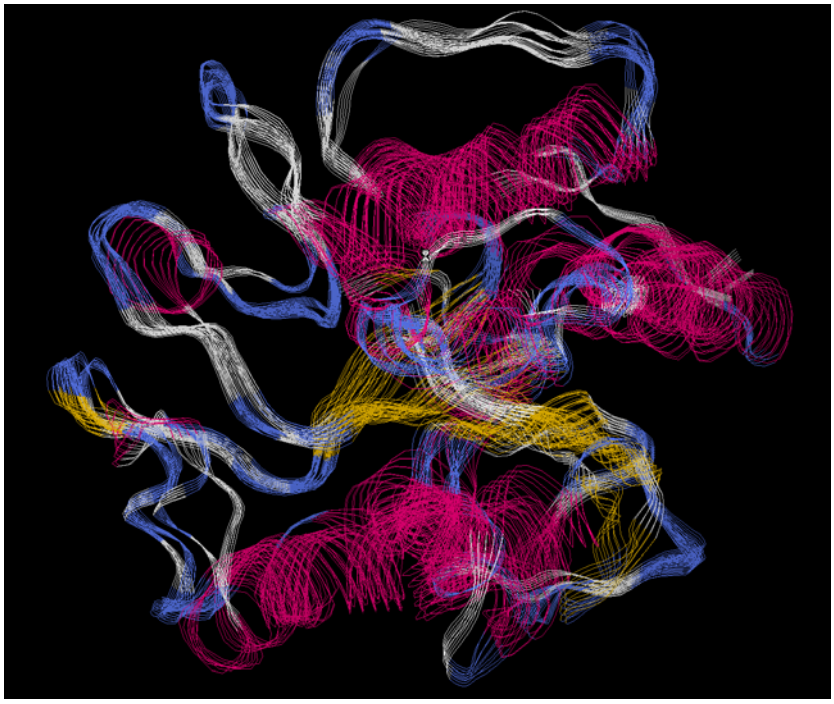
Sequential and Phylogenetic analysis of UDG from organisms of various temperature optima



Universitetet i Tromsø

Autumn 2005

Contents

Summary 3

Introduction 4

Aims 7

Materials and methods 8

Results 11

Discussion 20

Conclusion 22

References 23

Appendix: Full alignments

Summary

With this project we want to show how the proteins in the UNG family in bacteria has adapted to temperature differences. The UNG family of proteins is responsible for correcting DNA errors in the cell, and because of that they are essential for cell function. Different programs where used to align protein sequences and make phylogenetic trees, to look at the similarities between organisms.
Our results showed that the thermophiles had some differences in sequence compared to the psycrophile and mesophile organisms.

Introduction

Genomic DNA can become damaged by a variety of processes, some spontaneous, others catalyzed by environmental agents. Replication itself can occasionally damage the information content in DNA by leaving mispaired bases. Cells need several DNA repair systems to prevent fatal errors and in some cases permanent change in the gene expression.

One of the DNA repair pathways is the Base-Excision repair. Most cells have a class of enzymes called DNA glycosylase that recognize DNA mispairing, such as the products of cytosine and adenine deamination. Deamination is removal of the amino ($-NH_2$) groups from the DNA bases. DNA glycosylase remove the affected base by cleaving the *N*-glycosyl bond. Each DNA glycosylase is generally specific for one type of lesion. Deamination of cytosine to uracil is on of the major pro-mutagenic events in DNA, causing G:C to become A:T if not repaired. If not repaired, a change in gene expression occurs.

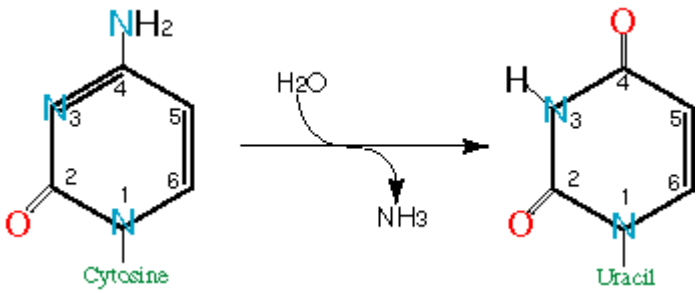


Fig 1. Deamination of cytosine.

Uracil-DNA glycosylase, found in most cells, specifically removes pro-mutagenic uracil from single and double stranded DNA. UDG was the first DNA glycosylase to be discovered. The uracil removing repair mechanism seems to be one of the most essential and common DNA repair system, since UNG gene homologues and their corresponding enzymes can be found a variety of organisms, including bacteria, viruses, archaea and eukaryotes. Different UDG activities have been identified in organisms from almost all kingdoms, and are likely to have arisen very early in cellular evolution.

Genes encoding uracil-excising enzymes can generally be classified into six families:

1) Uracil-DNA glycosylase (UDG)
1) Mismatch specific DNA glycosylase (MUG/TDG)
2) Single strand selective monofunctional UDG (SMUG)
3) Thermostable UDG (TmUDG)
4) UDG-B
5) MIG/EndoIII/MjUDG

Thymine-DNA glycosylase (TDG) removes uracil and thymine from G:U and G:T mismatches in dsDNA. Mismatch specific uracil-DNA glycosylase (MUG) is the prokaryotic homolog to TDG. SMUG removes uracil from U:G and U:A mismatches (greater activity toward U:G mismatches), and has a preference for ssDNA substrates. UDG excises uracil from dsDNA substrates.

Substrate recognition by Family-1 UDG: These enzymes bind uracil base in a “pocket” formed by residues that are very highly conserved throughout this family, and provide a uniquely selective and specific binding site for uracil, that disfavors the binding of another base. Selection between cytosine and uracil is achieved by a set of hydrogen bonding interactions provided by the bottom of the pocket. The O2 carbonyl of the bound pyrimidine receives a hydrogen bond from the NH of the peptide bond joining a conserved Gly-Gln sequence (Gly 143-Gln144 in human; Gly 62-Gln63 in *E.coli*). This interaction helps fix the orientation of the base in the pocket. The amide side-chain of a conserved aspargine (Asn204 I human; Asn 123 in *E.coli*) provides the primary interactions that are believed to distinguish between cytosine and uracil.

