bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Backbone brackets and arginine tweezers delineate class I and class II aminoacyl tRNA synthetases

Florian Kaiser[1,2]☯*, Sebastian Bittrich[1,2]☯, Sebastian Salentin[2], Christoph Leberecht[1,2], V. Joachim Haupt[2], Sarah Krautwurst[1], Michael Schroeder[2], Dirk Labudde[1]

1 University of Applied Sciences Mittweida, Technikumplatz 17, 09648, Mittweida, Germany
2 Biotechnology Center (BIOTEC), TU Dresden, Tatzberg 47-49, 01307, Dresden, Germany

☯These authors contributed equally to this work.
* florian.kaiser@hs-mittweida.de

Abstract

All living organisms share the machinery to translate RNA into amino acid sequences. One key component of this machinery are aminoacyl tRNA synthetases, which ligate tRNAs to amino acids. Sequence analyses revealed that these enzymes evolved to complementary classes, which can be characterized by several sequence motifs. However, there are no structural motifs, which capture the core function of the classes: high specificity ligand interaction. We identified backbone brackets and arginine tweezers and show that these two motifs optimize ligand recognition with complementary mechanisms. They are the most compact and simple characteristic to distinguish the aminoacyl tRNA synthetase class I from II. These findings support the hypothesis that the evolutionary convergence regarding function of aminoacyl tRNA synthetases was balanced by a divergence regarding ligand interaction.

Author summary

Aminoacyl tRNA synthetases (aaRS) are primordial enzymes essential for interpretation and transfer of genetic information. Disturbances in this fine-tuned system lead to severe malfunctions in organisms and to lethal diseases. The increasing amount of experimentally determined three-dimensional structures of aaRS opens up new avenues for high-throughput analyses of molecular mechanisms. In this study, we present an exhaustive structural analysis of the binding mechanisms of aaRS enzymes and discuss ligand recognition motifs. We unveil a divergent implementation of enzyme substrate recognition in each aaRS class. While class I binds via interactions mediated by conserved backbone hydrogen bonds, class II uses a pair of arginine residues to establish salt bridges. We show how evolution achieves binding of the same ligand species with completely different mechanisms. In addition, we demonstrate that sequence analysis for conserved residues may miss important functional aspects which can only be revealed by structural studies. Further detailed insights in aaRS substrate interaction and a manually curated high-quality dataset of aaRS structures serve as a rich resource for in-depth studies of these extraordinary enzymes.

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

The synthesis of proteins is fundamental to all organisms. It requires a complex molecular machinery of more than 100 entities to ensure efficiency and fidelity [1–3]. The ribosome pairs an mRNA codon with its corresponding anticodon of the tRNA molecule that delivers the cognate amino acid. Aminoacyl tRNA synthetases (aaRS) ligate amino acids to their corresponding tRNA, which is why they are key players in the transfer of genetic information. Individual aaRS and their mechanism to discriminate similar amino acids have been extensively studied on the structural level [4–7] to detect lethal or disease-relevant mutation spots. However, a comprehensive and comparative study of structural features in aaRS proteins is missing. There are no structural motifs known that capture differences of the ligand recognition mechanism.

Starting from a manually curated dataset of 972 individual aaRS protein molecules (448 chains for class I and 524 chains for class II) deposited in the Protein Data Bank (PDB) [8], we identified two functionally convergent but intrinsically different ligand recognition motifs, the *backbone brackets* and the *arginine tweezers*. These key interaction patterns are an outstanding example of evolutionary diversification and functional convergence, since they do not share structural features or conservation on sequence level (Fig 1).

Fig 1. Backbone brackets and arginine tweezers were identified as characteristic structural motifs from 972 aminoacyl tRNA synthetase 3D protein structures. Both motifs exhibit a complementary interaction mode and geometric characteristics. Backbone brackets are not conserved in sequence, but only in structure, while arginine tweezers are highly conserved in sequence.

The mere existence of proteins and nucleic acids is a chicken-and-egg dilemma. The sequential succession of amino acids in each protein is encoded by nucleic acid blueprints. In turn these proteins are indispensable to replicate and translate nucleic acids. It is debated how this self-referencing system came to be [9]. Three main theories have been proposed to explain the emergence of the self-encoding translational machinery, namely: coevolution [10], ambiguity reduction [11, 12], and stereochemical forces [13]. The interaction between amino acid and nucleic acid lies at the basis of each theory and is linked to the emergence of aaRS [14, 15]. There is strong evidence for two archaic proto-enzymes (urzymes) as the origin of all aaRS, which were among the earliest proteins that enabled the development of life [16–19]. Since then, these urzymes have evolved divergently into two classes I and II (Fig 2), where each is responsible for a distinct set of amino acids [20–22]. Every aaRS recognizes an amino acid and prevents misacylation of tRNAs by being as specific as possible.

It is still elusive how these two urzymes have evolved into 20 concrete realizations – referred to as aaRS types – observed in each organism today [23, 24]. One likely scenario is that amino acids were gradually incorporated into the genetic code and inefficient proteins were replaced by better versions over the course of evolution [14]. While similar amino acids were once processed by one aaRS, specificity was required to cope with increasing complexity. Some examples of such generic aaRS can still be found in organisms today [25, 26].

Since aaRS enzymes are essential to protein synthesis, they are under exceptional evolutionary pressure to maintain the necessary accuracy in amino acid recognition [29, 30].

aaRS enzymes can be separated based on their topology: the catalytic domain of class I adapts a common Rossmann fold [31], whereas class II possesses a unique fold [32–34]. The modular architecture of aaRS has evolved well-orchestrated and was optimized for its specific requirements [21, 35]. In principle, all aaRS have to conserve

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Fig 2. The two aaRS classes and amino acids they ligate to the cognate tRNA. Based on the physicochemical properties of the amino acids (colored according to [27]) no distinction can be made between the two classes. Lysine is mostly processed by class II aaRS, but in some archaic organisms a class I aaRS is responsible for lysine [28]. Prior to tRNA ligation the amino acid ligand is converted to its activated form: aminoacyl adenylate.

three functions: correct recognition of the tRNA identity and amino acid as well as ligation of both. The anticodon binding domain ensures the tRNA integrity by recognizing particular features of the anticodon [36, 37]. The identification and transfer of amino acids is then mediated by the catalytic domain. To minimize errors in protein biosynthesis, pre- and post-transfer editing mechanisms are conducted by approximately half of the aaRS [4, 38, 39].

Even though aaRS catalyze the same type of reaction, the exact mechanism depends on the aaRS class, the handled amino acid, and the host organism. By exploiting this diversity, the complex development of the genetic code [15], the phylogeny of organisms [34], and aaRS-associated diseases [40] were studied.

Mutations in aaRS-coding genes are usually lethal or lead to severe diseases, including neurological and metabolic disorders as well as cancer [40]. One of the most common diseases related to aaRS is Charcot-Marie-Tooth (CMT) disease, a dominantly inherited neuropathy [40, 41].

