

*Homing endonuclease like
proteins –
a computational and
biochemical analysis*

Master thesis

**Presented to the Scientific Council of the Weizmann
Institute of Science**

By Einat Sitbon

Supervisor Dr. Shmuel Pietrokovski

Department of Molecular Genetics

Weizmann Institute of Science

Submitted – 09/01/01

Acknowledgment

I would like to thank James Van-Etten's lab for sending clones and northern blot, Harry Burgess and Ksenia Sorokina from Orly Reiner's lab for helping me obtain a mutant protein, Roe Atlas for teaching me the western blot, Gil Amitai for lots of help on the experimental side, and all of Shmuel Pietrokovski's group for listening and discussing any problem. I would also like to thank Shmuel Pietrokovski for his patient guidance throughout my study.

Table of content

Abstract	3
Introduction	3
The HNH family.....	5
The GIY family.....	6
Paramecium bursaria Chlorella virus 1	7
Methods.....	8
Computational methods.....	8
Experimental methods.....	9
Results	11
Computational results.....	11
Experimental results.....	18
Discussion.....	21
References	22
Appendix 1	26
Appendix 2	27

Abstract

Many proteins are built from conserved domains – conserved sequence stretches, which are common to a protein family. These domains are often modular. They can occur in different contexts, and so append a function to a protein. In a way protein domains are the building blocks of evolution. Known examples are the ATP binding domain and the helix-turn-helix DNA binding domain. HNH and GIY-YIG (GIY) are two homing endonuclease families. Their catalytic domains were previously shown to be separate modular domains. Homing endonucleases (H-EN) are rare cutting enzymes, which occur in open reading frames inside several mobile elements. Their DNA binding and catalytic domains could be interesting and useful. In this work I studied the modularity of the HNH and GIY families. Using computational analysis I defined four new modular domains outside the catalytic domains. Two of those domains occur both in HNH and GIY proteins. The domains appear together in complicated patterns, in a way that suggests modular use. I experimentally studied uncharacterized proteins that have the homing endonuclease catalytic domains to start verifying my computational analysis. My preliminary results show that one of these proteins binds and probably degrades DNA.

Introduction

Homing endonucleases (H-EN) are rare-cutting enzymes whose recognition sequence is usually long – 12-40 BP of DNA, and can tolerate base changes. H-EN are present in many open reading frames (ORFs) in group I and II introns, and in inteins.

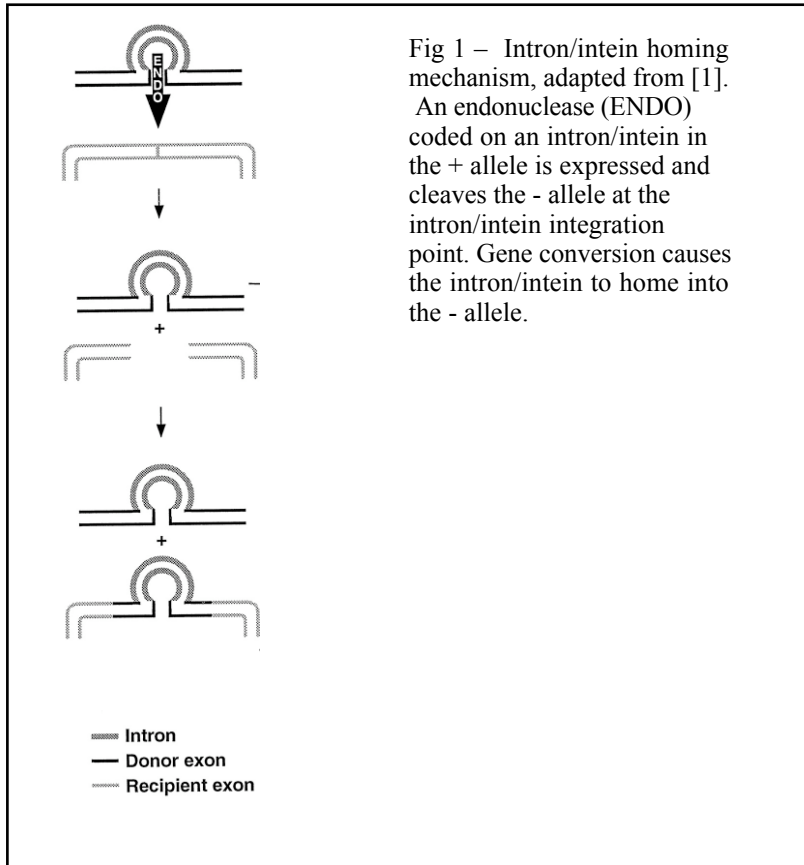
Four families of H-EN are known. Each family is characterized by several conserved short sequence motifs (termed LAGLIDADG, GIY-YIG, H-N-H, and His-Cys). Within each family, only the conserved motifs are common to all sequences [2]. Additional proteins have these sequence motifs, but are not known to have H- EN function. Some are proteins with known nuclease activity, and some are of unknown activity and function. Thus, the H-EN term should actually mean homing endonuclease like proteins.

Introns and inteins are mobile genetic elements found in species from all three domains of life

(*Eukaryota*, *Eubacteria* and *Archaea*). Apparently, they do not contribute any advantage to their host gene or organism. Rather, they survive by efficiently excising themselves from the host's RNA transcript (introns) or protein product (inteins). Many of these elements also encode an H-EN, as noted above. The H-EN mediates the horizontal transfer of the element into unoccupied integration points – '-' alleles – alleles that do not contain the particular intron or intein. This is called "homing". The H-EN makes a site-specific double stranded break in the '-' allele. In doing so, recombinogenic ends are created, which in turn engage in a gene conversion process with the '+' allele, that duplicates the intron or the intein [1], Fig. 1. Structural information is available on LAGLIDADG homing endonucleases (I-CreI [11], PI-Sce [6], I-DmoI [36], PI-PfuI [13]), on a His-Cys homing endonucleases (I-Ppo [8]), on HNH bacteriocin DNases [15, 16], and on a GIY-YIG homing endonucleases (I-TevI [17]). In the LAGLIDADG and His-Cys proteins the catalytic and DNA binding activities are in the same domain, while in the GIY-YIG and HNH proteins the activities are found in separate domains, suggesting modularity. The structural information implies that the known sequence motifs in all H-EN families define the catalytic (DNA cleaving) domain of the proteins.