UNG homologues have been characterized in many bacterial and eukaryotic species and these proteins have strikingly similar structures and functions. These proteins are enzymes, and enzymes represent a special case of protein function. Enzymes bind and chemically transform other molecules, that is they catalyze reactions. The molecules acted upon by enzymes are substrates, and the domain on the enzyme that binds with the substrate is called the catalytic domain. The catalytic domains are often conserved areas in the peptide chain, and proteins encoding the same function will often share similar sequences in these domains. Two catalytic domains are found in the UNG homologues.

Organisms adapted to cold environments are often termed psychrophiles, whereas their warm-active counterparts are referred to as mesophiles and thermophiles. Temperature is one of the most important factors for enzyme activity, and enzymatic reaction rates can be reduced 30-80 times when the temperature decreases from 37 to 0 °C. Psycrophile organisms grow in temperatures under 20°C, mesophiles grow in temperatures from 20 to 40°C, thermophiles grow in environments with temperatures above 40°C and finally hyperthermophiles grow in temperatures above 60°C.

To deal with this strong temperature dependence, cold-adapted enzymes usually have higher catalytic activity at low temperatures and decreased thermostability as compared with their mesophilic and thermophilic counterparts. Because cytosine hydrolysis is greatly enhanced at high temperatures, hyperthermophilic organisms might be expected to be more susceptible to this type of mutation. This does not appear to be the case, as the reported spontaneous mutation rates in thermophile bacteria are similar to those observed in *Escherichia coli*, a mesophile. Hyperthermophilic organisms are apparently more efficient at repairing damage to the DNA bases.

With this knowledge, we want to discover if the UNG homologue enzymes from different bacteria with varying temperature optima, vary in sequential or phylogentic analysis. We also want to know about the substrate specificity, if there is any point of distinction amongst the different UDG families and is there a change in the catalytic domain, and if so, which amino acid(s).

Aims

We want to gather protein sequences coding for the UNG protein family in bacteria. We then want to use these sequences to look at differences and similarities in substrate specificity and catalytic site residues in respect to organism temperature optimum. The methods we will be using is protein alignments and phylogeny.

Material and methods

1. Gathering sequences

To gather our sequences we first used SRS (srs.ebi.ac.uk/) to find different sequences of the different families. However, the number of sequences we were able to gather with SRS was not sufficient for our alignments. That's why we then took the sequences we found with SRS and used BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) to find similar sequences. We found a lot of sequences by doing this, so we could just pick the ones most suitable for our examination. We picked our sequences based on what temperature the organism optimally lived under. The way we found out which organism had which temperature optimum proved to be most effective with plain google (www.google.com) searches.

2. Alignment

We now wanted to examine the gathered sequences using multiple alignment methods. We wanted to use a global alignment tool, ClustalW (http://www.ebi.ac.uk/clustalw/), and one local alignment tool, T-COFFEE (http://www.ch.embnet.org/software/TCoffee.html). The difference between these two programs is that ClustalW looks for the best overall score, while T-COFFEE looks for small conserved regions, and tries to align these in the best possible way. This is why T-COFFEE is reckoned to be more efficient for sequences with less than 30% similarity than ClustalW (http://www.ch.embnet.org/software/TCoffee.html).

3. Phylogeny

Finally we wanted investigate the relationship between the different organisms proteins by using tree building methods.
The program we used to do this was MEGA2 (http://www.megasoftware.net/).
With Mega2 we were able to construct trees with the following methods (cut from Mega2 manual):

UPGMA -This method assumes that the rate of nucleotide or amino acid substitution is the same for all evolutionary lineages. An interesting aspect of this method is that it produces a tree that mimics a species tree, the branch lengths for two OTUs being the same after their separation. Because of the assumption of a constant rate of evolution, this method produces a rooted tree, though it is possible to remove the root for certain purposes.

Neighbour joining (NJ) - This method is a simplified version of the minimum evolution (ME) method. In the ME method, distance measures that correct for multiple hits at the same sites are used, and a topology showing the smallest value of the sum of all branches (S) is chosen as an estimate of the correct tree. However, construction of a minimum evolution tree is time-consuming, because in principle the S values for all topologies have to be evaluated. The number of possible topologies (unrooted trees) rapidly increases with the number of taxa. Therefore, it becomes very difficult to examine all topologies. In the case of the NJ method, the S value is not computed for all or many topologies, but the examination of different topologies is embedded in the algorithm, so that only one final tree is produced. NJ method produces an unrooted tree, because it does not require the assumption of a constant rate of evolution. Therefore, it requires an outgroup taxon to find the root. In the absence of outgroup taxa, the root is sometimes given at the midpoint of the longest route connecting two taxa in the tree. This is referred to as the mid-point rooting.

Minimum evolution (ME) - In the ME method, distance measures that correct for multiple hits at the same sites are used, and a topology showing the smallest value of the sum of all branches (S) is chosen as an estimate of the correct tree. However, construction of a ME tree is time-consuming, because in principle the S values for all topologies have to be evaluated. The number of possible topologies (unrooted trees) rapidly increases with the number of taxa. While the NJ tree is usually the same as the ME tree when the number of taxa is small; the difference between the NJ and ME trees can be substantial. In this case, the ME tree is preferable if a long DNA or amino acid sequence is used. When the number of nucleotides or amino acids used is relatively small, the NJ method generates the correct toplogy more often than the ME method (Nei et al. 1998, Takahashi and Nei 2000).

Results

Table 1. Found and used sequences.

