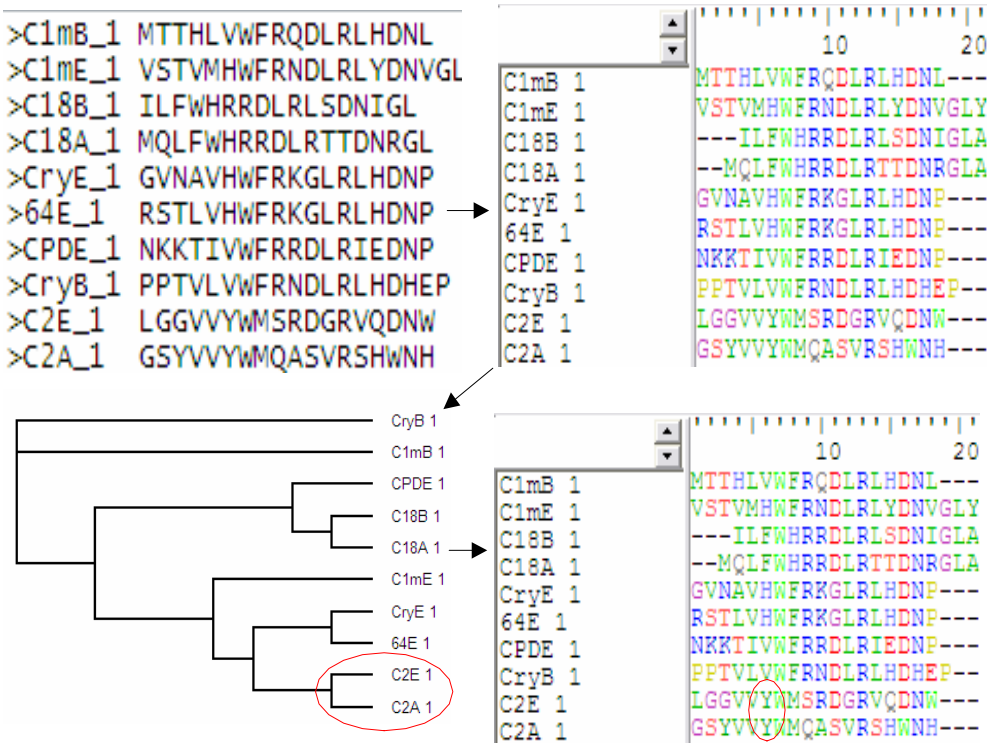
Bioinformatic approaches to search for functional differences yielding evolutionary branches, using the photolyase/blue-light photoreceptor family as an example case.

Bioinformatiske metoder for leting etter funksjonelle endringer som forårsaker evolusjonære forgreninger, med bruk av fotolyase/blå-lys fotoreseptor familien som et eksempel.



Utført av Harald Grove
2004
Ved institutt for naturforvaltning
Norges landbrukshøyskole, NLH

Preface

This thesis has been done as the finishing project of my degree at the Agricultural University of Norway at the Department of Ecology and Natural Resource Management, INA. The thesis was finished at the first of June 2004.

I would first of all like to thank my main supervisor Professor Manfred Heun at INA for his invaluable support with all parts of this work. I would also like to thank Dr. Anne Kristoffersen, currently at the Department of Biology, University of Oslo, acted partly as co-supervisor and also provided valuable help with the lab-work and the biochemical properties of the photolyase family.

The help and support for the PCR and sequencing reaction in addition to much of the DNA template were provided by Kari Vollan, without whose support things would have taken a lot more time. Additional DNA templates for the PCR reaction were provided by Ola Westereng (cotton), Linn Nilsen (cactus) and Eunice Olten (*Solanum nigrum*).

Seeds from *Arabidopsis thaliana* were kindly provided by Professor Atle Bones at NTNU.

Help with the computer programs and the algorithms for the alignment and phylogeny procedures were provided by Dr. Lars Snipen at IKBN and Jørn Henrik Sønsteby at INA.

Ås, 29.05.2004

_____

Harald Grove

Summary

This project was done to get an idea on how much information could be gained just by analysing sequence data. The photolyase gene family was used as an example. Since this family consists of five more or less different functional groups (three groups with repairing capability; (6-4) photolyase, class I and class II CPD photolyase and two groups with plant and animal blue light photoreceptrors), the goal was to figure out to what degree the difference between each functional group could be predicted with only the amino acid sequence as a reference. The amino acid sequence data was chosen instead of the nucleotide sequence data because the functionality of the gene lies in what amino acids are present in the final protein, and in what order.

The project consisted of two parts; laboratory work and computer work. The laboratory work was done to get some experience with the DNA amplification method, PCR (Polymerase Chain Reaction), and as an example on how a gene found in one species can be amplified from another species, using primers designed based on conserved areas in the already known gene sequences. If this part had worked fine then the PCR products were to be sequenced, and the sequence compared with the other sequences downloaded from the sequence databases. However, the first steps of the lab work ended up taking too much time to quite finish that part of the project. Even so, the results that were achieved showed that the method quite probably would have worked fine. The results from the PCR showed the presence of a probable photolyase homologue in cotton, tomato, cactus and maize. It would have been interesting to sequence the bands present to see if they actually showed the presence of a photolyase gene, or just a random hit in the genome.

The computer part was to analyse a number of sequences downloaded from a sequence database on the internet. With the photlyase/blue-light photoreceptor family as the source of the sequences, 38 sequences were downloaded. Then a multiple alignment was done, followed by the construction of a phylogenetic tree based on the alignment. These results corresponds well with what has previously been published, showing the groups of Cyclobutane Pyrimidine Dimer (CPD) photolyase (class I and II), 6-4 photolyase and cryptochromes (animals and plants).

The results showed that some indications of functional divergence could be gained from an analysis of the multiple alignment. Part of the gene showed little to no conservation, while other areas were to some extent fully conserved over all the sequences. Using this conservation as an indication for functionally importance several sites in the alignment were indicated as potentially important. Comparison with the known 3D structure of the enzyme showed that the gene function seemed to be able to tolerate a certain degree of variance in the protein sequence, even in the sites deemed important.

The multiple alignment was then searched to find potential sites for explaining the separation between the major groups shown in the phylogenetic tree. Several sites were detected, which could be important for the functional divergence. It was not possible within the scope of this thesis to get any proofs of the importance of these sites, but the possibilities for further research are good.

Sammendrag

Denne oppgaven ble utført for å få en ide om hvor mye informasjon som ligger tilgjengelig i sekvens data. Som et eksempel ble det valgt å bruke en genfamilie bestående av fotolyasegener. Denne familien besto av fem i varierende grad funksjonelt forskjellige grupper (tre grupper med repareringsmulighet; (6-4) fotolyase, klasse I og klasse II CPD (Cyclobutan Pyrimidin Dimer) fotolyase og to grupper bestående av blå-lys reseptorer fra plante og dyreriket), målet med oppgaven ble satt til å finne ut i hvor stor grad forskjellen mellom de funksjonelle gruppene kunne bli bestemt ut fra aminosyre rekkefølgen som referanse. Aminosyrer ble valgt i stedet for nukleinsyrer (DNA) for sekvensanalysen, fordi funksjonaliteten til genene ligger i rekkefølgen av aminosyrene i det ferdige proteinet.

Oppaven besto av to deler; laboratoriearbeid og datamaskinarbeid. Laboratoriearbeidet ble utført for å få erfaring med DNA amplifisering ved hjelp av PCR (Polymerase Chain Reaction). I tillegg skulle det vises hvordan et gen funnet i en eller flere andre arter kan fungere som basis til å amplifisere det samme genet i en annen art, ved hjelp av primere laget baser på konserverte områder i sekvensen fra de andre artene. Hvis denne delen hadde fungert etter planen, så ville neste steg være å sekvensere produktet fra DNA amplifiseringen. Det oppstod imidlertid problemer slik at den første delen av laboratoriearbeidet tok lengre tid enn beregnet. Dette førte til at resten av arbeidet med sekvenseringen måtte skrinlegges. Selv om denne delen av prosjektet ikke kunne gjennomføres fullt ut, viste det som ble gjort at det høyst sannsynlig ville fungert som planlagt. Resultatene fra PCR'en viste en mulig fotolyase homolog i bomull, tomat, kaktus og mais. Det ville vært interessant å se om dette faktisk var fotolyasegener eller bare et tilfeldig treff i genomet.

Resten av prosjektet var lagt opp basert på sekvenser hentet fra sekvensdatabaser på internett. Med fotolyase/blålys fotoreseptor familien som basis ble det funnet og lastet ned 38 sekvenser. Deretter ble det gjennomført en multippel sammenstilling av sekvensene, etterfulgt av en fylogenetisk analyse basert på denne sammenstillingen. Resultatene fra disse analysene stemt godt overens med tidligere publiserte artikler.

Resultatene av søket gjennom den multiple sammenstillingen ga indikasjoner på flere potensielt viktige funksjonelle områder. Et område av genet viste liten eller ingen konservering, mens andre var tilnærmet uforandret over alle sekvensene. Denne invariansen kan brukes som hint om at disse områdene inneholder funksjonelt viktige posisjoner. Sammenligning med den kjente 3D strukturen av enzymet, viste genet tydeligvis har mulighet til å tolerere en viss variasjon, selv i posisjoner som ser viktige ut.

Neste skritt var å bruke den samme sammenstillingen til å lete etter de posisjonene som er opphav til forgreningspunktene påvist i det fylogenetiske treet. Flere potensielt viktige posisjoner ble påvist. Det var ikke mulig innenfor rammene av denne oppgaven å påvise noen funksjonelle egenskaper til disse posisjonene, men resultatene viser at metoden har potensiale.

Index

1. Introduction

With the development of increasingly faster and more powerful sequencing techniques, more and more whole genomes have been sequenced during the last decade. This has led to an increasing interest in functional genomics, being able to compare the whole genome of one species with information gathered from other species. This also causes a number of possible genes to be proposed based on the finding of open reading frames in the genome. The identification of these hypothetical genes usually start with a match finding program like BLAST, thereby showing what known genes the new gene resembles. This gives an idea of what function the new gene may have. A phylogenetic method may then be used to gain an indication of the clustering of the new gene compared to the similar genes.

The phylogeny of the selected gene having divided into function can instead be used as a guideline to how the different variations of the gene evolved. The difference in function can be caused by several kinds of changes in the organism and the gene. One possibility is the mutations in the parts required for the expression of the protein; the promoter region, the enhancer areas or any transcription factors. Another is the possibility of an insertion or deletion (indel) causing a shift in the reading frame, leading to a protein with possibly quite different structure. The severity of this change depends on where the indel is located; early in the protein and the product will quite likely be unrecognisable and not functional, later on and the protein may keep some function although it will have acquired a different tail. Mutations within the intron parts of the sequence may also give an unrecognisable protein if it causes changes in the splicing of the mRNA.

All these cases have in common that they either cause the protein not to be produced or to be so changed as amounts to the same thing. This is going to make it, at worst, impossible to search for mRNAs or proteins from the affected organism. Any changes happened to the protein sequence will therefore be masked by the disappearance of the protein itself.

Since this thesis was about looking for changes in the protein structure, based on sequences derived mostly from mRNAs and proteins, any change that would lead to the disappearance or greatly transformation of the protein would not be detectable.

The changes in the gene expression that would be detectable for this thesis is the substitution of one or more bases in the gene sequence, causing a change in the amino acid composition of the final protein. Also, any indels causing the removal or addition of small segments of the protein will also be detectable as long as the majority of the protein is correct, and it is still translated. A rearrangement of the protein due to a different splicing pattern or the rearranging of the gene sequence may be discovered based on the search algorithm if scanning a sequence database, or the choice of primers or probes when working in the laboratory.

All the mutations and other changes mentioned will have to be non-deleterious to the organism to be able to detect the protein or mRNA. Otherwise, the organism wouldn't have been able to survive, and the mutations wouldn't be carried on. This leads to the assumption that all changes seen between different proteins have to either be neutral, slightly negative but not completely disrupting or positive. The positive changes can be either just an increase in effectiveness, or the development of a new function or new way of doing the old function. This thesis will look at examples of both these latter cases.

Based on an article by Kanai *et al.* (1995), a gene family were found that seemed a good start for this project; the photolyase gene. This gene repair damages to the DNA caused by UV-light.

Among other types of damage to the DNA, the UV-light can cause a binding between two adjacent pyrimidines. This gene family uses visible blue light to repair this damage. The photolyase family contains two types of photolyases repairing slightly different kinds of damage in addition to some similar but non-repairing relatives. This project will then have as a goal to try and suggest localisation of the residues important for this change in function between the two photolyases and the loss of function to those proteins devoid of repairing ability.

In the original article (see appendix A for easy reference) they operated with five groups; two from class I photolyase with cofactor MTHF and 8-HDF, blue-light photoreceptors, 6-4 photolyases and class II photolyase. The BLRs are classified as a group based on the fact that they all seem to contain a bit of sequence that is capable of receiving light energy and using as a reducing agent (FADH). As such there may be photolyases among this group that lacks a discovery of their functionality. The functional difference between class I and class II photolyase is not known at present, but there seems to be more prokaryotes in the class I version and more eukaryotes in the class II. This grouping is based on the alignment data and also on crystallographic analyses of the protein from E.coli (Park et.al 1995).

Especially one figure was used as inspiration for this work, shown in figure 1.1.

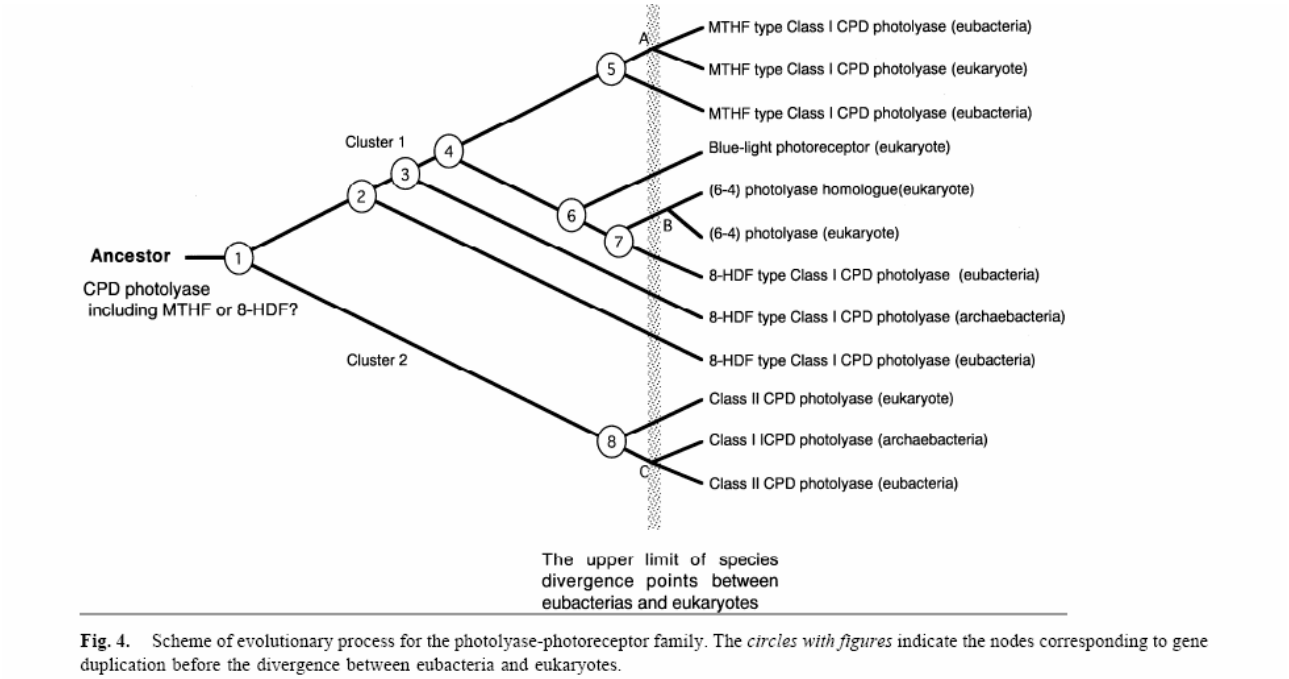


Fig. 4. Scheme of evolutionary process for the photolyase-photoreceptor family. The *circles with figures* indicate the nodes corresponding to gene duplication before the divergence between eubacteria and eukaryotes.

Figure 1.1: Figure 4 from the article by Kanai et al. (1995). Here they mention gene duplication events at the nodes. This thesis is about trying to find the changes following this duplication that lead to the change in function.

With this case as a guideline, several new photolyase genes were collected from the net, and added to the original 22 sequences. This was done by searching for articles describing the genes, or bye searching the sequence databases for matching genes. In the end 38 sequences were collected and analysed. The goal was to find the changes causing the split between, for example, (6-4) photolyase and the blue light photoreceptors at the node marked 6 in figure 1.1.

2. Theory

2.1. Sequence databases

The increase in the number of sequenced DNA, either small parts like genes and mRNA's or large parts like chromosomes and whole genomes, has led to the development of sequence databases. The most used are probably EMBL (http://www.ebi.ac.uk/embl/), NCBI (http://www.ncbi.nlm.nih.gov/) and PIR (http://pir.georgetown.edu/pirwww/search/textpsd.shtml). Here the full sequence of many genomes may be found, and the possibility of extensive sequence comparison has been made easier. The sequences are usually submitted with whatever information was known about them at the time. This could be anything from a full analysis of the gene expression with 3D structure and predicted or known functionality, to just the sequence itself. Lately, with the increasing number of sequenced genomes, a number of genes are submitted based solely on homologue analysis with known genes. These genes may have been found by screening the sequenced genome for homology matches to other known genes, or they may be the result of an analysis of the mRNA expressed in some tissue. In the latter case, if it came from an eukaryote, the intron information will be lost and only the post splicing sequence will be reported.

It is possible to try to infer the functional aspect of any newly proposed gene just by comparing it to other genes with known function and/or binding properties. The new gene can be scanned against the database using programs such as BLAST (found at the NCBI website), ParAlign (http://www.paralign.org/) or others. The resulting matches can give some indication of the function of the gene by observing what genes or fragments of genes it is found as homologue to.

If part of the new gene matches against part of an old gene shown to bind to FADH, then the new gene is also likely to be able to bind FADH. This inference should be made with some caution since homology does not necessarily mean identity, and a change in just a few bases might render the gene unable to do the proposed action, in this case; if the gene has lost one or more of the binding sites to FADH, while still retaining most of the primary structure, it may show up as a homologue but be unable to bind FADH. On the other hand, having some small changes in less important parts of the gene might not affect the function at all, and that's why this kind of search is made with a certain degree of tolerance against mismatches. The user has to specify how much difference is tolerated, or just manually inspect the suggested hits for possible matches.

2.2. Protein evolution

To analyse the difference in protein function, a model for the evolution of protein sequences is required. This is because the properties of the amino acids are in some cases overlapping; isoleucine may for example be substituted with leucine without affecting the protein in any serious way. A change from one amino acid to another in a certain part of the protein may have to be considered against this possibility. When aligning the sequences this model is used to minimize the total distance between all sequences, this will be further discussed later in the paper.

Since any changes to the protein is caused by changes in the DNA sequence, it would seem natural to analyse the DNA sequence to get a model of how the bases are either substituted or lost, thus creating a mutation which might lead to a new function. On the other hand, whether a protein is going to work or not is dependent on what amino acids it consists of and in what order they are present. If a protein loses its function due to a mutation, the cell may be unable to survive if the function was important enough. In this case the mutation will be lost. Any differences seen in the sequence of two functionally equivalent genes today must therefore be the result of either a beneficial mutation, or at least not a too detrimental mutation, or a non-essential mutation. In the latter case, this might be the substitution of one hydrophilic residue with another in a part of the protein only important for the solubility and/or structural properties. An example of the former case may be a different way to bind the substrate or being able to utilise a different cofactor in the reaction.

With the nucleotide data, the analysis will be one step removed from the function, since proteins are translated. Any information gained from the DNA sequence must be translated to the equivalent protein sequence before the functional importance of the observed changes is apparent. This is due to the matter of the redundancy in the code. Since there are 64 possible combinations with 3 sites of 4 nucleotides each and only 20 amino acids (21 if counting the stop codon), several 3-letter codes give the same amino acid. Valine is coded by the sets; GTT, GTC, GTA and GTG, which means a change in the third codon will be a silent mutation since you will still get the same protein composition. A mutation in the DNA which doesn’t result in any changes in the protein is called a synonymous mutation. The opposite will then be a nonsynonymous mutation. In a genome it is expected to find a certain amount of the former kind of mutation since it doesn't affect the organism in any significant way. On the other hand any nonsynonymous mutations will result in a different amino acid being inserted in the affected position of the protein. As mentioned above, if this change results in a change of function which is not for the good of the cell, the mutation is unlikely to be carried on to the next generation unless the cell can do without this gene. It has been shown that in general, for coding areas of the genome, the number of synonymous mutations outnumber the nonsynonymous mutations (Li *et al.* 1985).

To make a model of the amino acid substitution without using the nucleotide data, Dayhoff (1978) developed a protein substitution matrix, showing the probability that any amino acid mutates into any other. This matrix is based on a multiple alignment of sequences above a certain degree of identity. Later, several similar matrices have been suggested, depending upon the expected time of divergence between the species in the analysis. These include an improvement of the original Dayhoff matrix by Jones Taylor and Thornton (the JTT matrix, 1992), the Gonnet matrix (Gonnet *et al.* 1992) and the BLOSUM matrix (Henikoff and Henikoff 1992).

2.3. Alignment

The purpose of the alignment step is to generate a matrix where each row contains one sequence of either amino acids or nucleotides. The columns are then representing all the amino acids or nucleotides from the analysed sequences which have evolved from a common ancestral amino acid or nucleotide in the ancestral sequence. With an ideal alignment it will be possible to see how each site in the ancestral sequence has evolved. If the sequence hasn't had any deletions or inserts, the alignment is more or less the case of sliding each new sequence until it matches the previously aligned sequences. In its most simple form, this is what the sequence alignment programs are doing. They line up the sequences in such a way that as many character as possible matches. Whatever doesn't match is then considered to have changed sometime during the evolution of the sequences. Usually however, this only works for very closely related sequences or alternatively very much conserved sequences. What is more often the case is that each sequence needs the insertion of some gaps along the alignment to make as many amino acids or nucleotides match between each of the sequences. This is caused by another type of mutation; the gain or loss of one or more characters. This can be caused by the excision or insertion of nucleotides in the DNA sequence or, when considering proteins, a change in the preprocessing of the mRNA, changes in the excision of introns for example.

The general approach is to align the areas where both sequences are conserved, and introduce gaps in one or both of the sequences to make the rest match as good as possible. In the most liberal placement of gaps one could imagine that most, if not all, of the characters will be paired with a similar character as shown in figure 2.1.

```
Sequence 1 cgtgtagtcgtcatgt    cgtgtagtcgtcatgt-----
Sequence 2 cgtgttgtcgtcatgt    cgtgt--t-gtc--gtcatgt
                  ^
```

Figure 2.1: The alignment of two sequences with no restrictions on the placement of gaps. The single mutation from an 'a' to a 't' is marked.

The alignment in figure 2.1 shows what happens if gaps are placed with too little restriction. Then the single mutation marked at position '^' causes the subsequent nucleotides to be shifted away from their true position.

To avoid this problem, the insertion of one or more gaps in the alignment needs to be taken care of. That's why the alignment programs introduce two parameters called "gap open" and "gap extension". These indicate the penalty you would add to the final alignment score by adding respectively a gap and an additional gap. The difference between the two parameters stems from the assumption that as soon as an indel has happened, the chance for this event affecting more than one character is larger than for the indel effect in the first place.

Finding the right values for the two gap parameters is not immediately obvious. The amount of gaps tends to be correlated to how long it has been since the last common ancestor. The closer the time since the two split up the less chance there is that a mutation has happened. Accordingly the penalty for gap opening is increased the closer the common ancestor is supposed to have been.

Regarding ways of treating gap costs Durbin et al. (p44, 1998), references a paper by Vingron and Waterman (1994). One method is to choose the gap penalties based on the substitution model.

When the alignment is done, the final matrix can be used either as a way to generate a phylogeny, using the methods described earlier, or analysing the columns in the alignment matrix, looking for information on evolution between the different sequences.

When looking at the final alignment, some more or less obvious errors may be seen. This is due to the algorithm having been simplified to reduce computation time. One such error may be:

```
123456
LLDTGG -1
LLDTGG -2
LLDTDG -3
LL--DG -4
```

Here sequence 4 and 3 may have been aligned first and then added to the rest. Because the gaps inserted with the alignment of sequence 3 and 4 are not changed later, the D in position 5 will be kept in place, even though it may make more sense to place it at position 3. It may of course be that sequence 4 has lost the DT motif at position 3 and 4, and gotten a change from D to G at position 5. This shows that any manual editing of the alignment risks introducing a personal bias, which might be problematic when working with specific alignment positions. In the case of using an alignment for phylogenetic research, the impact of one small error like the one above will probably not matter much for the final result.

2.4. Phylogeny

After aligning the sequences, it is usually a good idea to generate a phylogeny for the sequences. There are several algorithms available for this task, and each algorithm can have some differences in the implementation. One method tries to find the most parsimonious tree structure based on the assumption that the solution that gives the least amount of evolution is to be preferred. Other methods calculates the pairwise distances between each pair of sequences and makes a clustering approach to the tree building, making the assumption that the closer the distance, the less time since a common ancestor. A third variant is to use maximum likelihood to calculate the probability for observing any tree given some predefined model of substitution, and then choose the tree which gives the best result. Within these groups there is also the possibility for different algorithms on how to actually calculate the relationships between sequences.

It has been shown that when generating a species tree, using too few or too short sequences, may lead to the wrong conclusions (Rokas *et al.* 2003). This result however, came from a study using eight different yeast types to show the general phylogeny between them. Since this thesis are concerned with the functional divergence of only one gene, the use of only one gene should not be a problem in this respect. All phylogenetic trees shown in this paper has to be regarded with this point in mind, they do not necessarily show the evolution between the species, but rather the evolution of the photolyase gene in question.

2.4.1. Distance based methods

All these methods have in common that they start with calculating a distance matrix P_{ij}, where each cell gives the evolutionary distance between species i and j. There are several proposed methods on how to calculate this distance matrix.

The easiest one is to just use the identity, where the number of equal character sites is divided by the total number of sites in the shortest sequence. This is similar to how distance is measured in marker analysis (AmplifiedFragmentLengthPolymorphism Vos *et.al* 1995), or when using phenotypic traits (eye colour, length etc.). The difference with sequence data is the possibility of grading all the non-identical sites based on the probability of the change having happened in a set amount of time. With nucleotide data, instead of setting the distance between two dissimilar bases to one, one could adjust the value dependent on if it was between for example A and T, A and G or A and C.

The simplest model of how nucleotides are substituted is presented by Jukes and Cantor (1969). Here they assume that all nucleotides have an equal probability of mutating into any other nucleotide, and that the frequency of each nucleotide is the same. This gives the estimation of the expected proportion of nucleotide differences between two sequences at a time t after their evolutionary separation as:

$k = - \frac{3}{4} \ln (1 - 4*d/3)$

where d is the number of differing nucleotides in the sequence and $k = \frac{3}{4}(8\alpha t)$ where α is the probability of mutating from one nucleotide to another.

With this model the number of observed differences reaches an asymptote at 0.75 as the time increases, which is to be expected since any two random sequences will be matching by chance in

25% of the sites. In practice this means that after a certain divergence, any further mutations have a high probability of hitting a previously mutated site. In that case it will either return to the original state, thereby hiding the fact that the site had mutated or it will change further, thereby replacing the first mutation with the second and giving rise to an underestimation of the total divergence.

Other models for nucleotide evolution have been developed to take into account that the nucleotides may have unequal frequency and that transversions and transitions may have unequal probability of happening (Kimura, 1981; Tamura & Nei 1993; among others).

After generating the distance matrix, a clustering method is usually applied such as the neighbor-join (NJ) (Saitou and Nei, 1987) or the Unweighted Pair Group Method with Arithmatic Mean (UPGMA).

