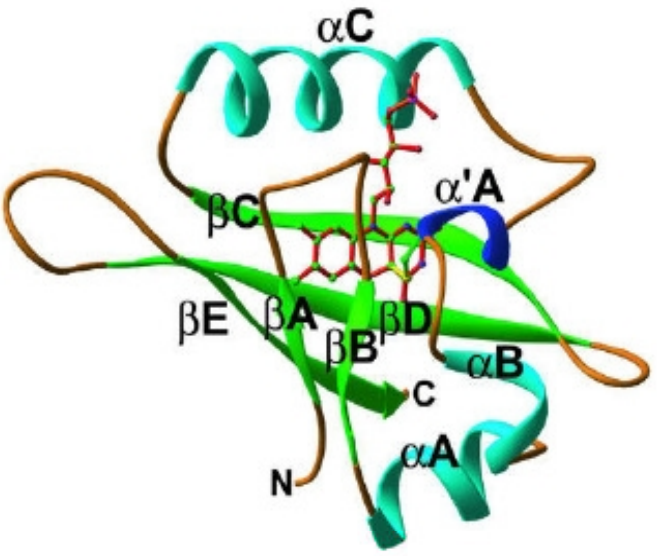


Figure 1.6 Ribbon diagram of the phy3 LOV2 structure under steady state illumination. The FMN cofactor is shown in the center of the fold with the bonds colored red. The sulfur atom of the conserved LOV-Cys residue is attached covalently to carbon 4a of FMN with a yellow adduct bond. Atoms of the Cys side chain and FMN are colored by elements: carbon, green; nitrogen, blue; phosphorus, pink, sulfur, yellow; and oxygen, red. N and C denote the N-terminal and C-terminal position of the domain, respectively [47].

In spite of the structural knowledge, only little information is available on how the signal is transmitted to the activation of kinase or partner output domains. Two mechanisms of LOV2 signal transmission have been proposed. Light-driven destabilization of the salt bridge that is present at the surface of the LOV domain has been hypothesized to play a role in phototropin kinase activation [50]. The second mechanism involves a conserved glutamine residue within the LOV domain that interacts with the FMN chromophore via a hydrogen bonding [51]. This glutamine has recently been shown to be involved in propagating light-induced protein conformational changes associated with LOV2 protein fragments [52;53], indicating that this residue may serve to transmit modifications from within the chromophore binding pocket to

protein changes at the LOV domain surface. In YtvA, this residue has been shown to be important for LOV to STAS signal transmission [54].

In particular, the βE sheet region of *Adiantum* phy3 LOV2 exhibits a significant conformational change upon cysteinyl adduct formation [52]. It contains the aforementioned conserved glutamine residue (Gln1029) that when mutated to leucine results in a loss of these light-driven protein changes [52;53]. Gln1029 forms hydrogen bonds with the FMN chromophore and undergoes side chain rotation upon cysteinyl adduct formation [42;47]. This residue therefore appears to be important for signal transmission from inside the chromophore-binding pocket to protein changes at the LOV2 surface. A possible model that explains how adduct formation can signal through the surrounding LOV domain to result in kinase activation for this process has been provided by solution NMR experiments with phototropin of LOV2, *A. sativa* [55]. The conserved α-helix (designated Jα) is associated with the surface of LOV2 in the dark state, exhibiting an amphipathic structure, consisting of polar and apolar sides, the latter of which docks against the five-stranded antiparallel β-sheet of the core LOV fold. It was also shown that the adduct formation triggers the unfolding of the helical domain Jα, which serves as a linker between the LOV2 domain and the kinase domain in LOV2 of *A. sativa*. That unfolding is believed to modulate the activity of the kinase domain, which is conductive to its autophosphorylation. Moreover, artificial disruption of the LOV2 Jα interaction through site directed mutagenesis results in activation of phot1 kinase activity in the absence of light [55], demonstrating that unfolding of Jα results in activation of the C-terminal kinase domain.

1.1.1.3 The biological role of LOV domain proteins

Plant phototropins play an important role in phototropism, hypocotyl development, chloroplast movement, stomata opening and leaf expansion in dependence on blue light [37]. NPH1 encodes a plasma membrane-associated protein known to be essential for most of the phototropic responses in *Arabidopsis*. Involvement of the blue light photoreceptor in hypocotyl extension in green seedlings was already reported in 1979 [56]. Response to changes in light conditions involves a

variety of photoreceptors that can modulate gene expression, enzyme activity and/or motility.

In *B. subtilis*, YtvA acts as a positive regulator of the sigma stress response pathway, which helps the cells to adapt to unfavourable environmental conditions [57]. The blue light LOV domain receptor from *Caulobacter crescentus* is involved in regulation of the cellular attachment factor which leads to their preferential attachment to particles in search of nutrients [3]. *Caulobacter crescentus* inhabits freshwater, the surface photic zone of which are oligotrophic but the benthic surface are rich in nutrients. Blue light perception and regulation is of particular benefit in such environments as only the blue light can penetrate the deeper water (Figure 1.7). In the mammalian pathogen *Brucella abortus*, the light inducible, LOV mediated kinase activity is responsible for virulence of the bacterium [58] and is activated by blue light.

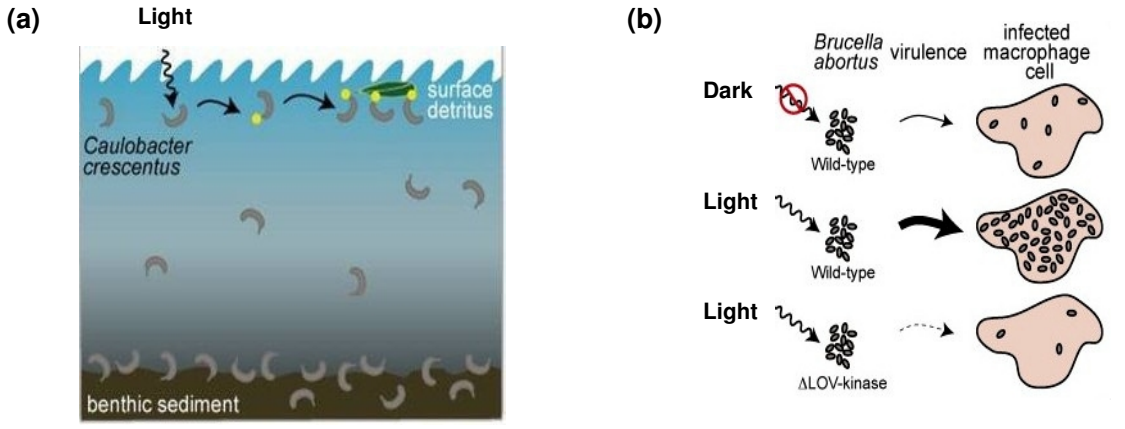


Figure 1.7 Demonstrated effects of blue light irradiation. (a) Light enhances cell-to-cell- and cells-to-surface attachment in *Caulobacter crescentus*. This property helps the bacteria living in the oligotrophic photic zone to move to the nutrient rich benthic sediments. (b) The virulence of the pathogen *Brucella abortus* is dependent of the light reception by the LOV domain kinase protein [3].

In addition, LOV domain-regulated blue light perception can be of great importance, as sensing this spectral range can help the bacteria to avoid the light radiation at the lower wavelengths (e.g., UV radiation) that may cause harmful effects at cellular and DNA level to the bacteria exposed to high intensity radiation lower than the blue spectrum in some environments. On the other hand, blue light is utilized in photolyases-mediated DNA repair processes. In either case, sensing blue light is important for the organisms.

1.2 Metagenomics

Microbes are the most important components on earth and it is estimated that their cell number exceeds 4-6 x 10^{30} [59]. Only few of these organisms cause diseases. Most of the microbes are important as they help to maintain the biogeochemical cycle on the earth and several processes on and in the body and tissues of different organisms. Microbes are present even in adverse and complex habitats such as hydrothermal vents [60;61], acidic environments with pH about 0.8 [62] and the gut of human or termites – only as far as they were identified so far [63]. The important question is "do we know all them?" The answer is "definitely not", we even do not know much about the microbial diversity inside our body itself.

"When we try to pick out anything by itself,
we find it is tied to everything else in the universe"

- John Muir (1911)

The above excerpt from John Muir [64] is relevant in the field of microbial ecology too. Many microbes in the complex environment live together and share to a great deal their metabolic products, which makes the isolation of a single species difficult.

1.2.1 Advent of metagenomics

The scientific community agrees that more than 99% of bacteria in the environment cannot be cultured using conventional methods applied to date, which implies that most species in the environment have never been described or

manipulated [65]. The extent of prokaryotic diversity can be also speculated on the basis of host symbiotic relationship of the microorganisms. There are millions of potential host organisms where microorganisms can live in, on or together with. A simple estimation on the basis of such diverse host organisms and the vast aquatic and terrestrial environment is enough to indicate that the microbial diversity is far greater than those established by cultivation based methods, which is less than 6000 species [65-67]. Alone the human gut contains 10^{14} microorganisms [68] and the gut of a termite has been described to have about 216 phylotypes [63]. These estimations and concepts indicate that a huge portion of microbial diversity remains silent to us and is not explored and understood yet.

A new branch of molecular biology that emerged to deal with the unseen and uncovered prokaryotic diversity has been named “metagenomics”. The term metagenomics was coined to refer the branch of molecular biology that deals with the habitat-based genomic analysis of mixed, uncultured microorganisms [69-73]. Metagenomics is synonymous for community genomics, microbial population genomics and environmental genomics [70;73]. It is also assumed as an umbrella descriptor for the approaches to the direct genomic analysis of the microbial community inhabiting in a given environment [74]. The metagenomic tools analyze the communities of the microbes in the environment, thus bypassing the need to isolate and cultivate the individual microbial species. It enables specific environmental questions to be answered by exploring and using the phenomenal resources of uncultivable and uncharacterized microorganisms [75].

1.2.2 Metagenomic approaches

The steps of metagenomic analysis involve isolation of DNA from an environmental sample, cloning the isolated DNA into a suitable vector, transformation of the clones into host bacteria to form a clone library and screening of the library [70;71]. A pre-cultivation approach can be also applied to overcome a major problem of cloning of environmental DNA, which is caused by the presence of polyphenolic

compounds co-purified with DNA that interfere with the enzymatic processes involved in cloning [71;76].

Metagenomic analysis can be divided into a function-based approach and a sequence-based approach (Figure 1.8) [71;72;77]. In the function-based approach metagenomic DNA is cloned and the clone libraries are screened for a particular phenotype [78-81]. Whereas, in the sequence-based approach the (random) metagenomic clones are sequenced directly [8;82] or screened for the highly conserved 16S rRNA genes to determine the phylogenetic affiliation [70;83]. Moreover, direct sequencing of the environmental DNA and then searching for the genes based on homology are other approaches that are being used nowadays aiming to understand the genomic and functional diversity of the environment [84]. Sequencing may either be random [85], targeted [86] or end sequencing of the large insert clones [87]. A recently introduced 454 pyrosequencing technique allows producing several sequence reads in a single reaction bypassing the need to clone the nucleic acid material [88;89].

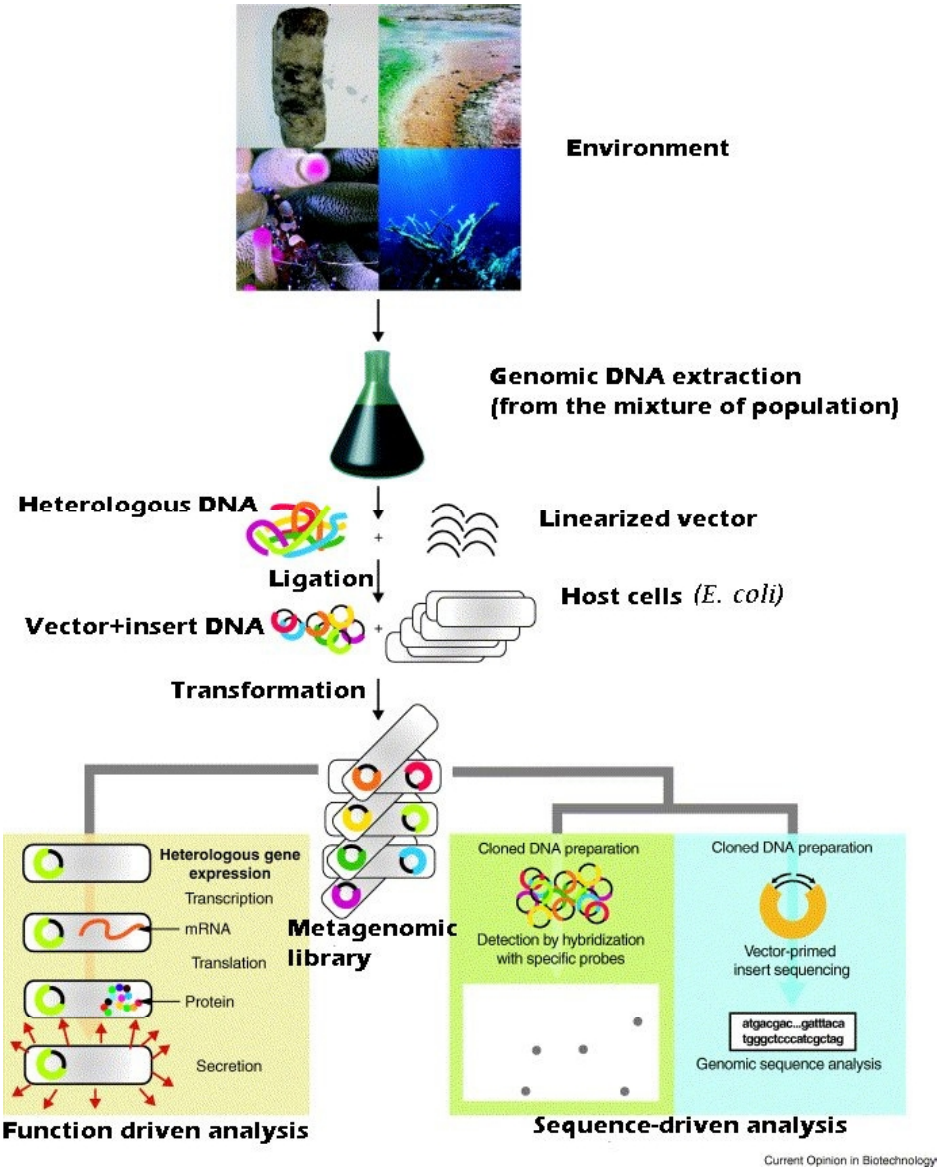


Figure 1.8 A schematic representation of the approaches applied in metagenomics. The DNA from the complex population of an environment is extracted and cloned in a suitable vector which is then transformed/transfected to the heterologous hosts to replicate and maintain. The entity of clones generated in such way is called a DNA library or bank. The library generated can be screened in two ways, one is function-based, in which the gene inside the clone is expected to code a protein which shows its function on the substrate; the other is sequence-based, where the cloned DNA is

sequenced to characterize the clones or hybridized with specific probes to test the clones to search for certain sequences [72].

The development of cloning vectors that accommodate large fragments of DNA for library construction, varieties of expression vectors and hosts for expression, high throughput DNA sequencing methods such as 454 pyro-sequencing, and developments in bioinformatics have helped metagenomics to develop further and gain widespread acceptance [71;73;89]. Progress in the identification tools of molecular microbiology, including 16S rRNA analysis, have made the metagenomic approach the most reliable and promising field in modern molecular microbiology and its dependent divisions [70;90].

1.2.3 Application and scope of metagenomics

The metagenomic approach of DNA analysis has received wide interest and applications in the recent years [70;74;76;90-93]. Though the metagenomics was formerly considered as a research technology for investigating microbial population ecology, the recent advancements show that this field is likely to open the door in the various fields related to human welfare. It has put forward a large scale of scientific explorations, the output of which can help addressing some of the complex medical, environmental, agricultural, and energy issues of today's world [94-96]. A number of publications and the initiation of large community genome sequencing projects in less than a decade after the emergence of the concept of metagenomics indicate that this field has got a wide interest among scientific communities.

Microorganisms have been important resources for novel enzymes since several years [97]. The rate of discovery of new drugs from the microbial secondary metabolites has significantly decreased in past 20 years, whereas a high number of synthetic molecules generated by combinatorial chemistry exhibit low chemical diversity [98]. To meet the enormous need of renovation of such molecules, more and more established chemical processes are switching to biotechnological routes and metagenomics has potential to fulfil this need [99]. Antibiotics [78] and novel biocatalysts, such as amidases, amylases and cellulases [85;100;101], biotin synthase

[79], lactonase [81] esterases and lipases [80;102-104] are just some of the examples of products with biotechnological importance obtained through metagenomics. Many of such novel outputs make metagenomics one of the most promising techniques for the discovery of drugs and novel biocatalysts with particular properties [105].

Metagenomic techniques have recently been applied to investigate the gut microbiomes from different termites, herbivores' rumens and even from humans of different ages, geography, ethnicity and dietary habits [68;106;107]. The human gut contains a dense, complex and diverse microbial community. It is estimated that the human gut contains 10^{13} to 10^{14} microorganism cells and the collective genome of the human gut microbiome is 100 times more than the gene content of our own genome [68]. Current estimates based on sequencing of 16S rRNA genes in DNA extracted from feces suggest that 800 to 1000 different microbial species and >7000 different strains inhabit the gastrointestinal tracts [108], and that the majority of these (>80%) have not yet been isolated or characterized. Each type of microbiota can employ various strategies to adapt the intestinal environment and as many as 104 gene families reported from the gut biomes were of unknown functions. Moreover, the presence of mobile elements among the human gut microbiomes emphasizes the fact that these microbiota are hot spots for horizontal gene transfers [106]. The composition of the human gut microbiome (low Bacteroides to Firmicutes ratio) has been shown to be associated with obesity [109;110]. The evidences suggest that the composition of gastrointestinal microbiota is also linked to inflammatory bowel diseases [111;112] and functions of the microbial communities in the gut influences host health, disease etiology and drug metabolisms [68;109]. It is no need to say here that the metagenomics of the microbiota associated with the human body will give a new perspectives about the probiotics and understanding and potentially curing dangerous diseases like cancer.

Hydrogen metabolism in different environments including hydrothermal vents [60] and similar researches in environmental community involved in energy cycling can be helpful to find important information on alternative energy to cope with the increasing energy demand globally. Many microbes drive various biogeochemical

processes, and clearly, knowledge of their metabolism is helpful for obtaining energy through white biotechnology. Harnessing the power of microbial communities might lead to more sustainable energy sources, since, e.g., certain microbes, when working together as a community, produce a variety of potential energy sources such as hydrogen, methane and even electric current [113;114].

Metagenomics has potential to discover novel cellulases [85;100;101] that are useful for production of biofuel from the cellulose found in common agricultural wastes as corn fiber, corn stalks, wheat straw, and other biomass like switchgrass and miscanthus.

Knowledge of metagenomics can be applied for waste treatment and reclamation of soil and water [115;116]. Besides, metagenomics helps to reveal some of the unknown processes in our environment; a more recent discovery is that of archaeal ammonia oxidizers. It was thought that only bacteria were responsible for aerobic ammonia oxidation, although their numbers often could not account for the observed rates of ammonia oxidation in many habitats. A fortuitous discovery of an ammonia monooxygenase gene next to an archaeal marker gene (encoding the small-subunit ribosomal RNA) implicated archaea as the main source of ammonia oxidation in many marine and terrestrial ecosystems [117].

A metagenomic approach to agriculture could help to develop advanced crops or find improved methods for detecting diseases in crops, livestock and food products. The microbial communities are important in agriculture as they manufacture the nutrients that plants need in order to grow, be it from atmospheric nitrogen or from decaying plants or animals. Moreover, the analysis of the metagenome of the disease-suppressive soil can reveal the anti-pathogenic products that can be used against the agricultural pests [95].

1.2.4 Some major metagenomic activities

Most of the metagenomic projects are focusing on sequencing of microbial genomes from various environments. The metagenome sequencing has made a great leap in recent years (Figure 1.9) and this trend is continuing. The major sequence data till date are from marine environment, which is the largest contiguous ecosystem on earth, occupying more than two thirds of the earth's surface with an average depth of 4 km [118]. Many of the metagenome sequencing projects have focused on this environment.

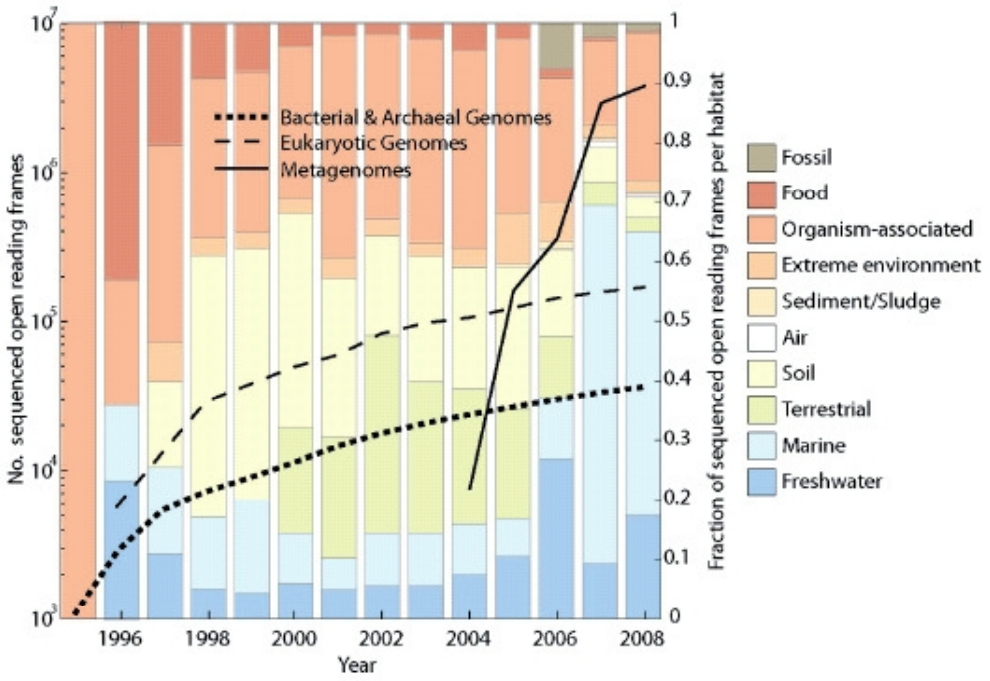


Figure 1.9 Trends of increase in sequence data in past few years [119]. The metagenomic sequence data is increasing in exponential form during the past five years and more than 184 habitats have already been sequenced by 2009. The color indicates the source type of the metagenome (shown with the boxes on the right). Most of the data deposited are from aquatic and organism-associated environments.

The first extensive large-scale environmental sequencing project was carried out by the J. Craig Venter Institute in 2004 in which they sequenced fragments of DNA derived from the entire microbial population of the nutrient-limited Sargasso Sea, an

intensively studied region of the Atlantic Ocean close to Bermuda [120]. The Sargasso Sea sequencing pilot project has already generated a vast amount of sequence data in the database with about one billion bp non redundant sequences, 1.2 millions ORFs (over one million previously unknown genes) estimated to be originated from at least 1800 genomic species [120]. Following, the Global Ocean Sampling expedition carried out after the Sargasso pilot project has yielded an extensive dataset consisting of more than six billion bp sequences and more than six million protein-coding genes [93;121]. Studies of deep-sea communities have been carried out in the Pacific gyre water column at "A Long-Term Oligotrophic Habitat Assessment station" (ALOHA, located 100 km north of Oahu, Hawaii) harvesting about 4.5 Gbp of DNA in fosmid clones from 10 to 4000 meter depth [87] and more recently at a depth of 3000 m a sinlge sample was collected in the Ionian Sea located south-east of Sicily in the deep Mediterranean [122].

Metagenomics has already extended its approach to study some extreme environments such as acid mine drainage (AMD) in the Richmond mine at Iron Mountain, California [62] and a hypersaline mat [123]. A complete reconstruction of the genomes was achieved from the metagenomic sequence data of the comparatively low species habitat of AMD [62].

Soil is regarded as the most biodiverse environment on earth; it is estimated to contain about 1000 Gbp of microbial genome sequences per gram of soil. This was clearly demonstrated by Tringle and colleagues [124] in their analysis of a soil metagenomic library where the soil metagenome constructed from 0.5 g of soil generated about 100 million bp of sequences. It was reported that only less than 1% of the nearly 150,000 sequence reads from independent clones generated from the soil library exhibited sequence overlap.

Other different habitats investigated using metagenomic techniques include air inside urban buildings [125], a carcass of the whale bone [124], a Neanderthal bone samples [126], a sludge community from sewage disposal treatment [127] and viral metagenomes [89] (Figure 1.10). Different sequencing technologies have been used in

the sequencing of metagenomes ranging from conventional shotgun sequencing to pyrosequencing.

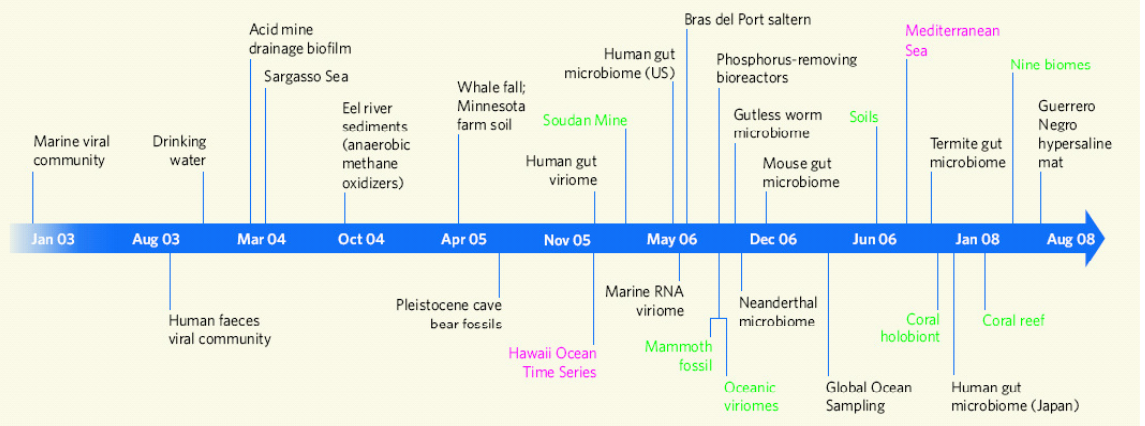


Figure 1.10 A detailed view of the metagenome sequence-based projects since 2002, in chronological arrangement. Data from the shotgun sequencing method are shown in black, pink indicates fosmid end sequencing and green refers to pyrosequencing the respective metagenomes [117].

Recently, metagenomic habitats are also studied by proteomic techniques to investigate the expression of the microbiomes in situ [128]. Another significant recent advancement in metagenomic research is the establishment of the Human Microbiome Project. This interdisciplinary, world wide effort aims to gain a greater appreciation for an understanding of the microbial communities inhabiting several regions of the human body and how they influence human health and diseases [129]. Similarly, the TerraGenome project, a massive metagenomic analysis of soil, comprising the scientists from 23 countries, has been initiated for the complete sequencing of a reference soil metagenome from Park Grass, Rothamsted, UK (http://www.terragenome.org).

The increasing numbers of metagenome sequencing projects have already added a tremendous amount of sequence data to the databases. After the onset of further large sequencing projects including TerraGenome and Human Microbiome Project, the public databases will be overwhelmed with the sequence deposits.

According to Nathan Blow [130], new DNA sequencing techniques for sequencing single molecules in real time predicting sequence reads of about 1500 bp is under development by VisiGen Biotechnologies (Houston, USA) and another effort of producing about 10 kilo base pair sequence read is targeted by Pacific Biosciences (California, USA). These technological advancements will expedite the deposition of bulk of data in the metagenomic database in the future.

The metagenomics is shifting from species and functional identification for individual datasets to comparative analysis [89]. Function-driven approaches, though promising for the discovery of novel biocatalysts, are slower and have limitations, e.g., a search for a host of expression, a vector, or the lack of knowledge about the required substrate [131]. The major focus of the recently completed and announced metagenomic projects is the sequencing of the microbiomes. The present focuses are more into the analysis and understanding of the potential functions of the genes from the metagenomes by relating them to the individual microorganisms using in silico methods [119].

For the identification of the novel genes from the metagenomes in order to address the biotechnological need and the understanding of in depth functional mechanisms of such genes, novel techniques are required. Researchers from every part of the world are working for the development of new techniques for the analysis of the metagenomic datasets and improving the tools to investigate the metagenomes [132-134].

1.3 Photoreceptors genes from the metagenome

The microbes in the various environments of our earth need to equip with photoreceptor genes for survival and maintenance as they are exposed to different and variable light conditions. Among the photoreceptor proteins, **proteorhodopsins** are the most widely investigated and reported ones from the marine environments [135]. The work of Oded Beja and colleagues in 2000 [8] where they reported the presence of this type of rhodopsins in a large insert clone (BAC clone) from a marine metagenomic DNA invented a new way of bacterially mediated, light driven energy

generation in oceanic surface waters. The marine environment-derived gene (product) was named “proteorhodopsin” because the source DNA was from a gammaproteobacteria which was confirmed by ribosomal RNA (rRNA) analysis and also by the homology of protein-coding regions upstream and downstream to the proteobacterial genes [8]. The wide presence of similar types of proteorhodopsin genes was demonstrated in the same environment showing their prevalence in marine surface waters [136]. In the meantime, proteorhodopsin-related genes have been reported in large numbers from the metagenome of ocean environments, e.g., the Sargasso Sea and the Global Ocean Sampling approaches list 782 and 2674 putative proteorhodopsins, respectively [93;120]. They are reported also from hypersaline water in a sea salt manufacturing facility [137]. A recent study on the marine microcosm using quantitative PCR has shown that the proteorhodopsin gene expression is strongly upregulated by light [138]. Apparently, daytime sensing is important for cells to select which uptake mechanisms to activate and which metabolic pathways to adopt. Moreover, as it is daytime when light induced DNA damage occurs in the cells, there is a need of a sensor for the cells inhabiting the exposed environments to switch an upregulation of DNA-repair proteins and mechanisms [135].

Similarly, rhodopsins have been recently reported from the freshwater metagenome dominated by actinobacteria and named as **actinorhodopsins** [139]. Presence of BLUF domains also have been reported by BLAST analysis from four metagenome assemblies [140]. The global analysis of the metagenome data bank using phylogenetic and neighborhood methods shows the presence of several novel branches of photoreceptors in the metagenomes [119]. DeLong’s group [141] has reported the presence of proteorhodopsins in fosmid clones derived from an archaeal metagenome from the North Pacific Subtropical Gyre, with evidence of lateral gene transfers between marine planktonic bacteria and archaea.

1.4 DNA Microarray

A microarray in its simple form is a compilation of several molecules of one type on a solid surface (glass or membrane) in an orderly manner. There are several types of microarray, e.g., DNA microarray, RNA microarray, protein microarray and tissue microarray, according to the probes immobilized on it. The most commonly used device is DNA microarray. A typical DNA microarray consists of several DNA samples immobilized on a glass slide.

In functional aspects, a DNA microarray is an evolved form of southern blots for colony filter and dot blot hybridization techniques. In the next step of miniaturization of the previous hybridization techniques robotic devices for spotting thousands of probes (sample immobilized on the solid surface) on a membrane or a glass slide were used making the array of probes [142;143]. It was initially used to monitor the expression of several genes in parallel in human or in *Arabidopsis thaliana* [142;144]. It is a high throughput technique which allows a hybridization-based analysis of several thousands of genes in parallel in a single experiment. DNA microarrays have already made important contribution in the analysis of cellular processes and gene level expression, but true potential of this technology is still far greater [145].

A DNA microarray can be categorized in different ways, but commonly there are three types [143]: (i) in situ synthesized array (Affymetrix types), (ii) double stranded DNA microarray and (iii) oligonucleotide DNA microarray. The Affymetrix types of array, which is popularly called gene chip, contains short oligonucleotide probes (about 20 bp in length), where the oligonucleotides are synthesized directly on the microarray slide [146;147]. This technology allows very high feature densities of up to 400,000 probes on a commercial array [148;149]. The latter two types of array contain either the PCR amplified probes (DNA or cDNA) or already synthesized oligonucleotides printed [150;151] or spotted [152] onto the slides containing chemically reactive groups (typically aldehydes or primary amines) that stabilize the DNA onto the slide, either by covalent bonds or electrostatic interactions (Figure 1.11). The spotted or printed probes are then hybridized to fluorescence-labeled DNA, and a positive hybridization is detected with the help of a fluorescence scanner. The probes

are arranged on the slides in such a way that their spatial position can be traced back and an individual probe can be identified.

Nowadays, a DNA microarray is a regular analysis tool for many laboratories involved in understanding biological processes of living organisms and also as a diagnostic tool to monitor serious diseases like cancer. It is being regularly used in microbiology with pure cultures to study biological processes and also for minisequencing of short nucleotide reads and mutation or SNP (single nucleotide polymorphism) analysis.

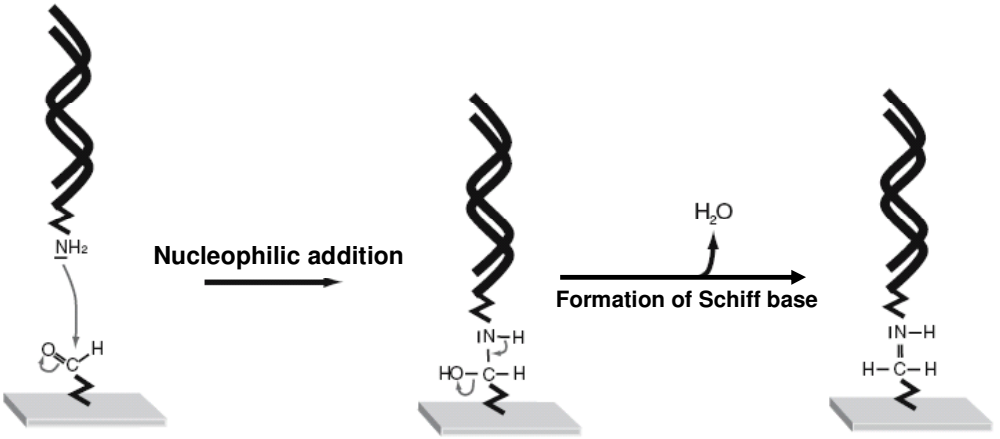


Figure 1.11 Immobilization (spotting) of amino-modified DNA to an active, aldehyde coated microarray slide. An amino linker is attached to the 5' end of the oligonucleotide or the PCR fragment and then spotted on the slide. The amino end group binds covalently to the slide via the prefabricated aldehyde group [143].

1.4.1 DNA microarray in microbial ecology

A critical and labor-intensive step in metagenomics is screening of several thousand heterogeneous clones in search of any specific genes. The already developed techniques such as polymerase chain reaction (PCR), reverse transcriptase PCR, real time PCR, fluorescence in situ hybridization (FISH) and southern blot hybridization techniques have made it possible to analyze the individual or simple groups of organisms but their application in metagenomic analysis involving several

thousand clones or a mixed population is less appropriate in terms of specificity and overall efficiency. DNA microarray is emerging as an alternative and robust tool in microbial ecology. The high throughput-analyzing capacity of the DNA microarray has already attracted the interests of environmental microbiologists. More recently, DNA microarrays have been applied for phylogenetic or functional characterization of the microbiomes from different environments using the community DNA [153-155]. Gentry and colleagues [156] have divided the microarrays that have been used in environmental microbiology into following five categories:

(i) Phylogenetic oligonucleotide arrays
(ii) Functional gene arrays
(iii) Community genome arrays
(iv) Metagenomic arrays
(v) Whole genome open reading frame arrays

The phylogenetic oligonucleotide arrays (i) are used for phylogenetic differentiation of the microbial population and detection of specific microorganisms [157]. They consist of short oligonucleotides as probes targeting specific regions of rRNA. Functional gene arrays (ii) are used to detect the genes involved in different biological processes in the environment which provide a certain degree of classification and also information about the gene activity in the environment under study [158]. Community genome arrays (iii) contain the entire genomic DNA as a single probe and were initially designed to detect specific organisms in a microbial community [159]. In metagenomic arrays (iv) the PCR-amplified DNA from the metagenome or the metagenomic clones are immobilized on the slides to generate a DNA microarray which is hybridized to the reference DNA samples allowing characterization of the metagenome [153;160]. Whole genome open reading frame arrays (v) contain the probes for all of the ORFs present in a genome and are used for the comparative genomics of different organisms particularly to understand the processes of lateral gene transfer [161;162]. The above DNA microarrays used in microbial ecology are employed in environmental profiling or the comparative analysis of the genes or the genomes.

1.5 Scope and outline of this thesis

The study of LOV domain proteins in prokaryotic domain is comparatively young. It was set off after revealing the bacterial counterpart of phototropins, YtvA from *Bacillus subtilis*, with its functional proof [163]. Few more LOV domain proteins have been characterized from the bacterial domains [46;140;164-168], but a metagenomic approach has not been applied in this field. Such approach requires high throughput tools to screen and identify the genes of interest, as several thousand clones are made from a single community and sometimes hundreds of thousands of clones need to be screened for the identification of a significant gene [102]. Moreover, a suitable tool to screen the clones for the presence of blue light receptor genes has not been reported yet.

In this background the major aim of this dissertation was to develop a robust tool based on the DNA microarray technique to screen-identify-retrieve novel LOV domains from metagenomic DNA libraries and to characterize the identified gene. Furthermore, to understand the ecological significance of the LOV domain in environmental aspects, habitat-based investigations and phylogenetic analysis of the different metagenome-derived LOV domains have been carried out.

Hence, the scope of this dissertation was multidisciplinary, covering methodological and biological objectives. Chapter 3 deals with design, standardization and application of a DNA microarray-based high throughput technique to screen for LOV domains in environmental samples and metagenomic libraries. In chapter 4, the functional characterization of metagenome derived LOV domains is presented. Chapter 5 is an attempt to shed light on the distribution of the LOV domain proteins in different metagenomes on the basis of available sequence data and their phylogeny.

Chapter 2
2. Materials and Methods

2.1 Bacterial strains, vectors and primers

2.1.1 Strains

Host bacterial strains used in this work are presented in table 2.1.

Table 2.1 Bacterial strains used in present work

Strain	Properties	Reference
E. coli		
VCS257	Derivative of DP50, supF	Stratagene, USA
XL1Blue	*rec*A1, *thi*-1, hsdR1, supE44, *relA*1, *lacF*, *proAB*, *laclq*, *lacZΔM15*, Tn10 (Tc)	[169]
BL21CODONPLUS®(DE3)RIL	*E. coli* B F^- *ompT hsdS*(r_B^- m_B^-) *dcm*+ Tet^r *gal*λ(DE3) *endA* Hte [*argU ileY leuW* Cam^r]	Stratagene, USA
BL21 AI	F^- , ompT, gal, dcm, lon, $hsdS_B$(r_B- m_B-) araB::T7RNAP-tetA	Invitrogen, USA
ER256	F- λ- fhuA2 [lon] ompT, lacZ::T7 gene 1 gal sulA11 Δ(mcrC-mrr)114::IS10 R(mcr-73::miniTn10-TetS)2 R(zgb-210::Tn10)(TetS) endA1 [dcm]	NEB, USA
P. syringae* pv. *tomato	Plant pathogen	TIGR
P. syringae* pv. *syringae	Plant pathogen	TIGR
P. putida KT2440		[170]
Bacillus subtilis	Wild type	[171]

2.1.2 Plasmid and cosmid vectors

The vectors used for cosmid library generation, cloning, sub-cloning and protein expression are listed in table 2.2.

Table 2.2 Plasmid/cosmid vectors and constructs

Plasmid	Properties	Reference
pWE15	Amp^r, Neo^r, cos	Stratagene
pHSG399	Cm^r, *lac*Z, MCS, high copy number	[172]
pJET1.2	*eco47IR*, Amp^r, MCS, blunt end cloning vector	Fermentas
pDRIVE	Amp^r, MCS, *lac*Z, TA cloning vector	Qiagen
pET28	His-tag, T7 tag, lacI, Km^r	Novagen
pET30	His-tag, S tag, lacI, Km^r	Novagen
pET52 3C/LIC	His-tag, Strep tag, *lacI*, Amp^r, ligation-independent cloning	Novagen
pWPPLOV	DNA from *P. putida* LOV core cloned into pWE15	This work
pWThLOV	Garden soil enrichment clone detected by microarray	This work
pHThLOV_HSG399	Subclone of pWThLOV in pHSG399	This work
pET19-LOV-HT-Met1	LOV domain from pWThLOV in pET19b vector	This work
pET28-LOV-HT-Met1	LOV domain from pWThLOV in pET28a vector	This work
pET30-LOV-HT-Met1	LOV domain from pWThLOV in pET30b vector	This work
pET52-LOV-HT-Met1	LOV domain from pWThLOV in pET52 vector	This work

Plasmid	Properties	Reference
pETElbeFusion	Fusion LOV domain from Elbe River metagenome and YtvA LOV	This work
pETElbeFusion:P	Amino acid replacement S to P in pETElbeFusion	This work
pETElbeFusion:R	Amino acid replacement P to R in pETElbeFusion	This work
pETWFLOV	Whale Fall LOV gene in pET28 a	This work
pETWFLOV-Trim	Whale Fall only LOV domain in pET28a	This work
pET28YtvA	YtvA in pET28a	[173]
pET28CauloLOV	Lov domain from *Caulobacter crescentus* cloned in pET28a	[173]
pET28PSTLOV	*P. syringae* pv tomato LOV in pET28a	[168]
pET28Psyrsyr	*P. syringae* pv syringae LOV in pET28a	[174]

2.1.3 Oligonucloetide primers

Oligonucleotide primers used for PCR amplification and sequencing reactions are given in table 2.3.

Table 2.3 Oligonucleotide primers used for PCR and sequencing reactions. K, R, Y and N are the IUB symbols used here for the degenerate bases (K= G, T; R= A, G; Y=C, T; N= A, C, G, T)

Name	Sequence (5'-3')
5784F	GGAATTCATATGGCCCTGATCGAGAAG
5784R	CGTACTCGAGTTACATTCCCGCTTC
5784Trim_For	TTATCCATGGCT TCG GTG ATC AGC AAT CC
5784Trim_Rev	TTTACTCGAG GAT CTC GAC CTG GGA ACC
Bsub_fus_for	TTG GCA TAT GGC TAG TTT TCA ATC ATT TGG
Bsub_fus_rev	TTA GTG GAT CCT AAA ATT TCC TCG GTC TC
Elb_fus_for	TAT TAG GAT CCA ACT GCC GGT TTT TGC AG
Elb_fus_rev	TAA TCT CGA GGT CAC GTC GTT CTG GA
Fus_52_For	CAG GGA CCC GGT ATG GCT AGT TTT CAA TC
Fus_52_Rev	GGC ACC AGA GCG TT G GTC ACG TCG TTC TG
ThMet_for1	TTT ACA TAT GTC GGG CAC CCA GGT GCT
ThermoLOV_rev	GTA ACT CGA GTT ACA GCT CCT CGC TCA CG
ThLOV_52_For	CAG GGA CCC GGT ATG TCG GGC ACC CAG
ThLOV_52_Rev	GGC ACC AGA GCG TT CAGC TCC TCG CTC AC
Chl_for	GCCGGAACTGCCGG TTY YTN CAR GG
Chl_rev	GGTCACGTCGTTC TGN ACN CCN AC
PSyrTom_rev	GAAGGTGGAGCCGTCC TTN CKR TAR TT
PsALL_rev2	GGGGCAGCCGTCC TTN CKR TAR TT

M13 for	GTT GTA AAA CGA CGG CCA G
M13 rev	CAC AGG AAA CAG CTA TGA C
RalD1_for	ATT ACC GVG CGG ARG AGG
B2D_1forSP	AAG GCG CGC CTG AAG GAA CAA C
B2Dfor_BackCheck	CTT CAG CAG CTC GCA GAC CAG
B2D_1revSP	ACG TAT TGC GAG TCG TCC AGT C
B2D_Seq_for2	GAG GAA CTC CGC CAG GTA G
B2D_Seq_for2a	TGA CTT CGG CGA TGA CCT TG
B2D_Seq_RD2	GCG AGC TGC AGA TCA TGC AC
B2D_Seq_D3F	CCG CTT CAT GCC GAT GTC G
B2D_Seq_D3FA	TCG CGG TGC AGC ACA AGA AC
B2D_Seq_D2R	GTA CTT GAT CGC GTT GGA GAT G
S5_1forSP	GAC AGA CGC GGT TTG CGT ATT G
S5_1revSP	CCT CTG TCA GGG AGG TTT CGT G
V1LOV_for	ACTGCCGGTTTCTGCAAG
V1LOV_rev	CTTGCGGTAGTTGAGTACTTCC
Fus_S2P_F	GACCGAGGAAATTTTAGGACCCAACTGCCGGTTTTTGCAGG
Fus_S2P_R	CCTGCAAAAACCGGCAGTTGGGTCCTAAAATTTCCTCGGTC
Fus_M_For	GACCGAGGAAATTTTAGGACGCAACTGCCGGTTTTTGCAG
Fus_M_Rev	CTGCAAAAACCGGCAGTTGCGTCCTAAAATTTCCTCGGTC

2.2 Chemicals, consumables and enzymes

All chemicals used were of analytical grade. The chemicals were obtained from Merck (Darmstadt, Germany), Sigma (Deisenhofen, Germany), Serva (Heidelberg, Germany), DIFCO (Detroit, USA), Gibco/BRL (Eggenstein), Biomol (Hamburg, Germany), Pharmacia (Freiburg, Germany), ICN Biomedicals (Aurora, USA) and BioRad (München). Restriction enzymes and DNA-modifying enzymes used in this work were purchased from Fermentas (St. Leon Rot, Germany). Oligonucleotides used as PCR and sequencing primers were synthesized by Metabion International AG (Martinsried, Germany). For DNA microarray construction oligonucleotide probes were synthesized by Operon Biotechnology (Cologne, Germany). Ni^{2+} IDA Metal Affinity Resin and disposable columns from Prochem (USA) were used for the purification of the His-tagged proteins. The anti-penta His antibodies were from Qiagen (Hilden) and the secondary antibodies (AP conjugates) were from Dako (Hamburg, Germany). dNTPs, Cy3- and Cy5-dCTPs were purchased from GE Healthcare (Freiburg, Germany). Random Hexamers used for labeling of target DNA for microarray hybridization were obtained from Roche (Germany).

2.3 Microbiological techniques

2.3.1 Media and supplements

Liquid media were prepared in the required volume in tri-distilled water and autoclaved at 121 °C (Systec, Germany). Millipore deionized and threefold-distilled water was used for preparing media and buffers. Solid media were prepared with the addition of 15 g/L bacteriological agar (Serva, Germany) before autoclaving. When required, heat labile components e.g., antibiotics, IPTG, X-Gal were filtered through 0.22 μm filter and added to the autoclaved media.

Media

LB:	10 g/l NaCl, 5 g/l yeast extract, 10 g/l tryptone
2xLB:	20 g/l NaCl, 10 g/l yeast extract, 20 g/l tryptone
LB-agar:	10 g/l NaCl, 5 g/l yeast extract, 10 g/l tryptone, 15 g/l agar
TBY:	16 g/l tryptone, 10 g/l yeast extract, 5 g/l NaCl
TBY-agar:	16 g/l tryptone, 10 g/l yeast extract, 5 g/l NaCl, 15 g/l agar

Antibiotics and supplement stock solutions

Ampicillin:	100 mg/ml in tdH_2O
Chloramphenicol:	25 mg/ml in 50% ethanol
Kanamycin:	50 mg/ml in tdH_2O
IPTG:	100 mg/ml in tdH_2O
X-Gal:	50 mg/ml in DMSO

Antibiotic solutions, IPTG and X-Gal were stored at -20 ^oC. The tubes containing the stock solution of X-gal was wrapped with aluminum foil to protect from light. Media supplements were added to the sterile media that were cooled below 50 ^oC.

2.3.2 Growth conditions

All *E. coli* strains except for pre-cultures were grown at 37 ^oC. The incubation period varied depending on the following application for the cells as mentioned in the related topics. *E. coli* cells were grown overnight on LB agar plates or in liquid LB media in test tubes or Erlenmeyer flasks. A single colony or an aliquot from glycerol culture stock was used to inoculate the culture medium. 1% (v/v) of a pre-culture was used to inoculate a large volume liquid culture. Pre-cultures (starter cultures) were prepared by overnight incubation of the cells at 30 ^oC. *Pseudomonas* strains and *B. subtilis* were grown at 30 ^oC on LB agar plates or LB liquid medium.

2.3.3 Culture maintenance

2.3.3.1 Maintenance on agar plates

Frequently required strains were maintained on LB agar plates for short term storage. The agar plates were sealed with paraffin sheet and stored at 4 ^oC for 4-6 weeks. For protein expression freshly streaked colonies were used.

2.3.3.2 Maintenance at -80 ^oC (glycerol stock)

In order to preserve cells for long term, glycerol stocks of the cells were made. 60% of overnight grown liquid culture was mixed with 40% of sterile glycerol (99.6%) in sterile vials to make glycerol stock. The vials with glycerol cultures were kept in -80 ^oC

for long term preservation while for short term preservation (about 1 year) the -20 ^{o}C deep freezer was used.

2.3.3.3 DMSO stock

Liquid cultures of *E. coli* were grown overnight at 37 ^{o}C in 96 well micro-plates with each well having 100 µl of LB selective medium. 5 µl of DMSO was added to each well of the micro titer plate containing overnight grown cells and mixed properly in order to make DMSO stock of cells. The DMSO cell stocks were frozen and preserved at -80 ^{o}C.

2.3.4 Cell harvesting

Cells in liquid cultures of up to 5 ml volume were sedimented using a tabletop centrifuge (Hettich, Germany), at 4 ^{o}C for 2-5 minutes at 6,000 to 13,000 rpm. Larger volumes of cells were pelleted using a rotor centrifuge (Beckman Coulter, USA) with JA-10 or JLA-16250 rotors at 4 ^{o}C at 6,000 rpm 10 to 20 minutes.

2.3.5 Measurement of growth using optical density

The growth of unicellular organisms could be rapidly determined by measuring the turbidity related to cell density. As microbial cell numbers increase, light passing through the culture decreases. The output is often expressed as absorbance or optical density which is a logarithmic expression of the amount of light that gets through the culture.

Cell density was determined by observing the optical density using spectrophotometer taking the pure medium as blank. Cells in specified solution or culture medium were placed on cuvettes with a layer thickness of 1 cm and observed at a wavelength of 600 nanometer (nm).

2.4 DNA manipulation

2.4.1 General techniques

All heat stable solutions and devices were autoclaved in order to inactivate nuclease activity and reduce contamination. Solutions were autoclaved at 121 ^{o}C at 20 psi for 20 minutes and instruments were autoclaved at 121 ^{o}C at 20 psi for 15 minutes. Devices that could not be autoclaved were rinsed with 70% ethanol and flamed to evaporate ethanol. Heat sensitive chemicals and solutions were filter-sterilized using 0.2 µm sterile filters.

2.4.2 DNA isolation

2.4.2.1 DNA from pure cultures using chloroform/isoamyl alcohol

DNA is isolated from cells by breaking the cell wall and removing the proteins. The cells are lysed with anionic detergents in the presence of a DNA stabilizer and peptide bonds are hydrolyzed to free the DNA from associated proteins.

A bacterial culture (1.5 ml) was centrifuged in an E-cup and re-suspended/washed in 1 ml of 0.9% NaCl using vortexing, and centrifuged again to sediment the cells. The supernatant was discarded. The pellet was re-suspended in 564 µl 1x TE buffer. 30 µl of 10% SDS was added to it and mixed gently. 6 µl of proteinase K (20 mg/ml) was added and mixed gently and the mixture was incubated for 1 hour at 37 ^{o}C. 100 µl 5 M NaCl was added to it, mixed gently and 80 µl prewarmed CTAB/NaCl (at 65 ^{o}C) was added to it, mixed gently and incubated for 10 minutes at 65 ^{o}C.

The chloroform-isoamyl alcohol (24:1 v/v) extraction was performed as a stand-alone DNA purification step. Chloroform extraction removes protein contaminations from the DNA samples, using the different reactivity of proteins in comparison to DNA with organic solvents. While the proteins are denatured by the organic solvent, and after phase separation are localized as a thin layer between the two phases, the DNA remains in soluble form and can be recovered from the aqueous phase. The DNA containing solutions (bacterial extracts, restriction reactions mixture) were mixed with an equal volume of chloroform-isoamyl alcohol and mixed vigorously. The mixture was

centrifuged at 4 °C at 13000 rpm for 15-20 minutes. The upper aqueous phase was carefully transferred into a new tube. The DNA in the aqueous phase, obtained after protein extraction can be precipitated using isopropanol or absolute ethanol. The clear nucleic acid phase obtained after chloroform/isoamyl alcohol extraction was precipitated using isopropanol or ethanol. For this purpose, 0.6 volumes of ice-cold isopropanol or 2-3 volume of ice-cold absolute ethanol was added to the aqueous phase containing DNA and mixed carefully. The mixture was incubated at -80 °C for 10-20 minutes. Precipitated DNA was collected after centrifugation at 13,000 rpm at 4 °C for 20-30 minutes. The DNA pellet was washed with 100 to 200 µl of 70% ethanol, centrifuged again at 4 °C for 5-10 minutes. The DNA pellet was then dried, resuspended in tdH_2O and stored at 4 °C.

Reagents and solutions

1XTE Buffer:	10 mM Tris-HCl (pH 8.0), 1 mM EDTA
SDS 10%:	SDS 10 mg, tdH_2O to 100 ml, filter sterilized
CTAB/NaCl:	NaCl 4.19 g, CTAB 10g, td H_2O 100 ml (filter sterilized)
Proteinase K:	Proteinase K 20 mg/ml in tdH_2O, freshly prepared

2.4.2.2 DNA from pure cultures using Qiagen Kits

To obtain pure DNA from pure culture strains, the "DNeasy® Blood & Tissue Kit" from Qiagen (Hilden, Germany) was used according to the manufacturer's manuals. All required reagents were contained in the kit.

2.4.2.3 Plasmid isolation by alkaline lysis

This method was modified from Birnboim & Doly [175]. This process was used to check the presence of recombinant plasmids from a large number of *E. coli* clones. Self-made solutions can be used in this method which is very cheap in comparison to the commercially available kits. The bacteria are lysed by SDS in presence of NaOH in cold reaction mixture. The reaction leads to denaturation of RNA, DNA and proteins. Following the neutralization reaction, chromosomal DNA and proteins co-precipitate whereas plasmid DNA remains in the solution. Plasmid DNA isolated in this way can be used directly for restriction analysis and PCR.

About 4 ml of culture was taken in an E-cup, centrifuged for 5 minutes at 13000 rpm to sediment the cells. Cell pellets were resuspended in 200 µl of P1 buffer. 200 µl of P2 buffer was added to it, mixed well and incubated at room temperature for 2 minutes. 200 µl of chloroform were added, the mixture was vortexed and incubated at room temperature for 5 minutes. 150 µl of P3 were added to it to precipitate the protein, vortexed for 2-3 seconds and incubated for 5 minutes at room temperature. The solution was centrifuged for 5 minutes at 13000 rpm at 4 ^{o}C, the upper clear phase was transferred to a new E-cup. The plasmid DNA was precipitated using isopropanol or ethanol, washed with 70% ethanol, dried at 37 ^{o}C and dissolved in 20 µl sterile H_2O.

Solutions and reagents

P1 Buffer: 50 mM Tris-HCl pH 8.0, 10 mM EDTA, 100 µg/ml RNaseA

P2 Buffer: 0.2 M NaOH, 1%SDS

P3 Buffer: 3.2 M potassium acetate, 11.5% acetic acid, pH5.5

2.4.2.4 Isolation of plasmid DNA using "QIAprep Spin Miniprep Kit"

To obtain high quality plasmid DNA for sequencing, the "QIAprep Spin Miniprep Kit" from Qiagen (Hilden, Germany) was used. About 25 µg of highly pure plasmid DNA (high copy number vectors) could be obtained from 3-5 ml of *E. coli* culture. The principle is based likewise on alkaline lysis of the cells that is followed by adsorption of negatively charged DNA to a Silica gel matrix in presence of high salt concentrations. The DNA was eluted in tdH_2O and the quality of DNA was appropriate for sequencing and other activities that required high quality DNA.

2.4.2.5 Rapid plasmid isolation from *E. coli* by "Cracking"

With this method, plasmid DNA from a single colony or pellet obtained by centrifugation of liquid culture could be isolated rapidly, in order to determine the size of the plasmids. This method was rapid and also bypassed the need of preparation of overnight liquid cultures. This method does not allow further manipulation of plasmid DNA i.e., for cloning or sequencing.

A colony from an agar plate was taken into an E-cup containing 25 µl of 10 mM EDTA (pH 8.0) and suspended. 25 µl of fresh cracking buffer was added to the suspension, mixed well by repeated pipetting and incubated for 5 minutes at 70 °C, then placed in ice to cool. 2 µl of cracking stain was added, mixed well with a micro pipette and incubated in ice for 5 minutes. The solution was centrifuged at 4 °C at 13,000 rpm for 10 minutes. 15-20 µl of supernatant was directly poured into agarose gel and the gel electrophoresis was run for 1 hour at 100 V.

Reagents and solutions

Cracking buffer:

2N NaOH	100 µl
10% SDS	50 µl
Sucrose	0.2 g
tdH_2O	850 µl

Cracking stain:

KCl (4 M)	1.5 µl
Bromophenol blue (0.4% w/v)	0.5 µl

2.4.2.6 DNA isolation from cosmid banks

Cosmid DNA from the library can be extracted using home-made alkaline lysis solutions (2.4.2.3) or 96 well BAC/cosmid DNA extraction kits. In order to apply the alkaline lysis procedure the clones can be pooled in batches of larger volumes. Here, the extracted DNA was obtained from groups of pooled clones.

When pooling strategies are applied, only aliquots of the contents of each well are pooled, allowing at a later step of the procedure to re-assign a positive clone to a position in the microplate.

(i) Pooling of clones and alkaline lysis

About 2 ml of an overnight grown culture from each well was pooled into a 50 ml polypropylene tube such that a pool consisted of the clones from a group of 32 wells.

The pooled culture was harvested by centrifugation at 6000 rpm. The alkaline lysis protocol was applied as described in 2.4.2.3.

(ii) Kit based extraction

A 96 well BAC or cosmid DNA isolation kit can be used to extract the cosmid or BAC DNA from the metagenomic library. The liquid culture in 96 deep well culture plates was harvested by centrifuging at 4000 rpm and subjected to the lysis procedure using a Favorprep™ 96 well plasmid isolation kit (Favorgen, Taiwan) following the kit supplier's manual. Final elution was made with 140 µl of tdH_2O added in each well. The extracted DNA was stored at 4 ^oC till further use or at -20 ^oC for longer time.

2.4.2.7 Agarose gel electrophoresis

Agarose gel electrophoresis is a method used to separate DNA strands by size in a porous agarose gel under an electric field. The electric field is applied to drag negatively charged DNA molecules through a gel matrix. The shorter DNA molecules move faster than the longer ones since they are able to slip through the gel more easily. The rate of migration is affected by a number of factors that are concentration of agarose, voltage applied and conformation of DNA. An agarose gel stock was made by adding agarose in 1% TAE buffer as required and solidified by pouring into a gel casting cassette. DNA was mixed with loading dye (0.2 volumes) in order to allow visualization and to sediment it in the gel pocket. In the SYBR Green (GE Healthcare) based staining method, the DNA was stained before loading to the gel by adding SYBR Green to it in a ratio of 1:5000. The gels were run at 5 V/cm until optimal separation was achieved. For ethidium bromide based staining the gels were stained with 0.1 µg ethidium bromide/ml (in tdH_2O or 1XTAE buffer) for about 10 to 15 minutes and de-stained for about 2-10 minutes in a water bath in order to reduce the background. The stained DNA (after SYBR Green or ethidium bromide staining) was visualized under UV light. Gene Ruler™ 1 kb or 100 bp DNA Ladder Plus (Fermentas) was used as a marker as required.

Reagents and buffers

Loading dye: Glycerol 30% (v/v), EDTA 50 mM, Bromophenol blue 0.25% (w/v), Xylene cyanol 0.25% (w/v)

50X TAE buffer: Tris 2M, EDTA 100 mM, tdH_2O to make final volume 1 L

2.4.2.8 DNA extraction from agarose gel

DNA fragments were recovered from the agarose gel using "QIAquick gel extraction kit" (Qiagen, Hilden) following manufacturer's protocol. Selection and recovery of a required fragment of DNA from the gel was possible with this method. The necessary reagents and DNA adsorption columns used were supplied with the kit.

2.4.3 Enzymatic modification of DNA

2.4.3.1 Restriction digestion

This method is based on type II restriction enzymes isolated from prokaryotes which are highly specific and act on recognition sites of double stranded palindromic (mostly) DNA. The enzyme hydrolyzes the phosphodiester bond between two bases in defined places of the strand making two incisions, one through each of the phosphate backbones of the double helix without damaging the bases and produces over-hangs (sticky) or smooth (blunt) ends.

For cloning of DNA into a circular plasmid or a cosmid vector, the vector DNA was first linearized using appropriate restriction enzymes. PCR products or DNA to be cloned were also digested using the appropriate restriction enzymes to produce complementary overhangs or blunt ends for cloning into the linear vector.

Analytical digestion reaction (conventional restriction enzyme)

DNA	7 µl (ca 0.5 µg)
10X Buffer	1.5 µl
Enzyme	0.3 µl
H_2O	to 15 µl

The reaction was incubated at 37 ^oC (or at the temperature recommended by the manufacturer) for 3 to 16 hours.

Analytical digestion reaction (fast digest enzyme)

DNA	7 µl (ca 0.6 µg)
10X Buffer	1.5 µl
Enzyme	1 µl
H_2O	to 15 µl

The digestion reaction was incubated for 10 to 30 minutes at 37 ^{o}C.

For reparative digestion reaction, the amount of the constituents was scaled up to required volume.

2.4.3.2 Dephosphorylation

In order to avoid re-ligation of empty cosmid and plasmid vectors during the ligation reaction, 5' phosphate residues at the end of linearized vectors were removed by alkaline phosphatase treatment. After preparative vector DNA digestion, the vector DNA was dephosphorylated by incubating with 1 µl (1 unit per pmol 5' ends) shrimp alkaline phosphatase (Fermentas, Germany), and 10x dephosphorylation buffer (10% of the total reaction volume) for 45 minutes at 37 ^{o}C. Volume adjustment was done by H_2O addition. The phosphatase enzyme was inactivated by incubation for 10 minutes at 65 ^{o}C. The DNA was purified using QIAquick PCR purification kit (Qiagen, Germany).

2.4.3.3 Ligation

DNA ligation involves creating a phosphodiester bond between the 3' hydroxyl end of one nucleotide and the 5' phosphate end of another. For cloning of DNA into plasmid/cosmid vector the complementary ends produced by the digestion of appropriate restriction enzymes were ligated into the linear vector prepared by the similar restriction enzyme that produce complementary overhangs or blunt ends.

Ligation reaction

DNA	6 to 8.5 µl (0.2-0.5 µg)
Vector DNA	4 to 8 µl (0.1-0.2 µg)
10x T4 buffer	2 µl
T4 ligase	0.5 µl
Sterile H_2O	up to 20 µl

The ligation reaction was incubated overnight at 16 ^{o}C.

For cloning into the ligation independent pET52 3C/LIC vector, primers were designed to introduce a 12 bp 5' extension into the sense and a 14 bp 5' extension to generate the vector specific complementary ends and the PCR product obtained was cloned into a linearized pET52 3C/LIC vector using the 3C/LIC cloning kit (Novagen) according to manufacturer's protocol.

2.4.4 Polymerase chain reaction

The PCR reaction was made in 0.2 ml plastic tubes with reaction volumes of 50 µl. For several or many parallel PCR reactions, a reaction master mix was prepared.

- PCR reaction:

Template DNA (max 0.1 µg/µl)	1 µl
10x polymerase buffer	5 µl
dNTP mix (10 mM each)	1 µl
Forward primer (50 pmol/µl)	1 µl
Reverse primer (50 pmol/µl)	1 µl
Taq/*Pfu* polymerase	0.5 µl
tdH_2O	40.5 µl

- Hot start PCR master mix

mi-Hot *Taq* mix (Metabion) was used to carry out analytical PCR reactions when several PCRs were to perform. The premix saves the times for pipetting of different components and reduces the risk of error in preparation of PCR components due to minimized pipetting. Moreover, the antibodies present in the hot start mix

prevent the polymerase activity at room temperature. The reactions mix was prepared as instructed in the manufacturer's manual.