Protein name	Organism	Substrate specificity	Sub familiy	Temperature optimum (°C)
UDG	*Vibrio fisheri*	DNA	UDG	30
UDG	*Streptococcus agalactiae*	DNA	UDG	37
UDG	*Clostridium perfringens*	DNA	UDG	37
UDG	*Pseudunomonas fluorescens*	DNA	UDG	22
UDG	*Pseudononas aeruginosa*	DNA	UDG	37
UDG	*Pseudoalteromonas atlantica*	DNA	UDG	20
UDG	*Photobacterium profundom*	DNA	UDG	<20
UDG	*Vibrio salmonidica*	DNA	UDG	<20
UDG	*Shewanella denitrificans*	DNA	UDG	4-30
UDG	*Colwellia psychrerythraea*	DNA	UDG	<5
UDG	*Heamophilus influenzae*	DNA	UDG	39
UDG	*Escherichia coli*	DNA	UDG	37
UDG	*Staphylococcus saprophyticus*	DNA	UDG	37
UDG	*Vibrio vulnificus*	DNA	UDG	37
UDG	*Salmonella enterica*	DNA	UDG	~35
UDG	*Thermosynechococcus elongates*	DNA	UDG	60
UDG	*Thermus thermophilus*	DNA	UDG	67
UDG	*Thermoanaerobacter tengcongensi*	DNA	UDG	
UDG	*Thermococcus kodakarensis*	DNA	UDG	85
UDG	*Thermoplasma volcanium*	DNA	MUG/TDG	60
MUG	*Serratia marcescens*	dsDNA	MUG/TDG	40
MUG	*E.coli*	dsDNA	MUG/TDG	37
TDG	*Homo sapiens*	dsDNA		37
TDG	*Schizosaccharomyces pombes*	dsDNA	TDG	~30
SMUG	*Rhodopirellu labaltica*	(ss)DNA	SMUG	n/a
SMUG	*Homo sapiens*	(ss)DNA	SMUG	37

Alignment ClustalW

Catalytic domain A.

```
Pseudomonasaeruginosa                KVVIIGQDP-YHGPGQAHGLCF---------SVQPGVPTPPSLQNIY-KE 102
Pseudomonasfluorescens               KVVILGQDP-YHGPGQAHGLCF---------SVQPGVPAPPSLVNIY-KE 101
Shewanelladenitrificans              KVVILGQDP-YHGLGQAHGLCF---------SVPEGIKIPPSLVNIY-KE 92
Colwelliapsychrerythraea             KVVILGQDP-YHGKDQAHGLAF---------SVQEGIKVPPSLVNIY-KE 94
Photobacteriumprofundum              KVVILGQDP-YHGANQAHGLSF---------SVLPGIKTPPSLANMY-KE 94
Vibriosalmonidica                    TVVILGQDP-YHGEGQAHGLSF---------SVLPGVKIPPSLRNMY-KE 93
Vibriovulnificus                     KVVILGQDP-YHGPNQAHGLCF---------SVLPGIKTPPSLVNMY-KE 103
Vibriofisheri                        KVVILGQDP-YHGPNQAHGLCF---------SVLPGIKTPPSLVNMY-KE 95
Pseudoalteromonasatlantica           KVVILGQDP-YHGPNQAHGLSF---------SVQKGVKTPPSLKNMY-KA 95
E.Coli                               KVVILGQDP-YHGPGQAHGLAF---------SVRPGIATPPSLLNMY-KE 95
Salmonellaenterica                   KVVILGQDP-YHGPGQAHGLAF---------SVRPGIAPPPSLVNMY-KE 95
Haemophilusinfluenzae                KVVILGQDP-YHGPNQAHGLAF---------SVKPEVAIPPSLLNIY-KE 92
Streptococcusagalactiae              KVVIIGQDP-YHGPQQAQGLSF---------SVPDNLPAPPSLQNIL-KE 93
Clostridiumperfringens               KVVILGQDP-YHGPNQAHGFSF---------SVKPGVQTPPSLRNMF-KE 96
Staphylococcussaprophyticus          KVVILGQDP-YHGPNQAHGLAF---------SVQPDAKFPPSLRNMY-KE 90
EColiMUG                             RVVFCGINPGLSSAGTGFPFAH---------PANRFWKVIYQAGFTD-RQ 50
SerratiamarcescensMUG                RVVFCGINPGLSSAHQGYPFAN---------GSNRFWKVIHQAGFTE-SQ 48
HumanTDG                             DIVIIGINPGLMAAYKGHHYPG---------PGNHFWKCLFMSGLSE-VQ 172
Thermusthermophilus                  KLMIVGEGP-GEE-EDKTGRPF---------VGKAGQLLNRILEAAG-IP 72
Thermosynechococcuselongates         KLMIIGEGP-GQA-EDESGRPF---------VGKAGQLLDKILASVN-LD 117
Thermoanaerobactertengcongensi       KVMFVGEGP-GRD-EDLQGRPF---------VGRAGQLLNKMLEDVG-LK 75
Thermococcuskodakarensis             KVMFVGEAP-GYW-EDQKGLPF---------VGRAGKLLDELLAEIG-LS 78
RhodopirellulabalticaSMUG            HVLFLGMNP-GPWGMAQSGVPFGEIDSVVQWMGLKGEVERPSNEHSK-RP 111
HumanSMUG                            EVLFLGMNP-GPFGMAQTGVPFGEVSMVRDWLGIVGPVLTPPQEHPK-RP 125
SchizosaccharomycespombeTDG          YAIIVGLNPGITSSLKGHAFAS---------PSNRFWKMLNKSKLLEGNA 191
Thermoplasmavolcanium                RLLIVGLAP-AATGGNRTGRVFT--------GDKSSDFLVSCLFEAGITN 97
```

Fig 2. Catalytic domain A, alignment made with ClustalW. The yellow background is showing the catalytic domain.