With respect to sequence or structure both classes have nothing in common [34]. Previous sequence-based studies demonstrated that the diversity results from fusion, duplication, recombination and horizontal gene transfer [34, 42].

Tracking the ancestry of aaRS is particularly difficult due to the low evolutionary stability [43]. Still, two pairs of class-specific and mutually exclusive sequence motifs were identified, which mediate interactions with the adenosine phosphate and catalysis [5, 20, 44]. Class I features the four-residue HIGH and five-residue KMSKS motifs [20]. The so-called motifs "2" and "3" are the class II equivalents, but here the essential mechanism can be broken down to two highly conserved arginine residues [20, 45, 46]. All these sequence motifs were described as well-conserved [5]. However, the KMSKS motif is located in a mobile loop [5] while the class II motifs are less conserved [43] and more variable in their relative arrangement [20].

A more comprehensive study of available aaRS structures was conducted by O'Donoghue and Luthey-Schulten [47]. Two major structural alignments were calculated for class I and class II, respectively, that revealed high structural similarity within each class with average sequence identity below 10% [47].

A structural study enriched with ligand interaction data allows to capture the core function of both classes in a previously unmatched compactness.

Results

This study presents a dataset of aaRS structures annotated with ligand information, which serves as a stepping stone to understand common and characteristic ligand interaction properties. It is composed of 972 individual chains containing 448 (524) class I (class II) catalytic aaRS domains and covers ligand-bound structures for each aaRS type. The dataset is provided in S1 File and S2 File. The class I chains originate from 256 biological assemblies and comprise 151 bacterial, 84 eukaryotic, 20 archaea, and one viral structure. The class II chain set corresponds to 267 biological assemblies where 102 are of bacterial origin, 104 from eukaryotes, and 61 from archaea (for a detailed overview see S9 Fig). The sequence identity is below 33% (29%) for 95% of all class I

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(class II) structures, while pairwise structure similarity is high with over 0.8 for 95% of the structures (S8 Fig) according to [48]. The high sequential diversity probably stems from the variety of covered organisms. In contrast, the low structural diversity can be seen as a result of conserved function.

Sequence positions of all structures in the dataset were unified using a multiple sequence alignment (MSA) and all further referenced positions are given in accordance to this renumbering (Section Mapping of binding sites, S5 File and S6 File).

Backbone brackets and arginine tweezers

In order to investigate the contacts between aaRS residues and their ligands, noncovalent protein-ligand interactions were annotated. This revealed two highly consistent interaction patterns between catalytic site residues and the adenosine phosphate part of the ligand: conserved backbone hydrogen bonds in class I and two arginines with conserved salt bridges and side chain orientations in class II.

Strikingly, the residues mediating the backbone interactions could be mapped in 441 of 448 (98%) class I renumbered structures at the two positions 274 and 1361. Closer investigation on the structural level revealed geometrically highly-conserved hydrogen bonds between the amino acid backbones and the adenosine phosphate part of the ligand (Fig 3A). These two residues mimic a bracket-like geometry (Fig 3B), enclosing the adenosine phosphate, and were thus termed backbone brackets. The interacting amino acids are not limited to specific residues as their side chains do not form any ligand contacts. Hence, position 274 of the class I motif is not apparent on sequence level while position 1361 exhibits preference for hydrophobic amino acids, e.g. leucine, valine, or isoleucine (Fig 3C).

In contrast, class II aaRS show a conserved interaction pattern of two arginine residues at positions 698 and 1786, which could be identified in 482 of 524 (92%) structures. The two arginine residues grasp the adenosine phosphate part of the ligand (Fig 3D) with their side chains resembling a pair of tweezers (Fig 3E) and were thus named arginine tweezers. These two arginines are invariant in sequence (Fig 3F). Glutamic acid is the most prevalent amino acid at position 700. This residue establishes hydrogen bonds to the adenine group of the ligand in all types except for AlaRS, AsnRS, GlyRS, and PheRS.

The backbone brackets and their counterpart, the arginine tweezers, are both responsible for constant interaction with the adenosine phosphate part of the ligand (all ligand interactions are shown exemplarily in S2 Fig). Mappings of the motif residues to original sequence numbers can be found in S7 File and S9 File. For some structures it was not possible to pinpoint the conserved motifs after unifying sequence positions (listed in S8 File and S10 File).

Further analysis of secondary structure elements for both motifs shows that residues of the backbone brackets are predominantly tied to unordered secondary structure elements (S3 Fig). However, the positions 275, 276, 277, 1359, and 1360 feature consistently unordered secondary structure. A mainly unordered state can also be observed for the first arginine tweezers residue 698, while the following three positions almost exclusively occur in strand regions (S4 Fig). Residue 1786 is always observed in α-helical regions, mostly at the third position of the α-helix element.

The high conservation of backbone or side chain geometry of these motifs suggests essential residues, indispensable for enzyme functionality. To substantiate this assumption, backbone brackets and arginine tweezers were characterized in greater detail and analyzed regarding their ligand interactions and geometric properties.

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Fig 3. (**A**) Structural representation of the backbone brackets motif interacting with Tryptophanyl-5'AMP ligand in TrpRS (PDB:1r6u chain A). The ligand interaction is mediated by backbone hydrogen bonds (solid blue lines). (**B**) The geometry of the backbone brackets motif resembles brackets encircling the ligand. (**C**) WebLogo [49] representation of the sequence of backbone brackets residues (274 and 1361) and three surrounding sequence positions. (**D**) Structural representation of the arginine tweezers motif in interaction with Lysyl-5'AMP ligand in LysRS (PDB:1e1t chain A). Salt bridges (yellow dashed lines) as well as π-cation interactions are established. (**E**) The arginine tweezers geometry mimics a pair of tweezers grasping the ligand. (**F**) Sequence of arginine tweezers residues (698 and 1786) and surrounding sequence positions. The backbone brackets show nearly no conservation on sequence level since backbone interactions can be established by all amino acids, while the arginine tweezers rely on salt bridge interactions, always mediated by two arginines.

Interaction patterns

Contacts between ligand and protein are established via a variety of noncovalent interaction types such as hydrogen bonds, π-stacking, or salt bridges. These interaction types were annotated using the Protein-Ligand Interaction Profiler (PLIP) [50] to investigate whether evolution adapted entirely different strategies or if some characteristics are shared between both aaRS classes.

Two sets of 29 and 40 representative complexes for class I and class II were composed to analyze adenosine phosphate-binding. For the comparison of commonly interacting residues between different aaRS types a matrix visualization was designed (Fig 4). This allows for the assessment of interaction preferences at residue level. Data for frequent interactions was available for 16 residues and 10 different aaRS types for class I as well as 13 residues and 11 aaRS types for class II.