The active sites of HNH and His-Cys were defined as structurally related [18]. However, there is no structural similarity beyond the catalytic site, and there is no sequence similarity we could detect. The similarity of the catalytic site could be the result of either convergent evolution or an early divergence event. These two families were reclassified as a single family and are called the $\square\square\square$ -Me family. Despite the similarity between the two catalytic sites, it should be noted that the characterized His-Cys proteins [8] work as homodimers while characterized HNH proteins work as monomers [16].

In this work I present computational analysis of two of the above families – HNH, and GIY-YIG (GIY), and an experimental analysis of some proteins from these families. These families were chosen because of their modularity – their catalytic domain is separate from their DNA binding domain. Following, are detailed descriptions of these two H-EN families.



The HNH family

The HNH proteins contain a conserved sequence domain of 30-33 amino acids. It contains two pairs of conserved histidines flanking a conserved asparagine. The domain forms a metal binding catalytic site of the nuclease [9, 35].

The HNH domains occur, both in free standing and in intron-encoded (group I and II) proteins, and in different contexts [2, 4]. The ability of HNH proteins to generate sequence specific double-stranded breaks was proven for an H-EN from a group I intron of a *Chlamydomonas* chloroplast [5]. Other experiments show that an HNH domain in yeast mitochondrial group II maturases is responsible for the cleavage of one DNA strand [44]. Interestingly, the other strand is cleaved by the spliced intron RNA molecule bound to the maturase. An HNH domain was found in a putative mismatch repair protein MutS of the mitochondria of a coral [29]. In

this context the HNH domain is proposed to carry the nicking function of MutH as analyzed in *E. coli* [25].

Bacteriocins (sometimes called colicins) are plasmid-encoded antibacterial proteins, secreted by bacteria. In this way they provide a competitive advantage against other closely related bacteria, competing for the same resources. Bacteriocins containing the HNH domain are DNases, cleaving DNA with no specific sequence recognition. [14, 31, 34]. The structure of two related HNH bacteriocins was recently determined [15, 16].

An HNH domain is also found in the commercially available type II restriction enzyme NlaIII from *Neisseria lactamica*. It is upstream to a methyltransferase gene [26], and therefore is probably a part of a restriction/modification system. The IceA1 protein from the gastric ulcer related bacteria *Helicobacter pylori* is very similar in sequence to NlaIII, and has the HNH domain. Expression of IceA1 is induced upon contact of *H. pylori* with epithelial cells, and therefor is believed to be related to ulcer formation. Carriage of IceA1 strains is associated with the presence of peptic ulcers [27, 39]. Different isolates of *H. pylori* have different variants of the IceA1 protein. The HNH domain is highly conserved in all strains, although in some strains a frame shift mutation caused a stop codon in the beginning of the gene [7].

These examples, with the addition of the known structure of colicins, imply that the HNH catalytic nuclease domain is separate, both in function and in structure, from the rest of the protein. It is probably an independent module used in several cases where DNA cutting/nicking is needed.

The GIY family

The GIY-YIG domain is 70-100 amino acids long, and in most cases contains 5 distinct motifs. The second motif includes an invariant arginine, and the third motif includes an invariant glutamate, both were found essential for the catalytic activity, but not for the structure of the domain [17]. The GIY family contains group I intronic ORFs from mitochondria and chloroplasts of fungi, algae and liverworts, a group I intronic protein from a phage (I-TevI), intergenic ORFs from phages (SEG A to E and END2 from BPT4), ORFs from *Chlorella* viruses (PBCV-1 and CIV) and ORFs from bacteria [17].

The GIY domain, like the HNH one, appears in different contexts. In the N-terminal domain of the UvrC subunit of bacterial and archaeal excision repair endonucleases the GIY domain is responsible for the 3' damage-specific incision [42].

In H-EN such as I-TevI the GIY domain has an endonuclease activity with a limited sequence specificity. The relationship between sequence, structure, and function of H-EN I-TevI was studied by NMR, sequence, and molecular analysis [17]. That work showed that the catalytic domain – which contains the GIY-YIG domain – is distinctly separate from the DNA binding domain. A flexible linker separates the two domains.

The actual homing ability of an intronic GIY H-EN in *Podospora* mitochondria was recently proven [30].

From all the research described above, a modularity of the nuclease domain in the HNH and GIY H-EN families emerges. The nuclease domain can be separated from the DNA binding domain in structure and in function. This modularity can be useful in engineering nucleases and endonucleases, since the nuclease domains seem amenable to function in different contexts. H-EN by themselves could be useful for gene therapy – for genome single incisions. In addition, H-EN mediate the horizontal transmission of genetic elements, thus, they might be the cause of horizontal gene transfer in nature. Finally, the exact function of many H-EN proteins is unknown, although they seem conserved in evolution.

Paramecium bursaria Chlorella virus 1 (PBCV-1) contains 12 H-EN ORFs, with no known function. These ORFs are apparently the result of gene expansion, since they are more similar to each other than to respective proteins in other organisms. The large number of uncharacterized H-EN like ORFs, the relative simplicity of a viral system, activity in an eukaryotic environment and interesting features of PBCV-1 (described below) led me to choose to experimentally study some PBCV-1 proteins.

Paramecium bursaria Chlorella virus 1

Paramecium bursaria Chlorella virus 1 (PBCV-1) is the prototype member of the *Phycodnaviridae* family. This family of large (190nm diameter) polyhedral viruses, replicate in

certain chlorella-like green algae, which live inside the protozoan *Paramecium bursaria*. This dsDNA virus has a large (330-Kb) linear nonpermuted genome [40].

Sequencing of the PBCV-1 genome [19, 21-24] revealed about 400 ORFs predicted to encode functional proteins [41]. PBCV-1 contains a high level of methylated DNA bases, and encodes multiple site specific restriction endonucleases and methyltransferases. Several of these endonucleases (that do not contain H-EN domains) have been characterized, and are frequent cutters – 4 base pairs or less. Degradation of host chloroplast and nuclear DNA occurs 1 to 2 hours post infection, by presumably a virus encoded protein(s). Since the total DNA levels increase 3 to 10 fold 5 hours post infection [40], the nucleotides released because of the DNA degradation could be used for recycling into the virus DNA.

