pharmaceuticals

Article

Rethinking Protein Drug Design with Highly Accurate Structure Prediction of Anti-CRISPR Proteins

Ho-Min Park [1,2], Yunseol Park [1], Joris Vankerschaver [1,3], Arnout Van Messem [4], Wesley De Neve [1,2] and Hyunjin Shim [1,*]

[1] Center for Biosystems and Biotech Data Science, Ghent University Global Campus, Incheon 21985, Korea; homin.park@ghent.ac.kr (H.-M.P.); yunseol.park@ghent.ac.kr (Y.P.); joris.vankerschaver@ghent.ac.kr (J.V.); wesley.deneve@ghent.ac.kr (W.D.N.)
[2] Department of Electronics and Information Systems, Ghent University, 9000 Ghent, Belgium
[3] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, 9000 Ghent, Belgium
[4] Department of Mathematics, University of Liège, 4000 Liège, Belgium; arnout.vanmessem@uliege.be
[*] Correspondence: hyunjin.shim@ghent.ac.kr

Abstract: Protein therapeutics play an important role in controlling the functions and activities of disease-causing proteins in modern medicine. Despite protein therapeutics having several advantages over traditional small-molecule therapeutics, further development has been hindered by drug complexity and delivery issues. However, recent progress in deep learning-based protein structure prediction approaches, such as AlphaFold2, opens new opportunities to exploit the complexity of these macro-biomolecules for highly specialised design to inhibit, regulate or even manipulate specific disease-causing proteins. Anti-CRISPR proteins are small proteins from bacteriophages that counter-defend against the prokaryotic adaptive immunity of CRISPR-Cas systems. They are unique examples of natural protein therapeutics that have been optimized by the host-parasite evolutionary arms race to inhibit a wide variety of host proteins. Here, we show that these anti-CRISPR proteins display diverse inhibition mechanisms through accurate structural prediction and functional analysis. We find that these phage-derived proteins are extremely distinct in structure, some of which have no homologues in the current protein structure domain. Furthermore, we find a novel family of anti-CRISPR proteins which are structurally similar to the recently discovered mechanism of manipulating host proteins through enzymatic activity, rather than through direct inference. Using highly accurate structure prediction, we present a wide variety of protein-manipulating strategies of anti-CRISPR proteins for future protein drug design.

Keywords: in-silico drug design; AlphaFold; anti-CRISPR proteins; prokaryotic defence mechanisms; bacteriophages; structural biology; protein drug

check for updates

Citation: Park, H.-M.; Park, Y.; Vankerschaver, J.; Van Messem, A.; De Neve, W.; Shim, H. Rethinking Protein Drug Design with Highly Accurate Structure Prediction of Anti-CRISPR Proteins. *Pharmaceuticals* **2022**, *15*, 310. https://doi.org/10.3390/ph15030310

Academic Editor: Osvaldo Andrade Santos-Filho

Received: 27 January 2022
Accepted: 1 March 2022
Published: 4 March 2022

1. Introduction

Proteins are macromolecules composed of amino-acid residues that perform diverse roles in biological entities, including catalysing biochemical reactions, providing cell/capsid structure, transporting molecules, replicating genetic material, and responding to stimuli. It is estimated there are over 25,000 functionally distinct proteins in the human body [1,2], and mutations or abnormalities in these proteins may result in diseases. Thus, modern medicine has focused on targeting such proteins to alleviate diseases, mostly through small-molecule therapeutic agents acting as competitive or non-competitive inhibitors [3]. However, it is estimated that only ~10% of the human proteome can be targeted with small-molecule drugs [3]. Since the introduction of human insulin as the first recombinant protein therapeutic in the 1980s [4,5], protein-based therapeutics have expanded the scope of "druggable proteins". Compared to small-molecule drugs, the major advantage of protein therapeutics is improved target specificity and reduced immunogenicity due to their proteinaceous nature [5]. Protein therapeutics can also serve complex functions that simple chemical compounds cannot achieve, such as replacing a deficient protein or providing a protein of novel function (Figure 1a). Furthermore, protein therapeutics can inhibit disease-related proteins that small-molecule drugs cannot target due to the lack of a cavity to bind. Currently, there are over 130 protein therapeutics commercially available and intense research efforts are ongoing to better design protein therapeutics [5].