Fig 4. Protein-ligand contacts in representative adenosine phosphate-binding complexes for aaRS class I and II. Residues are grouped according to the non-amino acid ligand fragment (phosphate, ribose, or adenine) they are interacting with. Preferred interaction types for each aaRS type and binding site residue are color-coded. Fields separated with triangles indicate two equally preferred interactions. The asterisk (*) indicates aaRS types incorporating noncanonical amino acids. Automatically retrieved [51, 52] mutation effects [53–59] are shown as centered shapes. In essence, class I interactions are mainly hydrogen bonds, while class II adenosine phosphate-binding is realized by an array of different interaction types.

While six different interaction types are used to bind the adenosine phosphate ligand, hydrogen bonds are the prevalent type of contact, especially for the recognition of the ribose moiety (see Fig 4). The aromatic ring system of adenine is recognized via hydrogen bonds and π-stacking interactions in complexes of both classes I and II. Class II aaRS bind this part of the ligand also forming π-cation interactions with the charge provided by one guanidinium group of the arginine tweezers (residue 1786). Residue 698 interacts predominantly with the negatively charged phosphate group of the ligand via salt bridges. This binding pattern is conserved in class II and handled by the other guanidinium group featured by the arginine tweezers. In class I, hydrogen bonding is essential for the recognition of phosphate. Here, residue 274 binds to the phosphate and is part of the backbone brackets motif which embraces the phosphate and the aromatic ring at the other end (residue 1361) using backbone hydrogen bonding.

Both motifs share the tendency to form electrostatic interactions with the α-phosphate of the ligand. In general, the phosphate group predominantly participates

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

in salt bridges and hydrogen bonds. The ribose moiety is almost exclusively stabilized by hydrogen bonds to its hydroxyl groups.

Geometric characterization

Backbone brackets and arginine tweezers were analyzed at the geometrical level to further substantiate the profound differences in adenosine phosphate recognition. One would expect the side chains of the backbone brackets' residues to exhibit higher degrees of freedom in comparison to the arginine tweezers. Furthermore, a significant change in alpha carbon distance of both motif residues would indicate a conformational change during ligand binding. The state complexed with adenosine phosphate (M1) and the state in which no adenosine phosphate is bound (M2) were analyzed separately in order to quantify these aspects (see S1 Fig). Structure alignments of both motifs in respect to their binding modes (provided in S7 Fig) visually support the differences in side chain orientation and variable amino acid composition of the backbone brackets.

The angle between side chains of the backbone brackets is continuously high (SD_{M1}=20.93°, SD_{M2}=20.13°) with averages of 144.90° for M1 and 141.40° for M2, respectively. This emphasizes that the side chain orientation is indistinguishable between M1 and M2 as only the backbone participates in ligand binding. The alpha carbon distance is conserved for the majority of the backbone brackets observations (SD_{M1}=0.86 Å, SD_{M2}=0.82 Å), with averages of 17.92 Å for M1 and 18.41 Å for M2, respectively. However, some observations (structures PDB:5v0i chain A, PDB:1jzq chain A, PDB:3tzl chain A, PDB:3ts1 chain A) exhibit higher alpha carbon distances of 20.54 Å, 19.74 Å, 19.10 Å, and 18.79 Å, respectively. In contrast, one occurrence of the backbone brackets motif in structure PDB:4aq7 chain A has a remarkably low alpha carbon distance of 16.50 Å. Nevertheless, alpha carbon distances between bound and unbound state differ significantly (p<0.01, S5 Fig). This indicates the substantial contribution of backbone interactions as well as the conformational change observed during adenosine phosphate-binding.

The side chain variation is marginal for the arginine tweezers motif if an adenosine phosphate ligand is bound. In contrast, the side chain angle of the apo form is highly variable (SD_{M1}=8.69°, SD_{M2}=21.67°) with averages of 91.82° for M1 and 79.81° for M2, respectively. The side chain angles between the bound and unbound state differ significantly (p<0.01, S6 Fig), reinforcing the pivotal role of highly specific side chain interactions during ligand binding. This effect cannot be observed for the alpha carbon distances of the arginine tweezers (SD_{M1}=0.66 Å, SD_{M2}=0.79 Å), with averages of 14.76 Å for M1 and 14.93 Å for M2, respectively.

Fig 5. Geometric analysis of the ligand recognition motifs responsible for the adenosine phosphate interaction for aaRS class I and class II representative and nonredundant structures. The alpha carbon distance is plotted in respect to the side chain angle θ. Binding modes refer to states containing an adenosine phosphate ligand (M1) or not (M2). Backbone brackets in M1 allow for minor variance with respect to their alpha carbon distance, constrained by the position of the bound ligand. In contrast, arginine tweezer in M1 adapt an orthogonal orientation in order to fixate the ligand.

Effect of mutagenesis experiments and natural variants

To estimate the importance of certain ligand interactions, one can exploit data derived from mutagenesis experiments and natural variants.

Fig 4 shows the effect of nine mutations on the enzymatic activity of aaRS. There is no obvious link between conserved interactions and outcomes of mutations. For

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

example, there are loss of function mutations occurring in regions with observed interactions and equally many cases where no interactions could be observed while the mutation still has a negative effect.

For class I TyrRS, mutations of any histidine of the HIGH motif [44] lead to a decrease in activity, since both residues contribute to the stabilization of the transition state of the reaction [53, 54]. The same holds true for Asp-1300 and Gln-1301 which interact with the ribose part of the ligand [57, 59]. Phe-1415 is part of a sequence motif responsible for ligand attachment to the tRNA; a mutation to leucine shows no effect on the binding affinity [59].

Cys-1458 in class II AlaRS is part of a four residue zinc-binding motif [60] and an exchange with serine results in no effect whatsoever. It is assumed that the other three amino acids can compensate the mutation [56]. The single-nucleotide polymorphism (SNP) with no known effect is associated to position 1703 in AspRS (rs1803165 in dbSNP [61]).

Ile-703 in class II GlyRS does not directly interact with the ligand – mutations, however, result in a negative effect and are most prominently linked to CMT disease as the amino acid is crucial for tRNA ligation [55]. Another SNP occurs at Gly-1783; the exchange with arginine prohibits ligand binding and was tied to a loss of activity as well as distal hereditary motor neuropathy type VA [58].

Relations to known sequence motifs

Fig 6 encompasses structure and sequence motifs as well as the sequence conservation scores of the underlying MSA. Amino acids interacting with the adenosine phosphate of the ligand (ordinate in Fig 4) are annotated.

For class I sequence motifs [22, 44, 62], the HIGH motif features sequence conservation and is located nine positions downstream of the first backbone brackets residue. The KMSKS motif exhibits no sequence conservation and can be observed downstream of the second backbone brackets residue. The five-residue motif is distributed within a corridor of around 70 aligned sequence positions and contains ligand binding site residues 1414, 1415, 1434, and 1441.

For the class II sequence motifs [22, 33, 45, 46, 62], motif "1" is moderately conserved in sequence. None of its positions interact with the ligand. Motif "2" is conserved around the first arginine tweezers residue and contains five additional ligand binding site residue of lower sequence conservation. The second arginine tweezers residue is part of motif "3" which exhibits high sequence conservation.