- *Taq polymerase PCR conditions:*

  Initial denaturation 95 ^{o}C 5 min.

  Three-step cycle:

  1. Denaturation 95 ^{o}C 1 min.

  2. Annealing ($T_{melting}$ – 5 to 10 ^{o}C) 1 min.

  3. Elongation 72 ^{o}C 1min/ kb.

  Number of cycles 25-35

  Final elongation 72 ^{o}C 20 min, store at 4 ^{o}C

- *Pfu polymerase PCR conditions*

  Initial denaturation 95 ^{o}C 4 min.

  Three-step cycle:

  1. Denaturation 95 ^{o}C 1 min.

  2. Annealing ($T_{melting}$ – 5 to 10 ^{o}C) 1 min.

  3. Elongation 72 ^{o}C 2min/ kb.

  Number of cycles 25-35

  Final elongation 72 ^{o}C 20 min, store at 4 ^{o}C

For the thermal cycling of the reaction, a Master cycler from Eppendorf (Eppendorf, Germany) was used. The obtained PCR products were analyzed by agarose gel electrophoresis.

2.4.5 Colony PCR

Colony PCR can be performed to quickly screen plasmid inserts directly from *E. coli* colonies or for the detection of genes from bacterial DNA. A colony of *E. coli* was mixed in 25 µl of 1XTE buffer, vortexed, boiled for 5 minutes and centrifuged for 5 minutes at 13000 rpm. 2 µl of the supernatant served as template for PCR using Hot *Taq* PCR protocol (2.4.4).

2.4.6 Site directed mutagenesis (point mutation)

Site directed mutation is a method to insert aimed point mutations in a plasmid. The point mutation is based on the principle of two complementary primers, which contain the desired mutation as false matching (Kunkel et al., 1985). The mutation primers should contain approx. 15 correct base matches with the template. A PCR reaction is accomplished, with the *Pfu* polymerase, which possesses an extremely low error rate and multiplies the entire plasmid. The PCR results in the production of the desired mutation.

2.5 Construction of cosmid library

2.5.1 DNA preparation

Cosmid vectors can accommodate a fairly large piece of DNA (about 20 to 40 kb) into it. Inserts of sizes smaller than 20 kb can not be packed in phage DNA for transfection into host. This method helps to avoid the cloning of small DNA fragments of less than 20 kb in size, hence the generated libraries are more uniform with large inserts. Large DNA digestion products can be obtained by partial digestion of metagenomic DNA using restriction enzyme *Bsp*143I. Genomic DNA was digested with the enzyme in dilution series 1:10000, 1:20,000, 1:40,000, 1:60,000 and 1:80,000. The reaction was incubated at 37 ^{o}C for 20 minutes, and stopped by adding 2 µl of loading buffer to each cup: the DNA was separated by agarose gel electrophoresis visualized by UV irradiation. The dilution that produced fragments at around 40 kb was determined and the same dilution stock was used to digest the genomic DNA to be cloned into cosmid vector. The digestion reaction was inactivated by heating at 70 ^{o}C for 10 minutes and the DNA was recovered by chloroform extraction and ethanol precipitation.

For cloning of genomic DNA into a cosmid vector, the complementary ends produced by the digestion of the DNA by *Bsp*143I were ligated into the linear cosmid vector pWE15 cut with *Bam*HI.

2.5.2 Packaging of cosmid DNA

Phages are used to deliver DNA to recipient bacteria via a protective capsid. Cosmid DNA with inserts can be packaged into λ phages in order to infect *E. coli*. The Gigapack® III Gold Packaging Extract (Stratagene, USA) was used to package the cosmid with metagenomic DNA inserts into phages using the manufacturer's manual.

2.5.3 Preparation of host bacteria

E. coli VCS257 was used as host strain. 1.5 ml of an overnight culture of VCS257 strain was added to 10 ml LB broth containing 10 mM $MgSO_4$ and 0.2% maltose. Cells were allowed to grow at 37 ^oC for about 4 hours in a rotary shaker (200 rpm) up to an OD_{600} = 1.0. Cells were pelleted by centrifugation at 5,000 rpm at 4 ^oC for 10 minutes and resuspended in sterile 10 mM $MgSO_4$. Cells were diluted to OD_{600} = 0.5 with sterile 10 mM $MgSO_4$.

2.5.4 Transfection of the host bacteria

100 µl of the cosmid packaging reaction was added to 400 µl of SM buffer to make 500 µl packaging mix at 1:5 dilution. It was mixed with 500 µl freshly prepared host cells (at OD_{600} = 0.5) in a 15 ml falcon tube and incubated for 30 minutes at room temperature. 4 ml of LB broth (without antibiotic) were added to it and incubated for about 1 to 1½ hour at 37 ^oC with gentle shaking at every 15 minutes. The culture was then centrifuged at 5,000 rpm at 4 ^oC for 10 minutes. Pellets were resuspended in 1 ml of LB broth, and 100 µl aliquots were spread on plates containing ampicillin (100 µg/ml) and incubated overnight at 37 ℃.

<u>Reagents and solutions</u>

SM buffer:

NaCl	5.8 g
$MgSO_4$	2.0 g
1 M Tris-HCl (pH 7.5)	50.0 ml
Gelatin (2% w/v)	5.0 ml

tdH_2O added to a final volume of 1 liter and autoclaved.

2.5.5 Multiplication and preservation of cosmid clones

Colonies were transferred from agar plates to 96 well micro titer plates containing 100 µl of LB broth with ampicillin (100 µg/ml) using sterile tooth picks. The microtiter plates were incubated overnight at 37 ^{o}C. The next day 5 µl of DMSO was added to each well, mixed properly with a pipette and plates were stored at -80 ^{o}C.

2.6 Transformation

2.6.1 Transformation by heat shock method

2.6.1.1 Preparation of competent *E. coli* cells

Preparation of competent *E. coli* cells in the presence of $CaCl_2$ was performed following the method developed by Mandel and Higa [176]. 200 ml LB medium were inoculated with 1% (v/v) overnight culture of *E. coli* and grown at 37 ^{o}C in rotary shaker up to OD_{600}= 0.4. The culture was kept in ice for 10 minutes. The cells were centrifuged at 4,000 rpm at 4 ^{o}C for 7 minutes. Cell pellets were re-suspended in 10 ml ice-cold $CaCl_2$ solution and centrifuged at 4,000 rpm at 4 ^{o}C for 5 minutes. The pellets were again re-suspended in 2 ml ice-cold $CaCl_2$ solution and the cell suspension was divided in different precooled E-cups into 200 µl aliquots. The competent cells prepared in this way were stored in -70 ^{o}C.

Reagents and solutions

$CaCl_2$ solution: $CaCl_2$ 100 mM, Glycerol 15% (v/v)

2.6.1.2 Transformation using the heat shock method

Competent *E. coli* cells were transformed using the heat shock method [177]. 200 µl competent cells were taken out from -70 ^{o}C and kept in ice for 10 minutes to thaw. 4 µl of ligation (10 - 150 ng DNA) complex was added and incubated for 30 minutes in ice. This mixture was placed at 42 ^{o}C for 90 seconds for the heat shock enabling the cells to uptake DNA. The cells were then incubated in ice for 3 minutes. Further, 800 µl LB broth was added to it and incubated at 37 ^{o}C for 1 hour shaking gently. 50 to 100 µl of transformation mix was poured on LB agar plates with appropriate antibiotics, spread with a sterile spatula and incubated overnight at 37 ^{o}C.

2.6.2 Transformation by electroporation

2.6.2.1 Preparation of electro-competent cells

About 200 ml cells were grown at 30 ^{o}C shaking at 180 rpm up to OD_{600} of 0.5-0.6. The cells were cooled on ice for 15 minutes and centrifuged at 5000 rpm at 4 ^{o}C. The pellet was washed in ice cold distilled water for 3-4 times by centrifuging at 4000 rpm at 4 ^{o}C. After the final wash, cells were resuspended in 2 ml of 10% glycerol (td H_2O) resulting in a total volume of about 3-4 ml and frozen and stored at -80 ^{o}C.

2.6.2.2 Electroporation

The competent cells stored in -80 ^{o}C were thawed on ice for about 15 minutes. About 100 ng (5-10 µl) of linear DNA were added to about 100 µl of competent host cells and kept on ice for about 2-5 minutes. The cells with added DNA were transferred to an electroporation cuvette (0.1 cm) (Biorad) pre-chilled on ice. The cells were pulsed at 25 mF, 2.4 kV and 200 ohm (5 milliseconds) using the GenePulser (Biorad). Immediately after electroporation, 1 ml of chilled LB media was added to the cuvette and the whole content was transferred to 2 ml eppendorf tubes. The tubes were incubated at 30 ^{o}C for 1 ½ hours with gentle shaking. After incubation the content was immediately spread on the LB plates containing appropriate antibiotics.

2.6.3 Selection of recombinant clones

2.6.3.1 Blue white screening

The X-Gal-plate-test was used for the selection of *E. coli* clones that contained recombinant plasmid DNA (cloned into a plasmid cyrrying a cloning site on *lac*Z gene). This test represents a further selection marker apart from the plasmid-coded antibiotic resistance. *E. coli* XL1Blue and DH5α used for the transformations are characterized by deletion of the *lac*Z gene causing a loss to form β-galactosidase. The multiple cloning sites (MCS) of plasmid vector pHSG399 are located in the *lac*Z gene which codes for active enzyme (Vieira and Brass, 1982). IPTG (Isopropyl-β-thiogalacto-pyranoside) induces the formation of β-galactosidase from the *lac*Z gene which converts the indicator compound X-Gal (5-bromo-4-chloro-3-indolyl-b-D-galactoside) in the agar plate into a blue product. The inactivation of *lac*Z site by insertion inhibits the

synthesis of this enzyme and the colony remains white. In this way, a selection of colonies harboring recombinant DNA was possible from the mixture of colonies. The pDRIVE PCR cloning vector is also based on blue white selection. pJET1.2 vector contains a lethal restriction enzyme gene that is disrupted by ligation of a DNA insert into the cloning site. As a result, only bacterial cells with recombinant plasmids are able to form colonies. Recircularized pJET1.2/blunt vector molecules lacking an insert express a lethal restriction enzyme which kills the host *E.coli* cell after transformation.

Reagents and solutions

X-Gal stock solution: X-Gal 200 mg, Dimethylformamide (DMF) to 10 ml

Filter sterilized, wrapped with aluminum foil to protect from light and stored at -20℃).

2.7 Protein chemical techniques

2.7.1 Heterologous protein expression

As a standard approach for protein expression, one liter of TBY media with an appropriate antibiotic was inoculated with 10 ml of an overnight grown *E. coli* culture, transformed with a recombinant gene in a baffled flask and grown at 30 to 37 ^{o}C and 180 rpm. At an OD_{600} between 0.5 and 1, 0.5 mM IPTG was added to the culture medium. After induction, cells were grown overnight at room temperature or at 30 ^{o}C and harvested by centrifugation at 4 ^{o}C for 10 to 12 minutes at 6000 rpm. Growth, induction and expression conditions were modified as necessary (see results). After determining the weight of the pellets, it was stored at -20 ^{o}C or directly used for the isolation of the expressed protein.

2.7.2 Cell lysis and protein purification

The cell pellet was thawed on ice, resuspended in lysis buffer, containing ß-mercaptoethanol (to a final concentration of 0.1%). Pefabloc (final concentration 200 mM) was added to prevent protease activation. The suspension was homogenized in liquid nitrogen with an ULTRA TURRAX (Milian, Geneva, Switzerland). After evaporation of the liquid nitrogen, the homogenate was stored at -80 ^{o}C or thawed on ice. The homogenate was first centrifuged for 25 to 30 minutes at 4 ^{o}C and 20,000

rpm, and the resulting supernatant was cleared by ultracentrifugation for 45 minutes at 4 ^{o}C and 50,000 rpm.

Reagents and solutions

Lysis buffer: 100 mM (pH 8.0) Tris, 600 mM NaCl, 10% Glycerol, 0.1% TritonX-100, 100 µg/ml Lysozyme, pH 8.0

2.7.2.1 His-tag purification

The attachment of a tag of six or ten histidine residues at the N- or C-terminus of recombinant proteins enables affinity purification via immobilized Metal Affinity Chromatography (IMAC). This method is based on the interaction between divalent metal ions (like Co^{2+}, Ni^{2+}, Cu^{2+} and Zn^{2+}) and the unprotonated imidazole ring of histidine. These metal ions have six coordination sites available for interaction with electron rich ligands. Ni-metal chelate affinity resin (Prochem, USA) was used for purification of recombinant protein fused with His-tag, which has tetradentate chelators that bind the metal ions to the chromatographic substrate, leaving two sites available for interaction with histidine residues. The resin was washed 2-3 times with water and further 2 times with wash buffer. Crude lysate solution was added to the resin (1 ml sedimented resin for 5-10 mg His-tagged protein) and was allowed to bind by incubation for about 1 hour by gentle shaking. The resin was transferred to a gravity column and allowed to settle. Then the column was washed with a buffer containing low concentrations of imidazole (20 to 30 mM) for 10 times. Finally, the His tagged protein was eluted with 2 ml of elution buffer. The purified protein, obtained by the His tag affinity chromatography was concentrated in a Centricon device using phosphate buffer.

Reagents and solutions

Wash buffer: 50 mM (pH 8.0) Tris, 300 mM NaCl, 5% Glycerol, pH 8.0, ß-mercaptoethanol (to a final concentration of 0.1%, freshly added) and Pefabloc (final concentration 200 mM, freshly added)

Elution Buffer: Wash buffer with 250 mM imidazole

Phosphate buffer: 10 mM NaCl, 10 mM Na_2PO_4, 10 mM NaH_2PO_4, pH 8.0

2.7.2.2 Gel purification

For protein purification, an Äcta^{TM}basic 10/100 system from Amersham Pharmacia Biotech (GE Healthcare) was used. Äcta^{TM}basic is an automated liquid chromatography system and consists of a compact separation unit and a personal computer running the UNICORN^{TM} control system version 3.0. His-tag purified protein was further gel purified using a Tricorn^{TM} Superdex^{TM}75 column, washed and run at a flow rate of 1 ml/minute with wash buffer (25 mM Tris-HCl, containing 10 mM NaCl, pH 8.0). The purified protein was finally eluted in the same wash buffer and concentrated further using a Centricon device with phosphate buffer.

2.7.3 Inclusion body solubilization and protein refolding

Overexpression of recombinant proteins often results in the accumulation of insoluble protein aggregates called inclusion bodies. In order to recover the protein from the inclusion bodies, the insoluble protein must be isolated, washed, and then solubilized. The final step in isolation of an active protein would be the refolding of the solubilized protein. To solubilize aggregated LOV domain protein from HT-Met1 LOV domain from the inclusion body, CelLytic^{TM} IB, Inclusion Body Solubilization Reagent (Sigma) was used according to the manufacturer's protocol. A dialysis protocol from the manufacturer was followed to refold the protein obtained as unfolded from inclusion body solubilization.

2.7.4 SDS-polyacrylamide gel electrophoresis (PAGE)

Purified protein preparations were separated and analyzed by standard SDS-PAGE [178]. NuPAGE minigels (Invitrogen) were used for SDS-PAGE separation. The precasted gels are 4-12% acrylamide/bis-acrylamide gradient gels. Protein samples were mixed with 0.2 volume of Laemmli-buffer and denatured by heating at 95 ^{o}C for 5 min. Gel electrophoresis was performed in the Minigel Electrophoresis Unit (GE Healthcare) with 1x SDS buffer at 140 V until the tracking dye reached the bottom of the gel. Prestained Protein Marker (Invitrogen Sea Blue 2) was used as the molecular weight marker. After electrophoresis, the gel was fixed by 2 x 5 min incubation in 40% ethanol, 10% acetic acid. Proteins were visualized by staining with Coomassie staining solution for 20 minutes. The gel was then placed in destaining solution (40% ethanol,

10% acetic acid) for 20 minutes and the used destaining solution was replaced with fresh one and left overnight. To prevent cracking of the gel during drying, the gel was washed for 5 min in 10% glycerol and rinsed in water, before allowing it to air-dry.

Reagents and solutions

2x Laemmli-buffer:	Tris 100 mM (pH 6.8), DTT 200 mM, SDS (w/v) 4%, Bromophenol blue (w/v) 0.2%, Glycerol 20%
Coomassie stock:	1 tablet of Phastgel blue R in 200 ml 60% (v/v) methanol
Coomassie stain:	1 volume Coomassie stock, 1 volume 20% acetic acid

2.7.5 Western blot

Proteins from a polyacrylamide gel were blotted onto a PVDF membrane via semi-dry electrophoretic transfer. The transfer sandwich was assembled under transfer buffer in the following way: starting from cathode, 2 pieces of soaking pads, 1 piece of filter paper (of same size as gel and membrane), gel, membrane (immersed in methanol until it does not float anymore, followed by transfer to transfer buffer), filter paper and soaking pads. All air bubbles were carefully removed from both sides of the gel. This sandwich was placed into a transfer cassette, which was then inserted into the tank, filled with transfer buffer. The proteins were transferred at a constant current of 25 V for 80-90 minutes. Pre-stained marker was used to control the completion of transfer. After washing two times for 10 minutes in TBST, and one time for 10 minutes in TBS, the membrane was incubated with the primary antibody (1:1000 diluted in blocking buffer) for one hour at RT. To remove unspecifically bound antibody, the blot was then washed two times for 10 minutes in TBST, and one time for 10 minutes in TBS. It was then incubated with the secondary antibody (1:1000 diluted in blocking buffer) for one hour at RT. Before developing the blot it was again washed four times for 10 minutes with TBST. To develop the blot, it was stained in freshly prepared AP buffer, and the color was allowed to develop in the dark. To stop the chromogenic reaction, the membrane was washed twice in water.

Reagents and solutions:

TBS:	10 mM tris HCl (pH 7.5), 150 mM NaCl
TBST:	20 mM tris HCl (pH 7.5), 500 mM NaCl, 0.05 % (v/v) tween-20
Block Buffer:	3 % (w/v) BSA in TBS buffer

2.7.6 Crystallization

The initial screening for crystallization conditions was done randomly so as to address a variety of screening conditions with varied parameters such as pH value, buffer type, type and concentration of precipitant etc. A series of crystallization buffers, Cryo-I, Crystal Screen, Wizard-I and Wizard-II (Emerald Biosystem, USA) were used for the initial screening in a 96-well plate, each well further containing a reaction well and reservoir buffer. The proteins were mixed to the buffer in 1:1 ratio and incubated at 4 ^{o}C for several weeks. The 96-well plates were studied under the microscope after few days to check whether any of the conditions was optimal for crystallization of the given protein. After the initial crystallization screens, the preparative crystallization was performed in 24 well preparative crystallization plates. The crystallization was performed at ambient light conditions.

2.8 DNA microarray technique

2.8.1 Preparation of microarray slide

Oligonucleotide-based DNA microarrays were prepared by spotting a 15 µM solution (100 mM sodium phosphate buffer pH 7.0) of oligonucleotides (carrying a 5' amino modification and an additional C6 aliphatic linker, Operon Biotechnologies, Cologne) onto CodeLink^{TM} activated slides (GE Healthcare, Freiburg). Covalent coupling of the oligonucleotides to the slide surface was achieved following the instructions of the manufacturer. Spotting was performed with a MicroGrid solid pin microarrayer (Zinsser Analytic, Frankfurt). Spotting and post processing of the spotted slide was carried out at the Department of Genomic and Applied Microbiology, University of Göttingen.

2.8.2 Fluorescence labeling of DNA

DNA from pure culture strains, metagenome or cosmid library were digested using *Alu*I and purified using the QIAquick PCR purification kit (Qiagen, Germany). The volume of the purified DNA was DNA adjusted to 30 µl with tdH_2O, 5 µl of 10x random hexamer, and 5 µl of Klenow buffer was added to it. The mixture was heated at 96 ^oC for 5 minutes to denature the DNA and snap chilled on ice. On ice, 5 µl of 10x dNTP mix (1.2 mM each dATP, dGTP and dTTP and 0.6 mM dCTP) was added to the denatured reaction mixture. About 2.5 µl of Cy5- or Cy3-dCTP and 2 µl of Klenow fragment (-exo) (Fermentas, Germany) were added and the reaction was incubated at 37 ^oC for 3 hours to overnight, during this time the samples were protected from light. The labeling reaction was stopped by addition of 5 µl of 0.5 M EDTA. Purification of the labeled DNA was done using illustra^{TM} CyScribe^{TM} GFX^{TM} purification kit (GE Healthcare, Freiburg, Germany) and finally eluted with tdH_2O. The labeled DNA was vacuum-concentrated using a Speedvac at room temperature and the volume was adjusted with tri-distilled water.

2.8.2.1 Determination of labeling efficiency

In order to determine the incorporation of fluorescent CyDye into the labeled DNA, the absorption pattern of the eluted DNA was recorded at various wavelengths. The absorption maximum of Cy5 labeled DNA is about 650 nm and using the extinction coefficient (250,000 M^{-1} cm^{-1}), the total amount of Cy5 molecules incorporated into DNA can be calculated. 10 µl of the purified, labeled DNA was diluted with tdH_2O and placed in a micro cuvette (d=1 cm). The absorption spectra were recorded from 700 to 500 nm using UV/VIS spectroscopy. The amount of the Cy5 incorporated into DNA can be calculated as given:

$$\text{pmoles of Cy5 in sample} = [A/E.d)] \times Z \times \text{dilution factor} \times 10^6$$

where, A = absorption of Cy5 at 650 nm

E= the extinction coefficient (250,000 M^{-1} cm^{-1} for Cy5)

Z= volume of labeled DNA after purification (in µl)

D= thickness of the cuvette (1 cm)

A similar relation can be used to calculate the incorporation of Cy3 molecules in the DNA by recording the absorption at 550 nm and using the extinction coefficient value of 150,000 M^{-1} cm^{-1} for Cy3.

2.8.3 Hybridization of the labeled DNA with microarray slide

The hybridization of labeled DNA to the microarray slide was performed in a GeneMachines® Hyb4 station (Genomic Solutions, MI, USA), controlled by the programmable Hyb4 editor software. Before hybridization, the labeled DNA was denatured at 95 ^{o}C for 5 minutes, snap chilled in ice, mixed with 3 volumes of Tom Freeman hybridization buffer (40% formamide, 5x Denhardt's solution, 5x SSC, 1 mM Na-pyrophosphate, 50 mM Tris pH 7.4 and 0.1% SDS), and kept warm at the initial hybridization temperature. Microarray slides were warmed at 75 ^{o}C for 2 minutes in the hybridization chamber and samples were injected at 65 ^{o}C. A 16 hours hybridization program was used. After the hybridization steps, the slides were immediately washed using wash solution 1 (1x SSC, 0.2% w/v SDS) at 45 ^{o}C, wash solution 2 (0.1x SSC, 0.2% w/v SDS) at 42 ^{o}C and wash solution 3 (0.1x SSC) at 42 ^{o}C. All wash procedures were carried out at a set program (flow for 10 s, hold for 20 s, 2 cycles) and slides were drained for 40 s.

Detailed hybridization program:

- Step 1 O-Ring conditioning, 75 ^{o}C, 2 min, Agitate: No
- Step 2 Introduce sample, 65 ^{o}C
- Step 3 Hybridization 1 65 ^{o}C, 4 hours, Agitate: Yes
- Step 4 Hybridization 2 60 ^{o}C, 4 hours, Agitate: Yes
- Step 5 Hybridization 3 55 ^{o}C, 4 hours, Agitate: Yes
- Step 6 Hybridization 4 50 ^{o}C, 4 hours, Agitate: Yes
- Step 7 Wash slides 1 42 ^{o}C, Flow for 10 s, hold for 20 s, 2 cycles
- Step 8 Wash slides 2 40 ^{o}C, Flow for 10 s, hold for 20 s, 2 cycles
- Step 9 Wash slides 3 40 ^{o}C, Flow for 10 s, hold for 20 s, 2 cycles
- Step 10 Drain slides 40 s

After draining, slides were taken out immediately from the hybridization chamber and further dried by centrifugation (600 rpm for 5 minutes).

Reagents and solutions:

Denhardt's solution:	PVP-40 2% (w/v), Ficoll 2% (w/v), BSA (fraction V) 2% (w/v), Filter sterilised and stored at -20 ^{o}C
20x SSC stock solution:	3.0 M NaCl, 0.3 M Na-citrate (pH 7.0)

2.8.4 Scanning and data analysis

The microarray slides were scanned immediately after hybridization on a GenePix 4100A scanner (Axon Instruments Inc, CA, USA) using the GenePix Pro6 software. The photomultiplier tube (PMT) gain was set uniform at 600 V for the 635 nm laser and at 540 V for the 532 nm laser throughout all experiments. The images were saved as 16-bit TIFF files. A grid of individual circles defining the location of each probe on the array was superimposed on the image to designate each fluorescent spot to the related probe. The signal intensity (SI = F635 mean - B635 for Cy5 labeled DNA, F532 mean - B532 for Cy3 labeled DNA), standard deviation of background and signal to noise ratio (SNR = signal intensity - background/standard deviation of background) were calculated using GenePix Pro6 and data were transferred to Microsoft Excel for further processing. The signal intensity and SNR from four replicate data sets were averaged to represent the image intensity and SNR of a particular probe.

2.9 Dot blot analysis

2.9.1 DIG-labeled probe synthesis

After a positive probe is determined from the microarray, the same oligonucleotide sequence can be used to synthesize a DIG-labeled probe for oligonucleotide-based dot blot hybridization. The plasmid clones can be hybridized with the DIG labeled oligonucleotide primer. In this test, a single probe can be hybridized to a number of individual clones simultaneously which is more cost effective than the microarray hybridization, and in addition it is time saving. After a positive hybridization signal was detected from a microarray probe, a 5' end DIG-labeled probe

(Dig-ACCGGCTACCGGGCGGAAGAGGTGCTGGGCCGCAACTGCCCTTCCTGCATTCG) was synthesized (Metabion, Germany).

2.9.2 Hybridization of the target to the DIG-labeled probe

Heat-denatured DNA (about 15 ng plasmid) was applied to a positively charged nylon membrane (Roche, Germany), air-dried and placed on filter paper, wetted in the denaturation solution (1.5 M NaCl, 0.5 M NaOH) for 5 min, in neutralisation solution (1.5 M NaCl, 0.5 M Tris/HCl, 1 mM EDTA pH 7.2) for 1 min, and in fixation solution (0.4 M NaOH) for 20 min. The membrane was incubated in hybridization buffer (5x SSC, 0.1% w/v N-Lauryl sarkosin, 0.02% SDS and 1% w/v blocking reagent) for 30 minutes at 42 ^{o}C, then the DIG-labeled oligonucleotide probe (4 pmol/ml) was added to the solution and hybridized overnight at 42 ^{o}C. The membrane was washed two times each with wash buffer 1 (2x SSC, 0.1% w/v SDS) at room temperature and wash buffer 2 (0.5x SSC, 0.1% SDS w/v) at 42 ^{o}C with gentle shaking.

2.9.3 Detection

The membrane was incubated with blocking solution (1% blocking reagent in DIG1 buffer) for 30 minutes at room temperature, which was replaced with 45 mU/ml of antibody conjugate (Roche) in DIG1 buffer, and incubation was continued for another 30 minutes. Afterwards, the membrane was washed with DIG1 buffer (Tris HCl 0.1 M, NaCl 0.1 M, pH 7.5) and equilibrated with detection buffer (Tris HCl 0.1 M, NaCl 0.1 M, pH 9.5). Finally, the membrane was incubated in 10 ml of detection buffer containing 200 µl of NBT/BCIP stock solution (Roche, Germany) in the dark until the hybridization spots were visible.

2.10 Bioinformatic applications

2.10.1 DNA sequencing

DNA sequencing was performed at the Automatic DNA Isolation and Sequencing Unit (ADIS), Max Planck Institute for Plant Breeding Research, Cologne. End sequencing of the plasmid clones was performed using standard primers. Further sequencing of DNA that was more than 1 kb in size was carried out by primer walking. Primer walking is a method for sequencing DNA fragments between 1.3 and 7 kb. The

DNA of interest may be either a plasmid insert or a PCR product. The initial sequencing is performed from each end using either universal primers or designated primers. In order to completely sequencing the region of interest, new primers from the prior identified region are designed according to the available partial sequence information.

2.10.2 Analysis of sequence data

The evaluation of the sequence data of from the clones were accomplished by using nucleotide-nucleotide blast (BLASTN), translated query vs. protein database (BLASTX) service offered by National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov).

2.10.3 Database mining for LOV domain

The BLAST search was performed against all deposited metagenomic proteins and nucleotide sequences available in NCBI, IMG/M [179] and CAMERA [180] databases using the LOV domain from *B. subtilis* as a reference sequence. For the hits obtained from the nucleotide databases, corresponding nucleotide sequences that showed features to LOV domains in BLAST search were extracted and open reading frames were identified using ORF finder of NCBI. When the BLAST search was done against the same sample deposits in different databases, the resultant sequences were cross checked to avoid redundancy and only one accession was taken in such cases. The available protein sequences obtained from the BLAST hits and the translated ORFs were analyzed manually to see whether they contain the residues of LOV domain proteins essential for flavin- (FMN) binding, photoadduct formation and functioning of blue light sensing LOV proteins. Information about the geographical and physicochemical conditions of the source environment, from where the DNA material originated, was extracted from the corresponding databases. Information about the total number of ORFs in each metagenome sample was obtained from the database.

2.10.4 Sequence alignment and editing

Alignment of two sequences was carried out using specialized BLAST-bl2seq of NCBI. Multiple sequence alignment was performed using clustalW2 web-based server at http://www.ebi.ac.uk/ [181]. BioEdit 7.0 [182] was used for editing of multiple nucleic acid and amino acid sequences.

2.10.5 Protein domain scan

Protein domains were determined with ScanPro [183] at http://www.expasy.org and SMART v5 [184] at http://smart.embl-heidelberg.de/, using both as complementary tools. The domain search was carried out at default parameters.

2.10.6 Phylogenetic tree

Phylogenetic trees were constructed with PhyML 3.0 [185] using the Jones-Taylor-Thornton (JTT) rate matrix [186]. For a highly refined tree an eight gamma estimated rate category and a bootstrap value of 100 replicates were used. The resultant trees were obtained as the maximum likelihood (mlk) output. All the trees were visualized using iTOL [187]. Each LOV domain sequence derived from the metagenome present in the phylogenetic tree was BLASTed to the non redundant protein sequences (nr_proseq) in NCBI database with e value > -15 to determine its nearest match. The environmental LOV domain sequences were assigned to the nearest group as determined by the BLAST hit. The percent similarity to the nearest match was used as a criterion to define the novel sequence. As a very conservative parameter, a sequence identity of less than 79% was assigned as a novel sequence.

2.10.7 Primer design

Most of the cloning primers were designed manually. However, few web-based tools were used to design primers to a target-specific region in certain DNA molecules for PCR based detection and mutagenesis. Primer 3 [188], a web based primer design tool was used to design specific primers to amplify a target region of a DNA. The CODEHOP web based server [189] was used to design the consensus-degenerate hybrid primers. Block processing was carried out at "Blocks WWW Server" of Fred

Hutchinson Cancer Research Center [190]. Primers for site directed mutagenesis were designed using QuickChange® Primer Design Program (Stratagene).

2.10.8 Other resources

Operon prediction and regulon prediction was carried out using www.http://www.microbesonline.org/, a product of the Virtual Institute for Microbial Stress and Survival (VIMSS) [191].

Thermodynamic properties of the DNA hybridization were calculated using "The DINAMelt web server" [192] at http://dinamelt.bioinfo.rpi.edu/. pDRAW32, a free DNA analysis software, was used for restriction analysis and virtual DNA cloning.

The PSIPRED secondary structure prediction method [193] was used to predict the secondary structure of a protein. The three-dimensional structure modeling and visualization was performed using SPDB 4.01 [194] and SWISSMODEL server http://swissmodel.expasy.org//SWISS-MODEL.html [195].

For the prediction of the theoretical molecular mass and the extinction coefficient of the proteins, the web based server at www.expasy.org was used.

2.11 Spectroscopy Methods

2.11.1 Ultraviolet/Visible (UV/VIS) spectrometry

UV-VIS spectroscopy is the measurement of the wavelength and the intensity of absorption of near-ultraviolet and visible light by a sample. The concentration of an analyte in solution can be determined by measuring the absorbance at certain wavelengths and applying the Beer-Lambert law:

$A = \varepsilon \times c \times d$

where, A is the measured absorbance, ε is a wavelength-dependent absorption coefficient (M^{-1} cm^{-1}), c is the concentration (M) of the solution, and d is the path length (in cm). All the measurements were done with a Shimadzu (UV-2401PC) spectrophotometer. The buffer or the eluent used to prepare the dilution of the sample was applied as a reference solution.

2.11.2 Fluorescence spectrometry

For the fluorescence analysis of LOV domain proteins two types of fluorescence measurement platforms were used. Regular fluorescence analysis was carried out in quartz cuvettes using Cary Eclipse Fluorescence spectrophotometer (Varian, USA) at our institute. To screen samples in 96 well micro-plates, TECAN Infinite® fluorescence plate reader (TECAN, Switzerland) at IWW Water Centre, Mülheim an der Ruhr (Germany) was used.

Chapter 3

Development of a high throughput technique as an approach to detect LOV domains from a metagenome

3.1 Function-based screening to detect LOV domains from library clones

As an initial approach to detect LOV domains of BL photoreceptors from a metagenomic library, a functional screen was applied. The rational for the experimental set up was based on the fluorescent properties of LOV domains that are solely dependent on the bound FMN chromophore. Several samples were used as positive controls for standardization of method and equipment calibration: the BL receptor YtvA as purified protein, crude cell lysate of an over night culture of *E. coli* BL21 cells containing *ytvA*-constructs in an expression vector and an overnight culture suspension of *E. coli* BL21 cells with an *ytvA*-construct (Table 3.1). The *E. coli* XL1Blue cells without the LOV domain expression plasmid served as negative control. Large insert clone libraries generated from diesel biofilm isolates (*Pseudomonas aeruginosa* strains) were used as reference samples, since sequencing of *P. aeruginosa* genomes has documented the absence of LOV domains on their chromosome [196]. The reference sample containing about 200 large insert clones generated from *P. aeruginosa* in pWE15 cosmid vector were available from the metagenomic library collection of University of Hamburg [197].

Table 3.1 Constructs and cells used for standardization of fluorescence based detection

Sample	Sample details
Pure YtvA	Purified YtvA protein (initial stock of 10 mg/ml)
E. coli-ytvA	BL-receptor *ytvA* gene cloned in pET28 expression vector and transformed into *E. coli* BL21 cells
E. coli XL1	*E. coli* XL1 cells without any construct
P. aeruginosa cosmid library in *E. coli* VCS2577	200 clones of various strains of *P. aeruginosa* derived from diesel biofilm

The samples were prepared in 96-well micro-plates. The micro-plates containing the samples were stored in dark until scanning. Scanning was performed in a TECAN Infinite® fluorescent plate reader after an excitement with a light beam of 486 nm. A linear increase of the fluorescent signal was observed for a dilution series of the pure protein in order of increasing concentrations (Figure 3.1, Table 3.2a). Similarly, the analysis of the crude extract of the expression cells containing *ytvA*-pET28 also showed a linear increase in the fluorescent signal in order of increasing concentration. However, a cell suspension containing the *ytvA*-pET28 did not produce a fluorescent signal that would be sufficient to discriminate the cells against a background of negative control clones from *P. aeruginosa.* On the other hand, few of the cells carrying *P. aeruginosa* clones produced a fluorescent signal that was of almost the same intensity as that one produced by the positive control of the *E. coli-ytv*A construct (Table 3.2b).

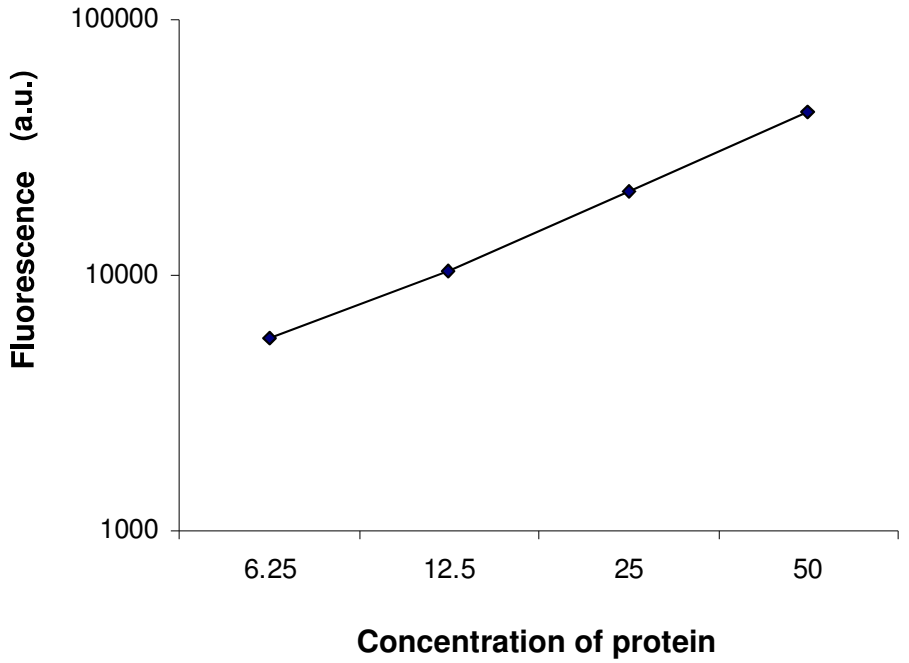


Figure 3.1 Increase of the fluorescence signal intensity of the pure YtvA protein with increase in protein concentration. The signal obtained from undiluted protein (100% concentration) was saturated.

In case of the purified protein the fluorescence decay was also observed dependent on the exposure to light. However, a significant fluorescence decay was not observed in the case of a cell suspension in vivo (cells with *ytv*A-construct) (Table 3.2a). The fluorescence decay of the pure protein exposed to light was dependent on the duration of the exposure. After a short room light exposure for 5, 10, 15 and 20 minutes the fluorescent value of the pure protein was found to decrease in accordance with the time of exposure. The linear relation of fluorescence with respect to the concentration of the protein was still observed (Figure 3.2).

Table 3.2 Fluorescence signal recorded from a 96 well micro-plate containing standard protein, cell suspensions and sample cell suspensions. (a) Column 1 and 2 of the micro-plate. Column 1 contains the pure YtvA protein and column 2 contains the *E. coli* BL21 cells with an *ytv*A-construct in the expression vector pET28a. Values from columns 1 and 2 were used as standards for calibration. Corresponding rows in column 1 and 2, marked with "0" (rows A and B, respcetively) refer to blank solution, which was used for preparing the standard solution. Phosphate buffer (10 mM NaH_2PO_4 and 10 mM NaCl, pH 8.0) was used as blank for protein and LB medium without cell suspension was used as blank for cell suspension. (b) Columns 3 to 8 contain an overnight culture suspension of the large insert clones of *P. aeruginosa* strains in the cosmid vector pWE15 maintained in *E. coli* VCS257. Sample cells with cosmid clones that produced high fluorescence signal comparable to the 75% standard cell suspension are highlighted in light blue.

(a)

		Conc.	YtvA Protein	Cells with *ytvA-construct*
		<>	1	2
Concentration	A	0	519	3049
	B	0	501	3177
	C	50	23926	3931
	D	50	24161	4192
	E	75	35782	*4828*
	F	75	35672	4497
	G	100	46629	4607
	H	100	48418	4713

(b)

	Sample cells					
	3	4	5	6	7	8
A	4515	4264	4247	4113	4112	4270
B	4370	3988	4228	3704	4112	3771
C	3952	3819	4219	3964	3756	3887
D	4135	3457	4243	3834	4167	3992
E	3979	4190	4180	3753	4102	3960
F	3979	4333	4110	3858	4151	3777
G	4106	4296	4329	3838	4059	3910
H	4487	4222	4258	4269	3931	4011

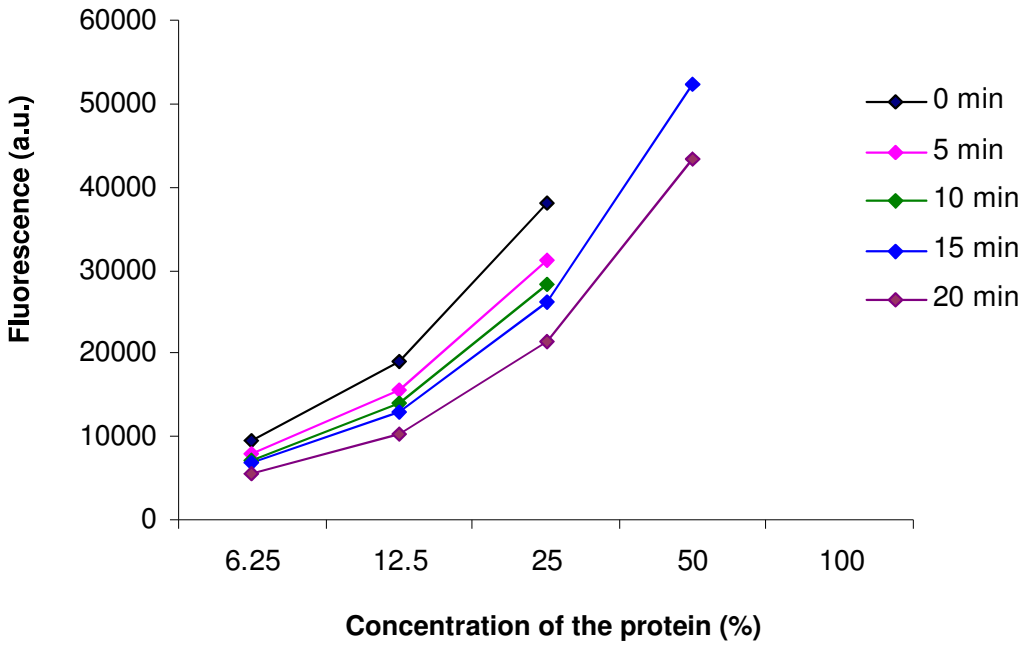


Figure 3.2 Fluorescence intensity decrease of the pure protein in different concentrations upon exposure to the room light for various time periods (see color coding in the figure legend). The signal obtained was saturated for undiluted protein (100% protein) in all exposure conditions. 100% concentration refers to 10mg/ml YtvA protein, which was diluted in phosphate buffer (10 mM NaH_2PO_4 and 10 mM NaCl, pH 8.0) to produce the given concentrations.

Following, it was attempted to test the potential of the crude lysate of expression cells to probe a crude lysate-based assay system for detection of LOV domain-containing clones based on the fluorescence signal intensity. Crude lysates of the *E. coli-ytv*A cells were prepared from overnight grown cultures induced with 0.5 mM IPTG. Similarly, crude lysates were prepared from three of the *P. aeruginosa* clones (PAII7B, PAIV6H and PAIV12A) that produced the highest fluorescent signals in the previous scanning. This was to test whether the crude-lysate based scanning could be applied for the detection of LOV domain carrying clones. However, in fluorescent scanning the crude lysates from *P. aeruginosa* clones produced stronger signals than that of *E. coli*-ytvA, despite no LOV domain being present (Table 3.3).

Table 3.3 Fluorescent signal obtained from the crude lysate of three selected clones.

Crude Protein/Control sample	Fluorescence intensity
YtvA construct	6822
PA II 7B	7164
PA IV 6H	7076
PA IV 12A	7511

3.2 Construction of a LOV domain-specific DNA microarray

The preliminary approach (3.1) to establish an in vivo fluorescence-based assay to detect LOV domain containing clones from a metagenomic library was found to be not efficient, which inferred that a more robust technique was required. In this background, a DNA microarray approach based on the conserved motif of LOV domains was regarded as a possible solution on the application of high throughput library screening.

The working principle of a DNA microarray is based on duplex formation (hybridization) between nucleic acid molecules to their complementary strands. In this technique the target DNA molecules hybridized to the complementary probes are detected using fluorescence labels. There are a number of advantages for such an approach: already the presence of a small fragment of a complementary DNA from a gene is sufficient for positive hybridization. Unlike in expression based systems, the DNA microarray-based screening of the library clones does not depend on the expression of a target gene. Moreover, an expression-based system requires the presence of a complete gene and accessory expression mechanisms in order to successfully express in a heterologous background, whereas a DNA microarray can detect a clone even containing only a fragment of a target gene.

For generating a LOV domain microarray, sequence information of possibly all putative LOV domain-containing genes is required. For that purpose, a BLAST search was performed against all genomic and metagenomic protein or DNA sequences

deposited in the NCBI database using the LOV domain sequence from *Bacillus subtilis* (YtvA) as query. Following, the protein sequences obtained from the sequence database or determined from the ORF finder of NCBI (in the case of nucleotide database) were aligned, identifying core regions from the consensus sequences that contained most of the conserved residues. As a result, a highly conserved motif consisting of eighteen most conserved amino acid residues of a LOV domain was selected to design the oligonucleotide probes for the LOV domain microarray (Figure 3.3). This part of the protein encompassed the core motif that includes the conserved, photo-adduct forming cysteine. This approach yielded a set of sequences that are 54 nucleotides in length.

Each single designed oligonucleotide was blasted in the NCBI database to verify its specificity and was also compared among each other to see if complete identity exists between any of the oligonucleotides. In the event of complete congruence between two or more oligonucleotides, only one of them was taken as a probe for the LOV microarray. Altogether, 149 54-mer oligonucleotide probes were designed and synthesized. The term “probes” has been used here to indicate the oligonucleotides immobilized on the microarray chip (slide). The variation of 149 selected (partial) sequences were derived from the LOV domain proteins of 45 proteobacteria, 8 firmicutes, 10 cyanobacteria, 1 planctomycetes, 3 plants and 4 archaea. The remaining 78 probes originated from metagenomic gene fragments. The GC content within the 54-mers ranged from 25 to 74 % with a T_m ranging from 66 to 86 ^{o}C. Similarity among the individual 54-mers probe sequences was $\leq$ 96%.

Each of the probes carried a 5’-amino group as modification to allow covalent attachment to the solid surface of the CodeLink activated microarray slide (2.8.1). In a first batch the probes were spotted on duplicate on the microarray slide, later all probes were spotted on quadruplicate. Spotting in four allows four independent hybridizations and readings of intensity, yielding more reliable data.

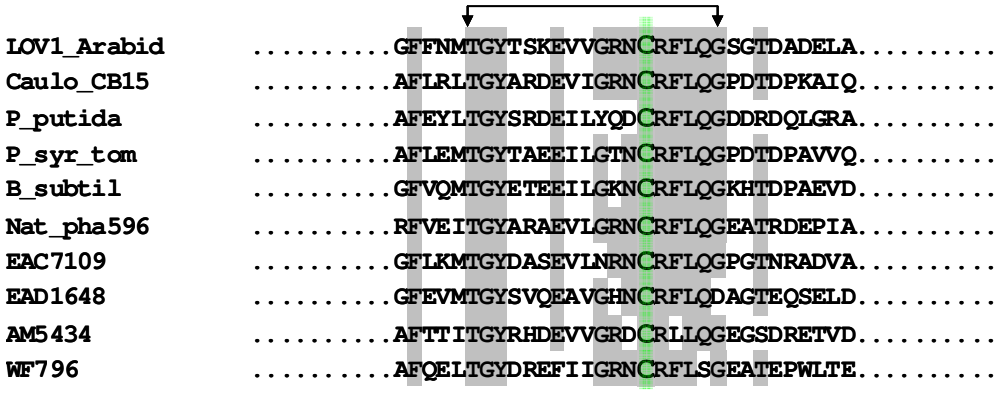


Figure 3.3 Sequence alignment of the core region exemplified for ten LOV domains. Indicated is the range of eighteen amino acids which were used to design 54-mer oligonucleotides to be spotted for the microarray approach. The fully conserved, functionally essential cysteine (involved in photo-adduct formation) is highlighted in green; LOV1_Arabid: *Arabidopsis thaliana* Phot1-LOV1; Caulo_CB15: *Caulobacter crescentus* CB15; P_putida: *Pseudomonas putida* KT2440, SB1; P_syr_tom: *Pseudomonas syringae* pv. *tomato*; B_subtil: *Bacillus subtilis*; Nat_pha596: *Natronomonas pharaonis* DSM2160; EAC7109: Sargasso metagenome EAC77109; EAD1648: Sargasso metagenome EAD41648; AM5434: Acid mine drainage metagenome (gi: 41580434); WF796: Whale fall metagenome (gi: 60178796). The horizontal bar indicates the region of highest conservation being used for oligonucleotide design for microarray construction.

Oligonucleotide probes from corresponding conserved regions of each amino acid sequence shown in figure 3.3 are given in table 3.4. A complete list of 149 probes used in LOV microarray is given in appendix (Appendix A).

Table 3.4 Oligonucleotide probes derived from the conserved region of ten amino acid sequences given in figure 3.3

Probes	Sequence
LOV1_Arabid	actggttacacttccaaagaagtcgtcggcagaaactgccgatttttacaagga
Caulo_CB15	acgggctatgcccgcgacgaagtgatcggccgcaattgccggttcctgcaggga
P_putida	accggctactgcgccgacgatattctctatcaggactgccgttttcttcagggc
P-syr_tom	acggggtatacggccgaagaaatccttggcaccaattgccggtttcttcagggg
B_subtil	accggctacgagaccgaggaaatttttaggaaagaactgtcgcttcttacagggg
Nat_pha596	accggctatgctcgtgctgaggtcctcggccggaactgccgcttcctccaaggt
EAC7109	acggggtacgacgcgagcgaggtcctgaaccgaaactgccggttcctgcagggt
EAD1648	acggggtacagcgtgcaagaagccgtgggtcacaactgtaggtttttgcaagac
AM5434	acggggtaccgccatgatgaggtcgtggggcgggactgccgacttctccagggg
WF796	accggctatgatcgtgagttcatcattggccgaaactgccgctttttgtcgggc

3.3 Optimization of labeling and hybridization conditions

3.3.1 Preparation and labeling of target DNA

The term “target” is used here to indicate the sample DNA that is to be labeled and hybridized to the array. After few trials, a standard procedure for DNA labeling was developed as described in methods (2.8.2). Genomic DNA was digested with a 4-base cutter restriction endonuclease (*Alu*I) before labeling to produce fragments of 500 to 1000 bp, since fragmentation of high molecular weight DNA improves labeling and hybridization efficiency. It is also helpful to reduce steric hindrance and target secondary structure formation [156]. Plasmid clones (size about 5-6 kb) were labeled in both forms, without or after restriction digestion. We did not find any significant difference when using unfragmented or fragmented plasmid DNA on the labeling efficiency. Even though, for the purpose of uniformity all the plasmid or cosmid clones were digested before the labeling. Cy5 dye incorporated in the DNA was determined on the basis of the absorbance at 650 nm (Figure 3.4). It was found that incorporation of more than 80 pmol of Cy5 to the labeled DNA was required for environmental applications.

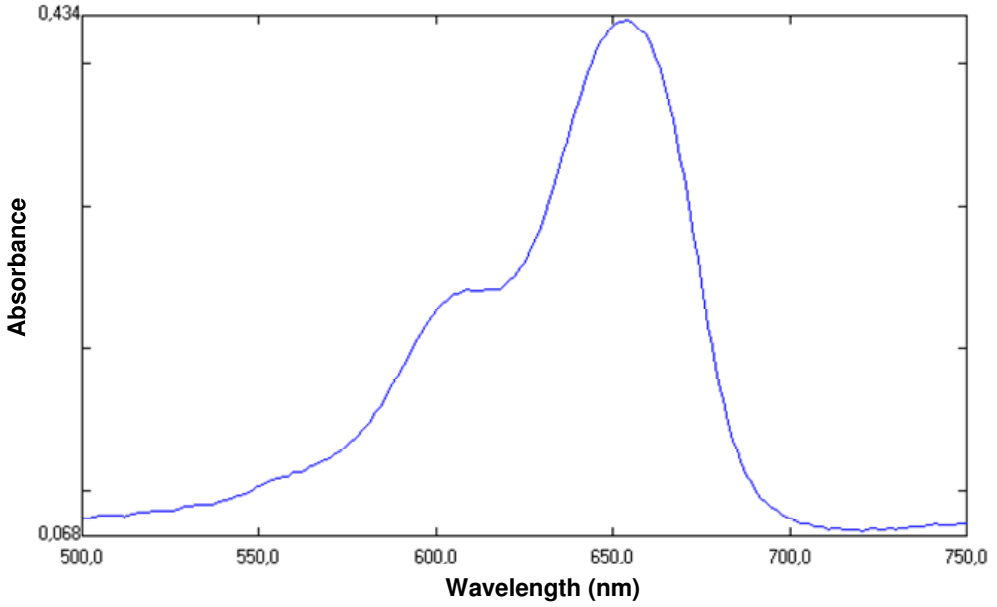


Figure 3.4 Absorption spectrum of Cy5 labeled DNA for the determination of incorporated fluorescent material

3.3.2 Optimization of hybridization temperature

One of the complex features of the LOV microarray was the presence of probes with a wide range of T_m (from 66 ^oC to 86 ^oC) that required a wide range of the theoretical hybridization temperature. Initially, several experiments were carried out at a uniform hybridization temperature using a target DNA. The melting temperature (T_m) of the 54 bp stretch of the target DNA, which is the same as T_m of the related probe, was calculated as 76.42 ^oC. An overnight (16 hours) hybridization at 50 ^oC was carried out, followed by a 55 ^oC repetition. Positive hybridization was detected in either case and there was no unspecific hybridization detected. However, due to a variation of 20 ^oC among the spotted probes, a single hybridization temperature can not properly satisfy the hybridization requirement of all probes when unknown heterogeneous DNA has to be screened. In the absence of satisfying hybridization temperature, a positive hybridization result could not be achieved.

As an alternative to the uniform hybridization temperature, a step down temperature program was tested. The program consisted of four different hybridization temperatures, each for four hours, starting from the highest one and lowering down in the order of decreasing temperature at each step (2.8.3**,** Figure 3.5).

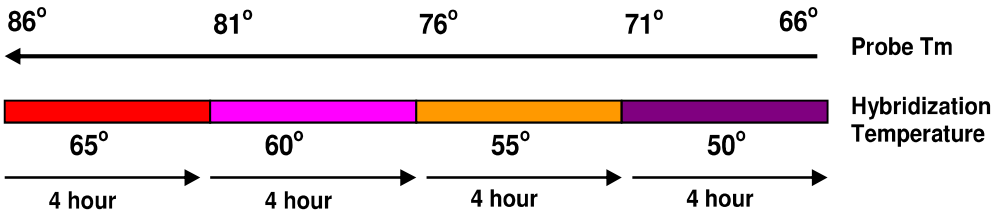


Figure 3.5 Optimized hybridization temperature program to screen heterogenous unknown target DNA using the LOV microarray constructed from oligonucleotide probes of various melting temperature (T_m).

The resulting hybridization images were of similar intensity to that of the uniform temperature program and no unspecific signal was detected. Afterwards, the step down hybridization temperature was applied in all experiments.

3.4 Hybridization with perfect match

In order to standardize and calibrate the newly constructed LOV-microarray, several control experiments were performed using the target DNA from pure cultures and plasmid clones that bear a 100% match to some of the probes spotted on the array. The target DNA was derived from putative and characterized LOV domains from different genomes.

3.4.1 Hybridization with pure culture DNA isolates

Genomic DNA from *Pseudomonas syringae* pv. *tomato* and *Bacillus subtilis* were used as positive control target DNA. *P. syringae* pv. *tomato* carries a single LOV-domain gene (PSPTO_2896) on its chromosome (positions 3260020 to 3258416) which in its consensus sequence is identical to the oligomer # 5038 spotted on the array. Similarly, the 54-mer oligonucleotide probe (# 5029) spotted on the array is

identical to a consensus region of the LOV-domain gene (*ytv*A: positions 3106210 to 3106995 on the chromosome) from *B. subtilis* genome. For the purpose of clarity the term “perfect match” has been used here to indicate targets that are 100% identical to the probes spotted on the array.

The genomic DNA was labeled using the Cy5-fluorescent dCTP and hybridized to the microarray individually or in various combinations. At first, genomes from *P. syringae* pv. *tomato*, *B. subtilis*, *P. syringae* pv. *syringae*, *E. coli* and *P. aeruginosa* were labeled with Cy5 and hybridized to the microarray individually. *E. coli* and *P. aeruginosa* genomes do not contain a LOV domain, which as expected did not produce any hybridization signal. DNA from both *E. coli* and *P. aeruginosa* genomes was thus taken as negative control. Both *P. syringae* pv. *tomato* and *B. subtilis* produced positive hybridization signals from the corresponding probes when hybridized individually or in combination of both (Figure 3.6).

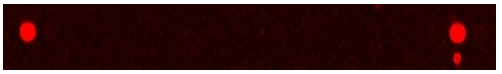
(a)  (b) 

Figure 3.6 Fluorescence image obtained after the hybridization of Cy5 labeled genomic DNA from (a) *P. syringae* pv. *tomato* position # 5038 and (b) *B. subtilis* position # 5029.

3.4.2 Hybridization with plasmid clones

Different LOV domain-containing genes from several prokaryotes and a metagenome sample cloned in plasmid vector (pET28a) were then used to test the hybridization efficiency of the LOV microarray. The plasmid, being of smaller size than genomic DNA, contains a higher concentration of the target gene even in low concentration of clones. Plasmid clones containing BL receptor genes from *B. subtilis*, *Caulobacter crescentus*, *P. syringae* pv. *syringae*, *P. syringae* pv. *tomato* and *Synechococcus elongatus* PCC7942 were available from our laboratory stock (2.1.2). One of the shotgun clones (AAGA01023330) that contains a putative LOV domain from the Whale fall metagenome project [124] was obtained by courtesy of Susan

Green Tringe (Joint Genome Institute/Diversa). The putative BL-coding frame was PCR-amplified from the Whale Fall shotgun clone (AAGA01023330) and cloned into a linearized pET28 expression vector using *Nde*I and *Xho*I restriction endonucleases. The plasmid clones containing the BL receptor genes from *B. subtilis*, *C. crescentus*, *P. syringae* pv. *syringae* and *P. syringae* pv. *tomato* were labeled with Cy5 or Cy3 dCTP and hybridized to the LOV array. All of the perfect match targets produced a positive hybridization signal (Figure 3.7). Similarly, the plasmid clones containing BL receptor genes from *S. elongatus* PCC7942 and from the Whale Fall metagenome also produced a positive hybridization signal as expected (Figure 3.7e and f).

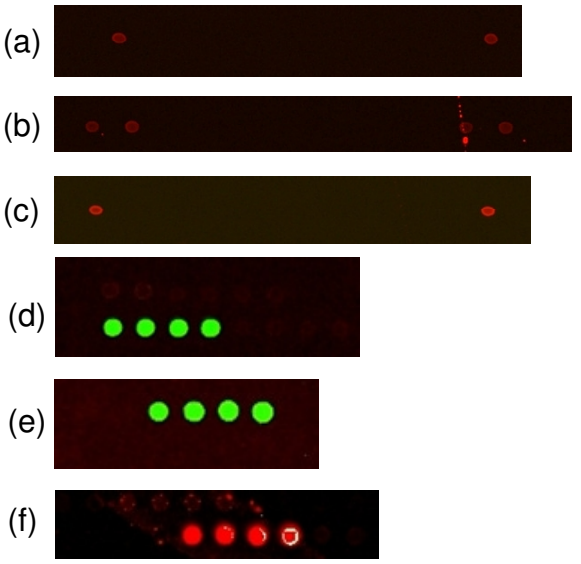


Figure 3.7 Positive hybridization signals produced by the perfect match target plasmid clones. The image corresponds to (a) *B. subtilis* (*ytv*A), (b) *P. syringae* pv. *syringae* (left), and *P. syringae* pv. *tomato* (right), (c) *Caulobacter crescentus*, (d) *P. syringae pv. syringae*, (e) putative BL receptor gene from Whale fall metagenome (AAGA01023330), and (f) *S. elongatus* PCC7942. The target DNA from (a), (b), (c) and (f) were labeled with Cy5 dCTP. The target from (d) and (e) were labeled with Cy3 dCTP. The probes on (a), (b) and (c) were spotted on the microarray slide on duplicate. The probes on (d), (e) and (f) were spotted on quadruplicate.

3.5 Specificity and sensitivity evaluation

3.5.1 Specificity and sensitivity evaluation with genomic DNA

Specificity is determined on the basis of the target and probe hybridization. It is difficult to predict the limit of cross hybridization for different probes without an experimental determination. Absolute specificity is the condition in which there is a positive hybridization only between probes that are completely identical to the targets. Although absolute specificity on a microarray is not possible due to a certain probability of unspecific hybridization, it is always important to determine the level of unspecific hybridization among the probe-target before proceeding to the sample screening. The hybridization of the labeled target from *P. aeruginosa* and *E. coli* did not produce any detectable signal from any of the probes. Genomic DNA from *P. syringae* pv. *tomato* and *B. subtilis* produced hybridization signals only from their perfect match probes, and there was no detectable unspecific hybridization from any of the rest of the probes.

Sensitivity is the parameter which determines the minimum amount of DNA required to generate a detectable hybridization signal. A serial dilution of the genomic DNA from *P. syringae* pv. *tomato* from 1000 ng to 500 ng, 100 ng and 50 ng was labeled with Cy5 dCTP and hybridized to the microarray. For all labeling reactions the total concentration of applied DNA was maintained to 2 µg using the genomic DNA from *P. aeruginosa* as background. In this experiment, about 50 ng of the genomic DNA from *P. syringae* pv. *tomato* against a background of 1950 ng of *P. aeruginosa* DNA was still detectable (Figure 3.8).

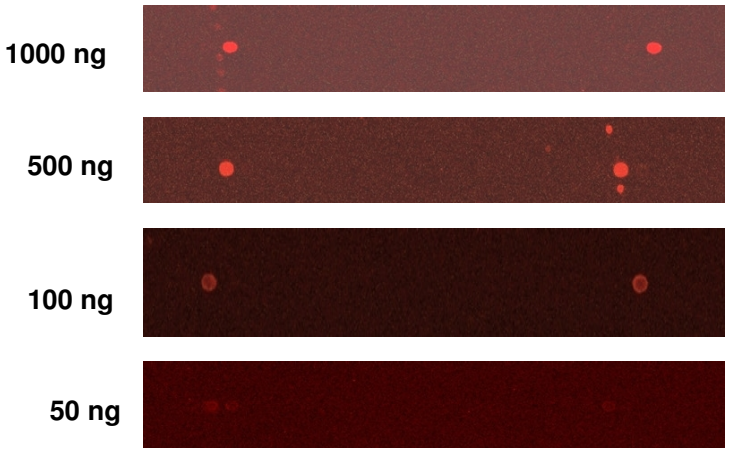


Figure 3.8 Fluorescence image obtained after the hybridization of different concentrations of labeled target DNA from *P. syringae* pv. *tomato* to the specific probe (probe no. 5038, probes in duplicate)

The image intensity values obtained from hybridization of different concentrations of target genomic DNA are given in table 3.5, which exhibits a large intensity variation dependent on the concentration of the DNA.

Table 3.5 Hybridization signal intensity obtained from the perfect match probe # 5038 using different concentrations of genomic DNA from *P. syringae* pv. *tomato*.

Target DNA Conc.	Average Signal intensity
1000 ng	1022.5
500 ng	272.5
100 ng	99.5
50 ng	14.5

3.5.2 Sensitivity evaluation with perfect match plasmid clones

After sensitivity determination for genomic DNA (3.5.1), the detection limit of a perfect match plasmid clone was studied, for which a plasmid clone containing the BL receptor gene from *P. syringae* pv. *tomato* was used. The DNA was probed in serial dilutions of 1:10, 1:100, 1:1000 and 1:10000 (v/v) in which the approximate concentration of the clone was determined to be 297 ng, 24.3 ng, 1.78 ng and 0.27 ng.

In all dilutions, about 4 to 5 µg of salmon sperm DNA was used as background DNA. The mixture of the DNA was labeled with Cy5 and hybridized to the array. The dilution of 1:10000 that contained approximately 0.27 ng of the plasmid DNA with a LOV domain was detectable. Only few unspecific hybridization signals were obtained in this test which was solely from salmon sperm DNA. When the clone was again mixed with *P. aeruginosa* and *E. coli* genomic DNA and hybridized, the unspecific signals were no more visible. Sequence information from salmon sperm DNA was not available to analyze the target probe identity. The salmon sperm DNA was not used as background in any further experiment because of the lack of sequence information and production of unspecific hybridization signal.