Catalytic domain B.

```
Pseudomonasaeruginosa                LKSAHPSPLSAYRGFLGNG----HFSRTNKFLEQNGKTPIDWSLPDL--- 231
Pseudomonasfluorescens               LTSVHPSPLSAYRGFLGCG----HFSRTNKFLEQNGEAPIEWRLPPVV-- 231
Shewanelladenitrificans              LQGPHPSPLSAYRGFFGCK----HFSTINGILLKQGDKAIDWQV------ 219
Colwelliapsychrerythraea             LAGPHPSPLSAYRGFFGCK----HFSQANAWLAKHEVNQVDWHLPLEA-- 225
Photobacteriumprofundum              LKSAHPSPLSAYRGFFGCK----HFSQVNQLLCDMNQSPIDWQPQRIQE- 226
Vibriosalmonidica                    LVAPHPSPLSARRGFFGCQ----HFSKTNELLREKNLSEINWSSITLD-- 224
Vibriovulnificus                     LTAPHPSPLSAHRGFLGCK----HFSQANQHLLDQGKQAIDWQLPLSL-- 234
Vibriofisheri                        LAAPHPSPLSAHRGFLGCK----HFSQANQLLASQGKEPINWHLPLTV-- 226
Pseudoalteromonasatlantica           LTAAHPSPLSAYRGFFECG----HFAQTNQILLDSGRTPIKWQI------ 222
E.Coli                               LKAPHPSPLSAHRGFFGCN----HFVLANQWLEQRGETPIDWMPVLPAES 228
Salmonellaenterica                   LKAPHPSPLSAHRGFFGCN----HFALTNQWLEQHGEKTIDWTPVLPAES 228
Haemophilusinfluenzae                LTAPHPSPLSAHRGFFGCR----HFSKTNSYLESHGIKPIDWQI------ 218
Streptococcusagalactiae              IESPHPSPLSAYRGFFGSR----PFSRTNHFLEEKGINEIDWLN------ 217
Clostridiumperfringens               IKSVHPSPLSATRGFFGSK----PFSKTNDFLVSINKEPIDWQIPNI--- 225
Staphylococcussaprophyticus          IQSPHPSPLSAYRGFFGSK----PYSQANTYLQANGKQPVNWCESEV--- 218
EColiMUG                             WVLPNPSGLSR--------------VSLEKLVEAYRELDQALVVRGR--- 168
SerratiamarcescensMUG                WVLPNPSGLNR--------------ATLEQLTASYRELFLALQ------- 162
HumanTDG                             LCYVMPSSSARCAQFPRAQDKVHYYIKLKDLRDQLKGIERNMDVQEVQYT 314
Thermusthermophilus                  GIKVFPMFHPAYLLRNPSRAP---GSPKHLTWLDIQEVKRALDALPPKER 193
Thermosynechococcuselongates         GRWCMPIFHPAYLLRNNSREP---GSPKWLTWQDIQAVRDRLRQLDS--- 235
Thermoanaerobactertengcongensi       GVKIIATYHPAAILRDPEK--------LQPAMEDFKKIKEELDKLV---- 186
Thermococcuskodakarensis             KIIIMPMYHPAAALYKPPL--------REELRKDFQKLKSLLENI----- 196
RhodopirellulabalticaSMUG            VKSMEQTSVSIHRILHPSPA----SPAANKDWAGKATSQLVEAQIW---- 252
HumanSMUG                            LAGLMP-EVQVEGLLHPSPR----NPQANKGWEAVAKERLNELGLLPLLL 269
SchizosaccharomycespombeTDG          FVGISSSGRAAGYSDEKKQ-------NLWNLFAEEVNRHREIVKHAV--- 325
Thermoplasmavolcanium                TFKLVPSYHPSPRNVNTGK------LKREDFVSLLQKVKALISE------ 227
```

Fig 3. Catalytic domain B, aligment created with ClustalW. The yellow background is marking the catalytic domain.

Alignment T-Coffee

Catalytic domain A

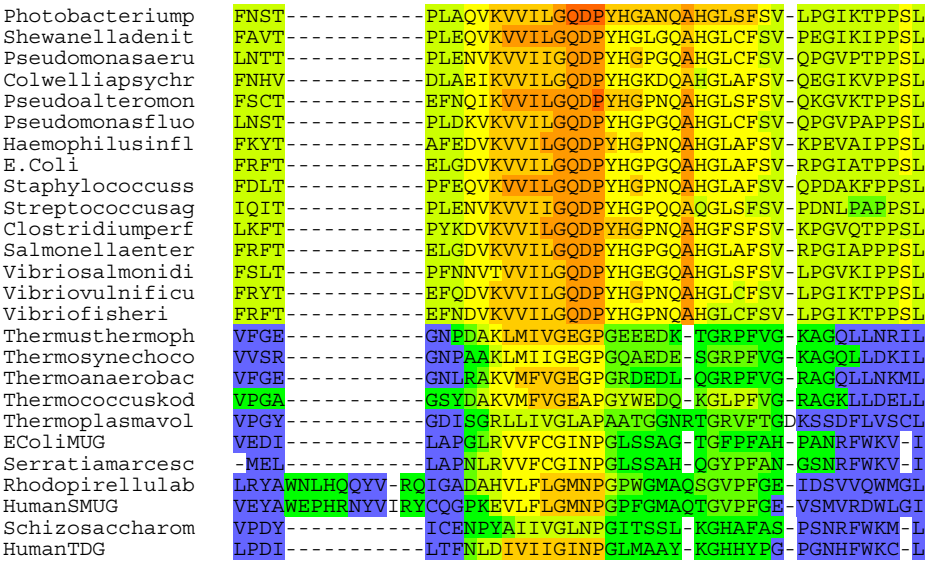


Fig 4. Catalytic domain A created with T-Coffee.