Further ligand binding site residues, which are not part of known sequence motifs, are mostly occurring in sequence conserved regions.

Fig 6. Integrative sequence view for aaRS class I (**A**) and II (**B**). Boxes delineate sequence motifs previously described in literature [44–46]. The trace depicts the sequence conservation score of each position in the MSA (S5 File and S6 File). These scores were computed with Jalview [27, 63], positions composed of sets of amino acids with similar characteristics result in high values. Furthermore all positions relevant for ligand binding (Fig 4) are depicted. Backbone brackets and arginine tweezers have been emphasized by their respective pictograms. Positions of low conservation or those not encompassed by sequence motifs were intangible to studies primarily based on sequence data. Especially backbone interactions might be conserved independently from sequence.

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

Conservation of function The self-referencing system of building blocks and building machinery implemented in aaRS is an intriguing aspect of the early development of living systems. There is evidence that proteins arose from an ancient set of peptides [64] and that these peptides were co-factors of the early genetic information processing by RNA.

Sequence-based analyses were one of the first tools to investigate the transfer of genetic information. DNA and protein sequences comprise the developmental history of organisms, their specialization and diversification [34]. However, following the "functionalist" principle in biology, sequence is less conserved than structure, which is in turn less conserved than function [65]. Later, structural features and molecular contacts have been recognized as key aspects in grasping protein function [66, 67] and their evolution. Only if the necessary function can be maintained by compatible interaction architectures, the global role of the protein in the complex cellular system is ensured [68].

Each amino acid of a protein fulfills a certain role and can be replaced by amino acids with compatible attributes [66]. By considering each amino acid in the context of its sequence, its structural surroundings, and finally its biological function, one can determine possible exchanges and the evolutionary pressure driving these changes [65, 69]. Up to this point, pure sequence or structure analysis methods – without including ligand interaction data – missed the functional relevance of the backbone brackets.

Evolutionary model Since the amino acid alphabet imposes constraints on evolutionary divergence and functional conservation a big part of the aaRS specificity developed before the separation of archaea and prokaryotes [47] and branching of prokaryotic and eukaryotic lineages [29]. This suggests that the translation apparatus evolved prior to the "Darwinian threshold" [47]. Additionally, there is evidence that bacteria are closer to origin of translation than archaea [18]. The analysis of aaRS phylogeny results in ancestries resembling the standard model of evolution, where horizontal gene transfer is considered as an important evolutionary phenomenon in early stages of life [21]. Even if there is a strong homology within the catalytic core of aaRS classes – but not between them [29, 47] – both classes evolved convergently towards the same function [29]. This is reflected in the intrinsically different but functionally identical nature of the backbone brackets and the arginine tweezers.

Backbone brackets and arginine tweezers

The analysis of backbone brackets geometry showed a high variance of side chain angles for both binding modes. The distinction between these modes is significantly manifested in a change of the alpha carbon distance, which supports that the conformational change during ligand binding previously observed in ArgRS [70] or TyrRS [71] is a general mechanism in class I aaRS.

In contrast to backbone brackets, the arginine tweezers are highly conserved in side chain orientation if a ligand is bound, which shows that this orientation is key to adenosine phosphate recognition. If no ligand is bound, the tweezers geometry is less limited, which is reflected in a higher variability of side chain orientations. Conclusively, the distinction between the two binding modes can be made by taking the geometry of the motifs into account: alpha carbon distances for backbone brackets and side chain angles for arginine tweezers.

The conserved arginine tweezers motif resembles a common interaction pattern for phosphate recognition [66], which usually features positively charged amino acids [72].

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

However, the conformational space of ATP ligands was shown to be large throughout diverse superfamilies [73] and hence the geometry of binding sites involved in ATP recognition is manifold. The uniqueness of aaRS compared to other ATP-binding proteins was shown for AspRS, where the ligand binds in a compact form with a bent phosphate tail instead of the usually found extended form [73]. This conformation of ATP is energetically unfavorable but allows easy access of the α-phosphate for tRNA binding [74]. This suggests that the arginine tweezers motif possesses a unique geometry and is not a generalizable pattern for ATP binding, such as the frequently occurring P-loop domain [72].

As the function of fixating the adenosine phosphate part is crucial in aaRS enzymes, mutations of the arginine tweezers residues result in loss of function [75, 76]. However, to our knowledge, the backbone brackets motif was not identified in earlier literature and is herein described for the first time. Both motifs reveal a key mechanism in aaRS ligand binding, which seems to have evolved convergently, but shows highly divergent implementations. In agreement to this, the whole catalytic core of aaRS is class specific and the most ancient part of all aaRS domains known today (e.g. editing domains, codon binding domains, or organism-specific modular domain attachments) [47, 77, 78].

The stunning balance of evolutionary diversification [77] and functional convergence is underlined by profoundly different implementation of ligand recognition in terms of adjacent sequence (Fig 3C and 3F), embedding secondary structure elements (S3 Fig and S4 Fig), geometrical properties (Fig 5), and interaction characteristics (Fig 4).

Backbone brackets are not conserved in sequence

The backbone brackets are remarkable, since backbone interactions are often neglected in structural studies. Nevertheless, backbone hydrogen bonds make up at least one-quarter of overall ligand hydrogen bonding [79]. In these cases, side chain properties may only play a minor role, e.g. for steric effects, and allow for larger flexibility in implementation of a binding pattern as long as the correct backbone orientation is ensured. There are examples of protein-ligand complexes where backbone hydrogen bonds are a major part of the binding mechanism, e.g. in binding of the cofactor NAD to a CysG protein from *Salmonella enterica* (PDB:1pjs) as determined with the PLIP [50]. In conclusion, the backbone brackets exhibit conservation on functional level rather than on sequence level, which renders sequence-based motif analysis infeasible. This motif is a prime example for conservation of function over structure or sequence [65].

Directed evolution

The dataset does not only describe the aaRS binding characteristics for adenosine phosphate but also provides a starting point to study the recognition of the cognate amino acid. Previous studies investigated how AlaRS prevents incorrect tRNA charging with glycine or serine [80] or how discrimination between aromatic amino acids is implemented [81], even without pre- or post-editing mechanisms involved. Insights into the toolkit of evolution are presented by research regarding the distinction between valine and threonine [4], tryptophan and tyrosine [5], glutamine and glutamic acid [6], or asparagine and aspartic acid [7]. We envision the possibility to unmask specific interaction patterns for each aaRS type. These patterns could serve as a starting point to design noncanonical aaRS enzymes by directed evolution experiments, which potentially enable the artificial extension of the genetic code [1].

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Disease implications

Due to the fundamental role of aaRS for protein biosynthesis, a systematic assessment of mutation effects in yeast was conducted by Cavarelli and coworkers [75]. Mutations of aaRS-coding genes can be drastic and may result in a variety of human diseases, even if the structural effect is unknown [82, 83].