3.6 Hybridization with cosmid/fosmid libraries

The main objective of the developed microarray was to screen large-insert metagenomic libraries. Before proceeding to experiments with the metagenomic samples, the procedure was calibrated and the detection efficiency was determined using large insert reference libraries.

3.6.1 Construction of a cosmid genomic library from *P. syringae* pv. *tomato*

As a positive background, a cosmid library from *P. syringae* pv. *tomato* genomic DNA was constructed. A cosmid clone ideally can accommodate between 25 kb and 45 kb of insert DNA. Before ligation into a cosmid vector, it is necessary to prepare the correct DNA fragment to be ligated into the vector such that the construct can be optimally packed into a phage and transfected to a host. To determine the optimal size of the DNA, a small portion of genomic DNA was subjected to partial digestion in different dilution series with restriction endonuclease *Bsp*143I (2.5.1) before ligation. After the determination of the optimal dilution of the restriction enzyme, the digested and purified fragments produced by the ideal concentration of the enzyme was ligated into the vector pWE15 (linearized with *Bam*HI).

After the phage packaging and transfection to the host, several colonies were seen on the agar plates containing selective medium. Randomly selected 20 colonies were digested using different restriction enzymes (6-base cutters) to determine the

insert size. All of the tested clones contained insert DNA with average sizes in the range of 35 to 45 kb. The clones were picked and transferred to 96 well micro-plates for multiplication and storage. The library consisted of 440 clones with approximately 2.7-times the genome coverage (provided each clone would contain an average of 40 kb of insert DNA).

3.6.2 Initial test with cosmid/fosmid DNA libraries

Initially, *P. syringae* pv. *tomato* and four other cosmid/fosmid libraries originated from *Burkholderia glumae*, *Erwinia* sp., *P. aeruginosa* and *Rhizobia* sp. were hybridized to the microarray. Hybridization was carried out with each individual library in a group of 200 clones. The clones from each library were pooled separately in a group of 200 clones using 20 ng of each clone, concentrated, labelled with Cy5 dCTP and hybridized. Hybridization experiments verified that only the corresponding oligomer (# 5038) gave a positive signal when hybridized to a fraction of *P. syringae* pv. *tomato* library that contain 200 cosmid clones. No false positive or unspecific hybridization signal was detected from any of the other libraries. This test was performed to track the hybridization pattern of the each fraction of the library clones to ensure the accuracy of further calibration using these reference libraries.

3.6.3 Relation between number of background clones and signal intensity

To further test the applicability of the LOV-microarray for the screening of DNA libraries and determine its sensitivity, a positive test clone was probed against an increasing amount of background DNA. A cosmid test clone was constructed by cloning a fragment of the BL receptor gene from *P. putida* KT2440 into pWE15. A 377 bp region containing the LOV core region was amplified from the *P. putida* genomic DNA using a pair of extended primers with *Bam*HI recognition sites on each 5'-end. The amplified DNA was cloned into the linearized cosmid vector pWE15, produced by *Bam*HI digestion. The test clone was named as pWPPLOV. The region from positions 115 to 168 of the insert in pWPPLOV was identical to oligonucleotide # 5040.

As background, cosmid libraries from *P. syringae* pv. *tomato*, *Burkholderia glumae*, *Erwinia* sp., *P. aeruginosa* and *Rhizobia* sp. were used. Only *P. syringae* pv.

tomato carries a LOV domain containing gene in the library corresponding to the oligomer # 5038.

For the detection of a specific target clone among a heterogeneous mixture of clones from cosmid libraries, 10 ng of the test clone pWPPLOV from *P. putida* KT2440 (identical to oligomer # 5040) were mixed with an increasing number of Cy5-labeled background clones from different cosmid libraries. Starting from 200 clones as background, the number of background clones was increased by 200 clones in each consecutive test reaching up to 1200 clones, maintaining the concentration of each individual background clone about 20 ng. The starting experiment “I” (Figure 3.9) contained the test clone pWPPLOV and 200 out of the 440 clones (randomly picked) of the *P. syringae* pv. *tomato* cosmid library. Experiment “II” contained 400 clones from *P. syringae* pv. *tomato* library (additional 200 clones to the experiment “I”). The third experiment (“III”) was set up like # “II”, except that in addition 200 clones from the *P. aeruginosa* library were added. Experiment “IV” contained all components of # “III” plus additional 200 clones from the *Erwinia* library. Experiments “V” and “VI” had additional 200 clones from the *Rhizobia* library (“V”), and in addition 200 clones from the *Burkholderia* library (“VI”), such that in this last experiment the positive clone was probed against a background 1200 additional cosmid clones.

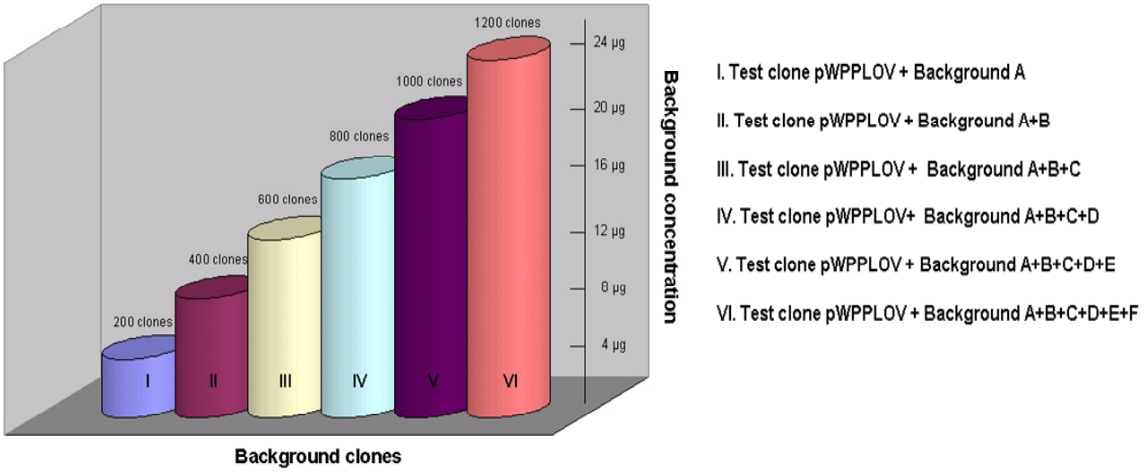


Figure 3.9 Number and concentration of clones used in consecutive tests to determine the sensitivity of the LOV microarray to screen a pool of library clones. 10 ng of Cy5 labeled test clone pWPPLOV was mixed with an increasing number of labeled heterogeneous background clones. In each consecutive test the number of background clone was increased by 200 clones. Background "A" consists of 200 cosmid clones of *Pseudomonas syringae* pv. *tomato*, background "B" indicates addition of other 200 cosmid (i.e., a total of 400) clones of *P. syringae* pv. *tomato*, background "C" consists of 200 clones from the cosmid bank of environmental isolates of *P. aeruginosa*, added to "B", background "D" consists of additional 200 cosmid clones from *Erwinia* sp. cosmid bank, background "E" consists of additional 200 clones from *Rhizobia* sp. cosmid bank, background "F" consists of again 200 clones added from *Burkholderia glumae* cosmid bank. Background B contains a positive control cosmid clone (a single cosmid clone from the library of *P. syringae* pv. *tomato* which is present randomly in the pool and is a perfect match to spot 5038) while background A, C, D, E and F are the negative clones that did not produce any hybridization signal in previous tests.

As expected, the pWPPLOV test clone produced a bright image (SNR 155, SI 6178) even in the presence of high levels (1200 clones corresponding to a total of about 24 µg DNA) of unspecific background DNA added into the hybridization assay (Figure 3.10). It is readily seen that the positive test clone gave four clearly identified signals (white arrows), and the positive clone included in the 400 background cosmid

clones from the *P. syringae* pv. *tomato* library yielded also four signals (yellow arrows). Variation in the image intensity in relation to the background clones for the test clone (pWPPLOV) and the positive control clone from *P. syringae* pv. *tomato* is given in figure 3.11.

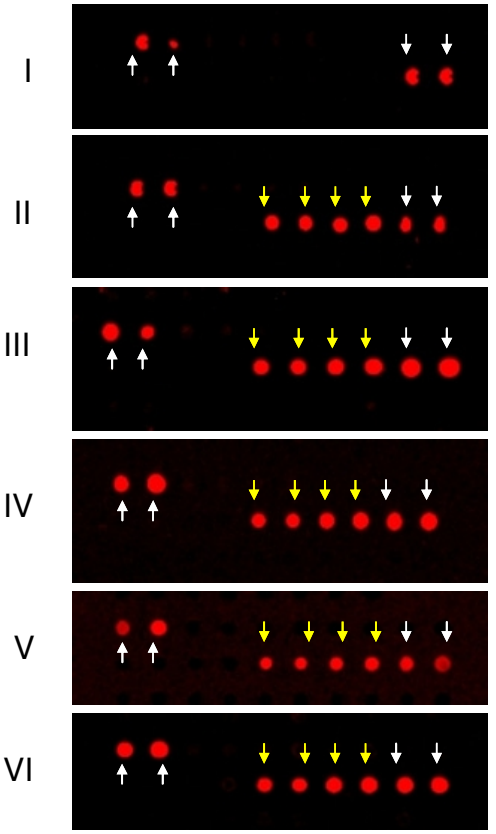


Figure 3.10 Fluorescent image obtained from the hybridization of 10 ng of the Cy5 labeled test clone (pWPPLOV) to the LOV-microarray in a different number of background clones. The number of the heterogeneous background clones starting with 200 at I is increased by 200 each time reaching upto 1200 in VI. (white arrows: spots # 5040 corresponding to the cosmid test clone pWPPLOV, yellow arrows: spots # 5038 corresponding to the *P. syringae* pv. *tomato* LOV gene, a positive control, from the background cosmid library clones). The image refers to experimental arrangement in I, II, III, IV, V and VI as shown in figure 3.9.

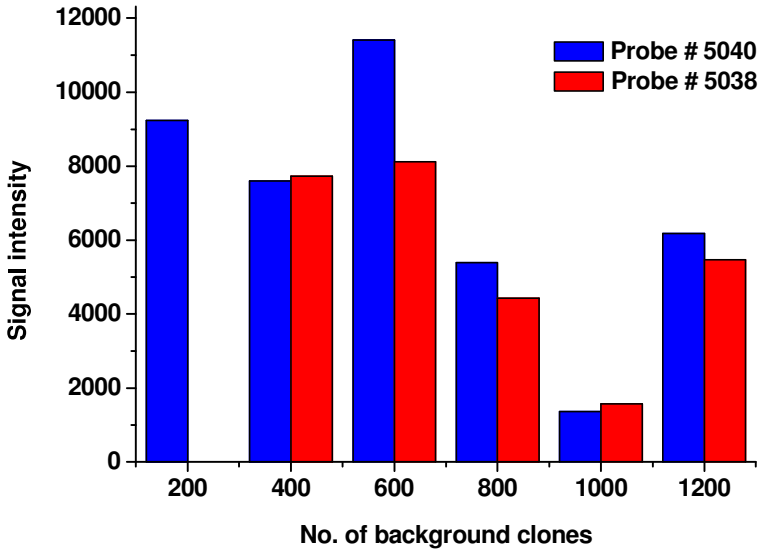


Figure 3.11 Graphical representation of the signal intensity obtained during the addition of the increasing number of background clones to a test clone. The blue bar indicates the signal obtained from probe # 5040 (related to test clone pWPPLOV) and the red bar indicates the signal obtained from probe # 5038 (the background clone from *P. syringae* pv. *tomato* library). For the experiments with 600 clones the signal intensity is high, even though the same experimental conditions were used. Probably the labeling condition and the amount of the successfully labeled positive target DNA have played role in different experiments.

3.6.4 Relation between DNA concentration and signal intensity

The relation between the concentration of a particular positive target clone and the hybridization image intensity of the LOV microarray was determined by lowering the concentration of the test clone pWPPLOV. Decreasing DNA amounts of the cosmid clone pWPPLOV ranging from 10 to 0.25 ng were mixed with 1,000 heterogeneous background cosmid clones (each clone approximately 10 ng), labeled with Cy5-dCTP and hybridized to the microarray. The background cosmid clones were from *P. syringae* pv. *tomato*, *Erwinia* sp., *P. aeruginosa*, and *Rhizobia* sp. Hybridization images obtained from the series of experiments using reduced amounts of positively determined DNA is shown in figure 3.12. Even an amount of 0.25 ng DNA for the test clone produced a bright signal (SI= 311 and SNR= 8.5) (Figure 3.12e).

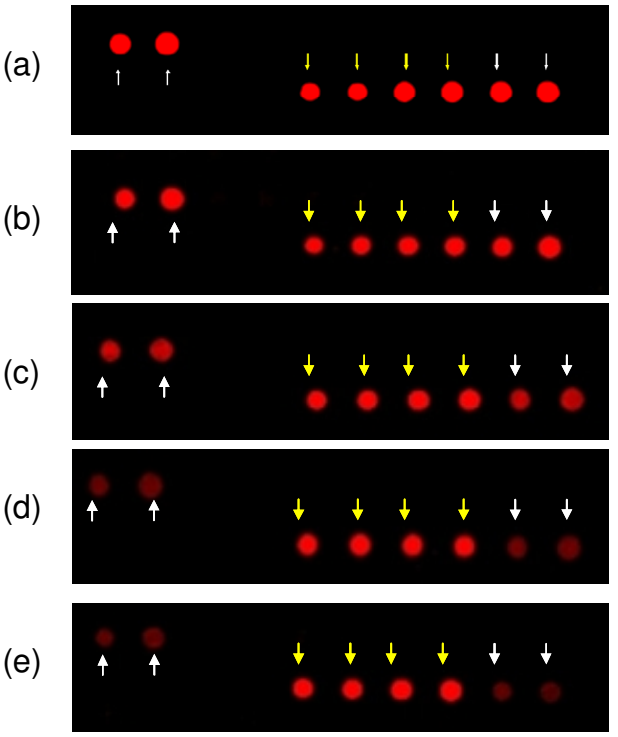


Figure 3.12 Fluorescent signal obtained after the hybridization of decreasing concentrations of test clone pWPPLOV (shown by white arrow) to the LOV microarray. Images (a), (b), (c), (d) and (e) correspond to 10 ng, 5 ng, 1 ng, 0.5 ng and 0.25 ng of Cy5 labeled test clones. About 1000 heterogeneous clones were used as background. The signals indicated by yellow arrows correspond to a randomly included perfect match clone corresponding to PSPTO-LOV present in the background library from *P. syringae* pv. *tomato*.

The hybridization results indicated a linear correlation (R^2= 0.99) between the concentration of Cy5-labeled test DNA and the fluorescent signal intensity (Figure 3.13a). The signal intensity produced from the hybridization of two targets to their corresponding probes is shown by a bar chart (Figure 3.13b). The signal from the positive background clone, the concentration of which was maintained uniform in all experiments, gave a standard deviation value of 970 (mean value 5285).

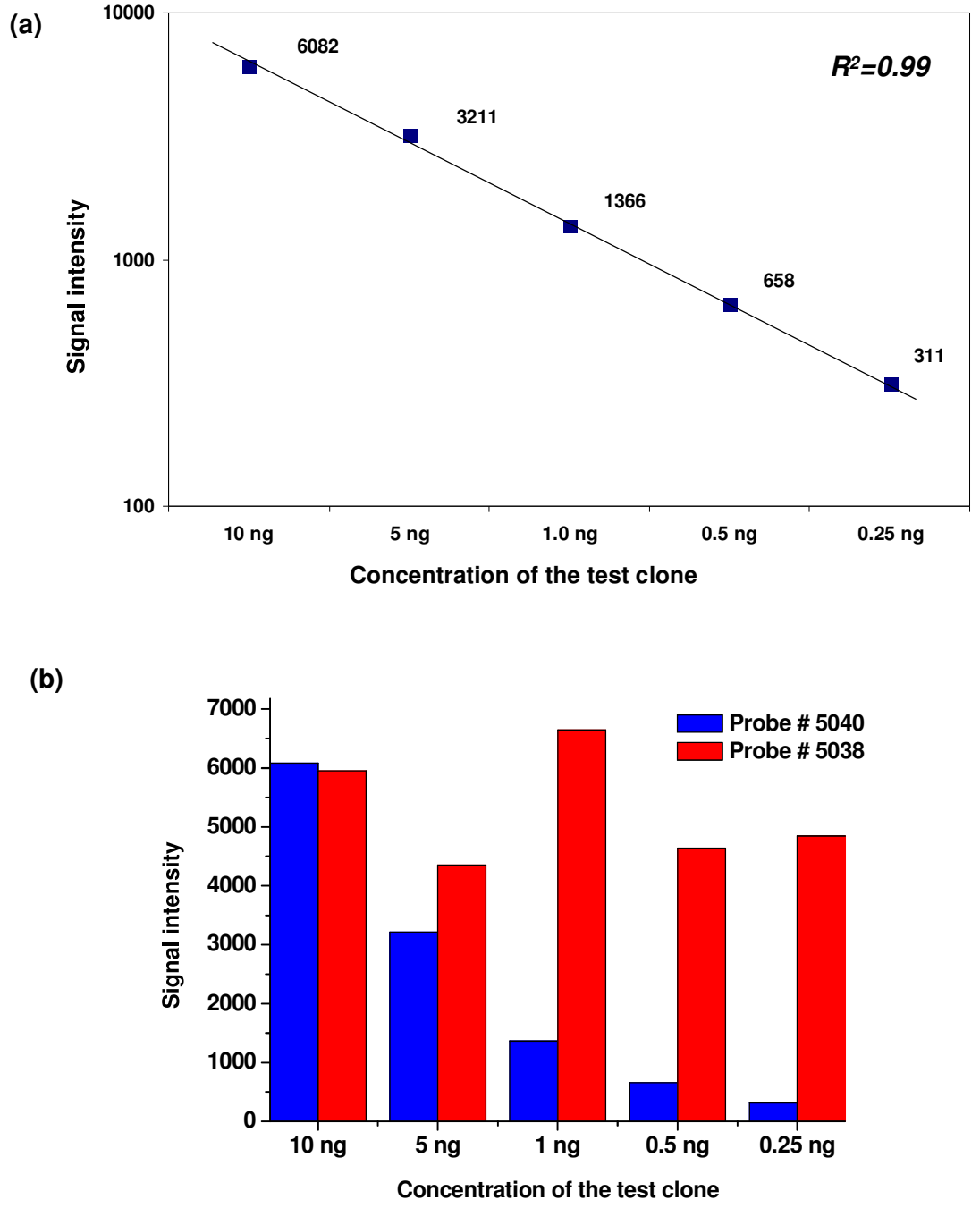


Figure 3.13 (a) Logarithmic relation between the concentration of a test clone and the hybridization signal intensity. (b) Signal intensity of the decreasing concentration of the test clone pWPPLOV in the uniform background of 1000 cosmid clones. The bar shown in blue corresponds to the signal intensity obtained from the test clone

(corresponding to # 5040). The bar shown in red corresponds to the signal intensity of a clone present in the background cosmid library (corresponding to # 5038).

3.6.5 Specificity evaluation with cosmid DNA

The LOV microarray slide consists of unique sequences having less than 85% similarity between most of the spotted probes. There are only few probes (about 10) which share a sequence identity of more than 85% (but less than 100%) among each other.

Pseudomonas putida KT2440 is one of the available targets in order to test for specificity. One of the LOV domains from *P. putida* KT2440 corresponding to probe # 5040 has a sequence similarity of 96% to the probe (# 5041) from *P. putida* F1B (Figure 3.14a). A probe derived from *Chromobacter salixigens* (# 5060) shares 81% identity to the probe # 5040 (corresponding to *P. putida* KT2440) (Figure 3.14b).

(a)

```
Pputida2440_5040      ACCGGCTACTGCGCCGACGATATTCTCTATCAGGACTGCCGTTTTCTTCAGGGC 54
PputidaF1B_5041       ACCGGCTATTGCGCCGACGATATCCTCTATCAGGACTGCCGTTTTCTTCAGGGC 54
                      ******** ************** ******************************
```

(b)

```
Pputida2440_5040      ACCGGCTACTGCGCCGACGATATTCTCTATCAGGACTGCCGTTTTCTTCAGGGC 54
Chrom_sal_5060        ACCGGCTACAGCGTCGACGAGATCCTTTACCGTGACTGCCGTTTCCTGCAGGGC 54
                      ********* *** ****** ** ** ** *  ************ ** ******
```

Figure 3.14 Alignment of three probes spotted on the microarray to show their sequence identity/similarity. P. putida2440_5040 refers to the probe derived from *P. putida* KT2440 which is a perfect match to # 5040. (a) Alignment of the perfect match probe # 5040 to the mismatch probe # 5041 (*P. putida* F1B). (b) Alignment of the perfect match probe # 5040 to the mismatch probe # 5060: Chrom_sal_5060 (derived from *Chromobacter salixigens*). The stars denote identity.

About 500 ng of the test clone (pWPPLOV) was labelled with Cy5 and hybridized to the microarray. A quite strong image was obtained from the specific probe (# 5040). The second probe with 96% identity also gave a good signal (Figure 3.15), but the signal intensity was about 17 times weaker than the perfect match probe (Table 3.6). There was no detectable signal from probe 5060 which shares only 81%

sequence identity to the perfect match. This experiment was carried out in duplicate and there was no significant difference in the signal intensity of the perfect match in the two different hybridizations.

Table 3.6. Signal intensity from the perfect match and the mismatch probe obtained after the hybridization of Cy5 labeled pWPPLOV clone. The signal intensity given here is the average of the probes spotted in duplicate on the microarray.

Probe	Image intensity in experiment 1	Image intensity in experiment 2	Sequence identity to the target
5040	21590	22442	100%
5041	1292	2614	96%
5089	23.5	28	38%

However, an unspecific signal was obtained from probe # 5089 (corresponding to Sargasso probe EAK54400). Hybridization with about 1 µg of pWE15 DNA (without any insert) also produced a weak signal from probe # 5089. The image from this probe was assigned as false positive and was not considered as a positive hybridization match in later hybridizations.

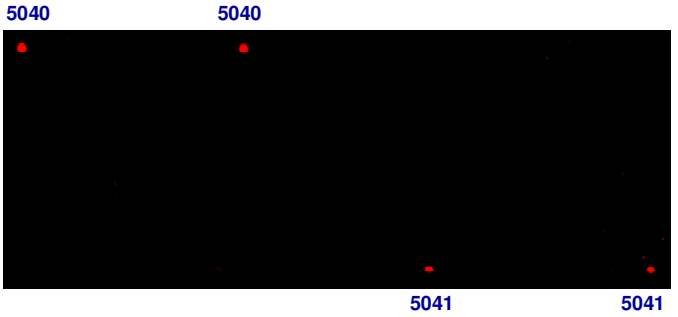


Figure 3.15 Fluorescent image obtained from the hybridization of the Cy5-labeled test clone pWPPLOV to the microarray slide. The test clone contains the perfect match for the probe # 5040 which shares 96% identity to the probe # 5041.

To test the relation between the concentration of the DNA and hybridization specificity, the data from the experiment in which the test clone was hybridized to the

microarray on decreasing concentration in a constant background of DNA were analyzed. Along with the perfect match probe, the test clone pWPPLOV had produced signals from probe # 5041 which is 96% identical to the test clone. The signal intensities from a mismatch probe in various concentrations of target DNA was compared to that of the perfect match probe. The signal intensity from the mismatch probe (probe # 5041) at a concentration of 10 ng was about 50% that of the perfect probe. The signal intensity of the mismatch probe also showed a relative decrease with respect to the decrease in the concentration of the target (Figure 3.16). The relative signal intensity, calculated on the basis of the signal intensity for the perfect match, which was taken as 1.0, was found in the range of 0.30 to 0.68 when various concentrations (0.25 ng to 5 ng) of the target clone were probed to the microarray.

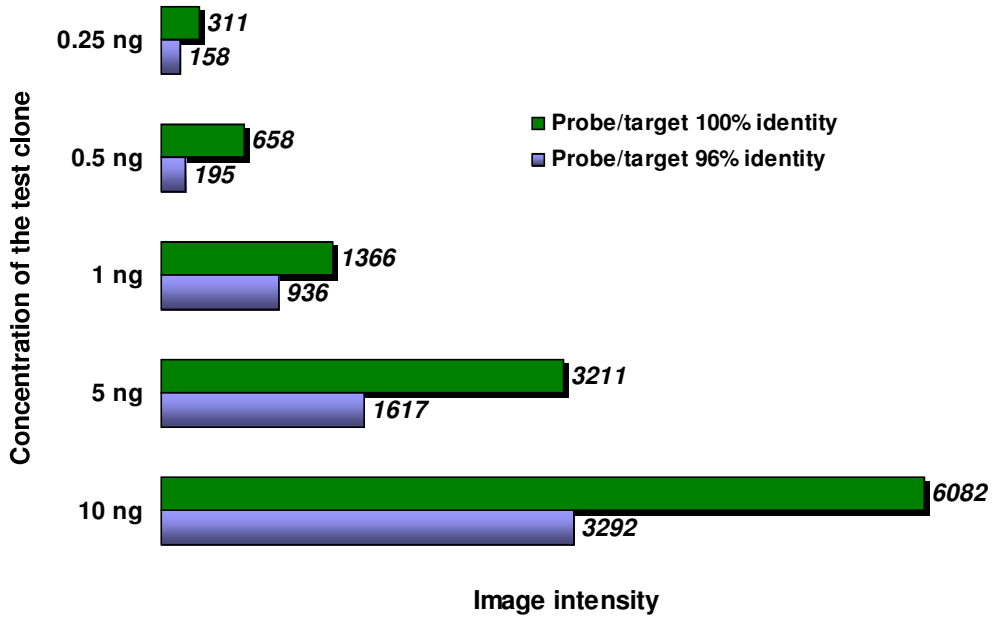


Figure 3.16 Signal intensities of the mismatch probe and perfect match probe obtained in the hybridization of different concentrations of the cosmid clone containing a target clone.

3.7 Screening of metagenomic libraries using LOV microarray

3.7.1 Drinking water biofilm library

After testing the reliability of the LOV microarray with genomic DNA, plasmid clones, cosmid clones and large insert reference library clones, it was used to screen metagenomic libraries. About 1600 cosmid clones from a metagenomic library prepared from a drinking water biofilm [198] were screened to search the presence of LOV domains. The cosmid library was derived from three biofilm samples collected from the rubber coated valve submerged in the drinking water networks of a town in the northwestern part of Germany in the state of North Rhine-Westphalia. The screening was done in groups of 800 clones in one batch. One of the batches showed a few weak positive hybridization signals to the LOV microarray. The signals were obtained from probes # 5058, # 5084 and # 5083. The signal from # 5058 was slightly stronger than the other ones, but SNR was >1(but <3), other images had SNR <1.0. The probes that gave a positive hybridization were derived from *Ralstonia solanacearum* (# 5084, # 5083) and *Novosphingobium aromaticivorans* (# 5058, Figure 3.17). After arriving to 100 clones from the group of 800 clones by subsequent reduction, the signal could not be traced further. It should be noted that a different batch of extracted DNA was used to prepare the pools of 100 clones in the final hybridization because of the consumption of the already extracted DNA in previous screenings.

Figure 3.17 Fluorescence image obtained after the hybridization of Cy5 labeled clones from a drinking water biofilm metagenome to the LOV microarray. The image shown corresponds to the probe # 5058 (derived from *Novosphingobium aromaticivorans*)

3.7.2 Thermophilic soil enrichment metagenomic library

In order to demonstrate the applicibility of this approach and to identify novel LOV domain proteins from metagenome samples, another environmental library

consisting of 2,500 cosmid clones was screened. The library was derived from a thermophilic enrichment culture that had originally been inoculated with soil samples from the botanical garden of the University of Hamburg and had been cultivated at 65 ^{o}C for several months in the laboratory. Each clone carried an insert of approximately 35-45 kb and 85% of all clones carried inserts. A phylogenetic analysis of the microbial community obtained prior to the library construction indicated a moderate biodiversity with mainly microbes from proteobacteria (alpha-, beta-, and gamma-subgroups) and some Flavobacterial species present in the consortium (Appendix B). The clones were initially screened in pools of 400 clones (as Cy5-labeled cosmid DNA), and few bright spots were identified after the hybridization test. One of these signals (# 5084) was traced down by subsequent hybridizations using DNA from pools with decreasing numbers of clones starting with pools of 200 clones and going down to two clones. Detection was then continued by using two different dyes to label the isolated cosmid DNAs. After several (nine) rounds of hybridization, the corresponding cosmid clone was detected. The detected clone was named pWThLOV. The clone reproducibly hybridized with the oligomer spot # 5084.

3.8 Subcloning of the cosmid clone and detection of the target

The identified cosmid clone contained an insert of about 40 kb. After the detection of this positive cosmid clone, the insert was fragmented by restriction enzymes and cloned in plasmid vector pHSG399 (Figure 3.18). This step was carried out to minimize the size of the insert for convenient handling and faster detection of the target gene. The clone was separately digested with the restriction enzymes *Bam*HI, *Eco*RI and *Sal*I, yielding respective populations of restriction fragments to determine the appropriate fragment sizes for subcloning. *Bam*HI and *Sal*I generated optimal fragments (Figure 3.18a) which were less than 15 kb in size.

The product from *Sal*I restriction digestion was purified directly from the reaction and randomly ligated with *Sal*I digested pHSG399 without separating the fragments from the mixture. After the transformation of the *Sal*I fragments into *E. coli* cells, 34 white colonies were picked randomly from the plates containing blue white colonies of the transformant cells to extract the plasmid DNA. The extracted plasmids were

digested with *Sal*I and inserts were analyzed to determine unique fragments. After the insert analysis, 19 possible unique clones were selected from these *Sal*I- generated clones. The fragments generated by *Bam*HI were separated by an agarose gel and bands were cut at four positions; all four gel pieces were purified separately. The purified DNA fragments were cloned into the plasmid vector pHSG399 and transformed separately. The largest gel fragment (the upper band in Figure 3.18) produced two types of inserts, slightly differing in size. Six colonies generated from the second, four colonies generated from the third and three colonies generated from the fourth gel fragments (first, second, third and fourth gel fragments refer to the boxes I, II, III, IV in Figure 3.18a) were checked for their insert size. Among 15 *Bam*HI generated colonies analyzed, only eight unique clones were left after discarding the inserts of similar size from two or more plasmids.

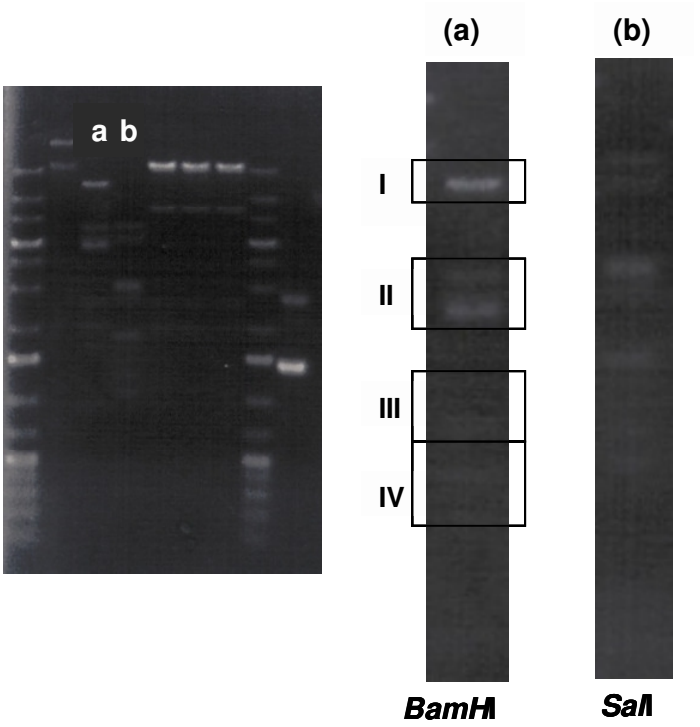


Figure 3.18 Restriction digestion of the positive cosmid clone (pWTHLOV) using various enzymes to determine the optimal fragments for subcloning. (a) *Bam*HI generated fragments: the areas marked by boxes were cut from the gel separately. (b) *Sal*I generated fragments: all the *Sal*I generated fragments were purified together from the reaction and ligated as such to the vector.

Dot blot hybridization was used to identify the positive plasmid clones from the group of unique subclones generated by *Bam*HI and *Sal*I digestion. In a dot blot assay, the plasmid clones immobilized on positively charged nylon membrane can be hybridized with a DIG (digoxigenin) labeled oligomer. In this test, a single probe can be hybridized to a number of individual clones simultaneously which is more cost effective and time saving than the microarray hybridization. Probe # 5084 was the one that gave the highest hybridization signal intensity in the initial screens from the pool of clones. The 54 bp oligonucleotide corresponding to the probe # 5084 on the microarray was synthesized and labelled with DIG-11-UTP at its 5' end.

Eight plasmid clones (unique in insert size to each other) generated from *Bam*HI and 19 plasmid clones generated from *Sal*I were immobilized on the positively charged nylon membrane and hybridized to the DIG-labelled oligomer probe and visualized using colorimetric method (2.9). Two dark spots, one each corresponding to the clone generated by *Bam*HI and *Sal*I digestion, were detected on the dot blot membrane (Figure 3.19).

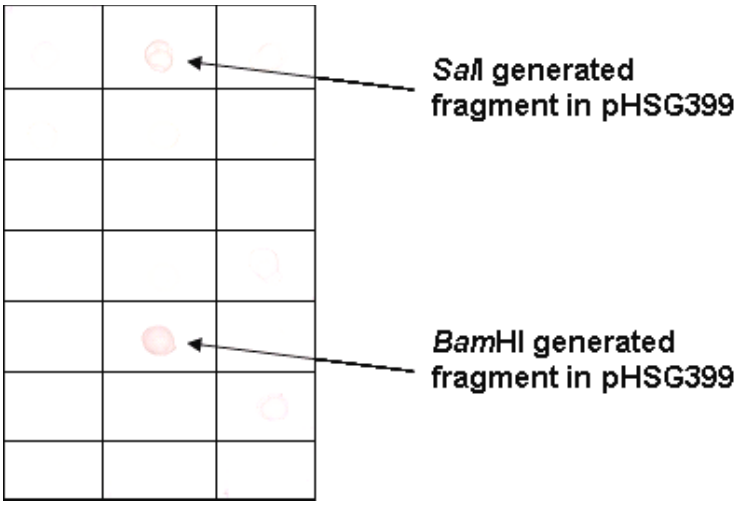


Figure 3.19 Hybridization signal from the clones containing the target genes probed with DIG-labeled oligonucleotide.

To confirm that the dot blot selection had identified the correct target both clones were hybridized to the LOV microarray. One of them was labeled with Cy5 and the other was labeled with Cy3, and they were hybridized together to the microarray to confirm the presence of the expected target. The bright yellow image (Figure 3.20a), from an overlay of both signals, indicative of a positive match from both parallel approaches, was observed as the ratio of both lasers (ratio 635 nm/532 nm) and also both separate screens by the red (635 nm) and the green (532 nm) laser produced bright red and green image (Figure 3.20b and c) in single channel scans confirming that both the clones contained the target gene.

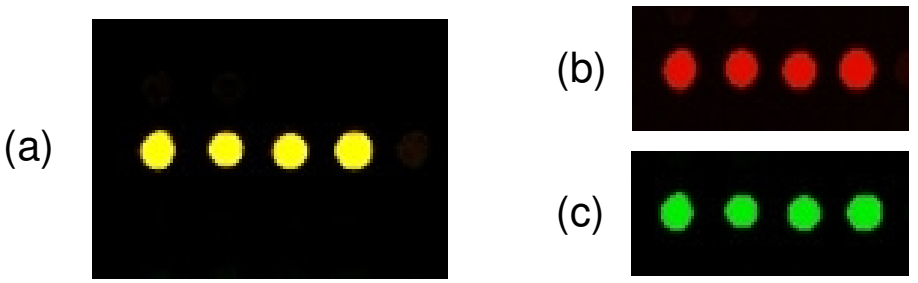


Figure 3.20 Image obtained after hybridization of Cy5- and Cy3-labeled clones, obtained from restriction digestion with *Bam*HI and *Sal*I. Both signals are overlaid, yielding the yellow color. Here, the *Bam*HI generated clone was labeled with Cy5 and the *Sal*I generated clone was labeled with Cy3. Both clones were detected first by dot blot hybridization from few clones and were subjected to microarray hybridization in order to confirm the data. (a) A bright yellow image was obtained as the ratio 635/532, (b) red and (c) green images were obtained in single wavelength scans using the 635 and 532 nm lasers separately.

3.9 Sequencing the target subclone

After the identification of the positive subclones was confirmed by microarray hybridization, the inserts in the subclones were sequenced. The first approach was to use the end sequencing M13 forward and M13 reverse primers. The second approach was to use the microarray probe-based sequencing primer which was designed on the nucleotide sequences from the microarray that produced the positive signal. Finally,

primer walking was applied to read further into the sequence (Figure 3.21). A sequence of about 5.2 kb was obtained from the *Bam*HI generated, positively identified plasmid clone. An open reading frame analysis showed that the 5.2 kb sequence contained 26 ORFs, out of these only three were related to the previously known genes. One of these ORFs included a LOV domain (start at position 146; stop at position 3760, giving a length of 3615 bp) and showed sequence motifs indicative for a multi domain signal transduction histidine kinase (HK) protein. The gene was named *ht-met*1 (Hamburg Thermophile Metagenome 1) denoting its place of origin and thermophilic nature of the source microorganisms (NCBI accession: EU934096. Expression and the functional characterization of the *ht-met*1 will be discussed in chapter 4.

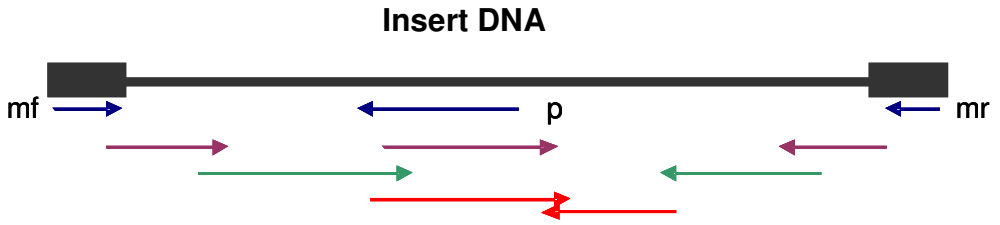


Figure 3.21 Sequencing strategy: The target domain which is at an unknown position in the insert region was sequenced using the microarray probe-based primer (p) and both ends were sequenced using M13 forward and reverse primers (mf and mr). Further sequencing was carried out using primer walking as shown by the coloured arrows.

3.10 Cross hybridization and thermodynamic properties

During screening a batch of 400 clones (from which the initial positive signal was obtained) from the thermophilic library, signals of various intensities were observed from six different probes. The signals were obtained from the probes # 5084, # 5083, # 5090, # 31GB, # Aus025 and # 5070. It was realized that the multiple signals were from a cross hybridization of a single LOV domain with these probes on the basis of their sequence similarity and identity on the level of contiguous stretch. All of the signaling probes shared 79 to 96% sequence identity between their highly

matching counterparts (Table 3.7), and also they shared more than 23 bp contiguous identities with each other (Figure 3.22a).

Table 3.7 Relative sequence similarity of the target detected from the thermophilic library to the probes that produced signals. The sources of the probes given on the table are: 5084: *Ralstonia solanacerum*; 5083: *R. solanacerum* UW1; 5090: *Kineococcus radiotolerans*; 31GB: Global Ocean Sampling; Aus025: EBPR sludge; 5070: Global Ocean Sampling.

	HT-LOV	5084	5083	5090	31GB	Aus25	5070
Target (HT-LOV)	100	81	83	85	85	79	75
5084	81	100	96	81	79	72	75
5083	83	96	100	85	87	72	75
5090	85	81	85	100	85	74	64
31GB	85	79	87	85	100	79	83
Aus25	79	72	72	74	79	100	72
5070	75	75	75	64	83	72	100

The alignment of the amino acid sequences from the target sequence and the signaling probes show that a positive detection is possible even with the sequences which vary up to 4 amino acid sequences in the conserved core region (Figure 3.22b).

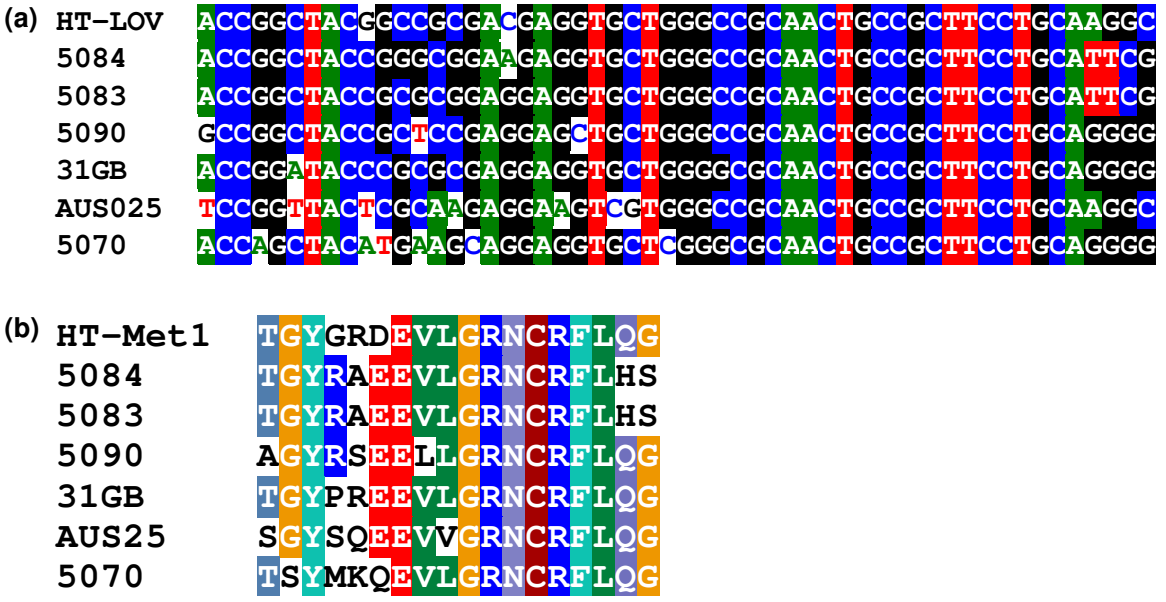


Figure 3.22 (a) Alignment of nucleotide sequences obtained from the LOV domain core of *ht-met*1 to the probes that had given a positive hybridization signal in the microarray screen. Probe 5084 contains 8 mismatches, 5083 contains 4 mismatches, 5090 contains 5 mismatches, 31GB contains 5 mismatches, AUS025 contains 9 mismatches and 5070 contains 10 mismatches on the aligned region. HT-LOV is the LOV domain from *ht-met*1 while the rest of the probes are as given on Table 3.7. (b) Alignment of amino acid sequences from the LOV domain core (region corresponding to the probe length on microarray) of the target sequence and the signaling probes.

Surprisingly, the probe that shares 81% sequence identity to the target *(ht-met*1 LOV) had produced the highest signal intensity. The signal intensity from the probe that shares 85% identity to the target had produced a signal intensity that was almost half of the former (Table 3.8), and this trend was consistent during downward screening and also with dual labeling too (with both Cy5 and Cy3 labeled clones).

The ultimately identified gene *ht-met*1 did not have a perfect match on the LOV microarray, yet, the detection was possible because of the cross hybridization obtained from the mismatch probe. However, the strength of the cross hybridization

was difficult to determine only on the basis of the sequence identity between the probes and targets.

Table 3.8 Fluorescent signal intensity from six different probes obtained during the hybridization of thermophilic soil metagenomic library. I1, I2, I3 and I4 are image intensities from four spots of a single probe (printed in quadruplicate).

S.N.	Probe	I1	I2	I3	I4	Average signal intesity
1	5084	2565	2129	1865	1595	**2038.5**
2	5083	1655	1384	1393	1512	**1486**
3	5090	859	952	854	795	**865**
4	31GB	450	441	394	375	**415**
5	Aus025	437	415	402	378	**408**
6	5070	316	311	335	310	**318**

To determine the effect of the thermodynamic properties on the hybridization signal, hybridization free energy between the target and each of the signaling probes were calculated using DINAMelt server, a web-based application (http://dinamelt.bioinfo.rpi.edu/). Probe 5084 that is 81% identical to the target but had produced the highest SI has a hybridization free energy of -65.75 kcal/mol, while probes 5083, 5090, 31GB, and Aus25 have -54.91, -45.71, -47.29 and -39.73 kcal/mol respectively. Probe 31GB, though yielding a slightly higher hybridization free energy than probe 5090, has lower signal intensity.

3.11 Hybridization with metagenomic DNA and environmental strains

To extend its applicability to screen the complex environment and unknown microbial strains, the LOV microarray was hybridized to the DNA material from such sources.

3.11.1 Metagenome from Elbe River

After the successful application of the microarray to detect the target from individual genomes, cloned DNA and libraries, it was used to search for LOV domain genes from complex DNA material extracted from various environmental sources. For this

purpose, metagenomic DNA from Elbe River sediments near Hamburg was made available (courtesy of Prof. Wolfgang Streit, Department of Microbiology, University of Hamburg). The DNA was originated from the natural microbial population and extracted from the river sediments directly. Approximately 2 µg of DNA from the Elbe River metagenome was labeled with Cy5-dCTP and hybridized overnight to the microarray using the step-down temperature program as described in 2.8.3. Several fluorescent spots with various signal intensities (Table 3.9) were seen during the image scan indicating the presence of LOV domain-coding sequences in that metagenome (Figure 3.23).

Figure 3.23 Fluorescence image obtained after the hybridization of Cy5 labeled DNA from the Elbe River metagenome. Each probe was spotted on the microarray slide in quadruplicate; hence the four fluorescent spots correspond to a single probe.

The highest signal intensities were 5083, 5084 and 5070. Signal to noise ratio of the probes 5083, 5084 and 5070 were 3.71, 3.28 and 3.01, respectively, while a few other probes produced lower SNR values (< 3.0). Some of the probes that produced signals share more than 80% sequence identity (probes 5083, 5084, 5090) to each other.

Table 3.9 Signal intensity (SI) and signal to noise ratio (SNR) obtained from the probes that showed positive hybridization to the Elbe River metagenome

Probes	Average SI	Average SNR
5070	108	3.01
5074	45	1.19
5083	134	3.71
5084	118	3.28
5085	100	2.85
5087	63	1.67
5090	81	2.21

In this approach, the purpose of the LOV array was to qualitatively assign the DNA sample rather than performing a quantification of the signal in relation to the species diversity or the functional diversity. The next step was identifying the microarray-detected LOV domain sequences from the complex metagenome. Because of the DNA quality (smaller size) and the limited amount of DNA, construction of a large insert library from the Elbe River metagenome was not taken into consideration. Instead, it was attempted to amplify the putative LOV domain using degenerate PCR. A pair of degenerate PCR primers was designed on the basis of the amino acid sequences of the related genes (similar to probe 5070, 5090) which targeted the most conserved regions of the LOV domain as annealing sites (Figure 3.24).

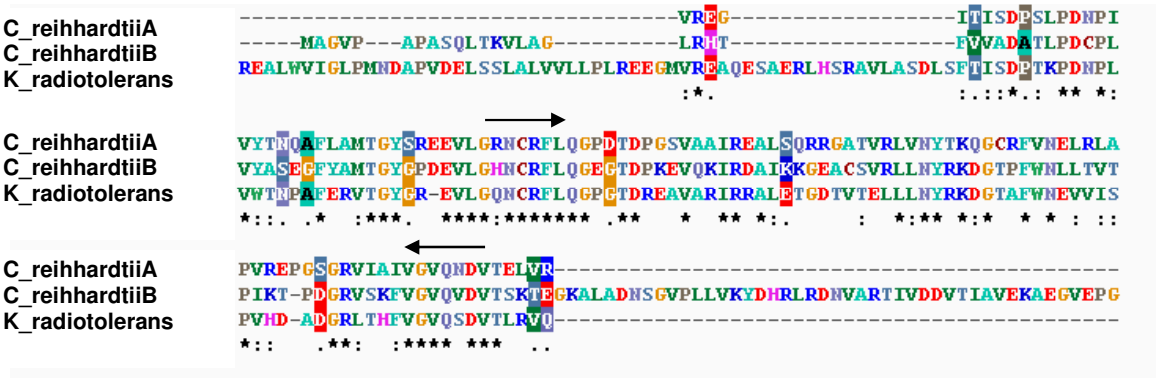


Figure 3.24 Amino acid sequence alignment used to determine the consensus region to design degenerate primers based on the conserved residues of three LOV domains. Arrows indicate the conserved residues which were targeted by the degenerate primers. The symbols indicate the amino acid identity based on clustal alignment.

Two partial amplification products, one with 207 bp (named Elbe1) and another with 123 bp (named Elbe2) were obtained from the PCR using consensus hybrid degenerate primer sets Chl_for (5'-GCC GGA ACT GCC GGT TYY TNC ARG G) and Chl_rev (5'-GGT CAC GTC GTT CTG NCA NCC NAC). Elbe1 was 69-amino acids long and Elbe2 was 41-amino acid long. Both of the sequences contained canonical conserved residues of a functional LOV domain including a reactive cysteine (Figure 3.25).

```
Elbe1                -------------------------------------RNCRFLQGSETDPAAVTELREAV 23
Elbe2                -------------------------------------NCRFFQGVETDAATLSGMRQAI 22
Ch_reinhardtii       VREGITISDPSLPDNPIVYTNQAFLAMTGYSREEVLGRNCRFLQGPDTDPGSVAAIREAL 60
LOV2_Arabidopsis     --KNFVITDPRLPDNPIIFASDSFLELTEYSREEILGRNCRFLQGPETDLTTVKKIRNAI 58
B_subtilis           --VGVVITDPALEDNPIVYVNQGFVQMTGYETEEILGKNCRFLQGKHTDPAEVDNIRTAL 58
P_syringaeDC3000     --MPMIVTDPNRPDNPIIFSNRAFLEMTGYTAEEILGTNCRFLQGPDTDPAVVQSIRDAI 58

Elbe1                RQGRECAVVIRNYCKDGTPFWNSLRVSPVFE-GN-QVTHYVGVQNDVT--- 69
Elbe2                AAGAPFLADLLNYRKSGEP---------------------------------- 41
Ch_reinhardtii       SQRRGATVRLVNYTKQGCRFVNELRLAPVREPGSGRVIAIVGVQNDVTELV 111
LOV2_Arabidopsis     DNQTEVTVQLINYTKSGKKFWNIFHLQPMRDQKG-EVQYFIGVQLDGS--- 105
B_subtilis           QNKEPVTVQIQNYKKDGTMFWNELNIDPMEIEDK---TYFVGIQNDIT--- 103
P_syringaeDC3000     AQRNDISAEIINYRKDGSSFWNALFISPVYNDAG-DLIYFFASQLDIS--- 105
```

Figure 3.25 Alignment of the partial LOV domains obtained from the Elbe metagenome using degenerate PCR primers (primers designed on the basis of information based on microarray detection). Residues from the Elbe River metagenome showing consensus to already characterized LOV proteins are indicated by shading. Elbe1 and Elbe2 are partial LOV domains of Elbe River metagenome; Ch_reinhardtii: LOV domain from *Chlamydomonas reinhardtii*; LOV2_Arabidopsis: LOV2 of Phot1 from *Arabidopsis thaliana*; B_subtilis: LOV domain from *Bacillus subtilis* (YtvA); P_syringaeDC3000: LOV domain from *Pseudomonas syringae* pv. *tomato*.

3.11.2 Selected cyanobacterial strains

Genomic DNA from eight individual cyanobacterial strains (unsequenced genomes) was obtained from Pasteur Institute, Paris (courtesy of Dr. Nicole Tandeau de Marsac). Genomic DNA from the following strains was screened using the LOV microarray.

1. *Calothrix* PCC7102
2. *Chlogloeopsis* PCC6912
3. *Chroococcidiopsis* PCC7436
4. *Fischerella* PCC7603
5. *Leptolynbya* PCC7376
6. *Microchete* PCC7126
7. *Myxosarcina* PCC7312
8. *Pleurocapsa* PCC8959
9. *Scytonema* PCC7110

As a preliminary screen, the DNA was mixed in groups of two to three strains together, labeled with Cy5 and hybridized to the LOV microarray. The strains PCC7102, PCC8959 and PCC7436 produced a hybridization signal when labeled and screened in a pool of three strains (Figure 3.26). After hybridization of the individual strains from the group of three strains, the signal was found to be related to PCC7102. Similarly, another strain, PCC6912, also produced the positive hybridization signal.

Figure 3.26 Signals obtained after the hybridization of Cy5 labeled genomic DNA from the mixture cyanobacterial strains PCC7102, PCC8959 and PCC7436.

A cosmid library with about 1000 clones was generated from strain PCC7102. The clones were screened by the microarray to detect the LOV domain containing clone. A positive hybridization signal was not obtained during the screen. A complete genome coverage and inclusion of a desired gene fragment depends also on the quality of the DNA material used to construct the library. Further analysis was not performed with the strains and generated library because of lack of the DNA material and time constraint.

3.11.3 Strains from high altitude Andean lakes in Argentina

High Altitude Andean Lakes (HAALs) have an extreme environment with high UV exposure. UV irradiation can cause DNA damages by the formation of cyclobutane pyrimidine dimers (CPD). One efficient method to repair the DNA damages caused by CPD is provided by photolyases, which act in the presence of blue light. On the other hand, cells need to respond to the blue light as light in the lower wavelength range can be harmful to them. It was thus of particular interest to investigate harvested material from HAALs due to the double-sided aspects of blue light. Few strains were already isolated from HAALs by the expedition of Dr. M.E. Farias and cultivated in her laboratory [199], from which ten strains were made available (courtesy of Dr. M.E. Farias, University of Tucuman, Argentina) for analysis (Appendix C). The genomic DNA of these isolates was labeled with Cy5 and hybridized to the microarray. Three of

the tested strains showed a positive hybridization signal. One of them (strain V1, see Appendix C) was further investigated with a set of degenerate primers to obtain partial sequence information. The PCR product was sequenced and a 117 bp sequence product showed evidence for a LOV domain. In BLASTX analysis, the nearest match was found to be putative singnal transduction histidine kinase protein (YP_497617) from *Novosphingobium aromaticivorans* DSM 12444 (52% identity), putative sensory box histidine kinase/response regulator protein (YP_003376125) from *Xanthomonas albilineans* (52% identity), histidine kinase protein (ZP_06492539) from *Pseudomonas syringae* pv. *syringae* FF5 (50% identity). Alignment of the amino acid sequences of the nearest matches is given in figure 3.27.

```
V1LOV         DCRFLQGPQT GKSALQRVQL AVERHHEVCL EVLNYRKD
X_albiline    NCRFLQGPDT DRETVRSVRD AIASHTEVAV EILNYRKD
P_syrFF5      NCRFLQGPDT DRAAVQSIRD AIEERVDIST EILNYRKD
N_aromat      NCRFLQGPGT DPAAVARIKA ALEREDVIVV ELLNYRKD
```

Figure 3.27 Amino acid sequence alignment of the partial LOV domain (V1LOV) to its nearest matches. *X_albiline:* YP_003376125 (*Xanthomonas albilineans); P_syrFF5:* ZP_06492539 (*Pseudomonas syringae* pv. *syringae* FF5) and *N_aromat* : YP_497617 (*Novosphingobium aromaticivorans* DSM 12444)

3.12 Discussion

3.12.1 Determination of an appropriate technique to screen metagenomic DNA

As discussed in chapter 1.2.2, an approach of detecting a gene of interest from a metagenome is based on two frequently applied and widely accepted methods. In this work, a functional approach was used as a first attempt to detect LOV domain-containing genes from a metagenome. Due to problems caused by fluorescence derived also from cellular metabolites, this approach was not followed further as in vivo detection tool to screen LOV domain containing genes in metagenomic libraries. As an alternative, a sequence-based screening method was applied that employed the modern and robust microarray technique. In either case, there was no previously available methodology or established technique that could be adapted to screen thousands of clones or a complex metagenome in search of LOV domains from BL receptors, requiring an extensive and careful calibration of the hereby developed methodology.

A major problem in metagenomics is the lack of tools and resources to screen already constructed metagenomic banks [90]. Although both the function- and sequence-based screening approaches have helped understanding the metagenome and discovering some novel genes [71;105;200], exploitation of the metagenomic potential could not be expedited. In working with metagenomic libraries one needs to handle thousands of large insert clones in parallel. The successful detection of a gene of interest from a large number of clones using a functional approach relies on a suitable substrate and systems for heterologous expression. To improve the function-based approach and to optimize the detection of a functional clone, a substrate-induced gene expression (SIGEX) system was suggested [201;202]. This proposal is based on the knowledge that a catabolic gene expression is generally induced by relevant compounds (substrates or metabolites) and, in many cases, it is controlled by regulatory elements situated proximate to catabolic genes. The problem of this system is that catabolic genes that are distant from a relevant transcriptional regulator cannot be traced. In addition, this system is applicable only with small insert clones (about 7 kb), and again, the knowledge of substrate is necessary [203].

A functional approach applied to detect the presence of LOV domain in metagenomic libraries based on the detection of the fluorescent signal generated from the expressed protein was not adopted further as preliminary calibration results could not generate enough confidence for the application of this approach here. The higher fluorescence of some of the library clones (from negative controls) than the signal generated by *E. coli* cells containing the *ytv*A construct was probably caused by a higher flavin accumulation in these background cells, such that the fluorescent scan could not discriminate between the different cellular flavin molecules and the FMN-LOV complex of *E. coli-ytv*A. Moreover, this approach was based on the expression of a foreign DNA in a heterologous system where presence of a complete coding frame is required. Alternatively, a degenerate PCR-based approach could be applied if sequence information is available from related taxa. However, a direct amplification of a previously unknown LOV-domain-containing gene from a pool of thousands of large insert clones is not feasible on the view of sequence variation and heterogeneity of the unknown environmental samples.

3.12.2 DNA Microarray: a sequence-based high throughput technique

A major significance of this dissertation was the application of the microarray technology in the field of metagenomics to screen unknown DNA libraries to find new genes. The development of a conserved motif-based DNA microarray, as applied here, demonstrated the complete workflow from designing an array, experimental laboratory work, microarray data analysis to gene identification.

The DNA microarray-based technology for the microbial fingerprinting and community genomic analysis has become popular in recent years because of their high-density and high-throughput capacity. The DNA microarray approach has revolutionized the fields of expression analysis and genotyping, but the true potential of this technique is still far away and waiting for innovation [145]. Besides these two common applications, use of a microarray in the field of environmental microbiology is now finally emerging. Initial stage microarrays used in the microbial ecology have primarily targeted functional genes or phylogenetic markers and the objectives of

these arrays were to assign the phylogeny or function of the microbial community in an environment by detecting the genes involved in the biological processes [156].

Once an array is designed, constructed and validated, it can be spotted again and again being a clear advantage for the succeeding research experiments [204-207].

3.12.2.1 Design criteria for generating the LOV microarray

As discussed in chapter 1 (1.4) a common DNA microarray platform makes use of PCR or oligonucleotide based probes. Construction of a comprehensive microarray based on PCR probes from all putative LOV domain-containing genes deposited in databases would be time consuming and difficult. In addition, an access to most of the clones from large metagenome projects is not possible as they are disposed after sequencing due to storage problems. Increasing use of oligonucleotide based microarrays and an improvement of their reliability have already made such microarrays promising tools [204]. In view of these facts it was such approach that was selected and developed here in the presented work.

LOV domains consist of a highly conserved core motif (Figure 3.3), including an essential cysteine residue which is required for their function. Selection of the nucleotides corresponding to the most conserved sequence motif (NCRFLQG in YtvA from *B. subtilis*) was the main design criteria of the LOV microarray so that any detection could lead more specifically and accurately to the correct identification. The array described here was designed to detect and isolate the detected LOV domain containing genes from an environmental library, a step ahead in such kind of application as previous DNA microarrays in microbial ecology were mostly based on environmental profiling or assumptive detection based on positive hybridization signals. In the absence of any functional activity and sequence information of target clones, identifying a new gene in complete form is a complicated task. We planned to “fish” a complete target gene using a short region that contains highly conserved and functional residues of a LOV domain. A positive hybridization in such a case would be due to the presence of the most conserved canonical region of a LOV domain because

of high specificity of DNA microarray as discussed later in this chapter. As a result, a positive hybridization would lead to the detection of a LOV domain protein and less prone to the detection of false positive. As the focus of the design was to include the same conserved motif of the LOV domain from all reference gene sequences, the design criteria of the probes for LOV microarray was not restricted by the GC content or the thermodynamic properties of the probes. This led to the broad range of GC content in the LOV array leading to the need of a wide variation in hybridization temperature. DNA microarrays with various sizes of oligomers have already been used in previous studies. The use of short oligonuculeotides probes (about 15 to 20 bp) for the construction of a phylogenetic oligonucleotide array [208] and a 40-mer oligonucleotide microarray to study gene expression profiling of pure culture microorganisms were already reported after the advent of microarray technology [209;210]. The environmental application of an oligonucleotide-based array to detect functional genes, however, began only with the design of 50-mer oligonucleotides that were targeted to detect genes involved in nitrogen cycling from a mixture of microbial species [158;211]. A DNA microarray constructed from 70-mer oligonucleotide probes, based on different regions of the genome of an environmental strain was applied by Rich and colleagues [212]. It was reported that there was no significant difference in the sensitivity of 50-mer oligonucleotide probes as compared to PCR probes 322-393 bases in length [213]. An additional study reported that the detection signal would be stronger in relation to the increasing length of the probe, and a shorter probe length would increase the specificity but decrease the signal and hybridization intensity [214]. One might assume, however, that inadequate uses of microarray-based techniques in environmental applications can lead to contrasting reports. It is clearly seen that oligonucleotide-based microarray probes are becoming more popular over PCR probes, also due to the challenging problem of obtaining the diverse clones or genomic DNA from various sources. Another technical problem would be the PCR amplification of the target gene and its immobilization on the glass slide, fulfilling the requirements on resolution and homogeneity of the probe material. Recently, a comprehensive microarray, named Geochip, was developed which contained 24,243 oligoucleotide (50-mers) probes that covered more than 10,000 genes from 150

functional groups [204]. In this background the probe length of 54 bp, which covered the region of 18 highly conserved amino acid residues, was considered to be ideal for the construction of the LOV microarray. The probes were designed manually as the available bioinformatics tools (e.g. PRIMEGENS [215]) are based on the physical parameters which design the primers from regions according to the set thermodynamic properties (e.g., T_m, hybridization free energy, GC content etc.) rather than deriving from conserved regions. It was shown that in silico prediction of the performance of particular probes in microarray experiments based on thermodynamic and physical properties is not possible [216]. Aim of the LOV microarray was also to target those sequences which might have slightly low similarity to the spotted probes on the array slide. In other words, a certain level of cross hybridization was expected and the detection potential was also based on the hybridization of probes to the complementary targets even with certain level of mismatch.

3.12.3 Proof of reliability: Sensitivity and specificity of LOV microarray

DNA microarrays require time and preliminary experiments for design and tests, but once a particular microarray has been established, processing of many samples can be performed rapidly. The most important potential of the microarray is to assess simultaneously several clones or a mixture of DNA to large number of probes. Although a DNA microarray holds a promise as a valuable tool for analyzing environmental samples; specificity, sensitivity and quantitative capabilities of the microarray technology for environmental application are still in the early stages of evaluation [217]. In many cases the reports are based on the individual experiments and specific applications, therefore, it is still difficult to draw a consensus conclusion.

Screening metagenomic libraries demands new tools that should on one hand be able to detect new genes and on the other hand have a high throughput capability in order to screen thousands of clones in short time. Moreover, the generation of minimum false positive information is required for which the technique needs to be sensitive and specific.

The key features, achieved in the standardization of the LOV microarray, were its sensitivity, specificity and high throughput screening capacity demonstrated by the detection of a perfect match target clone in a heterogenous pool of 1200 large insert clones in one single batch. Sensitivity is dependent not only on hybridization of the probe target but also on the concentration of probes spotted on the array. DNA preparation, labeling, repeated freeze and thaw of labeling components and also the washing conditions can affect the sensitivity. The LOV microarray was found to be specific and sensitive in several screenings, demonstrated by various hybridizations using DNA from various LOV domain containing genomes, plasmid and cosmid constructs. The determined sensitivity of about 50 ng of labeled genomic DNA from *Pseudomonas syringae* pv. *tomato* is consistent with previously reported studies where the sensitivity of a functional gene array using oligonucleotide probes were reported to be in the range of 25 to 100 ng [211].