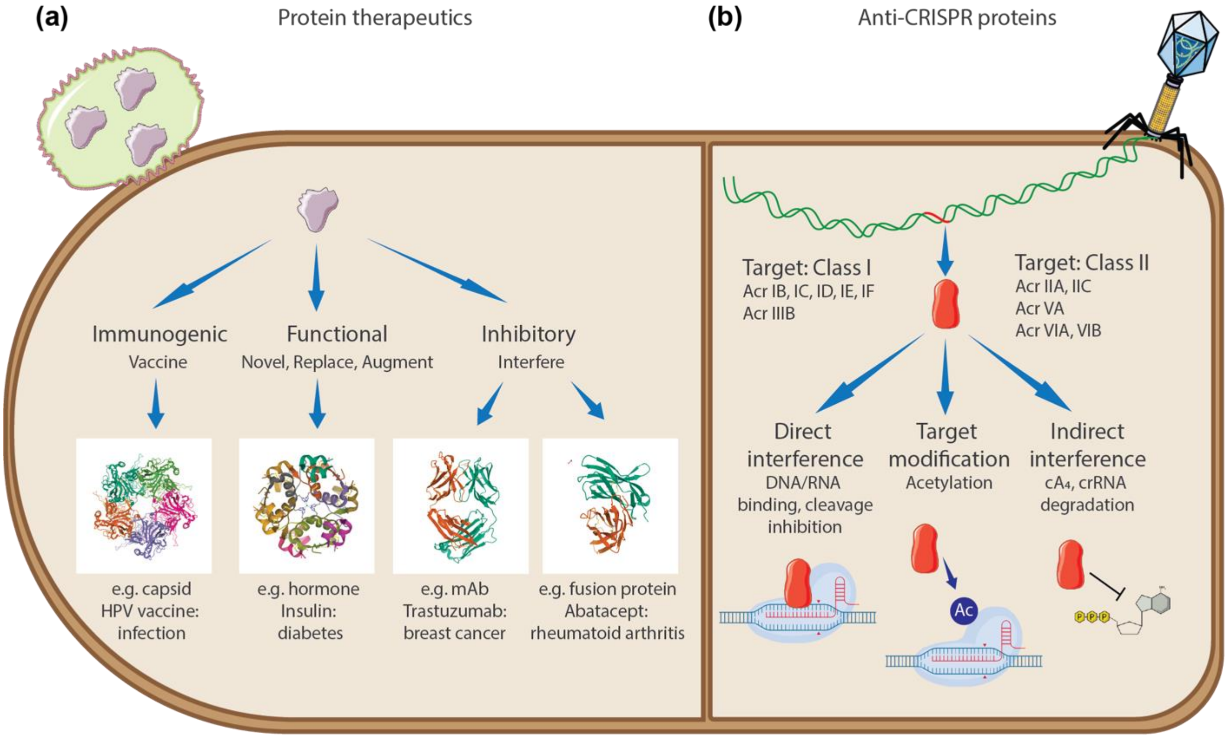

Figure 1. Mechanisms of protein therapeutics and anti-CRISPR proteins. (**a**) Mechanism of protein therapeutics. The first group consists of prophylactic or therapeutic vaccines that induce immunity against foreign or cancer cells. A highly successful protein vaccine against human papillomavirus (HPV) combining the capsids from four pathogenic HPV strains is given as an example (PDB: 2R5K [6]). The second group consists of protein therapeutics that provide novel functions, replace deficient or abnormal proteins, or augment existing activities. The approval of recombinant insulin in the 1980s to treat diabetes as the first abundant, inexpensive and low immunogenic therapeutic protein is given as an example (PDB: 4F8F [7]). The third group consists of proteins that interfere with target proteins through high binding specificity. Some monoclonal antibodies use antigen recognition sites or receptor-binding domains like Trastuzumab against breast cancer cells (PDB: 6MH2 [8,9]). Some fusion proteins inhibit target proteins by blocking interaction sites like Abatacept against rheumatoid arthritis (PDB: 1DQT [8,9]). (**b**) Mechanisms of anti-CRISPR proteins. Upon successful infection, phage genomes express anti-CRISPR proteins that neutralize host CRISPR-Cas immunity. Anti-CRISPR proteins target various types of both Class I and Class II CRISPR-Cas systems and the inhibitory mechanisms are highly diverse, including direct interference of DNA/RNA binding and cleavage of Cas complexes, enzymatic inhibition of the active site by acetylation, and nuclease activity of degrading the signalling molecule (cyclic nucleotide cA_4).

In this study, we present a group of naturally-occurring protein therapeutics, called anti-CRISPR (Acr) proteins, as a good example of how small proteins are used by invading bacteriophages (phages) in nature to control host proteins. Phages are the most abundant and diverse biological entities in the biosphere (estimated 10^{31} existing phages) that infect and replicate within host prokaryotes (such as bacteria or archaea) [10]. High selective pressures between these parasites and hosts drive dynamic coevolution of genomic and proteomic mechanisms and systems [11–13]. In particular, the evolutionary arms race between phages and prokaryotes has resulted in a vast arsenal of immune systems, including the prokaryotic adaptive immune system known as CRISPR-Cas [14,15]. CRISPR-Cas systems are defence mechanisms against phages (and other mobilomes) through a complex of RNA-guided Cas proteins (Figure 1b). Remarkably, prokaryotic genomes with CRISPR-Cas systems can acquire short fragments of foreign genetic sequences in their CRISPR arrays, which serve as RNA templates to recognize and cleave invading phages through the nuclease complex of Cas proteins [16]. Since the successful application of CRISPR-Cas systems as genome-editing tools [17,18], there has been a burst in the discovery of diverse CRISPR-Cas systems [19], followed by the discovery of Acr proteins that neutralize the activity of this prokaryotic adaptive immune system [20] (Figure 1b). A family of Acr proteins was first identified in the CRISPR-Cas-inactivating prophages of *Pseudomonas* genomes that disable Type I-F and Type I-E CRISPR-Cas systems [20,21]. A number of Acr proteins inhibiting type II CRISPR-Cas systems have since been applied as regulators of gene-editing activities [22]. The Acr protein families are known to have short sequences (<100 amino acids) with no common genetic features, and interact directly with Cas proteins to inhibit target DNA binding, DNA cleavage, CRISPR RNA loading and protein-complex formation [22,23]. A recent study reveals that AcrVA5 proteins inactivate Cas12a of Type V CRISPR-Cas systems enzymatically by acetylation of the active site, with structural similarity to an acetyltransferase protein [24].

In this study, we conducted a comprehensive analysis on the key characteristics of Acr proteins viewed from the perspective of naturally-occurring protein therapeutics that effectively inhibit host protein functions. Motivated by the observation that these Acr proteins are genetically diverse, we examined the protein structure of these diverse proteins using AlphaFold2 [25]. AlphaFold is a state-of-the-art deep learning-based approach that performs protein structure prediction, which takes a protein sequence as an input to predict its 3-D protein structure through an iterative exchange of information between its genetic representation and its structural representation. The recent release of AlphaFold2, the winner of CASP14, which achieves highly accurate protein structure predictions [25,26], is revolutionary for the field of life sciences and medicine, and is expected to accelerate critical research in a large number of fields ranging from structural biology to drug discovery. In this study, we first assessed the performance of AlphaFold2 in predicting the 3-D structures of Acr proteins based on similarity metrics against their experimentally reconstructed 3-D macromolecular structures. Using this performance as a basis, we further examined the Acr proteins without experimental structures with AlphaFold2, to predict the structural diversity of these genetically distinct proteins that are natural inhibitory proteins against prokaryotic CRISPR-Cas systems. We used AlphaFold-predicted structures of Acr proteins to infer a range of inhibition mechanisms through homology search and functional analysis, to demonstrate how bacteriophages exploit diverse strategies to manipulate host immune systems, with the long-term goal of providing a unique opportunity to learn from the evolution-optimized inhibitor proteins for future protein drug design.

2. Results

2.1. AlphaFold2 Prediction of Anti-CRISPR Protein Structures

The Acr protein datasets were acquired from various viral and prokaryotic genomes, including *Pseudomonas phage*, *Pseudomonas aeruginosa* and *Escherichia coli* [27], which were categorised into three sets: verified Acr proteins with experimental structure (Set A), verified Acr proteins without experimental structure (Set B), and putative Acr proteins with experimental structure (Set C) (Supplementary Table S1). From the AlphaFold-predicted structures of each set (Figure 2a), we compared the prediction performance between CASP14 (52 AlphaFold2 evaluation results from the CASP14 competition), Set A and Set C (Figure 2b,c), based on TM-scores, relative Z-errors (see Section 4 for details) and root mean square deviation (RMSD) (Supplementary Figure S1) against the true experimental structures (Supplementary Table S2), where Set B without experimental structures was excluded. According to the TM-score, CASP14 had a higher median than Set A and Set C

(0.925 vs. 0.895; 0.843, respectively), but Set A had the highest mean as compared to Set C and CASP14 (0.896 vs. 0.868; 0.882, respectively). Furthermore, Set A and Set C had significantly smaller standard deviations than CASP14 (0.095; 0.092, respectively, vs. 0.120), indicating that, according to the TM-score, their predictions are more accurate. According to the relative Z-error, CASP14 recorded a lower median than Set A and Set C (0.201 vs. 0.217; 0.230, respectively), but Set A recorded the lowest mean as compared to CASP14 and Set C (0.211 vs. 0.237; 0.259, respectively). Like the TM-score, Set A and Set C recorded smaller standard deviations than CASP14 (0.128; 0.129 vs. 0.141, respectively). For RMSD, Set A and Set C have more outliers and higher mean values than CASP 14 (2.206; 2.271 vs. 1.106, respectively). As can be seen in the boxplots (Figure 2b,c and Figure S1), there was no significant difference in prediction performance, validating that AlphaFold2 predicts 3-D structures of Acr proteins as accurately as the CASP14 dataset. Previous studies demonstrated that structures of identical proteins obtained from different experimental techniques had RMSD values of around 2.3 Å [28]. Thus, the average result of 2.271 Å in the RMSD distribution in Set C further validates the prediction performance of AlphaFold2 on Acr proteins.

