



Structural bioinformatics

In silico identification of archaeal DNA-binding proteins

Linus Donvil^{1,2}, Joëlle A.J. Housmans³, Eveline Peeters¹, Wim Vranken^{4,5,6,7,8} ,
Gabriele Orlando^{9,*} 

¹Research Group of Microbiology, Department of Bioengineering Sciences, Vrije Universiteit Brussel, Brussels B-1050, Belgium

²Center for Neurosciences (C4N), Vrije Universiteit Brussel, Research Group Experimental Pharmacology (EFAR), Jette 1050, Belgium

³Department of Food Technology, Safety and Health, Faculty of Bioscience Engineering, Research Unit VEG-i-TEC, Ghent University, Kortrijk 8500, Belgium

⁴Interuniversity Institute of Bioinformatics in Brussels, ULB/VUB, Brussels 1050, Belgium

⁵Structural Biology Brussels, Vrije Universiteit Brussel, Brussels 1050, Belgium

⁶AI Lab, Vrije Universiteit Brussel, Brussels 1050, Belgium

⁷Department of Chemistry, Vrije Universiteit Brussel, Brussels 1050, Belgium

⁸Department of Biomedical Sciences, Vrije Universiteit Brussel, Brussels 1050, Belgium

⁹Laboratory of Pathogens and Host Immunity, University of Montpellier, CNRS and INSERM, Montpellier 34095, France

*Corresponding author. Laboratory of Pathogens and Host Immunity, University of Montpellier, Bât 24 CC0107 Place Eugène Bataillon 34095 Montpellier, cedex 5, France. E-mail: gabriele.orlando@umontpellier.fr.

Associate Editor: Arne Elofsson

Abstract

Motivation: The rapid advancement of next-generation sequencing technologies has generated an immense volume of genetic data. However, these data are unevenly distributed, with well-studied organisms being disproportionately represented, while other organisms, such as from archaea, remain significantly underexplored. The study of archaea is particularly challenging due to the extreme environments they inhabit and the difficulties associated with culturing them in the laboratory. Despite these challenges, archaea likely represent a crucial evolutionary link between eukaryotic and prokaryotic organisms, and their investigation could shed light on the early stages of life on Earth. Yet, a significant portion of archaeal proteins are annotated with limited or inaccurate information. Among the various classes of archaeal proteins, DNA-binding proteins are of particular importance. While they represent a large portion of every known proteome, their identification in archaea is complicated by the substantial evolutionary divergence between archaeal and the other better studied organisms.

Results: To address the challenges of identifying DNA-binding proteins in archaea, we developed Xenusia, a neural network-based tool capable of screening entire archaeal proteomes to identify DNA-binding proteins. Xenusia has proven effective across diverse datasets, including metagenomics data, successfully identifying novel DNA-binding proteins, with experimental validation of its predictions.

Availability and implementation: Xenusia is available as a PyPI package, with source code accessible at <https://github.com/grogdrinker/xenusia>, and as a Google Colab web server application at [xenusia.ipynb](https://colab.research.google.com/github/grogdrinker/xenusia).

1 Introduction

The rise of next-generation sequencing technologies has brought us in a time where a lot of genetic information is being generated. This information, together with annotations such as domains and functions, is continuously added to big databases like UniProt (Coudert *et al.* 2023), which stores all the details about protein sequences. UniProt has grown significantly, now holding over 250 million sequences and still growing. However, upon closer examination, it becomes evident that there is an uneven distribution of information among various organisms. Widely studied organisms like *Homo sapiens*, *Escherichia coli*, and *Danio rerio* (zebrafish) are well represented, while other species suffer from neglectance, resulting in an underrepresentation. This is what experts call an “observational selection bias,” where most species end up with limited data and mainly inaccurate annotations, pushed to the sidelines without much detailed information (Orlando *et al.* 2016).

Currently, there is a big scientific debate about our ability to predict the fold of completely new proteins, especially

when dealing with proteins that have very limited or no evolutionary information available. Even AlphaFold (Jumper *et al.* 2021), which in recent years revolutionized structural biology, struggles to deal with proteins that have limited evolutionary information (Meng *et al.* 2023). A large part of the archaeal proteins belong to this group. By August 2024, UniProt contained only 3932 archaeal entries with evidence of existence at the protein level. The main problem includes the complete lack of information for a large portion of archaeal organisms. Recent studies on metagenomics data (Emerson *et al.* 2016, Parks *et al.* 2017) highlighted the inability of many archaeal species that live in extreme environments to grow *in vitro*. Fortunately, new sequencing technologies have been successful in extracting genomic and proteomic information from such extremophiles. While some of these proteins showed homology with both bacteria and eukaryotes, others were completely novel (Emerson *et al.* 2016). The lack of homologues makes novel proteins extremely challenging to annotate, requiring *ad hoc*

Received: 22 October 2024; Revised: 17 March 2025; Editorial Decision: 31 March 2025; Accepted: 31 March 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

bioinformatics approaches to assign their putative function. Given the extreme environments that many of these organisms live in, they likely have folds that we have never seen before, and for which there are almost no similar proteins available for comparison. Investigating these unknown areas could greatly improve our knowledge about protein sequence and structural space.