The LOV microarray contains probes with a wide range of T_m (from 66 ^oC to 86 ^oC) that requires a broad range of the theoretical hybridization temperatures. To overcome this problem a step down temperature program, in which four cycles of four hours each per hybridization temperature in decreasing order was set (2.8.3 and 3.3.2), proved to be successful. Also, hybridization times of three hours [218], four hours [219], eight hours [153], 16 hours [220] or 18 hours [218] with uniform temperature have been applied in previous studies employing microarrays in microbial ecology.

As the ultimate application of the microarray was screening a metagenomic library, optimum confidence was needed before it could be applied to the real samples containing unknown clones. As a microarray-based screening of a metagenomic library was not reported previously, the LOV microarray was subjected to tests with a library generated from a sequenced and characterized genome that contained a LOV domain. For these tests, the genome of *P. syringae* pv. *tomato* was selected. The selection of this organism to generate a reference library was ideal because the probe derived from it had the T_m (78 ^oC) of nearly average to all probes on the LOV microarray and its genome contains only a single LOV domain-coding gene. Moreover,

the preliminary hybridization results from a plasmid clone containing the LOV domain from *P. syringae* pv. *tomato* and also using its genomic DNA gave optimum results in terms of sensitivity and specificity. In the analysis of cosmid library clones, sufficient confidence was obtained by detecting the positive target in several combinations. The detection of the target LOV domain in the background library with a strong signal (SI 5472, SNR 130) even in the pool of 1200 Cy5-labeled clones, together with the test cosmid clone from *P. putida* was already a proof for the applicability of the LOV microarray to screen metagenomic clones for genes having identical sequence motifs to the oligonucleotide probes. Similarly, the detection of 0.25 ng of a test clone in the background of 1000 cosmid clones, altogether about 10 µg in concentration, was a good indication of its high detection sensitivity. Moreover, there was no significant unspecific hybridization in the presence of the 1200 background clones. A specificity problem, if present at all, could still be compensated by using more than one oligonucleotide probe for a single gene; this technique has been applied in our lab for the design of a hydrogenase microarray.

The quantitative capability of microarray-based hybridizations is another central issue for environmental applications. A good linear relationship was observed between the hybridization signal intensity and the target DNA concentration for the test clone-pWPPLOV. This is consistent with previous reports when pure cultures were tested [158] and also for PCR product-based functional gene arrays (FGA) [221], as well as with the findings of microarray studies of gene expression [222]. However, like other molecular approaches, the quantitative accuracy of the 50-mer-based FGA hybridization depends strongly on the probe specificity and target complexity. Pilot experiments with several targets are required before quantification of the results based on LOV microarray or similar DNA microarray.

For labeling and hybridization, pure DNA is required to ensure the enzymatic and hybridization activity. Pooling of material and the alkaline lysis method was used to extract the DNA from test and background libraries, which was proved to be efficient in terms of nucleic acid quality, still possible to be performed at reasonable costs.

A graphical representation of the signal intensity from positive control clones showed a slight variation between different experiments with the same concentration of background clones (Figure 3.12b). Here, it should be noted that the background clones were premixed and the test clones were added to the background clones for each single experiment. Though the concentration of the test clone was properly controlled, another perfect match (here named as positive control clone) from the background library (*P. syringae* pv. *tomato*) was at random. The calibration results with the serially diluted clones indicate that a small change in concentration can have a significant effect on signal intensity. It has been reported that dramatically different data can be obtained in microarray analysis when the same samples are processed in different laboratories or also in different batches in the same laboratory [223].

In five cases where hybridization was carried out with the target clones showing perfect matches to respective probes, no cross hybridization to any of the other probes was observed. A probe with a two nucleotide mismatch was clearly distinguishable from the perfect match probe in the signal strength (3.6.5) and there was no cross hybridization to the probe with 81% similarity. Cross hybridization to a related probe with 96% sequence similarity (4% mismatch) produced the signal intensity of 1/2 to 1/8 of the perfect match probe. Though the LOV microarray showed high specificity when tested with some target genes as discussed in 3.1.4, it gave a few cross hybridizations when probed to a batch of 400 clones from the thermophilic fraction of garden soil metagenome. All of the signaling probes share 79 to 96% sequence identity between their highly matching counterparts (Table 3.6), and also they share more than 23 bp contiguous identity with each other (Figure 3.22). It was reported that probes showing more than 15 bp of contiguous identity with each other could produce cross hybridization [213]. Only two of the probes producing a hybridization signal to the thermophilic library clone share more than 85% of sequence identity. Surprisingly, the probe that shares 81% sequence identity had produced the highest signal intensity while that sharing 85% identity had produced a signal intensity of 1/3 of the best match (Table 3.1), and this trend was consistent during downward screening and also with dual labeling (with both Cy5 and Cy3 labeled clones). In addition, probes 5090 and

31GB share 85% sequence identity to the target (HT-Met1 LOV core), however the signal intensity from probe 5090 is higher than that from 31GB. An experimental error could be excluded, as the same trend was observed in more than five to six hybridizations. This issue can probably be explained on the basis of thermodynamics and mismatch position. The secondary structure formation also may cause a different hybridization pattern [156]. Probe 5084 that is 81% identical to the target but had produced the highest SI has a hybridization free energy (-65.75 kcal/mol), while probes 5083, 5090, 31GB, and AU25 have -54.91, -45.71, -47.29 and -39.73 kcal/mol, respectively (3.10). Probe 31GB, though having a slightly higher hybridization free energy than probe 5090, showed lower signal intensity than latter. However, only thermodynamic properties are not enough to determine the hybridization behavior of the oligonucleotide probes in the microarray [216]. Three contiguous base mismatches in the case of probe 31GB in the inner core of the probe sequence (Figure 3.21) may cause a less stable hybridization and hence a lower signal than that of probe 5090 which has only a two base contiguous mismatch in the inner core. It was reported that a mismatch in the central region of the probe affects the signal intensity more strongly than the mismatch on the terminals [216]. As seen here, hybridization patterns are difficult to predict when the microarray is applied to screen unknown samples. Because of the thermodynamic unpredictability of hybridization, hybridization of all targets with a certain probe (perfect match targets plus mismatch targets with up to ~15% identity differences) can not be expected to exhibit the same capacity, and thus prediction of the tendency that was determined for the perfect match hybridization is not applicable to unknown mixtures [224].

Kane and colleagues [213] have reported the detection of synthetic targets having >75% sequence identity using a 50-mer oligonucleotide microarray. On the other hand, a sequence identity of 85% was set as threshold by He and colleagues [220;221] for cross hybridization. Geochip2.0, the most comprehensive environmental microarray is also based on the threshold of 85% sequence identity among the probes [204]. The result from the present work showed strong cross hybridization even in the

case of 79% sequence similarity while in one case there was no hybridization between 81% identical probes.

Our results suggest that one has to be very careful in a quantitative interpretation of microarray results if positive hybridizations are derived from probes sharing similarity more than 75%. In other words, prediction on the basis of a threshold value of >85% can not be always accurate.

To discriminate the cross hybridization, inclusion of more probes from different regions of the target gene may be a solution in which a signal generated by error or cross hybridization can be avoided by comparing the hybridization from other probes of the same gene. Inclusion of multiple oligonucleotide probes from a single gene on the array also results in a broader range of detection and helps to minimize the target escape. Some studies have suggested that a signal to noise ratio of >3 should be taken as a positive detection [220]. In the case of library screening which aims to retrieve the clone, a higher signal to noise ratio (>5) is preferable to minimize the risk of false positive detection. Hybridization trials with several targets are necessary to determine sensitivity and specificity for a newly developed microarray based on a conserved motif of a functional gene.

3.12.4 DNA Microarray: beyond the expression and phylogenetic analysis

We have shown for the first time that a DNA microarray can be confidently applied to detect and isolate a target clone from a DNA bank containing thousands of metagenomic library clones. This proof was an indication that the DNA microarray can be used for the identification of new genes from the environment. The approach presented here is an innovative application of the microarray technique in the field of metagenomics. Moreover, the results of the hybridization analysis were reproducible and consistent for repeated experiments. After testing the reliability of the LOV microarray on standard library clones, it was employed to screen metagenomic libraries. In one of the screens, a drinking water biofilm metagenome library [198] showed few weak positive hybridization signals to the LOV microarray. The probes that yielded a positive hybridization were derived from *Ralstonia solanacearum* and

Novosphingobium aromaticivorans. Previous analysis based on snapshot sequencing of random clones from the same metagenome had reported that 6.8% of ORFs (out of 1,344 ORFs) show high similarity (in the range of 37% to 84%) to the known proteins from *Ralstonia solanacearum* [198]. That strengthens the fact that the positive signal was more likely from the target LOV domain of *Ralstonia* sp., but the weak signal obtained was probably because of the sequence divergence.

A cosmid library made from thermophilic fraction of a garden soil metagenome was screened in pools of 400 clones which gave few hybridization signals and only two of them yielded a strong signal (Table 3.7). The downward hybridization with splitting the initial number of clones into two halves each time, and later by performing dual labeling yielded finally a single clone. An outline of the screening strategy is presented in figure 3.27.

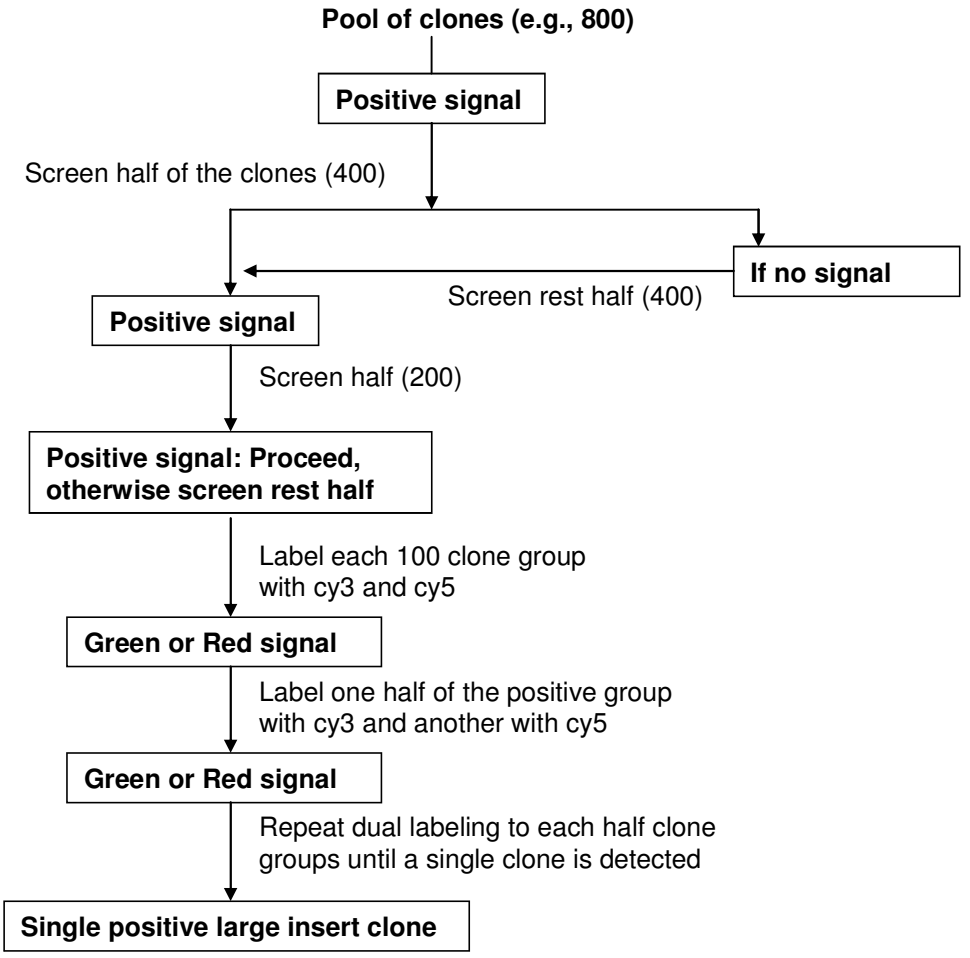


Figure 3.27 Overview of a screening approach for a metagenomic library to detect a clone containing a target gene. In the example presented here, the initial batch consisted of 800 clones which were reduced by half in each scan later. A dual dye approach was also used which helps to avoid false positive signals by experimental errors.

From the positively identified cosmid clone, small insert clones (plasmid cloning step) were generated, and dot blot-based detection was applied. Dot blot screening is optimal to apply on a small number of clones against a single probe and it is less cost intensive than microarray application. On the other hand, the use of a microarray is advantageous to screen clone banks, allowing screening several hundreds of clones against several thousands of probes in parallel. Sensitivity and specificity of the

microarray is far better than that of the dot blot hybridization and dot blot hybridization, as used here can not be taken as complementary screening tool over DNA microarray.

Since its advent [142], the DNA microarray has become a regular technique in large scale gene expression analysis. In environmental application this technique is still in juvenile state. Issues of reproducibility and predictability in microarray gene expressions have been raised in several studies [225-227]. Also some of the results published in renowned journals that were predicted on expression-based analysis were disputed [228;229]. Most of the applications of the microarray approach in metagenomic analysis are limited in the detection of the target microbes and functional pathways [153;158;204;211;224;230]. A metagenomic profiling module applied by Sebat and colleagues [153] is presented in figure 3.28. In the metagenomic profiling the microarray probes are derived from PCR amplification of metagenomic clones and spotted on the slides. Several types of reference DNA, genomic DNA from individual isolates and metagenomic DNA are hybridized to the array and affiliation is assigned based on the hybridization to the reference DNA. This technique can help to assign the cloned DNA to certain taxonomic or functional group but that approach is less feasible for the identification of a functional gene from metagenomic libraries.

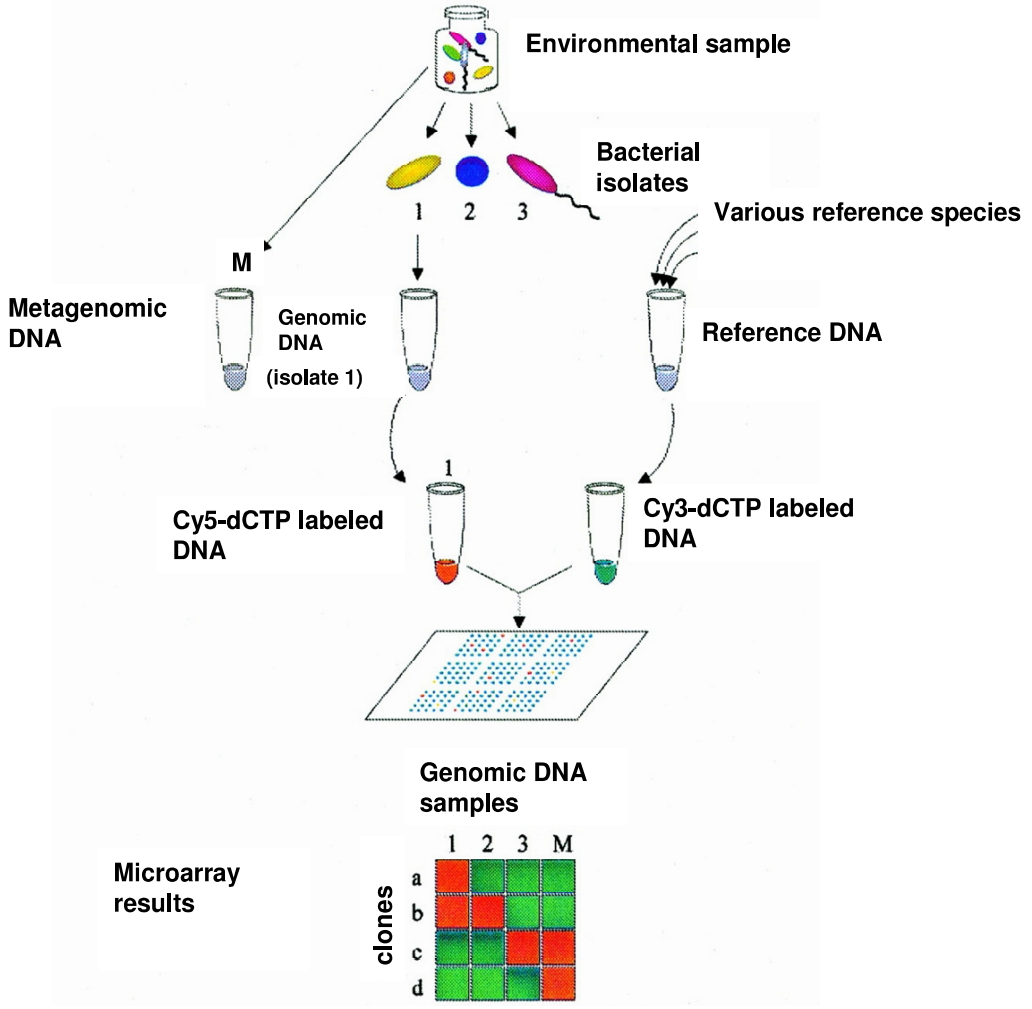


Figure 3.28 Schematic representation of metagenomic profiling adopted by Sebat and colleagues [153].

In contrast to studies using pure cultures, the target sequences in environmental DNA are diverse and complex, which make a microarray-mediated analysis of environmental nucleic acid material challenging. It is still uncertain whether a microarray-based detection can be quantitative [217]. Recently, Park and colleagues [231] used a metagenomic array in which metagenomic fosmid clones were spotted on the slide and the slide was hybridized with the known samples to detect the clones that

contain the specific genes (Figure 3.29). Immobilization of several thousands of unknown large insert clones on the array slide and their hybridization with a limited number of probes can be exhaustive in regular metagenomic library screening. Clearly, the approach developed in this work is convenient and cost effective in comparison to the array made of metagenomic clones. Also, the detection efficiency and image resolution of the system applied in this work can be better than that of the spotted clones besides uniformity and consistency.

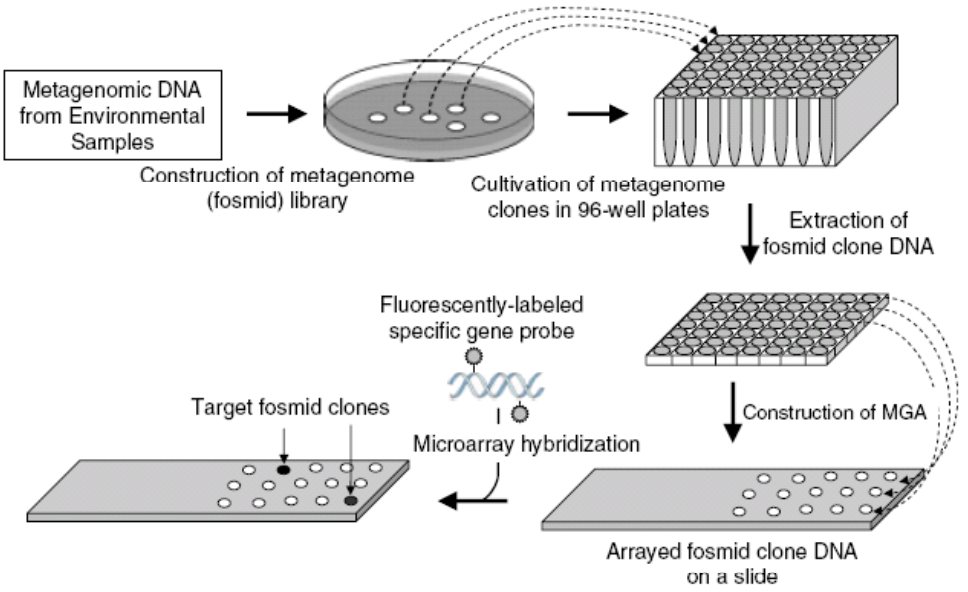


Figure 3.29 Schematic diagram of the construction of a metagenome microarray (MGA) and screening of target metagenomic clones adopted by Park and colleagues [231].

A schematic diagram for the detection of a clone containing a specific gene and its isolation as applied in this work is given in figure 3.30. As described in 3.7.2, dual labeling of the target is also an efficient way to speed up the identification process after a signal is detected.

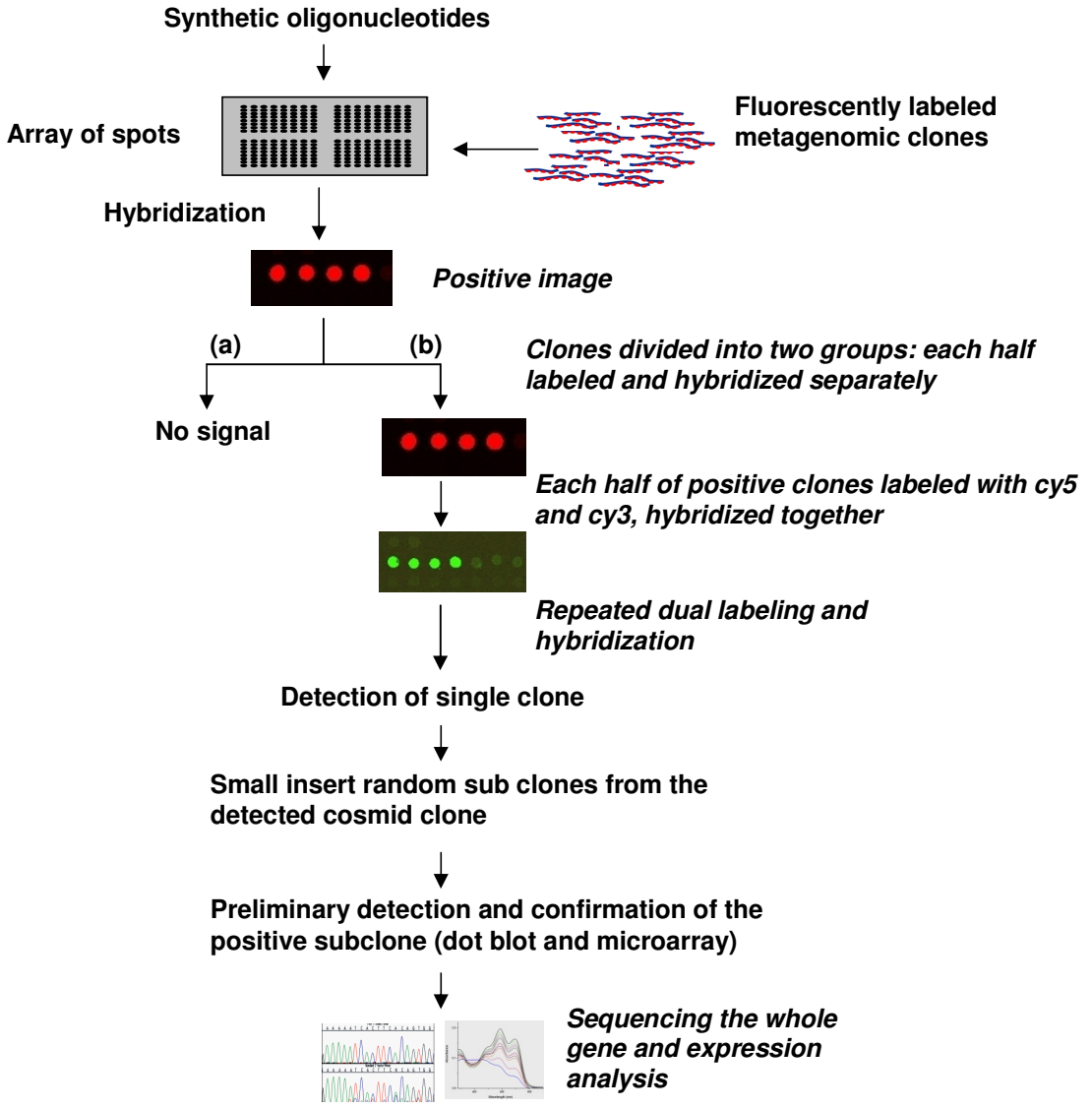


Figure 3.30 Schematic representation of the application of conserved motif based DNA microarray to detect and isolate a target gene applied in this work.

One of the possibilities offered by the high throughput sequencing technology [88] is sequencing of large amount of DNA in a single day. But sequencing whole DNA material in search of a single gene from a complex metagenome is not a meaningful

approach, because it is clearly too costly and needs further platforms for analysis too. Only four LOV domains were present in the sequence database derived from a major soil metagenomic sequencing project that generated 152,406,385 bp of sequences and nearly 150,000 reads [124]. Here, screening of just over 2000 cosmid clones with a microarray with 149 single probes could discover a previously unknown gene, which is a good indicator of the applicability of a microarray-based technique to identify a new gene from a metagenome in a short period of time bypassing the need of sequencing all clones of a library.

When DNA material from one environment is tested in a standard condition, the further tests in differential conditions (e.g., after addition of the substrate, stress etc) and the normalization of the data to the standard condition can give more accurate results in the environmental application of DNA microarray. Another advantage of the custom-made array to screen a library in search of a gene is its independence from expression data that are deposited in a public database. In-house standardization and calibration is sufficient for screening application.

In addition to the screening of DNA material from the environment, the DNA microarray can be used also to screen RNA from the environment. Application of the DNA microarray containing functional gene probes to detect mRNA in the environment provides direct information about active groups, genes expressed together in the environment and when they express. The results of such analysis can give insight into the interspecies processes including signaling and substrate exchange within the community. In the particular example outlined here, the LOV microarray could be combined with other functional gene probes to investigate transcription analysis in environmental communities, shedding more light on the role of LOV domains in microbial ecosystem.

3.12.5 Further application of the LOV microarray

As an extension of its applicability to screen complex environments and unknown microbial strains, the LOV microarray was hybridized to the DNA material from a metagenome obtained from river sediments and from few environmental

strains. Degenerate PCR amplification showed that the metagenome contained LOV domain sequences. Similarly, individual cyanobacteria strains and strains from HAALs were also screened using the LOV microarray. The signal obtained from more diverse strains isolated from unique environments like HAALs were weak. Since signal intensity is based on identity to the probes, the inclusion of more probes can extend the range of detection of the targets.

A gene-centric approach [117], which focuses a particular type of genes, can be applied to screen a complex DNA sample from an environment to get initial information of the presence of a gene of interest. It has been shown that certain environments are rich in specific gene diversity. For example, proteorhodopsins are very common in marine surface waters compared with other habitats [121], and enzymes that break down cellulose are more common in the termite hindgut than in other habitats [63]. Initial screening of complex DNA from any environment gives a preliminary idea about the presence of a target gene of a particular interest. After the positive identification, a large insert library can be made out of the environmental DNA and screening of the library can help to isolate the detected target.

Although applications to diagnostic and determinative studies can be well formulated [204;219;232], the microarray-based approach applied here should provide a powerful format for the systematic exploration of a metagenome for the detection of novel genes. It is increasingly evident that only a small fraction of the microbial world has been characterized, which remarkably challenges the biotechnological potential of this undescribed diversity [131]. However, the microarray based on a conserved motif of functional genes as described here has a potential to address this question and to accelerate the detection of genetic potential for biotechnological innovation.

Chapter 4

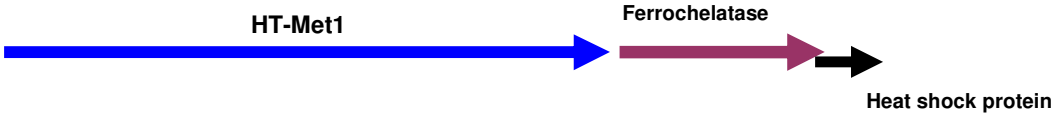
Sequential and functional characterization of metagenome-derived LOV domains

4.1 *ht-met1*: A novel BL receptor gene from a soil metagenome

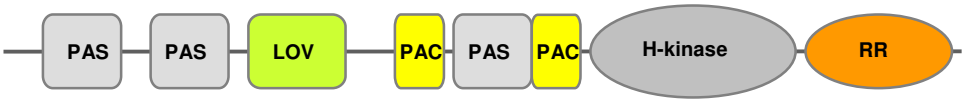
4.1.1 Genetic features of novel BL receptor gene

Among the 26 ORFs analysed from the thermophilic soil cosmid clone described in chapter 3, only three were related to known genes, a putative BL-receptor named *ht-met*1 (start 146 bp; stop 3760 bp, length 3615 bp), a ferrochelatase (start 3804 bp; stop 4853 bp, length 1050), and a ribosome-associated heat shock protein (start 4769 bp; stop 5065 bp, length 297 bp (Figure 4.1a). Compared to all reported LOV-domain genes, the novel BL gene showed highest similarity to a known sequence from *Kineococcus radiotolerans* SRS30216 (58% for the LOV domain only) and to a gene from *Methylibium petroleiphilum* PM1 (57% for the LOV domain only). The 1204 aa sequence of the metagenome-derived gene product was scanned for the presence of conserved domains using ScanProsite. The domain scan identified four PAS domains, three PAC domains (PAC: PAS-associated C-terminal motif), and a histidine kinase domain which was followed C-terminally by a CheY-type response regulator (Figure 4.1b). PAS domains are identified at aa170-240, aa310-380, aa428-502, and aa688-760, PAC domains at aa503-557, aa634-687, and aa764-817. The HK motif is at aa834-1054, and the response regulator is at aa1080-1197. A sequence comparison of the PAS domains identified the third PAS motif as a LOV domain (aa442-546) which exhibits all amino acids known to be essential in LOV domains for chromophore binding (Figure 4.1c). Also, two residues forming a highly conserved salt bridge, a glutamate and a lysine (aa473 and aa514) are detected at correct positions.

(a)



(b)



(c)

MTPAVPPDDEESGRLRVLRELRLLDSSREECFDAAVRLLARMARVPIATLSLVDAGRVWFKAIHGAELDE
LSRDQTCCTWTIAGDVPFLVDDLTADPRFAGVRTVTGAPYVRAYAGLPLTVEDQRIGALAVLDVQPRHFG
EAELAALHDVGRLVCELLKARLKEQRWRLQEARVRTASRAGSDWLWETDAQGRFTWASASMETHTGLRIE
DKLGKLPHEILLRRDDPEHARSYDRYLAAREQREPFFNLIADMRTPGGLVTVAVNGVPVFDSAGRFRGYR
GATRVVTAEIEAQRQAREAQALREAQLRESEARTAAVLQSLPDLWFVLDAEGRYVQCGTAAQHLLHMPFE
QMRGASLEEVLPPAVAVQHKNALQRALATGTLQRVEYELQPMGDPTTRTFEGRFMPMSGTQVLLIVRDLT
ELRTMERELQIMHRAIEGEASLPITVVDVLAPDHPVVYVNPAFERLTGYGRDEVLGRNCRFLQGPETDQP
ALGQVREALRHGRSTTVVLHNYRKDGSRFINELHIAPVHDSHGRVTHFIGVQTDITERSIAAERLRVSEE
LYRSVALSISDGLMVVTSGSYVVAVNPAACRILDATMEELVMAPGLELLTESLQPLPPAEHPVARALASG
EPVTERVYGLRRRDGQIRWLGVSAQPLRMAPHDKPLSAVVTFRDITQQRIAEQALSMSEQRWKFAIEGSG
DGVWDWDAVADRVYYSRRWKELLGYQEHELTDSPEEWSSRVHPEDLPRVLEKLELHLKDVVPVYHSEHRL
RHRSGRWLWVFERGRVVQRNEQGRALRIVGTYSDITRLKEAEKALRDKQAAELASRSKTEFLSRMSHEMR
TPLNAMIGFSQLLKLSATNAQPSASQVRHYADHMLQAGQHLLALINDVLDLQRVETGQLTINIETISLRR
TIEQALELLAPLAHASAVSFDNRIDEDVLVRADAQRLRQVLINIISNAIKYNRPDGVVRLTLGDRQGDAI
DVLIEDTGEGLRPDQLARLFQPFERLGRETSPIEGTGLGLIIAKRLIDEMKGSIEVSSEPGKGTCVTLMV
RLAEQAPAALPHEASAAPAPDPEVDLAPLRMLYVEDNPINALLFEEAMKLSGHIDLMIAEDGPQAFEIVR
DWRPDILVLDAHLPGMSGFEVLRELREVHGLRDVPAFMCSADAMPEDIQRALDAGFLGYWTKPIEVRKVI
AEVNQVAPRPVRVS*

Figure 4.1 (a) Arrangement of three ORFs in the microarray-detected cosmid clone. (b) Domain structure of the HT-Met1 protein. PAS: Per, Arnt, Sim domain; PAC: PAS-associated C-terminal motif; RR: response regulator; H-kinase: histidine kinase. The third PAS domain along with its fused PAC domain forms the LOV domain. (c) Amino acid sequences of the HT-Met1 full-length protein; the sequences encompassing the LOV domain are highlighted. Residues shown in red are active residues that interact with the FMN chromophore. The chromophore binding cysteine residue in the motif NCR is enlarged and shown in green.

For both, the protein derived from the thermophile metagenome and the protein from *M. petroleiphilum* PM1 (gene name Mpe_A2494), even the gene arrangement in

this genome region is conserved such that the gene coding for the LOV-domain protein is followed by an ORF coding for a ferrochelatase and the ribosome-associated heat shock protein. The nearest match for the over all, full length protein with 52% similarity at protein level of HT-Met1 is *Methylobium petroleiphilum* PM1. Also both of the other identified genes in the metagenome clone, encoding the putative ferrochelatase and the heat shock protein (S4 paralog) following HT-Met1, show similarity to the corresponding genes of *M. petroleiphilum*. The ferrochelatase from pWTHLOV clone was found to be conserved in many bacteria with *Acidovorax delafieldii* 2AN (similarity = 71%) and *M. petroleiphilum* PM1 (similarity = 69%) yielding the best match. The heat shock protein showed similarity to *M. petroleiphilum* PM1 (similarity = 75%) and *Polaromonas* sp. JS666 (similarity = 74%).

4.1.2 Phylogenetic analysis of HT-Met1

Though the prediction based on the sequence alignment has revealed the presence of LOV domains in diverse form of life, only few of them have been characterized functionally. Among the characterized ones, all have shown similar blue light-dependent photoactivity. It is difficult to taxonomically characterize metagenome-derived genes due to the limited information available. The best approach is to establish a phylogenetic relationship with potentially related proteins from known organisms.

Because of the high sequence divergence of LOV domain containing proteins, in part due to the variability of fused signalling domains, a phylogenetic analysis was performed in two steps. At first, sequences similar to the HT-Met1 full length protein were derived from a BLAST search of the NCBI protein database and aligned. Most of the identity obtained was in the histidine kinase and the response regulator domains rather than in the LOV domain or other sensor PAS modules (Figure 4.2). No significant alignment was found along the LOV domains, except with YP_001021684 from *M. petroleiphilum* PM1 (gi|124267680).

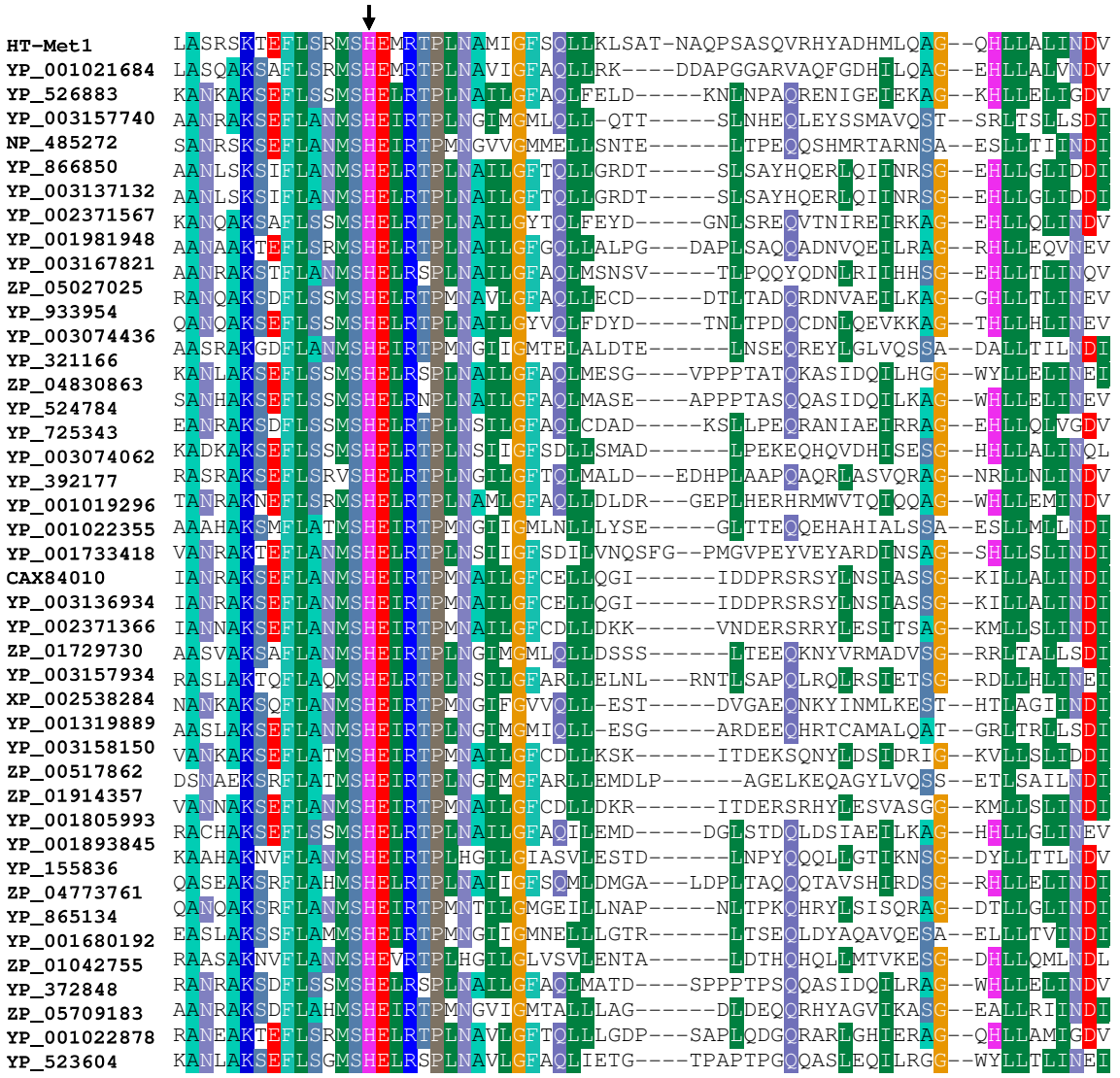


Figure 4.2 Alignment of HT-Met1 with similar sensor histidine kinase proteins, showing the conserved region that contains the function-involved histidine residue (indicated by an arrow).

The phylogenetic tree based on the full-length proteins from related genes is shown in figure 4.3. The bootstraps values are well supported along the generated branches although many of the proteins do not contain LOV domains. The HT-Met1 protein is grouped clearly along with the β-proteobacteria and is most closely positioned to *M. petroleiphilum* signal transduction protein (YP_001021684).

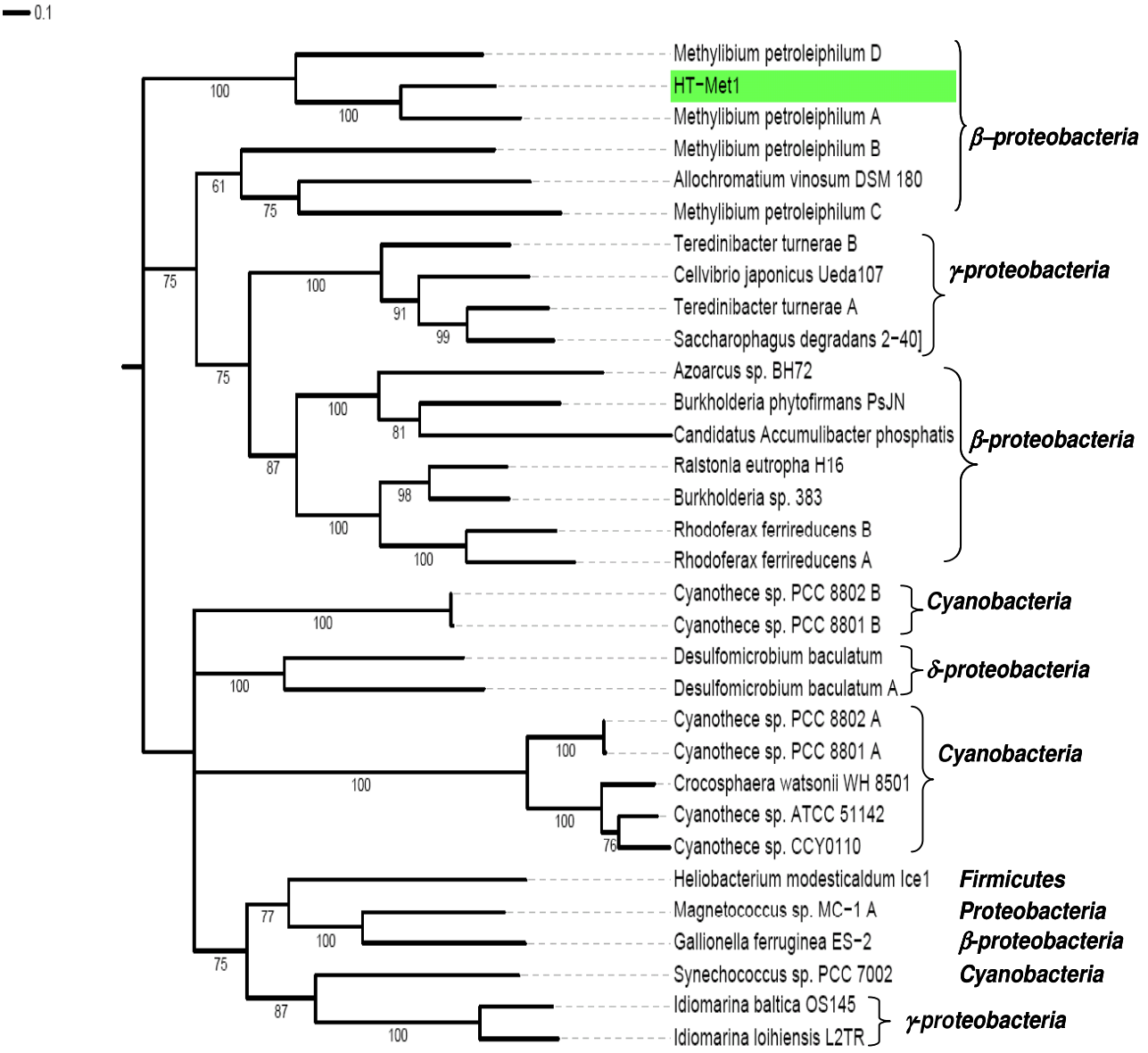


Figure 4.3 Maximum likelihood phylogenetic tree of HT-Met1 (shown on a green background) and related sensor proteins (PAS/PAC sensor histidine kinases). The mid-point rooted tree is based on a total of 100 bootstrap values; the bootstrap support is given at the nodes. A, B, C or D at the end of the leaf labels is used if there is more than one homologous protein related to HT-Met1 in the same species/strain. *M. petroleiphilum* A: YP_001021684; *M. petroleiphilum* B: YP_001022878; *M. petroleiphilum*

C: YP_001022355; *M. petroleiphilum* D: YP_001019296; *T. turnerae* A: YP_003074062; *T. turnerae* B: YP_003074436; *R. ferrireducens* A: YP_523604; *R. ferrireducens* B: YP_524784; *Cyanothece* sp. PCC 8801 A: YP_002371366; *Cyanothece* sp. PCC 8801 B: YP_002371567; *Cyanothece* sp. PCC 8802 A: YP_003136934; *Cyanothece* sp. PCC 8802 B: YP_003137132.

As an additional approach, only the LOV domain fraction of HT-Met1 sequence was subjected to a phylogenetic analysis, since the highly diverse nature of the sensor domains including the full length protein alignment interfered with a meaningful analysis. Amino acid sequences homologous to the LOV domain from HT-Met1 were obtained from the NCBI protein database. Only that part of the sequences spanning the entire LOV domain was aligned and rest of the sequences was removed. The constructed tree from the related LOV domains using 100 bootstrap values yielded a poor consensus among the replica trees. For that reason a maximum likelihood tree was generated without bootstrap analysis. The tree of the LOV domains related to HT-Met1 is shown in figure 4.4.

The clustering of the HT-Met1 LOV domains much closer to *M. petroleiphilum* and *Ralstonia solanacearum* than to any other available sequences is evident from the phylogenetic tree. Both approaches, performed with full length proteins and LOV domains only, clearly indicated that the thermophilic LOV fragment or the complete gene is related to the β-proteobacteria; interestingly, phylogenetic relatedness is more established at the level of the whole protein than on the level of the LOV domain only.

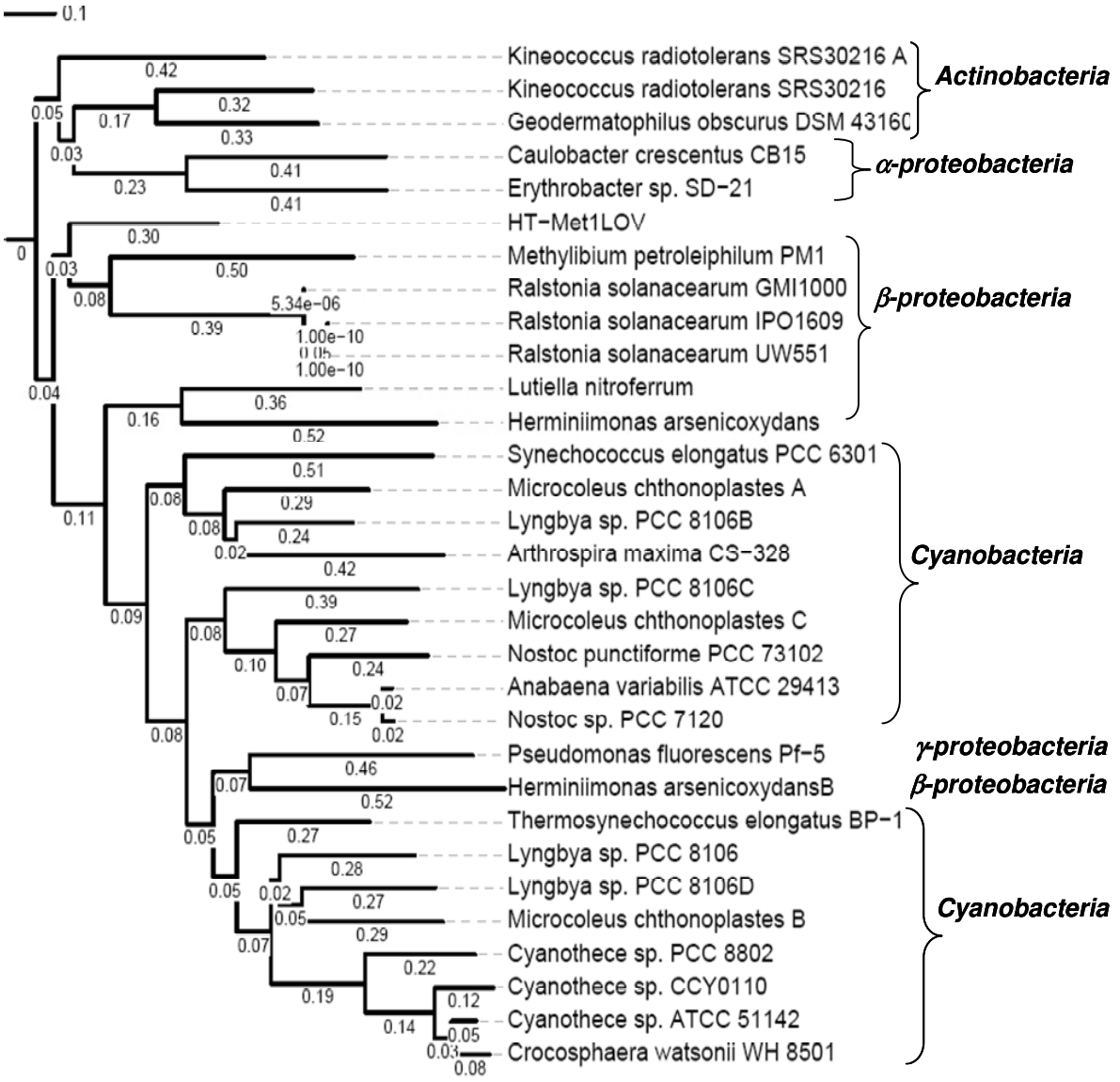


Figure 4.4 Maximum likelihood phylogenetic tree of the HT-Met1 LOV and related sensor LOV domains. The tree is mid-point rooted and the distance value is shown on the branch (ranging from 0 to 1, where 0 shows complete identity). A, B, C, or D at the end of the leaf labels is used if there more than one LOV domains were found from the same species/strain in the BLAST analysis.

4.1.3 Expression and purification of HT-Met1 LOV protein

For a functional test, the LOV-domain-encoding part (Figure 4.5) of the metagenomic gene (nucleotide positions 1273 to 1683, with respect to the start codon of HT-Met1) was cloned into an expression vector (pET52 3C/LIC); this cloning generates an N-terminal strep-tag and a C-terminal 10xhis-tag fusion protein.

MSGTQVLLIVRDLTELRTMERELQIMHRAIEGEASLPITVVDVLAPDHPVVYVNPAF
ERLTGYGRDEVLGRNCRFLQGPETDQPALGQVSEALRHGRSTTVVLHNYRKDDSR
FINELHIAPVHDSHGRVTHFIGVQTDITERSIAAERLRVSEEL

Figure 4.5 LOV domain portion of HT-Met1 gene cloned in pET52-3C/LIC for the purpose of expression and purification

The LOV domain construct in pET52 3C/LIC was expressed in *E. coli* BL21 RIL by inducing with 0.5 mM of IPTG at OD_{600} = 0.8 and overnight growth. After breaking the cells and centrifugation, the soluble fraction was devoid of the typical yellow color of FMN bound in a LOV domain. His-tag purification and western blot analysis did not show the presence of the HT-Met1-LOV protein in the soluble fraction. The expressed LOV domain protein from HT-Met1, when cloned alternatively in pET52 was accumulated in inclusion bodies as was evident by the bright yellow colour of the cell debris seen after removal of the soluble fraction. It was thus attempted to dissolve the protein for a study of its photochemical characteristics.

4.1.3.1 Solubilization of inclusion body and protein refolding

The cell debris was dissolved using CelLytic™ IB (Sigma), an inclusion body solubilizing reagent. A brilliant yellow solution was obtained which showed typical fluorescence spectra of unbound FMN. The solubilized fraction was dialyzed as described in 2.7.3 for in vitro refolding of the protein denatured during inclusion body solubilization. During dialysis all the FMN passed out through the dialysis membrane to the dialysis solution, which was indicative that the FMN was already dissociated from the protein during inclusion body solubilization. The protein obtained from the dialysis tube did not show any photochemical property.

4.1.3.2 Solubility enhancement of HT-Met1 LOV protein and purification

To obtain a soluble LOV domain protein from HT-Met1, various induction conditions, expression vectors and host systems were tested (Table 4.1). Among the various induction conditions applied, heat shock and salt stress slightly enhanced the solubility of the protein, but most of the protein was still found in the insoluble fraction.

In the heat shock method, cells were grown at 37 ^{o}C to an OD_{600} = 1.0. The cells were then transferred to another shaking incubator which was pre-heated to 42 ^{o}C. Heat shock was applied by incubating the cells at 42 ^{o}C for 30 minutes after which they were transferred to an incubator set at room temperature for expression of recombinant protein. The induction supplements (Table 4.1) were added to the culture medium and the cells were then grown for 20 hours at room temperature.

In the salt stress method, cells were grown at 37 ^{o}C to an OD_{600} = 1.0. Sterile NaCl solution was then added to the medium making a 0.4 M final concentration. The cells were further incubated under these conditions at 37 ^{o}C for 30 minutes, after which the induction supplements were added (Table 4.1). The cells were then grown at room temperature for 20 hours.

In addition, the LOV domain region was cloned into pET19b, pET28a and pET30b expression vectors and again tested for improved solubility. The constructs in pET19b and pET28a contained a His-tag at the N-terminal end, whereas the tag was C-terminal in pET30b. Only the LOV domain cloned in pET28a produced a soluble protein when grown overnight at room temperature (20-22 ^{o}C) after induction with 0.25 to 0.50 mM IPTG. The soluble fraction was purified using Ni-IDA resin.

The various constructs tested for solubility improvement are given in figure 4.6.

Construct in pET28a

MGSSHHHHHHSSGLVPRGSHMSGTQVLLIVRDLTELRTMERELQIMHRAIEGEASLPITVVDVLAPDHPV
VYVNPAFERLTGYGRDEVLGRNCRFLQGPETDQPALGQVSEALRHGRSTTVVLHNYRKDDSRFINELHIA
PVHDSHGRVTHFIGVQTDITERSIAAERLRVSEEL

Construct in pET52 3C/LIC

MASWSHPQFEKGALEVLFQGPGMSGTQVLLIVRDLTELRMMERELQIMHRAIEGEASLPITVVDVLAPDH
PVVYVNPAFERLTGYGRDEVLGRNCRFLQGPETDQPALGQVREALRHGRSTTVVLHNYRKDGSRFINELH
IAPVHDSHGRVTHFIGVQTDITERSIAAERLRVSEELNALVPRGSSAHHHHHHHHHH

Construct in pET19b

MGHHHHHHHHHHSSGHIDDDDKHMSGTQVLLIVRDLTELRTMERELQIMHRAIEGEASLPITVVDVLAPD
HPVVYVNPAFERLTGYGRDEVLGRNCRFLQGPETDQPALGQVSEALRHGRSTTVVLHNYRKDDSRFINEL
HIAPVHDSHGRVTHFIGVQTDITERSIAAERLRVSEEL

Construct in pET30b

MSGTQVLLIVRDLTELRTMERELQIMHRAIEGEASLPITVVDVLAPDHPVVYVNPAFERLTGYGRDEVLG
RNCRFLQGPETDQPALGQVREALRHGRSTTVVLHNYRKDGSRFINELHIAPVHDSHGRVTHFIGVQTDIT
ERSIAAERLRVSEELEHHHHHH

Figure 4.6 Recombinant HT-Met1 LOV constructs as generated in the given expression vector systems. The shaded amino acids are derived from the corresponding expression vectors. His-tags are underlined.

Table 4.1 Different conditions and expression systems used to enhance the solubility of the LOV domain from HT-Met1.

S.N.	Expression vector	Host cell	Growth temp.	Induction	Expression temp. (in °C)	Expression time	Cell pellets	Solubility
1	pET52	BL21RIL	37	1 mM IPTG	25	14 hrs	Yellow	Insoluble
2	pET52	BL21RIL	37	1 mM IPTG	30	6 hrs	Yellow	Insoluble
3	pET52	BL21RIL	37	150 µM	19	18	Yellow	Insoluble
4	pET30	BL21RIL	37	150 µM	19	18	Brown	No expression
5	pET52	BL21AI	37	1 mM	25	14	Yellow	Insoluble
6	pET52	BL21AI	37	1 mM	30	6 hrs	Yellow	Insoluble
7	pET19b	BL21RIL	37	0.5 mM	30	14 hrs	Brown	No expression
8	pET52	BL21AI	30	150 µM	19	18 hrs	Brown	No expression
9	pET30	BL21AI	30	150 µM	19	18 hrs	Brown	No expression
10	pET52	BL21RIL	30	9% Ethanol	RT	19	Brown	No expression
11	pET52	BL21AI	30	0.2% Arabinose 50 µM IMTG	RT	19 hrs	Yellow	Insoluble
12	pET52	BL21AI	30	0.2% Arabinose 1 mM IPTG	37	24 hrs	Yellow	Insoluble
13	pET30	BL21AI	30	0.2% Arabinose 1 mM IPTG	37	6 hrs	Brown	No expression
14	pET52	BL21RIL	30	1% Maltose	RT	19 hrs	Brown	No expression
15	pET52	BL21RIL	30	0.5 mM IPTG	RT	19 hrs	yellow	Insoluble
16	pET19b	ER256	30	0.5 mM IPTG*	21	2 hrs	Brown	No expression

17	pET52	BL21AI	37	Heat shock (42^{o} C) 0.1% Glycerine 0.1 mM Glutamine 0.05% Arabinose 100 µM IPTG	21	16 hrs	Yellow	Slightly soluble
18	pET52	BL21AI	30	0.4% Glycerine 3% Ethanol 3% 20 µM IPTG 1% Glucose	21	20 hrs	Yellow	Insoluble
19	pET52	BL21AI	37	Salt stress (0.4 M NaCl) 25 µM IPTG 0.1% Arabinose 0.1% Glycerine	21	20 hrs	Yellow	Slightly soluble
20	pET28	ER256	37	0.5 mM IPTG 0.4% Glycerine	30	12 hrs	Yellow	Insoluble
21	pET28	ER256	30	0.25 mM IPTG 0.4% Glycerine	RT (20 to 23)	14 hrs	Yellow	Soluble

* Induction was done at the OD_{600}= 2.0

For all other conditions induction was done at OD_{600}= 0.6 to 1.0

“No expression”: In that case there was no visual or spectroscopic evidence of bound FMN. RT = room temperature.

The predicted molecular weight of the recombinant LOV protein with 6x His-tag fused to N-terminal was 19.6 kDa. For photochemical studies the protein was affinity purified using Ni-IDA resin, and additional purification steps by gel filtration chromatography were performed using a Tricorn™ Superdex™75 column for crystallization trials. The purified protein showed the expected size of ca. 20 kDa (Figure 4.7).

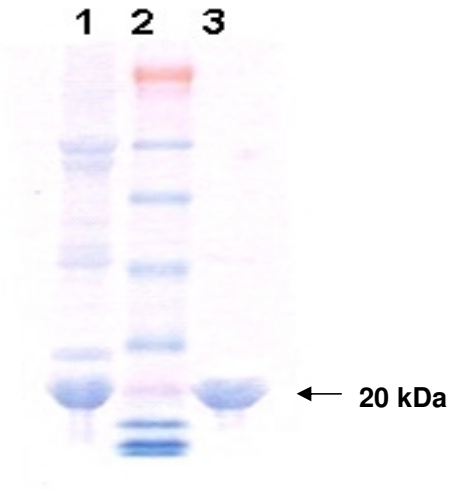


Figure 4.7 SDS-PAGE analysis of HT-Met1 LOV protein. Lane 1 shows the protein purification using affinity tag (His-tag), Lane 2 is the SeeBlue® protein marker (Novagen) and Lane 3 shows the protein further purified by gel filtration using Tricorn™ Superdex™75 column.

4.1.4 Photochemical properties of the HT-Met1 LOV domain

The photochemical properties of the purified recombinant protein were determined with UV/VIS absorption spectroscopy and fluorescence spectroscopy. The protein showed a fine-structured absorption spectrum with λ_{max} = 447 nm, characteristic in form and position for an FMN-binding LOV domain (Figure 4.8). After blue light irradiation, photobleaching of the absorption band was observed, concomitant with the formation of a short-wavelength maximum around 390 nm, indicative for the generation of a photoproduct. A complete photobleaching of the protein was not possible, since the protein exhibited a rapid dark recovery.

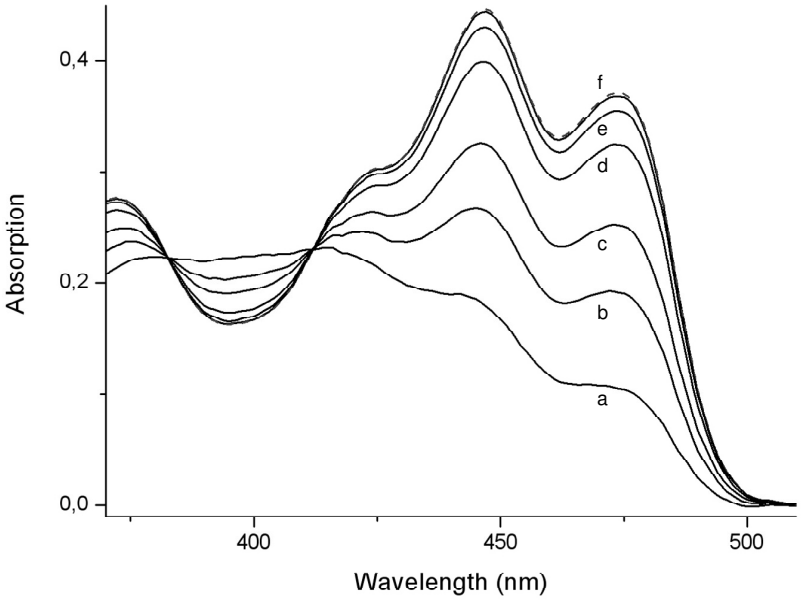


Figure 4.8 Photochemistry of recombinant HT-Met1-LOV; the dark state (λ_{max} = 447 nm, dashed curve) shows the characteristic three-peaked absorption band of an oxidized, protein-bound flavin. Curve a: maximal generation of lit state by continuous blue light irradiation (for about 4 minutes); curves b – f, 80 sec, 2 min, 4 min, 8 min, 12 min after irradiation, respectively (T = 20 ^{o}C).

Interestingly, and unreported for other prokaryotic LOV domains so far, this protein showed a fast dark recovery that was nearly complete after few minutes (t1 = 104 s at A1 = 40%; t2 = 226 s at A2= 60% at 20 ^{o}C) (Figure 4.9a).

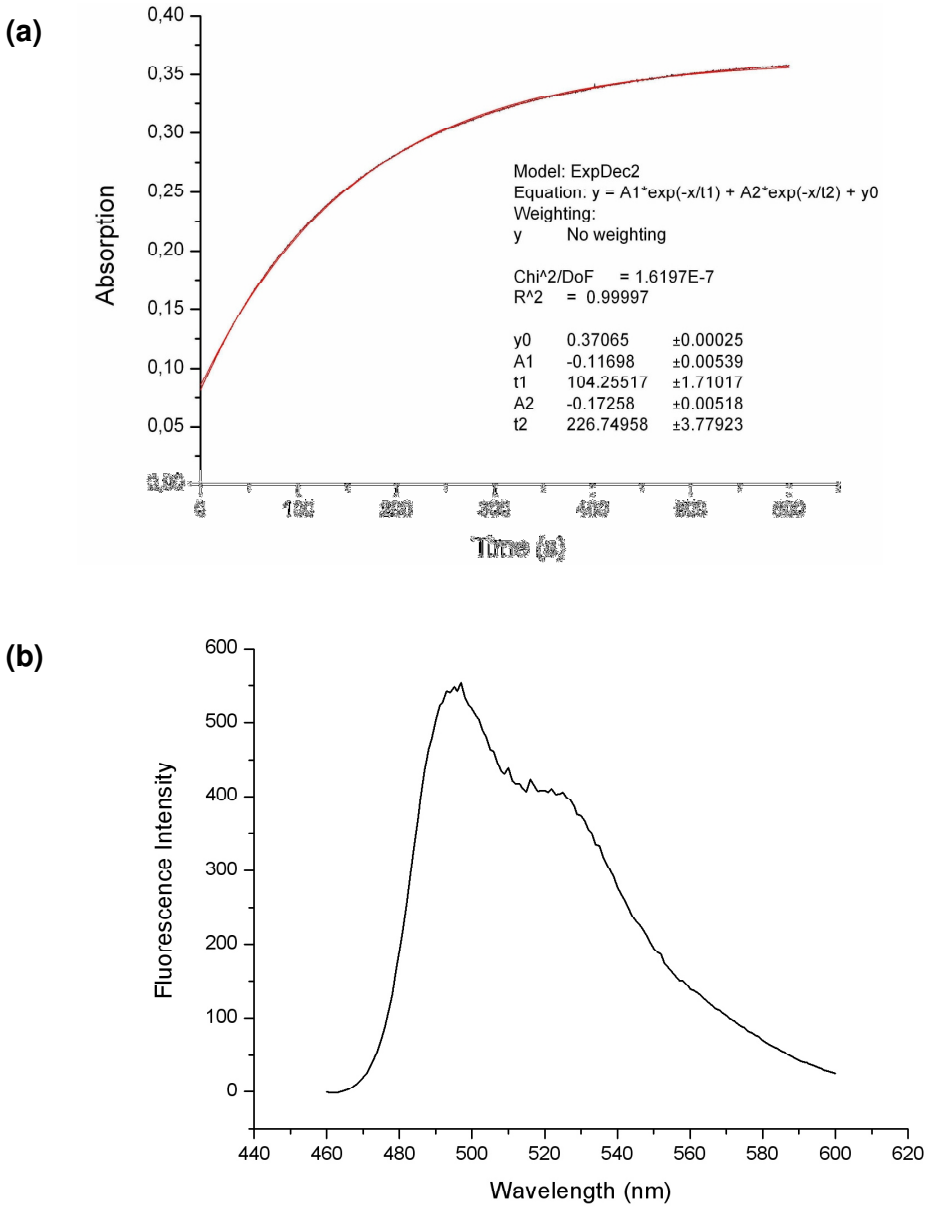


Figure 4.9(a) Dark recovery kinetics of HT-Met1 LOV. (b) Fluorescence spectrum of the HT-Met1 LOV domain (λ_{ex} = 450 nm).