Catalytic domain B

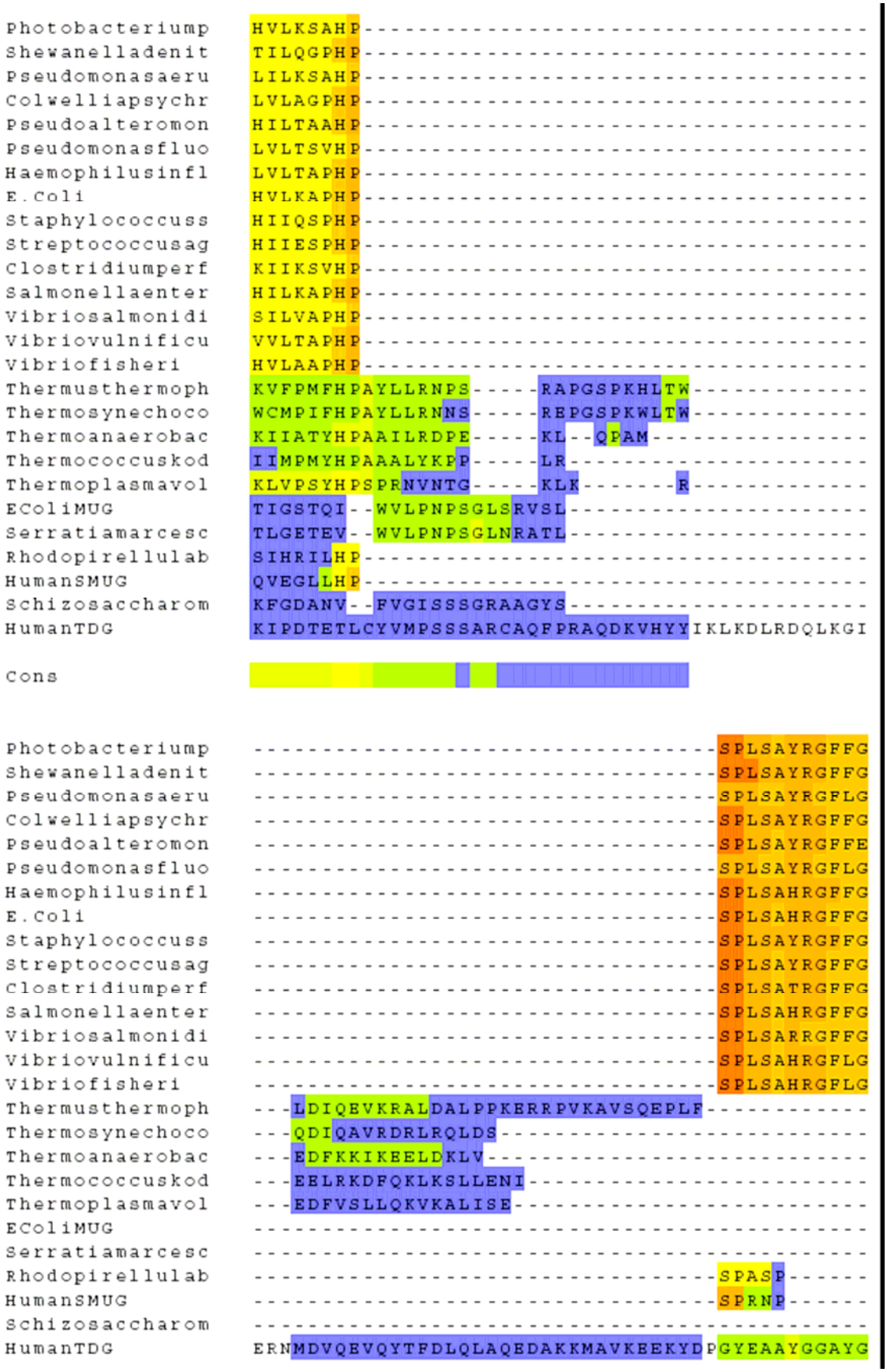


Fig 5. Catalytic domain B created with T-Coffee.