When calculating the distance, the presence of gaps is causing problems. One method used is to simply ignore them. This will lead to some information loss in the sites affected by gaps, although these sites probably are the more variable in the sequence and therefore is more likely to be less informative of the phylogeny.

2.4.2. Parsimony

Based on the principle that the easiest solution is usually the best, this method looks at each site and tries to figure out how this could have evolved with the smallest amount of changes. The total number of required mutations is used as a measure of how good the final tree is.

Since searching through all possible trees usually takes a long time, there has been devised a couple of methods for narrowing the search area. One such approach is to start with a pre designed topology, usually found via a distance based method such as NJ or UPGMA. Then the goal is to refine that tree by swapping branches in the tree and recalculate the number of mutations needed. If the new tree turns out to be better it is taken as a new starting location, otherwise the first tree is kept. The three steps of branch-swapping, recalculation and possibly finding a new starting tree is repeated until no further improvement can be found. Then it is possible to generate a new random tree and do the whole procedure over again as long as needed (wanted) until the best possible tree is found. The problem with this approach is that the best tree found may not be overall best, just the best within the local region, and there is no sure way of knowing whether it is or not unless one examines all the trees.

If one just counts mutations there will also be a problem if a sequence has had more than one mutation one a site. This could be a reversal to the original character, or a further mutation to another character. In this case, two sequences may appear to be closer related than they in fact are, and this may cause problems with the topography. Additionally one may have to take into account that certain mutations are more common than others; in nucleotide sequences transitions tend to occur more easily than transversions and with protein sequences one amino acid may have an easier time being changed to a similar amino acid than otherwise. There is also the question on how to treat gaps; are these just another mutation or should there be a different weighting scheme for these positions.

2.4.3. Maximum Likelihood

A second method to generate a phylogeny from an alignment is to use a maximum likelihood algorithm. The principle of this method is to calculate the probability of a certain phylogeny, based

on an assumed model of evolution, and then choose the tree which gives the highest probability given the model. The advantage of this method is that a full model of evolution is required, and the effect of each site can be calculated and added to the total probability. Since this happens each time two or more sequences are compared, no information is lost except for what the user decides to omit. This is in contrast with the distance based methods where the information from each site is used only once to calculate the pairwise distance with each of the other sequences, and then discarded.

This method can easily be the most computationally demanding and requires that either the data set is small with few and/or short sequences or the computer is fast. Although the computer speed has been greatly increased during the last decade, it is still difficult to make a full analysis. In this case heuristic approaches can simplify the analysis while still give adequate answers. As with parsimony, one approach is to first get a basic topology and then rearrange the branches, recalculating the probability for the tree each time. This is usually done with some main branches defined as certain, and the branch swapping happening on the exterior branches. This rearranging is then performed until a maximum value for the probability is reached. The limit is either defined by limiting the number of branches open to swapping, or by using a chain method. In the former case, all possible branches are rearranged and the probability calculated for each tree, while in the latter case, the probability of the new tree, after swapping, is compared to the probability of the old tree, and whichever has the highest score is chosen as the starting point for the next branch swapping. This continues until no further higher probability tree can be found.

Another way of doing ML without having to analyse all possible phylogenies, is the quartet puzzling method (Strimmer and von Haesler, 1996). Here they base the tree construction on quartet-puzzling, calculating the ML tree for each quartet among all the sequences in the data set. Then one tree is randomly chosen as a starting point, and the rest of the sequences are randomly drawn and added to the tree based on their quartet relationship with the sequences already present in the tree. The tree building steps are then repeated a number of times chosen by the user, each time recording the final tree. Finally, a consensus tree is presented with each cluster having a corresponding value showing how many times this cluster appeared among all the puzzling trees. This value can be used as a measure of how good the final tree is, much the same as with the bootstrapping method, although the two methods should not be confused with each other.

2.4.4. Bayesian methods

Another approach to the tree building is to use Bayes' theorem as a basis for inferring phylogeny. With this method the trees are ranked according to their posterior probability P(tree | data), whereas the likelihood method uses the probability P(data | tree). Bayes' theorem is in its simplest form

Prob (H | D) = Prob (H & D) / Prob (D)

Here H is the hypothesis and the data set is D. Bayes theorem then describes the probability of the hypothesis given the data. The Metropolis algorithm (Metropolis et al. 1953) is one of the most widely used algorithm as a basis for a Bayesian inference of phylogeny. The algorithm is in short; first a tree is chosen, then another tree is chose as the possible next step. The ratio of probabilities of the proposed new tree and the old tree is calculated. If this ratio is above one, the new tree is chosen. If the ratio is less than one, then a random number from 0 to 1 is drawn and the new tree is chosen if the random number is less than the ratio. Whichever of the two trees was chosen is set as the new starting point the procedure is repeated. The main difference between this method and the heuristic search used in parsimony and maximum likelihood is that this method can with a non-zero probability choose a new tree that is less probable than the old. In this way the method can work

away from a local maximum, to possibly find the global maximum.

A more detailed version of this method and a general review of Bayesian inference of phylogeny can be found in the book Inferring phylogenies (Felsenstein 2004 p.288).

A program which uses the bayesian approach to phylogeny is MRBayes (Huelsenbeck and Ronquist, 2001).

2.5. Photolyase characteristics

To have an example system to consider how looking for functional changes could be done, the family of photoreceptors was chosen. For one of the latest reviews of this family see Sancar (2003). This family consists of two main groups; the photolyases, which repair DNA damage, and the cryptochromes (Cashmore *et al.* 1999), which have an effect on the physiological cycles caused by changes in light level and day/night switching. This family has been studied to some degree and the manner of the photolyase mechanism is mostly known. While the cryptochromes are not so well classified, several tests have shown at least some of the properties for this group. Combined this gives a suitable project for a functional diversification since both groups utilise light as a source of energy, but their function is to some degree, quite different. The photolyases repair the bonding of two adjoining bases in the DNA, while the cryptochromes function as transcription factors.

2.5.1. Photolyases

The photolyases consist of two distinct groups, classified by which type of damage is repaired. These two groups are based on the two types of damage that, while apparently fairly similar, requires a different mode of repair. The damage is either a double bond between the C-6 and C-5 on adjacent pyrimidines, or the single bond between the C-6 and C-4. This kind of DNA damage is called photoproducts. This is shown in figure 2.2. For a more detailed description of the repair mechanism see Carell et al. (2001).

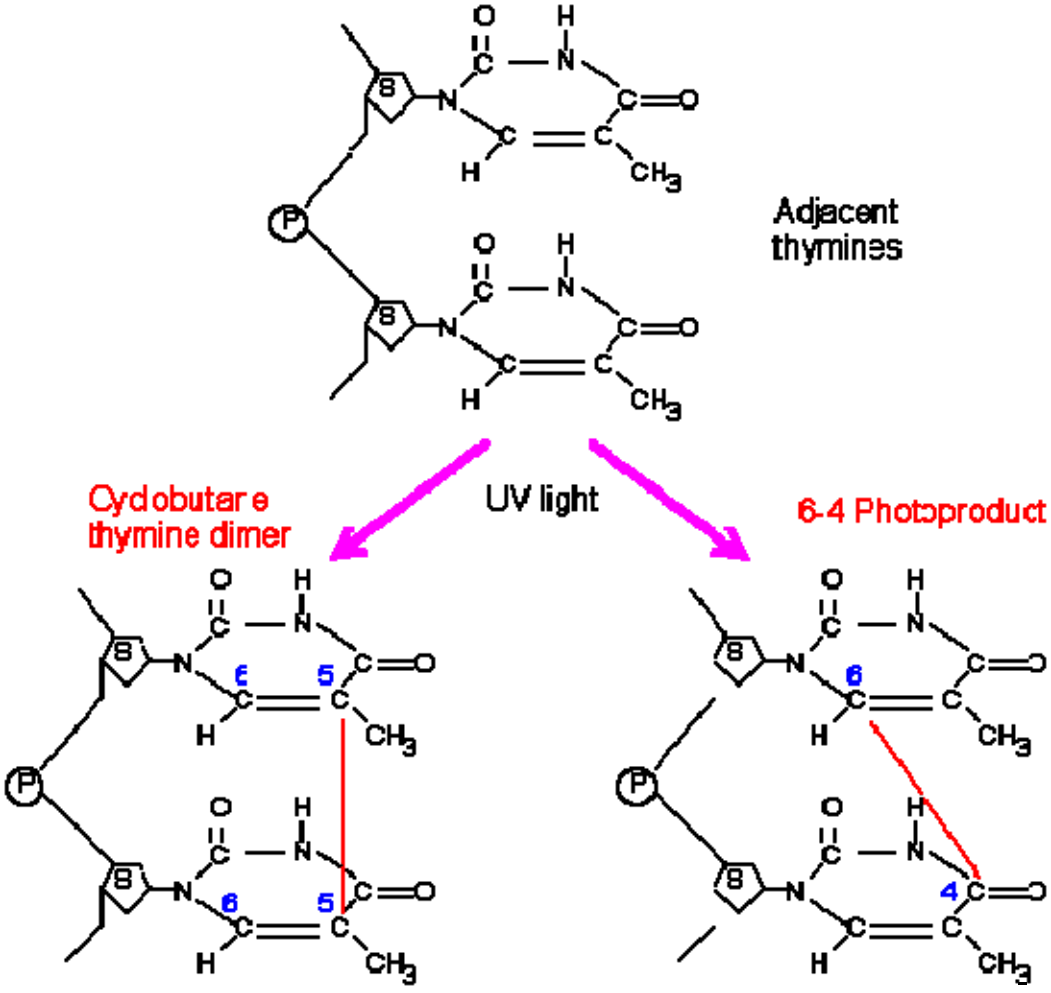


Figure 2.2: Structures of pyrimidine dimer photoproducts.

The two types of photolyase are then called cyclobutane pyrimidine dimer (CPD) photolyase and (6-4) photolyase (Todo *et al.* 1993). The CPD photolyase are further divided into two groups; class I and class II, and while the former grouping (CPD vs. (6-4)) is based on function, the latter is based on structure. As will be shown later the primary protein structure clearly shows class I and class II photolyase as two distinct evolutionary groups, in fact the class I photolyase is more closely related to (6-4) photolyase than it is to class II based only on the primary structure of the protein.

Both the CPD and the (6-4) photolyase have several things in common; both use FADH as cofactor to manage the electron transfer to and from the damaged DNA. In fact, the use of FADH is a common factor to all the photoreceptors. Additionally they both contain a second cofactor which has been shown to function as the light harvesting factor for the reaction. This second cofactor has been shown to be either 5,10-methenyltetrahydrofolate or 8-hydroxy-5-deazaflavin, depending on what type of organism coded for the protein. Last, they both bind to the affected area of the DNA prior to the electron transfer from FADH.

The actual mechanism of removing the damage however is different; the CPD damage requires only the electron transfer to the double bond, while the (6-4) damage needs a bit more rearranging. The functional difference, if any, between class I and class II photolyase is as yet unknown. Based on where the two types have been found it seems to be a case of class I appearing mostly in unicellular organisms (prokaryotes and eukaryotes), while class II mostly appears in multicellular organisms. This classification however does not cover all the observed results.

In regard to the possibility of detecting the functional differences, there are consequently three areas on the protein of special interest; the amino acids binding to the FADH cofactor, to the second cofactor and to the damaged DNA. In addition, the areas between the functional sites may also be of importance because they help the protein fold correctly and make sure the active sites are in the right position. Lastly there is also the matter of localization; if the protein is translated in a different compartment of the cell, it needs to be transported to the correct place. It has been shown that some photoreceptors have an N-terminal area coding for whether it should go to the mitochondria, chloroplasts or nucleus. In principal, any changes in one of these areas may cause the protein to stop functioning or reduce its effectiveness.

2.5.2. Cryptochromes

Cryptochromes are a group of photoreceptors first classified based on sequence homology with the photolyase group. Experiments with *Arabidopsis* have shown that defects in these proteins cause the plant to lose the ability to respond to blue light. It has been proposed that cryptochromes act as a transcription factor for the circadian clock, which is the biochemical response in organisms to varying levels of light (Lin 2002).

The term cryptochrome has been defined to mean a photolyase sequence homologue with no DNA repair activity but with blue-light-activated enzymatic functions.

The similarities to photolyase are several. Both bind FADH as a cofactor, and the cryptochrome found in humans have been shown to be able to bind MTHF (Özgur and Sancar 2003). Both proteins bind to DNA (Kleine *et al.* 2003) although with a slightly different preference when it comes to double stranded vs. single stranded DNA. The primary protein structure is also very similar, the cryptochromes clusters on some occasions closer together with the photolyases than the different photolyases do to each other. The main difference is in the function since while both proteins have the ability to bind to DNA, only the photolyase is capable of repairing the pyrimidine

dimer product. One notable difference is that the cryptochromes seem to have a C-terminal extension that is not present in the photolyases.

A third cryptochrome class has also been discovered, called DASH cryptochromes. This group seems to appear together with the (6-4) photolyase and the animal cryptochromes, and has the ability to bind to DNA and shows also some transcription control (Brudler *et al.* 2003). Differences in the cryptochromes were also shown with cryptochromes from zebrafish (Kobayashi *et al.* 2000). One of the cry genes were shown to have inhibitor activity for the circadian clock system, while the other did not.

2.5.3. Important residues for photolyase activity

Based on crystallographic analysis, the 3D structure of the class I CPD photolyases from *Escherichia coli* (Park *et al.* 1995) and *Thermus thermophilius* (Komori *et al.* 2001) have been described. From the *E.coli* structure, the FADH binding site has been shown to include 12 amino acids, located in the downstream part of the enzyme. Also, a triple tryptophan chain has been shown to mediate the electron transfer to the FADH molecule (Byrdin et al. 2003). This involves the three tryptophans W^{306}, W^{359} and W^{382}. From W^{382} the electron is transferred to the FADH molecule.

From (6-4) photolyase, two histidines, H^{354} and H^{358}, have been proposed to be important to the repair mechanism (Hitomi *et al.* 2001). It was shown that the enzyme lost almost all repair activity when these histidines were substituted with alanine. The numbers refer to the position these amino acids have in the protein from *Xenopus laevis*.

3. Materials and methods

3.1. Sequences

From the photolyase/blue-light photoreceptor family, sequences from all the most interesting groups were searched for in the NCBI database (http://www.ncbi.nlm.nih.gov/). This included the class I and II of the CPD photolyase, the cryptochromes (both animals and plants) and the 6-4 photolyase. The sequences were either discovered by searching the database for the different groups, or via references in several articles on the function of various isolated photolyase genes from different species. The sequences are shown in table 3.1.

The sequences were all downloaded from the NCBI sequence database, and the accession number, name and classification are shown in table. The classification is based on information from the database, or research articles where available. This is noted under the reference column.

Due to some restrictions in the programs used where the name has a limit of 10 letters, the sequences have gotten an ID to be used later in the analysis. This ID is made up as follows:
C1m: class I CPD photolyase with MTHF,
C18: class I CPD photolyase with 8-HDF,
C2: class II CPD photolyase,
64: 6-4 photolyase,
Cry: cryptochromes.

The next letter (E, B or A) indicates whether the gene came from respectively an eukaryote, a bacteria or an archebacteria. The last number is an index when more than one sequence have a similar ID.

The gene sequence from *Sinapis alba* have been shown to give a functional repair protein (Batschauer, 1993). However, other papers (Kanai, *et al.* 1997; Kobayashi, *et al.* 2000), lists it as a cryptochrome. The ID of this sequence is therefore slightly different from the rest to make it easier identifiable.

Table 3.1: The sequences downloaded from the NCBI database.

ID	*Name*	*Classification*	*Reference*	*NCBI acc. Nr.*
C1mB_1	*Escherichia coli*	ClassI, MTHF	Yes	P00914
C1mB_2	*Salmonella typhimurium*	ClassI, MTHF	Yes[15]	P25078
C1mE_1	*Saccharomyces cerevisiae*	ClassI, MTHF	Yes	P05066
C1mE_2	*Neurospora crassa*	ClassI, MTHF	Yes	P27526
C18B_1	*Synechococcus leopoliensis*	ClassI, 8-HDF	Yes	P05327
C18B_2	*Synechocystis sp.*	ClassI, 8-HDF	No*	AAB81109.1
C18B_3	*Streptomyces griseus*	ClassI, 8-HDF	Yes	P12768
C18A_1	*Halobacterium sp.*	ClassI, 8-HDF	Yes	NP_280191.1
64E_1	*Drosophila melanogaster*	6-4	Yes[3]	BAA12067.1
CryE_1	*Homo sapiens*	Cry	Yes[3]	BAA12068.1
64E_2	*Aphrocallistes vastus*	6-4	Yes[2]	CAD24679.1
CryE_2	*Homo sapiens*	Cry	No	NP_066940.1
64E_3	*Danio rerio*	6-4	Yes[1]	BAA96852.1
64E_4	*Arabidopsis thaliana*	6-4	Yes[4]	BAA34711.1
64E_5	*Xenopus laevis*	6-4	Yes[5]	BAA97126.1

ID	*Name*	*Classification*	*Reference*	*NCBI acc. Nr.*
CryE_3	*Gallus gallus*	Cry	Yes[6]	AAK61386.1
CryE_4	*Drosophilia melanogaster*	Cry	Yes[7]	BAA35000.1
CryE_5	*Danio rerio*	Cry	Yes[1]	NP_571864.1
CryE_6	*Chlamydomonas reinhardtii*	Cry	No	AAC37438.2
CPDE_1	*Sinapis alba*	CPD	Yes[13]	P40115
CryB_1	*Synechocystis sp.*	Cry	Yes[14]	S74805
CryE_7	*Arabidopsis thaliana*	Cry	No	Q96524
CryB_2	*Mesorhizobium loti*	Cry	No	NP_108272.1
CryB_3	*Xanthomonas campestris*	Cry	No	NP_636808.1
CryB_4	*Bradyrhizobium japonicum*	Cry	No	BAC50575.1
CryE_8	*Oryza sativa*	Cry	No	BAA82885.1
CryE_9	*Physcomitrella patens*	Cry	No	BAA83338.1
C2A_1	*Methanothermobacter thermautotrophicus*	ClassII	Yes[10]	P12769
C2E_1	*Drosophilia melanogaster*	ClassII	Yes[10]	S52047
C2E_2	*Carassius auratus*	ClassII	Yes[10]	A45098
C2E_3	*Oryzias latipes*	ClassII	Yes[10]	BAA05043.1
C2E_4	*Monodelphis domestica*	ClassII	Yes[11]	BAA06700.1
C2E_5	*Potorous tridactylus*	ClassII	Yes[10]	BAA05041.1
C2E_6	*Chlamydomonas reinhardtii*	ClassII	Yes[8]	AAD39433.1
C2E_7	*Spinacia oleracea*	ClassII	No	AAP31407.1
C2E_8	*Arabidopsis thaliana*	ClassII	Yes[12]	AAC08008.1
C2E_9	*Cucumis sativus*	ClassII	No	BAB91322.1
C2E_10	*Oryza sativa*	ClassII	Yes[9]	BAC76449.1

1) Kobayashi *et al.* (2000), 2) Schroder *et al.* (2003), 3) Todo *et al.* (1996), 4) Nakajima *et al.* (1998), 5) Todo *et al.* (1997), 6) Yamamoto *et.al* (2001), 7) Ishikawa *et.al* (1999), 8) Petersen *et al.* (1999), 9) Hirouchi *et al.* (2003), 10) Yasui *et al.* (1994), 11) Kato *et al.* (1994), 12) Ahmad *et al.* (1997) 13) Batschauer (1993) 14) Brudler *et al.* (2003) 15) Li YF, Sancar A. (1991) *) Kanai *et al.* (1997) indicates an 8-HDF cofactor

When all the sequences were ready, the next step was to make a multiple alignment and subsequently a phylogenetic tree, showing the evolution of the gene. Based on this tree, the sequences were analysed to find the sites which were important for both the topology of the tree, and the functional changes. These results were correlated to the known information about the photolyase family.

3.2. Computer programs

To do the global alignment, the Clustal program (Higgins & Sharp, 1988) was used. This can be found on the internet (http://www.ebi.ac.uk/clustalw/), or may also be downloaded and run from a local computer (ftp.embl-heidelberg.de). In this project the local version were used; ClustalX v1.81 (Thompson *et al.* 1997). The algorithm ClustalX uses is in short to first do a pairwise alignment between all sequences, and based on the calculated distance matrix generate a tree. This tree is then used as a basis for the multiple alignments. The two sequences with the smallest distance is aligned together, then the rest of the cluster, as indicated by the preliminary tree, is added, one sequence at the time adjusting the previously aligned sequences with gaps together as a block. Any sequences showing less than a preset value of homology to its group are delayed until the end. This value is by default set to 30%, but may be changed by the user. When the program is set to return all sequences in the order they were added, the delayed sequences can be seen at the bottom of the sequence order. In this way they can easily be identified and taken care of, if needed.

The subsequent phylogeny analysis was done with PHYLIP v3.6 (Felsenstein 2004) and Treepuzzle (Strimmer and von Haesler, 1996). The programs PAUP (http://paup.csit.fsu.edu/) and Molphy (Adachi and Hasegawa 1996) were used briefly to check the results of the phylogenetic analysis.

The editing and visualization of the sequences were done in the BIOEDIT (http://www.mbio.ncsu.edu/BioEdit/bioedit.html) program while the trees were drawn in Treeview (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html) and edited in the photoeditor of StarOffice.

To search for the specific sites in the alignment, a small program was written in Java.

The first alignment was done with the gap parameters set to its default values as used by ClustalX. These were a gap open/gap extension cost of 10/0.1 and 10/0.2 for the pairwise and multiple aligning respectively. The second alignment shows what happens when the penalty of introducing a gap is raised to its maximum value (100/1 and 100/2). The cost for the extension of a gap is scaled with the same factor as the cost for opening a gap. Then the first alignment was chosen to be the most representative and used for further analysis, unless otherwise stated.

After the alignment, the next step was to generate a phylogenetic tree of the sequences. This was done with the Trepuzzle program for both the generated alignments from above. The same was done using the NJ program (NEIGHBOR) with the JTT protein substitution model (Jones *et al.* 1992) from the PHYLIP package. The alignment was also analysed with the bootstrap method from PHYLIP using the programs SEQBOOT, PROTDIST, NEIGHBOR and CONSENSE.

The phylogenetic tree showed one interesting detail, the sequence from *S. alba* (CPDE_1) which has photolyase activity, is grouped together with the plant cryptochromes. Since none of the proposed plant cryptochromes gave any references to biological analysis of their function, an extra cryptochrome sequence from *A.thaliana* (accession number NP-568461), with no photolyase activity (Kleine et al. 2003) were acquired. This sequence was then given the ID CryE_10 and used in a phylogenetic analysis together with the class I photolyase and plant cryptochromes.

Based on the groups found from the phylogenetic analysis and the known functionality as presented in the sequence database, the sequence where first analysed to find the needed stringency for correctly predicting the important residues for the photolyase function.

Then the sites that best corresponded with the phylogenetic clustering were then searched for. To help with the analysis of the results, the N-terminal and C-terminal ends were shortened to where most of the sequences had characters. These parts were mostly very poorly conserved. The analysed segment of the multiple alignment was between and including position 200 and position 771 of the original alignment.

3.3. Laboratory work

In addition to the theoretical part with analysing sequences from the sequence database, and attempt was made to find the photolyase gene in cotton. By using primers designed based on conserved areas of known photolyase sequences, part of the gene were to be amplified with the PCR reaction, and potentially sequenced to try and observe any differences in the gene sequence.

3.3.1 Primers

Primers are constructed based on the information from mainly *Oryza sativa* (rice), and *Arabidopsis*

thaliana. In addition *Cucumis sativus* (cucumber) and *Spinaccia olacera* (spinach) were used to get an idea of the conservation of the chosen primer sites. Four primer pairs were designed; three based on the rice sequence and one from the *Arabidopsis* sequence. One of the rice primers were taken from an article on photolyase from rice (Hirouchi *et al.*, 2003).

The primer sequences were as follows:
Rice:

HG11 5'-CTAGAATGGACTAGTGGACCAGAAGA-3' fw
HG12 5'-ACTTGGAAAGCTTGACTACAGGATT-3' rw

HG21 5'-GCTGATAAAAGAGAGCATATCTATACGAG-3' fw
HG22 5'-AATTCATGTAACGTATCTTTCCAAATACTG-3 rw

Arabidopsis:

HG31 5'-ACTAATGGATCATGCTTCAGATAAGAG-3' fw
HG32 5'-ATATAGACCACATGCACCCAACATA-3' rw

The fourth primer pair was designed by Hirouchi, *et al.* (2003):

HG41 5'-GCGTCGGCGAAGATGGAGTAT-3' fw
HG42 5'-CATCTCCAACTGCGATGCATTCCA-3' rw

All primers were manufactured by Invitrogen.

3.3.2. DNA templates

The first PCR was done with Arabidopsis as a control, tomato and cotton. Later, maize, cactus, barley were also tried.

3.3.3. Expected PCR-product

The expected product is a part of the gene encoding the class II CPD photolyase. Due to having the primers anchored in conserved areas of the gene sequence, only a part of the gene were to be amplified. This part covers about 100-300 amino acids of the C-terminal end of the full gene. Crystallographic analysis of a similar gene from E.coli have shown this to be the area of DNA binding and also FADH binding. It is also the part of the gene that shows the most conservation within and between species, in addition to be least riddled by gaps during alignment.

3.3.4. PCR

A standard PCR was performed with reactants as described in table 3.1.

Table 3.1: Reactants for the PCR reaction

Name	Stock concentration	Final concentration
Buffer	10x	1x
dNTP	10 mM	0,2 mM
Primer HG31 (forward)	10 µM	0,5 µM
Primer HG32 (reverse)	10 µM	0,5 µM
Taq (Fermentas #EP0405)	5 U/µl	1,5 U

$MgCl_2$	25 mM	1,5 mM
dH_2O	*	*

*) Distilled water was added to adjust the final volume.

The PCR reactions were run on a PTC-200 (Peltier Thermal Cycler, MJ Research) with the program shown in table 3.2.

Table 3.2: PCR program run on the PTC-200

	Temperature [ºC]	Time [minutes]	Cycles
Pre-denaturation	95	5	1
Denatruation	94	1	35
	50	1	
Elongation	72	1	
Final elongation	72	10	1
Standby	4	Forever	1

The first PCR reaction was run with *Arabidopsis*, tomato and cotton in a 10 µl reaction (8 µl PCR mix + 2 µl template). The amount of template DNA was as follows: *Arabidopsis*; 50 ng, 10 ng and 1 ng, tomato; 50 ng and 10 ng, cotton; 100 ng and 50 ng.

The next reaction was done with the same parameters but with some additional species and also with the rest of the primers. The species, their amount of template DNA and the primer combination were; Arabidopsis (1 ng DNA, primer pair 2, 3 and 4), tomato (50 ng, primer pair 2, 3 and 4), cotton (50 ng, primer pair 2, 3 and 4), maize (~200 ng, primer pair 2 and 3), cactus (20 ng, primer pair 2, 3 and 4), Einkorn (20 ng, primer pair 1, 2, 3 and 4), barley (12 ng, primer pair 1, 2, 3, 4) and wheat (10 ng, primer pair 2, 3 and 4).

After the initial PCR reactions, it was decided to try and sequence the bands from *Arabidopsis*, cotton and tomato. Another PCR reaction was set up the same way as before, with four parallels of *Arabidopsis* (1 ng, primer pair 3) and five parallels each of cotton and tomato (50 ng, primer pair 3). Selected bands were removed from the gel and the DNA was extracted with the QIAquick Gel Extraction kit (Cat.no. 28704).

3.3.5. Gel electrophoresis

The PCR products were separated on a small (40 ml) 1.2% agarose gel (40 ml 1x TBE, 4,8g agarose, 0.5 µl EtBr), run at approximately 45 minutes at 80 V corresponding to a current of about 40-45 mA.

3.3.6. Sequencing

The sequencing PCR was performed with the BigDye Terminator kit (Applied Biosystems) version 2.0, using a dilution of ¼ in the PCR reaction.

The final sequencing was done at IKBN.

4. Results

4.1. Laboratory work results

The results from the first PCR are shown in figure 4.1.