Structural analysis of a GlyRS mutant (G526R) showed that CMT may be caused by blockage of the AMP binding site. Furthermore, this mutation induces a larger contact area in the homo-dimer interface, which stems partially from the anticodon binding domain [58]. Other mutations result in a wider range of diseases and symptoms such as hearing loss, ovarian failure, or cardiomyopathy [40, 84]. Even for cellular processes unrelated to translation, aaRS play a pivotal role, e.g. for angiogenesis [78]. Due to the highly individual characteristics of aaRS enzymes between organisms, it is possible to create precisely targeted antibiotics with minimal side effects [85–87].

Unfortunately, mutational data data does not cover the backbone brackets or arginine tweezers motif. One would expect mutations of the arginine tweezers to cause a strong decrease in enzyme activity. In contrast, the backbone brackets should be more resilient to mutational events. However, bridging the gap between mutational studies and key interaction patterns will require further analysis beyond this study and should be substantiated by *in vitro* experiments. The provided high-quality aaRS dataset can serve as the basis for such work.

Limitations

The method used to unify residue numbering in all structures relies on both, the quality of the used MSA as well as the quality of local structure regions. Hence, the backbone brackets and arginine tweezers could not be successfully mapped to all structures in the dataset. On the one hand, occasionally some binding site regions were not experimentally determined (e.g. PDB:3hri) or the mapping of the motif residues failed (e.g. PDB:4yrc) due to ambivalent regions in the MSA. On the other hand, some aaRS may have evolved different strategies to bind the ligand, even for the same aaRS type [88].

However, the conserved ligand interactions could be related to known sequence motifs (Fig 6). The sequentially high variance of the KMSKS-motif was described before [44] and explains why the MSA features many inserts in this region. The interacting residues 1352, 1360, and 1361 of class I are located upstream of the KMSKS motif. In case of class I the AIDQ-motif in TrpRS is known [82], yet no consensus for all types could be established. Class II sequence motifs exhibit high degeneracy and could not be identified without structural information [77]. Motif "1" is the only sequence motif to which could not be linked to any relevant ligand interaction site; its primary role lies in the stabilization of class II dimers [45].

The geometric characterization of the two ligand recognition motifs (see Fig 5) highlighted some observations of the backbone brackets, which exhibit a substantial increase or decrease of the residue alpha carbon distance. Chain A of an LeuRS of *Escherichia coli* (PDB:4aq7) is complexed with tRNA and the backbone brackets alpha carbon distance is about 1 Å below the average. Manual investigation of this structure showed that there is no obvious conformational difference to other structures. Likewise, the annotated interactions were checked for consistency using PLIP and showed usual interactions with the adenine and the sulfamate group (the phosphate analogue) of the ligand. For the backbone brackets with higher alpha carbon extent (structures of IleRS, TrpRS, and TyrRS) interaction analysis revealed that residue 274 interacts with the amino acid side chain, as all of these structures contain a single aminoacyl ligand (PDB:3tzl chain A, PDB:3ts1 chain A, PDB:1jzq chain A) or two separate ligands

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(amino acid and AMP, PDB:5v0i chain A). This suggests that the structures resemble a partially changed conformation prior to tRNA ligation and a possible role of the backbone brackets motif in amino acid recognition. Likewise, these effects could arise from low quality electron density maps in the structure regions of interest. However, these hypotheses should be addressed and validated in future work.

Interestingly, our analysis did not reveal a high count or conservation of interactions established with the well-known HIGH motif in class I. Despite irregularly occurring salt bridges, hydrogen bonds, and one π-cation interaction in GluRS (see Fig 4A), no interactions could be observed. This especially holds true for the first histidine residue of the HIGH motif, which only interacts with the ligand in GluRS. However, it was shown that the HIGH motif is mainly relevant for binding in the pre-acylation transition state of the reaction [44], i.e. HIGH interacts with the phosphate of ATP. This explains the irregular observations of interactions which are established only if an ATP ligand is present (e.g. GluRS PDB:1j09 chain A renumbered residue 281).

Prospects

Adaptations of the presented workflow to other protein families of interest, might allow to study binding mechanisms in a new level of detail and by using publicly available data alone. Even if the geometric characterization is dependent on the quality of local structure regions, the comparison of alpha carbon distances and side chain angles could be a simple yet valuable tool to separate different binding states and quantify levels of evolutionary divergence. Geometrical properties can reveal the importance of conserved side chain orientations, the degree of freedom in unbound state, or shifts in backbone arrangement. However, choosing these two properties to compare residue binding motifs depends on the specific use case.

In a similar way, the obtained interaction data proved as a valuable resource to understand fundamental aspects of aaRS ligand recognition. Despite the fact that interactions can not be determined for apo structures and do not take into consideration the dynamic nature of enzyme reactions, both, structure and interaction data conflates several aspects of evolution and proved to outperform pure sequence-based methods.

The designed approach was used to analyze aaRS from the different viewpoints: sequence backed by structure information, ligand interactions, and geometric characterization of essential ligand binding patterns. Additionally, this study provides the largest manually curated dataset of aaRS structures including ligand information available to date. This can serve as foundation for further research on the essential mechanisms controlling the molecular information machinery, e.g. investigate the effect and disease implications of mutations on crucial binding site residues.

Alongside the aaRS-specific results, the workflow could be a general tool for identification of significant ligand binding patterns and geometrical characterization of such. Further studies may adapt the presented methodology to study common mechanisms in highly variable implementations of ligand binding, i.e. for nonribosomal peptide synthetases as another enzyme family that is required to recognize all 20 amino acids [89].

Materials and methods

Dataset preparation

Proteins with domains annotated to belong to aaRS families according to Pfam 31.0 [90] were selected (see S1 Appendix for a detailed list of Pfam identifiers) and their structures were retrieved from PDB. Additionally, structures with Enzyme Commission

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(EC) number 6.1.1.- were considered and included in the initial dataset. Structures with putative aaRS function were excluded.

For each catalytic chain the aaRS class and type, resolution, mutational status, the taxonomy identifier of the organism of origin, and its superkingdom were determined. For chains where a ligand was present, these ligands were added to the dataset and it was decided if this ligand is either relevant for amino acid recognition (i.e. contains an amino acid or a close derivate as substructure), for adenosine phosphate-binding (i.e. contains an adenosine phosphate substructure), or for both (e.g. aminoacyl-AMP).

As the presented study focuses on the binding of the adenosine phosphate moiety, two binding modes referred to as M1 and M2 (S1 Fig) were defined. M1 features an adenosine phosphate-containing ligand (e.g. aminoacyl-AMP, ATP), whereas M2 does not contain any ligand that binds to the adenosine phosphate recognition region of the binding pocket (e.g. plain amino acid, empty pocket).

To avoid the use of highly redundant structures for analysis, all structures in the dataset were clustered according to >95% sequence identity using Needleman-Wunsch [91] alignments and single-linkage clustering. For each of these clusters a representative chain (selection scheme listed in S2 Appendix) was determined. The same procedure was used to define representative chains for the adenosine phosphate bound state M1 and no adenosine phosphate bound state M2. The final dataset is provided as formatted table in S1 File and as machine-readable JSON version in S2 File.