Phylogenetic tree

UPGMA

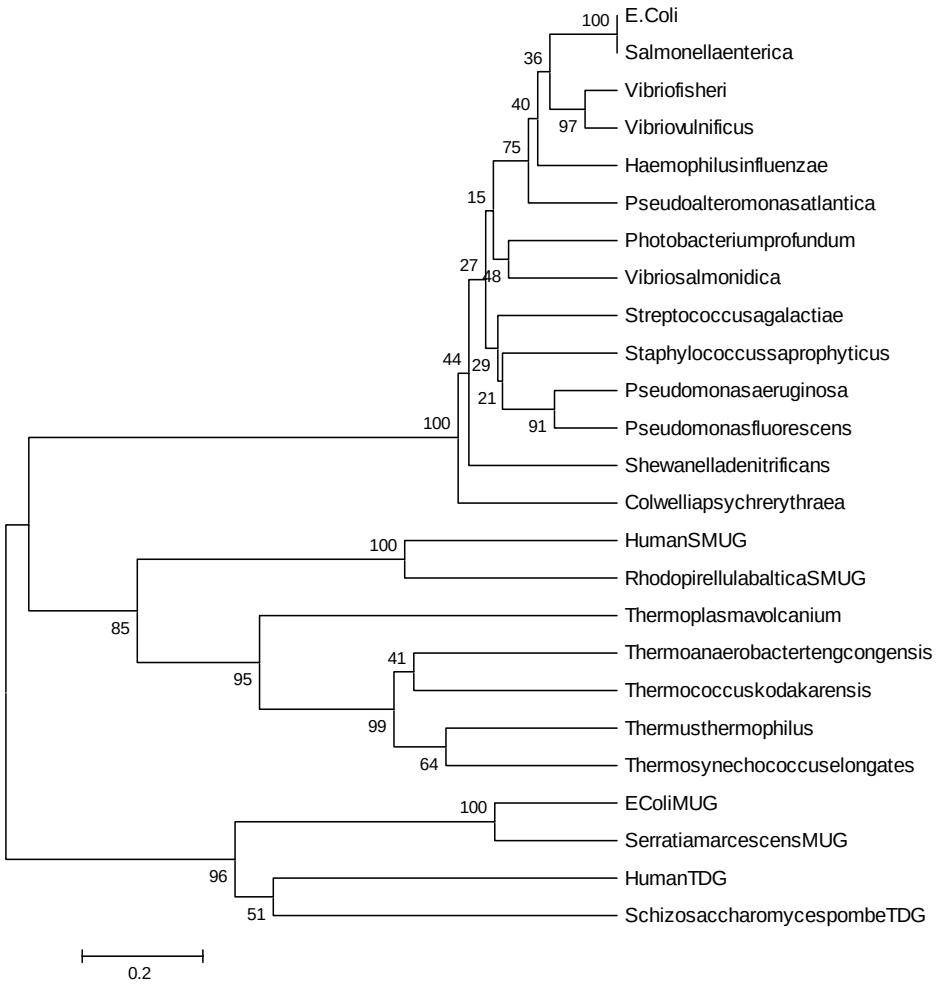


Fig 6. Phylogenetic tree made with the UPGMA method with complete deletion.

Neighbour joining

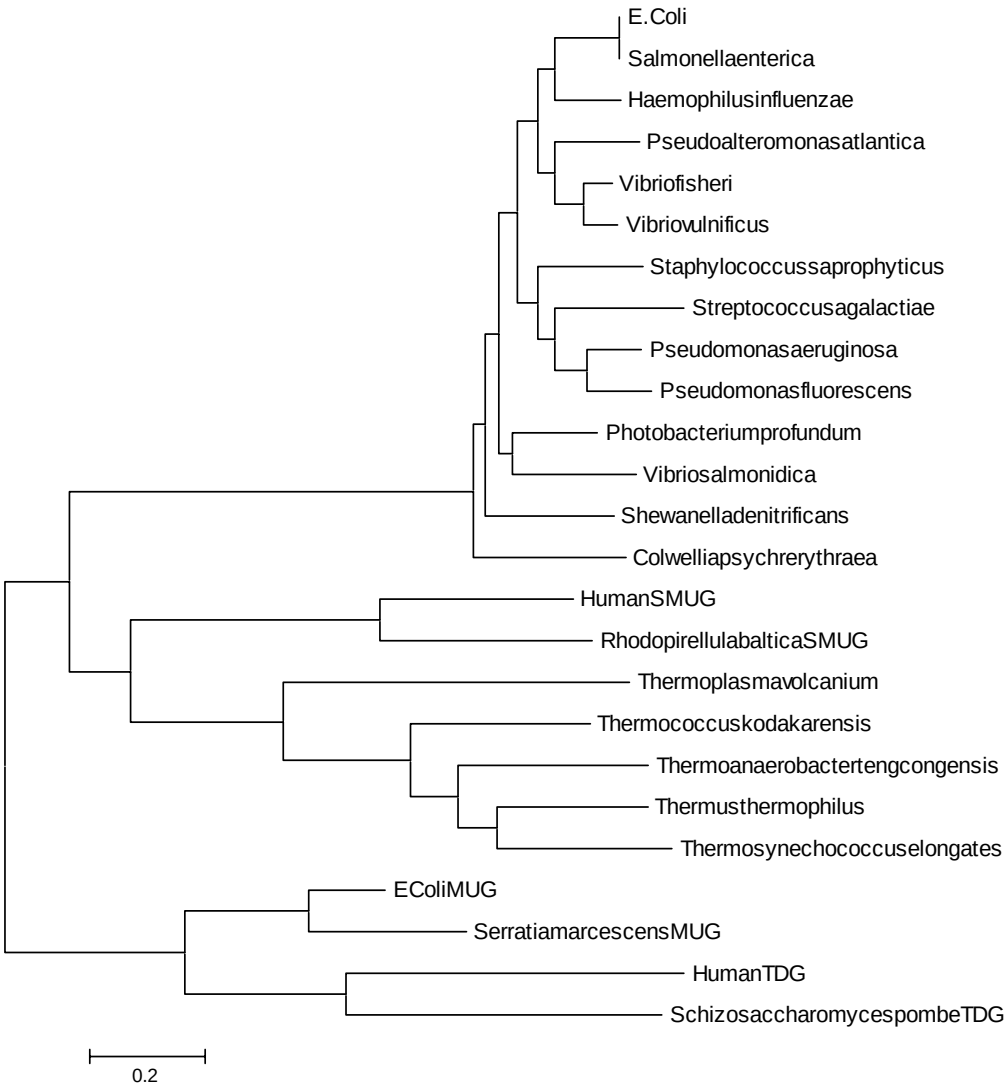


Fig 7. Phylogenetic tree made with neighbour joining method, complete deletion.