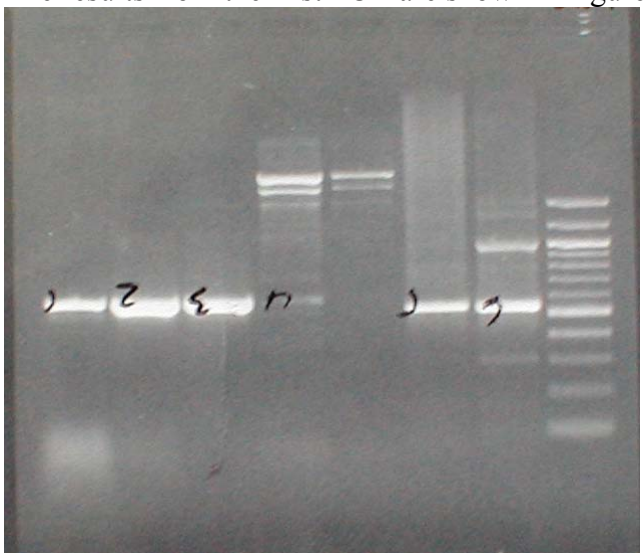


Figure 4.1: PCR products with primer pair HG31 and HG32. The lanes consist of, from left to right; Arabidopsis (50, 10 and 1 ng DNA template), tomato (50 and 10 ng DNA template), cotton (100 and 50 ng DNA template) and a 100 bp ladder.

Arabidopsis shows the expected band size in all tried dilutions of the template DNA. Both cotton and tomato have gotten bands which may be the wanted results.

The next reaction where more species were added and the number of primer combination increased, showed the same results as the first reaction in addition to positive indications from cactus and maize.

The different DNA amplified for sequencing are shown in figure 4.3.

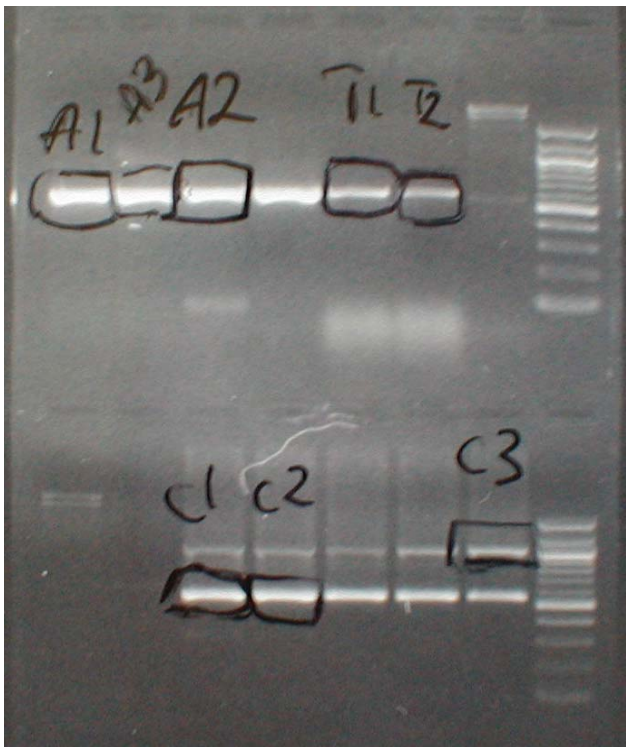


Figure 4.3: PCR with *Arabidopsis* (lanes 1-4), tomato (lanes 5-7 and 9-10) and cotton (lanes 11-15). Bands cut out for sequencing analysis are indicated.

Here the tomato sample has gotten a band at the same size as Arabidopsis, in addition to the previously discovered bands.

The control reactions from Arabidopsis matched perfectly with the original sequence from the database. There were some initial variation but a closer inspection of the raw data, as exemplified in figure 4.4, removed this variation.

The PCR product for tomato showed initially a slight variation to Arabidopsis. A closer look at the some of the ambiguous data points from the raw data, as shown in figure 4.4, resulted in the assumed product from tomato being identical to the sequence extracted from Arabidopsis.

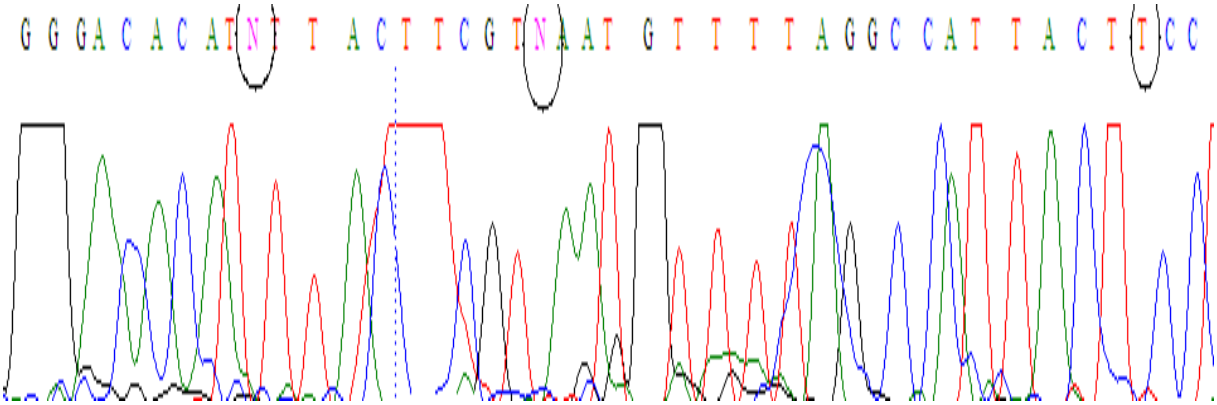


Figure 4.4: Raw data from sequencing reaction of PCR product from tomato. None of the letters N, N and T (marked) have any support in the raw data and are removed. The third G in the beginning is quite probably also not supported.

The sequencing results from the cotton bands had too low signal strength to make any conclusions from as shown in figure 4.5.

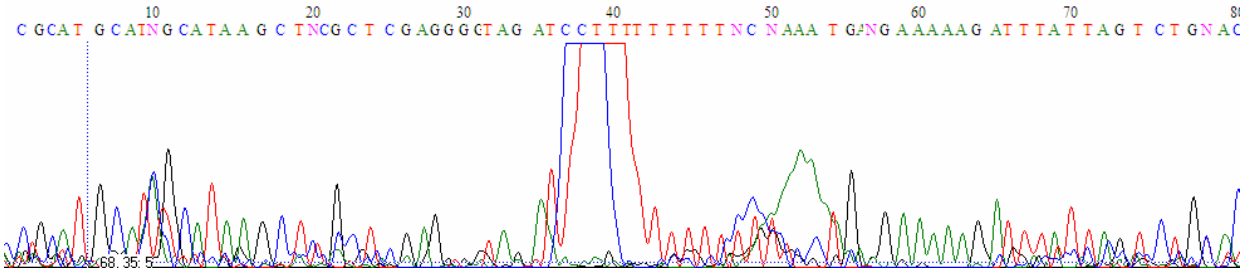


Figure 4.5: The sequencing results from cotton band C3.

4.2. Computer part

The results of the first alignments with differing gap cost parameters are shown in figure 4.6 and 4.7.

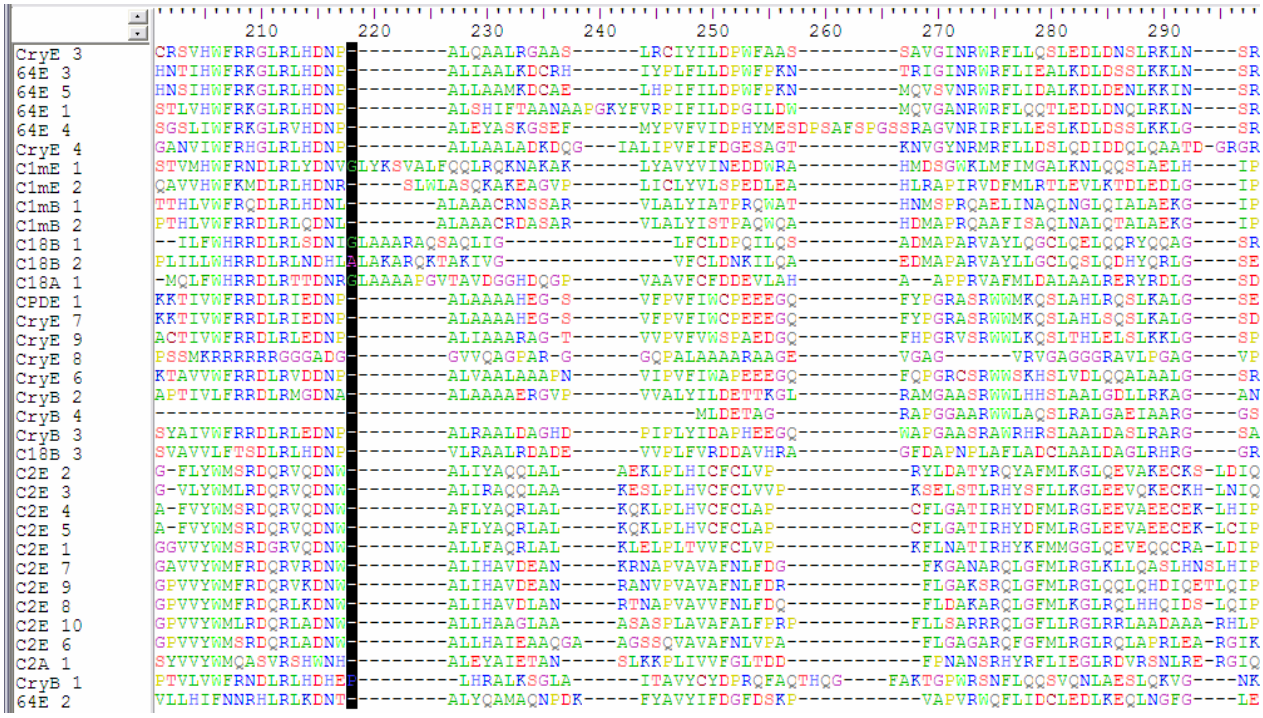


Figure 4.6: Multiple alignment of the downloaded sequences with parameters set to default values; pairwise gap open/gap extension: 10/0.1 and multiple gap open/gap extension 10/0.2. The figure includes the first 50-60 amino acids of the gene.

Figure 4.7: Multiple alignment of the downloaded sequences with the gap open parameters set to maximum values; pairwise gap open/gap extension: 100/1 and multiple gap open/gap extension

100/2. The figure includes the first 50-60 amino acids of the gene.

The black line in each figure marks the end of what in the first figure is a rather conserved area. The leftmost point of both figures is at the beginning of the conserved area.

Towards the C-terminal end of the alignment, the conservation was much better and the gaps were fewer and smaller as shown in figure 4.8.

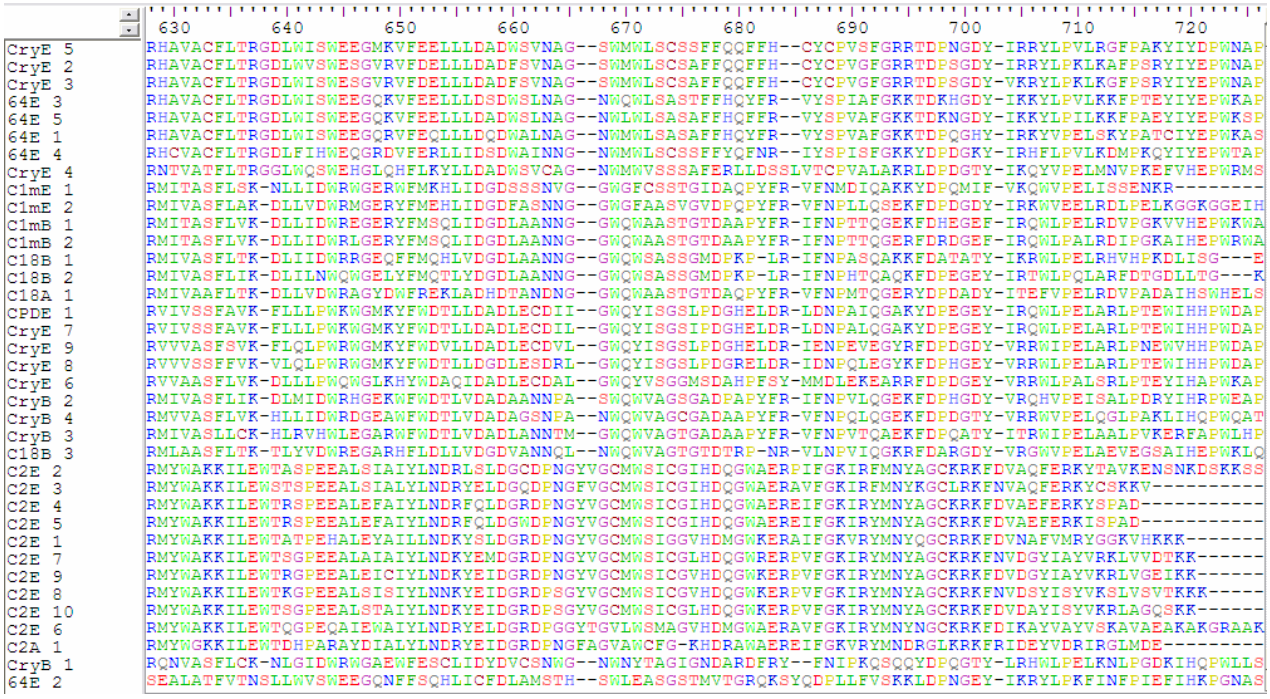


Figure 4.8: The C-terminal end of the alignment shown in figure 1 with default stringency in gap placement.

The full multiple sequence alignment from postion 200 to 771 are shown in figure 4.9.