Put together, PBCV-1 offers several advantages for studying H-ENs. Its H-ENs are active in eukaryotic cells (but not in organelles). Being a virus its life cycle is relatively simple. Finally, it encodes a large number of totally uncharacterized H-EN proteins. This led us to study those proteins experimentally, as well as computationally.

Methods

Computational methods

Building blocks

Our first step in sequence analysis is to define conserved areas. To do so, several similar sequences were aligned. Similar sequences can be identified by using sequence to sequence database searches such as the Blast program, which I used to search different NCBI sequence databases, and unfinished genomes (from the NCBI and TIGR centers).

Multiple alignment of the sequences was made by different local multiple alignment methods. The main method used is BlockMaker [12], which is an automated system that finds blocks in a group of protein sequences. BlockMaker uses and extends the Gibbs [20] algorithm based on Gibbs sampling, and the Motif [37] algorithm, which is based on identifying spaced triplets. Other programs I used are the interactive MACAW [32] program – based on Gibbs sampling algorithm, and the automated MEME program [10], which uses an expectation maximization algorithm.

After the initial blocks are built, they are used to search the NCBI non-redundant protein and nucleotide databases (using BLIMPS [43] and MULTIMAT [12] programs). Similar sequences are added to the blocks. The new blocks are in turn used to search the sequence databases again.

In order to find blocks in sub-families, two different approaches were used:

1Æ Defining subgroups of sequences with higher similarity using blast, and analyzing them separately.

2Æ Subtracting the block area, and analyzing the remaining sequence regions, termed ‘non-catalytic’ since they do not contain the catalytic H-EN domain.

The second approach was also used to define blocks common to the HNH and GIY sequences. After blocks were built, the LAMA program [28] was used to identify similar blocks by comparing the new blocks to those in the Blocks+ database.

Experimental methods

Over-expression constructs

James Van-Etten’s lab sent four Clones of PBCV-1 genes:

A267L, A315L, and A495R in pET23d(+) (Novagen), and A490L in pET23a(+) (Novagen).

These plasmids have a T7 promoter and a C-terminal 6xHis tag.

After failing to over-express the proteins, A315L and A495L were cloned into pET38b(+) at NcoI and XbaI sites. pET38b(+) has the more stringent, T7-Lac repressor promoter, a C-terminal cellulose binding domain (CBD) and 6xHis tag (Novagen). For plasmid DNA purification DH5α *E. coli* strain was used.

Induction of expression

The plasmids were introduced into host strain BL21 (DE3) pLysE that expresses the T7 RNA polymerase from the inducible lac UV 5 promoter [38]. Cells were grown in M9ZB + 0.4% glucose, with 30µg/ml kanamycin at 37°C to OD₆₀₀=0.6. Expression was induced by IPTG to a total concentration of 1mM, at 37°C. Cells were harvested after 2 1/2 hours, centrifuged and

kept in -20°C over-night.

Purification of A315 protein

Cell extract was thawed and resuspended in native lysis buffer (50mM NaH_2PO_4 ; 10mM NaCl pH 8.0) and sonicated. Insoluble material was removed by centrifugation (20' 15000 g). anti protease cocktail was added to the supernatant (Sigma P8465 protease inhibitor cocktail, used according to the manufacturer's instructions) to prevent proteolysis. The protein was extracted using nickel beads (CytoSignal Ni^{2+} charged resin) according to QIAGEN protocol, using QIAGEN 1ml polypropylene columns. The proteins were eluted from the column with 200 μl native elution buffer (50mM NaH_2PO_4 ; 10mM NaCl; 250mM imidazole pH 8.0).

Extracted protein was detected using western blot with mouse anti 6xHis antibodies (H1029 sigma 1:3000) and peroxidase-conjugated affinipure goat anti-mouse IgG (H+L) secondary antibodies (115-035-003 Jackson 1:10000). The substrate used was super signal (34080 pierce).

Activity assay

The digestion of a closed plasmid (pMAL C2 from New England BioLabs) by the extracted protein was performed in the following conditions:

2 μl of 200ng/ μl DNA was incubated with 1 μl eluted protein, 1 μl buffer Y tango (Fermentas, 66 mM K-Ac, 33 mM Tris-OAc, 7.9, 10 mM MgOAc, , 0.1 mg/ml) and 6 μl ddH₂O, for 1 hour. The reaction was stopped by the addition of 2 μl 6x Orange loading dye solution (Fermentas), and the samples were run on a 0.8% agarose gel.

Other DNA types also used as substrate were –

A linearized plasmid (pMAL C2 cut with XbaI), different plasmids (pET28c, pET23a+d pET38b, and pET14 from Novagen; and pTYB-2 from New England BioLabs) and *Saccharomyces cerevisiae* genomic DNA and *Schizosaccharomyces pombe* DNA libraries.

The activity was also examined in different buffers - 2xY, R, O, and B, buffers from Fermentas (**B**: 10mM tris-HCl, 7.5, 10mM MgCl₂, 0.1 mg/ml BSA; **R**: 100 mM KCl, 10mM tris-HCl, 8.5, 10mM MgCl₂, 0.1 mg/ml BSA; **O**: 100 mM NaCl, 50 mM tris-HCl, 7.5, 10mM MgCl₂, 0.1 mg/ml BSA; **Y 2x**: 132 mM K-Ac, 66 mM Tris-OAc, 7.9, 20 mM MgOAc, 0.2 mg/ml

BSA).

A Time-course analysis was made by doing the reaction in a volume 5 times larger from the above (increasing all ingredient's amount accordingly), and taking a 10 μ l sample in 10' intervals for 30', and an additional sample after 60'. Loading buffer was added to each sample and kept on ice, until all samples were run together on a gel.

As a control, I used an empty pET38b(+) vector – that codes for a protein of 20KD containing only the periplasmic signal, CBD, and 6xHis-tag, but not the H-EN gene.