This exploration is not only scientifically relevant but has also significant practical implications. Archaea survive in extreme environments, suggesting that their proteins may have unique molecular structures with valuable applications. Their resilience to harsh conditions could provide insights for industrial processes requiring protein stability, such as the development of heat-resistant enzymes.

A particularly important group of proteins are those that have the ability to interact with DNA. These proteins make up around 2%–5% (Zaman *et al.* 2017) of a species' proteome and are essential for various biological processes, including gene regulation and DNA repair (Iovine *et al.* 2011, Hudson and Ortlund 2014). While our current knowledge of archaeal protein structures is limited, the archaeal DNA-binding domains appear quite similar to those in bacteria, concerning structure and topology (Aravind and Koonin 1999). However, correctly identifying DNA-binding proteins in archaea is still an open issue. To address this challenge, we created a neural network-based tool called Xenusia. This tool operates without the need for multiple sequence alignment (MSA). The exclusion of MSA-based features has two significant effects: firstly, Xenusia provides extremely fast predictions, allowing proteome-wide analyses. Secondly, the lack of reliance on evolutionary features enhances Xenusia's resistance to overfitting while simultaneously allowing it to identify convergent evolution. We demonstrate its effectiveness on diverse datasets, including metagenomics data. Xenusia successfully identifies novel DNA-binding proteins, a finding confirmed through experimental examination. Xenusia is available as a PyPi package and as source code at <https://github.com/grogdrinker/xenusia>, and as a Google Colab web server application at <https://colab.research.google.com/drive/1c4eb4sEz8OsBqHL62XDFrQmwa7CxImww?usp=sharing>.

2 Materials and methods

2.1 Datasets

For this study, we constructed three distinct datasets from publicly available online resources. Two of these datasets were used exclusively for training and optimization, while the third dataset was reserved for final validation. The training datasets include: (i) the contact dataset, comprising protein sequences derived from PDB structures, where each residue is annotated with a binary label indicating whether the corresponding amino acid interacts with DNA and (ii) the domain dataset, which incorporates per-residue annotations from UniProt regarding DNA-binding domains. To reduce the risk of overfitting, both of these datasets contain only bacterial proteins.

The third dataset, designated as the validation dataset, contains solely archaeal proteins. This dataset was not used for any optimization or hyperparameter tuning to further mitigate overfitting risks. Additionally, all datasets were curated to ensure that both internal sequence identity and cross-dataset sequence identity are below 20% (obtained using CD-HIT; Li and Godzik 2006), ensuring minimal redundancy.

2.1.1 Contact dataset

To construct this dataset, we utilized the filtering function of the PDB website (2018 version) and selected proteins based on the following criteria: (i) the proteins must be of bacterial origin, (ii) the structure must have been resolved using X-ray diffraction, (iii) the structure resolution must be $<2\text{ \AA}$, and (iv) the structure must contain both protein and DNA cocrystallized.

The selected proteins were then annotated using a PyMOL script, which labeled each residue as interacting (1) if it has at least one atom within 1.5 \AA of a DNA atom, or noninteracting 0 otherwise. This process resulted in a dataset of sequences where each residue is annotated as either interacting or noninteracting with DNA. The final dataset comprises 78 annotated sequences. All datasets used in this study are available in the associated Git repository.

2.1.2 Domain dataset

This dataset was constructed using the UniProt filtering system by selecting bacterial proteins that have been manually annotated as DNA-binding proteins. To achieve this, we employed the keyword “DNA-binding” (KW-0238) within the UniProt search engine. UniProt also provides annotations for the regions corresponding to DNA-binding domains, enabling us to categorize the residues of each protein as either part of a DNA-binding domain (1) or not 0. The database consists of 162 DNA-binding proteins.

2.1.3 Validation dataset

This dataset is composed exclusively of archaeal proteins and was manually constructed and curated using data from the literature. Proteins labeled as positive were identified in the literature as having DNA-binding activity. In contrast, proteins labeled as negative were selected based on their extensive annotations in UniProt, with no reported DNA-binding activity or any other function involving interactions with nucleic acids. The final dataset comprises 211 proteins, of which 174 are negative hits and 37 are positive hits.

2.2 Computational approach

Xenusia is a neural network-based tool that implements a two-step prediction pipeline, where each step is carried out by a distinct neural network, designated as NN1 and NN2, respectively.

This architecture is specifically designed to predict DNA-binding proteins, aiming to integrate diverse data sources, such as protein structures and domain annotations, while minimizing the number of trainable parameters. This is particularly important in situations like the one addressed in this study, where the available data are very limited. Large-scale parameter optimizations and extensive grid searches of architectures pose a risk, as they can lead to hidden overfitting that is difficult to detect in the absence of large-scale experimental validations.