Recording the fluorescence of this recombinant protein (λ_{ex} = 450 nm) yielded two emission peaks, one at 492 nm and a shoulder with maximum at 522 nm (Figure 4.9b).

4.1.5 Crystallization of HT-Met1 LOV domain

The purified protein was concentrated to 10 mg/ml in phosphate buffer (pH 8.0) using a centrifugal filter device (10 kDa molecular-weight cutoff). The concentrated protein was centrifuged using 1.5 ml filter tubes to remove any insoluble or foreign particles. With this material, a crystallization screening was carried out using the sitting-drop vapour-diffusion method. As precipitants, Cryo-I, Crystal Screen, Wizard-Iand Wizard-II were used for initial screening. The protein droplets were prepared by mixing 1 µl protein solution and 1 µl reservoir buffer solution in a 96-well plate with 100 µl reservoir solution.

Crystals were visible in three buffer conditions after one month of incubation at 4 ^{o}C (Figure 4.10). The buffers that produced the crystals were Wizard-I 7, Wizard-I 48 and Cryo-I 48, all of which contain PEG and $Zn(OAc)_2$. Study of crystal diffraction analysis and structure solving are in progress.

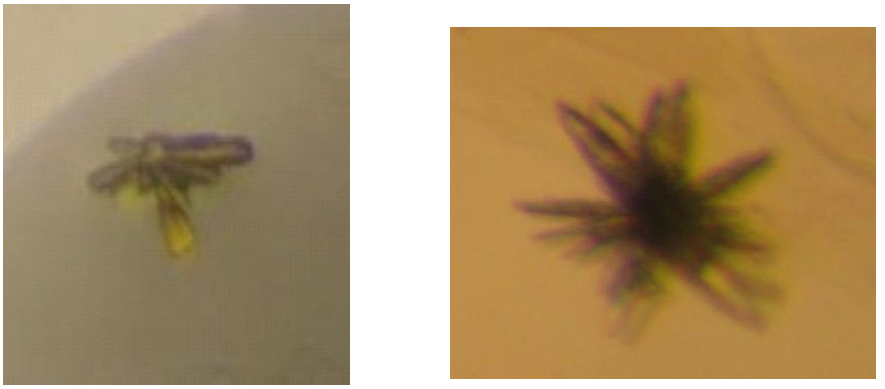


Figure 4.10 Crystals of LOV domain from the HT-Met1 protein.

4.1.6 Structure modelling

Four LOV domain proteins have been crystallized and characterized up to now in their three dimensional structure [47-49;233]. The secondary structure prediction of HT-Met1 LOV by PSIPRED reveals all typical features of LOV domains including the antiparallel ß-sheet and the α-helix bearing the amino acids that fix tightly the phosphate group of FMN (Figure 4.11).

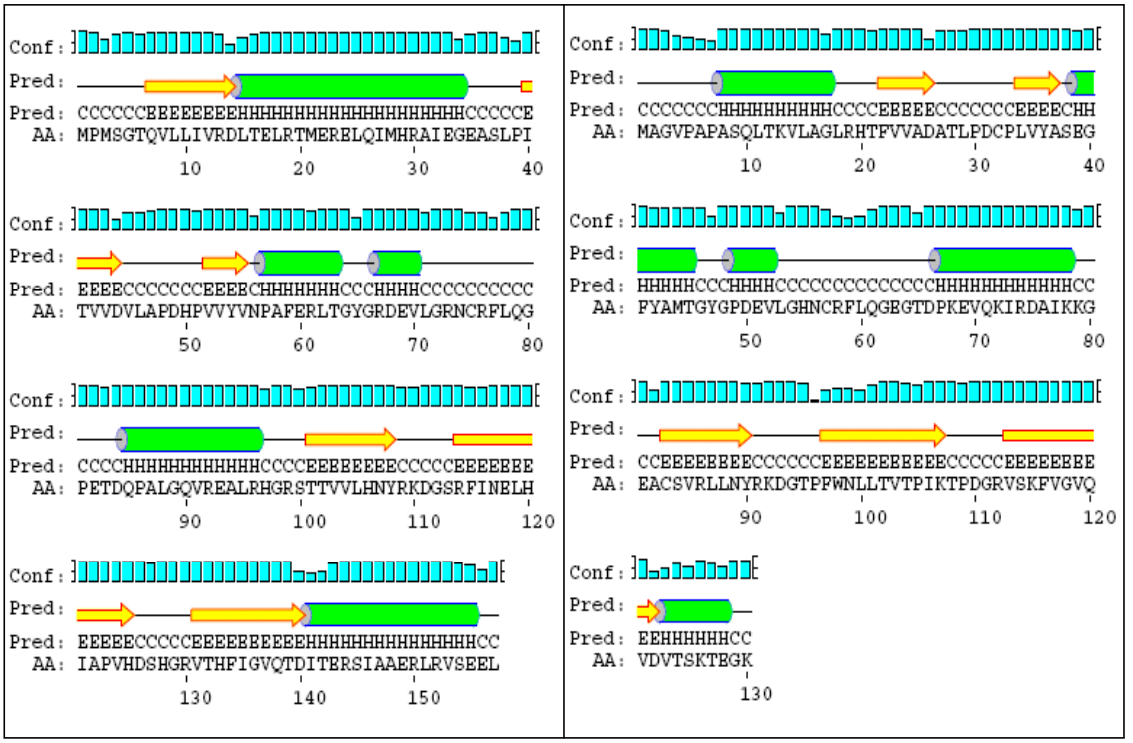


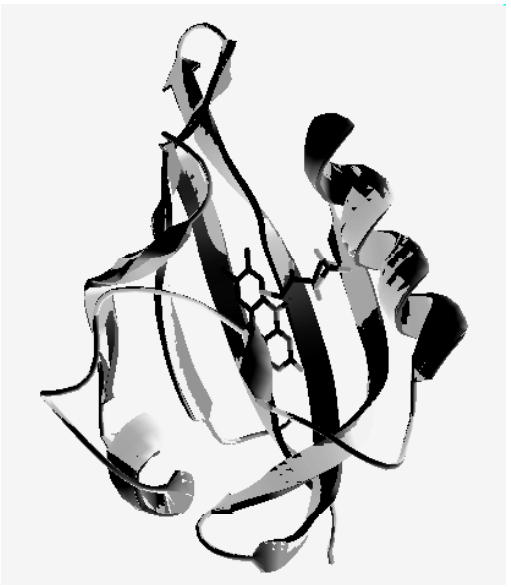
Figure 4.11 Secondary structure prediction of HT-Met1 LOV domain is shown in the left box and that of *C. reinhardtii* LOV1 is shown in the right box. A typical five β-strand arrangement comparable to *C. reinhardtii* LOV1 can be seen in HT-Met1 LOV.

Though the HT-Met1 LOV domain is highly diverse at the amino acid sequence level to any other characterized LOV proteins, structure modelling reveals a very good congruence between the three dimensional structure of the LOV domain from *C. reinhardtii* and the novel LOV domain from HT-Met1. The structure model built based on the sequence homology is shown in figure 4.12.

(a)

```
Ht-Met1L  15                        EAS LPITVVDVLA PDHPVVYVNP AFERLTGYGR
pdb1n91   17                        glr htfvvadatl pdcplvyase gfyamtgygp

Ht-Met1L                                 sssssss      ssssss  hhhhhh
pdb1n91                                  sssssss      ssssss  hhhhhh


Ht-Met1L  48    DEVLGRNCRF LQGPETDQPA LGQVREALRH GRSTTVVLHN YRKDGSRFIN
pdb1n91   50    devlghncrf lqgegtdpke vqkirdaikk geacsvrlln yrkdgtpfwn

Ht-Met1L                hh         hhh hhhhhhhhhh     sssssss ss    sssss
pdb1n91                 hh         hhh hhhhhhhhhh     sssssss ss    sssss


Ht-Met1L  98    ELHIAPVHDS HGRVTHFIGV QTDITE
pdb1n91   100   lltvtpiktp dgrvskfvgv qvdvts-

Ht-Met1L        sssssssss    ssssssss ssss
pdb1n91         sssssssss    ssssssss ssss
```

(b)



(c)

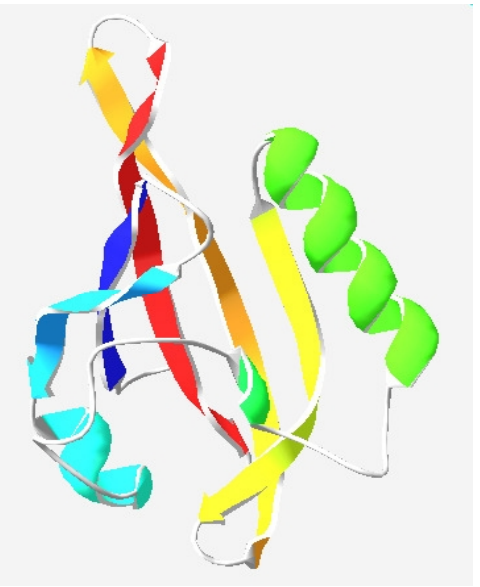


Figure 4.12 Modeling the three dimensional structure of the HT-Met1 LOV domain using the crystal structure of *Chlamydomonas reinhardtii* phototropin LOV1 (PDB entry: 1N9L) as template. (a) Sequence alignment of the LOV-core of HT-Met1 with the LOV1 domain of *Chlamydomonas reinhardtii* (Cr) phototropin (denoted as pdb1n91) with structure information ("h" denotes the helix and "s" denotes the sheet). (b) Overlay of the HT-Met1 LOV domain structural model (grey) and the template 1N9L (black). (c) Predicted 3-dimensional structure of HT-Met1 LOV. This structure is based on (b), after an additional energy minimization using the Gromos96 force field implementation of SwissPdbViewer.

4.2 A LOV domain from Elbe River metagenome

4.2.1 Construction of a fusion gene from a partial Elbe1 LOV domain

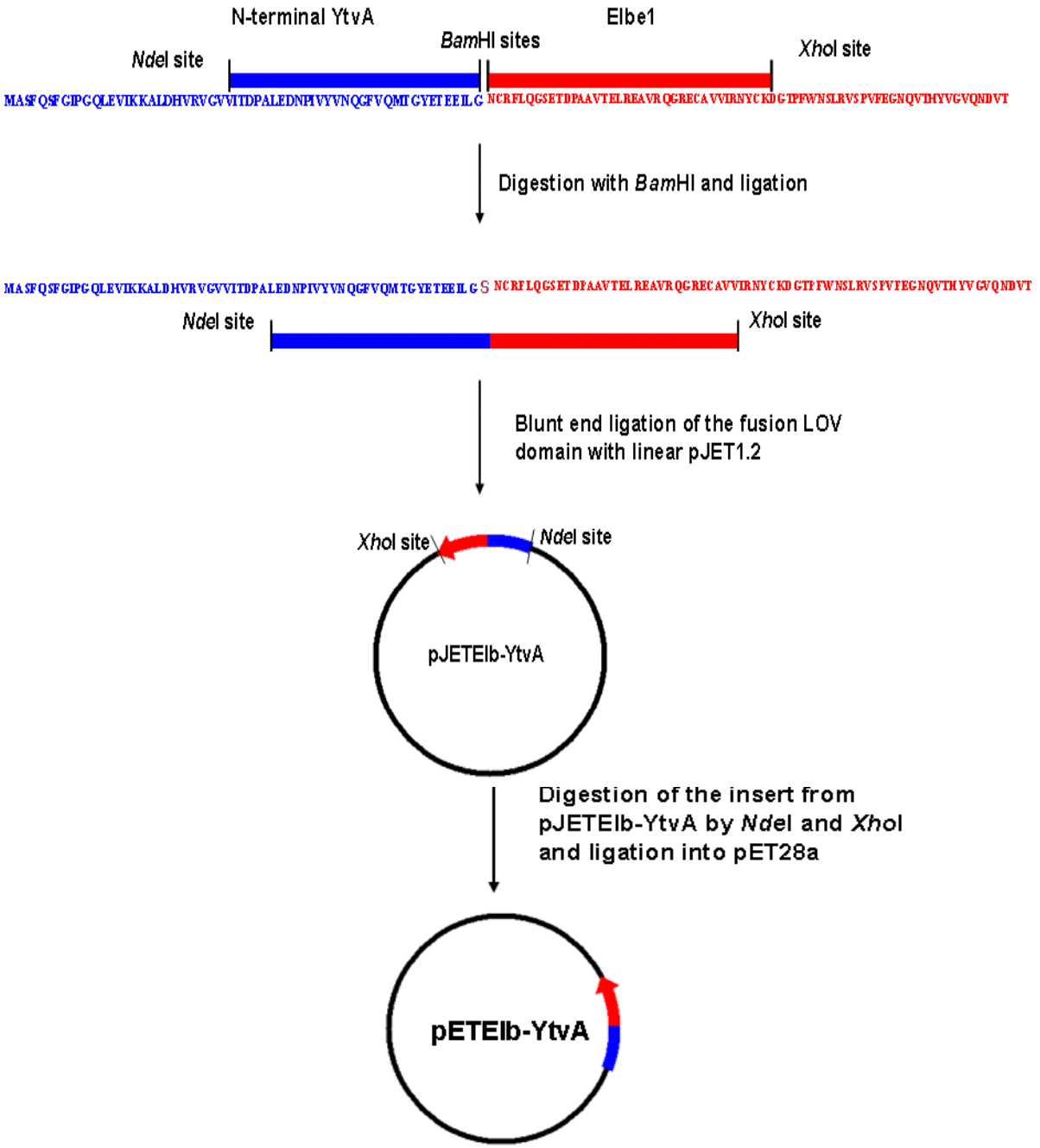
Between the two partial genes detected from and PCR-amplified from Elbe River metagenome (3.11.1), Elbe1 was selected to construct a complete synthetic LOV domain. Elbe1 comprised 69 amino acids, thus being longer than Elbe2 which was just 41 amino acids in length. The BLAST analysis showed that the nearest matches to Elbe1 available in the database were the multi-sensor hybrid histidine kinase from *Gemmata obscuriglobus* UQM 2246 (70% identity) and sensory box protein from *Methylococcus capsulatus* str. Bath (67% identity).

Since Elbe1 showed high sequence diversity in regions other than the conserved core region, an amplification of the whole gene from the Elbe River metagenome was not successful when attempted with several pairs of degenerate primers. However, a full-length LOV domain would be required to see its functional properties. For that reason, a complete LOV domain was constructed by fusing the N-terminal portion of the *ytv*A gene from *B. subtilis* with the Elbe1 fragment. Although YtvA was not its nearest match, the fusion of Elbe1 with a part of YtvA-LOV could retain all important amino acid residues responsible for the chromophore binding. In addition, the secondary structure prediction of the Elbe1-YtvA fusion revealed a secondary structure arrangement comparable to YtvA and other characterized LOV domains.

The DNA sequence from *ytv*A beginning with its start codon up to position 177 was amplified using a forward primer Bsub_fus_for and a reverse primer Bsub_fus_rev. The forward and reverse primers had 5'-end extensions including the restriction sites *Nde*I and *Bam*HI, respectively. The 207 bp sequence from Elbe1 was amplified using a forward primer Elb_fus_for and a reverse primer Elb_fus_rev. The forward primer had an extension on its 5'-end that included a *Bam*HI recognition site and an extension on the 5'-end of the reverse primer that included an *Xho*I site. The PCR products from both DNA templates were digested with *Bam*HI, purified and ligated to each other. The ligated product was blunt-end cloned into the vector pJET1.2. The insert DNA was digested from the vector and the linear fusion DNA

was excised from the gel and purified before ligating into the pET28a expression vector. In order to additionally clone it into the pET52 3C/LIC expression vector, the insert DNA was amplified from pJET1.2 using primers Fus_52_For and Fus_52_Rev, each having a 5'-end extension to create complementary overhangs for 3C/LIC cloning (2.4.3.3). A schematic presentation of the preparation of the fused gene is shown in figure 4.13. The fusion DNA contained a Serine (S) residue instead of commonly found Arginine (R) or Lysine (K) on the position -2 before the conserved cysteine because of the adjustment to accomodate a restriction site.

(a)



(b) MASFQSFGIPGQLEVIKKALDHVRVGVVITDPALEDNPIVYVNQGFVQMTGYETEEILGSNCRFLQ GSETDPAAVTELREAVRQGRECAVVIRNYCKDGTPFWNSLRVSPVFEGNQVTHYVGVQNDVT

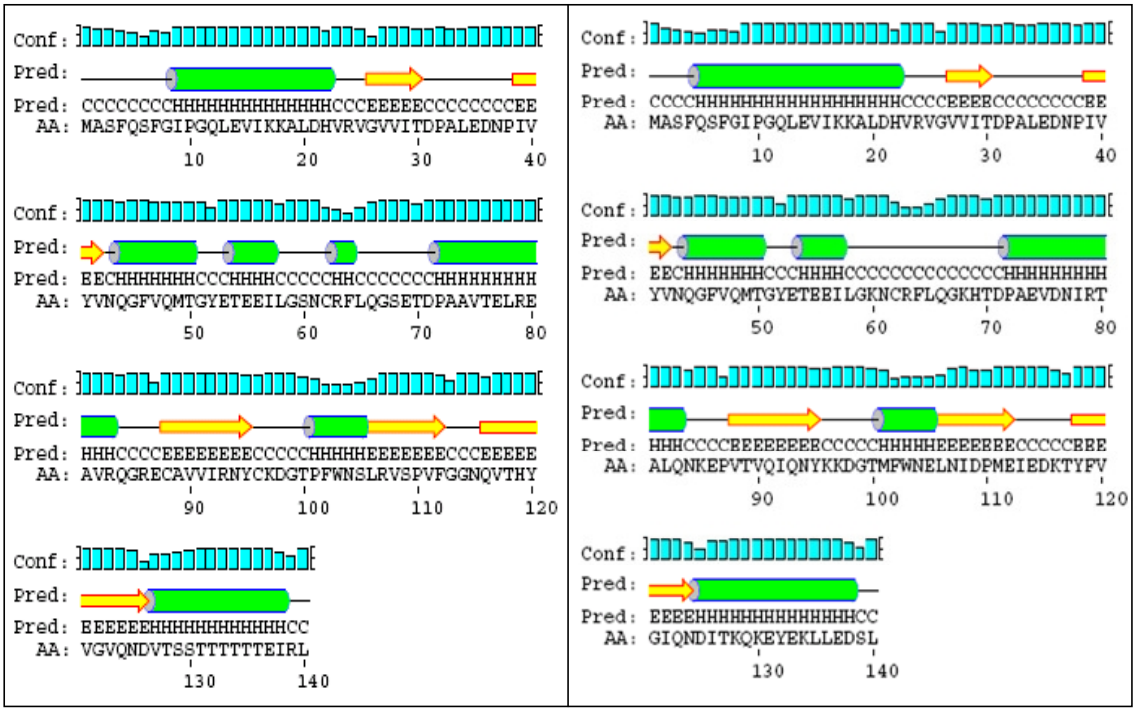
(c)

```
ElbeFusion      MASFQSFGIPGQLEVIKKALDHVRVGVVITDPALEDNPIVYVNQGFVQMTGYETEEILGS 60
B.subtilis      MASFQSFGIPGQLEVIKKALDHVRVGVVITDPALEDNPIVYVNQGFVQMTGYETEEILGK 60
                ************************************************************.

ElbeFusion      NCRFLQGSETDPAAVTELREAVRQGRECAVVIRNYCKDGTPFWNSLRVSPVFEGNQVTHY 120
B.subtilis      NCRFLQGKHTDPAEVDNIRTALQNKEPVTVQIQNYKKDGTMFWNELNIDP-MEIEDKTYF 119
                *******..**** * ::* *::: .  :* *:** **** ***.*.:.* :* :: *::

ElbeFusion      VGVQNDVT 128
B.subtilis      VGIQNDIT 127
                **:***:*
```

Figure 4.13 (a) Schematic outline of the construction of a fused gene from the N-terminal partial sequence of *ytvA* from *B. subtilis* and Elbe1. The line and sequences in blue correspond to *ytvA* while that colored in red correspond to Elbe1. The expression vector pET28a was linearized by *Nde*I and *Xho*I digestion and the fusion LOV domain construct was digested using the same enzyme pair from pJETElbe-YtvA and separated from agarose gel before ligation into pET28a. (b) Amino acid sequence of the LOV domain constructed from the N-terminal portion of YtvA-LOV (highlighted in yellow) fused with C-terminal Elbe1 (highlighted in gray). A serine (at position 60) residue which was not part of either fragment is shown without color coding. (c) Alignment of the amino acid sequences of Elbe fusion LOV domain to the LOV domain from *B. subtilis*-YtvA.

4.2.2 Expression of a synthetic fusion protein from the Elbe River metagenome

The fusion constructs cloned in pET28a and pET52 3C/LIC were attempted to express under standard conditions with IPTG induction. However, several attempts did not produce an FMN-binding functional protein. Although the two fragments of DNA used for the construction of the fusion protein were from diverse genes, the secondary structure prediction using PSIPRED (Figure 4.14) showed five β-sheets as in HT-Met1 and other characterized LOV domain proteins. All the functional residues are in correct position except for a serine (S) in place of an arginine (R) residue at amino acid position 60.



Legend:

Helix

Strand

Coil

Confidence of prediction

Pred: **Predicted secondary structure**

AA: **Target sequence**

Figure 4.14 Predicted secondary structure of the Elbe1 fusion LOV domain (left box). In the right box the highly similar secondary structure of the YtvA LOV domain is shown.

4.2.3 Conversion of serine into arginine

Since the construction of a synthetic protein from the part of Elbe1 and YtvA LOV yielded a serine residue (position 60) at a critical position for the function of a LOV domain, it was necessary to convert this residue into the functionally conserved residue. The position of that residue is taken by a lysine in YtvA, an arginine in HT-

Met1 LOV and a histidine in *C. reinhardtii* LOV1. The serine residue was first attempted to convert into argine by site directed mutagenesis (2.4.6). Conversion of serine into argenine requires changing of TCC (as present in the coding frame of Fusion LOV) into CGC. Mutation of both nucleotides in a single reaction was not successful, for that reason the residues were changed first into proline (CCC) using Fus_S2P_F and Fus_S2P_R primers. Following, proline was changed into argenine (CGC) using Fus_M_For and Fus_M_Rev primers.

Both of the fusion LOV constructs (in pET52) with the changed amino acid (proline or arginine) at position 60 were transformed into *E. coli* ER256 expression cells, grown overnight for expression supplemented with 0.5 mM of IPTG. The protein was purified using His-tag affinity resins. Expression of the recombinant protein was detected in the soluble and purified fraction with western blot analysis (Appendix D). However, the purified protein did not show an absorption pattern typical for functional LOV domains. The absorption behavior of the protein was same, even after a prolonged incubation at dark for three days.

4.3 Cloning and expression of a LOV domain protein from the Whale Fall metagenome

A BLAST search in the environmental sequence deposits of the NCBI database using the LOV domain of YtvA as a reference yielded several hits. Several of the project groups were contacted for the availability of the source clones. Among them was the Whale Fall metagenome [124] (see chapter 5), for which a number of putative LOV domains were found in the BLAST analysis. A related shot gun clone containing 1879 bp of DNA was obtained by courtesy of Susan Green Tringe (Joint Genome Institute, USA). A coding frame from position 1220 to 1801 (according to the nucleotide positions in the clone) contained a putative sensor protein that showed the sequential feature of a LOV domain. Following the LOV domain a helix-turn-helix-LuxR (HTH_LuxR) domain was present on its C-terminal as an output domain. The coding frame (whole gene) was amplified by PCR and cloned in pET28a expression vector using primer pairs 5784F and 5784R. However, the recombinant protein was not obtained in functional form as expected in a typical LOV domain case. In this

case also the western blot analysis (Appendix D) showed the expression of His-tagged recombinant protein in purified fraction but no photochemical reactivity could be induced. The further truncated LOV-domain-only region cloned in pET28a using primer pairs 5784Trim_for and 5784Trim_rev also did not bind the chromophore. Secondary structure prediction of the putative LOV domain protein from whale fall metagenome using PSIPRED revealed only three β-strands instead of five β-strands typical for the so far characterized LOV domains (Figure 4.15).

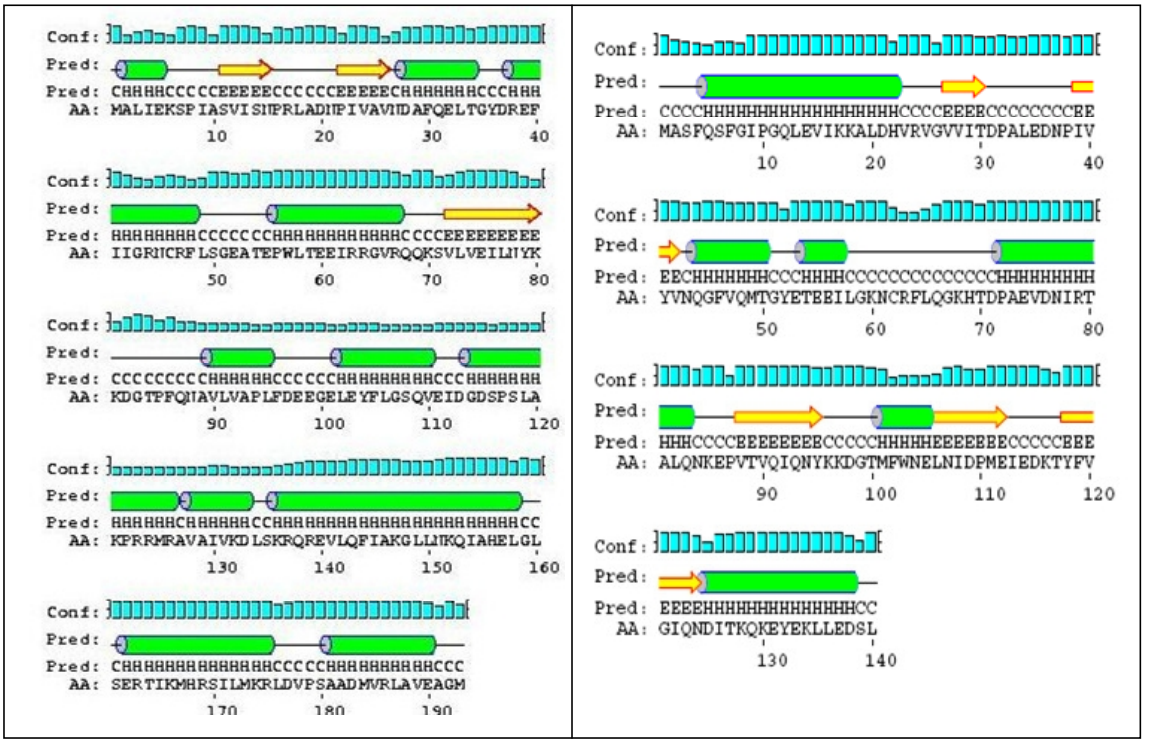


Figure 4.15 Predicted secondary structure of LOV domain containing gene from whale fall metagenome (on the left box) and that of YtvA LOV domain from *B. subtilis* (on the right box).

4.4 Discussion

4.4.1 HT-Met1 LOV: A novel LOV domain from a metagenomic approach

A novel multi domain protein (HT-Met1) with 1204 aa in length, which contained a LOV domain, several PAS domains, a histidine kinase and a response regulator domain, was obtained from the thermophilic fraction of a soil sample using the newly developed DNA microarray approach (Chapter 3). The LOV domain revealed a canonical conserved motif (GRNCRFLQG) and other important residues required for photoactivity of typical LOV domains. The gene was similar to the putative signal transduction histidine kinase protein from *Methylobium petroleiphilum* PM1. At the LOV domain level a similarity of 58% was found to *Kineococcus radiotolerans* SRS30216 and of 57% to that of *M. petroleiphilum* PM1.

None of the putative LOV domain-containing genes from the genera closely placed to HT-Met1 in the phylogenetic tree (Figure 4.4) has yet been characterized functionally. Not only the sequence and genetic composition, but also the photochemical properties of the HT-Met1 showed novel features. Due to the size of the protein, the expression and purification as a full length protein was considered difficult. Thus, only the putative LOV domain was cloned and expressed in *E. coli* yielding the typical three peaked absorption spectrum and the two peaked fluorescence spectrum of LOV domains. Surprisingly, the dark recovery of the photo-bleached protein from its light state was so fast that the generation of a fully photo-bleached stage was not possible. The dark recovery lifetime is much faster in comparison to YtvA-LOV from *B. subtilis* (3900 s) [234], and PpBS1 from *P. putida* (ca. 40 hours, [45]. This kinetics is comparable to PpSB2-LOV, the second LOV from *P. putida* (114 s) [164] and to the eukaryotic LOV1 domain from *Chlamydomonas reinhardtii* phototropin (204 s) [46].

A base catalyzed back reaction in which a proton is taken away from N(5) of the flavin isoalloxazine ring, thus breaking the Cys-S-C(4a) covalent bond to regenerate the C(4a)-N(5) double bond occurs in the interacting part of the LOV domain during dark recovery [41]. Surface exposed histidine residues in *Avena sativa* [235] and a histidine residue located near the conserved cysteine residue in *P.*

putida-PspSB2 (Figure 4.18) were supposed to be strong proton acceptors that could accelerate the recovery reaction [164]. The presence of eight histidine residues in HT-Met1-LOV might contribute to the fast dark recovery by collective or individual activity of any of the histidines as a proton accepting group. However, bacterial LOV proteins from *B. subtilis* YtvA and *R. sphaeroides* show slow recovery kinetics despite having several histidine residues in their LOV domains. One may propose that the position of the histidine is of importance for its ability to abstract the proton. Further studies are required for a verification of this proposal.

It was shown that a double mutation of lysine to argenine (position 2 of oat phot) and isoleucine to valine (at position 16 of oat phot) contributed greatly for the fast dark recovery [58]. As suggested by the authors, the isoleucine is probably more important for stabilizing the signaling intermediates because the isoleucine side chain provides direct steric support for generation of C-S bond between carbon (4a) and the sulfhydryl group of the reactive cysteine. This stability was believed to be a reason for longer dark recovery. At the position equivalent to I16 of oat one finds a valine in HT-Met1 LOV and it can be hypothesized that the presence of valine at this point in HT-Met1 LOV could have contributed for fast dark recovery. But if taken *P. putida* PpSB1 and PpSB2 LOV into account, both the domains have isoleucine at that position while their kinetic recovery is drastically different to each other, the former quite slow and the latter much faster [45;164]. It suggests that the explanation of the difference in the recovery kinetics only on the basis of a single amino acid, such that an isoleucine or valine residue at that position is insufficient, and more studies are needed to fully explain the mechanism behind it.

In order to gain more insight into the issue of fast dark recovery found with the LOV domain from the thermophile soil sample, HT-Met1 LOV could be used as a model system to apply mutagenesis on its histidine residues which have been proposed in other reports as being important for a fast dark recovery of LOV domains.

A recent study on proteorhodopsins (PRs) revealed an interesting effect such that proteins from a depth of 75-100 m in the Mediterranean Sea underwent a slower photocycle than those collected from a depth of 12 m from the same geographical region [135;236]. Interestingly, proteorhodopsins found in the surface water are tuned more to the green light whereas the ones found at the depth (75 to 120 m) have shown absorption maxima at blue region [237;238]. These blue-absorbing PRs required at least an order of magnitude more to complete their photocycle compared to the green-absorbing PRs. This slower rate was attributed to a reduced need to absorb photons owing to the low photon density of the deeper waters [237]. It has been found that a single amino acid (glutamine or leucine at position 105) is crucial to photocycle rate determination among green and blue absorbing proteorhodopsins [236;239]. The difference in photocycle rate of BPR and GPR has been explained as an adaptation to the different light conditions present in the marine environments [236]. Though more detailed investigations are still required, one might speculate that microbes that are more susceptible to a higher light exposure have a faster photocycle than those which experience a lower light intensity.

Another interesting feature of HT-Met1 LOV is that an otherwise fully conserved tryptophan residue is absent at position 114 (number according to the HT-Met1 LOV as in figure 4.16).

```
Ht-Met1   MSGTQVLLIVRDLTELRTMERELQIMHRAIEGEASLPITVVDVLAPDHPVVYVNPAFERL 60
PpSB2     MINAKLLQLMVEHSND--------------------GIVVAEQEGNESILIYVNPAFERL 40
C_reinh   APASQLTKVLAGLRHT---------------------FVVADATLPDCPLVYASEGFYAM 39
PpSB1     MINAQLLQSMVDASND--------------------GIVVAEKEGDDTILIYVNAAFEYL 40
B_sub     GQLEVIKKALDHVRVG---------------------VVITDPALEDNPIVYVNQGFVQM 39

HT-Met1   TGYGRDEVLGRNCRFLQGPETDQPALGQVREALRHGRSTTVVLHNYRKDGSRFINELHIA 120
PpSB2     TGYCADDILYQDCRFLQGEDHDQPGIAIIREAIREGRPCCQVLRNYRKDGSLFWNELSIT 100
C_reinh   TGYGPDEVLGHNCRFLQGEGTDPKEVQKIRDAIKKGEACSVRLLNYRKDGTPFWNLLTVT 99
PpSB1     TGYSRDEILYQDCRFLQGDDRDQLGRARIRKAMAEGRPCREVLRNYRKDGSAFWNELSIT 100
B_sub     TGYETEEILGKNCRFLQGKHTDPAEVDNIRTALQNKEPVTVQIQNYKKDGTMFWNELNID 99

HT-Met1   PVHDSHGRVTHFIGVQTDITERSIAAERLRVSEEL 155
PpSB2     PVHNEADQLTYYIGIQRDVTAQVFAEERV------ 129
C_reinh   PIKTPDGRVSKFVGVQ------------------- 115
PpSB1     PVKSDFDQRTYFIGIQKDVSRQV------------ 129
B_sub     PMEIEDK--TYFVGIQNDITKQKEYEK-------- 115
```

Figure 4.16 Alignment of the LOV domain from HT-Met1 to the LOV domains from *P. putida*, *Chlamydomonas reinhardtii* and *B. subtilis*. PpSB2: *P. putida* SB2 LOV; C_reinh: *C. reinhardtii* phot LOV1; Ppsb1: *P. putida* SB1 and B_sub: *B. subtilis* YtvA. Histidine residues are shown in red and the conserved tryptophan is shown in blue.

The HT-Met1 protein contains four PAS domains and three PAC (PAS associated C-terminal motif) domains followed by a histidine kinase (HK) and a response regulator. The role of multiple PAS domains in the same protein is unclear. The different PAS domains in HT-Met1 share less than 22% identity among each other. It was the third PAS domain with an immediately following PAC domain (from aa407 to aa561) which showed the blue light dependent photochemistry (4.1.5). The high similarity of the LOV-domain from HT-Met1 to the typical LOV domains' structure is also obvious from the overlay of the modelled three-dimensional structure of HT-Met1-LOV with the crystal structure of the LOV-domain from the *C. reinhardtii* phototropin (Figure 4.12) which yielded a practically complete congruence.

Similar to the Jα–helix of other LOV proteins [49;164;240;241], a 25 amino acid extension with α helical structure is found at the C-terminal end of HT-Met1 by secondary structure prediction using PSIPRED (Figure 4.17). The C-terminally

adjacent PAC domain of the HT-Met1 LOV is located 62 amino acids away and the histidine kinase domain is 262-amino acid away from its Jα-helix. In most cases, Jα–helices act as linker between the LOV core and the output domains, but no direct link of the Jα-helix to the kinase domain is assumed at the sequence level in HT-Met1 when the large distance between both domains (position of LOV-domain aa442-546 and position of HK aa834-1054) is taken into account. Moreover, both LOV and HK domains are flanked by further three (PAC-PAS-PAC) domains.

(a)

```
B_Sub    MASFQSFGIP----------GQLEVIKKALD-HVRVGVVITDPALEDNPIVYVNQGFVQM
HT_LOV   MSGTQVLLIVRDLTELRTMERELQIMHRAIEGEASLPITVVDVLAPDHPVVYVNPAFERL

B_Sub    TGYETEEILGKNCRFLQGKHTDPAEVDNIRTALQNKEPVTVQIQNYKKDGTMFWNELNID
HT_LOV   TGYGRDEVLGRNCRFLQGPETDQPALGQVREALRHGRSTTVVLHNYRKDGSRFINELHIA

B_Sub    PMEIED--KTYFVGIQNDITKQKEYEKLLEDSLTEITALSTPIVPIRNGISALPLVGNLT
HT_LOV   PVHDSHGRVTHFIGVQTDITERSIAAERLRVSEELYRSVALSISDGLMVVTSGSYVVAVN

B_Sub    EERFNSIVCTL
HT_LOV   PAACRILDATM
```

(b)

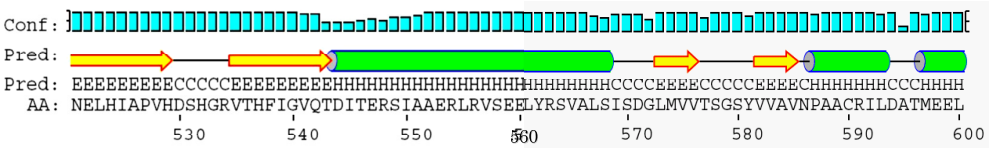


Figure 4.17 Jα helix from HT-Met1 LOV (a) Sequence alignment from the LOV domain from *B. subtilis* YtvA (B_sub) and HT-Met1 (HT_LOV) showing the α helical connector (underlined) (b) Secondary structure prediction of the C-terminal region of the LOV domain from HT-Met1 showing the connector helix as predicted by PSIPRED. The Jα helix ends at amino acid position 569 whereas the histidine kinase domain starts from amino acid position 834 flanked by 11 beta sheets and 10 helices (not shown in the figure).

4.4.2 The two component signaling module

The LOV domain of the HT-Met1 is followed by a PAC-PAS-PAC motif, a histidine kinase and finally a response regulator, which identifies HT-Met1 as a two component system. In prokaryotes, the hybrid kinases that contain also response regulators are rare, but in eukaryotes hybrids of kinase and response regulator are the common pattern [242]. Two component signal transduction is the primary signal transduction mechanism used to conduct global regulation of cell responses to changes in the environment which occurs in three consecutive processes, viz. autophosphorylation of the His-kinase, phosphotransfer to the response regulator and dephosphorylation of the response regulator. The role of multiple sensing domains in a single protein like HT-Met1 is difficult to explain. HT-Met1 was derived from a thermophilic soil fraction of the garden soil ca. 10 cm below the surface. Probably the different sensing domains integrate a variety of stimuli or changes around the environment, more importantly on the top soil where the environmental conditions are dynamic and changes are quite often. A LOV-mediated phosphorylation of the kinase domain was recently reported from *P. syringae* pv. *tomato* LOV [168] in which blue light was demonstrated as stimulus to upregulate the two component signal transduction system consisting of a LOV-Hiskin-RR module. In another investigation, the activation of a LOV domain-regulated kinase was shown to increase the infectivity of *Brucella abortus* [166]. Also the LOV domain from HT-Met1, for which already its blue light sensing properties have been demonstrated, might act as a stimulus in autophosphorylation of histidine kinase domain which in turn phosphorylates the regulator domain. Since a source species, from which HT-Met1 was derived, could not be identified, a deeper understanding of the biological role of HT-Met1 could be obtained by studying the biochemical property and the biological role of the similar protein Mpe_A2494 from *M. petroleiphilum* MP1.

4.4.3 Gene neighborhood of *ht-met1*

The nearest match for Ht-Met1 which has 52% similarity at the protein level is *M. petroleiphilum* PM1. Other genes, coding a putative ferrochelatase and a heat shock protein (S4 paralog) that follow HT-Met1 are also identical to corresponding

genes of *M. petroleiphilum* PM1. The ferrochelatase from the pWTHLOV clone was found to be similar and conserved in many bacteria, that of *Acidovorax delafieldii* 2AN (71%) and *M. petroleiphilum* PM1 (69%) being the nearest match. The heat shock protein showed highest similarity to *M. petroleiphilum* PM1 (75%) and *Polaromonas* sp. JS666 (74%). *M. petroleiphilum* is aquatic methylotroph and degrades MTBE (Methyl tert-Butyl Ether) [243;244], which contaminates water as it dissolves easily and is generally more resistant to natural biodegradation than other gasoline components.

An operon prediction analysis using MicrobesOnline Operon Predictions [245] shows that the putative signal transduction gene (that contains the putative LOV domain) in *M. petroleiphilum* is in the same operon with both, the ferrochelatase- and the heat shock protein-coding gene (Figure 4.18). The signal transduction histidine kinase gene (Mpe_A2494) is overlapped with ferrochelatase (-4 nucleotides). Similarly, the ferrochelatase gene is also overlapped with the heat shock protein gene (-4 nucleotides). In the metagenome derived clone that harbors the blue light receptor gene, the HT-Met1- and the ferrochelatase genes are not overlapping, but are separated by 44 nucleotides. This latter gene, on the other hand, is overlapped with the heat shock protein (-84 nucleotides). The presence of these three genes in the same operon suggests that they might be regulated together, HT-Met1 and probably the LOV domain being the main sensor to activate the protein regulation. The similar gene neighborhood module in *ht-met1* clone and *M. petroleiphilum* indicates that a similar functional pathway can exist in the environmental strain that carries the *ht-met1* gene. However, functional and sequence homology could not be determined for the other 23 ORFs predicted from the 5.2 kb clone from which *ht-met1* was isolated.

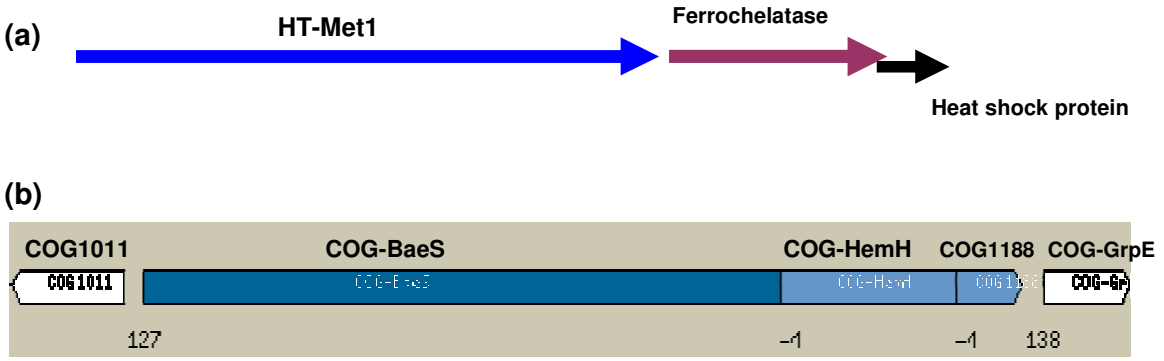


Figure 4.18 Gene arrangements in *M. petroleiphilum* signal transduction gene cluster and HT-Met1 clone (pWTHLOV). (a) Arrangement of HT-Met1, ferrochelatase and heat shock protein in metagenome clone pWTHLOV, (b) Graphical view of the predicted operon that consists of the signal transduction protein (Mpe_A2494), a ferrochelatase (Mpe_A2495) and a heat shock protein (Mpe_A249) in *M. petroleiphilum*; the arrow head shows the direction of the transcription of the operon (MicrobesOnline Operon Predictions, Price et al., 2005). COG1011: hydrolase (Mpe_A2493), COG-Baes: signal transduction protein (Mpe_A2494), COG-HemH: ferrochelatase (Mpe_A2495), COG1188: ribosome associated heat shock protein (Mpe_A2496) and COG-GrpE: putative heat shock protein (Mpe_A2497).

There are several reports about the role of a ferrochelatase to catalyse the addition of iron (II) to protoporphyrin IX to form a heme product [246;247](Figure 4.19). Protoporphyrins without a central metal ligand are strong photosensitizers [26], and the conversion of protoporphyrin IX to a heme product or less reactive iron-porphyrins can be an important survival strategy for the microorganisms.

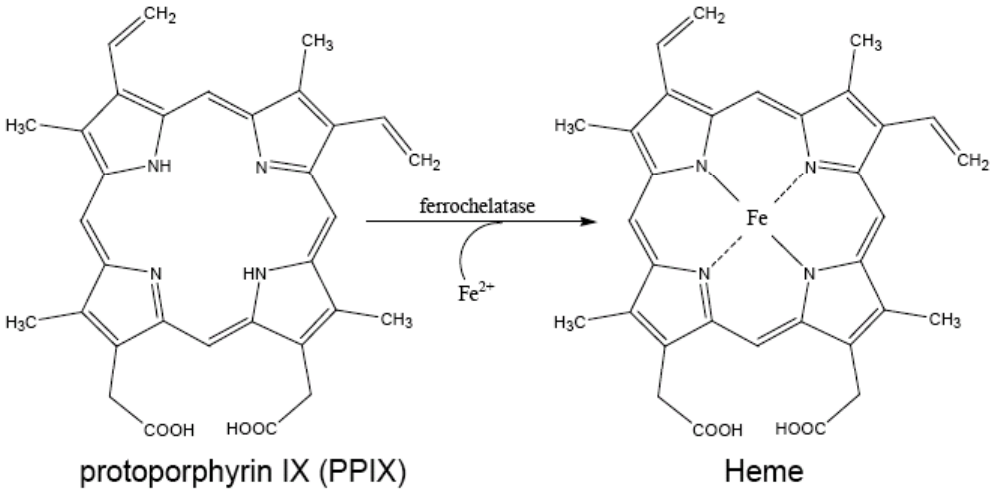


Figure 4.19 Catalytic step of the heme-biosynthesis pathway from protoprophyrin IX.

As supported by the gene arrangement and regulon prediction (Figure 4.20), one of the possible roles of the HT-Met1-ortholog Mpe_2494 protein (and its LOV domain) in *M. petroleiphilum* PM1 may be to regulate ferredoxin activity which takes part in the metabolism of porphyrins and is potentially regulated by the LOV-HK-RR protein. Such interaction/regulation would be advantageous, since ferredoxins mediate electron transfer required for various cellular activities and also participate in iron incorporation [248;249].

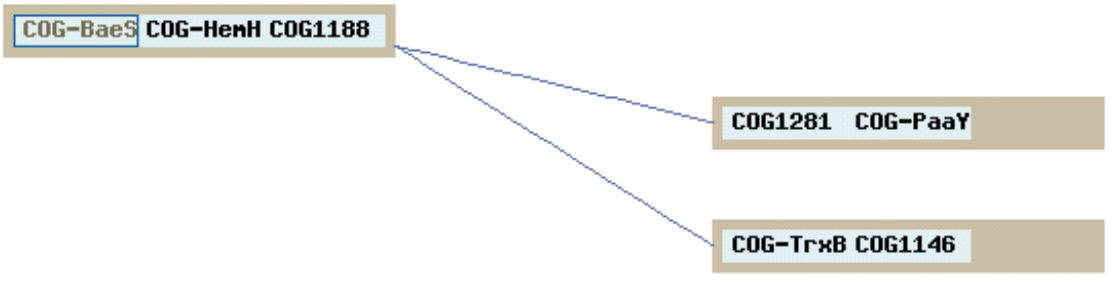


Figure 4.20 Association of the operon containing the coding frames for the signal transduction histidine kinase, ferrochelatase and heat shock proteins in *M. petroleiphilum* PM1 to other two operons. One of the operons contains ferredoxin (COG1146). (Regulon prediction: www.microbesonline.org)

COG (cluster of orhtologous groups) symbols:

COG-Baes: Signal transduction histidine kinase (Mpe_A2494)

COG-HemH: Ferrochelatase (Mpe_A2495)

COG1188: Putative heat shock protein (Mpe_A2496)

COG1281: Redox regulated molecular chaperon (Mpe_A2014)

COG-PaaY: Carbonic anhydrase/Acetyltransferase (Mpe_A2015)

COG-TrxB: Thioredoxin reductase (Mpe_A1496)

COG1146: Ferredoxin (Mpe_A1497)

4.4.5 Expression of HT-Met1 LOV in soluble form

HT-Met1 protein was deposited in inclusion bodies when over-expressed under standard conditions. Among several strategies applied, heat and salt stress and a construct in a different vector (pET28) improved solubility.

In vitro refolding has already been carried out for many overexpressed recombinant proteins that were found in inclusion bodies [250-252], but for a chromophore-bearing protein this is a complex process that often results in a very low efficiency of reconstitution [253;254]. For that reason it was attempted to obtain in vivo solubilization of HT-Met1 LOV protein, which in the present work was successful. The application of stress conditions and the generation of various types of constructs for the target recombinant protein in different expression vectors finally enhanced the production of the soluble recombinant protein.

Protein deposition in bacteria occurs in the form of inclusion bodies. The term inclusion body has been applied to the intracellular loci into which aggregated proteins are sequestered. In over-expression techniques the protein production occurs with a high translational rate and thereby provides the cells with a continuous supply of unfolded polypeptides [255]. Under these conditions, aggregation of protein dominates folding, and the stability of the aggregated form is often higher than that of the native structure. It is widely assumed that inclusion bodies form as a consequence of the self-assembly of non-native monomers into growing polymers (Figure 4.21a). An alternative model proposes (Figure 4.21b) that inclusion bodies

are aggregates of aggregates. In other words, inclusion bodies arise from the coalescence of individual aggregates into a single or limited number of loci [256].

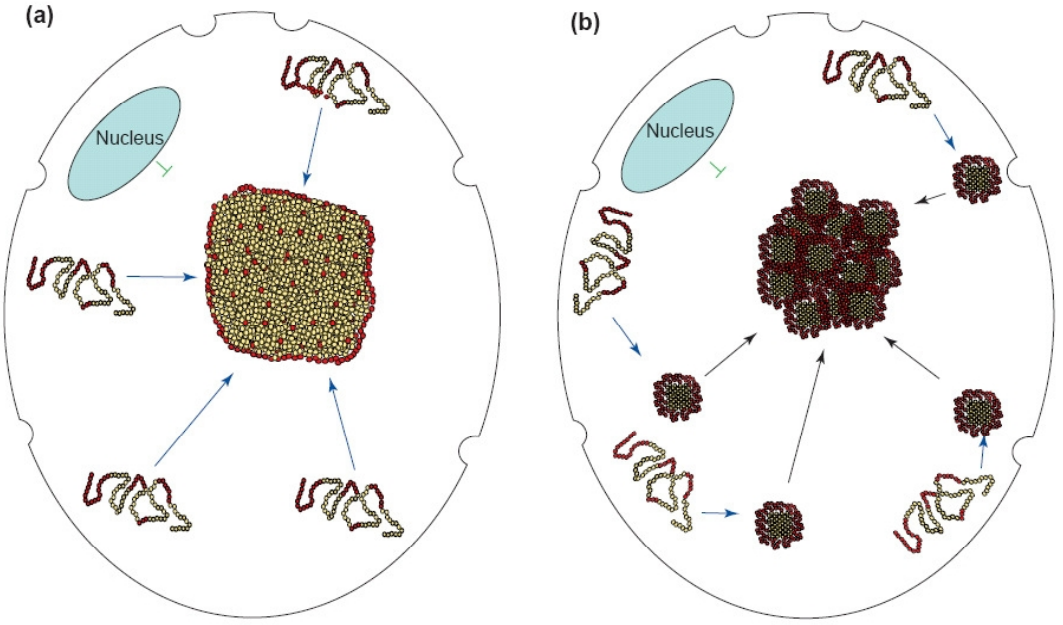


Figure 4.21 Models of inclusion body formation in a eukaryotic cell [256]; similar processes can be assumed for prokaryotic cells. (a) Direct deposition of the aggregated protein to the inclusion body. (b) Aggregation of the protein monomers in small aggregates which are then delivered to the inclusion body.

Bacterial inclusion bodies are generated by the failure of chaperones and preoteases to either fold or degrade unfolded or misfolded polypeptides, probably because their limiting amounts in recombinant cells when protein synthesis is directed at high rates [257] (Figure 4.22).

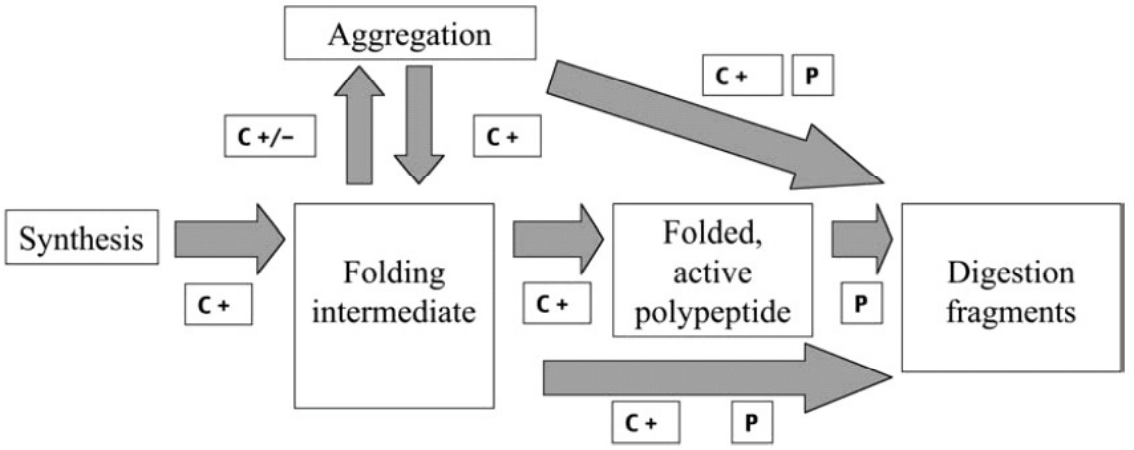


Figure 4.22 Fate of a cell protein after synthesis. Chaperons (C) play an important role in folding or unfolding of the synthesized proteins and proteases (P) degrade the misfolding-prone or excessive proteins. Chaperones have alternative but also antagonistic (+/–) physiological roles in protein managing [257].

Successful protein folding is a multistep process that results in a protein structure often representing the lowest free energy state. To achieve this process in a cellular environment, which predominantly favors aggregation and nonproductive folding, the assistance of protein complexes called chaperones is required [258;259].

Heat shock and salt stress techniques applied have improved the solubility of HT-Met1 LOV protein to some extent. A construct of the same DNA in pET28 was much more successful in obtaining a great proportion of the expressed protein in the soluble form. The success in improving the solubility of the expressed HT-Met1 LOV protein in pET28 was dependent on two factors, one was the expression temperature and the second was probably the position of the affinity tag (6X his-tag). The N-terminal 6X his-tag in pET28 is smaller compared to the C-terminal 10X his-tag in pET52. Moreover, a few extra fusion fragments were present on both sides of the protein derived from the expression vector in pET52 construct, whereas in pET28 construct the C-terminus was free from any externally fused amino acids.

Although a simple ‘‘heat-shock protocol’’ is sometimes very effective in enhancing the solubility of expressed proteins in cell extracts, it does not work on every protein that is overexpressed in *E. coli*. Indeed, even the same protein having mutations at different sites can behave quite differently. In comparison to other induction conditions, a heat shock and salt stress enhanced the solubility of the expressed HT-Met1 LOV protein in vivo. Bacterial chaperons, produced under heat and other conformational stress conditions, are involved in the conformational processing of a minor fraction of cell polypeptides and can also assist recombinant proteins [257].

The application of the heat shock protocol was followed since solubility of several recombinant proteins was shown to be enhanced in *E. coli* under heat stress conditions [260;261]. It has previously been shown that adding ethanol to the growth media stimulates the expression of heat-shock proteins in a manner similar to that of a 42 ^{o}C heat shock [262]. In fact, this approach has been reported to increase the solubility of an overexpressed protein [263]. However, when treating the media with 4% ethanol before induction, the HT-Met1 LOV protein was not obtained in soluble form. This suggests that the mechanism may be more complicated than a simple chaperone induction. Another method, called salt stress, applied here by the addition of NaCl to the medium aimed at generating a high external osmotic pressure. Bacteria adapt to high external osmotic pressure by accumulating small organic compounds known as osmolytes. These osmolytes can act as “chemical chaperones” by increasing the stability of native proteins and possibly assisting the refolding of unfolded polypeptides [261].

Chapter 5

Habitat-based analysis and phylogenetic relation of metagenomic LOV domains

5.1 LOV domains from a metagenome survey

Though a significant portion of sequenced genomes contain LOV domain proteins, a distinct functional role of the LOV domains in prokaryotes has not yet established. In fact, a community-level analysis can shed more light on the functional roles of the sensor proteins in the environment, as microorganisms living together in an interacting community share to a great deal their metabolic activities and functional pathways. Analysis of metagenomic sequence data can provide more insight into the ecological relevance and habitat-specific information in relation to LOV domain proteins. Environmental DNA samples' sequencing projects have already deposited a significant amount of information to the databases, which has been used as promising source for new enzymes, chemical compounds, various metabolic processes and pathways [264;265]. This chapter presents a sequence-based computational screen of the available metagenomic samples in order to obtain an insight into the distribution and significance of LOV domain containing proteins in various sample environments.

In order to find LOV domain-containing genes in various metagenomes, the reference LOV domain from *B. subtilis* was BLASTed (using tBLASTn or BLASTp) against more than 43 million ORFs and about 17 billion base pairs of metagenomic DNA sequences deposited in CAMERA, IMG/M and NCBI databases. More than 200 metagenomic sequencing projects have been listed in the Genomes OnLine Database v 3.0 by November 2009 (www.genomesonline.org), and sequence deposits of more than 52 metagenome projects are available in the databases, of which the Global Ocean Sampling Expedition (GOS) provides the largest amount of information.

Altogether, 82 GOS sites including the Sargasso Sea project and 52 microbial community samples sites were investigated. Among the available sequenced metagenome samples, 11 project samples from various geographical locations contained putative LOV domain sequences (Figure 5.1).

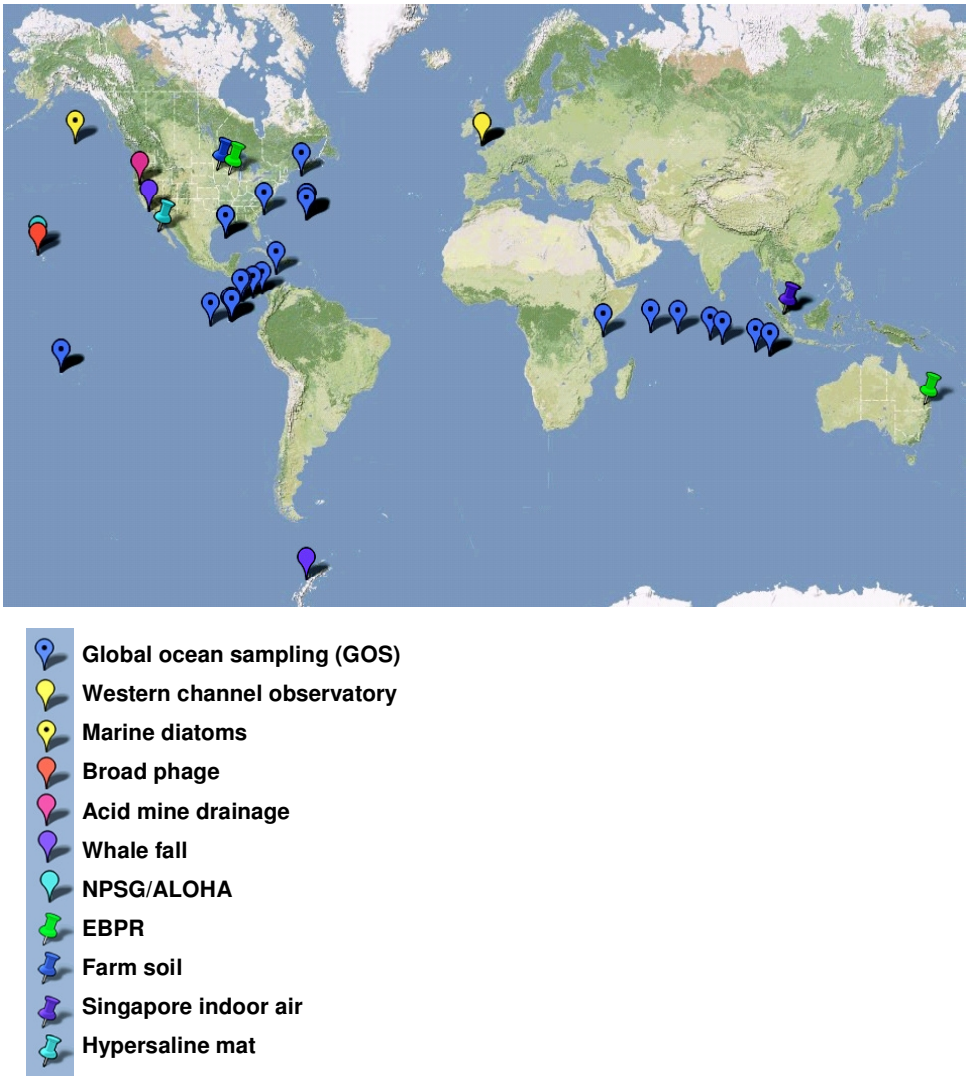


Figure 5.1 Geographical locations of various metagenomic sample sites from where LOV domain-containing genes could be identified. Colored marks denote the metagenome projects.

Initially, the BLAST analysis was carried out using an e-value >-10. Though 330 hits were obtained, the BLAST result could not include all putative sequences with that e-value as further putative LOV domain sequence hits were available when BLAST was carried out at default parameter. To include more sequences that contain most of the conserved residues of the typical LOV domain the BLAST results were relaxed to an e-value of >-5. Under all conditions, short sequences (<85 amino acids) and sequences that did not contain the conserved residues responsible for binding the FMN chromophore were removed. In this way, a total of 231 LOV domain-containing sequences were obtained (Appendix E). The numbers of LOV domains found in each of the metagenomic samples are given in table 5.1.

Table 5.1 Metagenome samples and number of LOV domains

S.N.	Metagenome samples	No. of LOV domains	Total ORFs	Reference
1	Acid mine drainage metagenome	6	12820	[62]
2	Enhanced biological phosphorus removal	6	65328	[127]
3	Farm soil (Diversa silage)	4	185274	[124]
4	Hypersaline Mat	6	136770	[123]
5	Singapore indoor air sample	6	92560	[125]
6	ALOHA/Gyre	4	449086	[87]
7	Western Channel observatory	6	NA	CAMERA database
8	Marine pennate diatoms	13	NA	CAMERA database
9	Broad phage	3	NA	CAMERA database
10	Whale fall metagenome	5	84453	[124]
11	Garden soil, University of Hamburg (pWThLOV)	1	NA	[266]

Table 5.2 Samples from the Global Ocean Expedition

S.N.	Sample(s)	Metagenomic sample location(s)	Environment	No. of LOV domains	No. of seq. reads	Total ORFs from the sample	Total sequence size in bp
1	GS000ab	Sargasso Station 11	Ocean	27	961731	4969170	979782003
2	GS000c	Sargasso Stations 3	Ocean	6	368835	1326538	371688861
3	GS000d	Sargasso Station 13	Ocean	1	1293971	1214791	335939509
4	GS001a	Sargasso, Hydrostation S	Ocean	8	142352	902869	143316448
5	GS001b	Sargasso, Hydrostation S	Ocean	8	325608	510218	90955161
6	GS003	Browns Bank, Gulf of Maine	Coastal	1	61065	227142	66907344
7	GS014	South of Charleston, SC	Coastal	4	128885	483025	139914998
8	GS015	Off Key West, FL	Coastal	2	127362	462767	138034062
9	GS016	Gulf of Mexico	Coastal	2	127122	475239	137479949
10	GS017	Yucatan Channel	Ocean	3	257581	927755	281259325
11	GS021	Gulf of Panama	Ocean	2	131798	540820	143454700
12	GS022	250 miles from Panama City	Ocean	1	121662	427726	131079270
13	GS023	30 miles from Cocos Island	Ocean	2	133051	473053	143626589
14	GS025	Dirty Rock, Cocos Island	Fringing Reef	35	120671	666439	129781299
15	GS027	Devil's Crown, Floreana Island	Coastal	3	222080	829079	237326008
16	GS029	North James Bay, Santigo Island	Coastal	1	131529	465644	143822814
17	GS032	Mangrove on Isabella Island	Mangrove	2	148018	628007	153341974
18	GS033	Punta Cormorant, Hypersaline	Hypersaline	32	692255	5265901	729708089
19	GS034	Galapagos Islands	Coastal	1	134347	574300	142199308
20	GS038	Tropical South Pacific	Ocean	2	741	2980	787340
21	GS048a	Moorea, Cooks Bay	Coral Reef	3	90515	3265	92813604
22	GS048b	Moorea, Cooks Bay	Coral Reef	5	47691	NA	NA
23	GS049	Moorea, Outside Cooks Bay	Coastal	4	735	2581	94424378
24	GS108a	Coccos Keeling, Inside Lagoon	Lagoon Reef	2	56789	135241	50891020
25	GS108b	Coccos Keeling, Inside Lagoon	Lagoon Reef	1	52201	194050	53530124
26	GS109	Indian Ocean	Ocean	1	62973	167702	62752808
27	GS112b	Indian Ocean	Ocean	1	56008	214095	55638894
28	GS113	Indian Ocean	Ocean	1	118264	308947	118339154
29	GS114	500 Miles west of the Seychelles	Ocean	5	408529	903527	345285679
30	GS117a	St. Anne Island, Seychelles	Coastal	3	383329	876774	339868195
31	GS149	West coast Zanzibar (Tanzania)	Harbor	2	118837	326413	111179909

5.2 Habitat-based analysis of LOV domains from the metagenomes

Among the various habitats investigated, eight distinct habitats were found where LOV domain containing genes were present. They were found in samples derived from extreme acidic environments, waste water treatment plants, hypersaline environments, soil, indoor air, sea coast, offshore sea, coral and fringing reefs (Figure 5.2, Table 5.2). The majority of metagenomic data is derived from the Global Ocean Sampling (GOS) Expedition, which was the largest encounter in the history of metagenomics and contributed thousands of new species [93]. The Western Channel, Whale Fall, Pacific Gyre, Broad phage metagenome and marine pennate diatoms, all of which are from marine environments, yielded 37 LOV domains. Altogether, the marine derivatives including the Global Ocean metagenome contain 208 putative LOV domains.