An additional control was a mutated A315L gene. Mutation to alanine was made at the catalytic arginine (R26A) [17] of the GIY domain. The mutation was made by QuikChange Site-Directed Mutagenesis using Turbo Pfu according to the Stratagene manual [3]. An additional silent mutation was inserted in order to create a new HindIII restriction site. In this method replication of the whole plasmid is made in a PCR machine, using complementary long (55BP) primers that contained the desired mutations. The template DNA is degraded by DmoI enzyme (Fermentas), which cuts only methylated DNA. Then the new plasmids, that contain the mutation, were transformed to *E. coli* (DH5 α). The mutations were confirmed by sequencing both strands of the resulting plasmid.

Results

In the computational analysis I used the HNH and GIY domains, in blocks format, to identify new HNH/GIY protein sequences. After adding these new sequences, and so refining the GIY and HNH blocks, I looked for new conserved domains outside the catalytic HNH/GIY domains.

Next I confirmed experimentally a sequence analysis prediction of the function of proteins that have an HNH or GIY domain. Following are details of these analyses and experiments.

Computational results

H-EN motifs and analysis

I used conserved sequence regions in 80 HNH and 54 GIY proteins to iteratively search sequence databases, as described in methods. I identified 35 HNH and 5 GIY protein sequences that were not previously described as H-EN in sequence databases (SwissProt,

GenBank) or in the literature [4, 17].

Selected organisms with numerous H-EN proteins

Organism	Total number of proteins in genome. ^a	Proteins with a H-EN domain ^b	
		HNH domains	GIY domains
<i>Escherichia coli</i>	4289	7 ^f	2 ^e
<i>Synechocystis sp.</i> <i>PCC6803</i>	3169	6	3 ^e
Chlorella virus PBCV-1	~400 ^c	5	7
Bacteriophage T4	279	5	7 ^d
<i>Podospira anserina</i> mitochondria	82	3 ^g	9 ^g

^a www3.ncbi.nlm.nih.gov/Genomes

^b Hypothetical proteins unless noted otherwise.

^c [41]

^d Including the intron encoded homing endonuclease I-TevI and 6 other known endonucleases.

^e Including UvrC (see introduction for details).

^f Including 4 Colicins, and 2 reverse transcriptases with an HNH domain.

^g Group I intron encoded proteins [17, 33].

Comparing HNH sequences, PBCV-1 sequences cluster together. The same result emerges when comparing GIY sequences. Similar subgroups occur in T4 bacteriophages as well. This suggests multiple gene duplication events in bacteriophage T4 and in PBCV-1, indicating a meaningful advantage for having several such proteins with a putative nuclease function. We do not yet know what is the role of this putative function but a few suggestions can be raised. These putative nucleases could be used for cutting DNA or RNA in different contexts. They could be used for instance- in host lysis, in virion release, in maturation of nucleic acids, and in protection against competing viruses/pathogens (restriction).

Additional motifs (Homing Endonuclease non-catalytic):

HNH and GIY sequences were further analyzed to characterize the non-catalytic domain (defined as the sequence regions outside the HNH and GIY domains), and hopefully to define

a DNA binding domain in some of these sequences. Several conserved motifs were found, and although no similarity was found between them and any other known protein, some interesting features do emerge.

Four new domains were defined. Some unique to a sub-group of sequences from one of the H-EN families, two include sequences from both families. Three of the domains are characterized by a single block and one by three blocks. Two of the domains appear either in single or multiple occurrences and two appear only in a single occurrence per protein. The domains appear together in a complicated pattern, as represented graphically in Fig 2. The composition of each domain is represented as sequence logos [12], Fig.3A, B.

HEnc1 – This domain consists of a single motif that is present in both HNH and GIY sequences. It appears C terminal to the catalytic blocks, in 16 sequences and is 34 amino acids long. With the HNH domain it appears only in Bacteriophage T4, in which it is repeated twice. It appears in 9 GIY proteins from PBCV-1, mitochondria of different fungi and of green algae, and is repeated three times in one sequence (o360_almmt).

HEnc2 – This domain also has a single motif. It is present both with HNH (in 2 bacteriophages), and with GIY (in mitochondrial proteins from two different *Ascomycota* fungi) domains. It appears C terminal to the catalytic domains, in four sequences, and is 19 amino acids long. Although it consists of only four sequences, it is significant – searching the SwissProt database (containing 80000 sequences) with this block, the sequences forming it were found with higher scores than any other sequence.

HEnc3 – This domain is present only in HNH sequences. It consists of three different motifs, in 11 sequences from 9 bacteriophages and from *Bacillus subtilis*. These motifs appear N terminal to the catalytic HNH blocks. The motifs are 19, 9 and 12 amino acids long.

HEnc4 – This single motif domain is present only in GIY sequences. It appears in 43 sequences from different Fungi mitochondria, PBCV-1, bacteriophage T4, and from the mitochondria of *Chlorogonium elongatum* (*chlorophyta*). It is 14 amino acids long.

After building these motif blocks, they were used to search the Blocks+ database using the LAMA block to block search method. LAMA search found no significant similarity between the non-catalytic blocks, and any known blocks. Since all characterized HNH/GIY proteins bind nucleic acids, and such binding is often modular in nature, the domains I identified could be new nucleic acid binding domains.

The proposed modularity of the non-catalytic domains is supported by the fact that HEnc1 and HEnc2 occur both with HNH and GIY domains. A domain similar in sequences that otherwise have no similarity could be the result of domain shuffling. Domain shuffling is a process in which functional domains associate to a protein independently, and so a protein can acquire an intact function.

Some of the newly defined motifs occur in PBCV-1, as described above. In addition, two

PBCV-1 proteins have a duplication of their HNH domain. This, the fact that more HNH/GIY proteins with no previously known function were found in PBCV-1 than in any other organism, the simple virus system in a eukaryotic environment and the interesting features of this virus led us to further examine the H-EN sequences from PBCV-1.

ORFs adjacent to the proteins containing H-EN domain in PBCV-1 were studied to find operons or methylase-endonuclease pairs. None were found.

Several possible reasons consist for multiple putative nucleases in PBCV. They could be important for the relation of the virus and its host, for instance the lysis of the host, cutting through the host DNA that might impede virion release and diffusion or, the nucleotides released as a result of the DNA degradation could be used for recycled into the virus DNA. These proteins could also be important for intrinsic virus functions such as the maturation of nucleic acids, or protection against competing viruses or other intracellular pathogens (restriction). In any case, multiple putative nucleases point to some kind of advantage they give to PBCV-1. The PBCV-1 ORFs with HNH/GIY domains are listed in appendix 1.