The first network (neural network 1—NN1) receives the initial feature vector as input and is trained to predict the residues directly interacting with DNA (refer to the contact dataset in the Methods section). The primary objective of NN1 is to identify the biophysical characteristics of amino acids that interact with DNA.

The second network (NN2) is trained on a dataset of DNA-binding domains (refer to the domain dataset in the Methods section). NN2 takes the output of NN1, comprising putative DNA-binding residues, as input and predicts which residues

belong to DNA-binding domains. This architecture enables the model to learn how the patterns of potentially DNA-interacting residues (predicted by NN1) contribute to the formation of a DNA-binding domain (predicted by NN2).

The rationale behind this design is closely linked to overfitting concerns. Extracting domain-specific information from a limited set of proteins lacking detectable evolutionary relationships is challenging. By incorporating intermediate tasks, such as predicting DNA-interacting residues, the network is better constrained, thereby reducing the risk of overfitting.

NN1 utilizes a recurrent neural network architecture, which encodes information from the entire protein sequence. In contrast, NN2 employs a sliding window-based feed-forward neural network, focusing on localized sequence features.

The final score for a protein, which estimates the likelihood that an archaeal protein interacts with DNA, is determined by taking the maximum value of a sliding window average over the residue-based predictions generated by NN2. This approach is based on the observation that a protein's DNA-binding capability is often attributed not to the entire protein, but rather to specific domains within it.

2.2.1 NN1 architecture: prediction of DNA-interacting residues

Every amino acid in the protein sequence is encoded with four features: (i) predicted backbone dynamics and secondary structure propensities for (ii) sheet, (iii) helix, and (iv) coil (using DynaMine; Cilia *et al.* 2014, similarly as described in Raimondi *et al.* 2017). The final input is defined as a window of 11 residues with the residue to be predicted as the central one (residue no. 6), resulting in 44 features (as we did in other tools; Orlando *et al.* 2017, Orlando *et al.* 2022). The feature vectors are then used to feed a 2-layer gated recurrent units (GRUs) neural network with a hidden size of 7, followed by a max-pooling layer and sigmoid activation.

The neural network was trained on the contact dataset for 300 epochs with learning rate of 0.001, batch of 5 and weight decay of 0.0001, using the ADAM optimizer. The network's training task was to predict which residues are in contact with DNA.

2.2.2 NN2 architecture: prediction of DNA-interacting residues

As for the development of NN1, every amino acid in the protein sequence is encoded with four features: (i) predicted early folding predictions (Raimondi *et al.* 2017), (ii) backbone dynamics predictions (Cilia *et al.* 2014), and (iii) a one-hot encoding vector defining the amino acid. The encoding scheme is then enriched with the addition of (iv) the predicted DNA-interacting residues from NN1, which is concatenated to obtain a feature vector describing every residue. The final input is defined by taking a sliding window of 11 contiguous residues and concatenating their feature vectors, as done in previous works (Cilia *et al.* 2014).

The final inputs are used to feed a feed-forward neural network consisting of three layers, with each 30 neurons and rectified linear unit activation. The last layer ends in a single neuron with sigmoid activation, providing the final per-residue prediction.

The neural network was trained on the domain dataset for 300 epochs with learning rate of 0.001, batch of five and weight decay of 0.0001, using the ADAM optimizer. The network's training task was to identify DNA-binding domains, taking the output of NN1 as input.

In order to obtain a single prediction per protein, we performed a sliding window averaging using a window size of 21 residues, and took the maximum average window value of the protein.

2.3 Experimental characterization of Asgard DNA-binding proteins

To evaluate Xenusia's ability to identify novel DNA-binding proteins in archaea, we conducted an experimental validation using metagenomic data from the Asgard superphylum. The following paragraphs describe the experimental pipeline.

2.3.1 Plasmid construction

The pET24a(+) expression vector was used to produce three transcriptional regulator proteins (OLS23561, OLS17986, and OLS31011). This vector allows IPTG-induced T7 polymerase overexpression through its T7 promoter. It also introduces a C-terminal His-tag, enabling protein purification by affinity chromatography. The synthetic DNA fragments of the chosen Asgard genes were inserted into the vector via a unidirectional restriction-ligation reaction, using the FastDigest restriction enzymes NdeI and XhoI (Thermo Fisher Scientific). The vectors were then transformed into competent DH5α *E. coli* cells and plated. After overnight growth at 37°C, multiple colonies were pooled to perform a colony PCR reaction for the identification of functional expression vectors. Sequencing of all positive candidates validated the sequences for each construct.