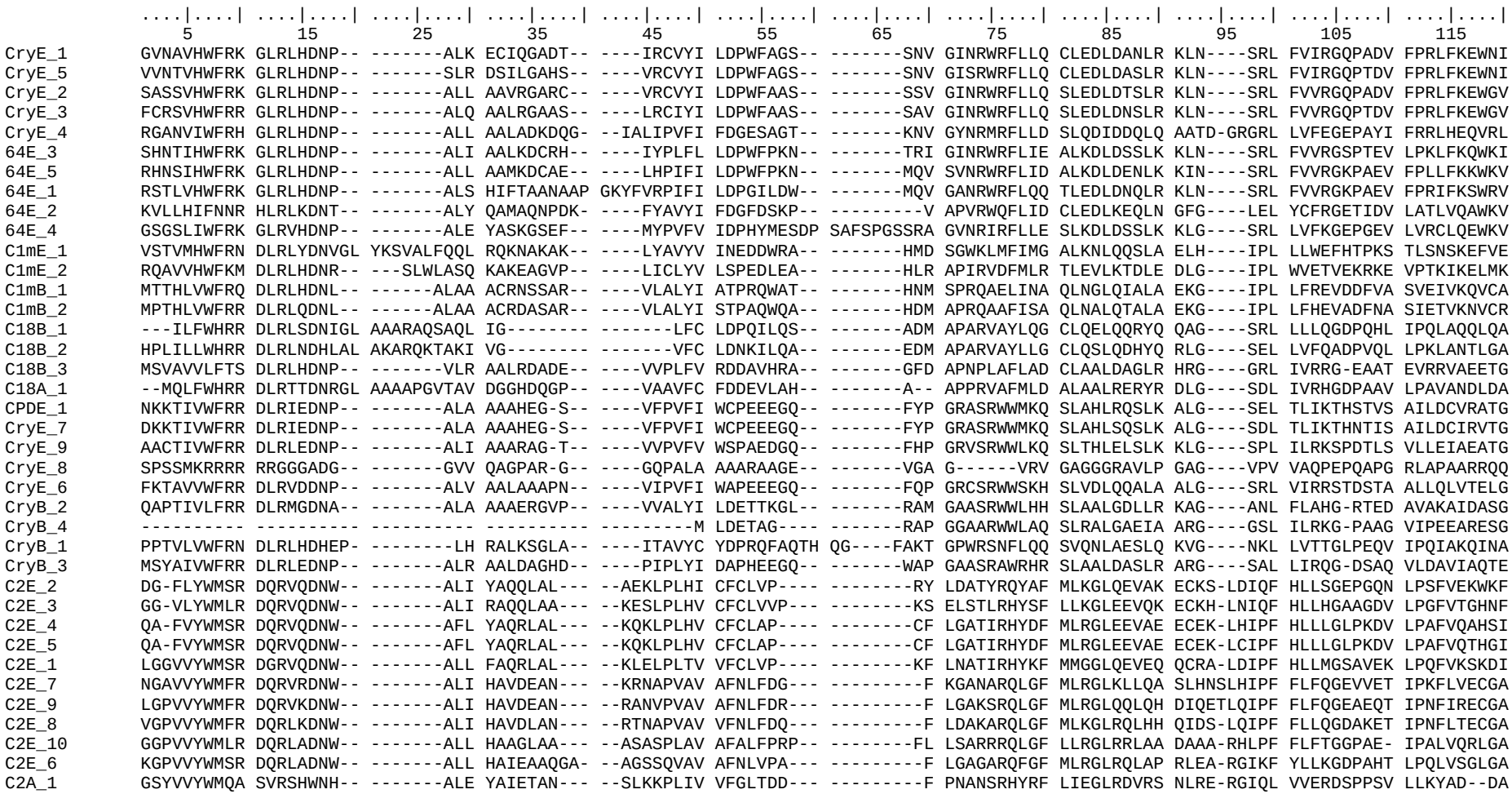


Figure 4.9: Multiple alignment of the 38 sequences from the photolyase/blue-light photoreceptor family

```
                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   125        135        145        155        165        175        185        195        205        215        225        235
CryE_1          TKLSIEY--- ---------- -DSEPFGKER -DAAIKKLAT EA----GVEV IVRISHTLYD --LDKIIELN GGQPPLTYKR FQTLISKMEP LEIP------ -VETITSEVI EKCTT-PLSD
CryE_5          NRLSYEY--- ---------- -DSEPFGKER -DAAIKKLAN EA----GVEV IVRISHTLYD --LDKIIELN GGQSPLTYKR FQTLISRMEA VVTP------ -AETITAEVM GPCTT-PLSD
CryE_2          TRLTFEY--- ---------- -DSEPFGKER -DAAIMKMAK EA----GVEV VTENSHTLYD --LDRIIELN GQKPPLTYKR FQAIISRMEL PKKP------ -VGLVTSQQM ESCRA-EIQE
CryE_3          TRLTFEY--- ---------- -DSEPFGKER -DAAIIKLAK EA----GVEV VIENSHTLYD --LDRIIELN GNKPPLTYKR FQAIISRMEL PKKP------ -VSSIVSQQM ETCKV-DIQE
CryE_4          HRICIEQ--- ---------- -DCEPIWNER -DESIRSLCR EL----NIDF VEKVSHTLWD --PQLVIETN GGIPPLTYQM FLHTVQIIGL PPRPT----- -ADARLEDAT FVELDPEFCR
64E_3           TRLTFEV--- ---------- -DTEPYSQSR -DKEVMKLAK EY----GVEV TPKISHTLYN --IDRIIDEN NGKTPMTYIR LQSVVKAMGH PKKP------ -IPAPTNEDM RGVST-PLSD
64E_5           TRLTFEV--- ---------- -DIEPYSRQR -DAEVEKLAA EH----DVQV IQKVSNTLYD --IDRIIAEN NGKPPLTYVR FQTVLAPLGP PKRP------ -IKAPTLENM KDCHT-PWKS
64E_1           EMLTFET--- ---------- -DIEPYSVTR -DAAVQKLAK AE----GVRV ETHCSHTIYN --PELVKAKN LGKAPITYQK FLGIVEQLKV PKVLG----- -VPEKLKKMP TPPKD-EVEQ
64E_2           KLLSINMD-- ---------- ---PDVNFTF FNEKIVKMCT IN---AVQLY NDMDSHRLLY LPPKYKSAIP MSKFRVLLAE AITAKQNNLE SEAKIQDITP ------PLNP EQLSDLGNKP
64E_4           KRLCFEY--- ---------- -DTDPYYQAL -DVKVKDYAS ST----GVEV FSPVSHTLFN --PAHIIEKN GGKPPLSYQS FLKVAGEPSC AKSEL----- -VMSYSSLPP IGDIG-NLG-
C1mE_1          FFKEKCMNVS SGTGTIITAN IEYQTDELYR -DIRLLE--- --NEDHRLQL KYYHDSCIVA ---PGLITTD RGTNYSVFTP WYKKWVLYVN NYKKSTSEI- ------CHLH IIEP---LKY
C1mE_2          SWGASHLFCA ---------- MEYEVDELRR -EAKLVKLLA EGEKGEKMAA DVVHDTCVVM ---PGALQSG SGGQYAVYSP WFRAWIKHIE ENPECLEIY- ------EKPG PNPPGTKEKH
C1mB_1          ENSVTHLFYN ---------- YQYEVNERAR -DVEVER--- ---ALRNVVC EGFDDSVILP ---PGAVMTG NHEMYKVFTP FKNAWLKRLR EG-------- ------MPEC VAAP----KV
C1mB_2          QHDVSHLFYN ---------- YQYEFNERQR -DAAVEK--- ---TLPSVIC EGFDDSVILA ---PGAVMTG NHEMYKVFTP FKNAWLKRLK ED-------- ------IPPC VPAP----KI
C18B_1          EAVYWN---- ---------- QDIEPYGRDR -DGQVAAALK ----TAGIRA VQLWDQLLHS ---PDQILSG SGNPYSVYGP FWKNWQAQPK PT-------- ------PVAT PTELVDLSPE
C18B_2          HGVTWT---- ---------- LDTEPYAQKR -DLAVAQALR ----ERGLAI ATEWDQLMHH ---PGEVLTQ AGSPYTVYTP FWKNWSQLPK TS-------- ------PVPT PKDLQGLTPA
C18B_3          AARVHIA--- ---------- AGVSRYAARR -EQRIREALA DS----GREL HVHDAVVTAL AP-GRVVPTG GKDHFAVFTP YFRRWEAEGV RGTQT----- ---------- -----APRTV
C18A_1          TRVVWN---- ---------- HDYSGLATDR -DAGVRDALD ----AAGVAH AQFHDAVHHR ---PGEIRTN AGDPYSVYTY FWRKWQDREK NP-------- ------PAPE P-EPADLAAD
CPDE_1          ATKVVFN--- ---------- HLYDPVSLVR -DHTVKEKLV ER----GISV QSYNGDLCMS ---PGRYTVK RANLLLVLIL TGKKCLDMSV ESVVL----- -------PPP WRLMPLS-AA
CryE_7          ATKVVFN--- ---------- HLYDPVSLVR -DHTVKEKLV ER----GISV QSYNGDLLYE ---PWEIYCE KGKPFTSFNS YWKKCLDMSI ESVML----- -------PPP WRLMPITAAA
CryE_9          ATQVFYN--- ---------- HLYDPVSLVR -DHRVKQGLS QR----GIVV HTFNGDLLYE ---PWEVYDE EGQAFTVYEA FWKKCMSMPF EPEAP----- -------LLP PR--RLTGPI
CryE_8          ARHPPLR--- ---------- RRRRRAHRAR PQHRRHASLL QPPLRPAVAG EGPPGEGAAD GRGHRRAVVQ RRPAVRAMGG GRRRRLPVHH VRAVLGQVPV HARPGGALLP PK--RIAPGE
CryE_6          AEAVFFN--- ---------- HLYDPISLMR -DHDCKRGLT AA----GVAH RTFNGDMLYE ---PWDVLDP NKQPYSTFDD FWNSVRAMPV PPPFP----- -------VSA PASMPAVPAA
CryB_2          ANCVFWN--- ---------- RRYDPSEVAV -DARLKAALR EK----GLTA LSFDGALLHE ---PSLLKTG SGGFYKVFTP FWKAMADKVD VRDPI----- ---------- DTPGQIDGWR
CryB_4          ARAVYWN--- ---------- GIAQAPHQAI -ERRLEAALA KL----GVDS QSFPGDLLVP ---PSAIRNK EGRGLRVFTP FWRRVLSLGD PPKPL----- ---------- PAPKQLRPGP
CryB_1          KTIYYHR--- ---------- ------EVTQ EELDVERNLV KQLTILGIEA KGYWGSTLCH ---PEDLPFS IQDLPDLFTK FRKDIEKKKI SIRPCF---- ---------- --FAPSQLLP
CryB_3          AVAVYWN--- ---------- RKYEPATQPR -DAQIKRSLR ER----GLEV QSCNAALLFE ---PWTLATQ QGRPYKVFTP FWRNALTQLR LPDAM----- ---------- PAPRSLPPLP
C2E_2           GAVVTDF--- ---------- ---NPLRIPL -QWIETVK-- -KHLPADVPF IQVDAHNVVP ---CWEASGK LEYGARTIRG KITKLLPEFL TEIPLV---- ---------- --DTHPH-SA
C2E_3           GAVVTDF--- ---------- ---SPLREPL -QWLEAVK-- -KGLPEDIPF IQVDAHNIVP ---CWVASPK LEYSARTIRG KITNLLSEFL TDFPLV---- ---------- --DKHPF-SA
C2E_4           GGIVTDF--- ---------- ---SPLLHHT -QWVKDVQ-- -DGLPKQVPF VQVDAHNIVP ---CWIASDK QEYGARTIRH KIHDRLPHFL TEFPPV---- ---------- --ICHPY-PS
C2E_5           GGIVTDF--- ---------- ---SPLLHHT -QWVKDVQ-- -DALPRQVPF VQVDAHNIVP ---CWVASDK QEYGARTIRH KIHDRLPHFL TEFPPV---- ---------- --ICHPY-TS
C2E_1           GAVVCDF--- ---------- ---APLRLPR -QWVEDVG-- -KALPKSVPL VQVDAHNVVP ---LWVASDK QEYAARTIRN KINSKLGEYL SEFPPV---- ---------- --VRHPHGTG
C2E_7           SLLVTDF--- ---------- ---TPLREIR -GFKEELC-- -KRVGDSVSI HEVDAHNVVP ---VWEASSK LEYGARTIRT KINKLLPTYL TDYPILQ--- ---------- --PPNCSWE-
C2E_9           SLLVTDF--- ---------- ---SPLREVR -KCKEEIC-- -KRVEESVKV HEVDAHNVVP ---TWVASEK LEYSAKTLRG KINKKLPDYL IDYPSMV--- ---------- --IPTRKWP-
C2E_8           SHLVTDF--- ---------- ---SPLREIR -RCKDEVV-- -KRTSDSLAI HEVDAHNVVP ---MWAASSK LEYSARTIRG KINKLLPDYL IEFPKLE--- ---------- --PPKKKWTG
C2E_10          STLVADF--- ---------- ---SPLRPVR -EALDAVVGD LRREAPGVAV HQVDAHNVVP ---VWTASAK MEYSAKTFRG KVSKVMDEYL VEFPEL---- ---------- --PAVVPWDR
C2E_6           GLLVTDY--- ---------- ---SPLRLGR -TWRDQVC-- --SALGSVPV HEVDAHNVVP ---VWAASEK REVGARTLRP KIHKALPEFL REFPEV---- ---------- --PTLPAWTP
C2A_1           AAAVTDR--- ---------- ---GYLDIQK -EWVDEAAG- ----ALHIPL TQVESNVIVP ---VETASDK EEYSAGTFKP KIKRHLKRFM VPLRMR---- ---------- ----TLKMDS
```

```
                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   245        255        265        275        285        295        305        315        325        335        345        355
CryE_1          DHDEKYGVPS LEELGFDTDG L--SSAV-WP GGETEALTRL ER--HLERKA WVANFERPRM NANS-LLASP TGLSPYLRFG CLSCRLFYFK LTDLYKKVKK NSSPP----- --LSLYGQLL
CryE_5          DHDEKFGVPS LEELGFDTEG L--SSAV-WP GGETEALTRL ER--HLERKA WVANFERPRM NANS-LLASP TGLSPYLRFG CLSCRLFYFK LTDLYRKVKK NSSPP----- --LSLYGQLL
CryE_2          NHDETYGVPS LEELGFPTEG L--GPAV-WQ GGETEALARL DK--HLERKA WVANYERPRM NANS-LLASP TGLSPYLRFG CLSCRLFYYR LWDLYKKVKR NSTPP----- --LSLFGQLL
CryE_3          NHDDVYGVPS LEELGFPTDG L--APAV-WQ GGETEALARL DK--HLERKA WVANYERPRM NANS-LLASP TGLSPYLRFG CLSCRLFYYR LWELYKKVKR NSTPP----- --LSLYGQLL
CryE_4          SLKLFEQLPT PEHFNVYGDN MGFLAKINWR GGETQALHLL DERLKVEQHA FERGFYLPNQ ALPN-IHDSP KSMSAHLRFG CLSVRRFYWS VHDLFKNVQL RACVRGVQMT GGAHITGQLI
64E_3           DHEEKFGIPT LEDLGLDTSS L--GPHL-FP GGEQEALRRL DE--HMERTN WVCKFEKPKT SPNS-LIPST TVLSPYVRFG CLSARTFWWR LADVYR-GKT HSDPP----- --VSLHGQLL
64E_5           SYDEKYGVPT LEELGQDPMK L--GPHL-YP GGESEALSRL DL--HMKRTS WVCNFKKPET EPNS-LTPST TVLSPYVKFG CLSARTFWWK IADIYQ-GKK HSDPP----- --VSLHGQLL
64E_1           KDSAAYDCPT IKQLVKRPEE L--GPNK-FP GGETEALRRM EE--SLKDEI WVARFEKPNT APNS-LEPST TVLSPYLKFG CLSARLFNQK LKEIIKRQPK HSQPP----- --VSLIGQLM
64E_2           RLDSPLPSEI PKLNALFTEE EIAKLNFIFQ GGERRTEDYL NE----YREA RLRDVSGDED ASPIAAK--A MGISPHLRFG CITPRHLFNF LVKTIKDANY SRIKIN---- ---KVLAGIM
64E_4           ----ISEVPS LEELGYKDDE Q--ADWTPFR GGESEALKRL TK--SISDKA WVANFEKPKG DPSAFLKPAT TVMSPYLKFG CLSSRYFYQC LQNIYKDVKK HTSPP----- --VSLLGQLL
C1mE_1          NETFELKPFQ YSLPDEFLQY IPKSKWCLPD VSEEAALSRL KD---FLGTK -SSKYNNEKD MLYLGG---T SGLSVYITTG RISTRLIVNQ AFQSCNGQIM SKALK--DNS STQNFIKEVA
C1mE_2          ENLFACSIPE APEGKRLRDD EKARYHSLWP AGEHEALKRL EK---FCDEA -IGKYAERRN IPAMQG---T SNLSVHFASG TLSARTAIRT ARDRNN---T KKLNG--GNE GIQRWISEVA
C1mB_1          RSSGSIE-PS PSITLNYPR- -QSFDTAHFP VEEKAAIAQL RQ---FCQNG -AGEYEQQRD FPAVEG---T SRLSASLATG GLSPRQCLHR LLAEQP---- -QALD--GGA GS-VWLNELI
C1mB_2          RVSGALSTPL TPVSLNYPQ- -QAFDAALFP VEENAVIAQL RQ---FCAQG -ADEYALRRD FPAVDG---T SRLSASLATG GLSPRQCLHR LLAEQP---- -QALD--GGP GS-VWLNELI
C18B_1          QLTAIAPLLL SELPTLKQLG FDWDGGFPVE PGETAAIARL QE---FCDRA -IADYDPQRN FPAEAG---T SGLSPALKFG AIGIRQAWQA ASAAHA---L SRSDE--ARN SIRVWQQELA
C18B_2          EKEKLAPLEP LAIPQLADLG FIWDQPLPLT PGEEAAEQRL DW---FVAHG -LEEYQQNRN FPALDG---T SQLSAALKFG VISPRTLWQT TLEAWE---Q SRSEE--ARA SIETWQQELA
C18B_3          RVPDGVASDP LPDRDCVENL SPG----LAR GGEEAGRKLV TS---WLNGP -MADYEDGHD DLAGDA---T SRLSPHLHFG TVSAAELVHR AREK------ -------GGL GGEAFVRQLA
C18A_1          T----ALADT SPLPSVQELG FAEPEAAVPD AGTAAARSLL DA---FRESG DIYRYEDRRD YPHEEP---T SRLSPHLKFG TIGIRTVYEA ARAAKS---D ADTDD--ERE NVAAFIGQLA
CPDE_1          ETVWACSVEE LGLENEAEKP SNALLTRAWS PGWSNADKIL NE---FIEKQ -LIDYAKNSK KVVGNS---T SLLSPYLHFG EISVRRVFQC ARMKQIIWAR DKNGE--GEE SADLFLRGIG
CryE_7          EAIWACSIEE LGLENEAEKP SNALLTRAWS PGWSNADKLL NE---FIEKQ -LIDYAKNSK KVVGNS---T SLLSPYLHFG EISVRHVFQC ARMKQIIWAR DKNSE--GEE SADLFLRGIG
CryE_9          GKIVGCNAEE LGLEDEFEKS SNALLARAWC PGWGFANKSL DS---FLRSP -LIDYARDRQ KADGASGTPT SLLSPHLHFG ELSVRKIFHE VRKRQITWAR EGNAG--GEA SVNMFLRALG
CryE_8          LPARRCPSDE LVFEDESERG SNALLARAWS PGWQNADKAL AA---FLNGP -LMDYSVNRK KADSAS---T SLLSPYLHFG ELSVRKVFHQ VRMKQLMWSN EGNHA--GDE SCVLFLRSIG
CryE_6          VPSMTVAEVD WFFTPEQE-A SSDQLKFKWK PGVGGAISEL EH---FLAER -LTEFEHDRA KVDRDS---T SRLSPWIHIG SISVRYIFYR VRQCQAEWLA AG-TD--RAQ SCDDFLQQMG
CryB_2          GELGGLRLDE LDLLPSKPDW AYG-LRETWT PGEKGAQARL GQ---FIEHG -LANYERQRD YPGQPS---T SRLSPHLTFG EITPFQIFAA LRR-----SK SS-------- GTSKFRAEIG
CryB_4          -KIVSDRLES WQLAPTKPDW AGG-LRERWT PGEASARARL RD---FLKTI -ARGYAGDRD RPDRVG---T SGLSPHLRFG ELSPRQVWHA ARF-----AA AEDAA--LGP GIEKFLSELG
CryB_1          SPNIKLELTA PPPEFFPQIN FDHRSVLAFQ GGETAGLARL QD--YFWHGD RLKDYKETRN GMVGAD--YS SKFSPWLALG CLSPRFIYQE VKRYEQERVS NDSTH----- ---WLIFELL
CryB_3          ASLDGVHVDA LNLLPT-PAW DQG-FWEHWQ PGEAGAHEML EI---FVDGA -LSGYRENRD RPDRVG---T SQLSPHLHFG EIAPWRIAST LEA-----QR SARN----GA DIDGYIRQLG
C2E_2           SRAAEPVDWE EVLSS-LEVE RSVGEVDWAQ PGTSGGMNML ES----FIDQ RLRLFATHRN NPNYD---AL SHLSPWIHTG QLSAQRVVKQ VKREKNASE- SVASF----I EELVVRRELA
C2E_3           TKTAKAVDWD KTLAS-LKVD RTVGEPKLAK PGTEAGLAML ES----FIDV RLKLFGTQRN DPNAA---AL SQLSPWLRFG QLSAQRVALQ VR--KNSSP- SVPAF----I EELVVRRELT
C2E_4           NIQAEPVDWN ACRAG-LQVD RSVKEVSWAK PGTASGLTML QS----FISQ RLPYFGSDRN NPNKD---AL SNLSPWFHFG QVSVQRAILE VQKHRSRYPD SVANF----V EEAVVRRELA
C2E_5           NVQAEPVDWN GCRAG-LQVD RSVKEVSWAK PGTASGLTML QS----FIAE RLPYFGSDRN NPNKD---AL SNLSPWFHFG QVSVQRAILE VQKHRSRYPD SVTNF----V EEAVVRRELA
C2E_1           CKNVNTVDWS AAYAS-LQCD MEVDEVQWAK PGYKAACQQL YE----FCSR RLRHFNDKRN DPTAD---AL SGLSPWLHFG HISAQRCALE VQRFRGQHKA SADAF----C EEAIVRRELA
C2E_7           -SSSPVIQWD QLIEDRLKKG AEVPEIDWCK PGETAALEVL KGSQNGFLTK RLKSYATDRN IPLKPG--AL SGLSPYLHFG QISAQRCAFE ARNVRKVAPE AVDAF----T EELIVRRELA
C2E_9           -SADKFIDWD RLIDDNLRKG ADVPELEWCK PGEKAAMEVL MGSKDGFLTK RLKGYAIDRN NPLKPK--GL SGLSPYLHFG QISAQRCALE ARSIRKLNPQ AVDVF----L EELIVRRELA
C2E_8           MMDKKLVDWD SLIDKVVREG AEVPEIEWCV PGEDAGIEVL MGNKDGFLTK RLKNYSTDRN NPIKPK--AL SGLSPYLHFG QVSAQRCALE ARKVRSTSPQ AVDIF----L EELIVRRELS
C2E_10          -EQPEGVDWD ALIARVCSEA ENVPEIDWCE PGEEAAIEAL LGSKDGFLTK RIKSYETDRN DPTKPR--AL SGLSPYLHFG HISAQRCALE AKKCRHLSPK SVDAF----L EELVVRRELA
C2E_6           AVAPEAVDWD GLISEVLSRG ADVPEVEWCT PGEEAALEAL TGPRGFLSPA RLSLYDTKRN DPATPS--AL SGLSPYLHFG QLAPQRAALE AAKHRAKYKA AVESY----L EELVVRRELA
C2A_1           LDLEPGPEFE DAVRD--FRA PEDLEPSVFR GGTSTALSIF SE----FLRE KLECFERYRN DPVKN---CL SNMSPYLHFG QISPLYLALR AS-----EAG ECPEF----L EELIVRRELS
```

```
                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                   365        375        385        395        405        415        425        435        445        455        465        475
CryE_1          WREFFYTAAT NNPRFDKME- GNPICVQIPW DK-NPEALAK WAEGRTGFPW IDAIMTQLRQ EGWIHHLARH AVACFLTRGD LWISWEEGMK VFEELLLDAD WSINAG--SW MWLSCSSFFQ
CryE_5          WREFFYTAAT NNPRFDKME- GNPICVQIPW DK-NPEALAK WAEGRTGFPW IDAIMTQLRQ EGWIHHLARH AVACFLTRGD LWISWEEGMK VFEELLLDAD WSVNAG--SW MWLSCSSFFQ
CryE_2          WREFFYTAAT NNPRFDRME- GNPICIQIPW DR-NPEALAK WAEGKTGFPW IDAIMTQLRQ EGWIHHLARH AVACFLTRGD LWVSWESGVR VFDELLLDAD FSVNAG--SW MWLSCSAFFQ
CryE_3          WREFFYTAAT NNPKFDRME- GNPICIQIPW DK-NPEALAK WAEGKTGFPW IDAIMTQLRQ EGWIHHLARH AVACFLTRGD LWISWESGVR VFDELLLDAD FSVNAG--SW MWLSCSAFFQ
CryE_4          WREYFYTMSV NNPNYDRME- GNEICLSIPW AKPNEDLLQS WRLGQTGFPL IDGAMRQLLA EGWLHHTLRN TVATFLTRGG LWQSWEHGLQ HFLKYLLDAD WSVCAG--NW MWVSSSAFER
64E_3           WREFFYTTAV GIPNFNKME- GNSACVQVDW DN-NPEHLAA WREARTGFPF IDTIMTQLRQ EGWIHHLARH AVACFLTRGD LWISWEEGQK VFEELLLDSD WSLNAG--NW QWLSASTFFH
64E_5           WREFYYTTGA GIPNFNKME- GNPVCVQVDW DN-NKEHLEA WSEGRTGYPF IDAIMTQLRT EGWIHHLARH AVACFLTRGD LWISWEEGQK VFEELLLDAD WSLNAG--NW LWLSASAFFH
64E_1           WREFYYTVAA AEPNFDRML- GNVYCMQIPW QE-HPDHLEA WTHGRTGYPF IDAIMRQLRQ EGWIHHLARH AVACFLTRGD LWISWEEGQR VFEQLLLDQD WALNAG--NW MWLSASAFFH
64E_2           ARDFALQVSQ LQTIPERIIS LNKICLPIPW DKNNNEIVEK LTDAQTGFPF FDAAITQLKT EGYVINEVSE ALATFVTNSL LWVSWEEGQN FFSQHLICFD LAMSTH--SW LEASGSTMVT
64E_4           WREFFYTTAF GTPNFDKMK- GNRICKQIPW NE-DHAMLAA WRDGKTGYPW IDAIMVQLLK WGWMHHLARH CVACFLTRGD LFIHWEQGRD VFERLLIDSD WAINNG--NW MWLSCSSFFY
C1mE_1          WRDFYRHCMC NWPYTSMGMP YRLDTLDIKW EN-NPVAFEK WCTGNTGIPI VDAIMRKLLY TGYINNRSRM ITASFLSK-N LLIDWRWGER WFMKHLIDGD SSSNVG--GW GFCSSTGIDA
C1mE_2          WRDFYKHVLV HWPYVCMNKP FKPTYSNIEW SY-NVDHFHA WTQGRTGFPI IDAAMRQVLS TGYMHNRLRM IVASFLAK-D LLVDWRMGER YFMEHLIDGD FASNNG--GW GFAASVGVDP
C1mB_1          WREFYRHLIT YHPSLCKHRP FIAWTDRVQW QS-NPAHLQA WQEGKTGYPI VDAAMRQLNS TGWMHNRLRM ITASFLVK-D LLIDWREGER YFMSQLIDGD LAANNG--GW QWAASTGTDA
C1mB_2          WREFYRHLMT WYPALCKHQP FIRWTKRVAW QE-NPHYFQA WQKGETGYPI VDAAMRQLNA TGWMHNRLRM ITASFLVK-D LLIDWRLGER YFMSQLIDGD LAANNG--GW QWAASTGTDA
C18B_1          WREFYQHALY HFPSLADGP- YRSLWQQFPW EN-REALFTA WTQAQTGYPI VDAAMRQLTE TGWMHNRCRM IVASFLTK-D LIIDWRRGEQ FFMQHLVDGD LAANNG--GW QWSASSGMDP
C18B_2          WREFYQHCLY SFPALAQGP- YRSPFQEFPW EE-NQDHFQA WCEGRTGYPI IDAAMAQLNQ TGWMHNRCRM IVASFLIK-D LILNWQWGEL YFMQTLYDGD LAANNG--GW QWSASSGMDP
C18B_3          WRDFHHQVLA DRPDASWSD- YRP--RHDRW RS-DADEMHA WKSGLTGYPL VDAAMRQLAH EGWMHNRARM LAASFLTK-T LYVDWREGAR HFLDLLVDGD VANNQL--NW QWVAGTGTDT
C18A_1          WREFYAQVLY FNQNVVSEN- FKAYEHPIEW RD-DPAALQA WKDGETGYPI VDAGMRQLRA EAYMHNRVRM IVAAFLTK-D LLVDWRAGYD WFREKLADHD TANDNG--GW QWAASTGTDA
CPDE_1          LRDYSRIICF NFPFTHEQS- LLSHLRFFPW DA-DVDKFKA WRQGRTGYPL VDAGMRELWA TGWMHNRIRV IVSSFAVK-F LLLPWKWGMK YFWDTLLDAD LECDII--GW QYISGSLPDG
CryE_7          LREYSRYICF NFPFTHEQS- LLSHLRFFPW DA-DVDKFKA WRQGRTGYPL VDAGMRELWA TGWMHNRIRV IVSSFAVK-F LLLPWKWGMK YFWDTLLDAD LECDIL--GW QYISGSIPDG
CryE_9          FREYSRYLSF HFPFTHERS- LLANLKSFPW RA-DEGYFKA WRQGRTGYPL VDAGMRELWA TGWAHNRIRV VVASFSVK-F LQLPWRWGMK YFWDVLLDAD LECDVL--GW QYISGSLPDG
CryE_8          LREYSRYLTF NHPCSLEKP- LLAHLRFFPW VV-DEVYFKV WRQGRTGYPL VDAGMRDSRA TGWLHDRIRV VVSSFFVK-V LQLPWRWGMK YFWDTLLDGD LESDRL--GW QYISGSLPDG
CryE_6          YREYSRYLAF HFPFIHERS- LLGHLRACPW RI-DQHAFKA WRQGQTGYPI VDAAMRQLWS SGWCHNRGRV VAASFLVK-D LLLPWQWGLK HYWDAQIDAD LECDAL--GW QYVSGGMSDA
CryB_2          WREFSYHLLF HNPDLSGRN- FRPEFDAMSW RD-DIRALRT WQRGLTGYPI VDAGMRELWR TGWMHNRVRM IVASFLIK-D LMIDWRHGEK WFWDTLVDAD AANNPA--SW QWVAGSGADP
CryB_4          WREFCRHLLH DHPDLATEN- LQTNFDGFPW QS-DGKVLAA WQRGRTGYPI VDAGLRELWH TGVMHNRVRM VVASFLVK-H LLIDWRDGEA WFWDTLVDAD AGSNPA--NW QWVAGCGADA
CryB_1          WRDFFRFVAQ KYGNKLFNR- GGLLNKNFPW QE-DQVRFEL WRSGQTGYPL VDANMRELNL TGFMSNRGRQ NVASFLCK-N LGIDWRWGAE WFESCLIDYD VCSNWG--NW NYTAGIGNDA
CryB_3          WRDFAYHLLH HFPDTTTQN- LNPRFAGFDW ATVDPVTLDA WQRGRTGIPI VDAGLRQLWH TGWMHNRVRM IVASLLCK-H LRVHWLEGAR WFWDTLVDAD LANNTM--GW QWVAGTGADA
C2E_2           DNFCFYNPSY DNISGAYDWA KKTLQDHAKD SRQYLYTKEQ LENAKTHDQL WNAAQRQLVS EGKMHGFLRM YWAKKILEWT ASPEEALSIA IYLNDRLSLD GCDPNGYVGC MWSICGIHDQ
C2E_3           DNFCFYNKNY DSVTGAYEWA QKTLKDHAKD KREYLYTREQ FEKAQTHDKL WNAAQIQMVT EGKMHGFLRM YWAKKILEWS TSPEEALSIA LYLNDRYELD GQDPNGFVGC MWSICGIHDQ
C2E_4           DNFCFYNKNY DKLEGAYDWA QTTLRLHAKD KRPHLYSLEQ LESGKTHDPL WNAAQMQMVQ EGKMHGFLRM YWAKKILEWT RSPEEALEFA IYLNDRFQLD GRDPNGYVGC MWSICGIHDQ
C2E_5           DNFCFYNKNY DKLEGAYDWA QTTLRLHAKD KRPHLYSLEQ LESGKTHDPL WNAAQMQTVK EGKMHGFLRM YWAKKILEWT RSPEEALEFA IYLNDRFQLD GWDPNGYVGC MWSICGIHDQ
C2E_1           DNFCFYNEHY DSLKGLSSWA YQTLDAHRKD KRDPCYSLEE LEKSLTYDDL WNSAQLQLVR EGKMHGFLRM YWAKKILEWT ATPEHALEYA ILLNDKYSLD GRDPNGYVGC MWSIGGVHDM
C2E_7           DNFCYYQPNY DSLMGAWEWA RKTLMDHASD KREHLYTREQ LEKAQTADPL WNASQLEMVH FGKMHGFMRM YWAKKILEWT SGPEEALAIA IYLNDKYEMD GRDPNGYVGC MWSICGLHDQ
C2E_9           DNYCYYQPHY DSLLGAWEWA RKTLMDHASD KREYIYTREQ LEKAQTADPL WNAAQLEMAH HGKMHGFMRM YWAKKILEWT RGPEEALEIC IYLNDKYEID GRDPNGYVGC MWSICGVHDQ
C2E_8           DNFCYYQPHY DSLKGAWEWA RKSLMDHASD KREHIYSLEQ LEKGLTADPL WNASQLEMLY QGKMHGFMRM YWAKKILEWT KGPEEALSIS IYLNNKYEID GRDPSGYVGC MWSICGVHDQ
C2E_10          DNFCYYQPQY DSLSGAWEWA RKTLMDHAAD KREHIYTREQ LENAKTHDPL WNASQLEMVH HGEMHGFMRM YWAKKILEWT SGPEEALSTA IYLNDKYEID GRDPSGYVGC MWSICGLHDQ
C2E_6           DNFCHYCPTY DSLEAAAEWA RDSLDKHRTD KREFLYTRDQ LECGATHDEL WNAAQLEMVH VGKMHGFMRM YWAKKILEWT QGPEQAIEWA IYLNDRYELD GRDPGGYTGV LWSMAGVHDM
C2A_1           MNFVHYSDSY SSISCLPEWA QRTLMDHVAD PREYEYSLRE LESASTHDPY WNAAQQEMVI TGKMHGYMRM YWGKKILEWT DHPARAYDIA LYLNDRYEID GRDPNGFAGV AWCFG-KHDR
```

```
            ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ..
               485        495        505        515        525        535        545        555        565
CryE_1      QFFH--CYCP VGFGRRTDPN GDY-IRRYLP VLRGFPAKYI YDPWNAPEGI QKVAK----C LIGVNYP-KP MVNHAEASRL NIERMKQIYQ QL
CryE_5      QFFH--CYCP VSFGRRTDPN GDY-IRRYLP VLRGFPAKYI YDPWNAPESV QKAAK----C IIGVHYP-MP MVHHAEASRL NIERMKQIYQ QL
CryE_2      QFFH--CYCP VGFGRRTDPS GDY-IRRYLP KLKAFPSRYI YEPWNAPESI QKAAK----C IIGVDYP-RP IVNHAETSRL NIERMKQIYQ QL
CryE_3      QFFH--CYCP VGFGRRTDPS GDY-VKRYLP KLKGFPSRYI YEPWNAPESV QKAAK----C IIGVDYP-KP MVNHAETSRL NIERMKQIYQ QL
CryE_4      LLDSSLVTCP VALAKRLDPD GTY-IKQYVP ELMNVPKEFV HEPWRMSAEQ QEQYE----C LIGVHYP-ER IIDLSMAVKR NMLAMKSLRN SL
64E_3       QYFR--VYSP IAFGKKTDKH GDY-IKKYLP VLKKFPTEYI YEPWKAPRSV QERAG----C IVGKDYP-RP IVDHEVVHKK NILRMKAAYA KR
64E_5       QFFR--VYSP VAFGKKTDKN GDY-IKKYLP ILKKFPAEYI YEPWKSPRSL QERAG----C IIGKDYP-KP IVEHNVVSKQ NIQRMKAAYA RR
64E_1       QYFR--VYSP VAFGKKTDPQ GHY-IRKYVP ELSKYPATCI YEPWKASLVD QRAYG----C VLGTDYP-HR IVKHEVVHKE NIKRMGAAYK VN
64E_2       GRQKSYQDPL LFVSKKLDPN GEY-IKRYLP KFINFPIEFI HKPGNASLEA QQAAN----C VIDIDYP-KP LFEYECRNGI CCKRLRVFME VV
64E_4       QFNR--IYSP ISFGKKYDPD GKY-IRHFLP VLKDMPKQYI YEPWTAPLSV QTKAN----C IVGKDYP-KP MVLHDSASKE CKRKMGEAYA LN
C1mE_1      QPYFR-VFNM DIQAKKYDPQ MIF-VKQWVP ELISSENKR- ---------- ---------- --PENYP-KP LVDLKHSRER ALKVYKDAM- --
C1mE_2      QPYFR-VFNP LLQSEKFDPD GDY-IRKWVE ELRDLPELKG GKGGEIHDPY GRGSEKVKKK LEEKGYP-RP IVEHSGARDR ALDAYKRGLA RD
C1mB_1      APYFR-IFNP TTQGEKFDHE GEF-IRQWLP ELRDVPGKVV HEPWKWAQKA G--------- -VTLDYP-QP IVEHKEARVQ TLAAYEAARK GK
C1mB_2      APYFR-IFNP TTQGERFDRD GEF-IRQWLP ALRDIPGKAI HEPWRWAEKA G--------- -VVLDYP-RP IVEHKQARIA TLSAYEAARK GA
C18B_1      KP-LR-IFNP ASQAKKFDAT ATY-IKRWLP ELRHVHPKDL ISG---EITP I--------E RR--GYP-AP IVNHNLRQKQ FKALYNQLKA AI
C18B_2      KP-LR-IFNP HTQAQKFDPE GEY-IRTWLP QLARFDTGDL LTG---KLTP G--------S RRSVNYP-EP IVDHNQQQRE FKRRYQLVK- --
C18B_3      RP-NR-VLNP VIQGKRFDAR GDY-VRGWVP ELAEVEGSAI HEPWKLQG-- --------LD RAGLDYP-DP VVDLAEARAR FERARGLD-- --
C18A_1      QPYFR-VFNP MTQGERYDPD ADY-ITEFVP ELRDVPADAI HSWHELSLSE R--------R RHAPEYP-DP IVDHSQRRED AIAMFERARG DE
CPDE_1      HELDR-LDNP AIQGAKYDPE GEY-IRQWLP ELARLPTEWI HHPWDAPLTV LKAS----GV ELGTNYA-KP IVVIDTAREL LTKAISRTRE AQ
CryE_7      HELDR-LDNP ALQGAKYDPE GEY-IRQWLP ELARLPTEWI HHPWDAPLTV LKAS----GV ELGTNYA-KP IVDIDTAREL LAKAISRTRE AQ
CryE_9      HELDR-IENP EVEGYRFDPD GDY-VRRWIP ELARLPNEWV HHPWDAPPSA LRAA----GV ELGTNYP-RP IVEIGAARER LQASLAEMWE RD
CryE_8      RELDR-IDNP QLEGYKFDPH GEY-VRRWLP ELARLPTEWI HHPWDAPESV LQAA----GI ELGSNYP-LP IVELDAAKTR LQDALSEMWE LE
CryE_6      HPFSY-MMDL EKEARRFDPD GEY-VRRWLP ALSRLPTEYI HAPWKAPASV LAAA----DV ELGCNYP-LP IITRSDAKAN VDYACGVLEK SA
CryB_2      APYFR-IFNP VLQGEKFDPH GDY-VRQHVP EISALPDRYI HRPWEAPAAV LKDK----GI VLGKTYP-NP IVDHGAARER ALIVYQSLKD --
CryB_4      APYFR-VFNP QLQGEKFDPD GTY-VRRWVP ELQGLPAKLI HQPWQATPNE LASA----GV TLGKTYP-QP IVDHARGRER ALSAYAKIRK G-
CryB_1      RDFRY--FNI PKQSQQYDPQ GTY-LRHWLP ELKNLPGDKI HQPWLLSATE QKQWG----V QLGVDYP-RP CVNFHQSVEA RRKIEQMGVI A-
CryB_3      APYFR-VFNP VTQAEKFDPQ ATY-ITRWIP ELAALPVKER FAPWLHPLSL AR-------- -LAPTYPRAP IIGLAEGRDA ALAAYAGTRG --
C2E_2       GWAERPIFGK IRFMNYAGCK RKFDVAQFER KYTAVKENSN KDSKKSSSKN ---------- ---------- ---------- ---------- --
C2E_3       GWAERAVFGK IRFMNYKGCL RKFNVAQFER KYCSKKV--- ---------- ---------- ---------- ---------- ---------- --
C2E_4       GWAEREIFGK IRYMNYAGCK RKFDVAEFER KYSPAD---- ---------- ---------- ---------- ---------- ---------- --
C2E_5       GWAEREIFGK IRYMNYAGCK RKFDVAEFER KISPAD---- ---------- ---------- ---------- ---------- ---------- --
C2E_1       GWKERAIFGK VRYMNYQGCR RKFDVNAFVM RYGGKVHKKK ---------- ---------- ---------- ---------- ---------- --
C2E_7       GWRERPVFGK IRYMNYAGCK RKFNVDGYIA YVRKLVVDTK K--------- ---------- ---------- ---------- --RKAEADIS --
C2E_9       GWKERPVFGK IRYMNYAGCK RKFDVDGYIA YVKRLVGEIK K--------- ---------- ---------- ---------- --RKPEETLE --
C2E_8       GWKERPVFGK IRYMNYAGCK RKFNVDSYIS YVKSLVSVTK KK-------- ---------- ---------- ---------- --RKAEEQLT RD
C2E_10      GWKERPVFGK IRYMNYAGCK RKFDVDAYIS YVKRLAGQSK K--------- ---------- ---------- ---------- --RNAEESPN PV
C2E_6       GWAERAVFGK IRYMNYNGCK RKFDIKAYVA YVSKAVAEAK AKGRAAKLPS AAAAGASGAA AAGATAAAAA AAAPGPSGAQ AAKAAKAKAE PK
C2A_1       AWAEREIFGK VRYMNDRGLK RKFRIDEYVD RIRGLMDE-- ---------- ---------- ---------- ---------- ---------- --
```

4.3. Clustering results

The resulting phylogenetic trees based on the multiple alignments are shown in figure 4.9.