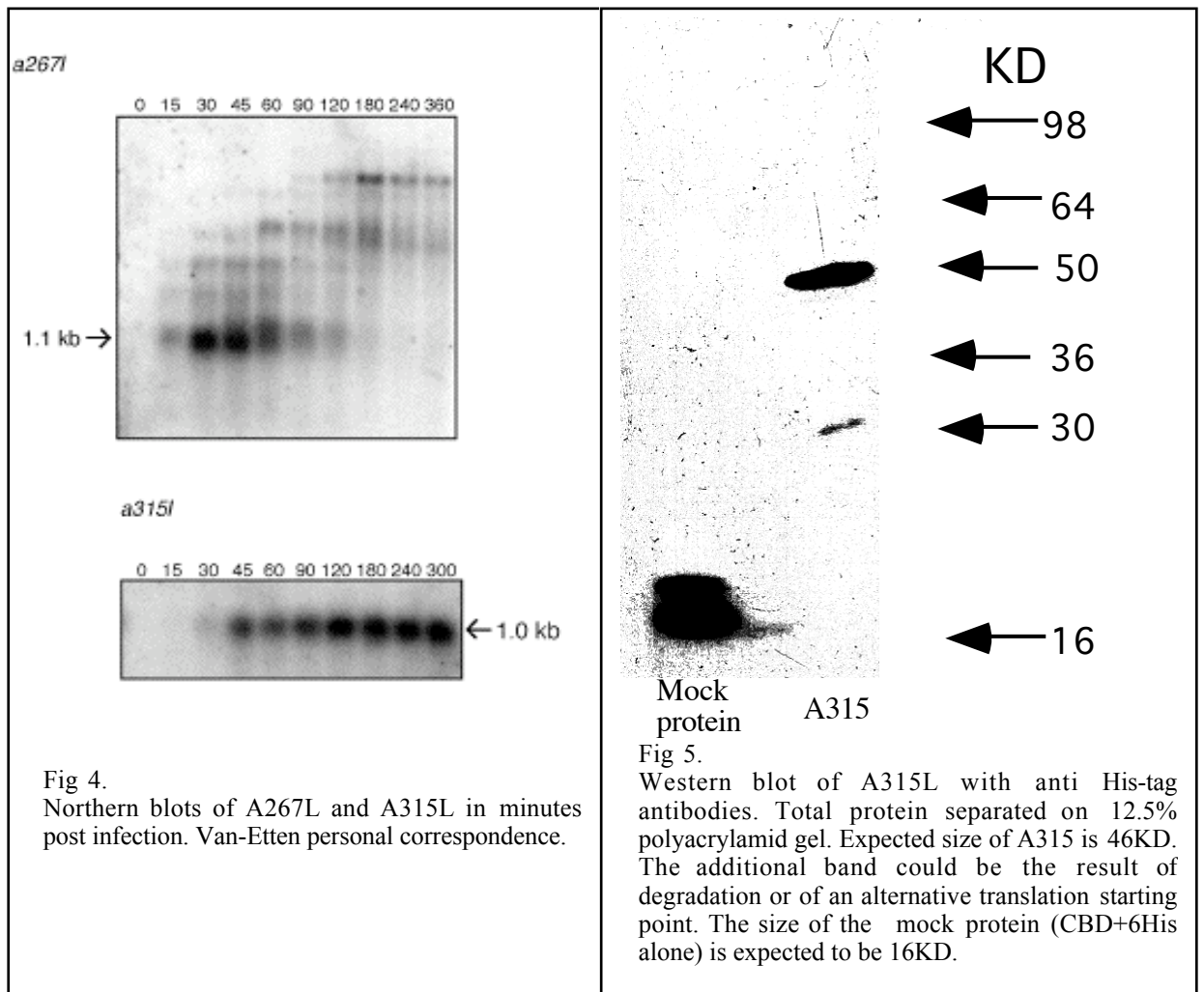
Experimental results

Expression of ORFs A315L and A267L in PBCV-1 was checked in James Van-Etten's lab (personal correspondence, Fig4) by northern blots. They were found to be early genes, expressed by 15 and 30 minutes respectively post infection. A267L transcript disappears at about 120', while A315L can be detected throughout the remainder of the infection cycle.

Van-Etten's lab cloned and sent us four PBCV-1 ORFs: A267L, A315L, A490L, and A495R. I was unable to express any of the four clones in *E. coli*. A possible reason for this might be the fact that the proteins are toxic to the expressing bacteria. To overcome this problem the proteins were re-cloned into a vector with a more stringent promoter, as described in methods. A315L was re-cloned, and expressed. Since no clear band could be seen on coomassie gel staining either of the total cell lysate or of the soluble fraction purified on a nickel beads column, a western blot was performed. A protein in the expected size (~46 KD) was detected by western blot using anti His tag antibodies Fig 5.

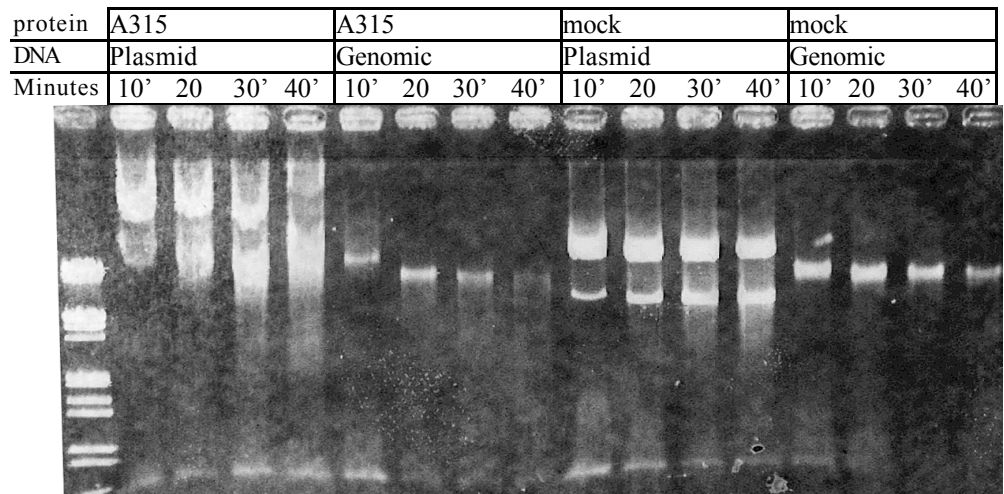
Protein cleaned on nickel beads was incubated with DNA to check its activity. In a time course