2.3.2 Protein expression and purification

The validated constructs were then transformed into competent Rosetta (DE3) *E. coli* cells. These cells, being derivatives of BL21 *E. coli* cells that contain additional genes for tRNAs recognizing rare *E. coli* codons, were explored for potentially accelerated translation of heterologous genes. A single colony preculture was grown overnight at 37°C, diluted 1:60 and allowed to grow until the OD600 reached 0.6. Protein expression was initiated through IPTG addition and incubated for 3 h at 37°C. Cells were harvested by centrifugation (10 min at 5000 rpm), washed in 0.9% NaCl and centrifuged again (10 min at 7000 rpm). The pellet was then lysed chemically and physically, using lysis buffer and sonication, respectively. The lysate was centrifuged (10 min at 9500 rpm) to obtain the soluble and insoluble protein fractions. The presence of the protein of interest in these fractions was assessed using SDS-PAGE. The proteins were purified from the soluble fraction through His-tag affinity chromatography, using an ÄKTA-fast protein liquid chromatography system with a HisTrap column (Thermo Fisher Scientific). The protein was eluted from the column by gradually increasing (40–500 mM) the imidazole concentration. The identity of the eluted protein was confirmed through SDS-PAGE analysis of selected peak fractions.

2.3.3 DNA-binding assay

To assess the *in vitro* DNA-binding activity of the purified proteins, a DNA-binding analysis was conducted using three ³²P-labeled dsDNA probes per protein, each approximately 100 bp in length. These probes were chosen to represent the promoter regions proximal to the regulator gene, as many prokaryotic transcription factors exhibit autoregulation, and their target genes are frequently situated in their immediate genomic vicinity.

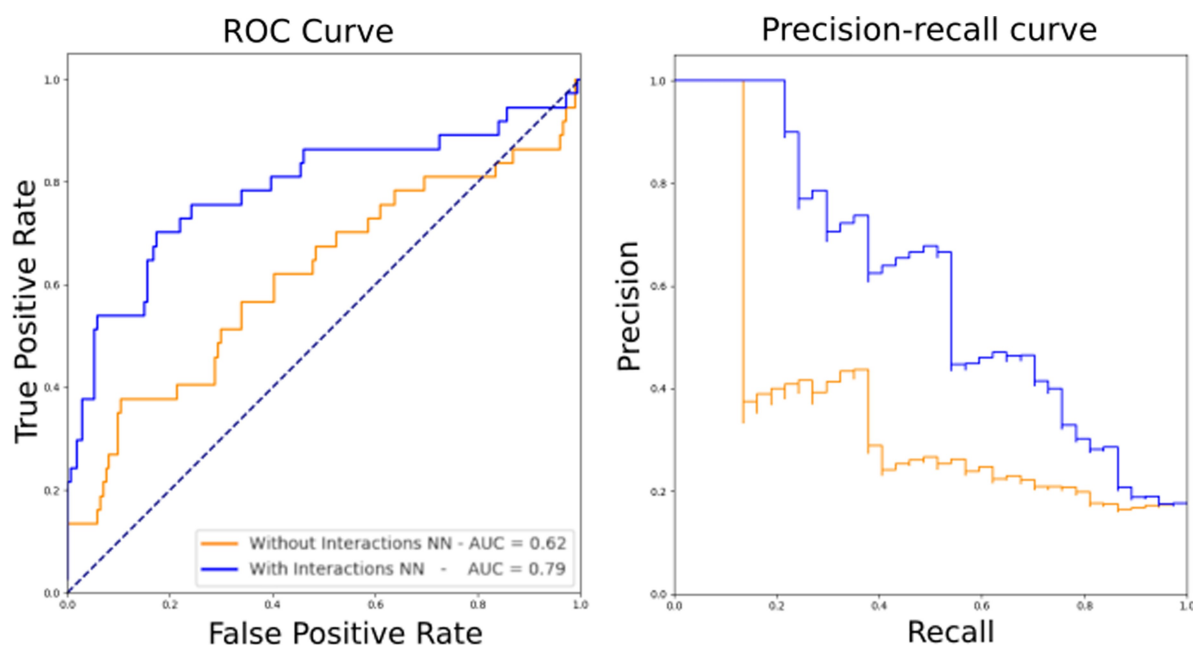


Figure 1. Performance of Xenusia. The tool's performance on the validation dataset, composed of archaeal proteins only, is compared for a neural network without (orange) and with (blue) the inclusion of an intermediate prediction on interaction residues. Left, ROC curve; right, precision–recall curve.

For OLS17986, probe 1 corresponds to the promoter region of a gene annotated as a putative citramalate synthase (cimA-1), an enzyme involved in the L-isoleucine biosynthesis pathway. Probe 2 corresponds to the promoter region of OLS17986 itself, allowing the exploration of potential autoregulation. Probe 3 contains the promoter region of an open reading frame housing a molybdenum cofactor guanylyl-transferase (mobA), a coenzyme F420: L-glutamate ligase, and a phosphoglycerate mutase.

For OLS31011, probe 1 includes the promoter region of OLS31011 and of a rhaT/eamA-type transporter. The other two probes contain promoter regions of hypothetical genes of unknown function.

For OLS23561, probe 1 includes the promoter of a putative alcohol dehydrogenase gene. Probe 2 contains the promoters of OLS23561 itself and a putative stress response protein. Probe 3 includes the promoter of a putative methyltransferase gene.