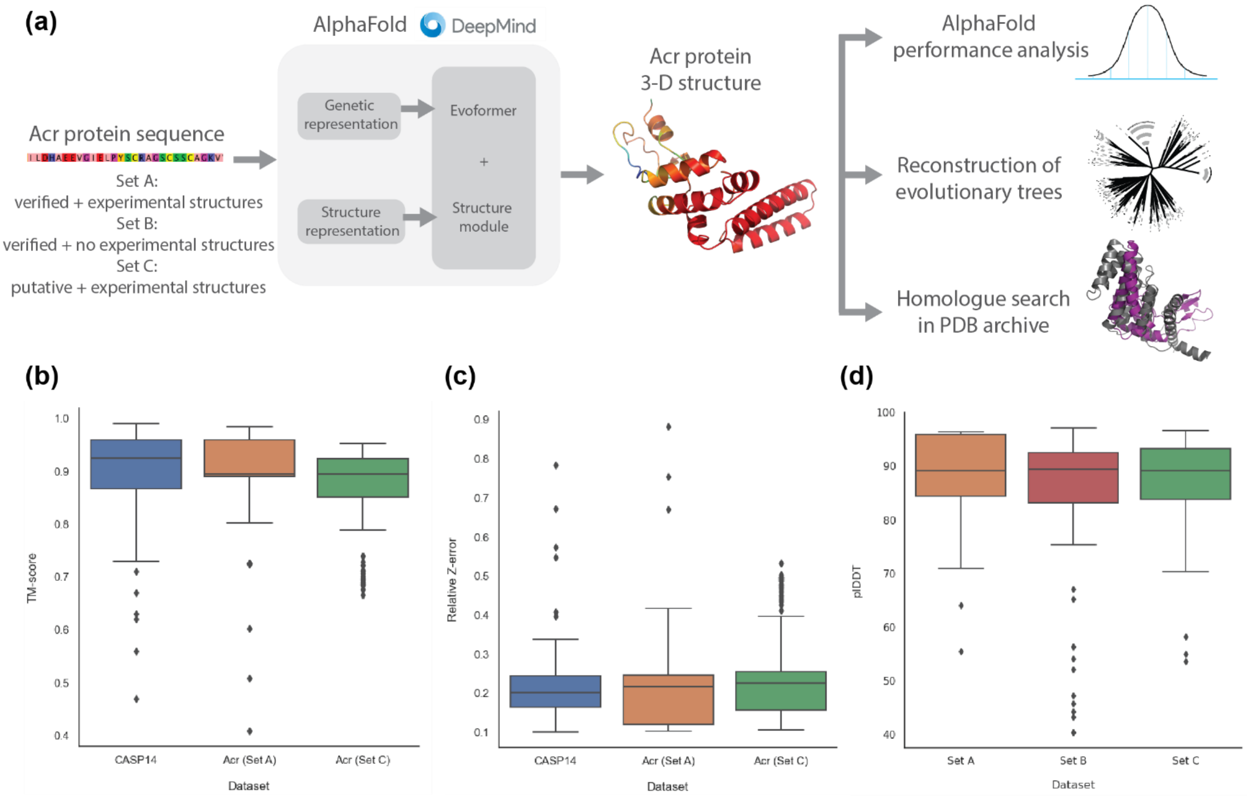

Figure 2. Performance analysis of AlphaFold2 on anti-CRISPR proteins in comparison to the CASP14 dataset. (**a**) Overall workflow to analyse the 3-D macromolecular structures of Acr protein sequences predicted with AlphaFold2. (**b**) The performance of AlphaFold2 on the Acr protein datasets in comparison to the CASP14 dataset using TM-scores. The closer the TM-score is to 1, the more similar the predicted structure is to its true experimental structure. (**c**) The performance of AlphaFold2 on the Acr protein datasets in comparison to the CASP14 dataset using relative Z-errors. The closer the relative Z-error is to 0, the more similar the predicted structure is to its true experimental structure. (**d**) The performance of AlphaFold2 on the Acr protein datasets using plDDT. The closer the plDDT is to 100, the higher the confidence level of prediction by AlphaFold2. (Set A: Verified Acr proteins with experimental structures, Set B: Verified Acr proteins without experimental structures, Set C: Putative Acr proteins with experimental structures).

For Set B without experimental structures, we calculated the predicted local distance difference test (plDDT) to check the confidence level of AlphaFold2 (Figure 2d). The median plDDT of Set B was almost identical to that of Set A and Set C (89.4 vs. 89.2; 89.1, respectively), but Set B had a slightly higher standard deviation (13.48 vs. 8.4; 7.84, respectively). However, as all the plDDT values of Set A, Set B and Set C (excluding a few outliers) were above the lower cut-off value of AlphaFold2 given as 70 [29], we concluded that the prediction quality of Set B was not significantly different from that of Set A and Set C.

2.2. Evolutionary Trees of Anti-CRISPR Proteins

The Acr proteins are known to be genetically diverse; this raises an intriguing question about the origin and evolution of Acr proteins. We reconstructed evolutionary trees of the verified Acr proteins (Set A + Set B; n = 207), using sequence-based and structure-based methods (see Section 4 for details). As expected, the phylogenetic tree built using genetic sequences (Figure 3a) shows high levels of variation, while forming consistent clades with the high bootstrap support. Analysis of the clades reveals some degree of clustering by the Acr family at the shallower nodes; however, this clustering is mostly due to the near-identical protein sequences. For instance, many of the AcrIIA proteins were derived from AcrIIA1 and AcrIIA2, driven by the technological interest to regulate CRISPR-Cas gene-editing activities in different cell types [22]. Otherwise, the phylogenetic tree shows absence of clustering by other biological features such as taxonomy and inhibition mechanism.

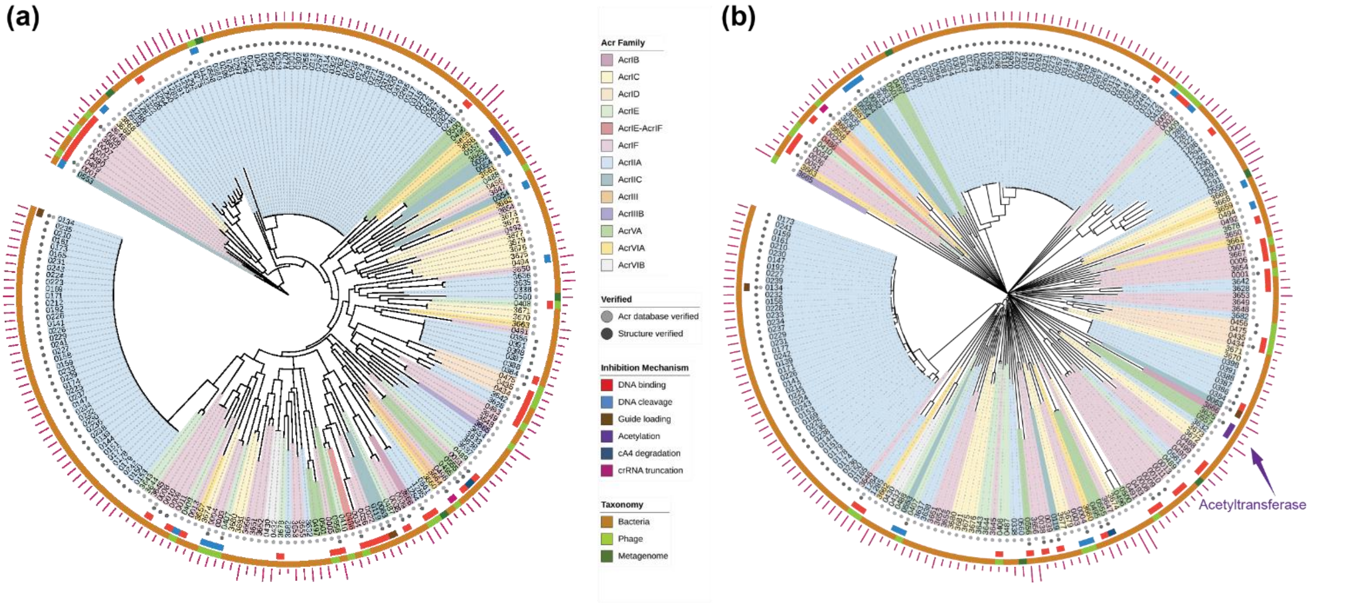

Figure 3. Evolutionary trees of anti-CRISPR proteins (numbered according to Anti-CRISPRdb [27]). (**a**) The phylogenetic tree of anti-CRISPR proteins reconstructed using sequence-based methods (Set A + Set B; n = 207). (**b**) The structural tree of anti-CRISPR proteins reconstructed using structure-based methods (Set A + Set B; n = 207). The clades of the two evolutionary trees were coloured by the Acr Family. The inner to outer rings display the Acr verification status, structural verification status, inhibition mechanism and source organism taxonomy. The outer magenta bars represent the genetic sequence length of each protein.