activity experiment it could be seen that the DNA is retarded and might be degraded, Fig 6A,B. These preliminary results suggest DNA binding and cleaving activities of A316L. In the first experiments my control was a mock protein – only the CBD+His tags. To get a better control to the protein’s nuclease activity I used a mutant. The mutation was made in the Arg 26 catalytic site of the GIY domain [17]. In preliminary activity assays the R26A mutant protein binds to DNA, Fig 6C. No clear difference in the amount of DNA was seen. Reasons for this are described in the discussion.



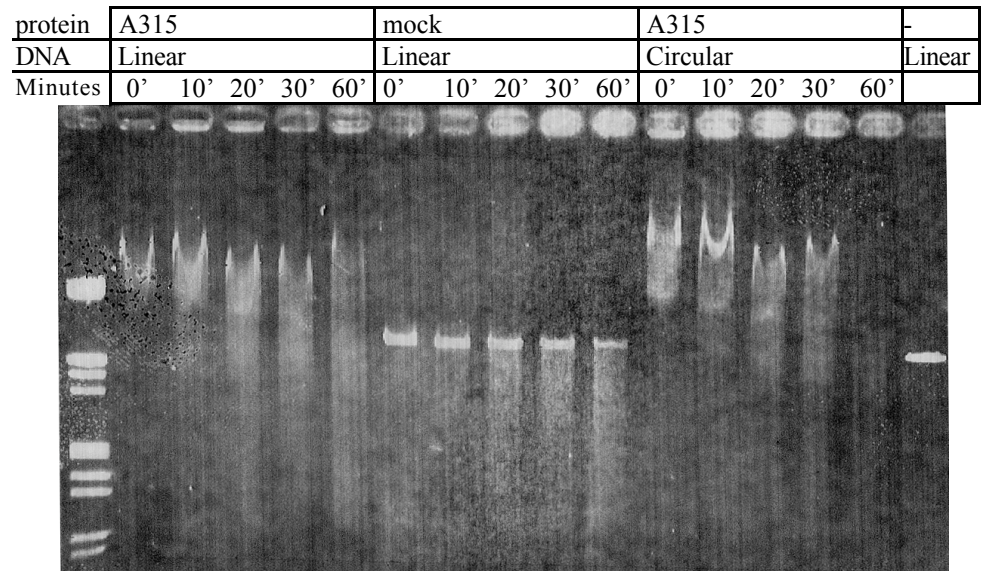
A

Genomic and plasmid DNA



Linear and plasmid DNA

B



C

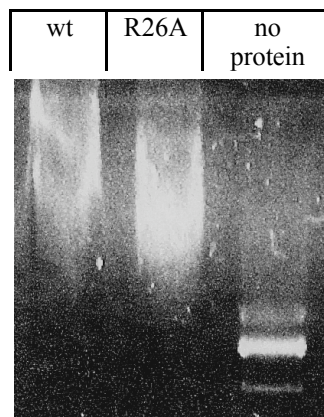


Fig 6. Activity analysis. All activity tests were in 37°C, all gels are 0.8% agarose gels. A. A315L protein with plasmid (pMAL C2) and genomic (Yeast) DNA as substrates. B. A315L protein with circular and linearized plasmid (pMAL C2) as substrates. C. wt and mutant (R26A) proteins incubated for 30' with plasmid (pMAL C2) DNA.

Discussion

Modularity of the HNH and GIY catalytic domains was previously shown. Here, I demonstrate modularity of new domains, which are assumed not to take part in the catalytic activity of the protein. These new domains can be responsible for the DNA binding function of these proteins, although this function should yet be tested.

The experimental part of my study shows that one of the proteins predicted as a nuclease binds DNA, but evidence regarding DNA degradation was inconclusive. This could be due to the DNA smear, that makes quantification the amount of degradation difficult, or to the specificity of the nuclease, that will cut only a specific sequence, or work only in specific conditions.

To confirm my finding, several more experimental studies should be made. To quantify DNA degradation the DNA could be separated from the protein by SDS before running the DNA on a gel. The activity of A315L should be tested in different conditions. The natural DNA substrate is unknown, so different DNA substrates, such as PBCV-1 genome, and the chlorella genome, should be tested as possible substrates. The protein's activity should be tested in different pH, temperatures and buffers, with different divalent ion and salt concentrations. In varying those different conditions a sequence specific DNA binding function might emerge. Finally, the function of each domain could be studied separately, to confirm their actual modularity.

The significance of this study is the strength of blocks multiple sequence analysis for finding protein function, and the modularity of the non-catalytic regions in H-EN, as shown by the newly defined conserved domains. As new genomes are sequenced, new H-EN proteins could be identified. With this new information more conclusions could be drawn about homing endonuclease like proteins.