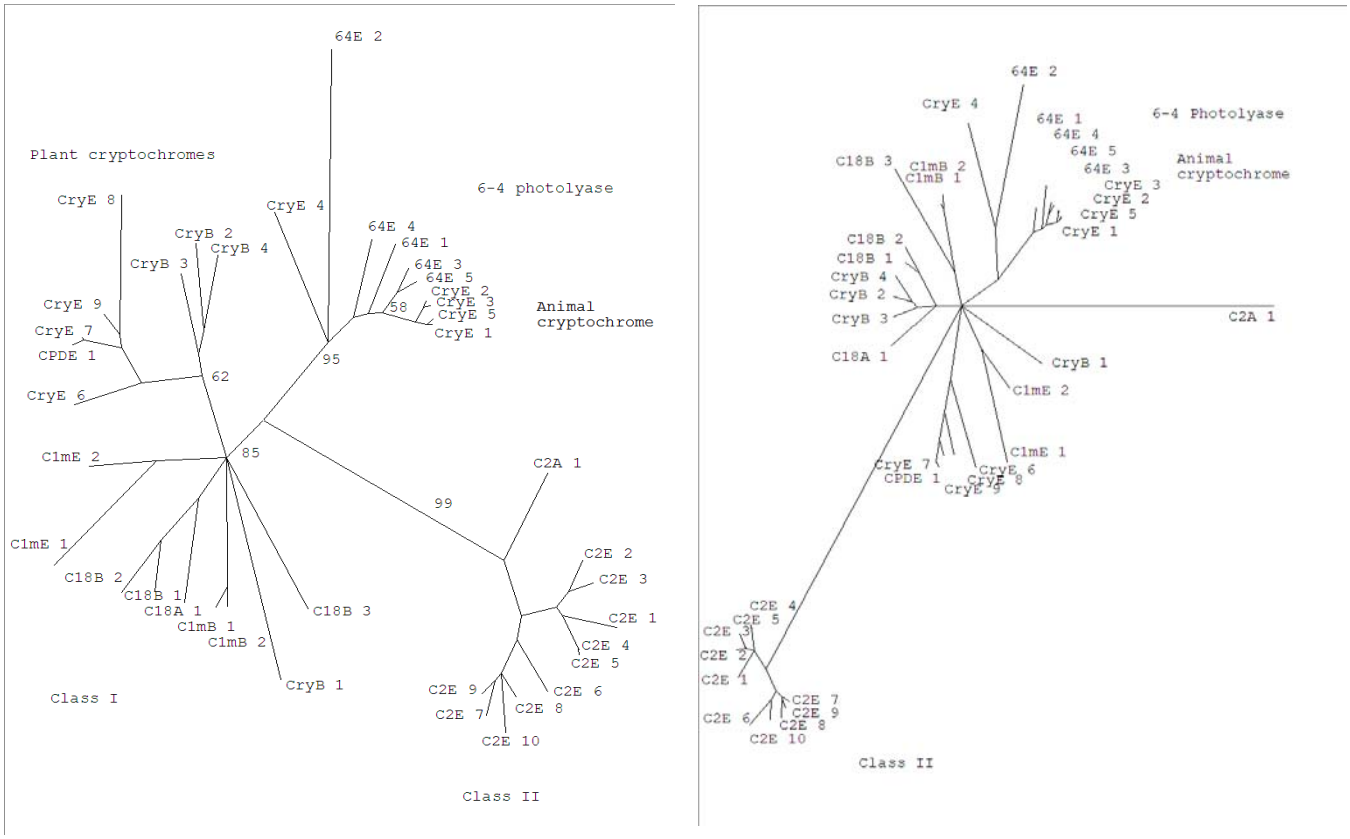


Figure 4.9: The phylogenetic trees from the alignments in figure 4.5 and 4.6 respectively. The trees were generated with the Treepuzzle program.

The reliability of the results shown in figure 4.9, left panel, was assessed with the bootstrap method in the PHYLIP package as shown in figure 4.10.

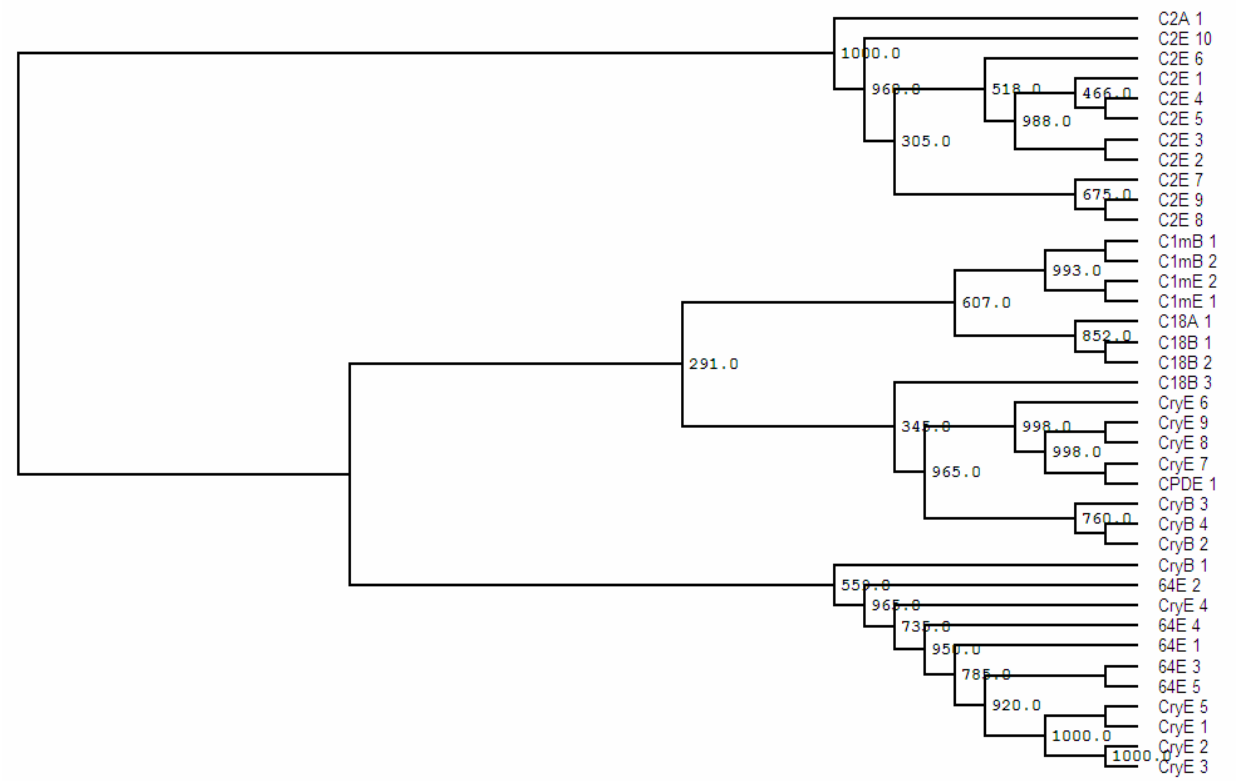


Figure 4.10: Bootstrap tree from the alignment shown in figure 4.5. The tree is generated with the SeqBoot, ProtDist, Neighbor, and Consense programs of the PHYLIP package. The bootstrap values are out of 1000 replicates.

This figure shows more or less the same clustering as figure 4.9. Two notable exceptions are the sequence C18B_3 which is closer to the plant cryptochromes instead of the class I photolyases and the sequence CryB_1 which are closer to the (6-4) photolyases than the plant cryptochromes. The same analysis done without the CryB_1 increased the bootstrap value for the class I group to 68%.

These results were checked with parsimony in PAUP and maximum likelihood (ML) in Molphy, the results were for the most part the same.

The phylogenetic tree with the additional cryptochrome from Arabidopsis showed the new sequence clustering together with the CryB_1 sequence and not together with the rest of the plant cryptochromes. Otherwise the phylogeny showed the same results.

4.4. Search for functional sites

The results of comparing number of conserved sites from the two analysed alignments are presented in table 4.4.

Table 4.4: The number of conserved sites in the alignment. The alignments refer to the ones mentioned under the alignment results above, and the A and B columns are, respectively, with and without the most divergent sequences as indicated by the ClustalX (CryB_1 and 64E_2). The strong and weak conservation are the same as used in the ClustalX program.

	Alignment 1	Alignment 2
Gap open parameters (pairwise/multiple)	10/10	100/100
All characters equal	4	0
Strong conservation[1)]	6	2
Weak conservation[2)]	11	4

1) STA, NEQK, NHQK, NDEQ, QHRK, MILV, MILF, HY, FYW
2) CSA, ATV, SAG, STNK, STPA, SGND, SNDEQK, NDEQHK, NEQHRK, FVLIM, HFY

The reduction of conserved sites from the first alignment to the second is obvious. Using alignment 1, the conservation within and between the different classes is shown in table 4.5.

Table 4.5: Number of sites conserved within each class. Numbers shows the accumulated sites, in parenthesis are the sites added with each group.

	Equal	Strongly conserved	Weakly conserved
1) (6-4) photolyase	87	178 (91)	235 (57)
2) animal crypochromes	195	299 (104)	354 (55)
3) Class I photolyase	76	132 (56)	146 (15)
4) Class II photolyase	121	187 (66)	204 (27)
5) Plant cryptochromes	58	100 (42)	122 (22)
6) Class I and plant cry	34	68 (34)	79 (11)
7) (6-4) and animal cry	66	103 (37)	180 (77)
8) Class I, (6-4) and both cry	18	36 (18)	50 (14)
9) Class I, plant cry and class II	7	18 (11)	31 (13)
10) (6-4), animal cry and class II	10	39 (29)	53 (14)

The positions of the fully conserved sites are shown in figure 4.11.

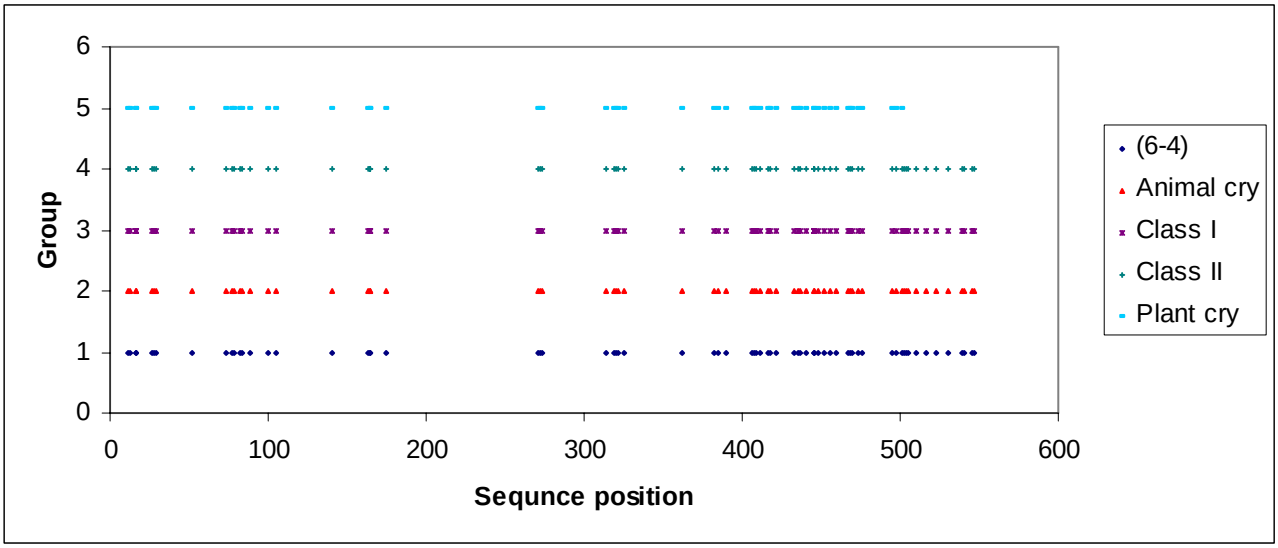


Figure 4.11: Graphical presentation of the results from table 4.5, showing the distribution of the conserved sites along the protein.

The amino acids suggested to bind to the FADH cofactors (Park *et al.* 1995) are; 310, 311, 312, 313, 314, 355, 362, 423, 458, 460, 463 and 464 (figure 4.12).

The positions of the conserved sites for the whole alginment are; fully conserved: position 314, 320, 406 and 460, strongly conserved: position 313, 322, 359, 412, 503 and 505, and weakly conserved: position 100, 280, 317, 362, 404, 417, 422, 433, 438, 469 and 498.

Another useful feature of the photolyase mechanism is the triple tryptophan chain analysed by Byrdin *et al.* (2003), these sites are at positions 390, 445 and 470. At all these three sites every sequence except the ones in class II shows a tryptophan residue.

The locations for the class I photolyase and (6-4) photolyase complete conservationon is shown below.
314, 320, 362, 390, 406, 407, 409, 412, 441, 445, 448, 460, 467, 468, 470, 498, 504, 546.

Figure 4.12: The five sites shown to bind FADH *in the photolyase protein, these are sites 310, 311, 312, 313, 314, 355, 362, 423, 458, 460, 463 and 464 from the full alignment (Appendix A).*

The results from the search for tree determinant positions for the node separating the three main groups yielded 9 sites; **310**, 404, **411**, **426**, 432, **454**, 469, **493** and 494. Visual inspection of the sites indicated that only those five in bold are useful for classification of the three groups.

This search method was also tried out on different parts of the tree as indicated in table 4.6.

Table 4.6: Number of tree determining sites based on the phylogenetic tree.

	Stringency level*			
	1 (high)	2	3	4 (low)
Class I vs. Class II	5	21 (16)	37 (8)	45 (10)
Class I vs. (6-4)	0	3 (3)	13 (10)	21 (8)
Class II vs. (6-4)	6	23 (17)	46 (23)	55 (9)
(6-4) photolyase vs. animal cryptochromes	0	2 (2)	13 (11)	48 (35)
Class I CPD photolyase vs. plant crytochromes	0	0 (0)	1 (1)	10 (9)
(6-4) photolyase and animal cryptochromes vs. class II CPD photolyase	6	22 (16)	36 (14)	45 (9)
Class II CPD photolyase vs. class I CPD photolyase and plant cryptochromes	1	11 (10)	16 (5)	23 (7)

*) 1: Both groups nonvariant and not equal
2: Both groups strongly conserved and not in the same strongly or weakly conserved category
3: Both groups at least weakly conserved and not in the same strongly or weakly conserved category
4: Both groups at least weakly conserved and not in the same strongly conserved category

The positions and residues of the first split shown in table 4.6 are presented in table 4.7.

Table 4.7. The amino acids at the positions indicated by the most stringent level in the analysis between class I and class II as shown in table 4.6.

	Position referring to the column in the multiple alignment				
Group name	310	390	415	435	482
Class I	T	W	M	F	P
Class II	L	D	Q	K	W

A last approach to possible important sites are those which show conservation in one group, but seems to be under neutral control in another group. These sites are shown in table 4.8

Table 4.8: Positions in the alignment showing conserved sites in one group and no conservation in another group.

Conserved group	Unconserved group*	Conservation level of first group		
		High	Middle	Low
Class I	Plant cryptochromes	27	61 (34)	67 (6)
(6-4) photolyase	Animal cryptochromes	10	25 (15)	42 (17)

*) The group is neither identical nor strongly or weakly conserved.

5. Discussion

5.1. Introduction

This thesis used the multiple sequence alignment of several photolyase and cryptochromes as a basis for suggesting functional sites. These sites would either be the ones important for the common function, such as the binding of the FADH cofactor, or they would be the ones making the difference in function between the groups. Due to looking at single sites within the alignment, and drawing conclusions on the specific amino acids present at that site, any mistakes made in the sequencing reaction of those who provided the original sequences to the database will be problematic. The analysis shown in the results part of this paper are based on the assumption that all the sequences downloaded are correct, and that any changes therefore are the result of a mutation in the respective gene. This lead to some difficulties when in a clearly conserved area, one or two proteins showed little or no homology with the rest. These sequences could then be taken as an indication that the function located in that area was not vitally important to the protein, or that those few proteins had developed a new way of doing that function. Since there was no clear indication that any of the sequences could be ignored, all the analysis presented here are the results of looking at all the downloaded sequences.

5.2. Laboratory work

From the sequence the primers were designed from the expected size was 596 bp. The single band in the lanes with Arabidopsis matched the size of the expected sequence fragment from the gene. The obtained sequence confirmed this conclusion and the principle worked. The PCR reaction with cotton also resulted in a band of about the same size as that for Arabidopsis. In addition to this band, cotton also showed a second band of about 1000 bp. It would have been interesting to sequence both bands and find out if they really were from a potential photolyase gene in cotton, or just a chance product based on the random matching of the primers to the template DNA. If the bands would have been shown to be from the photolyase gene, this could mean that cotton had two versions of the intron pattern in the gene sequence. Since the cotton used here is a tetraploid, it may be possible that the two genomes have different versions of the gene.

The work in the laboratory didn't yield as much as hoped, but the principle of amplifying a gene based on a primer pair designed from conserved areas at least worked out. With future research the primers can be optimized based on more closely related species to the one analysed, or via the use of degenerate primers. The primer pair designed from Arabidopsis gave a product in samples from cotton, tomato, cactus and maize. All the product sizes were between ~600 bp and ~1000 bp, which can be explained with a difference in the intron lengths if the products are all part of the photolyase gene.

5.3. Aligning the downloaded sequences

The figures from the first two alignments show why correct gap costs are important. When the gap penalty is set to high, on area in the beginning of the protein is not shown as being conserved over all the sequences. Whereby this area is clearly conserved when the gap cost are lowered. The reason for this comes from the fact that the area right after the first conserved block are very poorly

conserved and requires several gaps to line up properly. When the placement of gaps are too strict, the penalty for introducing the required gaps will not be outweighed by the conserved area since it's no more than about ten amino acids wide. The end of the alignment before the C-terminal extensions begin, the conservation are showed to be much greater, here the alignments are more similar since the positive effect of aligning conserved residues outweighs the gaps necessary to make the right matches, even when the gap penalty are high.

A closer inspection shows however that the structures from the first alignment are, with some exceptions, retained within each group in the second alignment. This is because the divergence times within the groups are small enough to not having caused too many changes, therefore the gaps needed are fewer and shorter, and the alignment score gets a lot higher with only these few gaps.

The alignments also show a couple of other things happening. The order of the species changes slightly in the output file. This is due to the fact that ClustalX orders the sequences after their pairwise similarity and subsequent clustering, which means that the similarities score between some species have been reduced. This is caused by the same effect as described above; when the sequences require a larger number of gaps to line up properly, a high gap penalty will tend to cause the lowering of the similarity score since the aligning of two almost similar amino acids with no gaps will be favoured over the aligning of two identical amino acids with a gap. This phenomenon can be seen when looking at the end of the sequence order where the sequences differing by more than 70% from any other sequence or group of sequences are placed. The alignment with high gap penalty contains nine of these sequences while the first only contains two.

The problems with exceptions in conserved areas are also clearly showed. Four sequences (64E_2, CryB1, C2A_1 and CryE_8) vary to a greater or lesser extent from the consensus sequence in the first 20 amino acids of the protein. One sequence (CryB_4) does not have that area at all. While the rest of the sequences show such conservation that without these exceptions it would have been taken as important, these five either shows that they have managed without that part, or that the conservation is not coupled to a specific function. An automated analysis of the sequences might well overlook this area, dependent on the level of stringency for the search, even though it seems to be potentially important. From the research done with the photolyases, this area has been shown to contain the second cofactor which, although it increases the effectiveness of the protein, is not strictly necessary. The explanation for these five proteins may therefore be that they don't use the second cofactor, or that they have acquired a different way of binding it.

While the liberal placement of gaps in some areas are needed to make the alignment correct, it makes it easier to slide residues by hand to improve the alignment. This might introduce an unwanted bias and should probably be avoided unless it's fairly obvious that the alignment program has made a mistake.

Looking at the types of gaps present indicates that most of them are caused by insertions. This is because most of the gaps are placed in all but a few sequences, usually only one. This is what one would expect if an insertion happened in one sequence during the evolution. A deletion would show up as a gap in the affected sequence only. In some cases however, the gap includes all sequences of a group and in this case, the gap might be an important factor for the functionality of that group.

When constructing phylogenies from sequences, the presence of gaps must be taken into consideration in some or other way. Unlike substitutions there has as yet not been given any clear probability measure of when gaps occur. Therefore, including gaps in distance calculations or tree building is quite difficult. One approach is to treat the gap as an extra character state, taking into consideration the problem with gap opening versus gap extension. This approach works best when only considering the number of mutations in a given tree as is the basis for parsimony. The other

approach is to ignore any site that contains a gap. This is probably the best when the gap obviously is caused by an insertion in one, or only a few sequences. In this case, only the information concerning the relationship between the sequences having the insert is reduced.

When one is trying to infer the functional difference between two sequences, the gaps are more important than with the phylogenetic analysis. Since a change in function can be caused by insertions and deletions, which distorts the active site of the protein, just as well as a substitution of the important residue(s), all the sites have to be considered for such an analysis.
When looking at the multiple alignment, any extensions to the proteins will show up before or after the main part of the gene. The class II CPD photolyse group are clearly shown to have an N-terminal extension of about 70 to 200 amino acids. Additionally the two eukaryote class I CPD photolyase sequences also have an N-terminal extension of ~70 and ~130 amino acids. The presence of an N-terminal extension seems therefore to be an indication of a CPD photolyase from an eukaryote.

At the other end, the C-terminal part of the protein, the cryptochromes from plants and animals have a clear extension. The length varies considerably from about 100 to about 400 amino acids. The 6-4 photolyase also seems to have an extension of somewhat smaller length, about 20-40 amino acids. The C-terminal extension of the cryptochromes has been proposed to affect the cellular transportation of the protein.

Because the alignment programs usually are not perfect, some editing by visual inspection may be necessary. This usually is the case in the vicinity of gap edges, especially in the areas of the alignment with most gaps. In this part of the alignment, the conservation is usually low, and the aligning of amino acids become more uncertain. If manually editing an alignment, it is necessary to be careful and just fix the most obvious problems, otherwise false positive results may be introduced in addition to the research being difficult to replicate. The presence of many gaps in an area may be taken as an indication that it does not contain any vital function to the protein. The results from the photolyase also show that the areas with fewer gaps contain more conserved sites.

When comparing the alignment from figure 4.9 with the alignment of figure 1 in the appendix, it can be seen that the alignment of the class II CPD photolyase is not the same. For some unknown reason the sequences of this class have been aligned slightly different to the rest of the sequences. This shows the problem with the reproducibility of sequence alignments. If any of the parameters of the alignment program, even the program itself, are changed, the resulting alignment may not be quite the same, and the results made based on this alignment is not directly comparable. In this case there are over 500 columns in the alignment when shortened to a practical level, and it is fairly impossible to manually check the alignments to find out exactly what the changes are.

5.4. Clustering results

All the phylogenetic trees show at least three main groups clearly indicated; the class II CPD photolyase, the 6-4 photolyase and animal cryptochromes, and the class I CPD photolyase and plant cryptochromes. Looking at figure 4.9 from the standpoint of not knowing the different functions and proposed evolution, the following points can be observed.

The same conclusions as were drawn from the Treepuzzle program apply to bootstrap analysis of the NJ method as well. The supporting values for each branch vary slightly but otherwise the tree is the same, with some exceptions. These exceptions are; CryB_1 which seems to appear closer to the 6-4 photolyase group, C18B_3 which are put together with the plant cryptochromes and the support values within the class I CPD photolyase group. The low bootstrap value for the class I group

(29%) is probably because of the CryB_1 which, according to the sequence database, should be placed with the class I and not with the 6-4 photolyase. This is supported by the fact that the bootstrap value for the 6-4 group with CryB_1 is 56% while the same group without CryB_1 is 96%. The NJ/bootstrap tree also shows the sequences using MTHF as a second cofactor as one group with a support of 99%, which in the Treepuzzle tree were shown as two groups. These differences show that several programs should be used to get a consensus on what sequences belong together.

One quite interesting observation is the placement of the sequence from *Sinapis alba*. This protein has been shown to have photolyase activity (Batschauer, 1993), yet it is consistently grouped together with the plant cryptochromes (Cashmore *et al.* 1999; Kobayashi *et al.* 2000; Kleine *et al.* 2003), specifically CryE_7, the cryptochrome from *Arabidopsis thaliana*. While this sequence seems to be referred to as a cryptochrome in all articles sited here, no mention of its photolyase activity has been found.

The identity between these two sequences is 74%, with most of the difference caused by a C-terminal extension in the *Arabidopsis* sequence. Without this extension, which seems to be a common property amongst cryptochromes but not in the photolyases (Ahmad *et al.* 1998), the sequence identity is 90%.

Possible conclusions of these results are; a mistake may have been made in the test of photolyase activity for the *S.alba* sequence, or else the phylogenetic clustering of cryptochromes is not able to separate between photolyases and cryptochromes in all cases. In the latter case, and if the cryptochrome from Arabidopsis really is non-repairing, then the 10% difference between the two sequences must contain the switch between the two functions (cryptochrome and photolyase).

When looking at just the different sites within the conserved region of the matchup between the sequence from Arabidopsis and the *S.alba* sequence, 10 sites were discovered

The phylogenetic results can also be analysed without the prior knowledge of function. In that case the results would look like the following. The group of class II CPD photolyase is obviously set apart from the rest. This group seems to contain three groups: C2E_1 through C2E_5, C2E_6 through C2E_10 and C2A_1. While the functional properties of the class II has yet to be determined, this grouping makes sense, since it corresponds to animals, plants and archaebacteria respectively. The 6-4 photolyase and animal cryptochrome group is not clearly different. Without knowing the difference, this group would have to be taken as one. The class I CPD photolyase and plant cryptochrome group consists mostly of small groups with no clear internal ordering. The only exception is the cryptochrome group that stands out, which in addition to plant cryptochromes also contains bacterial cryptochrome, although it lacks one of the cryptochrome sequences (CryB_1).

The phylogenetic analysis shows mostly what was expected based on the previous works on these sequences. The class II CPD photolyase is one group, the 6-4 photolyase and the animal cryptochromes form another group and the class I CPD photolyase and plant and bacterial cryptochromes form the last group. The main difference is that the 6-4 photolyase and the animal cryptochromes do not form distinct groups within their common group. This indicates that the difference in sequence causing the functional divergence is not caused by any clear changes, or that neutral evolution has masked the changes.

Compared to the article by Kanai et al. (1997), the clustering shown here is more clear on there being three groups, with the plant cryptochromes and the class I photolyase belonging to one group. Also the ML tree from this article shows the group of sequences using 8-HDF as spread out in the tree, with no clear differentiation between the (6-4) photolyase and the class I photolyase. Based on

the variability present in the sequences shown in this thesis, the differences in the ML tree from Kanai et al. (1997) may be due to the choice of sequences. The results from using the ML program from the Molphy package showed that three of the 8-HDF sequences grouped together, with the fourth appearing alone.