To perform the assay, reaction mixtures containing ^{32}P -labeled dsDNA probe: protein combinations (20 000 cps: 11.24, 5.20, and 64.22 μM for OLS17986, OLS31011, and OLS23561, respectively) and salmon sperm competitor DNA (10 mg/mL) were incubated for 25 min at 37°C. The mixtures were then separated onto an 8% acrylamide gel (180 V for 10 min followed by 130 V for 2 h). After the run, the gel was exposed overnight to an autoradiography gel and developed the next day.

3 Results

3.1 Computational validation

The performance of NN1, identifying DNA-binding residues, was validated by a five-fold cross-validation on the contact dataset (see Methods section), resulting in an area under the ROC curve (AUC) of 0.75, with an average precision (AP) of 0.245. Given the low performance, particularly in terms of the AP, this model cannot provide reliable information to experimentalists that are interested in, e.g. mutating potential

Table 1. Performance comparison of Xenusia with state-of-the-art methods for identifying DNA-binding proteins.^a

Method	Year	Acc	Sen	Spe	AUC	MCC
Xenusia	2025	77.73	70.27	79.31	0.79	0.414
IDNA-Prot	2011	76.08	60.00	79.31	–	0.330
DNA binder	2007	63.51	56.76	64.94	–	0.169
PseDNA-Pro	2015	64.93	72.97	63.22	–	0.278
Local DPP	2017	70.62	27.03	79.88	0.59	0.064

^a Since Xenusia outputs a probability score, a threshold was chosen for binarization, allowing comparison with state-of-the-art binary predictors. The threshold was set to equal the specificity value of the best performing state-of-the-art method in our validation procedure (IDNA-Prot).

DNA-binding residues. However, once NN1 is connected to NN2, the performances of Xenusia are significantly boosted. Figure 1 shows the increased performances in predicting archaeal DNA-binding domains (NN2) when including DNA-binding residue information (NN1). This highlights the importance of integrating information from various sources for reliable predictions.

The final performances of Xenusia result in an AUC of 0.79 and 0.60 for the ROC and precision–recall curve, respectively. In Table 1, the performance of Xenusia is compared to state-of-the-art methods. Since our tool is designed to prioritize archaeal DNA-binding proteins in large-scale proteome studies, it is essential that it provides a probability score rather than a simple binary classification. A probability output allows researchers to rank proteins by confidence, making it more useful for real-case applications. However, to ensure a fair comparison with existing binary classifiers, we set a threshold for our tool that matches the specificity of the best-performing state-of-the-art method. This approach was also applied to Local-DPP, as it is the only other predictor that provides probability scores.

Table 1 shows that Xenusia provides the best predictions compared to all other available methods. The Matthew's correlation coefficient (MCC), recognized as the best measure to

summarize a confusion matrix, surpasses the second best predictor, IDNA-Prot with 0.330, by 25%. It is important to note that most available methods lack quality scoring information, making them unsuitable for proteome-wide functional studies based on experiments.

3.2 Large scale analysis of Asgard metagenomics

Asgard is a proposed superphylum of archaea that has only recently been discovered (Zaremba-Niedzwiedzka *et al.* 2017). These uncultivated archaea show intermediate characteristics between eukaryotes and prokaryotes. They have been found in samples coming from the Arctic sea (Jorgensen *et al.* 2012) and other extreme environments, such as acid hot-springs (Zaremba-Niedzwiedzka *et al.* 2017). Homology analyses performed on a subclass of the Lokiarchaeota superphylum revealed that the protein phylogeny of these organisms is very uncommon: about one-third of the proteins is completely new, while two-thirds could be linked to other archaeal, bacterial, or eukaryotic proteins (Spang *et al.* 2015). This portion of completely new proteins is enormous, making standard homology-based computational annotation tools unsuitable for the analysis of this taxonomic group. Since these organisms live in the Arctic sea or hot springs, their proteins have likely evolved to maximize functionality in extreme environments. This means that functional annotation of their proteins could offer new insights into the relationship between sequence and structure. We ran Xenusia on a dataset consisting of 16 946 Asgard proteins. The 10 best scoring proteins are reported in Supplementary Table S1. The first hit, OLS30781, is a homologue of a bacterial metalloprotease. This class includes enzymes with various functions, i.e. DNA repair (Maté and Kleanthous 2004). Interestingly, one of their distant homologues was crystallized in a complex with DNA (PDB ID: 1V14). Additionally, it is worth mentioning that ABC transporters are also commonly linked to DNA repair processes (Davidson *et al.* 2008). The second and third hit, OLS30429 and OLS21705, respectively, are characterized proteins that are annotated as