References

1. Belfort M, Perlman PS: "Mechanisms of intron mobility" *J Biol Chem* , **270**:30237-40 (1995).
2. Belfort M, Roberts RJ: "Homing endonucleases: keeping the house in order" *Nucleic Acids Res* , **25**:3379-88 (1997).
3. Braman J, Papworth C, Greener A: "Site-directed mutagenesis using double-stranded plasmid DNA templates" *Methods Mol Biol* , **57**:31-44 (1996).
4. Dalgaard JZ, Klar AJ, Moser MJ, Holley WR, Chatterjee A, Mian IS: "Statistical modeling and analysis of the LAGLIDADG family of site- specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family" *Nucleic Acids Res* , **25**:4626-38 (1997).
5. Drouin M, Lucas P, Otis C, Lemieux C, Turmel M: "Biochemical characterization of I-cmoel reveals that this H-N-H homing endonuclease shares functional similarities with H-N-H colicins" *Nucleic Acids Res* , **28**:4566-72 (2000).
6. Duan X, Gimble FS, Quioco FA: "Crystal structure of PI-SceI, a homing endonuclease with protein splicing activity." *Cell* , **89**:555-64 (1997).
7. Figueiredo C, Quint WG, Sanna R, Sablon E, Donahue JP, Xu Q, Miller GG, Peek RM, Jr., Blaser MJ, van Doorn LJ: "Genetic organization and heterogeneity of the iceA locus of *Helicobacter pylori*" *Gene* , **246**:59-68 (2000).
8. Flick KE, Jurica MS, Monnat RJ Jr, Stoddard BL: "DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-PpoI." *Nature* , **394**:96-101 (1998).
9. Gorbalenya AE: "Self-splicing group I and group II introns encode homologous (putative) DNA endonucleases of a new family" *Protein Sci* , **3**:1117-20 (1994).
10. Grundy WN, Bailey TL, Elkan CP: "ParaMEME: a parallel implementation and a web interface for a DNA and protein motif discovery tool" *Comput Appl Biosci* , **12**:303-10 (1996).
11. Heath PJ, Stephens KM, Monnat RJ, Jr., Stoddard BL: "The structure of I-Crel, a group I intron-encoded homing endonuclease" *Nat Struct Biol* , **4**:468-76 (1997).
12. Henikoff S, Henikoff JG, Alford WJ, Pietrokovski S: "Automated construction and graphical presentation of protein blocks from unaligned sequences" *Gene* , **163**:GC17-26

(1995).

13. Ichiyanagi K, Ishino Y, Ariyoshi M, Komori K, Morikawa K: "Crystal structure of an archaeal intein-encoded homing endonuclease PI- PfuI" *J Mol Biol* , **300**:889-901 (2000).
14. James R, Kleanthous C, Moore GR: "The biology of E colicins: paradigms and paradoxes" *Microbiology* , **142**:1569-80 (1996).
15. Kleanthous C, Kuhlmann UC, Pommer AJ, Ferguson N, Radford SE, Moore GR, James R, Hemmings AM: "Structural and mechanistic basis of immunity toward endonuclease colicins" *Nat Struct Biol* , **6**:243-52 (1999).
16. Ko TP, Liao CC, Ku WY, Chak KF, Yuan HS: "The crystal structure of the DNase domain of colicin E7 in complex with its inhibitor Im7 protein" *Structure Fold Des* , **7**:91-102 (1999).
17. Kowalski JC, Belfort M, Stapleton MA, Holpert M, Dansereau JT, Pietrokovski S, Baxter SM, Derbyshire V: "Configuration of the catalytic GIY-YIG domain of intron endonuclease I- T_{ev}I: coincidence of computational and molecular findings" *Nucleic Acids Res* , **27**:2115-25 (1999).
18. Kuhlmann UC, Moore GR, James R, Kleanthous C, Hemmings AM: "Structural parsimony in endonuclease active sites: should the number of homing endonuclease families be redefined?" *FEBS Lett* , **463**:1-2 (1999).
19. Kutish GF, Li Y, Lu Z, Furuta M, Rock DL, Van Etten JL: "Analysis of 76 kb of the chlorella virus PBCV-1 330-kb genome: map positions 182 to 258" *Virology* , **223**:303-17 (1996).
20. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment" *Science* , **262**:208-14 (1993).
21. Li Y, Lu Z, Burbank DE, Kutish GF, Rock DL, Van Etten JL: "Analysis of 43 kb of the Chlorella virus PBCV-1 330-kb genome: map positions 45 to 88" *Virology* , **212**:134-50 (1995).
22. Li Y, Lu Z, Sun L, Ropp S, Kutish GF, Rock DL, Van Etten JL: "Analysis of 74 kb of DNA located at the right end of the 330-kb chlorella virus PBCV-1 genome" *Virology* ,

237:360-77 (1997).

23. Lu Z, Li Y, Que Q, Kutish GF, Rock DL, Van Etten JL: "Analysis of 94 kb of the chlorella virus PBCV-1 330-kb genome: map positions 88 to 182" *Virology* , **216**:102-23 (1996).
24. Lu Z, Li Y, Zhang Y, Kutish GF, Rock DL, Van Etten JL: "Analysis of 45 kb of DNA located at the left end of the chlorella virus PBCV-1 genome" *Virology* , **206**:339-52 (1995).
25. Malik HS, Henikoff S: "Dual recognition-incision enzymes might be involved in mismatch repair and meiosis" *Trends Biochem Sci* , **25**:414-8 (2000).
26. Morgan RD, Camp RR, Wilson GG, Xu SY: "Molecular cloning and expression of NlaIII restriction-modification system in E. coli" *Gene* , **183**:215-8 (1996).
27. Peek RM, Jr., Thompson SA, Donahue JP, Tham KT, Atherton JC, Blaser MJ, Miller GG: "Adherence to gastric epithelial cells induces expression of a Helicobacter pylori gene, iceA, that is associated with clinical outcome" *Proc Assoc Am Physicians* , **110**:531-44 (1998).
28. Pietrokovski S: "Searching databases of conserved sequence regions by aligning protein multiple-alignments" *Nucleic Acids Res* , **24**:3836-45 (1996).
29. Pont-Kingdon GA, Okada NA, Macfarlane JL, Beagley CT, Wolstenholme DR, Cavalier-Smith T, Clark-Walker GD: "A coral mitochondrial mutS gene" *Nature* , **375**:109-11 (1995).
30. Saguez C, Lecellier G, Koll F: "Intronic GIY-YIG endonuclease gene in the mitochondrial genome of Podospora curvicolla: evidence for mobility" *Nucleic Acids Res* , **28**:1299-306 (2000).
31. Sano Y, Matsui H, Kobayashi M, Kageyama M: "Molecular structures and functions of pyocins S1 and S2 in Pseudomonas aeruginosa" *J Bacteriol* , **175**:2907-16 (1993).
32. Schuler GD, Altschul SF, Lipman DJ: "A workbench for multiple alignment construction and analysis" *Proteins* , **9**:180-90 (1991).
33. Sellem CH, d'Aubenton-Carafa Y, Rossignol M, Belcour L: "Mitochondrial intronic open reading frames in Podospora: mobility and consecutive exonic sequence variations" *Genetics* , **143**:777-88 (1996).
34. Seo Y, Galloway DR: "Purification of the pyocin S2 complex from Pseudomonas