Mapping of binding sites

To allow a unified mapping of aaRS binding sites an MSA of 81 (75) representative wild type sequences of class I (class II) (S3 File and S4 File) aaRS was performed. The alignment was calculated with the T-Coffee expresso pipeline [92], which guides the alignment by structural information. Using the obtained MSA (S5 File and S6 File), residues in all aaRS structures were renumbered with the custom script "MSA PDB Renumber", available under open-source license (MIT) at `github.com/vjhaupt`. All renumbered structures are provided in PDB file format (S11 File and S12 File). Only protein residues were renumbered, while chain identifiers and residue numbers of ligands were left unmodified. Lists of structures where the backbone brackets or arginine tweezers could not be mapped are found in S8 File and S10 File.

Annotation of noncovalent interactions

Annotation of noncovalent interactions between an aaRS protein and its bound ligand(s) was performed with the PLIP [50] command line tool v1.3.3 on all renumbered structures with default settings. The renumbered sequence positions of all residues observed to be in contact with the ligand were extracted. This resulting set of interacting residues was used to determine the position-identical residues from all aaRS structures in the dataset, even if no ligand is bound.

Generation of interaction matrix

Information on noncovalent protein-ligand interactions from renumbered structure files (see above) was used to prepare separate interaction matrices for aaRS classes I and II. First, only representative structures for M1 were selected. Second, only residues which are in contacts with the non-amino acid part of the ligand (i.e. adenine, ribose moiety or the phosphate group) were considered. This was validated manually for each residue. Furthermore, residues relevant for only one aaRS type were discarded. For each considered residue, the absolute frequency of observed ligand interactions was

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

determined with respect to the PLIP interaction types (hydrophobic contacts, hydrogen bonds, salt bridges, π-stacking, and π-cation interactions). Additionally, the count of residues not interacting with any ligand ("no contact") was determined. In the interaction matrix (Fig 4), aaRS types are placed on the abscissa and renumbered residue positions on the ordinate. The preferred interaction type for each residue and ligand species is color-coded. If two interaction types occurred with the same frequency, a dual coloring was used. Residues were grouped in the figure according to the ligand fragment they are mainly forming interactions with.

Annotation of mutagenesis sites and natural variants

For each chain a mapping to UniProt [51] was performed using the SIFTS project [52]. Where available, mutation and natural variants data was retrieved for all binding site residues from the UniProt [51] database. In total, 32 mutagenesis sites and 8 natural variants were retrieved.

Analysis of core-interaction patterns

All motif occurrences in M1 and M2 representative chains were aligned in respect on their backbone atoms (S7 Fig) using the Fit3D algorithm [93]. Additionally, the alpha carbon distances and the angle between side chains were determined. The side chain angle θ between two residues was calculated by abstracting each side chain as a vector between alpha carbon and the most distant carbon side chain atom. If $\theta = 0°$ or $\theta = 180°$ the side chains are oriented in a parallel way. Side chain angles could not be calculated if one or both residues of the backbone brackets motif were glycine.

Furthermore, the sequential neighbors of the core-interaction patterns have been visualized with WebLogo graphics [49] regarding their sequence and secondary structure elements. Secondary structure elements were assigned according to the rule set of DSSP [94].

Acknowledgments

The authors thank Peter R Wills for initially approaching us with the intriguing topic of the origin of genetic coding. Further, we appreciated the continuous meetings and are grateful for guiding us the way through the entire project. Gratitude is owed to Hanna Siewerts and Alexander Eisold for proofreading the manuscript.

References

1. Mukai T, Lajoie MJ, Englert M, Soll D. Rewriting the Genetic Code. Annu Rev Microbiol. 2017;.

2. Lee JH, Choi SK, Roll-Mecak A, Burley SK, Dever TE. Universal conservation in translation initiation revealed by human and archaeal homologs of bacterial translation initiation factor IF2. Proc Natl Acad Sci USA. 1999;96(8):4342–4347.

3. Fox GE. Origin and evolution of the ribosome. Cold Spring Harb Perspect Biol. 2010;2(9):a003483.

4. Dock-Bregeon A, Sankaranarayanan R, Romby P, Caillet J, Springer M, Rees B, et al. Transfer RNA-mediated editing in threonyl-tRNA synthetase. The class II solution to the double discrimination problem. Cell. 2000;103(6):877–884.

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

5. Ibba M, Söll D. Aminoacyl-tRNA synthesis. Annu Rev Biochem. 2000;69:617–650.

6. Hadd A, Perona JJ. Coevolution of specificity determinants in eukaryotic glutamyl-and glutaminyl-tRNA synthetases. Journal of molecular biology. 2014;426(21):3619–3633.

7. Nair N, Raff H, Islam MT, Feen M, Garofalo DM, Sheppard K. The Bacillus subtilis and Bacillus halodurans Aspartyl-tRNA Synthetases Retain Recognition of tRNA Asn. Journal of molecular biology. 2016;428(3):618–630.

8. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235–242.

9. Di Giulio M. The origin of the genetic code: theories and their relationships, a review. Biosystems. 2005;80(2):175–184.

10. Wong JT. A co-evolution theory of the genetic code. Proc Natl Acad Sci USA. 1975;72(5):1909–1912.

11. Sonneborn T. Degeneracy of the genetic code: extent, nature, and genetic implications. Evolving genes and proteins. 1965; p. 377–397.

12. Woese CR. Order in the genetic code. Proceedings of the National Academy of Sciences. 1965;54(1):71–75.

13. Guimarães RC, Moreira CHC, de Farias ST. A self-referential model for the formation of the genetic code. Theory in Biosciences. 2008;127(3):249.

14. Wong JT. Coevolution theory of the genetic code at age thirty. Bioessays. 2005;27(4):416–425.

15. Wills PR. The generation of meaningful information in molecular systems. Phil Trans R Soc A. 2016;374(2063):20150066.

16. Brown JR, Doolittle WF. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proceedings of the National Academy of Sciences. 1995;92(7):2441–2445.

17. Schimmel P, Giege R, Moras D, Yokoyama S. An operational RNA code for amino acids and possible relationship to genetic code. Proceedings of the National Academy of Sciences. 1993;90(19):8763–8768.

18. Chandrasekaran SN, Yardimci GG, Erdogan O, Roach J, Carter CW. Statistical evaluation of the Rodin-Ohno hypothesis: sense/antisense coding of ancestral class I and II aminoacyl-tRNA synthetases. Mol Biol Evol. 2013;30(7):1588–1604.

19. Rodin SN, Ohno S. Two types of aminoacyl-tRNA synthetases could be originally encoded by complementary strands of the same nucleic acid. Orig Life Evol Biosph. 1995;25(6):565–589.

20. Eriani G, Delarue M, Poch O, Gangloff J, Moras D. Partition of tRNA Synthetases into Two Classes Based on Mutually Exclusive Sets of Sequence Motifs. Nature. 1990;347(6289):203.