An acid mine drainage metagenome sample was collected from a low-complexity microbial biofilm growing hundreds of feet underground within a pyrite (FeS_2) ore body from the Richmond mine at Iron Mountain, California, having a pH value of 0.83 [62] and temperature levels of up to 42 ^{o}C. The enhanced biological phosphorus removal (EBPR) sludge community DNA contains a metagenome sample from the Nine Springs Wastewater Treatment Plant in Madison (USEBPR), Wisconsin, USA and Thornside Sewage Treatment Plant in Brisbane, Queensland, Australia (AUEBPR) [127]. The hypersaline mat metagenome originates from the complex and stratified microbial mat of 49 mm thickness in Guerrero Negro hypersaline lakes in Mexico [123]. Singapore air metagenome contains environmental microbes harvested from the air of two densely populated urban buildings with roughly 80 million bases of DNA sequences from more than 140,000 sequence reads [125]. The whale fall metagenome originated from bones of two gray whale carcasses, one of these was artificially sunk at 1674 meters in the Santa Cruz Basin of the Pacific Ocean, while another metagenome sample derived from whale bones in the Southern Ocean at the depth of 560 meters off the West Antarctic Peninsula Shelf [124]. The diversa silage metagenome (also called farm soil) was originated from surface soil (0-10 cm) of a farm in Waseca County Minnesota [124].

The Western Channel metagenome is derived from the coastal waters of the Western English Channel (http://www.pml.ac.uk/). The ALOHA/Pacific Gyre metagenome was originated from vertically distributed microbial communities at different depths (from 10 to 4000 m) in the North Pacific Subtropical Gyre (NPSG) at the open-ocean time-series station Hawaii Ocean Time-series (HOT) station ALOHA [87].

The "Broad Phage Metagenome" was originated from diverse marine phage/virus genomes (http://www.broadinstitute.org/). Marine Pennate Diatoms metagenomes were originated from Pseudo-nitzschia (a genus of ubiquitous marine diatoms) derived from the subarctic North Pacific Ocean (http://camera.calit2.net/).

These findings already indicate the ubiquitous presence of LOV domains in most diverse habitats and microbial species. However, the sequence data from different sample sources are not uniform, as some of them contain hundreds of thousands of ORFs and some only very few. For that reason the data were normalized before further calculation of the distribution, such that the absolute number of LOV domains in each environment was divided by the total number of predicted ORFs of that particular sample as deposited in the database and converted into percentage. Marine environments were grouped into coastal, free marine and reef according to the information provided in the CAMERA database. The normalized data indicate (Figure 5.2a) that LOV domain-containing proteins are most abundant in the acid mine drainage (about 0.047% abundance) which is an extreme environment in terms of pH [62]. The total number of ORFs in that biofilm metagenome is 12,820 and six putative LOV domain sequences were identified in that metageome. In total numbers, however, most LOV-domain sequences are present in off-shore sea water metagenomes (Figure 5.2b), clearly due to the overwhelming total amount of ORFs identified by this survey. The metagenome samples from marine surface and planktonic community which contains the highest number of ORF deposits in the sequence database showed a fairly 0.0098 % abundance.

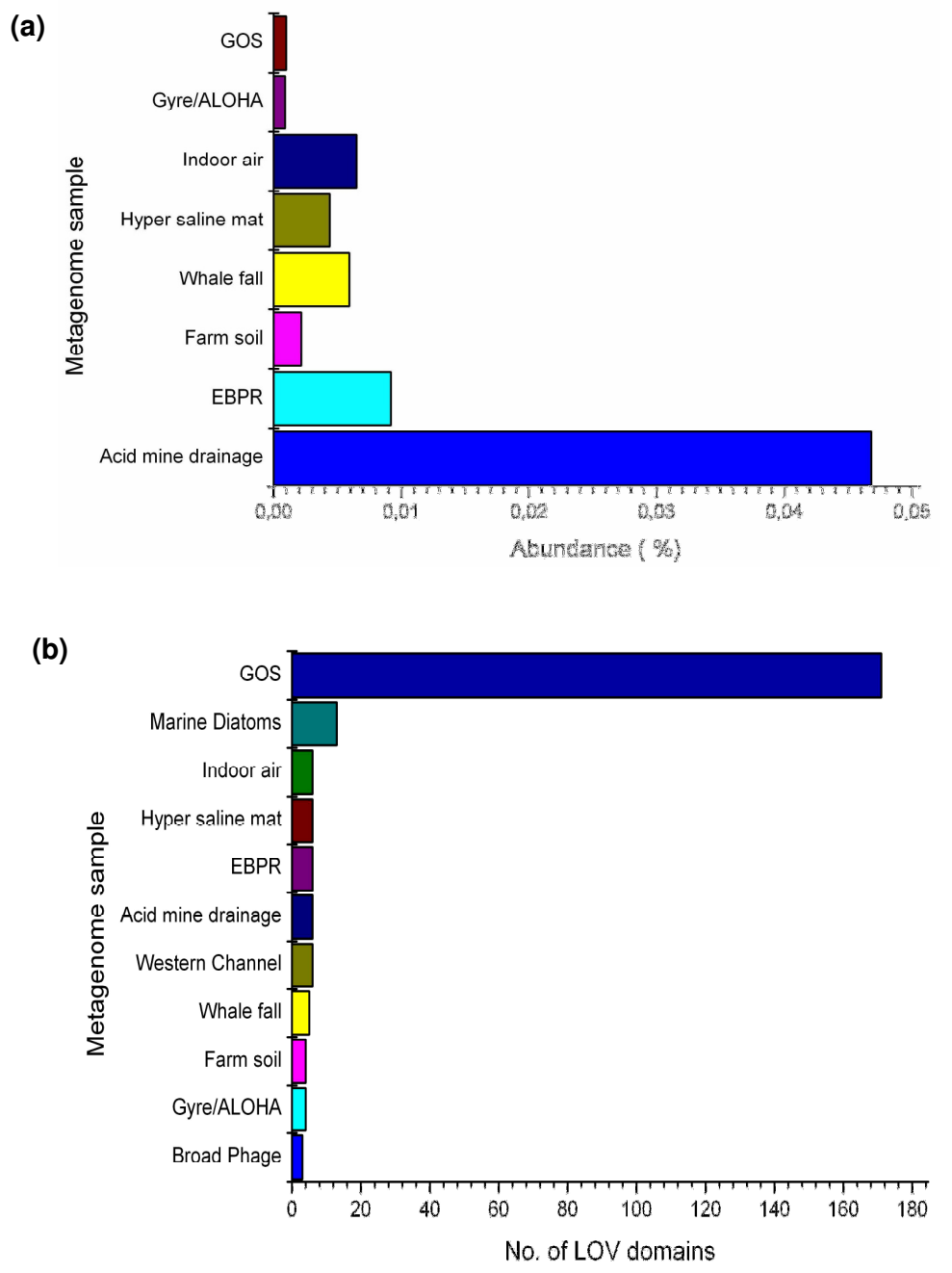


Figure 5.2 LOV domains in sequenced metagenomes. (a) Percentage abundance. The numbers of LOV domains present in each station were divided by the number of total ORFs in that particular station and the abundance is presented as percentage. (b) Absolute numbers of LOV sequences found in the various metagenome samples.

As the GOS metagenome contains a high number of sequence deposits and has the broadest coverage of worlds' oceanic environment, samples from this database were taken for an analysis, to obtain an insight on the distribution of LOV domains in global oceans. Here, the distribution and relative abundance was calculated based on each station which were located at different positions on the ocean. The abundance value from each station (a_i, a_{ii}, a_{iii}….) was added together resulting in "A". The relative abundance was calculated in terms of ratio (in percentage) of abundance of each station to the sum of abundances of LOV domain at all stations (Relative abundance of station-n = a_n/A x 100%, where n is a particular station i, ii, iii..). The relative abundance of putative LOV domain genes is high in sample stations GS049, GS048a and GS038 (Figure 5.3a; for an assignment to locations see Appendix F).

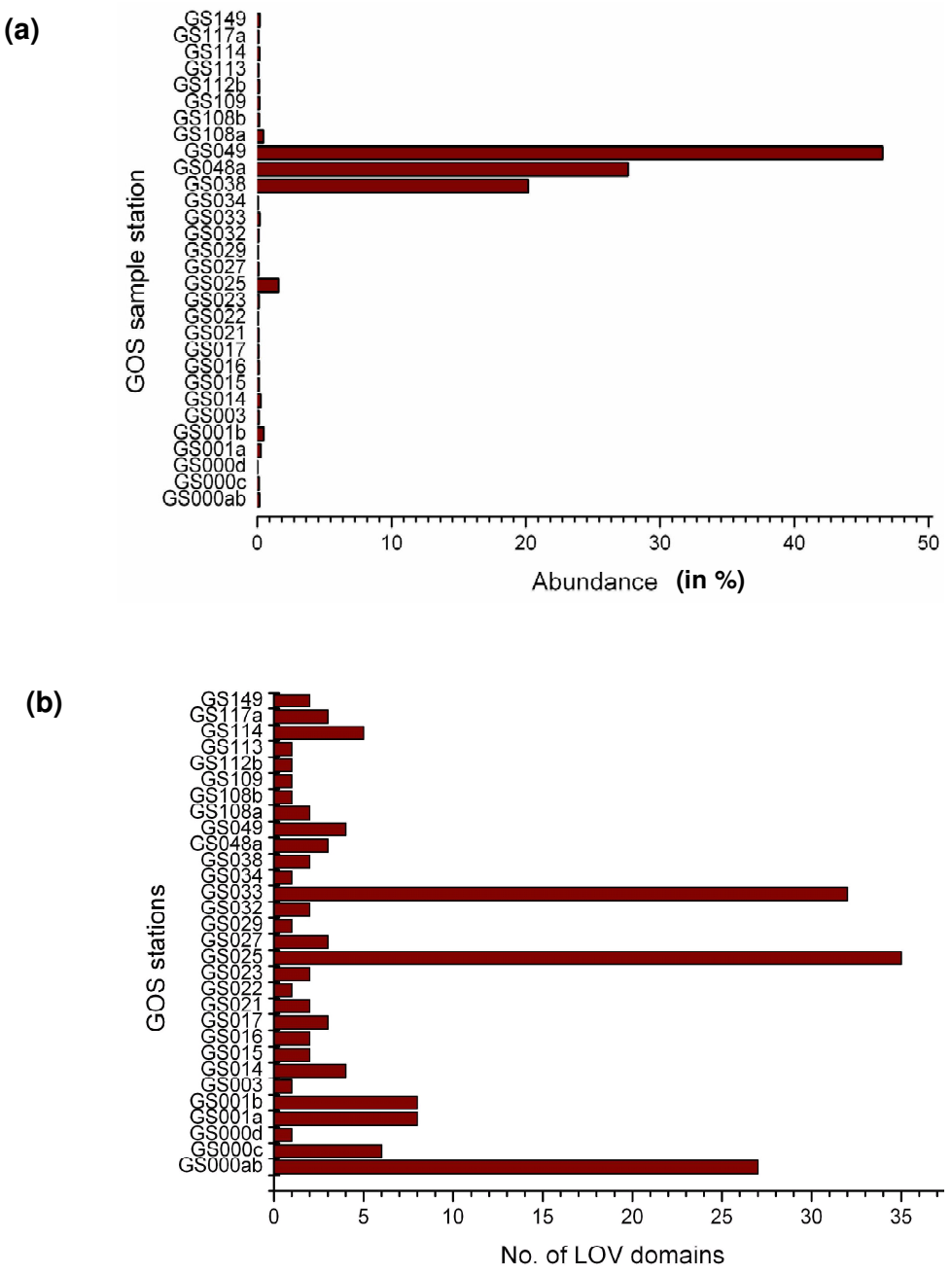


Figure 5.3 LOV domains in the Global Ocean Sampling (GOS) stations. (a) Relative abundance in GOS (in percentage). (b) Absolute numbers of LOV sequences in found in the different GOS samples.

In total numbers GS033, GS025 and GS000ab represent high numbers of LOV domain containing genes (Figure 5.3b). GS000ab represents Sargasso station 11 and 13 where samples were collected at the depth of 5 m from sea surface. GS025 represents the samples from fringing reef from 1.1 m below the surface water at Cocos Island at Eastern Tropical Pacific. GS033 was based on Punta Cormorant hypersaline lagoon at Floreana Island and the metagenome sample was collected from 0.2 m depth. GS039 was situated in international water at Tropical South Pacific where the sample was collected at the depth of 1.8 m. GS048 and GS049 were located in Moorea Cooks Bay, French Polynesia where samples were collected at the depth of 1.4 m. Surely, one has to take ino account that water movements by wind and waves causes an intermixing in the upper water layers.

A clear distribution of LOV domains in metagenome samples could be detected with respect to salinity: the majority of samples originated from sea water salinity of ca. 35 ppt (ca. 73 %), another significant amount of samples (ca. 13.8%) was obtained from salinity of 63.4 ppt, and a smaller community (ca. 2.6 %) was collected from an environment with an even three-fold higher salt content (ca. 90 ppt) than that of sea water. Altogether there were 38 LOV domain genes in the hypersaline environment (salinity >63.4 ppt). Among the ten layers in depth of the hypersaline mat that were individually investigated [123] the LOV domain containing proteins were found only in the upper first, second, fourth and the sixth layer. The upper two layers of the hypersaline mat were reported to contain a low oxygen concentration during day time whereas the rest was found anoxic. During the night, all layers became anoxic [123]. The layers of the hypersaline mats where the LOV domain sequences were found showed high numbers of proteins orthologus to *C. crescentus* (alphaproteobacteria), *Nostoc sp.* (cyanobacteria), *Porphyromonas gingivalis* (Bacteroides), *Bacteroides thetaiotaomicron* (Bacteroides), and *Gemmata obscuriglobus* (Plancomycetes) (Figure 5.4). In BLAST analysis the nearest matches for the hypersaline LOV domains were obtained from alphaproteobacteria, deltaproteobacteria, cyanobacteria and algae.

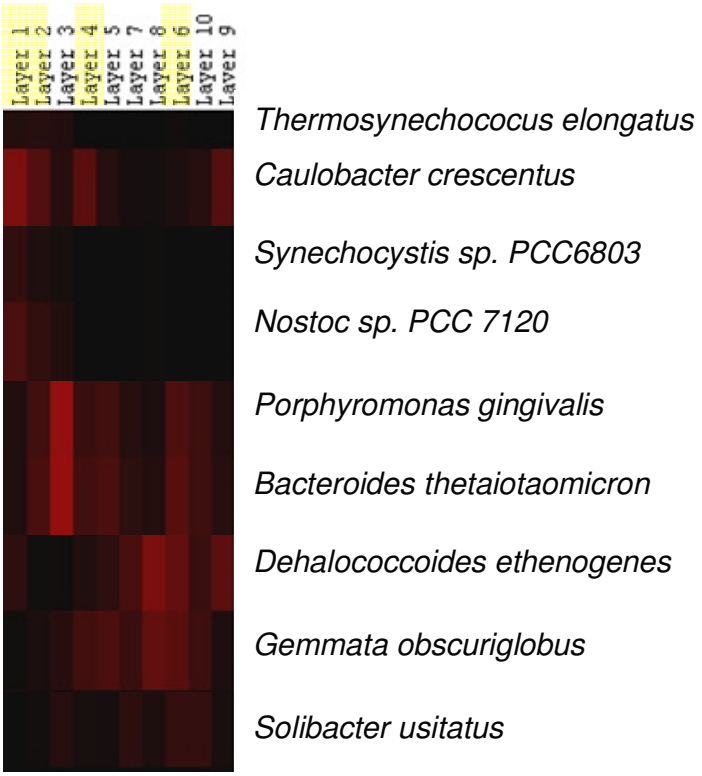


Figure 5.4 Intensity mapping of the orthologus proteins from different layers of hypersaline mat in the increasing order from dark to bright red. Dark indicates low and the bright red indicates highest number of layer-specific gene families assigned to individual species. Only dominant groups are shown. The layers of the hypersaline mats that contain LOV domain genes are highlighted in yellow. The clustering of the data was carried out using Cluster 3 and visualized by Tree View. The normalized data was obtained from Hypersaline mat metagenome project [123].

5.3 Temperature of the environment and LOV domains

A distribution with respect to different temperatures (Figure 5.5) yielded the majority in a mesophile environment, yet, positive hits were obtained for temperatures as high as 42 ^{o}C and as low as 1.46 ^{o}C. Information about the temperature of the source environment was available for 216 LOV domain sequences. For this reason only these out of the 231 putative LOV domain sequences obtained in the database search were plotted against the sample source environment. The graph demonstrates that LOV domains are present in wide range of sample environment, but most of them are clustered between 20 to 30 ^{o}C. One of the metagenome derived LOV domains detected in microarray based approach was derived from soil sample enriched at 65 ^{o}C.

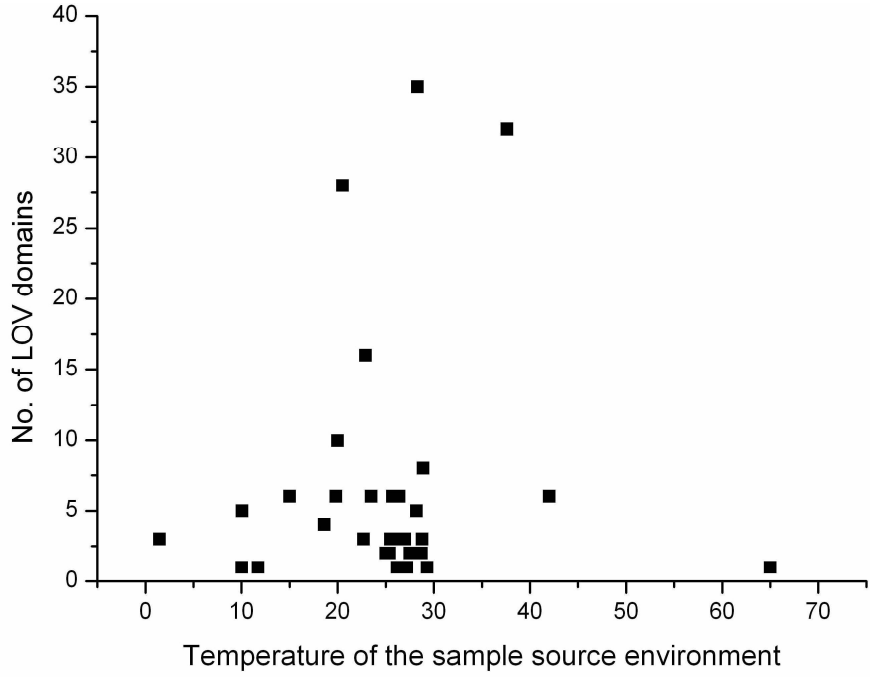


Figure 5.5 Temperature dependent clustering of LOV domain. Though most of the BL photoreceptor proteins are found between 20 and 30 ^{o}C, their distribution ranges from about 1.46 ^{o}C to 65 ^{o}C.

5.4 LOV domains in viral metagenome

Also, metagenome samples generated from viral nucleic acids were investigated for the presence of LOV domains. Among several viral metagenomes from various habitats available in the databases only the Broad Phage sequence database and mosquito viral metagenome gave positive results in a BLAST search.

Three partial LOV domain sequences were obtained in the BLAST search of the mosquito viral metagenome in NCBI using LOV domain from *B. subtilis* as a query sequence. The mosquito viral metagenome consists of DNA sequences derived from the viral community isolated from mixed species of mosquitoes collected in California, USA [89]. The partial sequences contained the inner conserved core of a typical LOV domain including the conserved cysteine and essential residues near to it. However, since the sequences were generated using 454 pyrosequencing techniques, creating only short read lengths, each of about 120 bp in length, it was not possible to identify a complete LOV domain sequence.

The Broad Phage sequencing project has initiated sequencing, assembling, and annotation of 200 diverse marine phage/virus genomes and led to sequences from approximately 50 viral metagenomes from an array of marine environments, (http://www.broadinstitute.org/). As a result of this approach, about 726,683,442 bp of sequence data from phage DNA have been already submitted to the CAMERA database. About 18 sequence reads containing putative LOV domains were found in the Broad Phage 454 sequence database in CAMERA platform in the tBLASTn analysis (with an $>e^{-10}$) against 2,696,129 sequence reads from marine cyanophages.

The sequence reads that contained LOV domains were further analyzed at the nucleotide level to control the identity among each other. Only one read was selected from the group of reads in cases when more than one fully identical sequence was present. The reads were translated to open reading frames using the ORF finder of NCBI, and the ORFs that contained putative LOV domains were selected. As in all other cases, sequences shorter than 90 amino acids were discarded, yielding here six unique reads (Table 5.3). Among them, only three sequences

(BroadPhage_READ_01585282, BroadPhage_READ_01587598 and BroadPhage_READ_01816240) contain the conserved cysteine residue responsible for binding FMN in the correct position.

Table 5.3 Sequence reads from the broad phage metagenome database in which LOV domain sequences were found

S.N.	Read	Sample source
1	BroadPhage_READ_01585282	Cyanophage_M4-247
2	BroadPhage_READ_01587598	Cyanophage_M4-247
3	BroadPhage_READ_01816240	Cyanophage_9303-10a
4	BroadPhage_READ_01950047	Cyanophage_9303-10a
5	BroadPhage_READ_01128448	Cyanophage_M4-259
6	BroadPhage_READ_00814651	Cyanophage_NATL2A-133

The ORFs from each of the six reads were then used as a query sequence and were further tBLASTxed against all Global Ocean Sampling (GOS) expedition [93] peptides data set. An e-value >-20 was set to find the nearest matches. In addition, the ORFs from the reads were pBLASTed against cyanobacterial genomes in the NCBI database and the nearest matches were selected. All nearest matches and the putative LOV domains from the phage metagenome were aligned and a consensus maximum likelihood tree was constructed from 100 bootstrap values using a JTT matrix. In the phylogenetic tree constructed using the nearest GOS matches, few nearest cyanobacterial genomes and the phage LOV domains, the five LOV domains derived from phage metagenome (READ_01950047, READ_00814651, READ_01128448, READ_01585282, READ_01587598) clearly clustered together with their GOS counterpart while the sixth one (READ_01816240) is grouped with *Cyanotheca* and *Crocosphaera* (Figure 5.6).

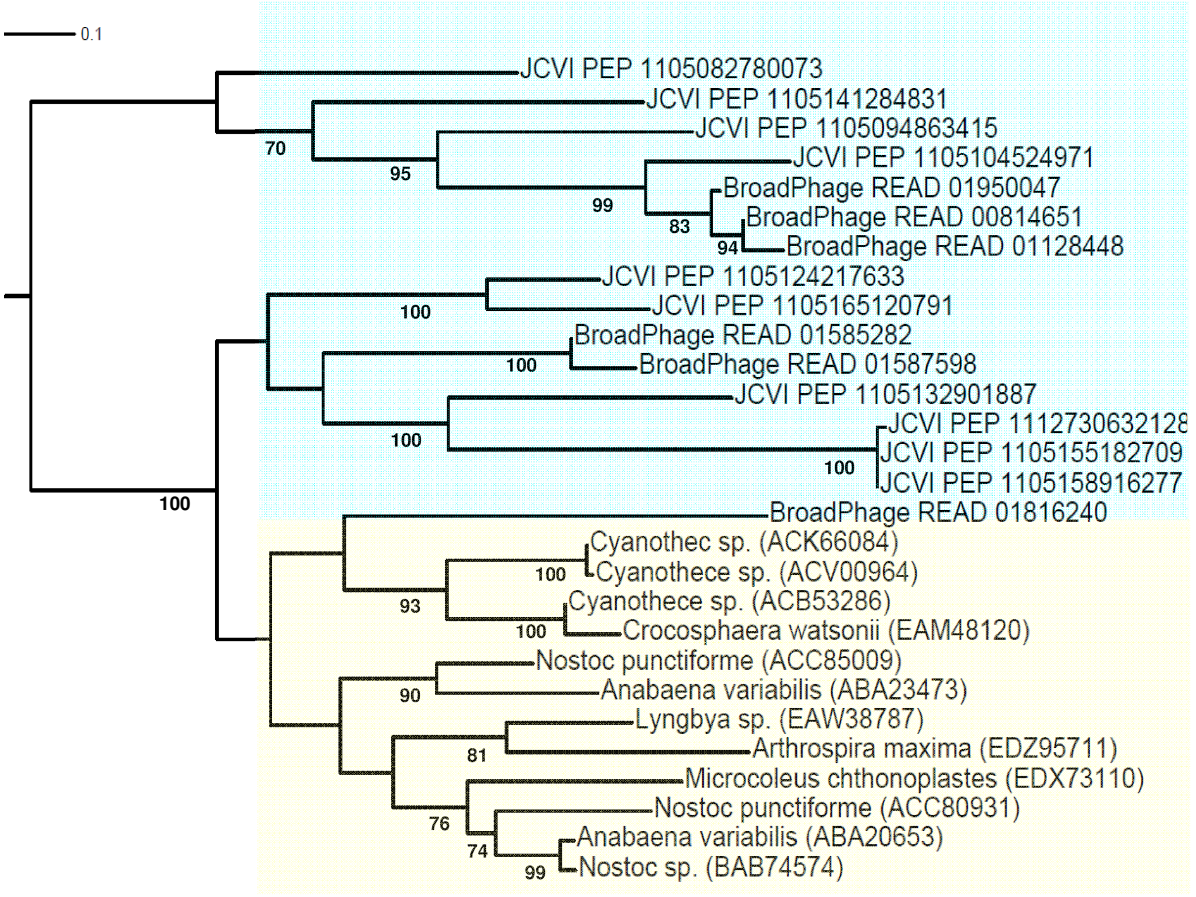


Figure 5.6 A maximum likelihood tree generated from cyanophage metagenome LOV domains (Broad Phage sequencing project) and the closely related LOV domains from sequenced genomes and the GOS metagenome. LOV domain sequences from sequenced genomes are shown against a light yellow background. The LOV domains from phage metagenome and GOS metagenomes are highlighted in light blue. The tree is based on the (90 to 100) shared amino acids. Sequences supported by less than 70% of bootstraps values have been removed from the tree.

5.5 Diversity and novelty of LOV domains in the metagenomes

In this encounter, the determination of novel LOV domain proteins was carried out by sequence homology check. The metagenomic LOV domains were BLASTed against protein sequences derived from individual genomes deposited in NCBI database using pBLAST. The percentage of similarity to the nearest match was identified, determining a sequence with a similarity value of less than 79% to be

assigned as a novel LOV domain. About 73% (169 out of 231) of the metagenome-derived putative LOV domains were classified as novel in that way (Figure 5.7). In addition, the affiliation of an individual LOV domain from the metagenome to different bacterial classes was determined according to the taxonomic position of its nearest match based on sequence homology (>40%). It can be seen that LOV domains from the metagenome are widely dispersed near the proteobacterial line, and most are close to alphaproteobacteria.

5.6 Phylogenetic analysis of LOV domains from the metagenomes

The complete LOV domain sequences from metagenome and sequenced genomes were aligned using the ClustalW web-based server. The alignment file was exported to the BioEdit sequence editor program to edit the sequences. Phylogenetic analysis was then carried out on the conserved LOV domains of already identified and putative LOV domain sequences obtained from all sequenced genomes and metagenomes. Only the conserved LOV domain region of the putative and characterized LOV domain proteins was used for the phylogenetic analysis, as the fused partner domains or the extensions of the LOV domains are of highly diverse nature. Sequences that did not contain the conserved photadduct forming cysteine in the conserved core region were manually removed. Similarly, sequences shorter than 85 amino acids were also withdrawn from the analysis. Altogether 461 LOV domain sequences, 231 from metagenome and 230 from the sequenced genomes were included in computing the phylogenetic tree. The computation was carried out using PhyML [185]. For reasons of available computer power, only several runs with 100 bootstraps value using the JTT matrix were performed; in each case, a highly similar tree topology was obtained. The consensus tree supported by maximum bootstrap values (created by PhyML) was used for the analysis of phylogenetic relation. The global phylogenetic tree of LOV domains is shown in figure 5.7. Few novel families of LOV domains have been identified on the basis of this phylogenetic analysis and the sequence homology based method. The relation between different subfamilies is poorly resolved but that within the subfamilies is strong, which is evident by the lower bootstrap at deeper branches and higher bootstrap values at the leaf labels.

Interestingly, the LOV domains from a similar sample environment are grouped together. The sample for the EBPR sludge metagenome was collected from waste water plants from two different geographical locations (Australia and the USA), but irrespective of the different geographic origin, the LOV domains from both metagenomes are grouped together. The same observation was found for oceanic environments, where most of the LOV domains are also grouped together in subfamilies.

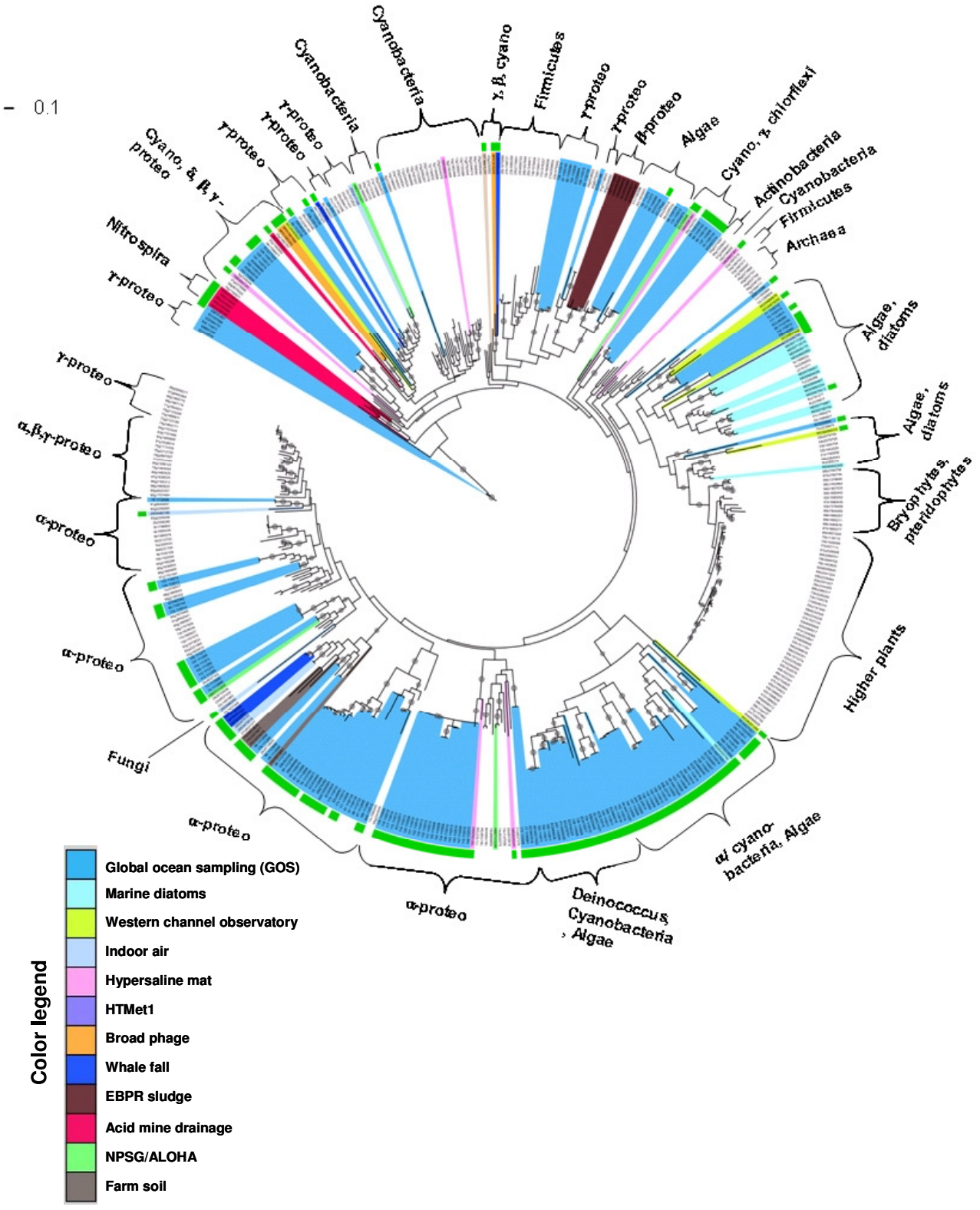
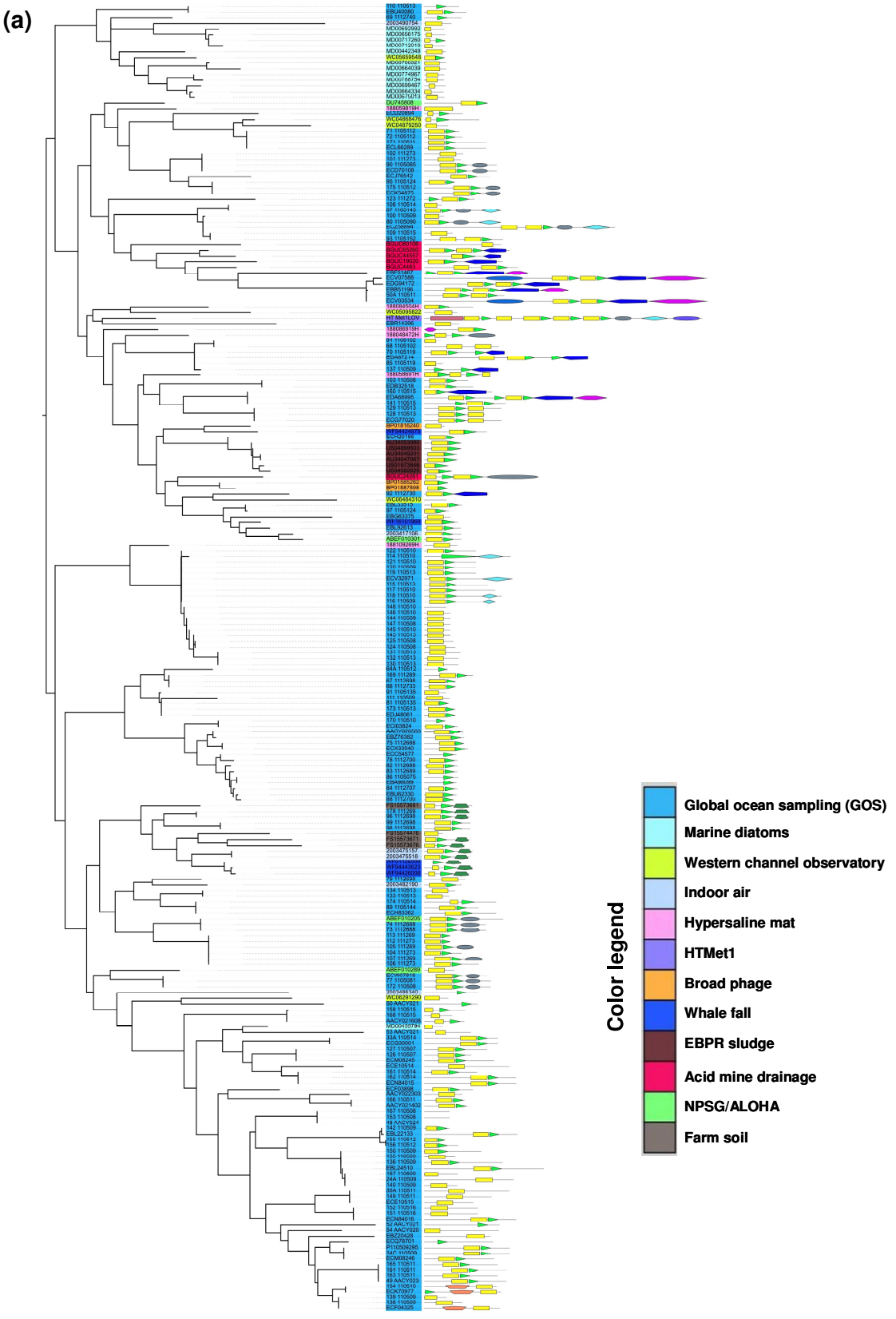

0.1
Cyanobacteria
γ, β, cyano
Firmicutes
γ-proteo
γ-proteo
β-proteo
Algae
Cyano, & chloroflexi
Actinobacteria
Cyanobacteria
Firmicutes
Archaea
Algae, diatoms
Algae, diatoms
Bryophytes, pteridophytes
Higher plants
α, cyano-bacteria, Algae
Deinococcus, Cyanobacteria, Algae
α-proteo
α-proteo
Fungi
α-proteo
α-proteo
α,β,γ-proteo
γ-proteo
Nitrospira
γ-proteo
Cyano, & β, γ-proteo
γ-proteo
γ-proteo
Cyanobacteria
Color legend
Global ocean sampling (GOS)
Marine diatoms
Western channel observatory
Indoor air
Hypersaline mat
HTMet1
Broad phage
Whale fall
EBPR sludge
Acid mine drainage
NPSG/ALOHA
Farm soil

Figure 5.7 Global phylogenetic tree of LOV domain proteins. The tree was constructed from 461 *BL* receptor LOV domains from the different metagenomes (231 sequences) and data base-deposited (230 sequences) genome sequences. The colored branches indicate the metagenomes (each color defines a particular metagenome, see illustration on the left). The uncolored branches refer to individual sequences from genome database. The green outer rim denotes those novel LOV domains that share less than 79% of sequence identity to the LOV domain protein from the known species deposited in genome database. Bootstrap values >70 are displayed with small circle on the node.

5.7 Domain architecture analysis

Most of the protein sequences from the metagenomes that contain putative LOV domains were found in partial form, which probably could not cover all the partner domains fused to the LOV domains. Only 56 protein sequences from Global Ocean Sampling metagenome and 25 protein sequences from the rest of the metagenomes were obtained in the complete form. All of the partial and complete ORFs were subjected to screening for the domain composition to analyze the domain architecture in putative LOV domain proteins. For this purpose, the protein sequences were scanned using the ScanPro [183] and SMART v5 [184] web-based domain detection softwares. As observed from the analysis, short LOV (PAS/PAC) domain modules are quite frequent in metagenomic putative BL receptor genes. A wide range of domain composition, from a single to multiple associated domains, was observed in the rest of the sequences (Figure 5.8). Interestingly, none of the LOV domain-containing metagenome sequences obtained from the databases was associated to a STAS domain, which is involved in stress regulation in some spore-forming firmicutes. The enhanced biological phosphate removal (EBPR) sludge metagenome has been found to contain only short LOV domains, whereas the soil metagenome shows a high degree of variation with several domain modules ranging from short single LOV (PAS/PAC) to long multi domain structures. LOV domains associated with a histidine kinase domain were seen mostly in aquatic environments (Whale fall, NPSG/ALOHA, Global Ocean).

Among the six AMD metagenome-derived LOV domains, three contain a fused GGDEF-output domain; one carries a histidine kinase output domain and the remaining two did not have any output domains. The AMD metagenome was dominated by a small number of species. Using random shotgun sequencing of DNA, a reconstruction of near-complete genomes of *Leptospirillum* group II and *Ferroplasma* type II could be carried out along with the partial recovery of three other genomes [62]. This approach was based on the raw metagenomic data. The LOV domain-containing genes identified in our sequence-based investigation of the AMD metagenomes were present in incomplete form. Because of the low species diversity and hence high possibility of the relatedness of the metagenome sequences to the already reconstructed genomes, reconstructed species can be used as a model to understand the genetic arrangements of the related metagenomic genes from AMD. To obtain deeper insight of the domain architecture of the LOV domain containing genes from the AMD metagenome, the homologous genes present in the reconstructed genomes from the acid mine drainage metagenome were BLAST-searched in the NCBI database. Using the LOV domains from the AMD metagenome as a query sequence, five different LOV domain-containing genes from reconstructed *Leptospirillum* sp. were obtained as BLAST hits. Only two of the LOV domain sequences from AMD metagenome were completely identical to the LOV domain-containing genes of reconstructed genomes at the amino acid level, however, they showed quite a significant difference at the nucleotide level (Figure 5.8b). The domain modules of the LOV domain containing genes of reconstructed genomes were further investigated. Two of the LOV domain genes from reconstructed genomes showed GGDEF+EAL, one showed histidine kinase+HATPase and the rest two contained GGDEF only as output domains (Figure 5.8c). In fact, the reconstructed genome provided information about the domain module arrangement and the phylogenetic relation of LOV domain-containing homologous genes from the AMD metagenome.


(a)
Color legend
Global ocean sampling (GOS)
Marine diatoms
Western channel observatory
Indoor air
Hypersaline mat
HTMet1
Broad phage
Whale fall
EBPR sludge
Acid mine drainage
NPSG/ALOHA
Farm soil

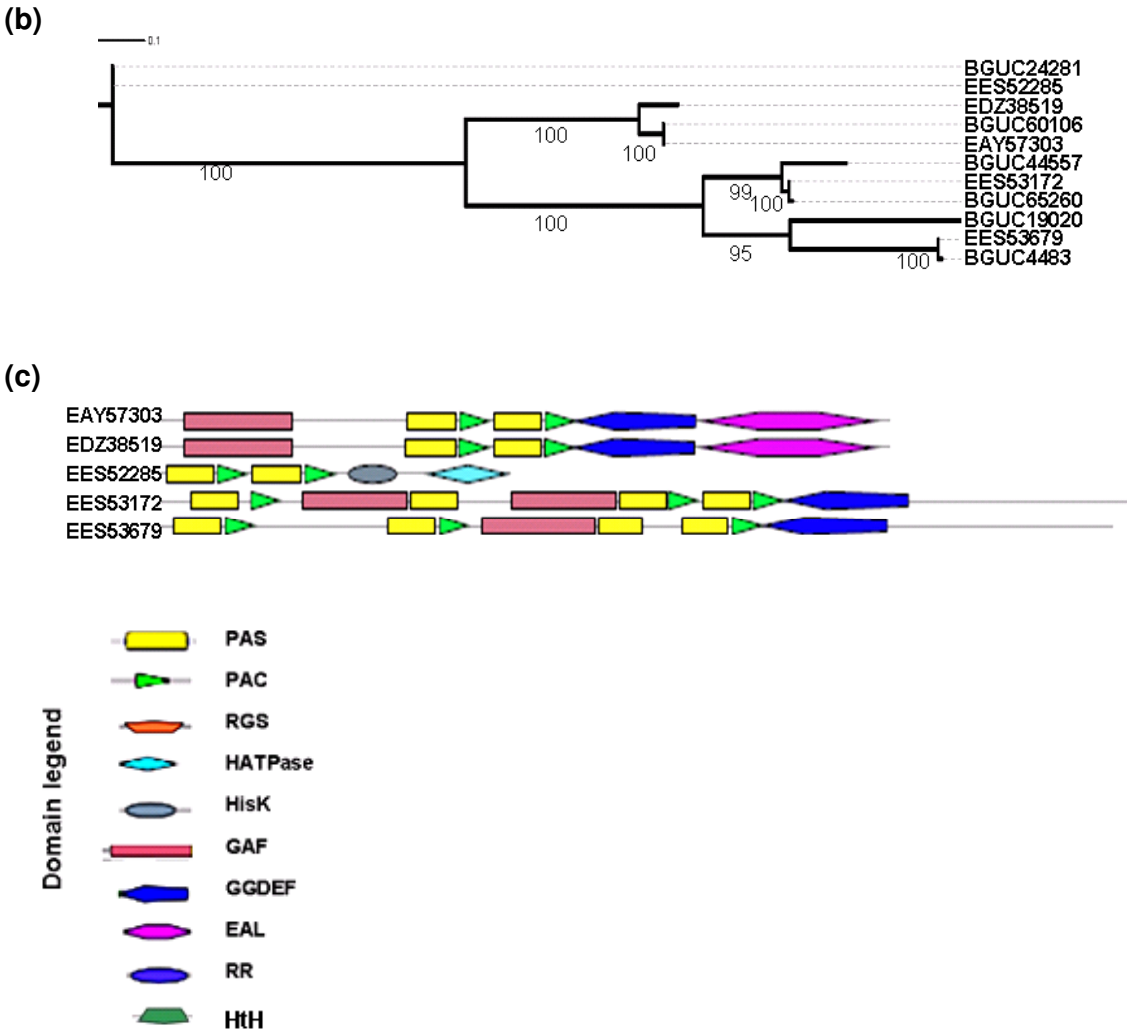


Figure 5.8 (a) Domain structures of putative BL receptor proteins from metagenomes. A variety of domain combinations are present, ranging from multi domain proteins to only PAS/PAC domain modules. The background color-coding indicates the grouping of the sequences according to the metagenome origin. The length of the protein is given as a slim line, and the various domains are depicted in different geometrical shapes (rectangles, pentagons, ellipses etc.), as clarified in the domain legend in the figure. (b) Phylogenetic relation of LOV domain-containing genes from the acid mine drainage metagenome and the reconstructed genome of *Leptosprillum* sps. (c) Domain structures of putative BL receptor proteins from reconstructed *Leptosprillum* genera: EES52285, EES53172 and EES53679 are from *Leptospirillum ferrodiazotrophum*,

EAY57303 is from *Leptospirillum rubarum* and EDZ38519 is from *Leptospirillum* sp. Group II. BGUC24281, BGUC60105, BGUC44557, BGUC65260, BGUC19020 and BGUC4483 are the sequences from acid mine drainage metagenome. PAS: Per-Arnt-Sim; PAC: PAS associated C-terminal motif; RGS: Regulator of G-protein signaling; HATPase: Histidine kinase-like ATPases; HisK: Histidine kinase; GAF: cGMP-specific and -regulated cyclic nucleotide phosphodiesterase, Adenylyl cyclase, and *E. coli* transcription factor FhlA; GGDEF: named after a conserved motif, this domain is likely to catalyze the synthesis of cyclic diguanylate; EAL: named after the conserved sequence EAL that is found to hydrolyze cyclic diguanylate; RR: response regulator; HtH: Helix-turn-Helix.

In most cases, GGDEF were the important partner domains for the LOV domain-containing genes in AMD. All of these proteins were long multi domain proteins and contained either two or more PAS sensor domains, one of the PAS being the LOV domain in each case.

5.8 Discussion

5.8.1 Phylogenetic and habitat-based analysis of putative LOV domain-containing genes

As a consequence from the reflection that most of the bacteria in the environment can not be cultivated by the conventional cultivation techniques, a large number of metagenomic approaches has been employed to study the genetic dynamics of various environments. As discussed in chapter 1 (1.2.5), several unique habitats have been investigated and historical expeditions have been made to investigate the metagenome from diverse locations. This resulted in the deposition of global metagenomic sequence data that have been expedited. Already nowadays, this number extends by more than an order of magnitude the data from sequenced genomes. The overwhelming sequence data in public databases from the different environments now have been analyzed from different perspectives to investigate the biological processes happening within the microbial communities present in such environments [119]. Referring to LOV domains, 13.5 % of the already sequenced genomes have been reported to contain proteins showing the signature of a LOV

domain [15]. Thus, the present study was dedicated to investigate the LOV domains in different environments and perform their habitat based and phylogenetic analysis.

An explanation of the presence of LOV domains on the basis of taxonomy seems less convincing because it is found that some but not all the species of any certain genus harbor LOV domain-containing genes [267]. In the analysis of different metagenomic databases, we found that the abundance of the genes that contain this sequence motif was more in extreme environments. LOV domains can be traced in every kingdom of life [267] and in almost all types of habitat. The present study also confirms the almost ubiquitous presence of LOV domain in water, air, soil or in extreme habitats. However, they have not been reported from some bacterial phyla. Among the investigated ones, metagenomes from human gut, lungs or termite guts did not contain any LOV domain. Human or termite gut flora are not exposed to light and most of the species in the human gut are from Flavobacteria and Bacteroides [109] which along with enterobacteria have not been reported to contain LOV domains [267]. However, *Brucella abortus*, a mammalian facultative intracellular pathogen, contains a LOV domain associated with histidine kinase which was shown to be important for host infection [166]. The interdisciplinary Human Microbiome Project [129] probably will provide more in-depth information about the light receptor genes from microbes inhabiting the human body.

The large number of LOV domain sequences from the metagenome was found to cluster in novel branches of the phylogenetic tree, and many of them were assigned to the alphaproteobacterial group. In accordance to former work [267], the phylogenetic tree derived from metagenomes, bacterial, fungal and plant LOV domains in this work places the LOV domain from higher plants on the higher level of the tree. One of the interesting features of several of the LOV domains from the metagenome is that many of these domains deriving from a similar habitat are grouped together. Possibly, the reason for that is the significant amount of horizontal gene transfer in these environments. Lateral gene transfers are quite common phenomena in marine environment. It was suggested that lateral gene transfer process is responsible for the presence of PYP in a halophilic Bacteroides,

Salinibacter ruber, which probably acquired the genes from halophilic proteobacteria that share the same ecological niche though being taxonomically diverse [268]. Similarly, the presence of proteorhodopsins in archaea has been reported which is believed to be due to the lateral gene transfer from proteobacteria to the archaea [141]. In this view, it is plausible that microbes sharing the same ecological niche can acquire genes via lateral gene transfers, which is also suggested by the clustering of the LOV domains from similar habitats together in the global phylogenetic tree of the LOV domain (Figure 5.7). It is obvious that the environment plays an important role for determining the acquirement of certain functional genes.

In comparison to proteorhodopsins, which were found to be more than 3000 in the Global Ocean Sampling database [93;135], LOV domains were found to be about 171 in that database. Although the short lengths of the reads are the greatest problem while analyzing the metagenome data, the low number of LOV domains was not only due to the short read lengths or incomplete sequence data. It suggests that the number of genes involved in energy harvesting is greater than that involved in sensing. It should be noted that proteorhodopsins are highly abundant in marine surface water and most probably they are involved in energy harvesting rather than sensory mechanisms. Probably those bacteria that do not carry LOV domain-encoding genes can use other blue light sensing mechanisms, like BLUF, PYP or even rhodopsins as suggested by several reports [135].

The analysis of the whale fall metagenome database revealed five putative LOV domain-containing genes (5.1). The age of the whale carcass from where the metagenomic nucleic acid was extracted was already more than 3 years [124]. Interestingly, proteorhodopsin genes that are quite frequent in marine environment and are supposed to drive the marine energy system were not reported in any of the whale fall samples that contained nearly 75 mbp of the sequences [124]. Whale fall communities are composed of diverse colorless sulfur bacteria including *Beggiatoa* mats and the whale bone provides habitats analogous to seeps and vents [269]. Sequence analysis revealed that *Beggiatoa* sps. do not contain a LOV domain on their genomes. A whale carcass is an important ecological niche spread in the oceanic

floor, as it is a continuing process where whales die, sink and remain on the sea floor, creating a nutrient-rich environment in the deep floor of the ocean. One hypothesis could be that the cells might have been attached to the sinking carcasses in the blue light receiving region and thereby reached the bottom of the sea.

Six genes from the sequence database of indoor air of two supermarkets in Singapore were also identified to contain LOV domains. Indoor air was shown to be a distinct habitat, the dominant organisms present being different from those from the nearby soil or the river, ruling out the possibility of direct transfer of microbes from the outside environment to the indoor environment [125]. Most abundant microbes from that environment were from Caulobacteriales and Xanthomonadales. Nearly 30% BLASTX hits from the generated sequence were asigned to *Caulobacter crescentus* alone [125]. But surprisingly, none of the LOV domain sequence obtained from that metagenome showed a significant similarity to the LOV domain from the database-deposited genome of *C. crescentus*. Airborne microbes are often attached to dust particles or water droplets forming aerosols, and upon evaporation of the water in the aerosol the microbes become droplet nuclei and clumps that can stay airborne indefinitely [270]. One of the functions of the LOV domain in the airborne metagenome could be to facilitate the attachment of cells to the dust particles as reported in the case of *C. crescentus*. Moreover, the air microbiota may encounter stresses like iron limitation, oxidative damage, and dessication [125], and LOV domains have already been demonstrated to be responsible in stress adaptation [57]. The role of the LOV domain in this indoor air environment can be attributed to act as stimulus for stress adaptation as well as cell attachment.

The investigation of the acid mine drainage (AMD) metagenome database, which was derived from a microbial biofilm growing in an abandoned mining site in Richmond, California (USA), revealed six genes that contain LOV domains (5.1). Five of them showed the highest level of homology to the signal transduction genes from the reconstructed *Leptospirillium* genome from AMD metagenome [62]. The source environment where the biofilm had been growing showed a pH of 0.83, a temperature of 42 ^{o}C, and high concentrations of Fe, Zn, Cu, and As [62]. The acidic mine drainage

results from bacterial iron oxidation, which leads to acidification of the medium due to the dissolution of pyrite in abandoned mines [271]. Genes involved in heterotrophic lifestyle and some other genes potentially responsible for microaerophilic survival, biofilm formation, acid tolerance, and metal resistance were observed in this AMD metagenome [62]. An annotation of the genes and the analysis of the gene neighborhood revealed the presence of heavy metal efflux pumps near to the putative BL receptor gene in *L. ferrodiazotrophum* (Figure 5.9). Possibly, the LOV domain can be involved in regulation of metal efflux and survival in the presence of heavy metal stress.

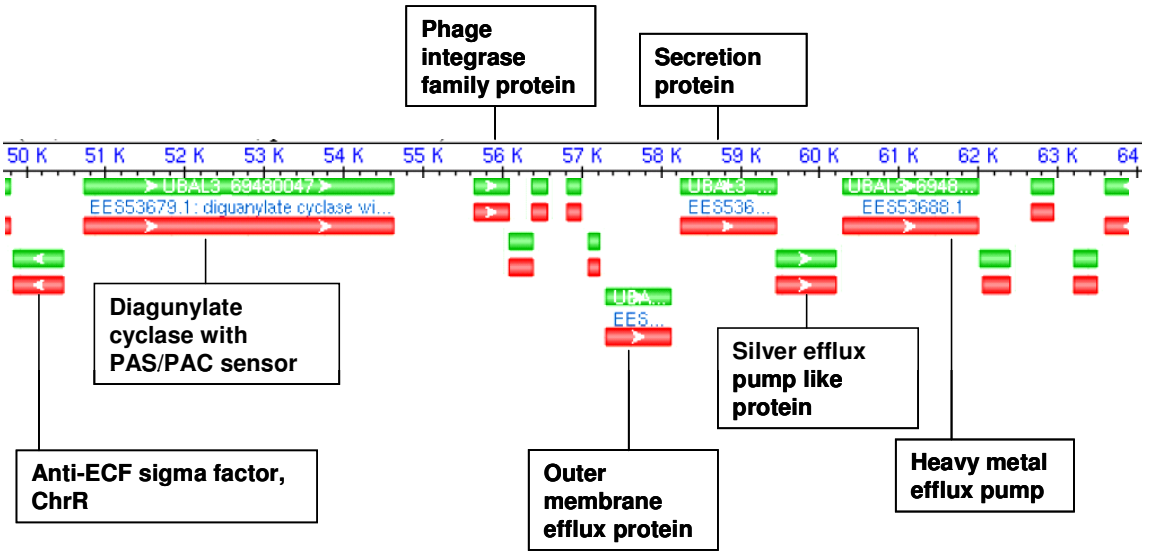


Figure 5.9 Graphical view of a part of the reconstructed genome of *Leptosprillum ferrodiazotrophum* showing the LOV domain-containing region with gene models (NCBI nucleotide Graphics). The LOV domain is found in the ORF annotated as diaguanylate cyclase (PAS/PAC sensor). Other annotated genes are shown in boxes.

The signal transduction genes from *Leptosprillum* sp. that contain the LOV domain exhibit a multidomain structure, containing a GGDEF motif in four cases and an additional EAL domain followed by GGDEF in two of them (Figure 5.8). GGDEF and EAL domains are involved in the synthesis and hydrolysis of c-di-GMP (cyclic dimeric guanosine monophosphate) [272], which in turn regulates the EPS

(extracellular polymeric substances) formation and hence participates in biofilm formation [273;274]. The GGDEF and EAL domains are often found to be amalgamated with N-terminal receiver domains [275].

Another metagenome from an extreme habitat that contained a significant abundance of LOV domains was the Guirero Negro Hypersaline mat, where the microbial mass developed in a biofilm [123]. Sequence reads from hypersaline mats in which LOV domain sequences were found, were less than 900 bp in length, which did not cover any output domain. The upper layers of the hypersaline mats were shown to be dominated by cyanobacteria and alphaproteobacteria according to the phylogenetic assignments of the sequences (Figure 5.10).

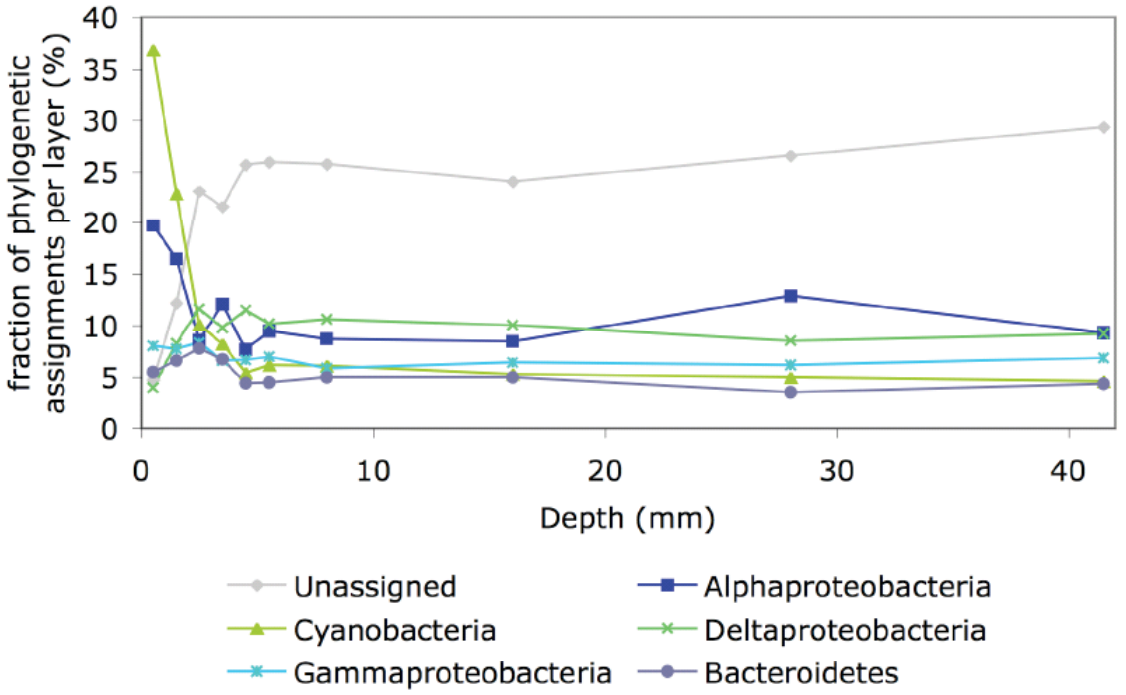


Figure 5.10 Phylogenetic assignments of sequences according to the depth of hyper saline mat layers. The affiliation is based on the 30% BLAST identity threshold against the IMG database. Only the most abundant phyla and proteobacterial classes are shown [123].

In the BLAST analysis it was also seen that the LOV domain sequences from the hypersaline mat were mostly affiliated to alphaproteobacteria, cyanobacteria and deltaproteobacteria (Figure 5.7). The presence of the LOV domain sequences in various biofilm environments in the analyzed metagenomes strengthens the involvement of this sensing domain in biofilm activities.

5.8.2 LOV domains in viruses

An investigation of the viral metagenomic databases revealed the presence of LOV domains in these collections of sequences derived from two sources: the mosquito viral metagenome and marine cyanophage metagenomes (5.3). Sequence reads from mosquito viral metagenome were short (ca 100 bp corresponding to 35 aa), therefore further analysis was not carried out. Cyanophage metagenome sequences showed LOV domain-specific sequential and phylogenetic features. This result confirms the presence of LOV domains in cyanophages. This is the first revelation of the presence of LOV domains in viral genomes. The phylogenetic tree constructed from the viral LOV domains, cyanobacterial LOV domains and selected metagenomic LOV domains (Figure 5.5) indicates that the cyanophage LOV are placed near to LOV domains from the marine metagenome. This branching could be explained by the fact that both these groups belong to marine habitats, where an infecting phage population can act as carrier for lateral gene transfer. A similar model of lateral gene transfer was proposed for photosystem genes, PSI and PSII, between cyanobacteria mediated by bacteriophages [276;277]. Such models are supported by phylogenetic and genetic evidences [278;279]. On the basis of the transfer mechanisms of photosystem genes and also considering the clustering of the phage LOV domains with those of the GOS metagenome, a lateral LOV domain gene transfer between the bacteria and phages is likely to take place in their natural environments. Due to their unique property of host infection and the capacity of integration into the host genome, it is quite possible that genes from the carrier phages can be transferred to the infected population.

Chapter 6

Conclusion and Outlook

6.1 Conclusion

This dissertation was an attempt to investigate metagenome samples for the presence of the blue light receptor LOV domains. Metagenomes are regarded as huge and unexplored reservoirs that harbor novel genes of ecological and biotechnological importance. Already today, sequence information obtained from metagenomes extends that from sequenced genomes by more than a factor of ten. However, there are several drawbacks related to metagenome research. One of the major problems for exploration and exploitation of such potential is the lack of techniques to detect and isolate genes and pathways in the absence of established cultivation methods.

While many researchers have employed metagenomics to identify novel biomolecules using functional assays, surprisingly few studies have so far employed microarray-based approaches to detect genes in community DNAs and to understand functional pathways. We have extended in this dissertation the application of a conventional DNA microarray and present for the first time an approach that includes the identification of a novel LOV protein-encoding gene from a soil metagenome without a prior sequence and functional information of the analyzed sample. This was made possible by including an additional search parameter on top of sequence similarities such that the presence of conserved and functionally essential residues indispensible for classification of a given sequence as a LOV domain sequence. This survey yielded 149 putative LOV domain containing gene sequences that could be used as probes for microarray fabrication (Chapter 3).

Furthermore, we have shown a functional proof of the encoded (recombinant) protein isolated from the metagenomic library. Chapter 4 presents the proof of functional novelty of the metagenome derived LOV domain with a remarkably fast dark recovery and some unique features in the amino acid sequence.

Biological photoreceptors are important biomolecules as they are involved in harvesting of the light energy and also in regulating pathways. Photoreceptor proteins have been significantly investigated in pure cell cultures or at the organism level, but metagenome level investigations of biological photoreceptors has been so far confined to proteorhodopsins. The investigation of LOV domain proteins in metagenomic samples that have been deposited in databases revealed that metagenome contains a large amount of novel LOV domain families (Chapter 5). Our investigation revealed the presence of LOV domains even in viruses, an observation that was not reported up to now. The presence of LOV domains in viruses indicates that LOV domains can have broad functional spectrum which could be unraveled by further analysis at the genomic and functional level. Moreover, it further stimulates for the consideration that blue light mediated phage infection and subsequent host cell lysis mechanisms might be triggered by LOV domain mediated sensing. This is quite possible in the deeper oceanic water where blue light is the main driving force for light mediated signaling processes.