Given the low sequence similarity among the Acr proteins, we built a structure-based tree using AlphaFold-predicted structures, which included 98 Acr proteins without experimental structures (Figure 3b). The structural tree showed an even higher level of diversity in the Acr proteins than the phylogenetic tree. In the structural tree, the Acr proteins share no common ancestor and display deep branches, consistent with earlier observations of how evolutionary pressure drives immunity-related mechanisms of hosts and parasites to coevolve rapidly [12,13]. The structural tree also shows some degrees of clustering by the Acr family, but the clusters do not always coincide between the two evolutionary trees. The visual analysis of the protein structures show that the branches of the structural tree are placed randomly in terms of representative structural forms and the functions are only related at the clade level (Supplementary Figure S2).

It is evident that the sequence-based and structure-based trees capture different evolutionary relationships between the Acr proteins. The 3-D structures of homologous proteins were previously shown to be better conserved than their corresponding genetic sequences, particularly when the sequence similarity was below 30% [30]. From the multiple sequence alignment, no site of the Acr proteins was conserved at 30% and only very few sites were conserved at 15% (Supplementary Figures S3 and S4). We calculated the congruence among distance matrices of the sequence-based and structure-based trees to be very low according to the measure of congruence (Kendall's coefficient of concordance, $W = 6.58 \times 10^{-1}$), confirming the correlation between these two types of evolutionary trees is poor among highly divergent proteins [30].

2.3. Structural Homology to Predicted Anti-CRISPR Structures

The characterised Acr proteins use diverse strategies to interfere directly with CRISPR-Cas systems, including inhibiting DNA binding, DNA cleavage, guide loading and ribonuclease activity [22]. To investigate the relation between the 3-D protein structures and their inhibitory functions, we identified homologous structures to the AlphaFold-predicted Acr proteins using structure-based distance measures [31] (see Section 4 for details).

First, we used a subset of the Acr dataset with experimental structures (Set A) to successfully validate the closest structural homologue to each AlphaFold-predicted Acr protein matched with its true experimental structure (Supplementary Table S3).

Second, we analysed the Acr proteins without experimental structures (Set B) by matching the AlphaFold-predicted structures to the closest homologues from the Protein Data Bank archive [31] (Supplementary Table S4). We first validated that Acr proteins retrieved their neighbours in the same clade as the closest homologue, for those cases where their neighbouring proteins had experimental structures in the Protein Data Bank. For example, the Acr proteins labelled '0434' and '0435' are in the same clade, and the closest homologue of the Acr protein '0435' matched with the experimental structure of the Acr protein '0434'. Functional analysis of the closest homologue to each Acr protein revealed a wide variety of protein functions, including polymerase, ligase, nuclease, regulation, and transport (Figure 4c). We acknowledge that some of the closest homologues have low structural similarity (Z-score below 4); however, it is intriguing that these Acr proteins have no close structural homologues in the Protein Data Bank. Some Acr proteins from the families AcrIC6, AcrVIA2, AcrIIA19, and AcrIF8 (Figure 4a) have no structural homologues (significance threshold for similarity: Z-score = 2).

Third, we cross-examined the Acr proteins whose inhibitory mechanism was experimentally characterised to verify that the homologues retrieved were functionally related (indicated as the middle ring in Figure 3b). The homologue functions of this subset were related to a wide variety of functional domains (Supplementary Table S4). For instance, the two Acr proteins in the same clade (labelled '3628' and '3642') were characterised to inhibit the DNA-binding of Cas proteins [32] and their structural homologues have functional domains of ligase. A few closest homologues with functions related to acetyltransferase drew particular attention, as a recent biochemical study revealed an unprecedented mechanism of inhibiting CRISPR-Cas systems through enzymatic activity rather than through direct interaction [24]. According to this study, the closest structural homologue to this Acr protein (labelled '3625') was found to be N-Alpha-Acetyltransferase from *Homo sapiens* (4U9W-C) (Figure 4b), despite their low sequence similarity. We found another homologue (1Y9W-A) with a better similarity score to the AlphaFold-predicted structure of this Acr protein, that had the functional annotation of acetyltransferase from *Bacillus cereus* (Table 1).

In addition, we found several uncharacterised Acr proteins in the same clade of the structural tree (between '0430' and '3681') related to Acetylglucosaminidase from various Acr families, including AcrIC, AcrIE, AcrIF, AcrIIA, and AcrVIB. Intriguingly, several proteins have homologues with the functional annotations of nuclease activity, which is reminiscent of the newly-discovered mechanism of nuclease activity against crRNAs and CRISPR-Cas signalling molecules [33,34].