21. Wolf YI, Aravind L, Grishin NV, Koonin EV. Evolution of aminoacyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome research. 1999;9(8):689–710.

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

22. Moras D. Structural and functional relationships between aminoacyl-tRNA synthetases. Trends Biochem Sci. 1992;17(4):159–164.

23. Burbaum JJ, Schimmel P. Structural relationships and the classification of aminoacyl-tRNA synthetases. J Biol Chem. 1991;266(26):16965–16968.

24. de Pouplana LR, Schimmel P. Aminoacyl-tRNA synthetases: potential markers of genetic code development. Trends in biochemical sciences. 2001;26(10):591–596.

25. Schulze JO, Masoumi A, Nickel D, Jahn M, Jahn D, Schubert WD, et al. Crystal structure of a non-discriminating glutamyl-tRNA synthetase. Journal of molecular biology. 2006;361(5):888–897.

26. Mailu BM, Ramasamay G, Mudeppa DG, Li L, Lindner SE, Peterson MJ, et al. A nondiscriminating glutamyl-tRNA synthetase in the Plasmodium apicoplast the first enzyme in an indirect aminoacylation pathway. Journal of Biological Chemistry. 2013;288(45):32539–32552.

27. Livingstone CD, Barton GJ. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput Appl Biosci. 1993;9(6):745–756.

28. Ambrogelly A, Söll D, Nureki O, Yokoyama S, Ibba M. Class I Lysyl-tRNA Synthetases. Landes Bioscience; 2013.

29. Nagel GM, Doolittle RF. Evolution and relatedness in two aminoacyl-tRNA synthetase families. Proceedings of the National Academy of Sciences. 1991;88(18):8121–8125.

30. Jakubowski H, Goldman E. Editing of errors in selection of amino acids for protein synthesis. Microbiological reviews. 1992;56(3):412–429.

31. Arnez JG, Moras D. Structural and functional considerations of the aminoacylation reaction. Trends in biochemical sciences. 1997;22(6):211–216.

32. Cusack S, Berthet-Colominas C, Hartlein M, Nassar N, Leberman R. A second class of synthetase structure revealed by X-ray analysis of Escherichia coli seryl-tRNA synthetase at 2.5 A. Nature. 1990;347(6290):249.

33. Cusack S, Hartlein M, Leberman R. Sequence, structural and evolutionary relationships between class 2 aminoacyl-tRNA synthetases. Nucleic Acids Res. 1991;19(13):3489–3498.

34. Woese CR, Olsen GJ, Ibba M, Söll D. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiology and Molecular Biology Reviews. 2000;64(1):202–236.

35. Schimmel P, Ripmaster T. Modular design of components of the operational RNA code for alanine in evolution. Trends in biochemical sciences. 1995;20(9):333–334.

36. Rould M, Perona J, Steitz T. Structural basis of anticodon loop recognition by glutaminyl-tRNA synthetase. Nature. 1991;352(6332):213.

37. Normanly J, Abelson J. tRNA identity. Annual review of biochemistry. 1989;58(1):1029–1049.

38. Martinis SA, Boniecki MT. The balance between pre- and post-transfer editing in tRNA synthetases. FEBS Lett. 2010;584(2):455–459.

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

39. Splan KE, Ignatov ME, Musier-Forsyth K. Transfer RNA modulates the editing mechanism used by class II prolyl-tRNA synthetase. J Biol Chem. 2008;283(11):7128–7134.

40. Datt M, Sharma A. Evolutionary and structural annotation of disease-associated mutations in human aminoacyl-tRNA synthetases. BMC Genomics. 2014;15:1063.

41. Motley WW, Griffin LB, Mademan I, Baets J, De Vriendt E, De Jonghe P, et al. A novel AARS mutation in a family with dominant myeloneuropathy. Neurology. 2015;84(20):2040–2047.

42. Diaz-Lazcoz Y, Aude J, Nitschke P, Chiapello H, Landes-Devauchelle C, Risler J. Evolution of genes, evolution of species: the case of aminoacyl-tRNA synthetases. Molecular biology and evolution. 1998;15(11):1548–1561.

43. Chaliotis A, Vlastaridis P, Mossialos D, Ibba M, Becker HD, Stathopoulos C, et al. The complex evolutionary history of aminoacyl-tRNA synthetases. Nucleic Acids Res. 2017;45(3):1059–1068.

44. Schmitt E, Panvert M, Blanquet S, Mechulam Y. Transition state stabilization by the 'high'motif of class I aminoacyl-tRNA synthetases: the case of Escherichia coli methionyl-tRNA synthetase. Nucleic acids research. 1995;23(23):4793–4798.

45. Åberg A, Yaremchuk A, Tukalo M, Rasmussen B, Cusack S. Crystal Structure Analysis of the Activation of Histidine by Thermus thermophilus Histidyl-tRNA Synthetase. Biochemistry. 1997;36(11):3084–3094.

46. Cusack S. Aminoacyl-tRNA synthetases. Current opinion in structural biology. 1997;7(6):881–889.

47. O'Donoghue P, Luthey-Schulten Z. On the evolution of structure in aminoacyl-tRNA synthetases. Microbiology and Molecular Biology Reviews. 2003;67(4):550–573.

48. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33(7):2302–2309.

49. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188–1190.

50. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M. PLIP: fully automated protein-ligand interaction profiler. Nucleic Acids Res. 2015;43(W1):W443–447.

51. Consortium TU. UniProt: the universal protein knowledgebase. Nucleic Acids Research. 2017;45(D1):D158. doi:10.1093/nar/gkw1099.

52. Velankar S, Dana JM, Jacobsen J, van Ginkel G, Gane PJ, Luo J, et al. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Research. 2013;41(D1):D483. doi:10.1093/nar/gks1258.

53. Vidal-Cros A, Bedouelle H. Role of residue Glu152 in the discrimination between transfer RNAs by tyrosyl-tRNA synthetase from Bacillus stearothermophilus. J Mol Biol. 1992;223(3):801–810.

54. Xin Y, Li W, Dwyer DS, First EA. Correlating amino acid conservation with function in tyrosyl-tRNA synthetase. J Mol Biol. 2000;303(2):287–298.

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

55. Griffin LB, Sakaguchi R, McGuigan D, Gonzalez MA, Searby C, Zuchner S, et al. Impaired function is a common feature of neuropathy-associated glycyl-tRNA synthetase mutations. Hum Mutat. 2014;35(11):1363–1371.

56. Miller WT, Hill KA, Schimmel P. Evidence for a "cysteine-histidine box" metal-binding site in an Escherichia coli aminoacyl-tRNA synthetase. Biochemistry. 1991;30(28):6970–6976.

57. Xin Y, Li W, First EA. Stabilization of the transition state for the transfer of tyrosine to tRNA(Tyr) by tyrosyl-tRNA synthetase. J Mol Biol. 2000;303(2):299–310.