In conclusion, development of new screening tools is highly awaited in metagenomics to unravel the hidden potential of metagenomes; in this aspect our work has demonstrated that the DNA microarray technique can be employed to detect unknown genes from such sources**.** The present work has further strengthened the potential of metagenomics as a source of novel genes and gene products for potential biotechnological applications.

6.2 Outlook

The DNA microarray-based high throughput techniques can be applied to conveniently screen thousands of metagenomic clones in parallel. Further functional characterization of the full length protein will help to understand the functional importance of, e.g., multi-domain sensor protein in signal transduction. Solving the crystal structures of the novel LOV domains can add more structural information that will be valuable for future investigations. Site directed mutagenesis will help understanding the fast dark recovery of HT-Met1 LOV, mostly considering its multiple histidine residues in that particular protein. LOV-microarrays can be used in future to

investigate additional samples from different environments. In addition, further probes from other light receptor or functional genes can be included in the microarray and applied to screen the environmental DNA yielding a more detailed knowledge of the functional composition of the microbial biomass of such environment. The DNA microarray approach based on the conserved motif as described in this work can be extended to investigate oxidoreductases, photolyases, hydrogenases, esterases and other proteins with conserved sequence motifs. We have already designed and applied a hydrogenase microarray based on similar technique successfully to screen for hydrogenase genes from the metagenome samples obtained from extreme environments. Inclusion of further probes to construct a comprehensive array will be a step further to screen multiple genes from metagenome samples in a single analysis.

Metagenomic analyses can offer insights into biogeographical distributions, community structure and ecological dynamics. Investigation of the metagenome will provide deeper insight into various photoregulatory pathways, which are still not understood in terms of complex natural environment. A metagenome level functional expression profiling of the photochemical properties of the LOV domains from various habitats can provide more insight on the light mediated ecological adaptation of these sensor domains. In addition, crucial genes for microbial adaptations in a given environment are likely to be moved by phages and can therefore be identified by analyzing the genomic content of viral communities. In this context, the analysis of the complete genome of phages that contain LOV domain genes will give important information about the characteristics of the viral LOV domain proteins and the roles they can play in viral life cycle, e.g. for host infectivity and lysis.

References

[1] Presti,D. & Delbrück,M. (1978) Photoreceptors for biosynthesis, energy storage and vision. *Plant Cell Environ* **1,** 81-100.

[2] Frohlich,H. (1983) Evidence for Coherent Excitation in Biological-Systems. *Int J Quantum Chem* **23,** 1589-1595.

[3] Purcell,E.B. & Crosson,S. (2008) Photoregulation in prokaryotes. *Curr Opin Microbiol* **11,** 168-178.

[4] Wilson,C., Caton,T.M., Buchheim,J.A., Buchheim,M.A., Schneegurt,M.A., & Miller,R.V. (2004) DNA-repair potential of *Halomonas* spp. from the Salt Plains Microbial Observatory of Oklahoma. *Microb Ecol* **48,** 541-549.

[5] Van der Horst,M.A. & Hellingwerf,K.J. (2004) Photoreceptor proteins, "star actors of modern times": A review of the functional dynamics in the structure of representative members of six different photoreceptor families. *Acc Chem Res* **37,** 13-20.

[6] Schegk,E.S. & Oesterhelt,D. (1988) Isolation of A Prokaryotic Photoreceptor - Sensory Rhodopsin from Halobacteria. *EMBO J* **7,** 2925-2933.

[7] Spudich,J.L., Yang,C.S., Jung,K.H., & Spudich,E.N. (2000) Retinylidene proteins: Structures and functions from archaea to humans. *Annu Rev Cell Dev Biol* **16,** 365-392.

[8] Beja,O., Aravind,L., Koonin,E.V., Suzuki,M.T., Hadd,A., Nguyen,L.P., Jovanovich,S., Gates,C.M., Feldman,R.A., Spudich,J.L., Spudich,E.N., & Delong,E.F. (2000) Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea. *Science* **289,** 1902-1906.

[9] Jung,K.H., Trivedi,V.D., & Spudich,J.L. (2003) Demonstration of a sensory rhodopsin in eubacteria. *Mol Microbiol* **47,** 1513-1522.

[10] Sineshchekov,O.A., Jung,K.H., & Spudich,J.L. (2002) Two rhodopsins mediate phototaxis to low- and high-intensity light in *Chlamydomonas reinhardtii*. *Proc Natl Acad Sci USA* **99,** 8689-8694.

[11] Bieszke,J.A., Spudich,E.N., Scott,K.L., Borkovich,K.A., & Spudich,J.L. (1999) A eukaryotic protein, NOP-1, binds retinal to form an archaeal rhodopsin-like photochemically reactive pigment. *Biochemistry* **38,** 14138-14145.

[12] Davis,S.J., Vener,A.V., & Vierstra,R.D. (1999) Bacteriophytochromes: Phytochrome-like photoreceptors from nonphotosynthetic eubacteria. *Science* **286,** 2517-2520.

[13] Jiang,Z.Y., Swem,L.R., Rushing,B.G., Devanathan,S., Tollin,G., & Bauer,C.E. (1999) Bacterial photoreceptor with similarity to photoactive yellow protein and plant phytochromes. *Science* **285,** 406-409.

[14] Purschwitz,J., Muller,S., Kastner,C., & Fischer,R. (2006) Seeing the rainbow: light sensing in fungi. *Curr Opin Microbiol* **9,** 566-571.

[15] Losi,A. & Gartner,W. (2008) Bacterial bilin- and flavin-binding photoreceptors. *Photochem Photobiol Sci* **7,** 1168-1178.

[16] Giraud,E., Fardoux,J., Fourrier,N., Hannibal,L., Genty,B., Bouyer,P., Dreyfus,B., & Vermeglio,A. (2002) Bacteriophytochrome controls photosystem synthesis in anoxygenic bacteria. *Nature* **417,** 202-205.

[17] Kehoe,D.M. & Grossman,A.R. (1997) New classes of mutants in complementary chromatic adaptation provide evidence for a novel four-step phosphorelay system. *J Bacteriol* **179,** 3914-3921.

[18] Wilde,A., Churin,Y., Schubert,H., & Borner,T. (1997) Disruption of a Synechocystis sp. PCC 6803 gene with partial similarity to phytochrome genes alters growth under changing light qualities. *FEBS Lett* **406,** 89-92.

[19] Kort,R., Vonk,H., Xu,X., Hoff,W.D., Crielaard,W., & Hellingwerf,K.J. (1996) Evidence for trans-cis isomerization of the p-coumaric acid chromophore as the photochemical basis of the photocycle of photoactive yellow protein. *FEBS Lett* **382,** 73-78.

[20] Meyer,T.E. (1985) Isolation and Characterization of Soluble Cytochromes, Ferredoxins and Other Chromophoric Proteins from the Halophilic Phototrophic Bacterium *Ectothiorhodospira halophila*. *Biochim Biophys Acta, Biophys* **806,** 175-183.

[21] Imamoto,Y. & Kataoka,M. (2007) Structure and photoreaction of photoactive yellow protein, a structural prototype of the PAS domain superfamily. *Photochem Photobiol* **83,** 40-49.

[22] Sprenger,W.W., Hoff,W.D., Armitage,J.P., & Hellingwerf,K.J. (1993) The Eubacterium *Ectothiorhodospira halophila* is Negatively Phototactic, with A Wavelength Dependence That Fits the Absorption-Spectrum of the Photoactive Yellow Protein. *J Bacteriol* **175,** 3096-3104.

[23] Kyndt,J.A., Meyer,T.E., & Cusanovich,M.A. (2004) Photoactive yellow protein, bacteriophytochrome, and sensory rhodopsin in purple phototrophic bacteria. *Photochem Photobiol Sci* **3,** 519-530.

[24] Kirk,J.T.O. (1994) Estimation of the Absorption and the Scattering Coefficients of Natural-Waters by Use of Underwater Irradiance Measurements. *Appl Opt* **33,** 3276-3278.

[25] Mitchell,D.L. (1988) The Relative Cyto-Toxicity of (6-4) Photoproducts and Cyclobutane Dimers in Mammalian-Cells. *Photochem Photobiol* **48,** 51-57.

[26] Redmond,R.W. & Gamlin,J.N. (1999) A compilation of singlet oxygen yields from biologically relevant molecules. *Photochem Photobiol* **70,** 391-475.

[27] Ahmad,M. & Cashmore,A.R. (1993) Hy4 Gene of *A. thaliana* Encodes A Protein with Characteristics of A Blue-Light Photoreceptor. *Nature* **366,** 162-166.

[28] Lin,C.T. & Todo,T. (2005) The cryptochromes. *Genome Biol* **6,** 220 .

[29] Iseki,M., Matsunaga,S., Murakami,A., Ohno,K., Shiga,K., Yoshida,K., Sugai,M., Takahashi,T., Hori,T., & Watanabe,M. (2002) A blue-light-activated adenylyl cyclase mediates photoavoidance in *Euglena gracilis*. *Nature* **415,** 1047-1051.

[30] Sancar,A. (2004) *Photolyase and cryptochrome blue-light photoreceptors*. ELSEVIER ACADEMIC PRESS INC, SAN DIEGO.

[31] Brudler,R., Hitomi,K., Daiyasu,H., Toh,H., Kucho,K.i., Ishiura,M., Kanehisa,M., Roberts,V.A., Todo,T., Tainer,J.A., & Getzoff,E.D. (2003) Identification of a New Cryptochrome Class: Structure, Function, and Evolution. *Molecular Cell* **11,** 59-67.

[32] Worthington,E.N., Kavakli,I.H., Berrocal-Tito,G., Bondo,B.E., & Sancar,A. (2003) Purification and characterization of three members of the photolyase/cryptochrome family blue-light photoreceptors from *Vibrio cholerae*. *J Biol Chem* **278,** 39143-39154.

[33] Gomelsky,M. & Klug,G. (2002) BLUF: a novel FAD-binding domain involved in sensory transduction in microorganisms. *Trends Biochem Sci* **27,** 497-500.

[34] Masuda,S., Hasegawa,K., & Ono,T. (2005) Light-induced structural changes of apoprotein and chromophore in the sensor of blue light using FAD (BLUF) domain of AppA for a signaling state. *Biochemistry* **44,** 1215-1224.

[35] Liscum,E. & Briggs,W.R. (1995) Mutations in the Nph1 Locus of *Arabidopsis* Disrupt the Perception of Phototropic Stimuli. *Plant Cell* **7,** 473-485.

[36] Christie,J.M., Salomon,M., Nozue,K., Wada,M., & Briggs,W.R. (1999) LOV (light, oxygen, or voltage) domains of the blue-light photoreceptor phototropin (nph1): Binding sites for the chromophore flavin mononucleotide. *Proc Natl Acad Sci USA* **96,** 8779-8783.

[37] Huala,E., Oeller,P.W., Liscum,E., Han,I.S., Larsen,E., & Briggs,W.R. (1997) *Arabidopsis* NPH1: A protein kinase with a putative redox-sensing domain. *Science* **278,** 2120-2123.

[38] Cheng,P., He,Q.Y., Yang,Y.H., Wang,L.X., & Liu,Y. (2003) Functional conservation of light, oxygen, or voltage domains in light sensing. *Proc Natl Acad Sci USA* **100,** 5938-5943.

[39] Huang,K.Y. & Beck,C.F. (2003) Photoropin is the blue-light receptor that controls multiple steps in the sexual life cycle of the green alga *Chlamydomonas reinhardtii*. *Proc Natl Acad Sci USA* **100,** 6269-6274.

[40] Nozue,K., Kanegae,T., Imaizumi,T., Fukuda,S., Okamoto,H., Yeh,K.C., Lagarias,J.C., & Wada,M. (1998) A phytochrome from the fern *Adiantum* with features of the putative photoreceptor NPH1. *Proc Natl Acad Sci USA* **95,** 15826-15830.

[41] Swartz,T.E., Corchnoy,S.B., Christie,J.M., Lewis,J.W., Szundi,I., Briggs,W.R., & Bogomolni,R.A. (2001) The photocycle of a flavin-binding domain of the blue light photoreceptor phototropin. *J Biol Chem* **276,** 36493-36500.

[42] Crosson,S. & Moffat,K. (2002) Photoexcited structure of a plant photoreceptor domain reveals a light-driven molecular switch. *Plant Cell* **14,** 1067-1075.

[43] Salomon,M., Christie,J.M., Knieb,E., Lempert,U., & Briggs,W.R. (2000) Photochemical and mutational analysis of the FMN-binding domains of the plant blue light receptor, phototropin. *Biochemistry* **39,** 9401-9410.

[44] Kottke,T., Heberle,J., Hehn,D., Dick,B., & Hegemann,P. (2003) Phot-LOV1: Photocycle of a blue-light receptor domain from the green alga *Chlamydomonas reinhardtii*. *Biophys J* **84,** 1192-1201.

[45] Jentzsch,K., Wirtz,A., Circolone,F., Drepper,T., Losi,A., Gaertner,W., Jaeger,K.E., & Krauss,U. (2009) Mutual Exchange of Kinetic Properties by Extended Mutagenesis in Two Short LOV Domain Proteins from *Pseudomonas putida*. *Biochemistry* **48,** 10321-10333.

[46] Losi,A. (2004) The bacterial counterparts of plant phototropins. *Photochem Photobiol Sci* **3,** 566-574.

[47] Crosson,S. & Moffat,K. (2001) Structure of a flavin-binding plant photoreceptor domain: Insights into light-mediated signal transduction. *Proc Natl Acad Sci USA* **98,** 2995-3000.

[48] Fedorov,R., Schlichting,I., Hartmann,E., Domratcheva,T., Fuhrmann,M., & Hegemann,P. (2003) Crystal structures and molecular mechanism of a light-induced signaling switch: The Phot-LOV1 domain from Chlamydomonas reinhardtii. *Biophys J* **84,** 2474-2482.

[49] Moglich,A. & Moffat,K. (2007) Structural basis for light-dependent signaling in the dimeric LOV domain of the photosensor YtvA. *J Mol Biol* **373,** 112-126.

[50] Crosson,S., Rajagopal,S., & Moffat,K. (2003) The LOV domain family: Photoresponsive signaling modules coupled to diverse output domains. *Biochemistry* **42,** 2-10.

[51] Jones,M.A., Feeney,K.A., Kelly,S.M., & Christie,J.M. (2007) Mutational analysis of phototropin 1 provides insights into the mechanism underlying LOV2 signal transmission. *J Biol Chem* **282,** 6405-6414.

[52] Nozaki,D., Iwata,T., Ishikawa,T., Todo,T., Tokutomi,S., & Kandori,H. (2004) Role of Gln1029 in the photoactivation processes of the LOV2 domain in *Adiantum* phytochrome3. *Biochemistry* **43,** 8373-8379.

[53] Iwata,T., Nozaki,D., Tokutomi,S., & Kandori,H. (2005) Comparative investigation of the LOV1 and LOV2 domains in *Adiantum* phytochrome3. *Biochemistry* **44,** 7427-7434.

[54] Buttani,V., Losi,A., Eggert,T., Krauss,U., Jaeger,K.E., Cao,Z., & Gartner,W. (2007) Conformational analysis of the blue-light sensing protein YtvA reveals a competitive interface for LOV-LOV dimerization and interdomain interactions. *Photochem Photobiol Sci* **6,** 41-49.

[55] Harper,S.M., Neil,L.C., Day,I.J., Hore,P.J., & Gardner,K.H. (2004) Conformational changes in a photosensory LOV domain monitored by time-resolved NMR spectroscopy. *J Am Chem Soc* **126,** 3390-3391.

[56] Gaba,V. & Black,M. (1979) Two separate photoreceptors control hypocotyl growth in green seedlings. *Nature* **278,** 51-54.

[57] Gaidenko,T.A., Kim,T.J., Weigel,A.L., Brody,M.S., & Price,C.W. (2006) The blue-light receptor YtvA acts in the environmental stress signaling pathway of *Bacillus subtilis*. *J Bacteriol* **188,** 6387-6395.

[58] Christie,J.M., Corchnoy,S.B., Swartz,T.E., Hokenson,M., Han,I.S., Briggs,W.R., & Bogomolni,R.A. (2007) Steric interactions stabilize the signaling state of the LOV2 domain of phototropin 1. *Biochemistry* **46,** 9310-9319.

[59] Whitman,W.B., Coleman,D.C., & Wiebe,W.J. (1998) Prokaryotes: The unseen majority. *Proc Natl Acad Sci USA* **95,** 6578-6583.

[60] Perner,M., Seifert,R., Weber,S., Koschinsky,A., Schmidt,K., Strauss,H., Peters,M., Haase,K., & Imhoff,J.F. (2007) Microbial CO2 fixation and sulfur cycling associated with low-temperature emissions at the Lilliput hydrothermal field, southern Mid-Atlantic Ridge (9.S). *Environ Microbiol* **9,** 1186-1201.

[61] Martin,W., Baross,J., Kelley,D., & Russell,M.J. (2008) Hydrothermal vents and the origin of life. *Nat Rev Microbiol* **6,** 805-814.

[62] Tyson,G.W., Chapman,J., Hugenholtz,P., Allen,E.E., Ram,R.J., Richardson,P.M., Solovyev,V.V., Rubin,E.M., Rokhsar,D.S., & Banfield,J.F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428,** 37-43.

[63] Warnecke,F., Luginbuhl,P., Ivanova,N., Ghassemian,M., Richardson,T.H., Stege,J.T., Cayouette,M., McHardy,A.C., Djordjevic,G., Aboushadi,N., Sorek,R., Tringe,S.G., Podar,M., Martin,H.G., Kunin,V., Dalevi,D., Madejska,J., Kirton,E., Platt,D., Szeto,E., Salamov,A., Barry,K., Mikhailova,N., Kyrpides,N.C., Matson,E.G., Ottesen,E.A., Zhang,X.N., Hernandez,M., Murillo,C., Acosta,L.G., Rigoutsos,I., Tamayo,G., Green,B.D., Chang,C., Rubin,E.M., Mathur,E.J., Robertson,D.E., Hugenholtz,P., & Leadbetter,J.R. (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450,** 560-565.

[64] Muir,J. (1911) *My First Summer in the Sierra*. Boston: Houghton Mifflin.

[65] Amann,R.I., Ludwig,W., & Schleifer,K.H. (1995) Phylogenetic Identification and In-Situ Detection of Individual Microbial-Cells Without Cultivation. *Microbiol Rev* **59,** 143-169.

[66] Curtis,T.P. & Sloan,W.T. (2005) Exploring microbial diversity - A vast below. *Science* **309,** 1331-1333.

[67] Curtis,T.P., Sloan,W.T., & Scannell,J.W. (2002) Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci* **99,** 10494-10499.

[68] Gill,S.R., Pop,M., DeBoy,R.T., Eckburg,P.B., Turnbaugh,P.J., Samuel,B.S., Gordon,J.I., Relman,D.A., Fraser-Liggett,C.M., & Nelson,K.E. (2006) Metagenomic Analysis of the Human Distal Gut Microbiome. *Science* **312,** 1355-1359.

[69] Handelsman,J., Rondon,M.R., Brady,S.F., Clardy,J., & Goodman,R.M. (1998) Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chem Biol* **5,** 245-249.

[70] Handelsman,J. (2004) Metagenomics: Application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* **68,** 669-685.

[71] Streit,W.R. & Schmitz,R.A. (2004) Metagenomics - the key to the uncultured microbes. *Curr Opin Microbiol* **7,** 492-498.

[72] Schloss,P.D. & Handelsman,J. (2003) Biotechnological prospects from metagenomics. *Curr Opin Biotechnol* **14,** 303-310.

[73] Delong,E.F. (2004) Microbial population genomics and ecology: the road ahead. *Environ Microbiol* **6,** 875-878.

[74] Ward,N. (2006) New directions and interactions in metagenomics research. *FEMS Microbiol Ecol* **55,** 331-338.

[75] Eyers,L., George,I., Schuler,L., Stenuit,B., Agathos,S.N., & El Fantroussi,S. (2004) Environmental genomics: exploring the unmined richness of microbes to degrade xenobiotics. *Appl Microbiol Biotechnol* **66,** 123-130.

[76] Daniel,R. (2005) The metagenomics of soil. *Nat Rev Microbiol* **3,** 470-478.

[77] Gabor,E., Liebeton,K., Niehaus,F., Eck,J., & Lorenz,P. (2007) Updating the metagenomics toolbox. *Biotechnol J* **2,** 201-206.

[78] Gillespie,D.E., Brady,S.F., Bettermann,A.D., Cianciotto,N.P., Liles,M.R., Rondon,M.R., Clardy,J., Goodman,R.M., & Handelsman,J. (2002) Isolation of Antibiotics Turbomycin A and B from a Metagenomic Library of Soil Microbial DNA. *Appl Environ Microbiol* **68,** 4301-4306.

[79] Entcheva,P., Liebl,W., Johann,A., Hartsch,T., & Streit,W.R. (2001) Direct cloning from enrichment cultures, a reliable strategy for isolation of complete operons and genes from microbial consortia. *Appl Environ Microbiol* **67,** 89-99.

[80] Elend,C., Schmeisser,C., Leggewie,C., Babiak,P., Carballeira,J.D., Steele,H.L., Reymond,J.L., Jaeger,K.E., & Streit,W.R. (2006) Isolation and biochemical characterization of two novel metagenome-derived esterases. *Appl Environ Microbiol* **72,** 3637-3645.

[81] Schipper,C., Hornung,C., Bijtenhoorn,P., Quitschau,M., Grond,S., & Streit,W.R. (2009) Metagenome-Derived Clones Encoding Two Novel Lactonase Family Proteins Involved in Biofilm Inhibition in *Pseudomonas aeruginosa*. *Appl Environ Microbiol* **75,** 224-233.

[82] Woyke,T., Teeling,H., Ivanova,N.N., Huntemann,M., Richter,M., Gloeckner,F.O., Boffelli,D., Anderson,I.J., Barry,K.W., Shapiro,H.J., Szeto,E., Kyrpides,N.C., Mussmann,M., Amann,R., Bergin,C., Ruehland,C., Rubin,E.M., & Dubilier,N. (2006) Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443,** 950-955.

[83] Stein,J.L., Marsh,T.L., Wu,K.Y., Shizuya,H., & Delong,E.F. (1996) Characterization of uncultivated prokaryotes: Isolation and analysis of a 40-kilobase-pair genome fragment front a planktonic marine archaeon. *J Bacteriol* **178,** 591-599.

[84] Sleator,R.D., Shortall,C., & Hill,C. (2008) Metagenomics. *Lett Appl Microbiol* **47,** 361-366.

[85] Rondon,M.R., August,P.R., Bettermann,A.D., Brady,S.F., Grossman,T.H., Liles,M.R., Loiacono,K.A., Lynch,B.A., MacNeil,I.A., Minor,C., Tiong,C.L., Gilman,M., Osburne,M.S., Clardy,J., Handelsman,J., & Goodman,R.M. (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol* **66,** 2541-2547.

[86] Courtois,S., Cappellano,C.M., Ball,M., Francou,F.X., Normand,P., Helynck,G., Martinez,A., Kolvek,S.J., Hopke,J., Osburne,M.S., August,P.R., Nalin,R., Guerineau,M., Jeannin,P., Simonet,P., & Pernodet,J.L. (2003) Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. *Appl Environ Microbiol* **69,** 49-55.

[87] DeLong,E.F., Preston,C.M., Mincer,T., Rich,V., Hallam,S.J., Frigaard,N.U., Martinez,A., Sullivan,M.B., Edwards,R., Brito,B.R., Chisholm,S.W., & Karl,D.M. (2006) Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. *Science* **311,** 496-503.

[88] Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z.T., Dewell,S.B., Du,L., Fierro,J.M., Gomes,X.V., Godwin,B.C., He,W., Helgesen,S., Ho,C.H., Irzyk,G.P., Jando,S.C., Alenquer,M.L.I., Jarvie,T.P., Jirage,K.B., Kim,J.B., Knight,J.R., Lanza,J.R., Leamon,J.H., Lefkowitz,S.M., Lei,M., Li,J., Lohman,K.L., Lu,H., Makhijani,V.B., Mcdade,K.E., McKenna,M.P., Myers,E.W., Nickerson,E., Nobile,J.R., Plant,R., Puc,B.P., Ronan,M.T., Roth,G.T., Sarkis,G.J., Simons,J.F., Simpson,J.W., Srinivasan,M., Tartaro,K.R., Tomasz,A., Vogt,K.A., Volkmer,G.A., Wang,S.H., Wang,Y., Weiner,M.P., Yu,P.G., Begley,R.F., & Rothberg,J.M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376-380.

[89] Dinsdale,E.A., Edwards,R.A., Hall,D., Angly,F., Breitbart,M., Brulc,J.M., Furlan,M., Desnues,C., Haynes,M., Li,L.L., McDaniel,L., Moran,M.A., Nelson,K.E., Nilsson,C., Olson,R., Paul,J., Brito,B.R., Ruan,Y.J., Swan,B.K., Stevens,R., Valentine,D.L., Thurber,R.V., Wegley,L., White,B.A., & Rohwer,F. (2008) Functional metagenomic profiling of nine biomes. *Nature* **452,** 629-632.

[90] Steele,H.L. & Streit,W.R. (2005) Metagenomics: Advances in ecology and biotechnology. *FEMS Microbiol Lett* **247,** 105-111.

[91] Noonan,J.P., Hofreiter,M., Smith,D., Priest,J.R., Rohland,N., Rabeder,G., Krause,J., Detter,J.C., Paabo,S., & Rubin,E.M. (2005) Genomic sequencing of Pleistocene cave bears. *Science* **309,** 597-600.

[92] Poinar,H.N., Schwarz,C., Qi,J., Shapiro,B., MacPhee,R.D.E., Buigues,B., Tikhonov,A., Huson,D.H., Tomsho,L.P., Auch,A., Rampp,M., Miller,W., & Schuster,S.C. (2006) Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA. *Science* **311,** 392-394.

[93] Rusch,D.B., Halpern,A.L., Sutton,G., Heidelberg,K.B., Williamson,S., Yooseph,S., Wu,D.Y., Eisen,J.A., Hoffman,J.M., Remington,K., Beeson,K., Tran,B., Smith,H., Baden-Tillson,H., Stewart,C., Thorpe,J., Freeman,J., ndrews-Pfannkoch,C., Venter,J.E., Li,K., Kravitz,S., Heidelberg,J.F., Utterback,T., Rogers,Y.H., Falcon,L.I., Souza,V., Bonilla-Rosso,G., Eguiarte,L.E., Karl,D.M., Sathyendranath,S., Platt,T., Bermingham,E., Gallardo,V., Tamayo-Castillo,G., Ferrari,M.R., Strausberg,R.L., Nealson,K., Friedman,R., Frazier,M., & Venter,J.C. (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PloS Biol* **5,** 398-431.

[94] Voget,S., Steele,H.L., & Streit,W.R. (2006) Characterization of a metagenome-derived halotolerant cellulase. *J Biotechnol* **126,** 26-36.

[95] van Elsas,J.D., Costa,R., Jansson,J., Sj÷ling,S., Bailey,M., Nalin,R., Vogel,T.M., & van Overbeek,L. (2008) The metagenomics of disease-suppressive soils - experiences from the METACONTROL project. *Trends Biotechnol* **26,** 591-601.

[96] Preidis,G.A. & Versalovic,J. (2009) Targeting the Human Microbiome With Antibiotics, Probiotics, and Prebiotics: Gastroenterology Enters the Metagenomics Era. *Gastroenterology* **136,** 2015-2031.

[97] Lorenz,P. & Eck,J. (2005) Metagenomics and industrial applications. *Nat Rev Microbiol* **3,** 510-516.

[98] Lefevre,F., Robe,P., Jarrin,C., Ginolhac,A., Zago,C., Auriol,D., Vogel,T.M., Simonet,P., & Nalin,R. (2008) Drugs from hidden bugs: their discovery via untapped resources. *Res Microbiol* **159,** 153-161.

[99] Langer,M., Gabor,E.M., Liebeton,K., Meurer,G., Niehaus,F., Schulze,R., Eck,J., & Lorenz,P.P.d. (2006) Metagenomics: An inexhaustible access to nature's diversity. *Biotechnol J* **1,** 815-821.

[100] Voget,S., Leggewie,C., Uesbeck,A., Raasch,C., Jaeger,K.E., & Streit,W.R. (2003) Prospecting for novel biocatalysts in a soil metagenome. *Appl Environ Microbiol* **69,** 6235-6242.

[101] Pottkamper,J., Barthen,P., Ilmberger,N., Schwaneberg,U., Schenk,A., Schulte,M., Ignatiev,N., & Streit,W.R. (2009) Applying metagenomics for the identification of bacterial cellulases that are stable in ionic liquids. *Green Chem* **11,** 957-965.

[102] Henne,A., Schmitz,R.A., Bomeke,M., Gottschalk,G., & Daniel,R. (2000) Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on Escherichia coli. *Appl Environ Microbiol* **66,** 3113-3116.

[103] Rhee,J.K., Ahn,D.G., Kim,Y.G., & Oh,J.W. (2005) New thermophilic and thermostable esterase with sequence similarity to the hormone-sensitive lipase family, cloned from a metagenomic library. *Appl Environ Microbiol* **71,** 817-825.

[104] Kim,Y.J., Choi,G.S., Kim,S.B., Yoon,G.S., Kim,Y.S., & Ryu,Y.W. (2006) Screening and characterization of a novel esterase from a metagenomic library. *Protein Expression Purif* **45,** 315-323.

[105] Steele,H.L., Jaeger,K.E., Daniel,R., & Streit,W.R. (2009) Advances in Recovery of Novel Biocatalysts from Metagenomes. *J Mol Microbiol Biotechnol* **16,** 25-37.

[106] Kurokawa,K., Itoh,T., Kuwahara,T., Oshima,K., Toh,H., Toyoda,A., Takami,H., Morita,H., Sharma,V.K., Srivastava,T.P., Taylor,T.D., Noguchi,H., Mori,H., Ogura,Y., Ehrlich,D.S., Itoh,K., Takagi,T., Sakaki,Y., Hayashi,T., & Hattori,M. (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* **14,** 169-181.

[107] Li,M., Wang,B., Zhang,M., Rantalainen,M., Wang,S., Zhou,H., Zhang,Y., Shen,J., Pang,X., Zhang,M., Wei,H., Chen,Y., Lu,H., Zuo,J., Su,M., Qiu,Y., Jia,W., Xiao,C., Smith,L.M., Yang,S., Holmes,E., Tang,H., Zhao,G., Nicholson,J.K., Li,L., & Zhao,L. (2008) Symbiotic gut microbes modulate human metabolic phenotypes. *Proc Natl Acad Sci* **105,** 2117-2122.

[108] Backhed,F., Ley,R.E., Sonnenburg,J.L., Peterson,D.A., & Gordon,J.I. (2005) Host-bacterial mutualism in the human intestine. *Science* **307,** 1915-1920.

[109] Ley,R.E., Turnbaugh,P.J., Klein,S., & Gordon,J.I. (2006) Microbial ecology: Human gut microbes associated with obesity. *Nature* **444,** 1022-1023.

[110] Tschop,M.H., Hugenholtz,P., & Karp,C.L. (2009) Getting to the core of the gut microbiome. *Nat Biotech* **27,** 344-346.

[111] Peterson,D.E., Bensadoun,R.J., & Roila,F. (2008) Management of oral and gastrointestinal mucositis: ESMO clinical recommendations. *Ann Oncol* **19,** 122-125.

[112] Dicksved,J., Halfvarson,J., Rosenquist,M., Jarnerot,G., Tysk,C., Apajalahti,J., Engstrand,L., & Jansson,J.K. (2008) Molecular analysis of the gut microbiota of identical twins with Crohn's disease. *ISME J* **2,** 716-727.

[113] Schluter,A., Bekel,T., Diaz,N.N., Dondrup,M., Eichenlaub,R., Gartemann,K.H., Krahn,I., Krause,L., Kromeke,H., Kruse,O., Mussgnug,J.H., Neuweger,H., Niehaus,K., Puhler,A., Runte,K.J., Szczepanowski,R., Tauch,A., Tilker,A., Viehover,P., & Goesmann,A. (2008) The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. *J Biotechnol* **136,** 77-90.

[114] Tsai,H.Y., Wu,C.C., Lee,C.Y., & Shih,E.P. (2009) Microbial fuel cell performance of multiwall carbon nanotubes on carbon cloth as electrodes. *J Power Sources* **194,** 199-205.

[115] Ono,A., Miyazaki,R., Sota,M., Ohtsubo,Y., Nagata,Y., & Tsuda,M. (2007) Isolation and characterization of naphthalene-catabolic genes and plasmids from oil-contaminated soil by using two cultivation-independent approaches. *Appl Microbiol Biotechnol* **74,** 501-510.

[116] Silva,s.S., Santos,E.d.C.d., Menezes,C.R.d., Faria,A.F.d., Franciscon,E., Grossman,M., & Durrant,L.R. (2009) Bioremediation of a polyaromatic hydrocarbon contaminated soil by native soil microbiota and bioaugmentation with isolated microbial consortia. *Bioresour Technol* **100,** 4669-4675.

[117] Hugenholtz,P. & Tyson,G.W. (2008) Microbiology: Metagenomics. *Nature* **455,** 481-483.

[118] Karl,D.M. (2007) Microbial oceanography: paradigms, processes and promise. *Nat Rev Microbiol* **5,** 759-769.

[119] Singh,A.H., Doerks,T., Letunic,I., Raes,J., & Bork,P. (2009) Discovering Functional Novelty in Metagenomes: Examples from Light-Mediated Processes. *J Bacteriol* **191,** 32-41.

[120] Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D.Y., Paulsen,I., Nelson,K.E., Nelson,W., Fouts,D.E., Levy,S., Knap,A.H., Lomas,M.W., Nealson,K., White,O., Peterson,J., Hoffman,J., Parsons,R., Baden-Tillson,H., Pfannkoch,C., Rogers,Y.H., & Smith,H.O. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304,** 66-74.

[121] Yooseph,S., Sutton,G., Rusch,D.B., Halpern,A.L., Williamson,S.J., Remington,K., Eisen,J.A., Heidelberg,K.B., Manning,G., Li,W.Z., Jaroszewski,L., Cieplak,P., Miller,C.S., Li,H.Y., Mashiyama,S.T., Joachimiak,M.P., van Belle,C., Chandonia,J.M., Soergel,D.A., Zhai,Y.F., Natarajan,K., Lee,S., Raphael,B.J., Bafna,V., Friedman,R., Brenner,S.E., Godzik,A., Eisenberg,D., Dixon,J.E., Taylor,S.S., Strausberg,R.L., Frazier,M., & Venter,J.C. (2007) The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein families. *PloS Biol* **5,** 432-466.

[122] Martin-Cuadrado,A.B., Rodriguez-Valera,F., Moreira,D., Alba,J.C., Ivars-Martinez,E., Henn,M.R., Talla,E., & Lopez-Garcia,P. (2008) Hindsight in the relative abundance, metabolic potential and genome dynamics of uncultivated marine archaea from comparative metagenomic analyses of bathypelagic plankton of different oceanic regions. *ISME J* **2,** 865-886.

[123] Kunin,V., Raes,J., Harris,J.K., Spear,J.R., Walker,J.J., Ivanova,N., von Mering,C., Bebout,B.M., Pace,N.R., Bork,P., & Hugenholtz,P. (2008) Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. *Mol Syst Biol* **4,** 198.

[124] Tringe,S.G., von Mering,C., Kobayashi,A., Salamov,A.A., Chen,K., Chang,H.W., Podar,M., Short,J.M., Mathur,E.J., Detter,J.C., Bork,P., Hugenholtz,P., & Rubin,E.M. (2005) Comparative Metagenomics of Microbial Communities. *Science* **308,** 554-557.

[125] Tringe,S.G., Zhang,T., Liu,X., Yu,Y., Lee,W.H., Yap,J., Yao,F., Suan,S.T., Ing,S.K., Haynes,M., Rohwer,F., Wei,C.L., Tan,P., Bristow,J., Rubin,E.M., & Ruan,Y. (2008) The Airborne Metagenome in an Indoor Urban Environment. *Plos One* **3,** e1862.

[126] Green,R.E., Krause,J., Ptak,S.E., Briggs,A.W., Ronan,M.T., Simons,J.F., Du,L., Egholm,M., Rothberg,J.M., Paunovic,M., & Paabo,S. (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature* **444,** 330-336.

[127] Martin,H.G., Ivanova,N., Kunin,V., Warnecke,F., Barry,K.W., McHardy,A.C., Yeates,C., He,S., Salamov,A.A., Szeto,E., Dalin,E., Putnam,N.H., Shapiro,H.J., Pangilinan,J.L., Rigoutsos,I., Kyrpides,N.C., Blackall,L.L., McMahon,K.D., & Hugenholtz,P. (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotech* **24,** 1263-1269.

[128] Ram,R.J., VerBerkmoes,N.C., Thelen,M.P., Tyson,G.W., Baker,B.J., Blake,R.C., II, Shah,M., Hettich,R.L., & Banfield,J.F. (2005) Community Proteomics of a Natural Microbial Biofilm. *Science* **308,** 1915-1920.

[129] Turnbaugh,P.J., Ley,R.E., Hamady,M., Fraser-Liggett,C.M., Knight,R., & Gordon,J.I. (2007) The Human Microbiome Project. *Nature* **449,** 804-810.

[130] Blow,N. (2008) Metagenomics: Exploring unseen communities. *Nature* **453,** 687-690.

[131] Schmeisser,C., Steele,H., & Streit,W.R. (2007) Metagenomics, biotechnology with non-culturable microbes. *Appl Environ Microbiol* **75,** 955-962.

[132] Raes,J., Foerstner,K.U., & Bork,P. (2007) Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr Opin Microbiol* **10,** 490-498.

[133] MacLean,D., Jones,J.D.G., & Studholme,D.J. (2009) Application of 'next-generation' sequencing technologies to microbial genetics. *Nat Rev Micro* **7,** 287-296.

[134] Huson,D.H., Richter,D.C., Mitra,S., Auch,A.F., & Schuster,S.C. (2009) Methods for comparative metagenomics. *BMC Bioinf* **10**.

[135] Fuhrman,J.A., Schwalbach,M.S., & Stingl,U. (2008) Proteorhodopsins: an array of physiological roles? *Nat Rev Micro* **6,** 488-494.

[136] Beja,O., Spudich,E.N., Spudich,J.L., Leclerc,M., & Delong,E.F. (2001) Proteorhodopsin phototrophy in the ocean. *Nature* **411,** 786-789.

[137] Papke,R.T., Douady,C.J., Doolittle,W.F., & Rodriguez-Valera,F. (2003) Diversity of bacteriorhodopsins in different hypersaline waters from a single Spanish saltern. *Environ Microbiol* **5,** 1039-1045.

[138] Lami,R., Cottrell,M.T., Campbell,B.J., & Kirchman,D.L. (2009) Light-dependent growth and proteorhodopsin expression by Flavobacteria and SAR11 in experiments with Delaware coastal waters. *Environ Microbiol* **11,** 3201-3209.

[139] Sharma,A.K., Sommerfeld,K., Bullerjahn,G.S., Matteson,A.R., Wilhelm,S.W., Jezbera,J., Brandt,U., Doolittle,W.F., & Hahn,M.W. (2009) Actinorhodopsin genes discovered in diverse freshwater habitats and among cultivated freshwater Actinobacteria. *ISME J* **3,** 726-737.

[140] Van der Horst,M.A., Key,J., & Hellingwerf,K.J. (2007) Photosensing in chemotrophic, non-phototrophic bacteria: let there be light sensing too. *Trends Microbiol* **15,** 554-562.

[141] Frigaard,N.U., Martinez,A., Mincer,T.J., & Delong,E.F. (2006) Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439,** 847-850.

[142] Schena,M., Shalon,D., Davis,R.W., & Brown,P.O. (1995) Quantitative Monitoring of Gene-Expression Patterns with A Complementary-Dna Microarray. *Science* **270,** 467-470.

[143] Ehrenreich,A. (2006) DNA microarray technology for the microbiologist: an overview. *Appl Microbiol Biotechnol* **73,** 255-273.

[144] Schena,M., Shalon,D., Heller,R., Chai,A., Brown,P.O., & Davis,R.W. (1996) Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci USA* **93,** 10614-10619.

[145] Hoheisel,J.D. (2006) Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet* **7,** 200-210.

[146] Fodor,S.P.A., Read,J.L., Pirrung,M.C., Stryer,L., Lu,A.T., & Solas,D. (1991) Light-Directed, Spatially Addressable Parallel Chemical Synthesis. *Science* **251,** 767-773.

[147] Lockhart,D.J., Dong,H.L., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Wang,C.W., Kobayashi,M., Horton,H., & Brown,E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14,** 1675-1680.

[148] Lander,E.S. (1999) Array of hope. *Nat Gen* **21,** 3-4.

[149] Ramsay,G. (1998) DNA chips: State-of-the-art. *Nat Biotechnol* **16,** 40-44.

[150] Lemmo,A.V., Rose,D.J., & Tisone,T.C. (1998) Inkjet dispensing technology: applications in drug discovery. *Curr Opin Biotechnol* **9,** 615-617.

[151] Okamoto,T., Suzuki,T., & Yamamoto,N. (2000) Microarray fabrication with covalent attachment of DNA using Bubble Jet technology. *Nat Biotechnol* **18,** 438-441.

[152] Hegde,P., Qi,R., Abernathy,K., Gay,C., Dharap,S., Gaspard,R., Hughes,J.E., Snesrud,E., Lee,N., & Quackenbush,J. (2000) A concise guide to cDNA microarray analysis. *Biotechniques* **29,** 548-562.

[153] Sebat,J.L., Colwell,F.S., & Crawford,R.L. (2003) Metagenomic profiling: Microarray analysis of an environmental genomic library. *Appl Environ Microbiol* **69,** 4927-4934.

[154] Zhou,J.H. (2003) Microarrays for bacterial detection and microbial community analysis. *Curr Opin Microbiol* **6,** 288-294.

[155] Wu,L.Y., Liu,X., Schadt,C.W., & Zhou,J.Z. (2006) Microarray-based analysis of subnanogram quantities of microbial community DNAs by using whole-community genome amplification. *Appl Environ Microbiol* **72,** 4931-4941.

[156] Gentry,T.J., Wickham,G.S., Schadt,C.W., He,Z., & Zhou,J. (2006) Microarray applications in microbial ecology research. *Microb Ecol* **52,** 159-175.

[157] Urakawa,H., El Fantroussi,S., Smidt,H., Smoot,J.C., Tribou,E.H., Kelly,J.J., Noble,P.A., & Stahl,D.A. (2003) Optimization of single-base-pair mismatch discrimination in oligonucleotide microarrays. *Appl Environ Microbiol* **69,** 2848-2856.

[158] Rhee,S.K., Liu,X.D., Wu,L.Y., Chong,S.C., Wan,X.F., & Zhou,J.Z. (2004) Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays. *Appl Environ Microbiol* **70,** 4303-4317.

[159] Wu,L.Y., Thompson,D.K., Liu,X.D., Fields,M.W., Bagwell,C.E., Tiedje,J.M., & Zhou,J.Z. (2004) Development and evaluation of microarray-based whole-genome hybridization for detection of microorganisms within the context of environmental applications. *Environ Sci Technol* **38,** 6775-6782.

[160] Park,S.J., Kang,C.H., Chae,J.C., & Rhee,S.K. (2008) Metagenome microarray for screening of fosmid clones containing specific genes. *FEMS Microbiol Lett* **284,** 28-34.

[161] Behr,M.A., Wilson,M.A., Gill,W.P., Salamon,H., Schoolnik,G.K., Rane,S., & Small,P.M. (1999) Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* **284,** 1520-1523.

[162] Barnett,M.J., Tolman,C.J., Fisher,R.F., & Long,S.R. (2004) A dual-genome Symbiosis Chip for coordinate study of signal exchange and development in a prokaryote-host interaction. *Proc Natl Acad Sci USA* **101,** 16636-16641.

[163] Losi,A., Polverini,E., Quest,B., & Gartner,W. (2002) First evidence for phototropin-related blue-light receptors in prokaryotes. *Biophys J* **82,** 2627-2634.

[164] Krauss,U., Losi,A., Gartner,W., Jaeger,K.E., & Eggert,T. (2005) Initial characterization of a blue-light sensing, phototropin-related protein from *Pseudomonas putida*: a paradigm for an extended LOV construct. *Phys Chem Chem Phys* **7,** 2804-2811.

[165] Briggs,W.R. (2007) The LOV domain: a chromophore module servicing multiple photoreceptors. *J Biomed Sci* **14,** 499-504.

[166] Swartz,T.E., Tseng,T.S., Frederickson,M.A., Paris,G., Comerci,D.J., Rajashekara,G., Kim,J.G., Mudgett,M.B., Splitter,G.A., Ugalde,R.A., Goldbaum,F.A., Briggs,W.R., & Bogomolni,R.A. (2007) Blue-light-activated histidine kinases: Two-component sensors in bacteria. *Science* **317,** 1090-1093.

[167] Purcell,E.B., Siegal-Gaskins,D., Rawling,D.C., Fiebig,A., & Crosson,S. (2007) A photosensory two-component system regulates bacterial cell attachment. *Proc Natl Acad Sci USA* **104,** 18241-18246.

[168] Cao,Z., Buttani,V., Losi,A., & Gartner,W. (2008) A blue light inducible two-component signal transduction system in the plant pathogen *Pseudomonas syringae pv. tomato*. *Biophys J* **94,** 897-905.

[169] Bullock,W.O., Fernandez,J.M., & Short,J.M. (1987) Xl1-Blue - A High-Efficiency Plasmid Transforming Reca *Escherichia coli* Strain with Beta-Galactosidase Selection. *Biotechniques* **5,** 376-379.

[170] Nelson,K.E., Weinel,C., Paulsen,I.T., Dodson,R.J., Hilbert,H., Martins dos Santos,V.A.P., Fouts,D.E., Gill,S.R., Pop,M., Holmes,M., Brinkac,L., Beanan,M., Deboy,R.T., Daugherty,S., Kolonay,J., Madupu,R., Nelson,W., White,O., Peterson,J., Khouri,H., Hance,I., Lee,P.C., Holtzapple,E., Scanlan,D., Tran,K., Moazzez,A., Utterback,T., Rizzo,M., Lee,K., Kosack,D., Moestl,D., Wedler,H., Lauber,J., Stjepandic,D., Hoheisel,J., Straetz,M., Heim,S., Kiewitz,C., Eisen,J.A., Timmis,K.N., Dusterhoft,A., Tummler,B., & Fraser,C.M. (2003) Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. *Environ Microbiol* **4,** 799-808.

[171] Buu,A., Menichi,B., & Heyman,T. (1981) Thiomethylation of Tyrosine Transfer Ribonucleic-Acid Is Associated with Initiation of Sporulation in *Bacillus subtilis* - Effect of Phosphate Concentration. *J Bacteriol* **146,** 819-822.

[172] Takeshita,S., Sato,M., Toba,M., Masahashi,W., & Hashimotogotoh,T. (1987) High-Copy-Number and Low-Copy-Number Plasmid Vectors for Lacz-Alpha-Complementation and Chloramphenicol-Resistance Or Kanamycin-Resistance Selection. *Gene* **61,** 63-74.

[173] Quest,B. Biochemische und spektroskopische charakterisierung zweier cyanobakterieller phytochrome aus *Calothrix* PCC 7601 (2003) Dissertation; Ludwig-Maximilians University of Munich.

[174] Cao,Z. Structural and functional characterization of bacterial LOV-domain-containing blue-light photoreceptors (2010) Dissertation; University of Düsseldorf.

[175] Birnboim,H.C. & Doly,J. (1979) Rapid Alkaline Extraction Procedure for Screening Recombinant Plasmid Dna. *Nucleic Acids Res* **7,** 1513-1523.

[176] Higa,A. & Mandel,M. (1970) Actinomycin Sensitive Mutants of Escherichia-Coli K-12. *Mol Gen Genet* **108,** 41-46.

[177] Cohen,S.N., Chang,A.C.Y., & Hsu,L. (1972) Nonchromosomal Antibiotic Resistance in Bacteria - Genetic Transformation of Escherichia-Coli by R-Factor Dna. *Proc Natl Acad Sci USA* **69,** 2110-2114.

[178] Laemmli,U.K. (1970) Cleavage of Structural Proteins During Assembly of Head of Bacteriophage-T4. *Nature* **227,** 680-685.

[179] Markowitz,V.M., Ivanova,N.N., Szeto,E., Palaniappan,K., Chu,K., Dalevi,D., Chen,I.M.A., Grechkin,Y., Dubchak,I., Anderson,I., Lykidis,A., Mavromatis,K., Hugenholtz,P., & Kyrpides,N.C. (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res* **36,** D534-D538.

[180] Seshadri,R., Kravitz,S.A., Smarr,L., Gilna,P., & Frazier,M. (2007) CAMERA: A community resource for metagenomics. *PloS Biol* **5,** 394-397.

[181] Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R., Thompson,J.D., Gibson,T.J., & Higgins,D.G. (2007) Clustal W and clustal X version 2.0. *Bioinformatics* **23,** 2947-2948.

[182] Hall,T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* **41,** 95-98.

[183] de Castro,E., Sigrist,C.J.A., Gattiker,A., Bulliard,V., Langendijk-Genevaux,P.S., Gasteiger,E., Bairoch,A., & Hulo,N. (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res* **34,** 362-365.

[184] Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J., & Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* **34,** 257-260.

[185] Guindon,S. & Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52,** 696-704.

[186] Jones,D.T., Taylor,W.R., & Thornton,J.M. (1992) The Rapid Generation of Mutation Data Matrices from Protein Sequences. *Comput Appl Biosci* **8,** 275-282.

[187] Letunic,I. & Bork,P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23,** 127-128.

[188] Rozen,S. & Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**.

[189] Rose,T.M., Schultz,E.R., Henikoff,J.G., Pietrokovski,S., McCallum,C.M., & Henikoff,S. (1998) Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic Acids Res* **26,** 1628-1635.

[190] Henikoff,S., Henikoff,J.G., Alford,W.J., & Pietrokovski,S. (1995) Automated Construction and Graphical Presentation of Protein Blocks from Unaligned Sequences. *Gene* **163,** 17-26.

[191] Dehal,P.S., Joachimiak,M.P., Price,M.N., Bates,J.T., Baumohl,J.K., Chivian,D., Friedland,G.D., Huang,K.H., Keller,K., Novichkov,P.S., Dubchak,I.L., Alm,E.J., & Arkin,A.P. (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res* **38,** 396-400.

[192] Markham,N.R. & Zuker,M. (2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res* **33,** 577-581.

[193] Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292,** 195-202.

[194] Guex,N. & Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* **18,** 2714-2723.

[195] Kiefer,F., Arnold,K., Kunzli,M., Bordoli,L., & Schwede,T. (2009) The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res* **37,** 387-392.

[196] Lee,D.G., Urbach,J.M., Wu,G., Liberati,N.T., Feinbaum,R.L., Miyata,S., Diggins,L.T., He,J.X., Saucier,M., Deziel,E., Friedman,L., Li,L., Grills,G., Montgomery,K., Kucherlapati,R., Rahme,L.G., & Ausubel,F.M. (2006) Genomic analysis reveals that *Pseudomonas aeruginosa* virulence is combinatorial. *Genome Biol* **7,** R90.

[197] Pathak,G. Phenotypic Characterisation of Diesel Biofilm Isolates and Isolation of Novel Biocatalysts using Metagenomic Approaches (2005) Dissertation; University of Duisburg-Essen.

[198] Schmeisser,C., Stockigt,C., Raasch,C., Wingender,J., Timmis,K.N., Wenderoth,D.F., Flemming,H.C., Liesegang,H., Schmitz,R.A., Jaeger,K.E., & Streit,W.R. (2003) Metagenome survey of biofilms in drinking-water networks. *Appl Environ Microbiol* **69,** 7298-7309.

[199] Dib,J., Motok,J., Zenoff,V.F., Ordonez,O., & Farias,M.E. (2008) Occurrence of resistance to antibiotics, UV-B, and arsenic in bacteria isolated from extreme environments in high-altitude (Above 4400 m) andean wetlands. *Curr Microbiol* **56,** 510-517.

[200] Simon,C. & Daniel,R. (2009) Achievements and new knowledge unraveled by metagenomic approaches. *Appl Microbiol Biotechnol* **85,** 265-276.

[201] Uchiyama,T., Abe,T., Ikemura,T., & Watanabe,K. (2005) Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nat Biotech* **23,** 88-93.

[202] Uchiyama,T. & Watanabe,K. (2007) The SIGEX scheme: high throughput screening of environmental metagenomes for the isolation of novel catabolic genes. *Biotechnol Genet Eng Rev* **24,** 107-116.

[203] Uchiyama,T. & Watanabe,K. (2008) Substrate-induced gene expression (SIGEX) screening of metagenome libraries. *Nat Protoc* **3,** 1202-1212.

[204] He,Z.L., Gentry,T.J., Schadt,C.W., Wu,L.Y., Liebich,J., Chong,S.C., Huang,Z.J., Wu,W.M., Gu,B.H., Jardine,P., Criddle,C., & Zhou,J. (2007) GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *ISME J* **1,** 67-77.

[205] Wang,F.P., Zhou,H.Y., Meng,J., Peng,X.T., Jiang,L.J., Sun,P., Zhang,C.L., Van Nostrand,J.D., Deng,Y., He,Z.L., Wu,L.Y., Zhou,J.H., & Xiao,X. (2009) GeoChip-based analysis of metabolic diversity of microbial communities at the Juan de Fuca Ridge hydrothermal vent. *Proc Natl Acad Sci USA* **106,** 4840-4845.

[206] Andersen,G.L., He,Z., DeSantis,T.Z., Brodie,E.L., & Zhou,J. (2010) The Use of Microarrays in Microbial Ecology. In Environ Mol Microbiol pp. 87-109. Caister Academic Press, UK.

[207] Van Nostrand,J.D., Wu,W.M., Wu,L.Y., Deng,Y., Carley,J., Carroll,S., He,Z.L., Gu,B.H., Luo,J., Criddle,C.S., Watson,D.B., Jardine,P.M., Marsh,T.L., Tiedje,J.M., Hazen,T.C., & Zhou,J.Z. (2009) GeoChip-based analysis of functional microbial communities during the reoxidation of a bioreduced uranium-contaminated aquifer. *Environ Microbiol* **11,** 2611-2626.

[208] Guschin,D., Yershov,G., Zaslavsky,A., Gemmell,A., Shick,V., Proudnikov,D., Arenkov,P., & Mirzabekov,A. (1997) Manual manufacturing of oligonucleotide, DNA, and protein microchips. *Anal Biochem* **250,** 203-211.

[209] Lessard,I.A.D., Domingo,G.J., Borges,A., & Perham,R.N. (1998) Expression of genes encoding the E2 and E3 components of the Bacillus stearothermophilus pyruvate dehydrogenase complex and the stoichiometry of subunit interaction in assembly in vitro. *Eur J Biochem* **258,** 491-501.

[210] Hughes,T.R., Mao,M., Jones,A.R., Burchard,J., Marton,M.J., Shannon,K.W., Lefkowitz,S.M., Ziman,M., Schelter,J.M., Meyer,M.R., Kobayashi,S., Davis,C., Dai,H.Y., He,Y.D.D., Stephaniants,S.B., Cavet,G., Walker,W.L., West,A., Coffey,E., Shoemaker,D.D., Stoughton,R., Blanchard,A.P., Friend,S.H., & Linsley,P.S. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* **19,** 342-347.

[211] Tiquia,S.M., Wu,L.Y., Chong,S.C., Passovets,S., Xu,D., Xu,Y., & Zhou,J.Z. (2004) Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples. *Biotechniques* **36,** 664-675.

[212] Rich,V.I., Konstantinidis,K., & Delong,E.F. (2008) Design and testing of 'genome-proxy' microarrays to profile marine microbial communities. *Environ Microbiol* **10,** 506-521.

[213] Kane,M.D., Jatkoe,T.A., Stumpf,C.R., Lu,J., Thomas,J.D., & Madore,S.J. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res* **28,** 4552-4557.

[214] Relogio,A., Schwager,C., Richter,A., Ansorge,W., & Valcarcel,J. (2002) Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Res* **30**.

[215] Xu,D., Li,G., Wu,L., Zhou,J., & Xu,Y. (2002) PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis. *Bioinformatics* **18,** 1432-1437.

[216] Pozhitkov,A., Noble,P.A., Domazet-Loso,T., Nolte,A.W., Sonnenberg,R., Staehler,P., Beier,M., & Tautz,D. (2006) Tests of rRNA hybridization to microarrays suggest that hybridization characteristics of oligonucleotide probes for species discrimination cannot be predicted. *Nucleic Acids Res* **34,** e66.

[217] Zhou,J.Z. & Thompson,D.K. (2002) Challenges in applying microarrays to environmental studies. *Curr Opin Biotechnol* **13,** 204-207.

[218] Peplies,J., Glockner,F.O., & Amann,R. (2003) Optimization strategies for DNA microarray-based detection of bacteria with 16S rRNA-targeting oligonucleotide probes. *Appl Environ Microbiol* **69,** 1397-1407.

[219] Yergeau,E., Kang,S., He,Z., Zhou,J., & Kowalchuk,G.A. (2007) Functional microarray analysis of nitrogen and carbon cycling genes across an Antarctic latitudinal transect. *ISME J* **1,** 163-179.

[220] He,Z.L., Wu,L.Y., Li,X.Y., Fields,M.W., & Zhou,J.Z. (2005) Empirical establishment of oligonucleotide probe design criteria. *Appl Environ Microbiol* **71,** 3753-3760.

[221] Wu,L.Y., Thompson,D.K., Li,G.S., Hurt,R.A., Tiedje,J.M., & Zhou,J.Z. (2001) Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl Environ Microbiol* **67,** 5780-5790.

[222] Bartosiewicz,M., Trounstine,M., Barker,D., Johnston,R., & Buckpitt,A. (2000) Development of a toxicological gene array and quantitative assessment of this technology. *Arch Biochem Biophys* **376,** 66-73.

[223] Yang,H.N., Harrington,C.A., Vartanian,K., Coldren,C.D., Hall,R., & Churchill,G.A. (2008) Randomization in Laboratory Procedure Is Key to Obtaining Reproducible Microarray Results. *Plos One* **3,** e3724.

[224] Ward,B.B., Eveillard,D., Kirshtein,J.D., Nelson,J.D., Voytek,M.A., & Jackson,G.A. (2007) Ammonia-oxidizing bacterial community composition in estuarine and oceanic environments assessed using a functional gene microarray. *Environ Microbiol* **9,** 2522-2538.

[225] Tan,P.K., Downey,T.J., Spitznagel,E.L., Xu,P., Fu,D., Dimitrov,D.S., Lempicki,R.A., Raaka,B.M., & Cam,M.C. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* **31,** 5676-5684.

[226] Ioannidis,J.P.A., Allison,D.B., Ball,C.A., Coulibaly,I., Cui,X.Q., Culhane,A.C., Falchi,M., Furlanello,C., Game,L., Jurman,G., Mangion,J., Mehta,T., Nitzberg,M., Page,G.P., Petretto,E., & van Noort,V. (2009) Repeatability of published microarray gene expression analyses. *Nat Genet* **41,** 149-155.

[227] Walker,M.S. & Hughes,T.A. (2008) Messenger RNA expression profiling using DNA microarray technology: Diagnostic tool, scientific analysis or un-interpretable data? *Int J Mol Med* **21,** 13-17.

[228] Cho,R.J., Huang,M., Campbell,M.J., Dong,H., Steinmetz,L., Sapinoso,L., Hampton,G., Elledge,S.J., Davis,R.W., & Lockhart,D.J. (2001) Transcriptional regulation and function during the human cell cycle. *Nat Genet* **27,** 48-54.

[229] Shedden,K. & Cooper,S. (2002) Analysis of cell-cycle-specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization. *Proc Natl Acad Sci USA* **99,** 4379-4384.

[230] Gebert,J., Stralis-Pavese,N., Alawi,M., & Bodrossy,L. (2008) Analysis of methanotrophic communities in landfill biofilters using diagnostic microarray. *Environ Microbiol* **10,** 1175-1188.

[231] Marchetti,A., Parker,M.S., Moccia,L.P., Lin,E.O., Arrieta,A.L., Ribalet,F., Murphy,M.E.P., Maldonado,M.T., & Armbrust,E.V. (2009) Ferritin is used for iron storage in bloom-forming marine pennate diatoms. *Nature* **457,** 467-470.

[232] Pritchard,L.l.a.u., Liu,H., Booth,C., Douglas,E., Francois,P., Schrenzel,J., Hedley,P.E., Birch,P.R., & Toth,I.K. (2009) Microarray Comparative Genomic Hybridisation Analysis Incorporating Genomic Organisation, and Application to Enterobacterial Plant Pathogens. *PLoS Comput Biol* **5**.

[233] Ogata,H., Cao,Z., Losi,A., & Gartner,W. (2009) Crystallization and preliminary X-ray analysis of the LOV domain of the blue-light receptor YtvA from *Bacillus amyloliquefaciens* FZB42. *Acta Crystallogr, Sect F: Struct Biol Cryst Commun* **65,** 853-855.

[234] Losi,A., Quest,B., & Gartner,W. (2003) Listening to the blue: the time-resolved thermodynamics of the bacterial blue-light receptor YtvA and its isolated LOV domain. *Photochem Photobiol Sci* **2,** 759-766.

[235] Alexandre,M.T.A., Arents,J.C., van Grondelle,R., Hellingwerf,K.J., & Kennis,J.T.M. (2007) A base-catalyzed mechanism for dark state recovery in the Avena sativa phototropin-1 LOV2 domain. *Biochemistry* **46,** 3129-3137.

[236] Wang,W.W., Sineshchekov,O.A., Spudich,E.N., & Spudich,J.L. (2003) Spectroscopic and photochemical characterization of a deep ocean proteorhodopsin. *J Biol Chem* **278,** 33985-33991.

[237] Man,D.L., Wang,W.W., Sabehi,G., Aravind,L., Post,A.F., Massana,R., Spudich,E.N., Spudich,J.L., & Beja,O. (2003) Diversification and spectral tuning in marine proteorhodopsins. *EMBO J* **22,** 1725-1731.

[238] Sabehi,G., Kirkup,B.C., Rozenberg,M., Stambler,N., Polz,M.F., & Beja,O. (2007) Adaptation and spectral tuning in divergent marine proteorhodopsins from the eastern Mediterranean and the Sargasso Seas. *ISME J* **1,** 48-55.

[239] Man-Aharonovich,D., Sabehi,G., Sineshchekov,O.A., Spudich,E.N., Spudich,J.L., & Beja,O. (2004) Characterization of RS29, a blue-green proteorhodopsin variant from the Red Sea. *Photochem Photobiol Sci* **3,** 459-462.

[240] Losi,A., Ghiraldelli,E., Jansen,S., & Gartner,W. (2005) Mutational effects on protein structural changes and interdomain interactions in the blue-light sensing LOV protein YtvA. *Photochem Photobiol* **81,** 1145-1152.

[241] Hendrischk,A.K., Moldt,J., Fruhwirth,S.W., & Klug,G. (2009) Characterization of an Unusual LOV Domain Protein in the alpha-Proteobacterium Rhodobacter sphaeroides. *Photochem Photobiol* **85,** 1254-1259.

[242] Stock,A.M., Robinson,V.L., & Goudreau,P.N. (2000) Two-component signal transduction. *Annu Rev Biochem* **69,** 183-215.

[243] Hristova,K.R., Schmidt,R., Chakicherla,A.Y., Legler,T.C., Wu,J., Chain,P.S., Scow,K.M., & Kane,S.R. (2007) Comparative transcriptome analysis of Methylibium petroleiphilum PM1 exposed to the fuel oxygenates methyl tert-butyl ether and ethanol. *Appl Environ Microbiol* **73,** 7347-7357.

[244] Schmidt,R., Battaglia,V., Scow,K., Kane,S., & Hristova,K.R. (2008) Involvement of a Novel Enzyme, MdpA, in Methyl tert-Butyl Ether Degradation in *Methylibium petroleiphilum* PM1. *Appl Environ Microbiol* **74,** 6631-6638.

[245] Price,M.N., Huang,K.H., Alm,E.J., & Arkin,A.P. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res* **33,** 880-892.