helix-turn-helix (HTH)-motif transcription regulators (Brinkman *et al.* 2000, 2002). These two proteins have relatively distant homologues, sharing only 46% and 37% sequence identities with their closest nonhypothetical relatives. The fifth and sixth hits, OLS17998 and OLS16705, respectively, are also homologues of archaeal HTH transcription factors. Regarding the protein homologue of the seventh hit, OLS19303 or also known as the response regulator SaeR, several studies have been performed on the equivalent protein of *Staphylococcus aureus*, defining the presence of a DNA-binding domain that acts as transcription regulator (Steinhuber *et al.* 2003). Interestingly, the ninth highest-scoring protein, OLS30097, is an unannotated sequence, lacking any known annotated homologues. Despite this, the AlphaFold predicted structure reveals a distinct HTH structure. In Fig. 2, a comparison between this model and the crystal structure of an archaeal HTH DNA-binding protein (5 box, chain A: 1:111) is shown. Notably, the unannotated Asgard protein displays a reversed topology compared to the crystal structure: the former begins with a globular domain and concludes with a long alpha helix, while the latter starts with a long alpha helix and finishes with the globular domain. This suggests independent, convergent evolution. It also clarifies why standard alignment algorithms and even structural alignment tools struggle to detect similarities between these two proteins. Supplementary Figure S1 shows the result of the structural alignment of the two proteins. It is important to note that in this case AlphaFold was used solely to obtain a structure for a protein with no experimentally solved structure. The Xenusia pipeline does not include AlphaFold and does not rely on homology detection or multiple sequence alignments for its predictions.

3.3 Experimental characterization of Asgard DNA-binding proteins

As mentioned earlier, working with Asgard proteins is often challenging due to weak homology links with proteins of known functions. Consequently, transferring annotations can

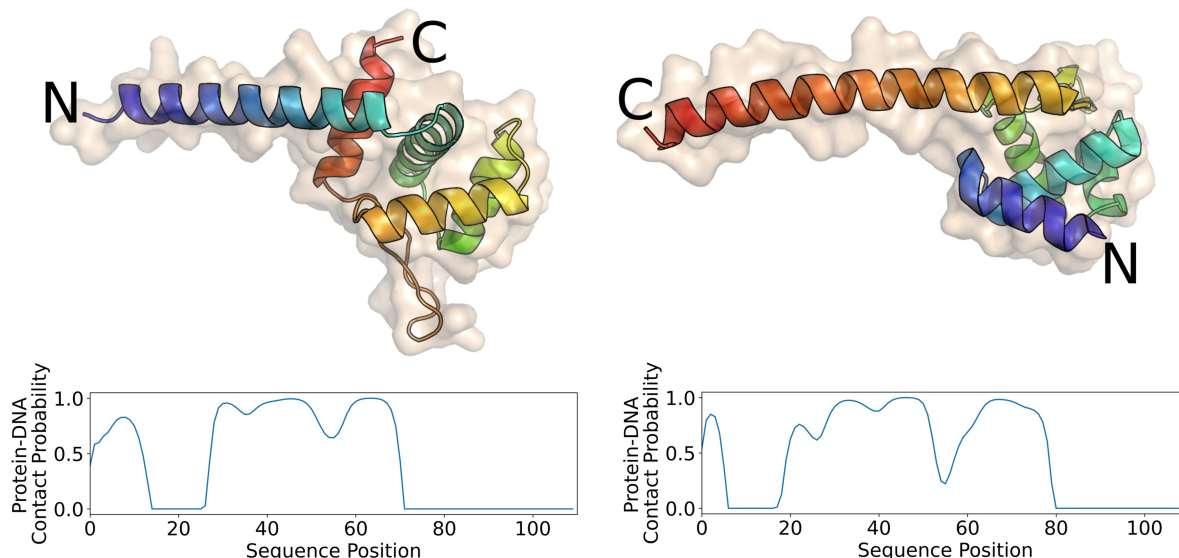


Figure 2. Comparison between an archaeal HTH domain (5 box chain A: 1–111) and an unannotated Asgard sequence. The image displays a crystal structure of an archaeal HTH DNA-binding protein (left) and the AlphaFold structure of an unannotated Asgard protein (OLS30097, right). Xenusia predicts the latter with high confidence to be a DNA-binding protein. The color gradient represents the amino acid positions in the sequence, transitioning from blue at the beginning (N-terminal) to red at the end (C-terminal). Although the two proteins share a remarkably similar structure, their orientation is reversed, indicating a potential result of convergent evolution. The plots below each structure represent the predicted probability of each residue to interact with DNA.

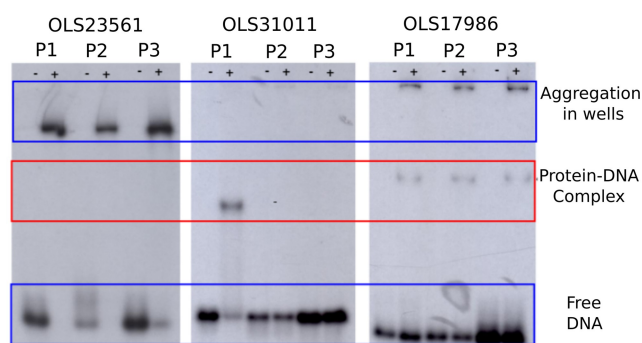


Figure 3. Experimental characterization of three Asgard DNA-binding proteins. The interaction of the three Asgard proteins (OLS23561, OLS31011, and OLS17986) with each three labeled dsDNA probes (P1, P2, and P3) was assessed via gel electrophoresis followed by autoradiography. Wells indicated with a “+” contain protein and DNA, while those indicated with a “-” are the negative control and only contain the DNA probe. OLS23561 does not show monomeric protein–DNA complex formation, only aggregated complexes are detected. OLS31011 only forms a complex with P1, while OLS17986 forms a complex with all three probes. The red box indicates where the band of the monomeric protein–DNA complex should appear.

be risky. To address this, we chose three Asgard proteins (OLS23561, OLS17986, and OLS31011) to experimentally investigate.