- aeruginosa PAO1: analysis of DNase activity" *Biochem Biophys Res Commun* , **172**:455-61 (1990).
35. Shub DA, Goodrich-Blair H, Eddy SR: "Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns" *Trends Biochem Sci* , **19**:402-4 (1994).
36. Silva GH, Dalgaard JZ, Belfort M, Van Roey P: "Crystal Structure of the Thermostable Archaeal Intron-encoded Endonuclease I-DmoI." *J Mol Biol* , **286**:1123-1136 (1999).
37. Smith HO, Annau TM, Chandrasegaran S: "Finding sequence motifs in groups of functionally related proteins" *Proc Natl Acad Sci U S A* , **87**:826-30 (1990).
38. Studier FW, Rosenberg AH, Dunn JJ, Dubendorff JW: "Use of T7 RNA polymerase to direct expression of cloned genes" *Methods Enzymol* , **185**:60-89 (1990).
39. van Doorn LJ, Figueiredo C, Sanna R, Plaisier A, Schneeberger P, de Boer W, Quint W: "Clinical relevance of the cagA, vacA, and iceA status of Helicobacter pylori" *Gastroenterology* , **115**:58-66 (1998).
40. Van Etten JL, Lane LC, Meints RH: "Viruses and viruslike particles of eukaryotic algae" *Microbiol Rev* , **55**:586-620 (1991).
41. Van Etten JL, Meints RH: "Giant viruses infecting algae" *Annu Rev Microbiol* , **53**:447-94 (1999).
42. Verhoeven EE, van Kesteren M, Moolenaar GF, Visse R, Goosen N: "Catalytic sites for 3' and 5' incision of Escherichia coli nucleotide excision repair are both located in UvrC" *J Biol Chem* , **275**:5120-3 (2000).
43. Wallace JC, Henikoff S: "PATMAT: a searching and extraction program for sequence, pattern and block queries and databases" *Comput Appl Biosci* , **8**:249-54 (1992).
44. Zimmerly S, Guo H, Eskes R, Yang J, Perlman PS, Lambowitz AM: "A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility" *Cell* , **83**:529-38 (1995).

Appendix 1

ORFs from PBCV-1 that have HNH or GIY domains

ORF	HNH/GIY domains	Additional domains	Comments
A134L	GIY	-	
A267L	2 x HNH		In both copies of the HNH domain a position with usually a highly conserved His is Asn.
A287R	GIY	3 x HEnc4 , HEnc1	
A315L	GIY	3 x HEnc4 , HEnc1	Similar to A651L
A351L	GIY	-	
A422R	HNH	-	
A478L	2 x HNH	-	Similar to A490L
A490L	2 x HNH	-	Similar to A478L
A495R	GIY	3 x HEnc4	
A539R	GIY	-	
A651L	GIY	3 x HEnc4 , HEnc1	Similar to A315L
A87R	HNH	-	

Appendix 2

Sequence information

Accession	gi	organism	description ¹
A287R_pbcv1	1181450	Paramecium bursaria Chlorella virus 1	
A315L_pbcv1	1181478	Paramecium bursaria Chlorella virus 1	
A651L_pbcv1	2447115	Paramecium bursaria Chlorella virus 1	
COBi1_ceumt	2865254	Chlamydomonas eugametos mt	Intronic ORF
COBi1_chemt	2193888	Chlorogonium elongatum	COB I ORF3
COBi1_necmt	13116	Neurospora crassa mt	COB I1
COBi1_sadmt	13617	Sacchoromyces douglasii	COB I1
COBi2_poamt	1334531	Podospira anserina mt	COB I2
COIi14_poamt	1334547	Podospira anserina mt	COI I14
COIi4_agamt	2738528	Agrocybe aegerita mt	COI I4
Y03E_BPT4	141153	coliphage T4	
o296_CVK2	2190279	Chlorella virus strain:CVK2	
o360_almmt	459018	Allomyces macrogynus mt	
ND1i1_necmt	14129	Neurospora crassa mt	ND1 I1
ND1i1_pocmt	1743352	Podospira comata mt	ND1 I1
e20_BPbIL170	3282299	bacteriophage bIL170	
en_BpphiE	495456	Bacteriophage phi-E	
IHmulI_BPSP82	1085761	phage SP82	
O36.1_BPSP1	1090456	Bacteriophage SPP1	
O38_BPPLLH	945381	Bacteriophage LL-H	
YG31_BPSP1	465641	Bacteriophage SPO1	
e11_BPbIL170	3282308	bacteriophage bIL170	
e37_BPbIL170	3282283	bacteriophage bIL170	
o193_bpt5	g15420	Bacteriophage T5	HNH-endonuclease like ORF
o194_bpt5	g15420	Bacteriophage T5	HNH-endonuclease like ORF
o41_BPr1t	1353558	Bacteriophage r1t	
yosQ_bacsu	2634396	Bacillus subtilis	
A495R_pbcv1	1620166	Paramecium bursaria Chlorella virus 1	
AT6i1_poamt	83822	Podospira anserina mt	ATPase6 I1 protein
COBi1_trpmt	732979	Trimorphomyces papilionaceus mt	COB I1
COBi1b_poamt	578862	Podospira anserina mt	COB I1b
COIi1_poamt	1334559	Podospira anserina mt	COI I1
COIi2_chemt	4379168	Elongatum mt intronic	
ND1i2_almmt	2147548	Allomyces macrogynus mt	
ND1i4b_poamt	1334568	Podospira anserina mt	ND1 I4b
SEGA_BPT4	417766	coliphage T4	endonuclease segA
SEGC_BPT4	2506234	coliphage T4	endonuclease segC
SEGD_BPT4	730735	coliphage T4	endonuclease segD
TEV1_BPT4	119333	coliphage T4	intron associated endonuclease 1
o211a_almmt	2144206	macrogynus	

1)
 COB- mitochondrial apocytochrome b
 COI/II cytochrome oxidase c subunit I/II
 ND1- in NADH dehydrogenase subunit 1
 I# - intron number #