Minimum evolution

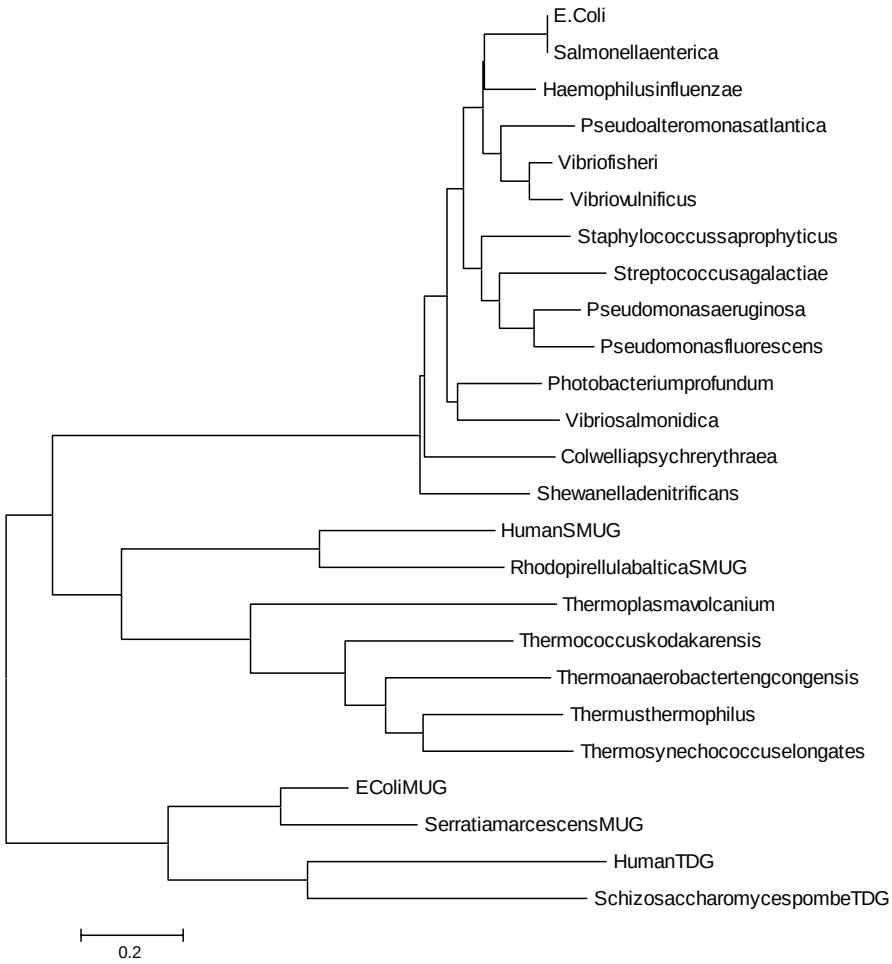


Fig 8. Phylogenetic tree made with minimum evolution, complete deletion.

Consensus sequences
ClustalW

```
Mesophile         ------------------------------------WHDVLEYFNFEYSGI-VYPPKDVFN 24
Psycrophile       ------------------------------------MWQ--IEYFQVT-RGI---FPEDVF- 19
Thermophile       ----------------------------------------LLCKCLRTVGG---------- 11
                                                          :     .   *

Mesophile         AFFTPLDVKVVILGQDPYHGPQAHGLAFSVPGPPSLNMYKELDIG-PGLES---WARQGV 80
Psycrophile       -AFTPL-VKVVILGQDPYHG-QAHGLFSVGIKPPSLN-YKELATDIEGFPHGLSWA-QGV 74
Thermophile       -------AKLMIVG-EPGE----DGRPFVG-KAGLLLGLRE-----------------EV 41
                         .*::*:* :* .    .*       .  *   :*                 *

Mesophile         LLLNTVLTVRGQAHSHALGWETFTDVILNQHEVVFILWGAQKKALIDHHIIKSP---HPS 137
Psycrophile       LLLNTVLTV--EAHSHAKGWETFTD-ILNG----IFLWGAHAK---KGKIKHLP---HPS 121
Thermophile       YITNIVKCRPPNRPTEIKACP-YLQ-IIPK-----IILGAAL-------IRGWE----GK 83
                   : * *      :  :.  .   : : *:       :: **         *         .

Mesophile         PLSA-RGFFGSPFSTNLEGPIDW------------ 159
Psycrophile       PLSAYRGFFG---CKHFSNLIDW------------ 141
Thermophile       PYHP--ALRP----LDFQKLL-------------- 98
                  *  .  .:        .:.  :
```

Fig 9. Alignment with consensus sequences. Alignment is made in ClustalW, and the consensus sequences were made based on the sequences used previously in bioedit.

T-COFFEE

```
Thermophile    :  28
Mesofil        :  45
Psycrophile    :  46

Thermophile    --LLCKCLRTVG-----------------GAKLMIVGEPGEDGRPFVGKAGLLLGL
Mesofil        WHDVLEYFNFEYSGIVYPPKDVFNAFFTPLDVKVVILGQDPYHGP----QAHGLAF-
Psycrophile    -MWQIEYFQVT-RGIF--PEDVF--AFTPL-VKVVILGQDPYHG-----QAHGLFS-

Thermophile    REEVYITNIVKCRPPNR------PTEIKACPY-LQII-PKIILGAALI---------
Mesofil        ---------SVPGPPSLNMYKELDIG----PGLESWARQGVLLLNTVLTVRGQAHSH
Psycrophile    ---------VGIKPPSLN-YKELATDIEGFPHGLSWA-QGVLLLNTVLTV--EAHSH

Thermophile    -RGWE-----------------------GKPYH----PALRPLD----------FQK
Mesofil        ALGWETFTDVILNQHEVVFILWGAQKKALIDHHIIKSPHPSPLSA-RGFFGSPFSTN
Psycrophile    AKGWETFTDILNG-----IFLWGAHAKKGKIKHL---PHPSPLSAYRGFFGCKHFSN

Thermophile    L---L--
Mesofil        LEGPIDW
Psycrophile    L---IDW
```

Fig 10. Consensus sequence alignment made in T-COFFEE, The same sequences as used in fig 8.

Overview

Table 2. Conserved domains overview

	Domain A: conserved sequence	**Domain B: conserved sequence**
Thermophile	G--PG	HPA
mesophile	G--PY	HPS
psychrophile	G--PY	HPS
MUG	G--PG	NPS
TDG	G--PG	MPS/GIS
SMUG	G--PG	HPS

Discussion

We observe that proteins with similar substrate specificities appears to come out on the same branch, because entities belonging to a subfamily (MUG, TDG or SMUG) comes out on the same branch, independent of temperature specificity (fig 6, 7 and 8).