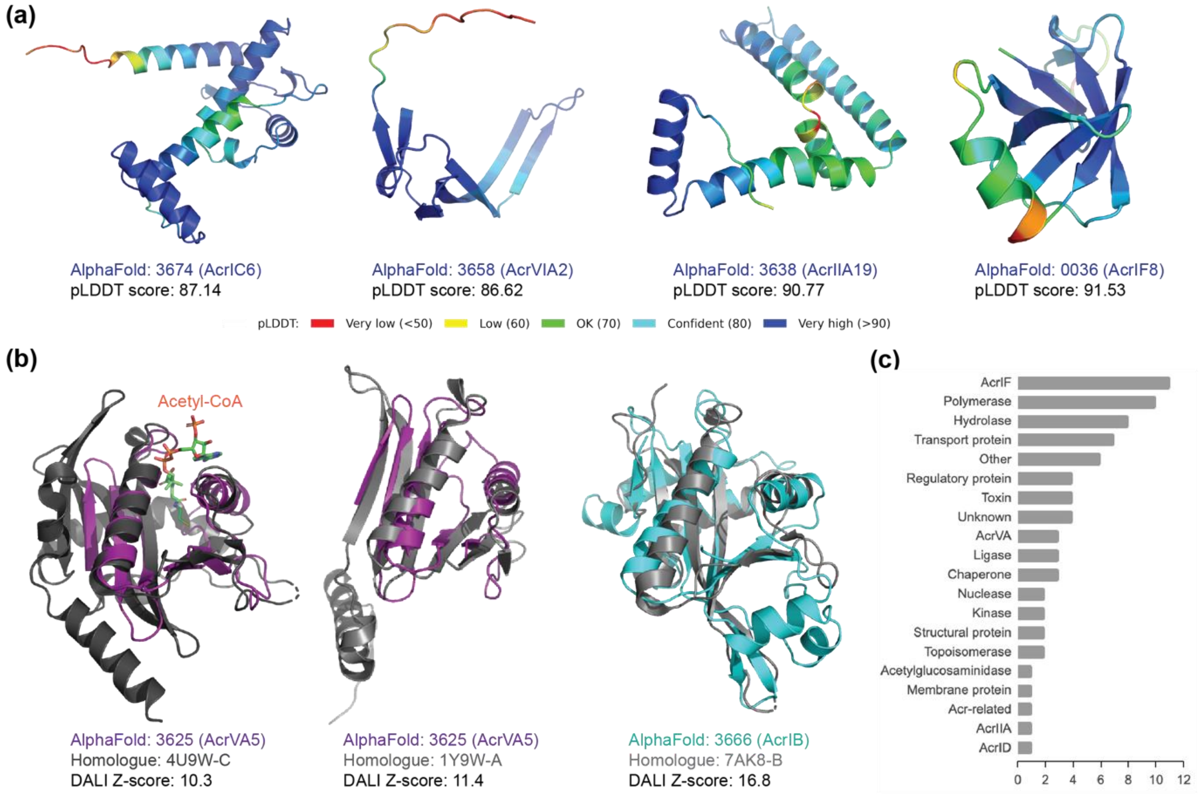

Figure 4. AlphaFold-predicted 3-D structures of outlier anti-CRISPR proteins. (**a**) The AlphaFold-predicted structures of the Acr proteins without homologues in the Protein Data Bank archive (Dali Z-score < 2). The 3-D protein structures are coloured according to the b-factor spectrum in PyMol, with a per-residue estimate of the AlphaFold2 confidence on a scale from 0–100 (high pLDDT accuracy in blue, low pLDDT accuracy in red). (**b**) Superimposition of the AlphaFold-predicted structures of Acr proteins and their closest structural homologues retrieved from the Protein Data Bank archive. All the closest structural homologues in grey have functional annotations related to acetyltransferase. The homologue to the Acr protein labelled '3625' has a cofactor (acetyl-CoA) bound, revealing the functionally critical site of the enzyme. (**c**) Functional analysis of the closest homologues to the AlphaFold-predicted Acr proteins without experimental structures (Set B). Only the functional annotations above the significance threshold of Dali Z-score (>4) were included.

Table 1. Closest homologue to the AlphaFold-predicted structure of Acr proteins with acetyltransferase annotations.

Acr ID	Family	Type	Length	Homologue	Dali Z-Score	Annotation	%ID Structure	%ID Sequence
3625	AcrVA5	V-A	92	4U9W-C	10.3	N-Alpha-Acetyltransferase	14	6.8
3625	AcrVA5	V-A	92	1Y9W-A	11.3	Acetyltransferase	22	16.8
3666	AcrIB	I-B	193	7AK8-B	16.8	Acetyltransferase	21	17.6

%ID structure: percentage identity in structure. %ID sequence: percentage identity in sequence.

2.4. New Anti-CRISPR Family of Acetylation Inhibition

Previously described Acr proteins of the Acr family VA5 disable Type V Cas12a by acetylation, which leads to a complete loss of the DNA-cleavage activity [24]. We found that this AcrVA5 protein (labelled '3625') structurally aligned closer to *Bacillus cereus* acetyltransferase than the previous structural homology of *Homo sapiens* acetyltransferase (Figure 4b). We further identified other Acr proteins that are related to this AcrVA5 protein on the evolutionary trees of the Acr proteins (Figure 3b). Notably, there are two Acr proteins in the same clade as the AcrVA5 protein on the structural tree, one of which was lacking experimentally validated structure or function. Analysis of its function using the structural homologues reveals that this Acr protein (labelled '3666') is related to acetyltransferase. Interestingly, this Acr protein belongs to a different Acr family (AcrIB) than the previously identified AcrVA5. Its superimposition with the closest structural homologue reveals a similar structural alignment at the functionally critical site of the acetyltransferase where acetyl-CoA binds (Figure 4b). The sequence identity of the AcrIB protein to its closest homologue was found to be 17.6% (Supplementary Table S4), while the structural identity between these two proteins in 3-D was found to be higher at 21% (Table 1 and Supplementary Figure S5). On the phylogenetic tree, this AcrIB protein was not placed close to the other two Acr proteins of acetyltransferase function (labelled '3625' and '0557') (Figure 3a), demonstrating that these two types of Acr proteins have close structural similarity but not genetic similarity. This finding suggests that for proteins with low sequence similarity, structure-based trees cluster proteins with most similar biochemical functional properties perform better than sequence-based trees [30]. Using the structural tree, we discovered a new family of Acr proteins belonging to AcrIB that was structurally similar to acetyltransferase from a different organism (gram-negative bacteria *Salmonella enterica*), whereas the previously characterised AcrVA5 matched to acetyltransferase from gram-positive bacteria *Bacillus cereus*.

3. Discussion

We show that the 3-D structures of Acr proteins predicted with AlphaFold2 achieve high accuracy. The structural tree reconstructed from these AlphaFold-predicted structures display more diversity of Acr proteins with no common evolutionary origin as compared to the phylogenetic tree. On the structural tree, the Acr proteins form small clades by their unique structural similarity, which are also related by the inhibition mechanism. The functional annotations of the Acr protein homologues are extremely diverse, relating to a wide range of enzymatic and regulatory activities from different organisms. Most characterised Acr proteins inhibit host CRISPR-Cas systems by direct interference; we show that this category of Acr proteins displays various functional annotations and unique structural forms in the multiple branches of the structural tree.

Specifically, we found a number of Acr proteins with homologue annotations related to acetylation. A recent discovery of Acr proteins that manipulate CRISPR-Cas systems through enzymatic activities demonstrates extensive phage defence mechanisms driven by the intense host-parasite arms race [24,33]. Through the AlphaFold-predicted structural analysis, we found a novel family of Acr proteins (AcrIB) from the genome of a human pathogen (*Leptotrichia buccalis* C-1013-b) that shows more structural similarity to acetyltransferase than the previously characterised AcrVA protein. Intriguingly, other Acr proteins on the multiple branches of the structural tree have homologues related to different types of acetyltransferase enzymes from heterologous species. As Acr proteins with acetyltransferase activities permanently disable Cas proteins by covalent modification [24], other Acr proteins with enzymatic activities such as acetylglucosaminidase are expected to have similar inactivation functions of biochemically modifying CRISPR-Cas systems. The Acr proteins could evolve independently from various host genomes and mobile genetic elements, exploiting a vast inventory of protein structures as the basis for their counter-defence advantage [10,13,35].

More broadly, Acr proteins are exceptional examples of coevolution dynamics optimizing the phage genomes to manipulate host systems and maximize survival. As phages can only replicate within host cells and are void of metabolic capacity to synthesize small molecules, their counter-defence machinery against the sophisticated and extensive prokaryotic anti-phage systems is protein-based. Nonetheless, phages are the most abundant biological entities in the biosphere [36], and their successful protein-based viral arsenals such as Acr proteins provide an important insight on how to expand the potential of protein therapeutics. Specifically, we could get inspiration from these phage-derived protein structures that resemble segments of related target proteins, to design highly-specialised protein inhibitors with diverse protein-manipulating strategies [37]. In future, utilizing the ability of small proteins to engage in indirect interference through enzymatic activities is to be explored against disease-causing proteins (such as in cancer) and disease-causing organisms (such as drug-resistant bacteria) [36].

4. Materials and Methods

4.1. Curation of Anti-CRISPR Datasets

The anti-CRISPR dataset contained 443 Acr proteins (207 verified, 236 putative) that inhibit a wide range of CRISPR-Cas systems including I, II, III and VI from the Anti-CRISPRdb [27]. The term ‘verified’ indicates that the protein was validated as CRISPR-Cas-inactivating Acr (either by the database or other published papers), while ‘putative’ indicates the protein was predicted to be Acr without sufficient experimental support. The anti-CRISPR dataset was further curated into Set A, Set B, and Set C for AlphaFold2, according to the availability of experimentally reconstructed 3-D macromolecular structures (hereby referred to as “experimental structures”) (Supplementary Table S1). Each protein was annotated with the Acr family, type of inhibited CRISPR-Cas systems, NCBI accession, genetic sequence, source organism, taxonomy, and inhibitory mechanism when available (Supplementary Tables S3 and S4).