5.5. Search for functional sites

The final step in this analysis, and the final goal, was to search the alignment for the functional sites causing the different functions of the protein groups. The first part here is the division into the functional groups; class I photolyase, plant cryptochromes, (6-4) photolyase, animal cryptochromes and class II photolyase. From the phylogenetic analysis it seems clear that only four of these five groups can with any certainty be classified based solely on the phylogeny without any prior knowledge. The separation that has been shown between (6-4) photolyase and animal cryptochromes are not possible to infer from the figures shown in this report. This could however be fixed with the addition of more sequences since the phylogenetic results also show that each class has one or more sequences that differs from the rest with an appreciable amount. In the sampling of the (6-4) photolyases, the most differing sequences may have been collected, leading to the results shown in this report, where the (6-4) photolyase sequences are not shown to be more similar to each other than to the animal cryptochromes.

Since all sequences are expected to use FADH as a cofactor, this would probably mean that the binding site should be conserved in all sequences. Using the conservation scheme found in the ClustalX program based on the positive scoring groups of the Gonnet250 substitution matrix, the multiple alignment were screened for sites showing either full conservation or some conservation. The requirement of allowing even weakly conserved sites may not be suitable from a biological viewpoint, since the requirement of active sites in some cases are rather strict. An amino acid functioning as an electron donor may not be easily replaced, unless the new amino acid also has the ability to transfer electrons. The low level of stringency were added however to see how many indicated sites appeared with each level of stringency.

The results when looking at conserved sites show 21 potential sites total with any level of conservation. When compared to the sites shown to bind FADH through crystallographic mapping of the enzyme and mutational research, only 5 out of the 10 functional sites are found among the 21 conserved sites; position 310, 313, 314, 362 and 460. This result shows that, for these genes at least, it would be difficult to pinpoint the active sites just by looking at conservation of amino acids. Particularly since several sites not included in the binding sites show more conservation than those that are included. These sites were picked based on all the amino acids being at least weakly conserved.This number does not increase significantly when looking at subsets of the total alignment, although the number of possible sites increases.

The method of electron transport to the FADH molecule has been determined to be the three tryptophan residues at positions 390, 445 and 470. These sites corresponds to position 307, 360 and 383 local to the *E.coli* sequence, which is off by one position compared to the coordinates mentioned by Byrdin et al. (W306, W359, W382). Unlike the FADH sites, these three sites were completely conserved across every sequence except for the sequences in the class II photolyase group. These results indicate that the electron transfer chain is more conserved than the residues binding to FADH.

This leads to the conclusion that the sequences may either have had a change in the structural conformation making the structural information from the *E.coli* (class I CPD photolyase) uninformative for any of the other groups, or it could be that the similarity between especially the

class II photolyase group and the rest that the alignment becomes too unreliable. Otherwise, the conservation groups of the amino acids may be wrong for this kind of analysis, or mistakes may be appearing in the sequences. This is indicated with the comparing of the alignment in this report (figure 4.9), and the alignment made by Kanai et al. (1997) (figure 1 in the appendix).This shows that the class II photolyase, at least, are not aligned in the same way. There are quite possibly other differences, but it shows the importance of having a good alignment when doing this kind of analysis.

Another point of interest when looking at a multiple alignment like this, are the positions that caused the split at each node. The first thing to look for will be positions which are conserved within each group originating from the particular node, but not conserved across the groups. The stringency for this search could at its lowest practical level be set to contain all sites where each group is at least weakly conserved, and the conservation between groups at most strongly conserved.

Searching for sites showing potentially functional divergence between the different groups, gave a list of possible sites. The total number varied based on the criteria used in the search. The least stringent criterion used was that each proposed group should at least be weakly conserved, and that there should be no strong conservation between each group. The results from the main node separating the three groups showed five sites which could, to a certain degree, be used to separate the groups. One of these five sites corresponded with a FADH binding site and could in that respect be the cause of changes in those proteins. The further analysis of the splits yielded a good deal more sites, depending on what level of conservation is expected or wanted. One example is the difference between the class I and class II photolyase groups. Five sites in the alignment are completely conserved with a different amino acid for each group. The five changes are; threonin (hydrophilic) changes to leucine (hydrophobic), tryptophan (hydrophobic) changes to aspartate (hydrophilic), methionine (hydrophobic) changes to glutamine (hydrophobic), phenylalanine (hydrophobic) changes to lysine (hydrophilic) and proline (slightly hydrophilic) changes to tryptophan (hydrophobic). Since four out of five of these changes are from a hydrophilic residue to a hydrophobic (or the opposite), there is at least a possibility of these sites being the cause of a change in function. This, as has been mentioned before, assumes that the two compared groups have been aligned properly.

The last possibility for functionally interesting sites is the case when one group shows conservation at a position while another group, which has lost the function, shows no conservation. For the latter group the site should show characteristics of being under a different regulation than for the former group. The search for this kind of change is based on the assumption that a mutation at a site resulting in the loss of function is not necessary the new active site for the new function. In that case the site may be less regulated than the corresponding position in the original protein.

6. Conclusion

The goal of this thesis was to look at the possibilities of finding the sites in a protein sequence that either were directly involved in the function of the protein, or else caused the change in function between to functionally different proteins. This was exemplified by the photolyase blue-light photoreceptor (cryptochrome) family. Part of the thesis was working in the laboratory trying to find the photolyase gene in a species based only on the sequence information of the gene as found in other species. The results clearly indicated that this principle quite likely would work. It also indicated a possible photolyase homolog in cotton, tomato, maize and cactus. Further research is needed to establish whether these are false indications or not.

The principle of searching for functional importance from a multiple sequence alignment seems to be a possibility, at least from a theoretical point of view. It is quite natural to assume that if an amino acid shows a higher degree of conservation than most of the others across several duplication events and speciation events, then there is a good chance this site has an important function. Finding these sites is a matter of screening the columns in the multiple alignment matrix.

However, the results from this project show that the photolyase and blue-light photoreceptor family may be too divergent for this method to work properly. While the three residues important for the electron transfer (Byrdin *et al.* 2003) were conserved in all but the sequences of the class II photolyase group, this was not the case for the residues binding to FADH. Here, some of the functional sites showed less conservation than other presumably more neutral sites. There is also the question if this kind of analysis would require a higher level of sequence identity, making the alignment less difficult and increasing the chance that all groups line up correctly with each other.

When it comes to figuring out the sites for functional divergence, the level of mutations also makes the assumption underlying the algorithm less good. It was assumed that the interesting sites for functional divergence were sites that would be strongly conserved in each group. With the amount of changes shown in all parts of the sequence, it seems the stringency for such a search has to be set so low that it will come up with too many possibilities.

7. Future possibilities

Future work on this project would, at one time or another, need some biologic testing to see if the sites indicated are important or not. This could be done by causing point mutations at specific sites and observe any changes in the response to UV damage. This method have been used by Hitomi et al. (2001), to show that two histidines in the (6-4) photolyase plays an important role, since the repair ability dropped to almost nothing when these sites were mutated to alanine. Doing the same thing to the proposed sites from this thesis could give an indication on the usefulness of these methods for searching the alignment.

One example where this kind of analysis might be useful is the classification of resistance genes. These genes contain several different kinds of resistance, depending on what's attacking at any given time. This might be intruders such as fungal, bacterial, nematode and viral pathogens. Since the intruders come in many different sizes and shapes, the resistence genes has to be able to detect several types of danger. One of the detection modes is a Nucletide Binding Sequence (NBS) that is a common characteristic for the resistance genes (Lee et al. 2003). Shirasu and Schulz-Lefert (2003) have shown that the resistance genes contain areas which are quite similar across different types of resistance genes. If the method from this project would work out, these resistance genes could be classified by a couple of amino acids in important spots for either functional divergence or tree determinant positions. Especially the resistance mechanism of killing the affected cell (the often called hypersensitive responses (HR) cell death) seems to appear in rather different resistance genes. Since the mechanism is so alike, it is possible that the underlying sequence information will be similar as well. Homologous genes isolated from species where the genome still is unknown, could be easily classified by sequencing the part of the gene where the required DNA is present. By making and analysing a multiple alignment it should be possible to show the sites which are conserved across species and also between functions thereby giving a hint to potential sites for doing point mutations later on.

8. Literature

Adachi J, Hasegawa M (1996) MOLPHY (programs for molecular phylogenetics) 2.3b3. Institute of Statistical Mathematics, Tokyo

Ahmad M, Jarillo JA, Smirnova O, Cashmore AR (1998) The CRY1 blue light photoreceptor of *Arabidopsis* interacts with phytochrome A *in vitro*. Molecular Cell 1: 939-948.

Ahmad M, Jarillo JA, Klimczak LJ, Landry LG, Peng T, Last RL, Cashmore AR (1997) An enzyme similar to animal type II photolyases mediates photoreactivation in *Arabidopsis*. Plant Cell 9 (2):199-207

Batschauer A. (1993) A plant gene for photolyase: an enzyme catalyzing the repair of UV-light-induced DNA damage. Plant Journal 4(4):705-709.

Brudler R, Hitomi K, Daiyasu H, Toh H, Kucho K, Ishiura M, Kanehisa M, Roberts VA, Todo T, Tainer JA, Getzoff ED. (2003) Identification of a new cryptochrome class. Structure, function, and evolution. Molecular Cell 11 (1):59-67

Byrdin M, Eker APM, Vos MH, Brettel K (2003) Dissection of the triple tryptophan electron transfer chain in Escherichia coli DNA photolyase: Trp382 is the primary donor in photoactivation. PNAS 15:8676-8681

Carell T, Burgdorf LT, Kundu LM, Cichon M (2001) The mechanism of action of DNA photolyases. Current Opinion in Chemical Biology 5:491-498

Cashmore AR, Jarillo JA, Wu Y-J, Liu D (1999) Cryptochromes: blue light receptors for plants and animals. Science 284:760-765

Dayhoff MO (1978) Survey of new data and computer methods of analysis. Pp 2-8 in M.O. Dayhoff, ed. Atlas of protein sequence and structure. National Biomedical Research Foundation, Silver Springs, Md.

Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis, probabilistic models of proteins and nucleic acids. Cambridge University Press, United Kingdom.

Felsenstein J (2004) Inferring phylogenies. Sinauer Associates Inc. Sunderland, Massachusetts

Felsenstein J (2004) PHYLIP (phylogeny inference package) version 3.6. Department of genetics, University of Washington, Seattle.

Gonnet GH, Cohen MA, Benner SA (1992) Exhaustive matching of the entire protein sequence database. Science 256:1443-1445

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. PNAS 89:10915-10919

Higgins DG, Sharp PM. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73,237-244.

Hirouchi T, Nakajima S, Najrana T, Tanaka M, Matsunaga T, Hidema J, Teranishi M, Fujino T,

Kumagai T, Yamamoto K (2003) A gene for a Class II DNA photolyase from Oryza sativa: cloning of the cDNA by dilution-amplification. Molecular Genetics and Genomics 269:508-516

Hitomi K, Nakamura H, Kim S-T, Mizukoshi T, Ishikawa T, Iwai S, Todo T (2001) Role of two histidines in the (6-4) photolyase reaction. The Journal of Biological Chemistry 13:10103-10109

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees BIOINFORMATICS 17 (8): 754-755

Ishikawa T, Matsumoto A, Kato T(Jr), Togashi S, Ryo H, Ikenaga M, Todo T, Ueda R, Tanimura T (1999) DCRY is a *Drosophilia* photoreceptor protein implicated in light entrainment of circadian rhythm. Genes to Cells 4:57-65

Jones DT, Taylor WR, Thornton JM. (1992) The rapid generation of mutation data matrices from protein sequences. Computer Applications in the Biosciences (CABIOS) 8: 275-282.

Kanai S, Kikuno R, Toh H, Ryo H, Todo T (1997) Molecular evolution of the Photolyase-Blue-Light Photoreceptor Family. Journal of Molecular Evolution 45:535-548

Kato T Jr., Todo T, Ayaki H, Ishizaki K, Morita T, Mitra S, Ikenaga M. (1994) Cloning of a marsupial DNA photolyase gene and the lack of related nucleotide sequences in placental mammals. Nucleic Acids Research 22 (20), 4119-4124

Kimura M (1980) A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences. Journal of molecular evolution 16:111-120

Kleine T, Lockhart P, Batschauer A (2003) An *Arabidopsis* protein closely related to *Synechocystis* cryptochrome is targeted to organelles. The Plant Journal 35:93-103

Kobayashi Y, Ishikawa T, Hirayama J, Hiromi D, Kanai S, Toh H, Fukuda I, Tsujimura T, Terada N, Kamei Y, Yuba S, Iwai S, Todo T (2000) Molecular analysis of zebrafish photolyase/cryptochrome family: two types of cryptochromes present in zebrafish. Genes to Cells 5:725-738

Komori H, Masui R, Kuramitsu S, Yokoyama S, Shibata T, Inoue Y, Miki K (2001) Crystal structure of thermostable DNA photolyase: Pyrimidine-dimer recognition mechanism. PNAS (98) 24:13560-13565

Lee SY, Seo JS, Rodriguez-Lanetty M, Lee DH (2003) Comarative analysis of superfamiles of NBS encoding disease resistance gene analogs in cultivated and wild apple sedes. Molecular Genetics and Genomics 269: 101-108.

Li YF, Sancar A. (1991) Cloning, sequencing, expression and characterization of DNA photolyase from Salmonella typhimurium. Nucleic Acids Research. 19(18):4885-4890.

Lin C (2002) Blue light receptors and signal transduction. The Plant Cell Supplement S207-S225

Nakajima S, Sugiyama M, Iwai S, and more (1998) Cloning and characterization of a gene (UVR3) required for photorepair of 6-4 photoproducts in *Arabidopsis thaliana*. Nucleic Acids Research 26:638-644

Park H-W, Kim S-T, Sancar A, Deisenhofer J (1995) Crystal structure of DNA photolyase from

Escherichia coli. Science 268:1868-1872

Petersen JL, Lang DW, Small GD (1999) Cloning and characterization of a class II DNA photolyase from *Chlamydomonas*. Plant Molecular Bology 40:1063-1071

Rokas A, Williams BL, King N, Carroll SB (2003) Genome scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798-804

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 7:406-425

Sancar A (2003) Structure and function of DNA photolyase and cryptochrome blue-light photoreceptors Chemical Review 103:2203-2237

Schroder HC, Krasko A, Gundacker D, Leys SP, Muller IM, Muller WE (2003) Molecular and functional analysis of the (6-4) photolyase from the hexactinellid *Aphrocallistes vastus*. Biochim. Biophys. Acta 1651 (1-2) 41-49

Shirashu K, Schulze-Lefert P (2003) Complex formation, promiscuity and multi-functionality: protein interactions in desease-resistane pathways. TRENDS in Plant Science 8 (6) -

del Sol Mesa A, Pazos F, Valencia A (2003) Automatic methods for predicting functionally important residues. Journal of Molecular Biology 326: 1289-1302.

Strimmer K, von Haesler A (1996) Quartet Puzzling: A quartet Maximum-Likelihood method for reconstructing tree topologies. Molecular and Biological Evolution 13 (7) : 964-969.

Tamura Nei (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. Molecular Biology Evolution 10 (3): 512-526

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research, 24:4876-4882.

Todo T, Kim ST, Hitomi K, Otoshi E, Inui T, Morioka H, Kobayashi H, Ohtsuka E, Toh H, Ikenaga M (1997) Flavin adenine dinucleotide as a chromophore of the Xenopus (6-4) photolyase. Nucleic Acids Research 25 (4) 764-768

Todo T, Ryo H, Yamamoto K, Toh H, Inui T, Ayaki H, Ikenaga M (1996) Similarity among the *Drosophilia* (6-4) photolyase, a human photolyase homolog, and the DNA photolyase-blue-light photoreceptor family. Science 272:109-112

Todo T, Takemori H, Ryo H, Ihara M, Matsunaga T, Nikaido O, Sato K, Nomura T (1993) A new photoreactivating enzyme that specifically repairs ultraviolet light induced (6-4) photoproducts. Nature 361:371-374

Vingron M, Waterman MS (1994) Sequence alignment and penalty choice: review of concepts, case studies and implications. Journal of Molecular Biology 235: 1-12.

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Freijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP: a new concept for DNA fingerprinting. Nucleic Acids Research 23 (21): 4407-441

Yamamoto K, Okano T, Fukada Y (2001) Chicken pineal Cry genes: light-dependent up-regulation of cCry1 and cCry2 transcripts. Neuroscience Letters 313 (1-2), 13-16

Yasui A, Eker AP, Yasuhira S, Yajima H, Kobayashi T, Takao M, Oikawa A. (1994) A new class of DNA photolyases present in various organisms including aplacental mammals. EMBO Journal 13(24):6143-6151

Özgür S, Sancar A (2003) Purification and properties of human blue-light photoreceptor cryptochrome 2. Biochemistry 42:2926-2932

Appendix

Appendix A: The article by Kanai *et al.* (1997) that inspired this thesis.

J Mol Evol (1997) 45:535–548

JOURNAL OF MOLECULAR EVOLUTION

© Springer-Verlag New York Inc. 1997

Molecular Evolution of the Photolyase–Blue-Light Photoreceptor Family

Satoru Kanai,[1] Reiko Kikuno,[1] Hiroyuki Toh,[1] Haruko Ryo,[2] Takeshi Todo[3]

[1] Department of Bioinfomatics, Biomolecular Engineering Research Institute, 6-2-3, Furuedai, Suita, Osaka, 565 Japan
[2] Department of Radiation Biology, Faculty of Medicine, Osaka University, 2-2 Yamada-oka, Suita, Osaka, 565 Japan
[3] Radiation Biology Center, Kyoto University, Yoshida-konoecho, Sakyo-ku, Kyoto 606-01 Japan

Received: 23 October 1996 / Accepted: 1 April 1997

Abstract. The photolyase–blue-light photoreceptor family is composed of cyclobutane pyrimidine dimer (CPD) photolyases, (6-4) photolyases, and blue-light photoreceptors. CPD photolyase and (6-4) photolyase are involved in photoreactivation for CPD and (6-4) photoproducts, respectively. CPD photolyase is classified into two subclasses, class I and II, based on amino acid sequence similarity. Blue-light photoreceptors are essential light detectors for the early development of plants. The amino acid sequence of the receptor is similar to those of the photolyases, although the receptor does not show the activity of photoreactivation. To investigate the functional divergence of the family, the amino acid sequences of the proteins were aligned. The alignment suggested that the recognition mechanisms of the cofactors and the substrate of class I CPD photolyases (class I photolyases) are different from those of class II CPD photolyases (class II photolyases). We reconstructed the phylogenetic trees based on the alignment by the NJ method and the ML method. The phylogenetic analysis suggested that the ancestral gene of the family had encoded CPD photolyase and that the gene duplication of the ancestral proteins had occurred at least eight times before the divergence between eubacteria and eukaryotes.

Key words: Evolution — Protein — DNA photolyase — Blue-light photoreceptor — Phylogenetic analysis — Photoreactivation

Introduction

UV light induces DNA damage such as CPDs and (6-4) photoproducts. Corresponding to the damage, there are two types of DNA photolyases. One of them specifically exerts the activity on CPD whereas another repairs only (6-4) photoproducts (reviewed by Sancar 1996). The former is called CPD photolyase and the latter (6-4) photolyase. The amino acid sequences of the two enzymes are similar to each other (Todo et al. 1996), although the chemical structures of their substrates are quite different (Brash 1988; Mitchell and Nairn 1989; Taylor 1995).

CPD photolyases are divided into two subclasses, class I and II, based on the sequence similarity, which is here referred to as class I photolyases and class II photolyases, respectively. CPD photolyases contain two types of cofactors. One of them is two-electron–reduced FAD, which acts as the active-site cofactor. Another cofactor is MTHF, or 8-HDF, which acts as the photoantenna. Based on the kind of photoantenna, class I photolyases are further divided into two subgroups, which are here denoted as MTHF-type photolyase and 8-HDF–type photolyase, respectively (Yasui et al. 1994). The photoreactivation mechanism of MTHF-type or 8-HDF–type photolyases has been investigated in great detail: (1)

Abbreviations: UV, ultraviolet; CPD, cyclobutane pyrimidine dimer; (6-4) photoproducts, pyrimidine (6-4) pyrimidone photoproducts; FAD, flavin-adenine dinucleotide; MTHF, 5,10-methenyltetrahydrofolate; 8-HDF, 8-hydroxy-5-deazaflavin; NJ, neighbor-joining; ML, maximum likelihood; AIC, Akaike information criterion

Correspondence to: S. Kanai; e-mail skanai@beri.co.jp

the photoantenna (MTHF or 8-HDF) absorbs a blue-light photon; (2) the excitation energy is transferred from the photoantenna to active-site cofactor (FAD) by dipole–dipole interaction; (3) the excited FADH radical donates an electron to a CPD, splitting the cyclobutane ring; and (4) finally, the electron is transferred back to FADH, accompanied by the generation of the two canonical bases (Hearst 1995; Kim et al. 1991, 1992). The efficiency of energy transfer from 8-HDF to FAD is 98% in eubacterium *Anacystis nidulans* 8-HDF–type photolyase, while that from MTHF to FAD is 63% in an *Escherichia coli* MTHF-type photolyase (Kim et al. 1992). Therefore, it was suggested that 8-HDF–type photolyases are able to eliminate CPD more rapidly than MTHF-type photolyases (Malhotra et al. 1992). The crystal structure of MTHF-type photolyase from *E. coli* was determined (Park et al. 1995). The structure consists of two domains, an N-terminal α/β domain and a C-terminal helical domain, connected by a loop of 72 residues. MTHF binds in a cleft between the two domains, whereas FAD is included in the helical domain and is accessible through a hole in the surface of the domain. CPD binding sites have not been identified, but the hole in the helical domain has shape and polarity suitable for CPD binding. Therefore, the hole is considered to correspond with putative CPD binding site (Park et al. 1995). In contrast, investigation of the photoreactivation mechanism of class II photolyases has not advanced. However, the cofactors of a few class II enzymes have been elucidated. The class II photolyase from *Drosophila melanogaster* has FAD and MTHF (Kim et al. 1996a), while FAD and 8-HDF are included in the enzyme from an archaebacterium, *Methanobacterium thermoautotrophicum* (Kiener et al. 1989). Yasui et al. (1994) reported that the enzyme from a eukaryote, *Potorous tridactylis,* includes FAD, although the second cofactor has not been detected in the enzyme.

(6-4) photolyases have been identified recently (Todo et al. 1993), but the presence of cofactors has not been confirmed and the precise photoreactivation mechanism has not been well elucidated. The fluorescence and action spectra of *Xenopus laevis* (6-4) photolyase suggested that the cofactor composition of the enzyme is different from those of CPD photolyases (Kim et al. 1996b). On the other hand, it was suggested that the enzyme includes FAD as the cofactor of the *X. laevis* enzyme (Todo et al. 1997). Further studies are required for the identification of the cofactors for (6-4) photolyases. Two (6-4) photolyase homologues have been found in *Homo sapiens* (Todo et al. 1996; Hsu et al. 1996). However, the homologues, which contain FAD and MTHF as the cofactors, show neither (6-4) nor CPD photolyase activity (Hsu et al. 1996). The function of the homologues has been unknown, although Hsu et al. (1996) suggested that the homologues may act as blue-light photoreceptors.

Blue-light photoreceptors are essential light detectors for the early development of plants (Ahmad and Cashmore 1993). In addition, they mediate phototropism, hypocotyl elongation, stomatal opening, and expression of specific genes (Kaufman 1993; Short and Birggs 1994; Liscum and Hangarter 1994). The nucleotide sequences of the cDNAs for the receptors were recently determined (Ahmad and Cashmore 1993; Batschauer 1993; Malhotra et al. 1995). The deduced amino acid sequences were unexpectedly similar to those of photolyases (Ahmad and Cashmore 1993; Batschauer 1993; Malhotra et al. 1995). The receptors from the eukaryotes *Sinapis alba* and *Arabidopsis thaliana* overexpressed in *E. coli* also contain FAD and MTHF as the cofactors, although the receptors do not show photoreactivation activity (Malhotra et al. 1995).

Table 1 summarizes the classification of the family. The functions carried by the members of the protein family are highly divergent. However, the evolutionary process of the functional divergence of the family has not been sufficiently studied. Yasui et al. (1994) reported the molecular phylogeny of CPD photolyases by the NJ method (Saitou and Nei 1987), although they used only 13 sequences, which did not include (6-4) photolyases or blue-light photoreceptors. Now, 22 sets of sequence data of this family are available. We compared the amino acid sequences to reveal the evolutionary relationship among the members of the photolyase–blue-light photoreceptor family. The evolutionary history of the family will be discussed based on the reconstructed phylogenetic trees.

Materials and Methods

The 22 amino acid sequences used in this study are listed in Table 1. A multiple sequence alignment was constructed with the program Clustal W 1.4 (Higgins et al. 1991; Thompson et al. 1994), which was then modified by visual inspection with an alignment editor, BIORESEARCH/AE 3.0 (Fujitsu Ltd. 1995). To examine the conservation of aligned sites, amino acid residues were classified into six physicochemically similar groups based on the criteria of Schwartz and Dayhoff (1978): (1) positively charged group: Lys, Arg, His; (2) negatively charged group: Asp, Asn, Glu, Gln; (3) small hydrophilic group: Gly, Pro, Ser, Thr; (4) hydrophobic group: Leu, Ile, Met, Val; (5) aromatic group: Phe, Tyr, Trp; (6) Cys. The last category consists of only Cys residue.

All the sites containing gaps were excluded from the alignment for the following phylogenetic analyses. At first, we performed the NJ inference to obtain an overview of the evolutionary relationship of the photolyase–blue-light photoreceptor family. The genetic distance for each aligned pair was calculated with the program PROTDIST in PHYLIP 3.5c (Felsenstein 1993, 1996), where the amino acid substitution model, PAM001 (Dayhoff et al. 1978), was used. Then, an unrooted NJ tree was constructed with NEIGHBOR in PHYLIP 3.5c. The statistical significance of each cluster in the tree was evaluated with 1,000 iterations of bootstrap resamplings and tree reconstructions (Felsenstein 1985) using PROTDIST, NEIGHBOR, SEQBOOT, and CONSENSE in PHYLIP 3.5c.

To further examine the phylogenetic relationship obtained by the NJ method, we employed the ML method (Felsenstein 1981; Kishino et

Table 1. List of proteins used in the current analyses[a]

Type of protein (abbreviation)	Species	Kingdom	Abbreviated name	Accession No. (DB)/References	Characterized cofactors
MTHF type class I CPD photolyase (C1MPHR)	*Bacillus firmus*	Eubacterium	PHR_BACF	Q04449 (sp)	FAD, MTHF (Malhotra et al. 1994)
	Escherichia coli	Eubacterium	PHR_ECOL	P00914 (sp)	FAD, MTHF (Johnson et al. 1988)
	Salmonella typhimurium	Eubacterium	PHR_SALT	P25078 (sp)	FAD, MTHF (Li and Sancar 1991)
	Saccharomyses cerevisiae	Eukaryote	PHR_SCER	P05066 (sp)	FAD, MTHF (Sancar et al. 1987; Johnson et al. 1988)
	Neurospora crassa	Eukaryote	PHR_NEUC	P27526 (sp)	FAD, MTHF (Eker et al. 1994)
8-HDF type class I CPD photolyase (C18PHR)	*Anacystis nidulans*	Eubacterium	PHR_ANAN	P05327 (sp)	FAD, 8-HDF (Eker et al. 1990)
	Synechocystis sp.	Eubacterium	PHR_SYNE	U51943 (gb)	
	Streptomyces griseus	Eubacterium	PHR_STRG	P12768 (sp)	FAD, 8-HDF (Eker et al. 1981)
	Halobacterium halobium	Archaebacterium	PHR_HALH	P20377 (sp)	FAD, 8-HDF (Iwasa et al. 1988)
(6-4) photolyase (64PHR)	*Drosophila melanogaster*	Eukaryote	64_DROM	D83701 (gb)	
(6-4) photolyase homologue (64PHR)	*Homo sapiens*	Eukaryote	64_HUMA1	D83702 (gb)	FAD, MTHF (Hsu et al. 1996)
	Homo sapiens	Eukaryote	64_HUMA2	Hsu et al. 1996	FAD, MTHF (Hsu et al. 1996)
Blue-light photoreceptor (BLR)	*Chlamydomonas reinhardtii*	Eukaryote	BLR_CHRE	S57795 (pir)	
	Sinapis alba	Eukaryote	BLR_SIAL	P40115 (sp)	FAD, MTHF (Malhotra et al. 1995)
	Arabidopsis thaliana	Eukaryote	BLR_ARTH	S66907 (gb)	FAD, MTHF (Malhotra et al. 1995)
Class II CPD photolyase (C2PHR)	*Myxococcus xanthus*	Eubacterium	PHR_MXU	U44437 (gb)	
	Methanobacterium thermoautotrophicum	Archaebacterium	PHR_METH	P12769 (sp)	FAD, 8-HDF (Kiener et al. 1989)
	Drosophila melanogaster	Eukaryote	PHR_DM	S52047 (pir)	FAD, MTHF (Kim et al. 1996a)
	Carassius auratus	Eukaryote	PHR_CA	A45098 (pir)	
	Oryzias latipes	Eukaryote	PHR_OLAP	D26022 (gb)	
	Monodelphis domestica	Eukaryote	PHR_OPPO	D31902 (gb)	
	Potorous tridactylis	Eukaryote	PHR_PTRI	D26020 (gb)	FAD (Yasui et al. 1994)

[a] ''DB'' denotes the symbol of database : sp, SWISS-PROT release 33.0; gb, GenBank release 95.0; pir, PIR protein sequence database release 49.0

al. 1990). However, it was difficult to apply the ML analysis to all the 22 sequences, since a large number of possible trees are generated. We, therefore, selected the proteins derived from nodes with less than 50% bootstrap probabilities in the NJ analysis, as well as the representative proteins of each subgroup. Then, nine amino acid sequences were selected, which were subjected to the ML analysis with PROTML in MOLPHY 2.3b3 (Adachi and Hasegawa 1996). In the analysis, AIC of each possible tree was calculated based on the four different amino acid substitution models—the JTT (Jones et al. 1992), Dayhoff (Dayhoff et al. 1978), JTT-F, and Dayhoff-F models (Adachi and Hasegawa 1995). AIC is defined as $-2 \times$ (log-likelihood) $+ 2 \times$ (number of free parameters) (Akaike 1974). The tree with minimal AIC was considered to be the most appropriate tree. The statistical significance of each cluster in the ML tree was evaluated by the bootstrap analysis with 1,000 iterations.