The proteins were selected from the top 50 highest-ranked candidates in the Asgard superphylum metagenomics dataset. Among these, three proteins were manually chosen based on their weak sequence similarity to known DNA-binding proteins. The selected proteins exhibit sequence identities of 41%, 43%, and 43%, respectively, with previously characterized DNA-binding proteins. The homologs were not included in our training dataset.

These proteins are confidently predicted by Xenusia to be DNA-binding proteins. We then conducted experimental investigations on their capability of interacting with DNA. [Supplementary Table S2](#) provides details about these selected proteins.

The three proteins were produced in *E. coli* and purified via His-tag affinity chromatography. The capability of these proteins to bind DNA was evaluated with gel electrophoresis followed by autoradiograph, shown in [Fig. 3](#). For each protein, we tested three different ^{32}P -labeled dsDNA probes, named P1, P2, and P3 (see [Supplementary Table S3](#)). These probes consist of the promoter regions of the genes encoding the putative DNA-binding proteins and the promoter regions of genes in the proximity of the target gene (see Methods section DNA-binding assay). We made this choice because in archaea, transcription factors frequently influence genes that are proximate in the sequence space. This assay should provide a first estimation of the binding specificity, as it reveals the migration speed of the protein–dsDNA mixture with respect to the free dsDNA probe. Upon binding, the protein–dsDNA complex will be heavier and therefore migrate slower. [Figure 3](#) clearly shows that OLS17986 binds to all three probes, both in its monomeric form as in an aggregated form. OLS31011 forms only a monomeric complex with its P1 probe. The concentration of OLS23561 appeared to be too high as only aggregated probe–protein complexes could be observed. Therefore, a crescent protein concentration of OLS23561 was assessed to test for monomeric interaction to the probes, as was observed for probes 1 and 2 (see [Supplementary Fig. S2](#)).

3.4 Limitations of the experimental validation

The experimental validation performed in this study had two main objectives. The first objective was to evaluate the effectiveness of our experimental protocol in the context of archaeal metagenomics. This approach is based on the assumption that DNA-binding proteins will interact with promoters of nearby genes, reflecting a localized effect when the binding is specific. The second objective was to assess Xenusia’s ability to identify archaeal DNA-binding proteins.

To maximize the likelihood of identifying actual DNA-binding proteins and to test the protocol effectively, we selected only proteins that were scored highly by Xenusia and exhibited some sequence similarity to known DNA-binding proteins. All tested proteins yielded positive results, supporting both the protocol and Xenusia’s predictions. However, a notable limitation of this validation is the absence of experimental testing on sequences that are completely evolutionarily uncorrelated. Additional experimental studies will be necessary to fully evaluate Xenusia’s performance, as well as that of other state-of-the-art tools, on entirely novel proteins.

To facilitate further research, we have provided a ranked list of Asgard proteins in the Git repository, sorted by their predicted likelihood of being DNA-binding proteins. This resource allows other researchers to expand the experimental validation efforts presented in this article.

4 Conclusions

In this paper, we introduce Xenusia, a neural network-based method for predicting DNA-binding proteins in archaea, an understudied domain of life. Xenusia addresses key limitations of existing tools, offering significant improvements in both performance and usability.

The architecture of Xenusia is specifically designed to predict DNA-binding proteins while handling the limited amount of available data. This tailored approach results in a performance improvement of approximately 25% in MCC compared to the best state-of-the-art methods. Additionally, in contrast to most of the state-of-the-art methods, Xenusia offers more than a simple binary classification by providing a probability score for each protein’s likelihood of binding DNA. This probabilistic output is particularly valuable in real-world applications, such as ranking proteins from entire proteomes for further experimental validation.

Xenusia is intended to assist researchers in focusing their experimental efforts, which are often time- and resource-intensive, on a smaller subset of proteins that are more likely to exhibit DNA-binding activity. To make the tool accessible to a broad range of users, including those with limited computational expertise, we provide two deployment options. First, Xenusia is available as a pip-installable package, allowing local installation with a single line of code. Second, we offer a Google Colab web application with a user-friendly graphical interface, enabling predictions to be performed online without requiring any local setup or file downloads.

Acknowledgments

G.O. is grateful to R. Gilbert and R. Sanchez for inspiration.