4.2. Prediction of Anti-CRISPR Protein Structure with AlphaFold2

We predicted the 3-D protein structures of each set with AlphaFold2, using the Acr protein sequence as the input to AlphaFold2 (Figure 2a). AlphaFold2 creates genetic and structural representations by comparing the protein sequence with several pre-installed databases. Those representations are used as input to five prediction models to generate five candidate 3-D structures. The result with the highest per-residue confidence score (pLDDT: per residue estimate of confidence on a scale from 0 to 100 [29]) among the five results was determined as the final structure and saved in a protein data bank (PDB) format (Supplementary Figure S6). For the Acr datasets, we used the PDB archive until 31 December 2012 as templates in AlphaFold2 to exclude the true experimental structures of the Acr proteins. The details of the experiments related to the hardware specification and to the processor performance are given in the Supplementary Material (Supplementary Table S5 and Supplementary Table S6, respectively).

4.3. Comparison of AlphaFold2 Performance on Anti-CRISPR against CASP14

To validate the performance of AlphaFold2 for predicting Acr structures, we benchmarked the CASP14 dataset against Set A and Set C of Acr proteins with the corresponding true experimental structures available. We excluded predicted structure and experimental structure pairs for which the TM-score and/or the Z-score were too low. Finally, 52 pairs of CASP14, 99 pairs of Set A, and 207 pairs of Set C were used for the comparison study. We used the TM-score [38,39] and Dali Z-score [40] as similarity measures between the predicted and the experimental structures. Unlike traditional metrics (e.g., root-mean-square deviation), the TM-score is length-independent and more sensitive to the global similarity than to the local variation. The Dali Z-score is the sum of the equivalent residue-wise intermolecular distances among two proteins, and does not have a fixed upper bound [40]. We then used the following relative Z-error to calculate the relative difference:

$$Z_{error} = \frac{Z_{gt} - Z_{pd}}{Z_{gt}} \tag{1}$$

where Z_{gt} is the self Dali Z-score between experimental structure and itself, and Z_{pd} is the Dali Z-score between experimental structure and predicted structure. We obtained the Z_{pd} and TM-score for the CASP14 set of AlphaFold2 from the CASP14 assessment scores [41], whereas Z_{gt} of CASP14 and Z_{gt} and Z_{pd} of Set A and Set C were calculated using DaliLite.v5 [40]. Finally, a protein structure comparison and clustering tool called MaxCluster [42] was used to calculate the TM-scores of Set A and Set C. Both distance metrics have values between 0 and 1, with 1 as the best score for TM-scores and 0 as the best score for relative Z-errors.

4.4. Reconstruction of Evolutionary Trees of Anti-CRISPR Proteins

We reconstructed the evolutionary trees of the anti-CRISPR dataset (Set A and Set B) using sequence-based and structure-based inference. Set C was excluded from the evolutionary analysis due to the absence of functional verification and due to sequence variation (Supplementary Figure S3). The sequence-based tree of the Acr proteins was built by aligning the amino acid sequences using a multiple alignment program, MAFFT (version 7.471, -auto option) [43]. The multiple sequence alignment of the Acr proteins was then visualized using Jalview (version 2.11.1.3) with a conservation visibility of 15% (Supplementary Figure S2) [44]. Subsequently, a phylogenetic tree of the Acr proteins was built with IQ-Tree using ModelFinder (-auto option) to find the best-fit model among the supported range of protein substitution models [45,46] (Supplementary Table S7). Using the best-fit substitution model, 1000 ultrafast bootstrap replicates were run to check bootstrap support of the reconstructed tree topology [47].

The structure-based tree of the Acr proteins was built by calculating the similarity matrix between the Dali Z-scores of the AlphaFold-predicted structures and its corresponding experimental structures. We used the Dali server [40] for generating structural trees from hierarchical clustering of the similarity matrix. The structural tree of the Acr proteins was generated from distance matrices, where the pseudo-distance between two structures Q and T was defined as [48]:

$$D_{QT} = Z_{QQ} + Z_{TT} - 2Z_{QT} \tag{2}$$

The hierarchical clustering of the similarity matrix was outputted as a Newick formatted dendrogram. The phylogenetic tree and the structural tree of the Acr proteins were visualized with iTOL (version 4) and iTOL annotation editor [49,50] with the following labels: Acr Family, Taxonomy, and Inhibition Mechanism.

4.5. Congruence among Distance Matrices of Sequence-Based and Structure-Based Trees

We measured the congruence among distance matrices of the reconstructed trees from the sequence-based and structure-based methods using Kendall's coefficient of concordance, W, which ranges from 0 (no congruence) to 1 (complete congruence) [51]. First, we computed the cophenetic value of pairwise distances between the terminals from a phylogenetic tree using its branch lengths with the function cophenetic.phylo from ape-package (version 5.0) [52]. Then, we used the function CADM.global to calculate the coefficient of concordance among the distance matrices of the sequence-based and structure-based trees of the Acr proteins through a permutation test.

4.6. Visualization of Protein Structure Superimposition

For functional analysis, the AlphaFold-predicted structures with functional annotations of interest were superimposed with their structural homologues using PyMol (version 2.5.2) to visualize the overlap in structure of the functionally active sites. The inhibitory mechanism of the Acr proteins without experimental structure was inferred through examining functional annotations of the structural homologues to the AlphaFold-predicted structure, with the significance threshold of Z-score > 4.

4.7. Code Availability

Protein structures were predicted with AlphaFold2, available under an open-source license at https://github.com/deepmind/alphafold, accessed on 27 September 2021. For protein structure similarity metrics, we used MaxCluster (http://www.sbg.bio.ic.ac.uk/~maxcluster/index.html, accessed on 13 October 2021) for TM-score and DaliLite.v5 (http://ekhidna2.biocenter.helsinki.fi/dali/README.v5.html, accessed on 24 October 2021) for the Dali Z-score. For MSA, we used MAFFT.v7 (https://mafft.cbrc.jp/alignment/server, accessed on 29 October 2021) and Jalview.v2 (https://www.jalview.org, accessed on 29 October 2021) for visualization. For phylogenetic tree reconstruction, we used IQ-Tree (http://www.iqtree.org, accessed on 14 November 2021) with ModelFinder and UFBoot options. For structural tree reconstruction, we used the Dali server (http://ekhidna2.biocenter.helsinki.fi/dali, accessed on 16 November 2021) for building dendrograms. The 3-D Structure visualizations were created in Pymol v.2.5.2 (https://pymol.org, accessed on 4 November 2021) and Py3DMol v.1.7.0 (https://pypi.org/project/py3Dmol, accessed on 26 October 2021) with Jupyter v.1.0.0 (https://jupyter.org, accessed on 26 October 2021). For data analysis, Python v.3.6.4 (https://www.python.org, accessed on 27 November 2021), NumPy v.1.17.5 (https://github.com/numpy/numpy, accessed on 27 November 2021), SciPy v.1.1.0 (https://www.scipy.org, accessed on 27 November 2021), seaborn v.0.9.0 (https://github.com/mwaskom/seaborn, accessed on 25 November 2021), Matplotlib v.3.3.4 (https://github.com/matplotlib/matplotlib, accessed on 24 November 2021), pandas v.0.22.0 (https://github.com/pandas-dev/pandas, accessed on 24 November 2021) were used.