[246] Ferreira,G.C. (1999) Ferrochelatase. *Int J Biochem Cell Biol* **31,** 995-1000.

[247] Dailey,H.A., Dailey,T.A., Wu,C.K., Medlock,A.E., Wang,K.F., Rose,J.P., & Wang,B.C. (2000) Ferrochelatase at the millennium: structures, mechanisms and [2Fe-2S] clusters. *Cell Mol Life Sci* **57,** 1909-1926.

[248] Lange,H., Muhlenhoff,U., Denzel,M., Kispal,G., & Lill,R. (2004) The Heme Synthesis Defect of Mutants Impaired in Mitochondrial Iron-Sulfur Protein Biogenesis Is Caused by Reversible Inhibition of Ferrochelatase. *J Biol Chem* **279,** 29101-29108.

[249] Schurmann,P. & Buchanan,B.B. (2008) The ferredoxin/thioredoxin system of oxygenic photosynthesis. *Antioxid Redox Signaling* **10,** 1235-1273.

[250] Franzmann,T.M. (2006) Matrix-assisted refolding of oligomeric small heat-shock protein Hsp26. *Int J Biol Macromol* **39,** 104-110.

[251] Lyukmanova,E.N., Shulepko,M.A., Tikhonov,R.V., Shenkarev,Z.O., Paramonov,A.S., Wulfson,A.N., Kasheverov,I.E., Ustich,T.L., Utkin,Y.N., Arseniev,A.S., Tsetlin,V.I., Dolgikh,D.A., & Kirpichnikov,M.P. (2009) Bacterial production and refolding from inclusion bodies of a "Weak" toxin, a disulfide rich protein. *Biochemistry (Moscow)* **74,** 1142-1149.

[252] Goudarzi,G., Sattari,M.S.a.i., Roudkenar,M.H., Montajabi-Niyat,M., Zavaran-Hosseini,A., & Mosavi-Hosseini,K. (2009) Cloning, expression, purification, and characterization of recombinant flagellin isolated from *Pseudomonas aeruginosa*. *Biotechnol Lett* **31,** 1353-1360.

[253] Salomon,M., Eisenreich,W., Durr,H., Schleicher,E., Knieb,E., Massey,V., Rudiger,W., Muller,F., Bacher,A., & Richter,G. (2001) An optomechanical transducer in the blue light receptor phototropin from Avena sativa. *Proc Natl Acad Sci USA* **98,** 12357-12361.

[254] Durr,H., Salomon,M., & Rudiger,W. (2005) Chromophore exchange in the LOV2 domain of the plant photoreceptor phototropin1 from oat. *Biochemistry* **44,** 3050-3055.

[255] de Groot,N.S., Sabate,R., & Ventura,S. (2009) Amyloids in bacterial inclusion bodies. *Trends Biochem Sci* **34,** 408-416.

[256] Kopito,R.R. (2000) Aggresomes, inclusion bodies and protein aggregation. *Trends Cell Biol* **10,** 524-530.

[257] Villaverde,A. & Carrio,M.M. (2003) Protein aggregation in recombinant bacteria: biological role of inclusion bodies. *Biotechnol Lett* **25,** 1385-1395.

[258] Zhang,Y.X., Zhu,Y., Xi,H.W., Liu,Y.L., & Zhou,H.M. (2002) Refolding and reactivation of calf intestinal alkaline phosphatase with excess magnesium ions. *Int J Biochem Cell Biol* **34,** 1241-1247.

[259] Vallejo,L. & Rinas,U. (2004) Strategies for the recovery of active proteins through refolding of bacterial inclusion body proteins. *Microb Cell Fact* **3,** 11.

[260] Chen,J.Q., Acton,T.B., Basu,S.K., Montelione,G.T., & Inouye,M. (2002) Enhancement of the solubility of proteins overexpressed in *Escherichia coli* by heat shock. *J Mol Microbiol Biotechnol* **4,** 519-524.

[261] Oganesyan,N., Ankoudinova,I., Kim,S.H., & Kim,R. (2007) Effect of osmotic stress and heat shock in recombinant protein overexpression and crystallization. *Protein Expression Purif* **52,** 280-285.

[262] van Bogelen,R., Kelley,P., & Neidhardt,F. (1987) Differential induction of heat shock, SOS and oxidation stress regulons and accumulation of nucleotides in Escherichia coli. *J Bacteriol* **169,** 26-32.

[263] Donnelly,M.I., Zhou,M., Millard,C.S., Clancy,S., Stols,L., Eschenfeldt,W.H., Collart,F.R., & Joachimiak,A. (2006) An expression vector tailored for large-scale, high-throughput purification of recombinant proteins. *Protein Expression Purif* **47,** 446-454.

[264] Kannan,N., Taylor,S.S., Zhai,Y.F., Venter,J.C., & Manning,G. (2007) Structural and functional diversity of the microbial kinome. *PloS Biol* **5,** 467-478.

[265] Foerstner,K.U., Doerks,T., Creevey,C.J., Doerks,A., & Bork,P. (2008) A Computational Screen for Type I Polyketide Synthases in Metagenomics Shotgun Data. *Plos One* **3**.

[266] Pathak,G.P., Ehrenreich,A., Losi,A., Streit,W.R., & Gartner,W. (2009) Novel blue light-sensitive proteins from a metagenomic approach. *Environ Microbiol* **11,** 2388-2399.

[267] Krauss,U., Minh,B.Q., Losi,A., Gaertner,W., Eggert,T., von Haeseler,A.a.v.h., & Jaeger,K.E. (2009) Distribution and Phylogeny of Light-Oxygen-Voltage-Blue-Light-Signaling Proteins in the Three Kingdoms of Life. *J Bacteriol* **191,** 7234-7242.

[268] Kumauchi,M., Hara,M.T., Stalcup,P., Xie,A.H., & Hoff,W.D. (2008) Identification of six new photoactive yellow proteins - Diversity and structure-function relationships in a bacterial blue light photoreceptor. *Photochem Photobiol* **84,** 956-969.

[269] Smith,C.R. & Baco,A.R. (2003) *Ecology of whale falls at the deep-sea floor*. TAYLOR & FRANCIS LTD, LONDON.

[270] Streifel,A. (2004) Design and maintenance of hospital ventilation systems and prevention of airborne nosocomial infections. (Mayhall C, ed), pp. 1577-1589. Philadelphia, PA: Lippincott Williams & Wilkins.

[271] Edwards,K.J., Bond,P.L., Druschel,G.K., McGuire,M.M., Hamers,R.J., & Banfield,J.F. (2000) Geochemical and biological aspects of sulfide mineral dissolution: lessons from Iron Mountain, California. *Chem Geol* **169,** 383-397.

[272] Romling,U., Gomelsky,M., & Galperin,M.Y. (2005) C-di-GMP: the dawning of a novel bacterial signalling system. *Mol Microbiol* **57,** 629-639.

[273] Jenal,U. & Malone,J. (2006) Mechanisms of cyclic-di-GMP signaling in bacteria. *Annu Rev Genet* **40,** 385-407.

[274] Nakhamchik,A., Wilde,C., & Rowe-Magnus,D.A. (2008) Cyclic-di-GMP regulates extracellular polysaccharide production, biofilm formation, and rugose colony development by *Vibrio vulnificus*. *Appl Environ Microbiol* **74,** 4199-4209.

[275] Galperin,M.Y. (2004) Bacterial signal transduction network in a genomic perspective. *Environ Microbiol* **6,** 552-567.

[276] Mann,N.H., Cook,A., Millard,A., Bailey,S., & Clokie,M. (2003) Marine ecosystems: Bacterial photosynthesis genes in a virus. *Nature* **424,** 741.

[277] Sharon,I., Alperovitch,A., Rohwer,F., Haynes,M., Glaser,F., tamna-Ismaeel,N., Pinter,R.Y., Partensky,F., Koonin,E.V., Wolf,Y.I., Nelson,N., & Beja,O. (2009) Photosystem I gene cassettes are present in marine virus genomes. *Nature* **461,** 258-262.

[278] Lindell,D., Sullivan,M.B., Johnson,Z.I., Tolonen,A.C., Rohwer,F., & Chisholm,S.W. (2004) Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc Natl Acad Sci USA* **101,** 11013-11018.

[279] Mann,N.H., Cook,A., Millard,A., Bailey,S., & Clokie,M. (2003) Marine ecosystems: Bacterial photosynthesis genes in a virus. *Nature* **424,** 741.

Appendix

Appendix A: Oligonucleotide sequences spotted on the microarray and names/labeling of microorganisms, from which the genes originated.

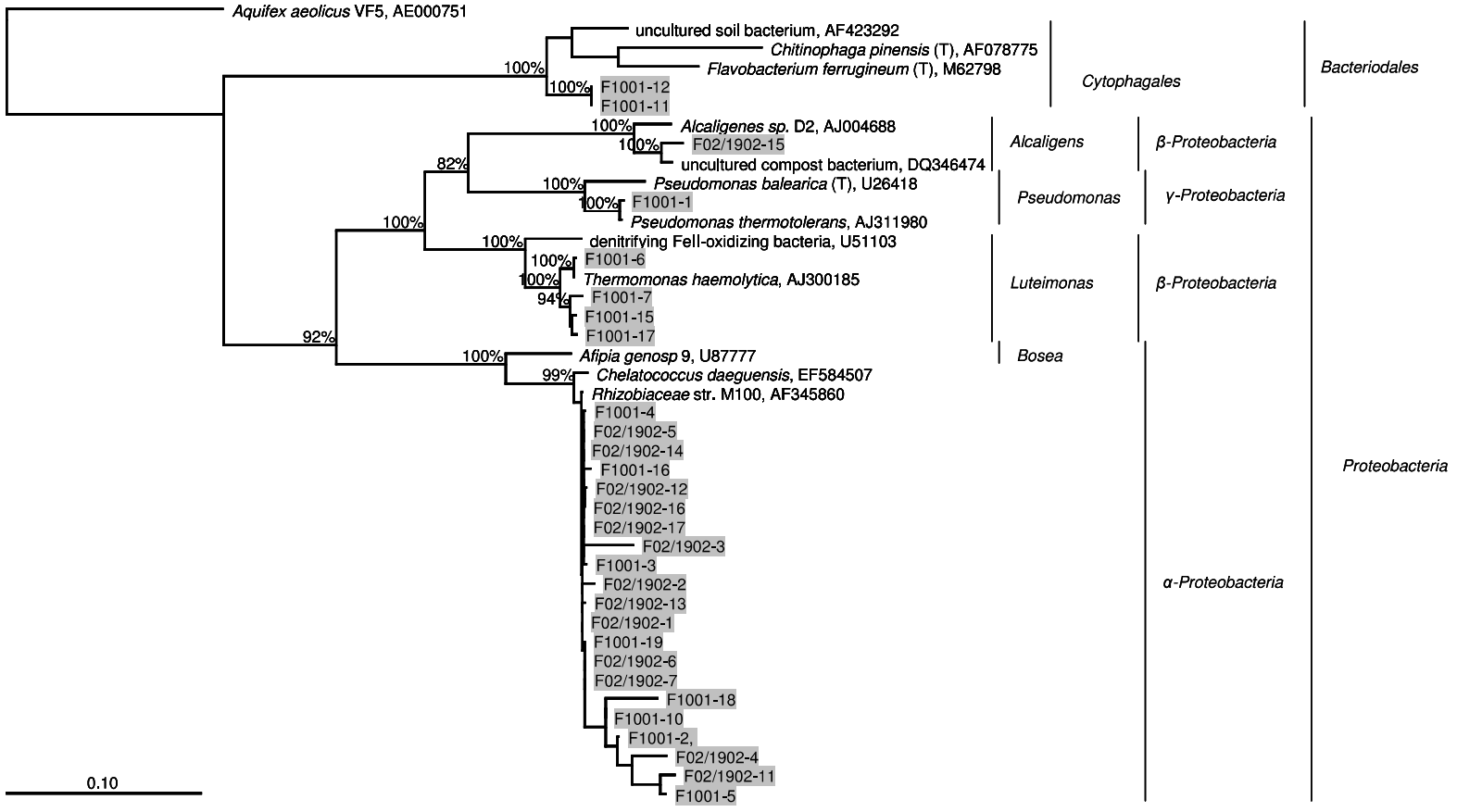
S.N.	Probe ID	Organism/Source	Oligonucleotide sequence
1	5001	*Thiomicrospira denitrificans* ATCC 33889	tttagatataaatatgaagaggtaatcggtaaaaactgtcgtttttacatagt
2	5002	*Nitrosococcus oceani* ATCC 19707	accggatataaaaaagaagaaattcttggcaaaaactgtcgtattttacagggt
3	5003	*Arabidopsis thaliana* LOV2	acggaatatagccgtgaagaaattcttggaagaaattgcaggtttctacaaggt
4	5004	*Crocosphaera watsonii* WH 8501	acaggatattccttgcaggaagtaactggaaaaaactgtcgtttttacaagga
5	5005	*Oceanobacillus iheyensis* HTE831	actggttatgaagaacacgaaataattgggaagaactgtagattcttacagggc
6	5006	Sargasso Metagenome	actggatacagctcgtcagatataattggtgaaacacctagaatttacaaggc
7	5007	Sargasso Metagenome	actggctattcaaaaaaatttgcgactggtaaaaagcctgcttttctacaagga
8	5008	Sargasso Metagenome	actggttactcaagggatgaaatcataggtaaaaactgtcgttttattgaagcc
9	5009	*Oceanobacter* sp. RED65	actggatattctgaatctgaagttatcgggaagaattgtcgattccttcaaggt
10	5010	*Listeria innocua*	actggctatactaaagaagaagctattggctctaattgtcacttttgcaagga
11	5011	*Synechococcus elongatus* PCC 7942	acgggttattctgaagctgaaatgctaggcagaaactgcagaatattacaaggc
12	5012	Sargasso Metagenome	accgggtgggagattggcgaagttaaaggtaagaaacctggatcatttttacaa
13	5013	*Listeria monocytogenes* str. 1/2a F6854	accggctatgctaaagaagaagcacttggctccaattgtcacttttacaagga
14	5014	*Listeria monocytogenes str. 4b H7858*	accggctatgctaaagaagaagcacttggctccaattgtcacttttacaagga
15	5015	*Synechocystis* sp. PCC 6803	accggttttaatcgggaggaagtgattgggcaaaactgtcgttttctactcaat
16	5016	*Arabidopsis thaliana* LOV1	actggttacacttccaaagaagtcgtcggcagaaactgccgatttttacaagga
17	5017	*Anabaena variabilis* ATCC 29413	acaggttacaaagctacagaaatcgtcggacgcaactgtcgtttttgcaaggt
18	5018	*Anabaena variabilis* ATCC 29413	actggctactcggctggtgaagttattggacaaaactgccgatttttacaagct
19	5019	*Nostoc* sp. PCC 7120	actggctactctgccggtgaagttattgggcaaaactgccgatttttacaaact
20	5020	*Nitrosospira multiformis* ATCC 25196	acgggatatacccaggaagaaactatcggaaagaattgccggtttttgcaaggt
21	5021	Sargasso Metagenome	actggctatgcgcctgaggaagtatttggtcaaaattgtcgattccttcaaggg
22	5022	Sargasso Metagenome	actggctattcgcaagctgagaccataggacgtaactgtcgatttctacagtca
23	5023	Sargasso Metagenome	acaggttactcgcgggaagagttacttgggcaaaatccacgcttatttaattcc
24	5024	*Nostoc* sp. PCC 7120	acaggttataaagcgacagaagttgtcggacgcaactgtcgtttttgctaggt
25	5025	*Nostoc punctiforme* PCC 73102	acaggctacactgctgctgatgtgattgggcaaaactgtcgatttttgcaacgc
26	5026	*Nostoc punctiforme* PCC 73102	acagggtactgtgccagtgaagtcattgggcgtaactgtcgtttttgcaagga
27	5027	Sargasso Metagenome	gggtgggagctggacgaagtaaaaggcaagaaacctggagcatttttacaaggt
28	5028	*P. syringae pv. phaseolicola* 1448A	acgggctattcgtcagaagaaatcatcggtacaaactgccggtttcttcagggg
29	5029	*Bacillus subtilis*	accggctacgagaccgaggaaattttaggaaagaactgtcgcttcttacagggg
30	5030	*Alteromonas macleodii* 'Deep ecotype'	actggctatacccgcgaagaaatcatcgggcataattgccgttctatgcaggga
31	5031	Sargasso Metagenome	acaggctacagtcgtgaagaaataatcggtcataactgcaggtcgatgcagggc
32	5032	Sargasso Metagenome	accggctactcaagggacgagattatcggtaaaaattgccgtttcctcgaggct
33	5033	Sargasso Metagenome	accggctactcaagggatgagattatcggtaaaaattgccgtttcctcgaggcc
34	5034	*Pseudomonas syringae* pv. *syringae* B728a	accggctatgcgtcagaggaaatcatcggtagcaactgccgatttcttcaaggg
35	5035	*Rhodopirellula baltica* SH 1	actgggttctccgagcaagagatcctcggtcgcaattgtcgattcttgcaagga
36	5036	Sargasso Metagenome	acggggtacagcgtgcaagaagccgtgggtcacaactgtaggtttttgcaagac
37	5037	*Parvularcula bermudensis* HTCC2503	acgggctattccctcagttatgttctcggtcggaattgccgctttttgcagcgg
38	5038	*Pseudomonas syringae* pv. *tomato* str.DC3000	acggggtatacggccgaagaaatccttggcaccaattgccggtttcttcagggg
39	5039	*Pseudomonas putida* F1	actggctacagccgtgacgagattctctatcaggattgccggttcctgcaggct

40	5040	*Pseudomonas putida* KT2440	accggctactgcgccgacgatattctctatcaggactgccgttttcttcagggc
41	5041	*Pseudomonas putida* F1	accggctattgcgccgacgatatcctctatcaggactgccgttttcttcagggc
42	5042	*Synechococcus elongatus* PCC 6301	actggctacaacgccgccgaagcgctcggtaaaagttgtcgttttctgcaggga
43	5043	*Erythrobacter litoralis* HTCC2594	acggggtacgaggaaaaggacatcatcgggcgcaattgccgcttcctgcaaggt
44	5044	*Thermosynechococcus elongatus* BP-1	acggggtaccgggcaacggaggtcattggcaaaaattgccgcttttgcagggg
45	5045	*Pseudomonas putida* KT2440	actggctacagccgtgacgagattctctaccaggattgccggttcctgcagggt
46	5046	*Erythrobacter litoralis* HTCC2594	accggctattccgaagaagaatgcgtcggccgcaattgccgattcctggcaggt
47	5047	*Parvularcula bermudensis* HTCC2503	actggctatgcggctgaggtgtccctcgggcgtaattgccggttcttacaagga
48	5048	*Adiantum* Phy3 LOV2	acggagtataccagggaggaggtgctgggaaacaactgccgtttcctgcagggc
49	5049	*Brucella abortus* biovar 1 str. 9-941	acgggttacgaggctgacgaggttatggggcgcaattgccgctttcttcagggg
50	5050	*Polaromonas naphthalenivorans* CJ2	accggctatgagccccaggatgtgatcggaaacgattgccggtttctccagggc
51	5051	*Xanthomonas axonopodis* pv. citri str. 306	accggctatgccgccgatgaagtcatcggcaacaactgccgcttcctgcagggg
52	5052	*Xanthomonas oryzae* pv. *oryzae* KACC10331	accgggtattcggccgaggaggtcatcggcaacaactgccgctttctgcagggt
53	5053	*X. camp pv. campestris* str. ATCC 33913	accgggtacgctgccgacgaaatcatcggcaacaactgccgcttcctgcagggc
54	5054	*X. campestris pv. vesicatoria* str. 85-10	accggctatgccgccgatgaggtcatcggcaacaactgccgtttcctgcagggg
55	5055	*Bradyrhizobium* sp. BTAi1	accggctacgagctcaacgagatcgtcggcaccaattgccgtttcctgcagggg
56	5056	*Erythrobacter litoralis* HTCC2594	accggctatgcgcgcgaggagatcgtcggcaggaattgccgctttctacagggc
57	5057	*Sphingomonas* sp. SKA58	acgggctatgaggaaagcgagattgtcgggcgcaactgccggttcctgcaaggg
58	5058	*Novosphingobium aromaticivorans* DSM 12444	accggctatgacgaacacgaggtcgtcggccgcaactgccggtttctgcaaggc
59	5059	*Aurantimonas* sp. SI85-9A1	atcggctacgaccgcgacgagatcatcggtcgcaactgccggttcttgcagggg
60	5060	*Chromohalobacter salexigens* DSM 3043	accggctacagcgtcgacgagatcctttaccgtgactgccgtttcctgcagggc
61	5061	*Pseudomonas fluorescens* Pf-5	accggctacagcgccgacgaagtgctgtaccaggattgccgcttcctgcaatcc
62	5062	*Haloarcula marismortui* ATCC 43049	actggctactcgctcccggagctgctcgggaagaactgccgaatacttcagggc
63	5063	Sargasso Metagenome	accggctacgatcccgctgaagtgatcggccacgattgccgtgtgctgcaacgc
64	5064	*Erythrobacter* sp. NAP1	accggatacagcgcagagatgacggtggggcgcaattgccgcttcctgcaaggc
65	5065	*Erythrobacter litoralis* HTCC2594	accggctattcgcgatcttccgtggtcggtcgcaactgccgcttcctgcaaggc
66	5066	*Sphingomonas* sp. SKA58	acgggctataggcgcgaagaggtgataggccgcaattgccgctttatgcgcggc
67	5067	*Caulobacter crescentus* CB15	acgggctatgcccgcgacgaagtgatcggccgcaattgccggttcctgcaggga
68	5068	*Natronomonas pharaonis* DSM 2160	accggctatgctcgtgctgaggtcctcggccggaactgccgcttcctccaaggt
69	5069	Sargasso Metagenome	acggggtacgacgcgagcgaggtcctgaaccgaaactgccggttcctgcagggt
70	5070	Sargasso Metagenome	accagctacatgaagcaggaggtgctcgggcgcaactgccgcttcctgcagggg
71	5071	Sargasso Metagenome	acaggctacagccctgaggaggtgattggccgtaacccgcgtctgctgagcgcc
72	5072	*Sphingopyxis alaskensis* RB2256	accggataccggcccgacgaaatcatcggccgcaactgccgctttctgaccggc
73	5073	*Rhodobacter sphaeroides* 2.4.1	accggctataccgaagggcggatcctcgggttcaactgccggttcctccagcgc
74	5074	*Magnetospirillum magnetotacticum* MS-1	accggctacacccgcgaggaggtgatcggacgcaattgtcgcctcctgcagggc
75	5075	*Natronomonas pharaonis* DSM 2160	accggctacgagagacacgaggtgctggggcggaactgtcggttcctgcagggc
76	5076	Sargasso Metagenome	acggggtatgacaagcacgactgcgaggggcgcaactgccgcttcctgcagggc
77	5077	*Novosphingobium aromaticivorans* DSM 12444	accggctatgcgcgcgaggagatcattggccgcaactgccgcttccttcgcggg
78	5078	*Rubrobacter xylanophilus* DSM 9941	accggctactcgaggcgggaggtggtgggccgcaactgccgctttctgcagcgg
79	5079	*Magnetospirillum magnetotacticum* MS-1	accggctatacgcgcctggaggtgacgggccgcaattgccggttcctgcagggg
80	5080	*Haloarcula marismortui* ATCC 43049	accggctacacggagtctgaggtccgcggccggaactgccggttcctccagggc
81	5081	*Rubrivivax gelatinosus* PM1	agcggcctgccgctcgaccaggtgatcgggcgcaactgccgctttctgcaggcc
82	5082	*Oceanicola granulosus* HTCC2516	acgggctacgtctcgcacatggcggtggggcgcaattgccggttcctgcagggc
83	5083	*Ralstonia solanacearum*	accggctaccgcgcggaggaggtgctgggccgcaactgccgcttcctgcattcg
84	5084	*Ralstonia solanacearum* UW551	accggctaccgggcggaagaggtgctgggccgcaactgccgcttcctgcattcg
85	5085	*Aurantimonas* sp. SI85-9A1	acccgctattccgccgacgcggtgatcggccgcaattgccgcttcctgcagggc
86	5086	*Kineococcus radiotolerans* SRS30216	gtcaccggctacggccgcgaggtcctgggccagaactgccgcttcctgcagggc

87	5087	*Burkholderia fungoforum*	accggctacgacgccgccgaggcgatcggccacgattgccgcctgctacagcgc
88	5088	Sargasso Metagenome	acggggtacagcgcggcagaggcggtggggcgcaactgccgcttcctgcaaggc
89	5089	Sargasso Metagenome	accggctacacggtggacgagctggtgggccagacgccgcgcctgctgcactcg
90	5090	*Kineococcus radiotolerans* SRS30216	gccggctaccgctccgaggagctgctgggccgcaactgccgcttcctgcagggg
91	1.GS	Global Ocean Expedition/Sargasso	acagggtattcaatcgaagaatctttgggtaaaaattgtagatttttacaggga
92	3.GS	Global Ocean Expedition/Sargasso	acaggatatagtgttgaagagtcattaggtaaaaactgtagattccttcaaggc
93	4.GS	Global Ocean Expedition/Sargasso	actggttatacaattgaagaatctcttggaaaaaattgtagatttcttcaagga
94	6.GS	Global Ocean Expedition/Sargasso	acgggctatgcgcgcagctcggccatcggccgcaactgccgctttctccaaggc
95	7.GS	Global Ocean Expedition/Sargasso	accgggtactcggtcgaggaggccaccggccagaactgccggttcctgcaaggc
96	9.GS	Global Ocean Expedition/Sargasso	actgggtataccgtggatgagacggttgggaagaactgtcgtttttacaaggc
97	10.GS	Global Ocean Expedition/Sargasso	acaggttactcaattaaggaggcaataggaaaaaattgcaggtttttacaagga
98	11.GS	Global Ocean Expedition/Sargasso	actggctatagtgttaaggaatctttagggaaaaattgcagatttctccagggt
99	12.GS	Global Ocean Expedition/Sargasso	accggctatgcgcgagatgcgatcatcggtcggaactgccggttcctgcagggc
100	13.GS	Global Ocean Expedition/Sargasso	accggatactccatcgaagagaccgttggacataactgcaggttcctgcagggt
101	15.GS	Global Ocean Expedition/Sargasso	acgggttacgccagcgaagacatcctttatcaggattgtcgttttctgcaggct
102	16.GS	Global Ocean Expedition/Sargasso	accggctacagtgcagacgaaattctgtatcaggattgtcgtttcctgcagaca
103	18.GS	Global Ocean Expedition/Sargasso	acaggctattccgcggccgaggccatcggcaagaactgccggtacctgcagggc
104	19.GS	Global Ocean Expedition/Sargasso	accggatactctgctgaagaagctgttggtcgtaattgcagatttttacagggg
105	20.GS	Global Ocean Expedition/Sargasso	accggctatagcgttgaggaatcattaggtaaaaactgtagatttcttcaaggt
106	21.GS	Global Ocean Expedition/Sargasso	acgctctacgaccgcgccttcgccaccggccgcaattgccgctttctgcaaggg
107	22.GS	Global Ocean Expedition/Sargasso	acgggctacgagacccacgagctgcttggcttcaactgccgcatgctgcagcgc
108	23.GS	Global Ocean Expedition/Sargasso	acggggtacccgaaggagcacacgcagggccgcaactgccgcttcctgcagggc
109	25.GS	Global Ocean Expedition/Sargasso	accggatactccaagaaagaagcacagggcaggaactgtcgcttcctccaggga
110	26.GS	Global Ocean Expedition/Sargasso	acgggatacagccgcgaggaggtgctcgggcgcaactgccgcttcttgcacggg
111	28.GS	Global Ocean Expedition/Sargasso	gtcggttacacccgcaccgagatgctcggccgcaactgccgcttcctacagggg
112	29.GS	Global Ocean Expedition/Sargasso	acgggatacagtcgggatgaagtgctcggatacaactgccgattccttcagggc
113	30.GS	Global Ocean Expedition/Sargasso	accggctacgagagggaggaggtccttggccgcaactgccgcttcttacagggc
114	31.GS	Global Ocean Expedition/Sargasso	accggatacccgcgcgaggaggtgctggggcgcaactgccgcttcctgcagggg
115	35.GS	Global Ocean Expedition/Sargasso	acgggcttcagaaagagggaagtggtcggcctcaactgtcggctgctgcagggc
116	37.GS	Global Ocean Expedition/Sargasso	acggggtacgacaagggagacatatatgggcgcaactgtcgcttcttgcagggc
117	38.GS	Global Ocean Expedition/Sargasso	acggagtacccgcgcgaggaggccgtggggcggaactgtcgcttcttgcaaggg
118	39.GS	Global Ocean Expedition/Sargasso	acaggatatgcgaagaaagaggccatgggtcgcaactgtcgcttcctgcagggt
119	41.GS	Global Ocean Expedition/Sargasso	acagagtactgcttcgacgagtgtgttggccgcaattgcaagttcctgcagggc
120	46.GS	Global Ocean Expedition/Sargasso	acgggctacgggaagacggaggcgcagggccgcaactgccgcttcctgcagggc
121	48.GS	Global Ocean Expedition/Sargasso	actggccactcgaagcgcgacatggagggcacgaactgccgcattctgcagggg
122	49.GS	Global Ocean Expedition/Sargasso	actggctacactccagacgaaattctctatcgtgactgccgtttcttacagggc
123	50.GS	Global Ocean Expedition/Sargasso	gtcatgtacagcgaggcggagcagtgcggcaagaactgcaagctcctgcagggc
124	51.GS	Global Ocean Expedition/Sargasso	accgagtacagcaagcacgaggccaacggccagaactgccgctttcttcagggc
125	53.GS	Global Ocean Expedition/Sargasso	actggctatgcgaagaaggaggcacaagggcgcaactgccgcttcctgcaaggc
126	54.GS	Global Ocean Expedition/Sargasso	acaggcttctcgaagagagaggcagaaggcaccaactgccgcattattcaaggg
127	57.GS	Global Ocean Expedition/Sargasso	acgggctactcgcgggaggaggtgctggggcggaactgccgcttcctgcagggc
128	58.GS	Global Ocean Expedition/Sargasso	acgggctactcgaaatcggacgccgtcggaaaaagttgccgcttcctgcagggg
129	59.GS	Global Ocean Expedition/Sargasso	accggttacagcaaatcggaggcacagggccgcaactgccgcttcttgcagggt
130	60.GS	Global Ocean Expedition/Sargasso	accggctatagtattgaggaatctttagggaaaaattgtagatttcttcaaggc
131	62.GS	Global Ocean Expedition/Sargasso	acaggggatgttccggacgaggccgttggtaaaacttgccgcattttgaaatcg
132	64.GS	Global Ocean Expedition/Sargasso	tgcggctacgacgccgaggaggcgctcggcaagacgtgcaagatgctgcaggga
133	66.GS	Global Ocean Expedition/Sargasso	acgggttattcccgagaggatattattggtaaaaactgccgttttatagaggct
134	AUS688	EBPR Sludge Metagenome	accggttatacacccgaagaagtgattggcaagaattgtcgtttcttgcagggc
135	AUS644	EBPR Sludge Metagenome	tccggttattcgcaaaaggaagtcgtgggccgcaactgccgcttcctccagggc
136	AUS025	EBPR Sludge Metagenome	tccggttactcgcaagaggaagtcgtgggccgcaactgccgcttcctgcaaggc

137	AM434	Acid Mine Drainage Biofilm Metagenome	acggggtaccgccatgatgaggtcgtggggcgggactgccgacttctccagggg
138	AM614	Acid Mine Drainage Biofilm Metagenome	agcggatactcccttgatgaaatcctggacaaaaactgtcggttcttgaaccgc
139	AM093	Acid Mine Drainage Biofilm Metagenome	acaggatattcccgggcggagacagtcgggaagaactgtcgttttttgcagggg
140	GLW929	Gutless Worm Metagenome	accggctacggcaccgatgaacttatcggcaaatcctgccgactcttgaattgc
141	GLW746	Gutless Worm Metagenome	accgggtataccgccgacgaattggtcggccaaaattgccgcaccttgaattgc
142	FS816	Waseca County Farm Soil Metagenome	accggctactcgcgcgacgaaattttgggaaaaaactgccgtttccttgccggc
143	FS256	Waseca County Farm Soil Metagenome	accggttacgccgaggcagagattatccaccgcaattgcagattccttgcaggg
144	FS774	Waseca County Farm Soil Metagenome	acgggatatgatgagcaggagatcgtcgcccagaattgccggttcctcgctgcg
145	FS654	Waseca County Farm Soil Metagenome	acgggaggtgcggaggccgatgtcntcggccggaattgccggtttctcgcccgt
146	WF793	Whale Fall Metagenome	accggctacgagact gacgagattctctatcaggattgccgattccttcagaag
147	WF796	Whale Fall Metagenome	accggctatgatcgtgagttcatcattggccgaaactgccgctttttgtcgggc
148	WF430	Whale Fall Metagenome	acaggatttaacaaagaagaaatattgggtagaaactgcaatatgttgcaatct
149	WF137	Whale Fall Metagenome	agtggttatagtagtaaagaacttataggtaaaaattgtaaagagttacgacat

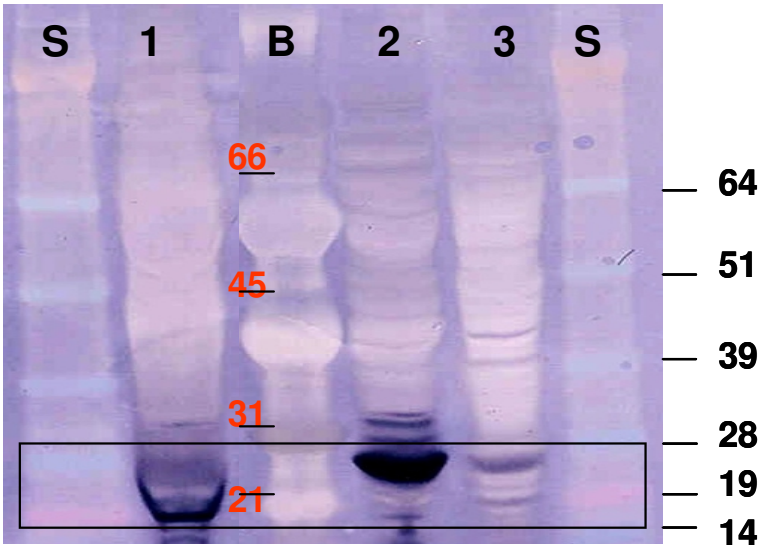
Appendix B: Phylogenetic relation of thermophilic fraction of garden soil metagenome (Ulrich Köhler, University of Hamburg). F1001 and F02/1902 refers to the clones containing 16S DNA from garden soil metagenome, sequenced for phylogenetic analysis.



Appendix C: Strains from High Altitude Andean Lakes (HAALs) screened using LOV microarray.

S.N.	Strains	Source Lake	Phylogenetic affiliation (16S)
1	A2	Azul	*Staphylococcus sp.*
2	Ap13	Aparejos	*Brevibacterium sp.*
3	A5	Azul	*Nocardia sp.*
4	N24	Negra	*Stenotrophomonas sp.*
5	N38	Negra	*Stenotrophomonas sp.*
6	N40	Negra	*Acinetobacter sp.*
7	V1	Vilama	*Pseudomonas/Xanthomonas sp.*
8	Ver3	Verde	*Acinetobacter sp.*
9	Ver5	Verde	*Acinetobacter sp.*
10	Ver7	Verde	*Acinetobacter sp.*

Appendix D: Western blot analysis of few LOV domain proteins that failed to show photochemical property.



S: SeeBlue Marker (in kDa)

1: Elbe-YtvA fusion synthetic construct (predicted molecular weight = 19.6 kDa)

B: Bio Rad Marker (in kDa)

2: Whale fall LOV domain protein with N and C-terminal 6X histidine (predicted molecular weight = 25 kDa)

3: Whale fall protein with N-terminal 6X histidine (predicted molecular weight = 23.6 kDa)

Appendix E: LOV domain containing genes found in metagenome databases

S.N.	Accession Number	Sequence Database	Source	Temperature of sample environment (°C)
1	JCVI_ORF_1096698682082	CAMERA	Global Ocean Sampling	37.6
2	JCVI_ORF_1105126376196	CAMERA	Global Ocean Sampling	18.6
3	JCVI_ORF_1105085123422	CAMERA	Global Ocean Sampling	28.4
4	JCVI_ORF_1096684715374	CAMERA	Global Ocean Sampling	20.5
5	JCVI_ORF_1105124753400	CAMERA	Global Ocean Sampling	20.5
6	JCVI_ORF_1105080898860	CAMERA	Global Ocean Sampling	29.3
7	JCVI_ORF_1096687332572	CAMERA	Global Ocean Sampling	37.6
8	JCVI_ORF_1096686573116	CAMERA	Global Ocean Sampling	37.6
9	JCVI_ORF_1096686282170	CAMERA	Global Ocean Sampling	20.5
10	JCVI_ORF_1105148840438	CAMERA	Global Ocean Sampling	28.3
11	JCVI_ORF_1096668445268	CAMERA	Global Ocean Sampling	28.3
12	JCVI_ORF_1096668445262	CAMERA	Global Ocean Sampling	28.3
13	JCVI_ORF_1096670714512	CAMERA	Global Ocean Sampling	28.3
14	JCVI_ORF_1105075290969	CAMERA	Global Ocean Sampling	28.3
15	JCVI_ORF_1096674403350	CAMERA	Global Ocean Sampling	19.8
16	JCVI_ORF_1096692846140	CAMERA	Global Ocean Sampling	22.9
17	JCVI_ORF_1096667325048	CAMERA	Global Ocean Sampling	18.6
18	JCVI_ORF_1105145233270	CAMERA	Global Ocean Sampling	11.7
19	JCVI_ORF_1105101567846	CAMERA	Global Ocean Sampling	22.9
20	JCVI_ORF_1096687818004	CAMERA	Global Ocean Sampling	27
21	JCVI_ORF_1105128122152	CAMERA	Global Ocean Sampling	20
22	JCVI_ORF_1105132901886	CAMERA	Global Ocean Sampling	25.5
23	JCVI_ORF_1096674168822	CAMERA	Global Ocean Sampling	22.9
24	JCVI_ORF_1105098984396	CAMERA	Global Ocean Sampling	28.3
25	JCVI_ORF_1105107608674	CAMERA	Global Ocean Sampling	28.3
26	JCVI_ORF_1105116671992	CAMERA	Global Ocean Sampling	28.3
27	JCVI_ORF_1105143100324	CAMERA	Global Ocean Sampling	28.3
28	JCVI_ORF_1105085223494	CAMERA	Global Ocean Sampling	19.8
29	JCVI_ORF_1105123932226	CAMERA	Global Ocean Sampling	28.3
30	JCVI_ORF_1105098583890	CAMERA	Global Ocean Sampling	26.4
31	JCVI_ORF_1096665878132	CAMERA	Global Ocean Sampling	26.4

32	JCVI_ORF_1105161266870	CAMERA	Global Ocean Sampling	27.6
33	JCVI_ORF_1105081255630	CAMERA	Global Ocean Sampling	22.9
34	JCVI_ORF_1105116019292	CAMERA	Global Ocean Sampling	26.2
35	JCVI_ORF_1105132161616	CAMERA	Global Ocean Sampling	37.6
36	JCVI_ORF_1105113646308	CAMERA	Global Ocean Sampling	22.9
37	JCVI_ORF_1105165120790	CAMERA	Global Ocean Sampling	27.6
38	JCVI_ORF_1105124217632	CAMERA	Global Ocean Sampling	28.7
39	JCVI_ORF_1105092271932	CAMERA	Global Ocean Sampling	28.3
40	JCVI_ORF_1105119874452	CAMERA	Global Ocean Sampling	25.4
41	JCVI_PEP_1105082692919	CAMERA	Global Ocean Sampling	37.6
42	JCVI_ORF_1096700206154	CAMERA	Global Ocean Sampling	20.5
43	JCVI_ORF_1105129645378	CAMERA	Global Ocean Sampling	25
44	JCVI_ORF_1105085155300	CAMERA	Global Ocean Sampling	20.5
45	AACY023493122	NCBI	Global Ocean Sampling	20.5
46	JCVI_ORF_1105109033888	CAMERA	Global Ocean Sampling	22.9
47	JCVI_ORF_1105080088700	CAMERA	Global Ocean Sampling	20.5
48	JCVI_ORF_1105080087070	CAMERA	Global Ocean Sampling	20.5
49	JCVI_ORF_1105153755710	CAMERA	Global Ocean Sampling	22.9
50	AACY021402583	NCBI	Global Ocean Sampling	20.5
51	JCVI_ORF_1105138050468	CAMERA	Global Ocean Sampling	22.9
52	JCVI_ORF_1105142186790	CAMERA	Global Ocean Sampling	22.9
53	JCVI_ORF_1105142355268	CAMERA	Global Ocean Sampling	27.5
54	JCVI_ORF_1112698808521	CAMERA	Global Ocean Sampling	28.9
55	JCVI_ORF_1112698808517	CAMERA	Global Ocean Sampling	28.9
56	JCVI_ORF_1105124217636	CAMERA	Global Ocean Sampling	28.7
57	JCVI_ORF_1112698347823	CAMERA	Global Ocean Sampling	28.9
58	JCVI_ORF_1105124525108	CAMERA	Global Ocean Sampling	18.6
59	JCVI_ORF_1105102734280	CAMERA	Global Ocean Sampling	20.5
60	JCVI_ORF_1105152260514	CAMERA	Global Ocean Sampling	20.5
61	JCVI_ORF_1112730632127	CAMERA	Global Ocean Sampling	26.4
62	JCVI_ORF_1105135527030	CAMERA	Global Ocean Sampling	37.6
63	JCVI_ORF_1105085223492	CAMERA	Global Ocean Sampling	19.8
64	JCVI_ORF_1105144971616	CAMERA	Global Ocean Sampling	27
65	JCVI_ORF_1112700543983	CAMERA	Global Ocean Sampling	27.2
66	JCVI_ORF_1105145351304	CAMERA	Global Ocean Sampling	20.5
67	JCVI_ORF_1105075921520	CAMERA	Global Ocean Sampling	25

68	JCVI_ORF_1105119180118	CAMERA	Global Ocean Sampling	20.5
69	JCVI_ORF_1112707117557	CAMERA	Global Ocean Sampling	
70	JCVI_ORF_1112689556339	CAMERA	Global Ocean Sampling	28.2
71	JCVI_ORF_1112688503741	CAMERA	Global Ocean Sampling	28.2
72	JCVI_ORF_1105135527020	CAMERA	Global Ocean Sampling	37.6
73	JCVI_ORF_1105090285476	CAMERA	Global Ocean Sampling	20.5
74	JCVI_ORF_1112698820899	CAMERA	Global Ocean Sampling	28.9
75	JCVI_ORF_1112700820459	CAMERA	Global Ocean Sampling	27.5
76	JCVI_ORF_1105081642194	CAMERA	Global Ocean Sampling	37.6
77	JCVI_ORF_1112688867317	CAMERA	Global Ocean Sampling	28.2
78	JCVI_ORF_1112688627123	CAMERA	Global Ocean Sampling	28.2
79	JCVI_ORF_1112688627127	CAMERA	Global Ocean Sampling	28.2
80	JCVI_ORF_1105112209512	CAMERA	Global Ocean Sampling	19.8
81	JCVI_ORF_1105112209510	CAMERA	Global Ocean Sampling	19.8
82	JCVI_ORF_1105119180116	CAMERA	Global Ocean Sampling	20.5
83	JCVI_ORF_1112740824577	CAMERA	Global Ocean Sampling	26.6
84	JCVI_ORF_1105102734276	CAMERA	Global Ocean Sampling	20.5
85	JCVI_ORF_1112698063915	CAMERA	Global Ocean Sampling	25.8
86	JCVI_ORF_1112733814917	CAMERA	Global Ocean Sampling	25.8
87	JCVI_ORF_1105122892786	CAMERA	Global Ocean Sampling	25.4
88	JCVI_ORF_1105111481394	CAMERA	Global Ocean Sampling	20.5
89	JCVI_ORF_1105148434592	CAMERA	Global Ocean Sampling	22.9
90	JCVI_ORF_1105092952048	CAMERA	Global Ocean Sampling	28.3
91	JCVI_ORF_1105096310770	CAMERA	Global Ocean Sampling	28.3
92	JCVI_ORF_1112698347821	CAMERA	Global Ocean Sampling	28.9
93	JCVI_ORF_1105124525106	CAMERA	Global Ocean Sampling	18.6
94	JCVI_ORF_1105144971614	CAMERA	Global Ocean Sampling	27
95	JCVI_ORF_1105135527028	CAMERA	Global Ocean Sampling	37.6
96	JCVI_ORF_1105081642196	CAMERA	Global Ocean Sampling	37.6
97	JCVI_ORF_1105112209508	CAMERA	Global Ocean Sampling	19.8
98	JCVI_ORF_1105101567844	CAMERA	Global Ocean Sampling	22.9
99	JCVI_ORF_1112698129339	CAMERA	Global Ocean Sampling	25.8
100	JCVI_ORF_1105153755714	CAMERA	Global Ocean Sampling	22.9
101	JCVI_ORF_1105085155306	CAMERA	Global Ocean Sampling	20.5
102	JCVI_ORF_1105110473978	CAMERA	Global Ocean Sampling	28.3
103	JCVI_ORF_1105111505266	CAMERA	Global Ocean Sampling	28.3

104	JCVI_ORF_1105111505270	CAMERA	Global Ocean Sampling	28.3
105	JCVI_ORF_1105111505268	CAMERA	Global Ocean Sampling	28.3
106	JCVI_ORF_1105142667664	CAMERA	Global Ocean Sampling	28.3
107	JCVI_ORF_1105142667666	CAMERA	Global Ocean Sampling	28.3
108	JCVI_ORF_1105158916276	CAMERA	Global Ocean Sampling	20.5
109	JCVI_ORF_1105153755712	CAMERA	Global Ocean Sampling	22.9
110	JCVI_ORF_1105096310774	CAMERA	Global Ocean Sampling	28.3
111	JCVI_ORF_1105128774228	CAMERA	Global Ocean Sampling	28.3
112	JCVI_ORF_1105128774226	CAMERA	Global Ocean Sampling	28.3
113	JCVI_ORF_1105109024408	CAMERA	Global Ocean Sampling	22.9
114	JCVI_ORF_1105085155304	CAMERA	Global Ocean Sampling	20.5
115	JCVI_ORF_1105161346770	CAMERA	Global Ocean Sampling	28.3
116	JCVI_ORF_1105161346774	CAMERA	Global Ocean Sampling	28.3
117	JCVI_ORF_1105090611388	CAMERA	Global Ocean Sampling	28.3
118	JCVI_ORF_1105116671994	CAMERA	Global Ocean Sampling	28.3
119	JCVI_ORF_1105107419960	CAMERA	Global Ocean Sampling	37.6
120	JCVI_ORF_1105083047918	CAMERA	Global Ocean Sampling	37.6
121	JCVI_ORF_1105107005220	CAMERA	Global Ocean Sampling	37.6
122	JCVI_ORF_1105100325920	CAMERA	Global Ocean Sampling	37.6
123	JCVI_ORF_1105094283744	CAMERA	Global Ocean Sampling	37.6
124	JCVI_ORF_1105130584246	CAMERA	Global Ocean Sampling	37.6
125	JCVI_ORF_1105092271938	CAMERA	Global Ocean Sampling	28.3
126	JCVI_ORF_1105096310772	CAMERA	Global Ocean Sampling	28.3
127	JCVI_ORF_1105098984400	CAMERA	Global Ocean Sampling	28.3
128	JCVI_ORF_1105097504288	CAMERA	Global Ocean Sampling	20.5
129	JCVI_ORF_1105090611390	CAMERA	Global Ocean Sampling	28.3
130	JCVI_ORF_1105136820831	CAMERA	Global Ocean Sampling	22.9
131	JCVI_ORF_1105136820837	CAMERA	Global Ocean Sampling	22.9
132	JCVI_ORF_1105130922146	CAMERA	Global Ocean Sampling	37.6
133	JCVI_ORF_1105130922148	CAMERA	Global Ocean Sampling	37.6
134	JCVI_ORF_1105130922144	CAMERA	Global Ocean Sampling	37.6
135	JCVI_ORF_1105132901884	CAMERA	Global Ocean Sampling	25.5
136	JCVI_ORF_1105132901888	CAMERA	Global Ocean Sampling	25.5
137	JCVI_ORF_1105075290971	CAMERA	Global Ocean Sampling	28.3
138	JCVI_ORF_1105075290973	CAMERA	Global Ocean Sampling	28.3
139	JCVI_ORF_1105083047912	CAMERA	Global Ocean Sampling	37.6

140	JCVI_ORF_1105083047914	CAMERA	Global Ocean Sampling	37.6
141	JCVI_ORF_1112726389575	CAMERA	Global Ocean Sampling	
142	JCVI_ORF_1105107005216	CAMERA	Global Ocean Sampling	37.6
143	JCVI_ORF_1105100325916	CAMERA	Global Ocean Sampling	37.6
144	JCVI_ORF_1105094283740	CAMERA	Global Ocean Sampling	37.6
145	JCVI_ORF_1105130584242	CAMERA	Global Ocean Sampling	37.6
146	JCVI_ORF_1105107005214	CAMERA	Global Ocean Sampling	37.6
147	JCVI_ORF_1105100325914	CAMERA	Global Ocean Sampling	37.6
148	JCVI_ORF_1105094283738	CAMERA	Global Ocean Sampling	37.6
149	JCVI_ORF_1105130584240	CAMERA	Global Ocean Sampling	37.6
150	JCVI_ORF_1105107419950	CAMERA	Global Ocean Sampling	37.6
151	JCVI_ORF_1112697680445	CAMERA	Global Ocean Sampling	28.9
152	JCVI_ORF_1112732656761	CAMERA	Global Ocean Sampling	28.8
153	JCVI_ORF_1105099124700	CAMERA	Global Ocean Sampling	37.6
154	JCVI_ORF_1105132161614	CAMERA	Global Ocean Sampling	37.6
155	JCVI_ORF_1105152260516	CAMERA	Global Ocean Sampling	20.5
156	JCVI_ORF_1105145351306	CAMERA	Global Ocean Sampling	20.5
157	JCVI_ORF_1112697680455	CAMERA	Global Ocean Sampling	28.9
158	JCVI_ORF_1112732656765	CAMERA	Global Ocean Sampling	28.8
159	JCVI_ORF_1112697680457	CAMERA	Global Ocean Sampling	28.9
160	JCVI_ORF_1112732656763	CAMERA	Global Ocean Sampling	28.8
161	JCVI_ORF_1105085123420	CAMERA	Global Ocean Sampling	28.4
162	JCVI_ORF_1112730911523	CAMERA	Global Ocean Sampling	26.4
163	JCVI_ORF_1112730911519	CAMERA	Global Ocean Sampling	26.4
164	JCVI_ORF_1105090285478	CAMERA	Global Ocean Sampling	20.5
165	JCVI_ORF_1105090611386	CAMERA	Global Ocean Sampling	28.3
166	JCVI_ORF_1105098984398	CAMERA	Global Ocean Sampling	28.3
167	JCVI_ORF_1105155182708	CAMERA	Global Ocean Sampling	20.5
168	JCVI_ORF_1105092952046	CAMERA	Global Ocean Sampling	28.3
169	JCVI_ORF_1096684275532	CAMERA	Global Ocean Sampling	20.5
170	JCVI_ORF_1096685113010	CAMERA	Global Ocean Sampling	20.5
171	JCVI_ORF_1096669084170	CAMERA	Global Ocean Sampling	20.5
172	WesternChannelOMM_READ_04868476	CAMERA	Western Channel Observatory	10.05
173	WesternChannelOMM_READ_05659548	CAMERA	Western Channel Observatory	10.1
174	WesternChannelOMM_READ_05095822	CAMERA	Western Channel Observatory	10.05
175	WesternChannelOMM_READ_06291290	CAMERA	Western Channel Observatory	10.05

176	WesternChannelOMM_READ_06484310	CAMERA	Western Channel Observatory	10.05
177	WesternChannelOMM_READ_04879250	CAMERA	Western Channel Observatory	10.05
178	NCBI_READ_1610196915	NCBI	Whale Fall	20
179	NCBI_READ_1594424875	NCBI	Whale Fall	20
180	NCBI_READ_1594426009	NCBI	Whale Fall	20
181	NCBI_READ_1594443623	NCBI	Whale Fall	20
182	NCBI_READ_1594426008	NCBI	Whale Fall	20
183	1569701381	CAMERA	Acid Mine Drainage	42
184	1569712663	CAMERA	Acid Mine Drainage	42
185	1569870459	CAMERA	Acid Mine Drainage	42
186	1569752263	CAMERA	Acid Mine Drainage	42
187	1569871995	CAMERA	Acid Mine Drainage	42
188	1569862503	CAMERA	Acid Mine Drainage	42
189	ABPP01003740	NCBI	Hypersaline Mat	15
190	ABPS01002685	NCBI	Hypersaline Mat	15
191	ABPQ01005096	NCBI	Hypersaline Mat	15
192	ABPQ01003968	NCBI	Hypersaline Mat	15
193	ABPS01005100	NCBI	Hypersaline Mat	15
194	ABPU01010779	NCBI	Hypersaline Mat	15
195	1557368101	CAMERA/NCBI	Minnesota Farm Soil	20
196	1557367607	CAMERA/NCBI	Minnesota Farm Soil	20
197	1557367152	CAMERA/NCBI	Minnesota Farm Soil	20
198	1557447683	CAMERA/NCBI	Minnesota Farm Soil	20
199	ABEF01020575	NCBI	ALOHA/Pacific Gyre	1.46
200	ABEF01028922	NCBI	ALOHA/Pacific Gyre	1.46
201	ABEF01030149	NCBI	ALOHA/Pacific Gyre	1.46
202	DU745808	NCBI	ALOHA/Pacific Gyre	26.4
203	BroadPhage_READ_01585282	CAMERA	Cyanophage	22.7
204	BroadPhage_READ_01587598	CAMERA	Cyanophage	22.7
205	BroadPhage_READ_01816240	CAMERA	Cyanophage	22.7
206	2003417106	IMG/M	Indoor Air	25.7
207	2003490754	IMG/M	Indoor Air	25.7
208	2003486340	IMG/M	Indoor Air	25.7
209	2003482190	IMG/M	Indoor Air	25.7
210	2003475518	IMG/M	Indoor Air	25.7
211	2003475157	IMG/M	Indoor Air	25.7

212	MarinePennateDiatoms_READ_00706921	CAMERA	Marine Pennate Diatoms	
213	MarinePennateDiatoms_READ_00774967	CAMERA	Marine Pennate Diatoms	
214	MarinePennateDiatoms_READ_00664334	CAMERA	Marine Pennate Diatoms	
215	MarinePennateDiatoms_READ_00675013	CAMERA	Marine Pennate Diatoms	
216	MarinePennateDiatoms_READ_00699467	CAMERA	Marine Pennate Diatoms	
217	MarinePennateDiatoms_READ_00656175	CAMERA	Marine Pennate Diatoms	
218	MarinePennateDiatoms_READ_00788754	CAMERA	Marine Pennate Diatoms	
219	MarinePennateDiatoms_READ_00664039	CAMERA	Marine Pennate Diatoms	
220	MarinePennateDiatoms_READ_00692992	CAMERA	Marine Pennate Diatoms	
221	MarinePennateDiatoms_READ_00442349	CAMERA	Marine Pennate Diatoms	
222	MarinePennateDiatoms_READ_00717260	CAMERA	Marine Pennate Diatoms	
223	MarinePennateDiatoms_READ_00712019	CAMERA	Marine Pennate Diatoms	
224	MarinePennateDiatoms_READ_00455794	CAMERA	Marine Pennate Diatoms	
225	JGI_SLUDGE_AUS_READ_1434053080	CAMERA	EBPR Sludge	23.5
226	JGI_SLUDGE_US_READ_1404899503	CAMERA	EBPR Sludge	23.5
227	JGI_SLUDGE_AUS_READ_1434047067	CAMERA	EBPR Sludge	23.5
228	JGI_SLUDGE_AUS_READ_1434049231	CAMERA	EBPR Sludge	23.5
229	JGI_SLUDGE_US_READ_1404982625	CAMERA	EBPR Sludge	23.5
230	JGI_SLUDGE_US_READ_1401973846	CAMERA	EBPR Sludge	23.5
231	pWThLOV-HT-Met1 (EU934096)	NCBI	Garden Soil Enrichment	65

Appendix F: Global Ocean Sampling Expedition (GOS) sample stations and their locations

S.N.	Sample Dataset	Sample Location	Country	Latitude	Longitude
1	GS000a	Sargasso Station 11	Bermuda	31°10'30"N	64°19'27.6"W
2	GS000a	Sargasso Station 13	Bermuda	31°10'30"N	64°19'27.6"W
3	GS000b	Sargasso Station 11	Bermuda	31°10'30"N	64°19'27.6"W
4	GS000b	Sargasso Station 13	Bermuda	31°10'30"N	64°19'27.6"W
5	GS000c	Sargasso Stations 3	Bermuda	32°10'29.4"N	64°00'36.6"W
6	GS000d	Sargasso Station 13	Bermuda	31°10'30"N	64°19'27.6"W
7	GS001a	Hydrostation S	Bermuda	32°10'00"N	64°30'00"W
8	GS001b	Hydrostation S	Bermuda	32°10'00"N	64°30'00"W
9	GS001c	Hydrostation S	Bermuda	32°10'00"N	64°30'00"W
10	GS002	Gulf of Maine	Canada	42°30'11"N	67°14'24"W
11	GS003	Browns Bank, Gulf of Maine	Canada	42°51'10"N	66°13'2"W
12	GS004	Outside Halifax, Nova Scotia	Canada	44°8'14"N	63°38'40"W
13	GS005	Bedford Basin, Nova Scotia	Canada	44°41'25"N	63°38'14"W
14	GS006	Bay of Fundy, Nova Scotia	Canada	45°6'42"N	64°56'48"W
15	GS007	Northern Gulf of Maine	Canada	43°37'56"N	66°50'50"W
16	GS008	Newport Harbor, RI	USA	41°29'9"N	71°21'4"W
17	GS009	Block Island, NY	USA	41°5'28"N	71°36'8"W
18	GS010	Cape May, NJ	USA	38°56'24"N	74°41'6"W
19	GS011	Delaware Bay, NJ	USA	39°25'4"N	75°30'15"W
20	GS012	Chesapeake Bay, MD	USA	38°56'49"N	76°25'2"W
21	GS013	Off Nags Head, NC	USA	36°0'14"N	75°23'41"W
22	GS014	South of Charleston, SC	USA	32°30'25"N	79°15'50"W
23	GS015	Off Key West, FL	USA	24°29'18"N	83°4'12"W
24	GS016	Gulf of Mexico	USA	24°10'29"N	84°20'40"W
25	GS017	Yucatan Channel	Mexico	20°31'21"N	85°24'49"W
26	GS018	Rosario Bank	Honduras	18°2'12"N	83°47'5"W
27	GS019	Northeast of Colon	Panama	10°42'59"N	80°15'16"W
28	GS020	Lake Gatun	Panama	9°9'52"N	79°50'10"W
29	GS021	Gulf of Panama	Panama	8°7'45"N	79°41'28"W
30	GS022	250 miles from Panama City	Panama	6°29'34"N	82°54'14"W
31	GS023	30 miles from Cocos Island	Costa Rica	5°38'24"N	86°33'55"W
32	GS025	Dirty Rock, Cocos Island	Costa Rica	5°33'10"N	87°5'16"W

33	GS026	134 miles NE of Galapagos	Ecuador	1°15'51"N	90°17'42"W
34	GS027	Devil's Crown, Floreana Island	Ecuador	1°12'58"S	90°25'22"W
35	GS028	Coastal Floreana	Ecuador	1°13'1"S	90°19'11"W
36	GS029	North James Bay, Santigo Island	Ecuador	0°12'0"S	90°50'7"W
37	GS030	Warm seep, Roca Redonda	Ecuador	0°16'20"N	91°38'0"W
38	GS031	Upwelling, Fernandina Island	Ecuador	0°18'4"S	91°39'6"W
39	GS032	Mangrove on Isabella Island	Ecuador	0°35'38"S	91°4'10"W
40	GS033	Punta Cormorant, Hypersaline Lagoon, Floreana Island	Ecuador	1°13'42"S	90°25'45"W
41	GS034	North Seamore Island	Ecuador	0°22'59"S	90°16'47"W
42	GS035	Wolf Island	Ecuador	1°23'21"N	91°49'1"W
43	GS036	Cabo Marshall, Isabella Island	Ecuador	0°1'15"S	91°11'52"W
44	GS037	Equatorial Pacific TAO Buoy	International	1°58'26"S	95°0'53"W
45	GS038	Tropical South Pacific	International	2°34'55"S	97°51'5"W
46	GS039	Tropical South Pacific	International	3°20'36"S	101°22'26"W
47	GS040	Tropical South Pacific	International	4°29'56"S	105°04'12"W
48	GS041	Tropical South Pacific	International	5°55'48"S	108°41'13"W
49	GS042	Tropical South Pacific	International	7°06'27"S	116°07'9"W
50	GS043	Tropical South Pacific	International	7°39'40"S	120°24'8"W
51	GS044	600 miles from F. Polynesi	International	8°24'54"S	124°14'23"W
52	GS045	400 miles from F. Polynesi	International	9°01'3"S	127°46'2"W
53	GS046	300 miles from F. Polynesi	International	9°34'16"S	131°29'30"W
54	GS047	201 miles from F. Polynesi	French Polynesia	10°7'53"S	135°26'58"W
55	GS048a	Moorea, Cooks Bay	Fr. Polynesia	17°28'33"S	149°48'44"W
56	GS048b	Moorea, Cooks Bay	Fr. Polynesia	17°28'33"S	149°48'44"W
57	GS049	Moorea, Outside Cooks Bay	Fr. Polynesia	17°27'11"S	149°47'56"W
58	GS050	Tikehau Lagoon	Fr. Polynesia	15°16'40"S	148°13'28"W
59	GS051	Rangirora Atoll	Fr. Polynesia	15°8'37"S	147°26'6"W
60	GS108a	Coccos Keeling, Inside Lagoon	Australia	12°5'33"S	96°52'54"E
61	GS108b	Coccos Keeling, Inside Lagoon	Australia	12°5'33"S	96°52'54"E
62	GS109	Indian Ocean	International	10°56'37"S	92°3'32"E
63	GS110a	Indian Ocean	International	10°26'46"S	88°18'10"E
64	GS110b	Indian Ocean	International	10°26'46"S	88°18'10"E
65	GS111	Indian Ocean	International	9°35'49"S	84°11'51"E
66	GS112a	Indian Ocean	International	8°30'18"S	80°22'32"E

67	GS112b	Indian Ocean	International	8°30'18"S	80°22'32"E
68	GS113	Indian Ocean	International	7°0'27"S	76°19'53"E
69	GS114	500 Miles west of the Seychelles in the Indian Ocean	International	4°59'25"S	64°58'36"E
70	GS115	Indian Ocean	International	4°39'45"S	60°31'23"E
71	GS116	Outside Seychelles, Indian Ocean	Seychelles	4°38'6"S	56°50'10"E
72	GS117a	St. Anne Island, Seychelles	Seychelles	4°36'49"S	55°30'31"E
73	GS117b	St. Anne Island, Seychelles	Seychelles	4°36'49"S	55°30'31"E
74	GS119	International Water Outside of Reunion Island	International	23°12'58"S	52°18'22"E
75	GS120	Madagascar Waters	Madagascar	26°2'6"S	50°7'23"E
76	GS121	International water between Madagascar and South Africa	International	29°20'56"S	43°12'56"E
77	GS122a	International water between Madagascar and South Africa	International	30°53'54"S	40°25'13"E
78	GS122b	International water between Madagascar and South Africa	International	30°53'54"S	40°25'13"E
79	GS123	International water between Madagascar and South Africa	International	32°23'57"S	36°35'31"E
80	GS148	East coast Zanzibar, offshore Paje lagoon	Tanzania	6°19'S	39°33"E
81	GS149	West coast Zanzibar, harbour region	Tanzania	6°7'S	39°07'E
82	MOVE858	Chesapeake Bay, MD	USA	38°8'N	76°23'W

Die hier vorgelegte Dissertation habe ich eigenständig und ohne unerlaubte Hilfe angefertigt. Die Dissertation wurde in der vorgelegten oder in ähnlicher Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Mülheim an der Ruhr

Gopal Prasad Pathak