Author contributions

Linus Donvil (Data curation [equal], Methodology [supporting]), Joëlle A.J. Housmans (Writing—original draft [equal], Writing—review & editing [equal]), Eveline Peeters (Conceptualization [equal], Funding acquisition [lead], Methodology [lead], Supervision [lead]), Gabriele Orlando (Conceptualization [equal], Data curation [lead], Formal analysis [lead], Investigation [lead], Methodology [lead], Project administration [lead], Software [lead], Supervision [supporting], Validation [lead], Writing—original draft [lead], Writing—review & editing [lead]), and Wim F. Vranken (Conceptualization [equal], Funding acquisition [lead], Methodology [supporting])

Supplementary data

[Supplementary data](#) are available at *Bioinformatics* online.

Conflict of interest: None declared.

Funding

This research was funded by the Vrije Universiteit Brussel (Strategic Research Program SRP91) and the Research Foundation Flanders (FWO)—project no. G.0328.16N. G.O. is funded by the University of Montpellier and by an ANR project (project no. ANR-23-CPJ1-0061).

Data availability

All the data related to this article is available at <https://github.com/grogdrinker/xenusia>

References

- Aravind L, Koonin EV. DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res* 1999; 27:4658–70.
- Brinkman AB, Bell SD, Lebbink RJ *et al.* The *Sulfolobus solfataricus* Lrp-like protein LysM regulates lysine biosynthesis in response to lysine availability. *J Biol Chem* 2002;277:29537–49.
- Brinkman AB, Dahlke I, Tuininga JE *et al.* An Lrp-like transcriptional regulator from the archaeon *Pyrococcus furiosus* is negatively autoregulated. *J Biol Chem* 2000;275:38160–9.
- Cilia E, Pancsa R, Tompa P *et al.* The DynaMine webserver: predicting protein dynamics from sequence. *Nucleic Acids Res* 2014; 42:W264–70.
- Coudert E, Gehant S, de Castro E *et al.* UniProt Consortium. Annotation of biologically relevant ligands in UniProtKB using ChEBI. *Bioinformatics* 2023;39:btac793. <https://doi.org/10.1093/bioinformatics/btac793>
- Davidson AL, Dassa E, Orelle C *et al.* Structure, function, and evolution of bacterial ATP-binding cassette systems. *Microbiol Mol Biol Rev* 2008;72:317–64.
- Emerson JB, Thomas BC, Alvarez W *et al.* Metagenomic analysis of a high carbon dioxide subsurface microbial community populated by chemolithoautotrophs and bacteria and archaea from candidate phyla. *Environ Microbiol* 2016;18:1686–703.
- Hudson WH, Orlund EA. The structure, function and evolution of proteins that bind DNA and RNA. *Nat Rev Mol Cell Biol* 2014; 15:749–60.
- Iovine B, Iannella ML, Bevilacqua MA. Damage-specific DNA binding protein 1 (DDB1): a protein with a wide range of functions. *Int J Biochem Cell Biol* 2011;43:1664–7.
- Jorgensen SL, Hannisdal B, Lanzén A *et al.* Correlating microbial community profiles with geochemical data in highly stratified sediments from the Arctic Mid-Ocean ridge. *Proc Natl Acad Sci U S A* 2012; 109:E2846–55.
- Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;596:583–9.
- Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006; 22:1658–9.
- Maté MJ, Kleantous C. Structure-based analysis of the metal-dependent mechanism of H-N-H endonucleases. *J Biol Chem* 2004; 279:34763–9.
- Meng Q, Guo F, Tang J. Improved structure-related prediction for insufficient homologous proteins using MSA enhancement and pre-trained language model. *Brief Bioinform* 2023;24:bbad217. <https://doi.org/10.1093/bib/bbad217>
- Orlando G, Raimondi D, Codicè F *et al.* Prediction of disordered regions in proteins with recurrent neural networks and protein dynamics. *J Mol Biol* 2022;434:167579.
- Orlando G, Raimondi D, Khan T *et al.* SVM-dependent pairwise HMM: an application to protein pairwise alignments. *Bioinformatics* 2017;33:3902–8.
- Orlando G, Raimondi D, Vranken WF. Observation selection bias in contact prediction and its implications for structural bioinformatics. *Sci Rep* 2016;6:36679.
- Parks DH, Rinke C, Chuvochina M *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol* 2017;2:1533–42.
- Raimondi D, Orlando G, Pancsa R *et al.* Exploring the sequence-based prediction of folding initiation sites in proteins. *Sci Rep* 2017; 7:8826.
- Spang A, Saw JH, Jørgensen SL *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* 2015;521:173–9.
- Steinhuber A, Goerke C, Bayer MG *et al.* Molecular architecture of the regulatory locus SAE of *Staphylococcus aureus* and its impact on expression of virulence factors. *J Bacteriol* 2003;185:6278–86.
- Zaman R, Chowdhury SY, Rashid MA *et al.* HMMBinder: DNA-binding protein prediction using HMM profile based features. *Biomed Res Int* 2017;2017:4590609.
- Zaremba-Niedzwiedzka K, Caceres EF, Saw JH *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* 2017; 541:353–8.