5. Conclusions

The high biodegradability issue of protein therapeutics has partially been solved by the recent success of mRNA vaccine delivery using lipid nanoparticles [53], making low risk protein therapeutics ever more attractive to the industry. From the AlphaFold-predicted structures, we accelerated the structural and functional analysis of the Acr proteins whose experimental 3-D structures remain to be resolved. In conclusion, we wonder whether there is a vast repertoire of unexplored protein structural configurations that can be exploited for protein drug design, given the number of Acr proteins without homologues in the current protein structure domain.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ph15030310/s1. Figure S1. The performance of AlphaFold2 on the Acr protein datasets in comparison to the CASP14 dataset using RMSD; Figure S2. The structural tree of anti-CRISPR proteins reconstructed using structure-based methods; Figure S3. Multiple sequence alignment of anti-CRISPR proteins using MAFFT, visualized with Jalview; Figure S4. The structural tree of anti-CRISPR proteins reconstructed using structure-based methods; Figure S5. A scatter plot of percentage identity in structure against percentage identity in sequence of the closest homologue to each protein; Figure S6. Outlier AlphaFold structures with bad scores; Table S1. Acr protein types in Anti-CRISPRdb; Table S2. Detailed statistics and scores on AlphaFold prediction performance; Table S3. Closest homologue (highest TM score) to the AlphaFold-predicted structure of Acr proteins with experimentally reconstructed 3-D macromolecular structures (Set A) from the Protein Data Bank archive; Table S4. Closest homologue (highest DALI Z-score) to the AlphaFold-predicted structure of Acr proteins without experimentally reconstructed 3-D macromolecular structures (Set B) from the Protein Data Bank archive; Table S5. Hardware specification for AlphaFold experiments; Table S6. Time to create a 3-D structure from one protein sequence (in seconds); Table S7. The best-fit-model of the phylogenetic tree of each Acr set using IQ-TREE ModelFinder.

Author Contributions: Conceptualization, H.S.; methodology, H.-M.P. and H.S.; software, H.-M.P. and H.S.; validation, H.-M.P., Y.P. and H.S.; formal analysis, H.-M.P., Y.P. and H.S.; investigation, H.-M.P., Y.P. and H.S.; resources, J.V. and W.D.N.; data curation, H.-M.P. and H.S.; writing, H.-M.P., Y.P., J.V., A.V.M., W.D.N. and H.S.; visualization, H.-M.P. and H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ghent University Global Campus grant number [RC4].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. All the input protein sequences analysed in this study are available in Anti-CRISPRdb at https://doi.org/10.1093/nar/gkx835 (accessed on 13 November 2021). As well as our project GitHub page which can be found here: https://github.com/powersimmani/ACR_alphafold, accessed on 28 November 2021. All the 3-D structures used as ground truth for calculating Dali Z-scores and TM-scores; and superimposing structures in this study are available in Protein Data Bank at https://www.rcsb.org/ accessed on 13 November 2021. CASP (Critical Assessment of Structure Prediction) competition datasets were used for measuring AlphaFold performance. This data can be found here: https://predictioncenter.org/casp14/, accessed on 29 September 2021.

Acknowledgments: We thank members of the Center for Biosystems and Biotech Data Science at GUGC for helpful discussions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; et al. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921. [PubMed]
2. Venter, J.C.; Adams, M.D.; Myers, E.W.; Li, P.W.; Mural, R.J.; Sutton, G.G.; Smith, H.O.; Yandell, M.; Evans, C.A.; Holt, R.A.; et al. The sequence of the human genome. *Science* **2001**, *291*, 1304–1351. [CrossRef]
3. Hopkins, A.L.; Groom, C.R. The druggable genome. *Nat. Rev. Drug Discov.* **2002**, *1*, 727–730. [CrossRef]
4. Goeddel, D.V.; Kleid, D.G.; Bolivar, F.; Heyneker, H.L.; Yansura, D.G.; Crea, R.; Hirose, T.; Kraszewski, A.; Itakura, K.; Riggs, A.D. Expression in Escherichia coli of chemically synthesized genes for human insulin. *Proc. Natl. Acad. Sci. USA* **1979**, *76*, 106–110. [CrossRef] [PubMed]
5. Leader, B.; Baca, Q.J.; Golan, D.E. Protein therapeutics: A summary and pharmacological classification. *Nat. Rev. Drug Discov.* **2008**, *7*, 21–39. [CrossRef] [PubMed]
6. Bishop, B.; Dasgupta, J.; Klein, M.; Garcea, R.L.; Christensen, N.D.; Zhao, R.; Chen, X.S. Crystal structures of four types of human papillomavirus L1 capsid proteins: Understanding the specificity of neutralizing monoclonal antibodies. *J. Biol. Chem.* **2007**, *282*, 31803–31811. [CrossRef]
7. Fávero-Retto, M.P.; Palmieri, L.C.; Souza, T.A.C.B.; Almeida, F.C.L.; Lima, L.M.T.R. Structural meta-analysis of regular human insulin in pharmaceutical formulations. *Eur. J. Pharm. Biopharm.* **2013**, *85*, 1112–1121. [CrossRef]
8. Luthra, A.; Langley, D.B.; Schofield, P.; Jackson, J.; Abdelatti, M.; Rouet, R.; Nevoltris, D.; Mazigi, O.; Crossett, B.; Christie, M.; et al. Human antibody bispecifics through phage display selection. *Biochemistry* **2019**, *58*, 1701–1704. [CrossRef]
9. Ostrov, D.A.; Shi, W.; Schwartz, J.C.; Almo, S.C.; Nathenson, S.G. Structure of murine CTLA-4 and its role in modulating T cell responsiveness. *Science* **2000**, *290*, 816–819. [CrossRef] [PubMed]
10. Dion, M.B.; Oechslin, F.; Moineau, S. Phage diversity, genomics and phylogeny. *Nat. Rev. Microbiol.* **2020**, *18*, 125–138. [CrossRef] [PubMed]
11. Koonin, E.V.; Dolja, V.V.; Krupovic, M. Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* **2015**, *479–480*, 2–25. [CrossRef] [PubMed]
12. Watson, B.N.J.; Steens, J.A.; Staals, R.H.J.; Westra, E.R.; van Houte, S. Coevolution between bacterial CRISPR-Cas systems and their bacteriophages. *Cell Host Microbe* **2021**, *29*, 715–725. [CrossRef]
13. Hampton, H.G.; Watson, B.N.J.; Fineran, P.C. The arms race between bacteria and their phage foes. *Nature* **2020**, *577*, 327–336. [CrossRef] [PubMed]
14. Jansen, R.; van Embden, J.D.A.; Gaastra, W.; Schouls, L.M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* **2002**, *43*, 1565–1575. [CrossRef] [PubMed]
15. Mojica, F.J.M.; Díez-Villaseñor, C.; García-Martínez, J.; Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **2005**, *60*, 174–182. [CrossRef] [PubMed]
16. Barrangou, R.; Fremaux, C.; Deveau, H.; Richards, M.; Boyaval, P.; Moineau, S.; Romero, D.A.; Horvath, P. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **2007**, *315*, 1709–1712. [CrossRef]
17. Deltcheva, E.; Chylinski, K.; Sharma, C.M.; Gonzales, K.; Chao, Y.; Pirzada, Z.A.; Eckert, M.R.; Vogel, J.; Charpentier, E. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **2011**, *471*, 602–607. [CrossRef]