The obtained phylogenetic trees were drawn with TREETOOL 2.0.2 (Maciukenas and McCaughey 1994).

Results and Discussion

Functional Implications Derived From Multiple Sequence Alignment

Figure 1 shows a multiple alignment of the photolyase–blue-light photoreceptor family. Since the N- and C-terminal regions of the members were highly diverged in residue and length, they were excluded from the alignment. The crystal structure of class I photolyase from *E. coli* has been determined recently (Park et al. 1995), which revealed MTHF and FAD binding sites of the enzyme. In addition, the CPD binding sites were putatively assigned to a hole of the crystal structure. These sites play important roles for the function of the enzyme. Therefore, it is expected that the comparison of the residues at these sites would provide much information about the functional divergence of the protein family. Table 2 shows the summary of the comparative study, although the alignment in Fig. 1 also includes the same information. As shown in Table 2, the protein family was divided into five groups based on the functional difference as follows: group A, MTHF-type photolyases; group B, 8-HDF–type photolyases; group C, (6-4) photolyases; group D, blue-light photoreceptors; and group E, class II photolyases. For simplicity, (6-4) photolyase homologues were included in group C, although their functions have remained unknown. The functional classification roughly corresponded with the phylogenetic clustering as described below. In the current approach, class I photolyases were classified into two groups, based on the available second cofactors (groups A and B). In contrast, group E, or the group of class II CPD photolyases, was not further classified, because identification of the second cofactors of the photolyase has not advanced (see Table 1).

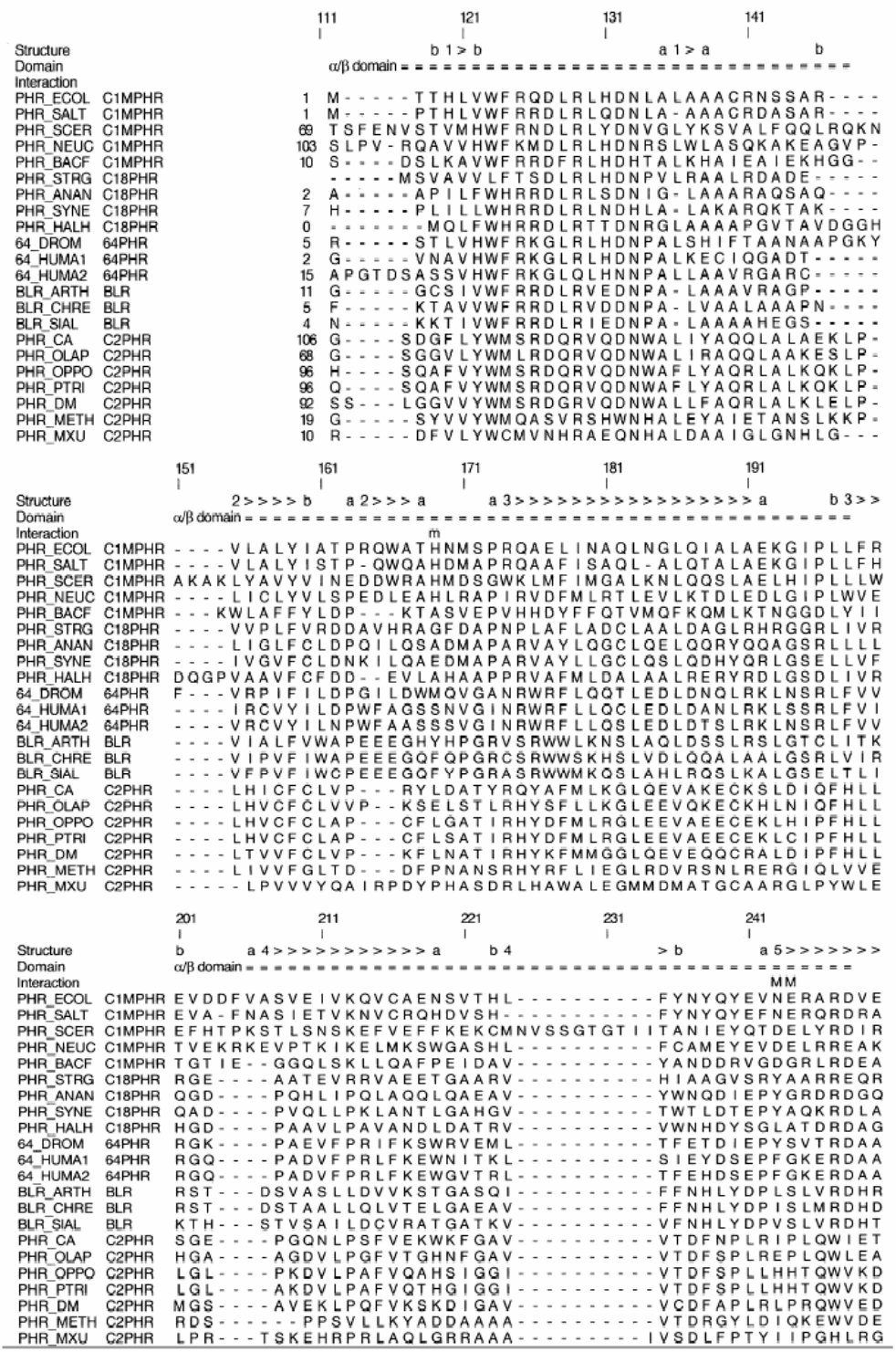


Fig. 1. Alignment of the amino acid sequences of DNA photolyases and blue-light photoreceptors. The *residue number* of beginning amino acid of each sequence is shown at the left side of the sequence. The information about the secondary structures, domains, interactions with cofactors, binding with CPD, and electron transfer derived from the crystal structure of DNA photolyase from *E. coli* (Park et al. 1995) is also shown in the alignment. The line ''*Structure*'' shows the secondary structures (a: α helix, b: β strands, 3: 3_{10} helix). For example, in the case of ''an>>>>a,'' where ''an'' means the beginning of the *n*-th α helix, ''>'' indicates the continuation of the structure. The last ''*a*'' denotes the end of the structure. The line ''*Domain*'' indicates the regions of the two domains and an interdomain loop in the DNA photolyase. The line ''*Interaction*'' denotes the residue sites which interact with the cofactors. The characters ''*f*'' and ''*F*'' indicate residues which interact with FAD, while ''*m*'' and ''*M*'' indicate residues which interact with MTHF. The *capital letter* means the direct H bond to the cofactors. The *lowercase letter* indicates the water-mediated H bond to the cofactors. ''*C*'' shows putative CPD binding site. Abbreviated name of each sequence is shown in Table 1.

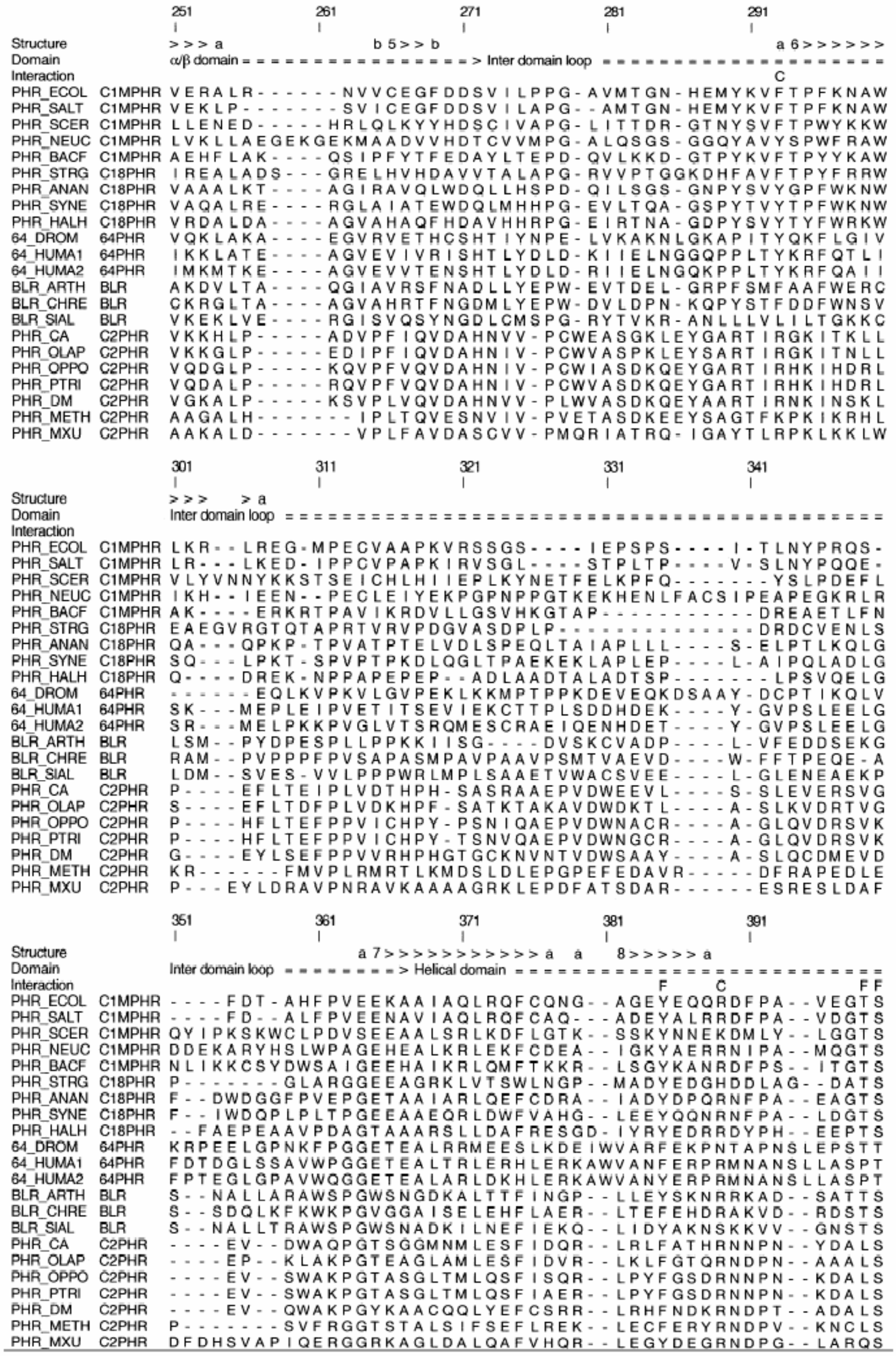


Fig. 1. Continued.

Table 2(a) shows the comparison of the MTHF binding residues. X-ray crystallographic analysis revealed that seven residues of MTHF-type photolyase from *E. coli* (PHR_ECOL in Table 2) are involved in the MTHF binding function of the enzyme. Surprisingly, five out of the seven residues were not conserved among the members of group A. Only Glu residue at alignment site 536 was invariant, and alignment site 243 was occupied by the residues of the negatively charged group. However, the remaining five sites included physicochemically different residues. Payne et al. (1990) reported that MTHF-type photolyase from *E. coli* shows photolyase activity,

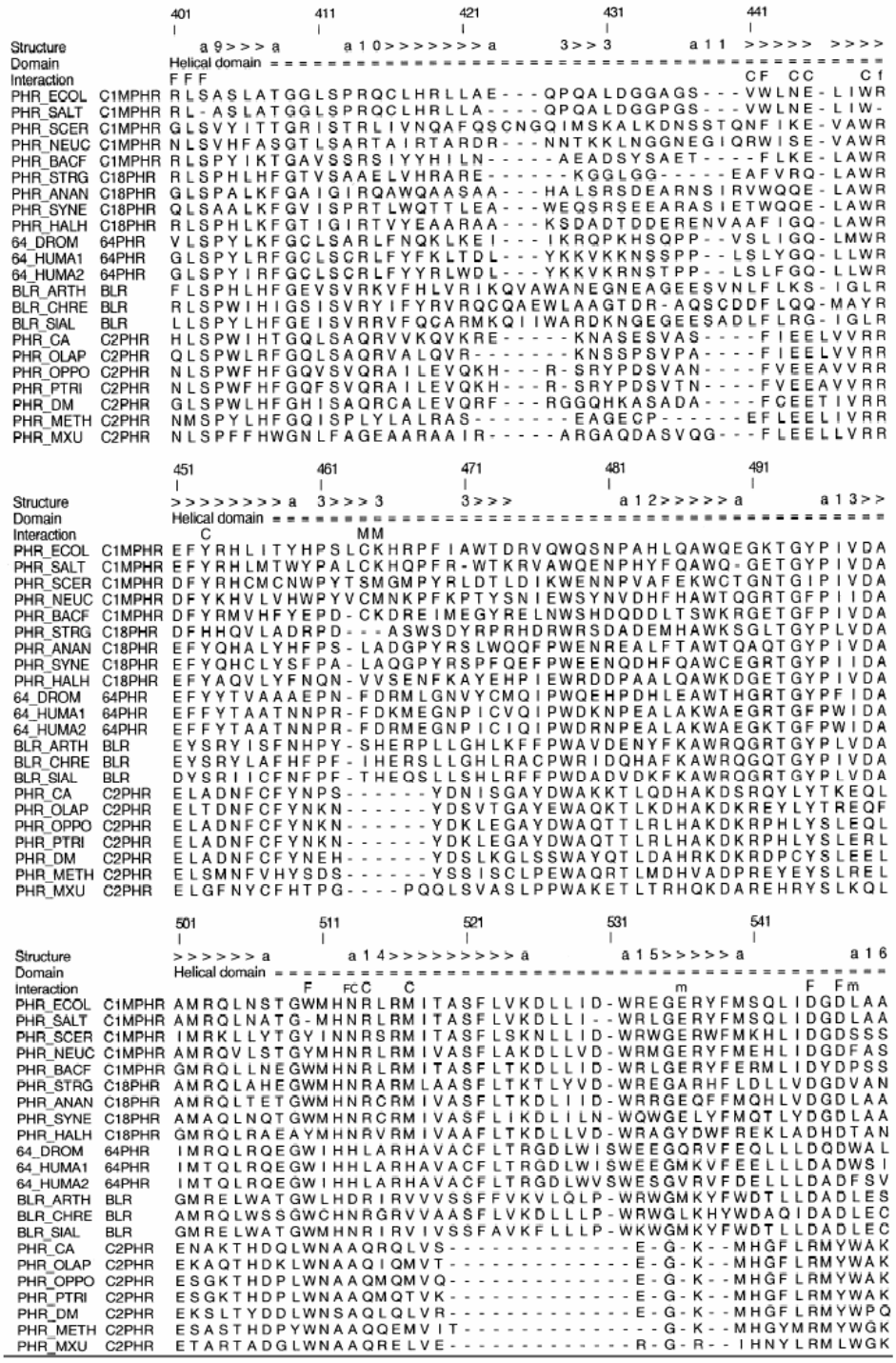


Fig. 1. Continued.

even without MTHF. The observation suggested that the second cofactor or photoantenna is not essential for the photolyase activity, which could explain the weak conservation of the MTHF binding sites among the members of group A. Blue-light photoreceptors utilize MTHF as the second cofactor as well. However, these sites of the members of group D were occupied by physicochemically different residues from those of PHR_ECOL, even at the alignment sites 243 and 536. The high variability of MTHF binding sites in the members of group and D could be explained by the weak functional constraint described above. Another possible interpretation of the

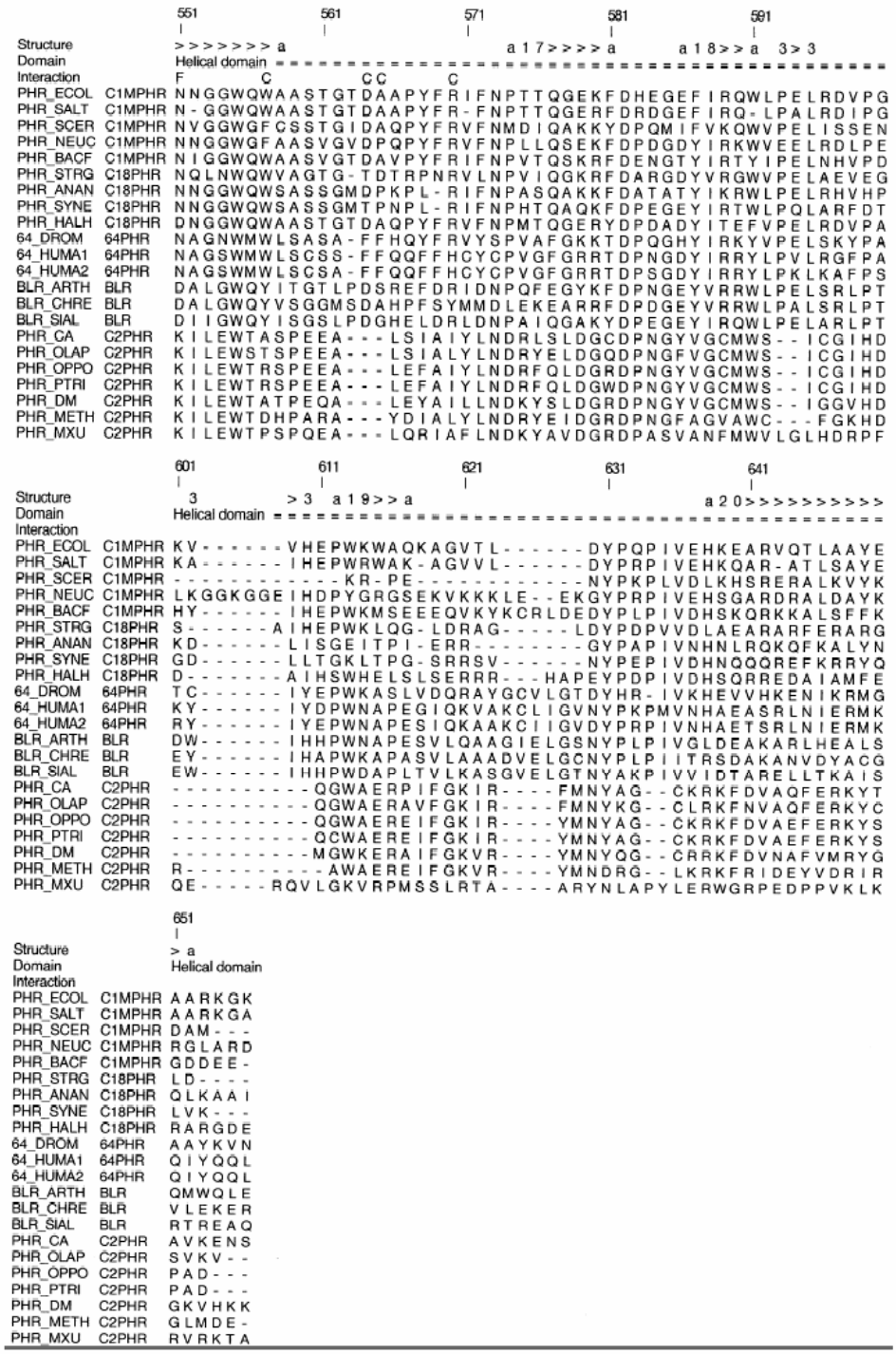


Fig. 1. Continued.

observation is that the members of group D may have a different MTHF binding mechanism than those of group A. It is difficult to definitively describe the problem at this stage, and further study is required. There were no data available to identify the 8-HDF binding sites of the members of group B. Therefore, we just examined the residues of the members of group B, which correspond to the seven MTHF binding sites. As expected, these sites were not conserved among the members of group B. In addition, the physicochemical characters of the amino acid residues at the five alignment sites, 169, 243, 244, 464 and 465, of the members of group B were quite

Table 2. Comparisons of residues of (a) MTHF, (b) FAD, and (c) CPD binding sites[a]

(a) MTHF binding sites

		Site	169	243	244	464	465	536	548
		Interaction	m	M	M	M	M	m	m
Group A	PHR_ECOL	C1MPHR	H	N	E	C	K	E	L
	PHR_SALT	C1MPHR	H	N	E	C	K	E	L
	PHR_SCER	C1MPHR	H	D	E	S	M	E	S
	PHR_NEUC	C1MPHR	H	D	E	C	M	E	F
	PHR_BACF	C1MPHR	A	D	G	C	K	E	P
Group B	PHR_STRG	C18PHR	G	Y	A	–	–	A	V
	PHR_ANAN	C18PHR	A	Y	G	L	A	E	L
	PHR_SYNE	C18PHR	E	Y	A	L	A	E	L
	PHR_HALH	C18PHR	A	L	A	V	V	Y	T
Group C	64_DROM	64PHR	M	Y	S	F	D	Q	W
	64_HUMA1	64PHR	S	F	G	F	D	M	W
	64_HUMA2	64PHR	S	F	G	F	D	V	F
Group D	BLR_ARTH	BLR	Y	L	S	S	H	M	L
	BLR_CHRE	BLR	F	I	S	I	H	L	L
	BLR_SIAL	BLR	F	V	S	T	H	M	L
Group E	PHR_CA	C2PHR	L	I	P	–	–	–	W
	PHR_OLAP	C2PHR	E	E	P	–	–	–	W
	PHR_OPPO	C2PHR	L	H	H	–	–	–	W
	PHR_PTRI	C2PHR	L	H	H	–	–	–	W
	PHR_DM	C2PHR	L	L	P	–	–	–	W
	PHR_MXU	C2PHR	P	I	I	–	–	–	W
	PHR_METH	C2PHR	P	I	Q	–	–	–	W

(b) FAD binding sites

		Site	385	399	400	401	402	403	442	450	510	513	545	547	550	551
		Interaction	F	F	F	F	F	F	F	f	F	F	F	F	f	F
Group A	PHR_ECOL	C1MPHR	Y	T	S	R	L	S	W	R	W	N	D	D	A	N
	PHR_SALT	C1MPHR	Y	T	S	R	L	–	W	–	–	N	D	D	A	N
	PHR_SCER	C1MPHR	Y	T	S	G	L	S	F	R	Y	N	D	D	S	N
	PHR_NEUC	C1MPHR	Y	T	S	N	L	S	W	R	Y	N	D	D	S	N
	PHR_BACF	C1MPHR	Y	T	S	R	L	S	F	R	W	N	D	D	S	N
Group B	PHR_STRG	C18PHR	Y	T	S	R	L	S	F	R	W	N	D	D	N	N
	PHR_ANAN	C18PHR	Y	T	S	G	L	S	W	R	W	N	D	D	A	N
	PHR_SYNE	C18PHR	Y	T	S	Q	L	S	W	R	W	N	D	D	A	N
	PHR_HALH	C18PHR	Y	T	S	R	L	S	F	R	Y	N	D	D	N	D
Group C	64_DROM	64PHR	F	T	T	V	L	S	L	R	W	H	D	D	L	N
	64_HUMA1	64PHR	F	P	T	G	L	S	L	R	W	H	D	D	I	N
	64_HUMA2	64PHR	Y	P	T	G	L	S	L	R	W	H	D	D	V	N
Group D	BLR_ARTH	BLR	Y	T	S	F	L	S	F	R	W	D	D	D	S	D
	BLR_CHRE	BLR	F	T	S	R	L	S	F	R	W	N	D	D	C	D
	BLR_SIAL	BLR	Y	T	S	L	L	S	F	R	W	N	D	D	C	D
Group E	PHR_CA	C2PHR	F	L	S	H	L	S	F	R	W	A	R	Y	K	K
	PHR_OLAP	C2PHR	F	L	S	Q	L	S	F	R	W	A	R	Y	K	K
	PHR_OPPO	C2PHR	F	L	S	N	L	S	F	R	W	A	R	Y	K	K
	PHR_PTRI	C2PHR	F	L	S	N	L	S	F	R	W	A	R	Y	K	K
	PHR_DM	C2PHR	F	L	S	G	L	S	F	R	W	A	R	Y	Q	K
	PHR_MXU	C2PHR	Y	Q	S	N	L	S	F	R	W	A	R	L	K	K
	PHR_METH	C2PHR	F	L	S	N	M	S	F	R	W	A	R	Y	K	K

different from those of group A, although the several residues at remaining two sites, 536 and 548, were similar or identical between the members of groups A and B. The residues of the seven sites of the members of group B seemed to be rather similar to those of group D, although the functional meaning of the similarity was not clear. Here, we refrain from discussion of the second cofactor binding sites of groups C and E, because the

Table 2. Continued

(c) Putative CPD binding sites

			Hydrophobic							Polor						
		Site	293	441	449	442	517	557	565	389	444	445	513	514	564	570
Group A	PHR_ECOL	C1MPHR	F	V	W	Y	M	W	A	R	N	E	N	R	D	R
	PHR_SALT	C1MPHR	F	V	W	Y	M	W	A	R	N	E	N	R	D	R
	PHR_SCER	C1MPHR	F	N	W	Y	M	F	A	K	K	E	N	R	D	R
	PHR_NEUC	C1MPHR	Y	R	W	Y	M	F	P	R	S	E	N	R	D	R
	PHR_BACF	C1MPHR	F	–	W	Y	M	W	A	R	K	E	N	R	D	R
Group B	PHR_STRG	C18PHR	F	A	W	H	M	W	D	H	R	Q	N	R	T	R
	PHR_ANAN	C18PHR	Y	V	W	Y	M	W	P	R	Q	E	N	R	D	R
	PHR_SYNE	C18PHR	Y	T	W	Y	M	W	P	R	Q	E	N	R	T	R
	PHR_HALH	C18PHR	Y	A	W	Y	M	W	A	R	G	Q	N	R	D	R
Group C	64_DROM	64PHR	Y	S	W	Y	H	W	F	N	G	Q	H	L	F	R
	64_HUMA1	64PHR	Y	S	W	F	H	W	F	R	G	Q	H	L	F	H
	64_HUMA2	64PHR	Y	S	W	F	H	W	F	R	G	Q	H	L	F	H
Group D	BLR_ARTH	BLR	F	L	L	S	V	Y	S	R	K	S	D	R	D	R
	BLR_CHRE	BLR	F	D	Y	S	V	Y	A	R	Q	Q	N	R	D	Y
	BLR_SIAL	BLR	L	L	L	S	V	Y	G	S	R	G	N	R	D	R
Group E	PHR_CA	C2PHR	R	–	R	A	L	A	–	R	E	E	A	Q	–	I
	PHR_OLAP	C2PHR	R	–	R	T	M	T	–	R	E	E	A	Q	–	L
	PHR_OPPO	C2PHR	R	–	R	A	M	R	–	R	E	E	A	Q	–	I
	PHR_PTRI	C2PHR	R	–	R	A	T	R	–	R	E	E	A	Q	–	I
	PHR_DM	C2PHR	R	–	R	A	L	A	–	R	E	E	A	Q	–	I
	PHR_MXU	C2PHR	R	–	R	G	L	P	–	R	E	E	A	Q	–	A
	PHR_METH	C2PHR	K	E	R	S	M	D	–	R	E	E	A	Q	–	L

[a] The information is derived from the alignment in Fig. 1. See legend of Fig. 1 for notation of the characters in the line ''*Interaction*''

second cofactors of (6-4) photolyases and the most of class II enzymes have not been identified.