The phylogenetic trees made from three different treebuilding-methods are similar, only the branch-lengths seem to be different. The general pattern observed, is that thermophiles comes out on the same branch, while the psychrophiles are mixed with the mesophiles. This result is different from what we had expected: thermophiles, psychrophiles and mesophiles appearing on three different branches. The result indicates that the mesophiles and psychrophiles are more closely related to each other, than to thermophiles.

We expect the catalytic domains to be highly conserved in the different organisms. In domain A, especially three residues are mostly the same in all sequences: a glycine (G), a proline and another glycine (or tyrosine). In the global alignment made by Clustal W, the first G and the P are aligned in all organisms. Next, all mesophiles and psychrophiles have a conserved tyrosine (Y) where the thermophiles have a conserved G. One exception is found: *Termoplasma volcanium* has an Alanine (A) residue instead of G or Y. The amino acids Tyrosine and Glycine have very different properties. G has a very small sidechain (-H), and is non-polar and hydrophobic. Y has a large, aromatic sidechain, and is polar and hydrophobic. This indicates that the observed substitution in domain A between the thermophiles and the other organisms may have a great effect on the enzyme activity (fig 2).
The thermophiles have changes in the catalytic site (in both domain A and domain B), not found in the other organisms. Maybe this change is responsible for the high efficiency of repairing deaminations, found in termophiles. Cytosine hydrolysis is greatly enhanced at high temperatures, so the repair systems in thermophiles need to be more efficient than in their mesophilic and psychrophilic counterparts.

In the local alignment made by T-COFFEE, the result of aligning the catalytic domain A was similar to the result from the global alignment (Clustal W): the first G and the P aligned in all organisms, Y aligned in mesophiles and psychrophiles, G in termophiles, and A in *T.volcanum* (fig 4).

In domain B, three residues are conserved in the different subfamilies: H, P and S. The local and global alignment showed the same results in conserved residues of domain B. The mesophiles and SMUGs have HPS. The termophiles have an Alanine (A) instead of Serine (S). In the other subfamilies: MUG and TDG, the conserved residues are NPS, and GIS or MPS, respectively. The human TDG has a sequence that is totally different from the fungus TDG (fig 3).
The fact that the other subfamilies have different catalytic sites than UDG, is as expected, since they have different substrate specificities and functions.

Clustal W aligned the conserved residues in catalytic domain B much better than the alignment by T-COFFEE. We started by making a T-COFFEE alignment with only the UDG subfamily, and then the catalytic B domain was aligned correctly. When we later added sequences from more subfamilies, a big gap was inserted in domain B, destroying the correct alignment. Some of the new sequences added, TDG and MUG, were much longer than the UDG sequences, this could have caused the problems in the T-COFFEE alignment.
In general, we observed that the local alignment by T-COFFEE, arranged the domains of the catalytic site in a more correct position than the global alignment by Clustal W did (except some mistakes in domain B). T-COFFEE aligned all the conserved residues, and put gaps in the correct positions, giving us the most reliable result, compared to previous findings. Therefore, we used the alignment from T-COFFEE to build trees from. Without some biological knowledge of the catalytic sites of the enzyme, it would have been difficult to compare the results of the different methods. In this study, local alignment (T-COFFEE) was found to be the best method for comparing short sequences and looking for conserved regions. A global alignment (Clustal W) is looking for the best overall score for the whole sequence, and could leave out the aligning of important short sequences, such as catalytic domains.

Conclusion

The UDG protein from different bacteria has a conserved sequence, especially in the catalytic domains. The alignments show that organisms within a group having the same temperature optimum (termophiles, mesophiles or psychrophiles), have homologous sequences in the catalytic domains. Comparing the different temperature groups, we observe some important differences in the catalytic domains. The termophiles have amino acid changes in the catalytic domain not found in the mesophiles or psychrophiles, which probably have an effect on the UDG enzyme activity. Proteins belonging to the same subfamily (MUG, TDG, SMUG) come out on the same branch of the phylogenetic trees. The thermophiles also come out on the same branch, while the psychrophiles are mixed with the mesophiles, indicating that mesophiles and psychrophiles are more closely related to each other, than to thermophiles.

References

L.Snyder, W.Champness, Molecular Genetics of Bacteria, 2nd edition (2003), 371-372

L.H.Pearl, Structure and function in the uracil-DNA glycosulase superfamily (2000)

E.Moe, I.Leiros, A.O.Smalås, S.McSweeney, The Crystal Structure of MUG from Deinococcus Radiodurans reveals a novel catalytic residue and broad substrate specificity. (2002)

J.H.Chung, E.K.Im, H-Y.Park, J.H.Kwon, S.Lee, J.Oh, K-I.Hwang, J.H.Lee, Y.Jang, A novel uracin-DNA glycosylase family related to the helix-hairpin-heix DNA glycosylase superfamily.

L.Aravind, E.V.Koonin, The α/β fold uracil DNA glycosylases: a common origin with diverse fates. (2000)

David L. Nelson, Michael M. Cox (2000). *Lehninger Principles of biochemistry*, third edition.

Magne Olufsen, Arne O. Smalås, Elin Moe, Bjørn O. Brandsdal (2005). *Increased flexibility as a strategy for cold adaption.*

Ingeborg Knævelsrud, Peter Ruoff, Hilde Ånensen, Arne Klungland, Svein Bjelland, Nils-Kåre Birkeland (2001). *Excision of uracil from DNA by the hyperthermophilic AfUNG protein is dependent on the opposite base and stimulated by heat-induced transition to a more open structure.*

Ingar LeiroS, Elin Moe, Olav Lanes, Arne O. Smalås, Nils P. Willassen (2003). *The structure of uracil-DNA glycosylase from Atlantic cod (Gadhus morhua) reveals cold-adaptation features.*

Alessandro A. Sartori, Primo Schär, Sorel Fitz-Gibbon, Jeffrey H. Miller, and Josef Jiricny (2001). *Biochemical characterization of uracil processing activities in the hyperthermophilic archaeon Pyrbaculum aerophilum.*