18. Jinek, M.; Chylinski, K.; Fonfara, I.; Hauer, M.; Doudna, J.A.; Charpentier, E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **2012**, *337*, 816–821. [CrossRef] [PubMed]
19. Makarova, K.S.; Wolf, Y.I.; Iranzo, J.; Shmakov, S.A.; Alkhnbashi, O.S.; Brouns, S.J.J.; Charpentier, E.; Cheng, D.; Haft, D.H.; Horvath, P.; et al. Evolutionary classification of CRISPR-Cas systems: A burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **2020**, *18*, 67–83. [CrossRef] [PubMed]
20. Bondy-Denomy, J.; Pawluk, A.; Maxwell, K.L.; Davidson, A.R. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature* **2013**, *493*, 429–432. [CrossRef] [PubMed]
21. Pawluk, A.; Bondy-Denomy, J.; Cheung, V.H.W.; Maxwell, K.L.; Davidson, A.R. A new group of phage anti-CRISPR genes inhibits the type I-E CRISPR-Cas system of *Pseudomonas aeruginosa*. *MBio* **2014**, *5*, e00896:1–e00896:7. [CrossRef] [PubMed]
22. Marino, N.D.; Pinilla-Redondo, R.; Csörgő, B.; Bondy-Denomy, J. Anti-CRISPR protein applications: Natural brakes for CRISPR-Cas technologies. *Nat. Methods* **2020**, *17*, 471–479. [CrossRef]
23. Pawluk, A.; Staals, R.H.J.; Taylor, C.; Watson, B.N.J.; Saha, S.; Fineran, P.C.; Maxwell, K.L.; Davidson, A.R. Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species. *Nat. Microbiol.* **2016**, *1*, 16085:1–16085:6. [CrossRef] [PubMed]
24. Dong, L.; Guan, X.; Li, N.; Zhang, F.; Zhu, Y.; Ren, K.; Yu, L.; Zhou, F.; Han, Z.; Gao, N.; et al. An anti-CRISPR protein disables type V Cas12a by acetylation. *Nat. Struct. Mol. Biol.* **2019**, *26*, 308–314. [CrossRef] [PubMed]
25. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef]
26. Callaway, E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* **2020**, *588*, 203–204. [CrossRef] [PubMed]
27. Dong, C.; Hao, G.F.; Hua, H.L.; Liu, S.; Labena, A.A.; Chai, G.; Huang, J.; Rao, N.; Guo, F.B. Anti-CRISPRdb: A comprehensive online resource for anti-CRISPR proteins. *Nucleic Acids Res.* **2018**, *46*, D393–D398. [CrossRef] [PubMed]
28. Kufareva, I.; Abagyan, R. Methods of protein structure comparison. *Methods Mol. Biol.* **2012**, *857*, 231–257. [PubMed]
29. Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; et al. Highly accurate protein structure prediction for the human proteome. *Nature* **2021**, *596*, 590–596. [CrossRef] [PubMed]
30. Balaji, S.; Srinivasan, N. Comparison of sequence-based and structure-based phylogenetic trees of ho-mologous proteins: Inferences on protein evolution. *J. Biosci.* **2007**, *32*, 83–96. [CrossRef]
31. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef] [PubMed]
32. Watters, K.E.; Shivram, H.; Fellmann, C.; Lew, R.J.; McMahon, B.; Doudna, J.A. Potent CRISPR-Cas9 inhibitors from Staphylococcus genomes. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 6531–6539. [CrossRef] [PubMed]
33. Athukoralage, J.S.; McMahon, S.A.; Zhang, C.; Grüschow, S.; Graham, S.; Krupovic, M.; Whitaker, R.J.; Gloster, T.M.; White, M.F. An anti-CRISPR viral ring nuclease subverts type III CRISPR immunity. *Nature* **2020**, *577*, 572–575. [CrossRef] [PubMed]
34. Knott, G.J.; Thornton, B.W.; Lobba, M.J.; Liu, J.J.; Al-Shayeb, B.; Watters, K.E.; Doudna, J.A. Broad-spectrum enzymatic inhibition of CRISPR-Cas12a. *Nat. Struct. Mol. Biol.* **2019**, *26*, 315–321. [CrossRef] [PubMed]
35. Shim, H. Feature learning of virus genome evolution with the nucleotide skip-gram neural network. *Evol. Bioinform.* **2019**, *15*, 1176934318821072:1–1176934318821072:10. [CrossRef] [PubMed]
36. Frost, L.S.; Leplae, R.; Summers, A.O.; Toussaint, A. Mobile genetic elements: The agents of open source evolution. *Nat. Rev. Microbiol.* **2005**, *3*, 722–732. [CrossRef] [PubMed]
37. Shim, H.; Shivram, H.; Lei, S.; Doudna, J.A.; Banfield, J.F. Diverse ATPase proteins in mobilomes constitute a large potential sink for prokaryotic host ATP. *Front. Microbiol.* **2021**, *12*, 691847:1–691847:11. [CrossRef] [PubMed]
38. Xu, J.; Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **2010**, *26*, 889–895. [CrossRef]
39. Zhang, Y.; Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **2004**, *57*, 702–710. [CrossRef]
40. Holm, L. Using Dali for protein structure comparison. *Methods Mol. Biol.* **2020**, *2112*, 29–42. [PubMed]
41. Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; Moult, J. Critical assessment of methods of pro-tein structure prediction (CASP)-Round XIV. *Proteins* **2021**, *89*, 1607–1617. [CrossRef] [PubMed]
42. Siew, N.; Elofsson, A.; Rychlewski, L.; Fischer, D. MaxSub: An automated measure for the assessment of protein structure prediction quality. *Bioinformatics* **2000**, *16*, 776–785. [CrossRef] [PubMed]
43. Katoh, K.; Rozewicki, J.; Yamada, K.D. MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **2019**, *20*, 1160–1166. [CrossRef]
44. Taylor, W.R. Residual colours: A proposal for aminochromography. *Protein Eng.* **1997**, *10*, 743–746. [CrossRef]
45. Nguyen, L.T.; Schmidt, H.A.; von Haeseler, A.; Minh, B.Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **2015**, *32*, 268–274. [CrossRef]
46. Kalyaanamoorthy, S.; Minh, B.Q.; Wong, T.K.F.; von Haeseler, A.; Jermiin, L.S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **2017**, *14*, 587–589. [CrossRef]
47. Hoang, D.T.; Chernomor, O.; von Haeseler, A.; Minh, B.Q.; Vinh, L.S. UFBoot2: Improving the Ultra-fast Bootstrap Approximation. *Mol. Biol. Evol.* **2018**, *35*, 518–522. [CrossRef]
48. Holm, L. DALI and the persistence of protein shape. *Protein Sci.* **2019**, *29*, 128–140. [CrossRef] [PubMed]

49. Letunic, I.; Bork, P. Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **2019**, *47*, W256–W259. [CrossRef] [PubMed]
50. Letunic, I.; Bork, P. Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics* **2007**, *23*, 127–128. [CrossRef] [PubMed]
51. Legendre, P. Species associations: The Kendall coefficient of concordance revisited. *JABES* **2005**, *10*, 226–245. [CrossRef]
52. Paradis, E.; Schliep, K. ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **2019**, *35*, 526–528. [CrossRef]
53. Jackson, N.A.C.; Kester, K.E.; Casimiro, D.; Gurunathan, S.; Derosa, F. The promise of mRNA vaccines: A biotech and industrial perspective. *npj Vaccines* **2020**, *5*, 11:1–11:6. [CrossRef] [PubMed]