FAD is considered to be included in all the members of the protein family, and 14 FAD binding amino acid residues have been found in the crystal structure of PHR_ECOL. Alignment sites 385, 400, and 402 were occupied by physicochemically similar residues, respectively. Alignment sites 403 and 450 were invariant except for the deletion in MTHF-type photolyase from *Salmonella typhimurium* (PHR_SALT). Alignment site 510 was occupied by aromatic residues, which was also deleted in PHR_SALT. Alignment site 442 was also occupied by aromatic residues, except for three members of group C. Thus, FAD binding sites seemed to be highly conserved, comparing to the case of MTHF binding sites. However, alignment sites 399, 513, 545, 547, and 551 were occupied by the residues with different physicochemical characters between group E and the other groups. As discussed below, group E was distantly related to the other members of the family. In addition, the members of group E included a long deletion in the C-terminal helical domain (alignment sites 520 to 539). Such a deletion was not found in the other members of the family. The different conservation pattern between class II photolyase and the others may reflect the change of FAD binding mechanism caused by the deletion in class II photolyase or the insertion in the others.

CPD binding sites were putatively assigned in the crystal structure of PHR_ECOL. The putative sites constitute a hole in the structure, to which CPD is supposed to be bound. One face of the hole consists of seven hydrophobic residues, while seven polar residues form another surface of the hole. Table 2(c) summarizes the comparison of the residues at the putative CPD binding sites. CPD is the substrate for the members of both group A and B. Alignment sites 449, 513, 514, 517, and 570 were invariant among the members of groups A and B. In addition, four sites, 293, 389, 445, and 557, were occupied by physicochemically similar residues, and two sites, 442 and 564, were nearly invariant. Site 565 was occupied by physicochemically similar residues, except for 8-HDF–type photolyase from *Streptomyces griseus* (PHR_STRG). The observation suggested that these sites are important for CPD binding activity of the members of groups A and B. The members of group D follow a conservation pattern similar to those of groups A and B, although the members of group D do not have photolyase activity. The members of group C share residues similar or identical to those of groups A, B, and D at the alignment sites 293, 389, 513, and 570. However, the amino acid residues at the alignment sites 444, 513, 514, 517, and 565 were conserved among the members of group C. These residues were physicochemically different from corresponding residues of the members of groups A, B, and D. (6-4) photolyase from *D. melanogaster* (64_DROM) uses (6-4) photoproduct as substrate instead

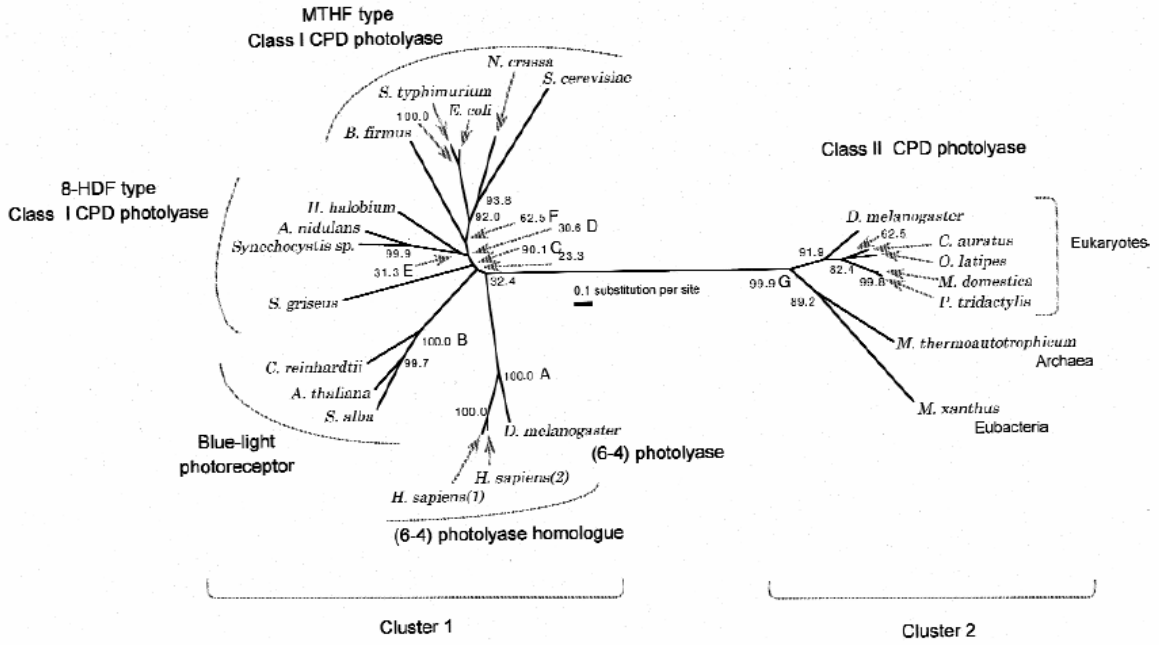


Fig. 2. An unrooted phylogenetic tree obtained by the neighbor-joining method. The *numbers* at the nodes indicate the bootstrap probabilities.

of CPD, and the difference in the conservation pattern may be correlated with the substrate specificity of the enzymes, although (6-4) photolyase homologues from *H. sapiens* (64_HUMA1 and 2) show neither CPD nor (6-4) photolyase activity (Hsu et al. 1996). On the other hand, the members of group E also utilize CPD as their substrate. However, they showed a conservation pattern similar to those of groups A and B at only the three sites 389, 445, and 517, and the remaining 11 sites diverged highly from those of groups A and B. As described above, group E was distantly related to the other members of the protein family, and the C-terminal helical domain of class II photolyase included a long deletion. The deletion was observed close to and within the putative CPD binding sites. The observation suggested that the CPD binding mechanism of class II photolyase is different from those of MTHF-type photolyase and 8-HDF–type photolyase, although they share CPD as their substrate.

Phylogeny of the Photolyase–Blue-Light Photoreceptor Family

Figure 2 shows an unrooted phylogenetic tree by the NJ method, which was divided into two clusters. The clusters are here referred to as clusters 1 and 2. Cluster 1 was composed of (6-4) photolyases, blue-light photoreceptors, and class I photolyases, while cluster 2 consisted of only class II photolyases. The phylogenetic clustering roughly corresponded to a functional classification of the family in the previous section, except for the 8-HDF–type photolyase.

In cluster 1, MTHF-type photolyases, (6-4) photolyases, and blue-light photoreceptors constituted three distinctive subclusters, which corresponded to groups A, B, and D in the previous section. On the other hand, 8-HDF–type photolyases did not form a single subcluster. That is, the functional classification of group B had no evolutionary meaning. As shown in Fig. 2, the roots for the subclusters (6-4) photolyase, blue-light photoreceptor, and MTHF-type photolyase were located at nodes A, B, and F, respectively. The tree topology suggested that present (6-4) photolyases have derived from an ancestral enzyme at node A, which carried (6-4) photolyase activity. Similarly, an ancestral protein corresponding to node B was considered to carry blue-light photoreceptor activity, from which current photoreceptors have evolved. The present MTHF-type photolyases originated from node F, which corresponded to an ancestral CPD photolyase with MTHF. This view is supported by the high bootstrap probabilities for nodes A, B, and F, which were 100.0%, 100.0%, and 62.5%, respectively. On the other hand, the tree topology suggested that there are two independent lineages to 8-HDF–type photolyase. One of them branched off at node C, including only one enzyme from *S. griseus,* whereas another lineage consisted of two enzymes from eubacteria: *Synecocystis* sp.; *A. nidulans;* and an enzyme from an archaebacterium, *Halobacterium halobium.* The branching point for the latter lineage was node D. Thus, the topology of the two lineages was not consistent with the pattern of ordinary species divergence and suggested two independent lineages to 8-HDF–type photolyase. The lineage consisting of the enzyme from *S. griseus* was statistically significant,

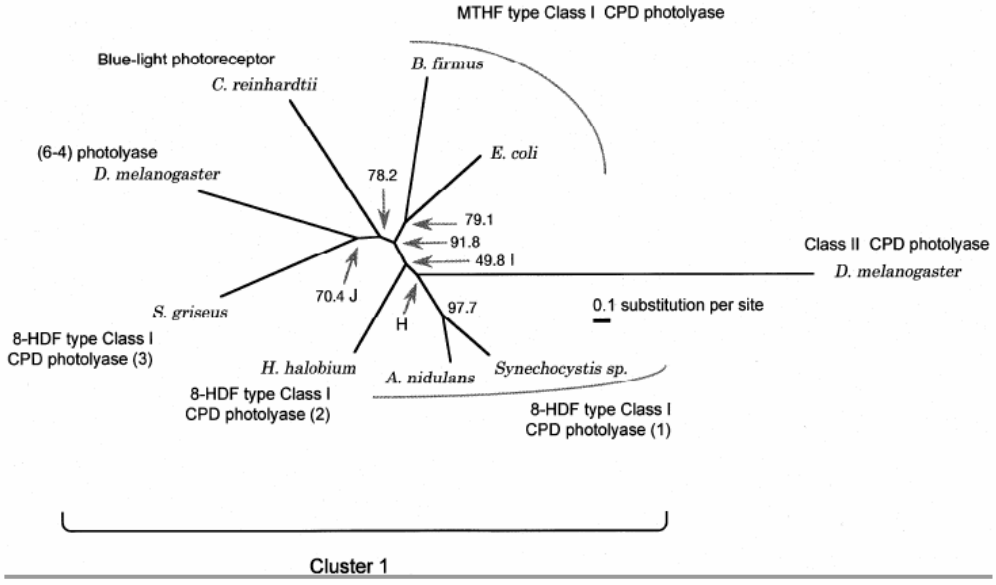


Fig. 3. An unrooted phylogenetic tree obtained by the maximum likelihood method. The *numbers* at the nodes indicate the bootstrap probabilities.

since the bootstrap probability for nodes C is high (90.1%). On the other hand, the bootstrap probability for nodes D is quite low (30.6%). In addition, the bootstrap probability for node E, which corresponds with the divergence point between eubacteria and archaebacterium in the former lineage, is only 31.3%. Thus, the topology about the divergence of 8-HDF–type photolyases from *Synecocystis* sp., *A. nidulans,* and *Halobacterium halobium* was not statistically significant.

In cluster 2, class II photolyase was divided into two subclusters at the node G. One of them includes the enzymes from eukaryotes, whereas another consists of the enzymes from an archaebacterium and a eubacterium. Therefore, node G is considered to correspond to the gene duplication of the enzymes before the divergence between eubacteria and archaebacteria. The tree topology in cluster 2 is considered to be significant because high bootstrap probabilities are evaluated for most of the nodes in the cluster.

As described above, the NJ tree included several nodes with low bootstrap probabilities. However, some of these nodes were related to positions in the tree crucial to a description of the evolutionary history of the family. To check the tree topologies at the nodes, ML analysis was applied to the aligned sequences. However, it requires enormous computational time to examine all of the 22 sequences. Therefore, we selected the following nine sequences, which were related to the crucial nodes or representatives of each subcluster: four 8-HDF–type photolyases from *A. nidulans, Synechocystis* sp., *S. griseus,* and *H. halobium;* MTHF-type photolyase from *E. coli* and *Bacillus firmus;* class II photolyase from *D. melanogaster;* (6-4) photolyase from *D. melanogaster;* and blue-light photoreceptor from *Chlamydomonas reinhardtii.* Only one enzyme from cluster 2 was included in the selection, since the topology of cluster 2 was shown to be statistically significant by NJ analysis. The enzyme was used as an outgroup of the remaining eight sequences from cluster 1.

We examined four amino acid substitutions models for the ML analysis, among which the JTT-F model produced a tree with minimal AIC. Figure 3 shows the unrooted tree. The differences between the minimal AIC and the AICs of the other possible trees were greater than 1.0, which suggested that the tree topology shown in Fig. 3 is statistically significant.

In the ML tree, eight sequences from cluster 1 of the NJ tree formed a cluster against class II photolyase from *D. melanogaster.* Hereafter, the cluster was also referred to as cluster 1. As in the case of the NJ analysis, (6-4) photolyase, blue-light photoreceptor, and MTHF-type photolyase also constituted distinctive subclusters in cluster 1. However, the relative position of each subcluster in the tree was slightly different from that of the NJ tree. Contrary to the result by NJ analysis, most of the nodes in cluster 1 of the ML tree contained high bootstrap probabilities, greater than 70%. We, therefore, considered that the topology of cluster 1 in the ML tree was more reliable than that in the NJ tree. The evolutionary divergence of the three subclusters will be discussed in the next section.

The NJ tree suggested two independent lineages to 8-HDF–type photolyase. In contrast, the ML analysis suggested that there are three independent lineages to 8-HDF–type photolyase. In addition, the relative positions of the enzymes were different from those in the NJ tree. The enzymes from *A. nidulans* and *Synechocystis* sp. branched at node H, while the enzymes from *H. ha-*

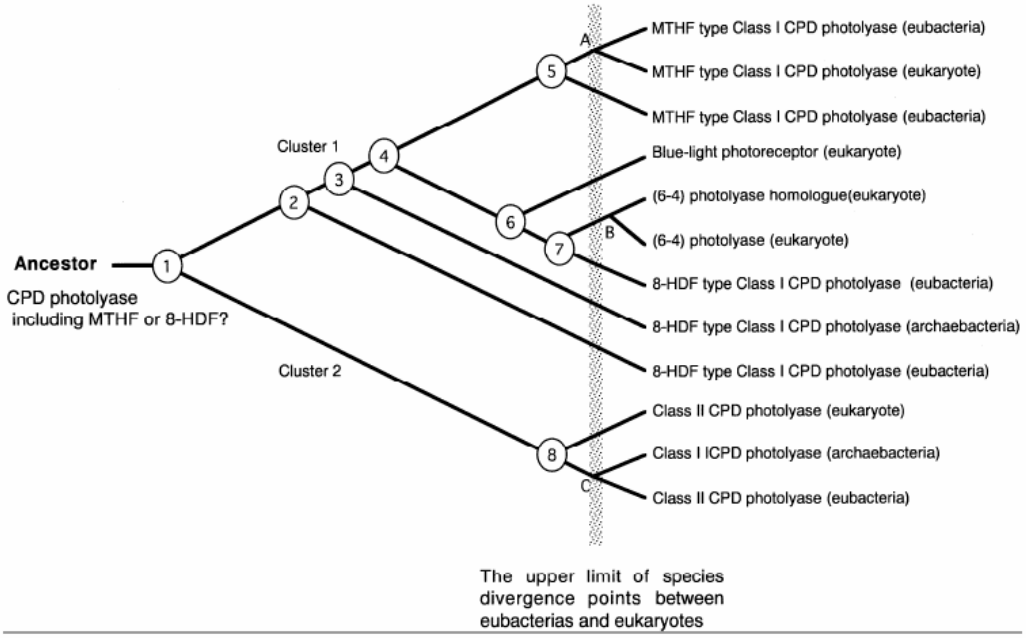


Fig. 4. Scheme of evolutionary process for the photolyase-photoreceptor family. The *circles with figures* indicate the nodes corresponding to gene duplication before the divergence between eubacteria and eukaryotes.

lobium diverged at node I. The locations of the nodes suggested early and independent divergence of the enzymes in cluster 1, while the enzyme from *S. griseus* diverged from the lineage to the subcluster to (6-4) photolyase at node J. The bootstrap probability for node J was high, 70.4%, while the bootstrap probability for node I was only 49.8%. The probability for node H is not shown in the figure, because it was identical to that for the node I. Thus, the ML analysis, as well as the NJ analysis, suggested several independent lineages to 8-HDF–type photolyases, although the branching pattern of the lineages in the ML tree was different from that in the NJ tree. Therefore, the tree topology among the lineages could not be uniquely determined in the current approach.

Evolutionary Scheme of Photolyase–Blue-Light Photoreceptor Family

As described above, the NJ tree is composed of two clusters. The two clusters are connected by the longest branch in the tree. In addition, both clusters include the enzymes from three primary domains of organisms, archaebacteria, eubacteria, and eukaryotes. The observations suggest that the root of the family may be placed on the longest branch. Introducing the putative root, the schematic phylogenetic tree is redrawn (Fig. 4) where the topologies of cluster 1 and 2 were reconstructed based on those in the ML and NJ tree, respectively. The node 1 is the putative root of the tree, where the first gene duplication occurs. Both clusters include CPD photolyases, which are derived from eukaryotes, eubacteria, and archaebacteria. The observation suggests that the ancestral protein at node 1 was a CPD photolyase. However, we could not identify the second cofactor of the ancestral enzyme, because both MTHF and 8-HDF are used as the second cofactors for CPD photolyases belonging to clusters 1 and 2.

We searched for such nodes as correspond to the divergence between proteins from eubacteria and those from eukaryotes or archaebacteria in the schematic phylogenetic tree. Then, nodes A and C were selected as putative species divergence points between eubacteria and eukaryotes (see ''The upper limit of species divergence points between eubacteria and eukaryotes'' in Fig. 4). Here, we assumed that archaebacteria are evolutionarily closer to eukaryotes than eubacteria (Iwabe et al. 1989). Nodes 1 to 8 (Fig. 4) in the upstream to the putative divergence points were, then, considered to correspond to the gene duplications, which had occurred before the species divergence between eubacteria and eukaryotes. As described above, node 1 corresponded to the putative root of the tree, where the ancestral genes for class I photolyases and class II photolyases diverged. The gene duplication of the ancestral enzyme is considered to have occurred at least eight times, and the prototypes of the current genes formed before the divergence between eubacteria and eukaryotes. The genomic redundancy of the photolyase genes in the ancestral organisms may be an adaptation against the high UV radiation on the ancient Earth.

It is interesting and important to investigate which

was the second cofactor of the ancestral CPD photolyase at node 1, MTHF or 8-HDF. In addition, the problem is related to the change in the second cofactor during the evolution of the protein family. As described above, 8-HDF–type photolyase derived from several different lineages. Figure 4 shows a model with three lineages based on the ML analysis. However, the interpretation of the model differs depending on the second cofactor of the ancestral enzyme. MTHF is widely distributed over the current living organisms, while 8-HDF is rare. Therefore, it seems likely that MTHF was used as the second cofactor in the ancestral enzyme. If so, 8-HDF–type photolyase had independently appeared three times from the enzymes in cluster 1 (nodes 2, 3, and 7 in Fig. 4). In this case, MTHF-type photolyase was present on the line connecting nodes 4 to 7 via node 6. Blue-light photoreceptor had functionally diverged from the CPD photolyase at node 6. The gene duplication at node 7 was followed by the functional divergence of each copy. One of them evolved to be (6-4) photolyase, while another remained CPD photolyase, although the second cofactor changed from MTHF to 8-HDF. In cluster 2, 8-HDF–type enzyme appeared at node C. On the other hand, there is a report that 8-HDF is a more efficient photoantenna than MTHF, whose ratio of electron transfer is one and a half times higher than MTHF, including enzyme (Kim et al. 1992). Thus, 8-HDF seemed more suitable to adapt to the high UV environment on the ancient Earth, and the ancestral enzyme may have used 8-HDF as the photoantenna. Then, 8-HDF gradually replaced MTHF as UV radiation on Earth decreased, because MTHF was more abundant than 8-HDF. This idea also seemed as likely as the former one. In this case, the ancestral proteins on the line, connecting nodes 1 to 4 through nodes 2 and 3, were considered to have been 8-HDF–type photolyase. Similarly, 8-HDF–type photolyase were present on the line connecting nodes 4 and 7 via node 6. The change of the second cofactor from 8-HDF to MTHF occurred on three lines in cluster 1. One of them connected nodes 4 and 5, from which MTHF-type photolyase appeared. The second line connected node 6 to blue-light photoreceptor. The third one was a line connecting nodes 7 and B. In cluster 2, the change of the second cofactor from 8-HDF to MTHF occurred at node 8. As described above, it was difficult to identify the second cofactor of the ancestral enzyme in the current approach. Future study of this problem will reveal the adaptation strategy of ancestral organisms on ancient Earth.

References

Adachi J, Hasegawa M (1995) Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. J Mol Evol 40:622–628

Adachi J, Hasegawa M (1996) MOLPHY (programs for molecular phylogenetics) 2.3b3. Institute of Statistical Mathematics, Tokyo

Ahmad M, Cashmore AR (1993) HY4 gene of *A. thaliana* encodes a protein with characteristics of a blue-light photoreceptor. Nature 366:162–166

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Contr 19:716–723

Batschauer A (1993) A plant gene for photolyase: an enzyme catalyzing the repair of UV-light-induced DNA damage. Plant J 4:705–709

Brash DE (1988) UV mutagenic photoproducts in *Escherichia coli* and human cells: a molecular genetics perspective on human skin cancer. Photochem Photobiol 48:59–66

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein sequence structure. National Biomedical Research Foundation, Washington, DC, p 345

Eker APM, Dekker RH, Berends W (1981) Photoreactivating enzyme from *Streptomyces griseus*—IV. On the nature of the chromophoric cofactor in *Streptomyces griseus* photoreactivating enzyme. Photochem Photobiol 33:65–72

Eker APM, Kooiman P, Hessels JK, Yasui A (1990) DNA photoreactivating enzyme from the cyanobacterium *Anacystis nidulans*. J Biol Chem 265:8009–8015

Eker APM, Yajima H, Yasui A (1994) DNA photolyase from the fungus *Neurospora crassa*. Purification, characterization and comparison with other photolyases. Photochem Photobiol 60:125–133

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Felsenstein J (1993) PHYLIP (phylogeny inference package) version 3.5c. Department of Genetics, University of Washington, Seattle

Felsenstein J (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. In: Doolittle RF (ed) Methods in enzymology, vol 266. Academic Press, San Diego, p 418

Hearst JE (1995) The structure of photolyase: using photon energy for DNA repair. Science 268:1858–1859

Higgins DG, Bleasby AJ, Fuchs R (1991) CLUSTAL V: improved software for multiple sequence alignment. Comput Appl Biosci 8:189–191

Hsu DS, Zhao X, Zhao S, Kazantsev A, Wang RP, Todo T, Wei YF, Sancar A (1996) Putative human blue-light photoreceptors hCRY1 and hCRY2 are flavoproteins. Biochemistry 35:13871–13877

Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA 86:9355–9359

Iwasa T, Tokutomi S, Tokunaga F (1988) Photoreactivation of *Halobacterium halobium:* action spectrum and role of pigmentation. Photochem Photobiol 47:267–270

Johnson JL, Hamm-Alvarez S, Payne G, Sancar GB, Rajagopalan KV, Sancar A (1988) Identification of the second chromophore of *Escherichia coli* and yeast DNA photolyases as 5,10-methenyltetrahydrofolate. Proc Natl Acad Sci USA 85:2046–2050

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282

Kaufman LS (1993) Transduction of blue-light signals. Plant Physiol 102:333–337

Kiener A, Husain I, Sancar A, Walsh C (1989) Purification and properties of Methanobacterium thermoautotrophicum DNA photolyase. J Biol Chem 264:13880–13887

Kim S-T, Heelis PF, Okumura T, Hirata Y, Mataga N, Sancar A (1991) Determination of rates and yields of interchromophore (folate → flavin) energy transfer and intermolecular (flavin → DNA) electron transfer in *Escherichia coli* photolyase by time-resolved fluorescence and absorption spectroscopy. Biochemistry 30:11262–11270

Kim S-T, Heelis K, Sancar A (1992) Energy transfer (Deazaflavin →

$FADH_2$) and electron transfer ($FADH_2 \rightarrow T\langle\rangle T$) kinetics in *Anacystis nidulans* photolyase. Biochemistry 31:11244–11248

Kim S-T, Malhotra K, Ryo H, Sancar A, Todo T (1996a) Purification and characterization of *Drosophila melanogaster* photolyase. Mutat Res 363:97–104

Kim S-T, Malhotra K, Taylor JS, Sancar A (1996b) Purification and partial characterization of (6-4) photoproduct DNA photolyase from *Xenopus laevis*. Photochem Photobiol 63:292–295

Kishino H, Miyata H, Hasegawa M (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol Evol 31:151–160

Li YF, Sancar A (1991) Cloning, sequencing, expression and characterization of DNA photolyase from *Salmonella typhimurium*. Nucleic Acids Res 19:4885–4890

Liscum E, Hangarter RP (1994) Mutational analysis of blue-light sensing in *Arabidopsis*. Plant Cell Environ 17:639–648

Maciukenas M, McCaughey M (1994) TREETOOL 2.0.2. Ribosomal RNA Database Project, University of Illinois

Malhotra K, Kim S-T, Walsh C, Sancar A (1992) Roles of FAD and 8-hydroxy-5-deazaflavin chromophores in photoreactivation by Anacystis nidulans DNA photolyase. J Biol Chem 267:15406–15411

Malhotra K, Kim S-T, Sancar A (1994) Characterization of a medium wavelength type DNA photolyase: purification and properties of photolyase from *Bacillus firmus*. Biochemistry 33:8712–8718

Malhotra K, Kim S-T, Batschauer A, Dawut L, Sancar A (1995) Putative blue-light photoreceptors from *Arabidopsis thaliana* and *Sinapsis alba* with a high degree of sequence homology to DNA photolyase contain the two photolyase cofactors but lack DNA repair activity. Biochemistry 34:6892–6899

Mitchell DL, Nairn RS (1989) The biology of the (6-4) photoproduct. Photochem Photobiol 49:805–819

Park H-W, Kim S-T, Sancar A, Deisenhofer J (1995) Crystal structure of DNA photolyase from *Escherichia coli*. Science 268:1868–1872

Payne G, Wills M, Walsh C, Sancar A (1990) Reconstruction of *Escherichia coli* photolyase with flavins and flavin analogues. Biochemistry 29:5706–5711

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sancar A (1996) No ''end of history'' for photolyases. Science 272: 48–49

Sancar GB, Smith FW, Heelis PF (1987) Purification of the yeast PHR1 photolyase from an *Escherichia coli* overproducing strain and characterization of the intrinsic chromophores of the enzyme. J Biol Chem 262:15457–15465

Schwartz RM, Dayhoff MO (1978) Matrices for detecting distant relationships. In: Dayhoff MO (ed) Atlas of protein sequence and structure. National Biomedical Research Foundation, Washington, DC, p353

Short TW, Birggs WR (1994) The transduction of blue light signals in higher plants. Annu Rev Plant Physiol Plant Mol Biol 45:143–171

Taylor JS (1995) DNA, sunlight and skin cancer. Pure Appl Chem 67:183–190

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Todo T, Takemori H, Ryo H, Ihara M, Matsunaga T, Nikaido O, Sato K, Nomura T (1993) A new photoreactivating enzyme that specifically repairs ultraviolet light-induced (6-4) photoproducts. Nature 272:109–112

Todo T, Ryo H, Yamamoto K, Toh H, Inui T, Ayaki H, Nomura T, Ikenaga M (1996) Similarity among the *Drosophila* (6-4) photolyase, a human photolyase homologue, and the DNA photolyase-blue-light photoreceptor family. Science 272:109–112

Todo T, Kim S-T, Hitomi K, Otoshi E, Inui T, Morioka H, Kobayashi H, Ohtsuka E, Toh H, Ikenaga M (1997) Flavin adenine dinucleotide as a chromophore of the *Xenopus* (6-4) photolyase. Nucleic Acids Res 25:764–768

Yasui A, Eker APM, Yasuhira S, Yajima H, Kobayashi T, Takao M, Oikawa A (1994) A new class of DNA photolyases present in various organisms including aplacental mammals. EMBO J 13: 6143–6151