58. Xie W, Nangle LA, Zhang W, Schimmel P, Yang XL. Long-range structural effects of a Charcot-Marie-Tooth disease-causing mutation in human glycyl-tRNA synthetase. Proceedings of the National Academy of Sciences. 2007;104(24):9976–9981.

59. Xin Y, Li W, First EA. The 'KMSKS' motif in tyrosyl-tRNA synthetase participates in the initial binding of tRNA(Tyr). Biochemistry. 2000;39(2):340–347.

60. Berg JM. Potential metal-binding domains in nucleic acid binding proteins. Science. 1986;232(4749):485–487.

61. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–311.

62. Carter CW. Cognition, mechanism, and evolutionary relationships in aminoacyl-tRNA synthetases. Annu Rev Biochem. 1993;62:715–748.

63. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25(9):1189–1191.

64. Alva V, Soding J, Lupas AN. A vocabulary of ancient peptides at the origin of folded proteins. Elife. 2015;4:e09410.

65. Najmanovich RJ. Evolutionary studies of ligand binding sites in proteins. Curr Opin Struct Biol. 2016;45:85–90.

66. Gutteridge A, Thornton JM. Understanding nature's catalytic toolkit. Trends in biochemical sciences. 2005;30:622–629.

67. Salentin S, Haupt VJ, Daminelli S, Schroeder M. Polypharmacology rescored: protein-ligand interaction profiles for remote binding site similarity assessment. Prog Biophys Mol Biol. 2014;116(2-3):174–186.

68. Samish I, Bourne PE, Najmanovich RJ. Achievements and challenges in structural bioinformatics and computational biophysics. Bioinformatics. 2015;31(1):146–150.

69. Caetano-Anolles G, Wang M, Caetano-Anolles D, Mittenthal JE. The origin, evolution and structure of the protein world. Biochem J. 2009;417(3):621–637.

70. Delagoutte B, Moras D, Cavarelli J. tRNA aminoacylation by arginyl-tRNA synthetase: induced conformations during substrates binding. The EMBO Journal. 2000;19(21):5599–5610.

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

71. Kobayashi T, Takimura T, Sekine R, Kelly VP, Vincent K, Kamata K, et al. Structural snapshots of the KMSKS loop rearrangement for amino acid activation by bacterial tyrosyl-tRNA synthetase. J Mol Biol. 2005;346(1):105–117.

72. Barelier S, Sterling T, O'Meara MJ, Shoichet BK. The Recognition of Identical Ligands by Unrelated Proteins. ACS Chem Biol. 2015;10(12):2772–2784.

73. Stockwell GR, Thornton JM. Conformational diversity of ligands bound to proteins. J Mol Biol. 2006;356(4):928–944.

74. Schmitt E, Moulinier L, Fujiwara S, Imanaka T, Thierry JC, Moras D. Crystal structure of aspartyl-tRNA synthetase from Pyrococcus kodakaraensis KOD: archaeon specificity and catalytic mechanism of adenylate formation. EMBO J. 1998;17(17):5227–5237.

75. Cavarelli J, Eriani G, Rees B, Ruff M, Boeglin M, Mitschler A, et al. The active site of yeast aspartyl-tRNA synthetase: structural and functional aspects of the aminoacylation reaction. EMBO J. 1994;13(2):327–337.

76. Navarre WW, Zou SB, Roy H, Xie JL, Savchenko A, Singer A, et al. PoxA, yjeK, and elongation factor P coordinately modulate virulence and drug resistance in Salmonella enterica. Mol Cell. 2010;39(2):209–221.

77. Giege R, Springer M. Aminoacyl-tRNA Synthetases in the Bacterial World. EcoSal Plus. 2016;7(1).

78. Mirando AC, Francklyn CS, Lounsbury KM. Regulation of angiogenesis by aminoacyl-tRNA synthetases. Int J Mol Sci. 2014;15(12):23725–23748.

79. Gallina AM, Bork P, Bordo D. Structural analysis of protein-ligand interactions: the binding of endogenous compounds and of synthetic drugs. J Mol Recognit. 2014;27(2):65–72.

80. Guo M, Chong YE, Shapiro R, Beebe K, Yang XL, Schimmel P. Paradox of mistranslation of serine for alanine caused by AlaRS recognition dilemma. Nature. 2009;462(7274):808–812.

81. Yang XL, Otero FJ, Skene RJ, McRee DE, Schimmel P, Ribas de Pouplana L. Crystal structures that suggest late development of genetic code components for differentiating aromatic side chains. Proc Natl Acad Sci USA. 2003;100(26):15376–15380.

82. Guo LT, Chen XL, Zhao BT, Shi Y, Li W, Xue H, et al. Human tryptophanyl-tRNA synthetase is switched to a tRNA-dependent mode for tryptophan activation by mutations at V85 and I311. Nucleic Acids Res. 2007;35(17):5934–5943.

83. Simons C, Griffin LB, Helman G, Golas G, Pizzino A, Bloom M, et al. Loss-of-function alanyl-tRNA synthetase mutations cause an autosomal-recessive early-onset epileptic encephalopathy with persistent myelination defect. Am J Hum Genet. 2015;96(4):675–681.

84. Stum M, McLaughlin HM, Kleinbrink EL, Miers KE, Ackerman SL, Seburn KL, et al. An assessment of mechanisms underlying peripheral axonal degeneration caused by aminoacyl-tRNA synthetase mutations. Mol Cell Neurosci. 2011;46(2):432–443.

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

85. Randall CP, Rasina D, Jirgensons A, O'Neill AJ. Targeting Multiple Aminoacyl-tRNA Synthetases Overcomes the Resistance Liabilities Associated with Antibacterial Inhibitors Acting on a Single Such Enzyme. Antimicrob Agents Chemother. 2016;60(10):6359–6361.

86. Pham JS, Dawson KL, Jackson KE, Lim EE, Pasaje CF, Turner KE, et al. Aminoacyl-tRNA synthetases as drug targets in eukaryotic parasites. Int J Parasitol Drugs Drug Resist. 2014;4(1):1–13.

87. Chopra S, Palencia A, Virus C, Schulwitz S, Temple BR, Cusack S, et al. Structural characterization of antibiotic self-immunity tRNA synthetase in plant tumour biocontrol agent. Nat Commun. 2016;7:12928.

88. Merritt EA, Arakaki TL, Gillespie JR, Larson ET, Kelley A, Mueller N, et al. Crystal structures of trypanosomal histidyl-tRNA synthetase illuminate differences between eukaryotic and prokaryotic homologs. J Mol Biol. 2010;397(2):481–494.

89. Challis GL, Ravel J, Townsend CA. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. Chem Biol. 2000;7(3):211–224.

90. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44(D1):D279–285.

91. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48(3):443–453.

92. Armougom F, Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, et al. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res. 2006;34(Web Server issue):W604–608.

93. Kaiser F, Eisold A, Bittrich S, Labudde D. Fit3D: a web application for highly accurate screening of spatial residue patterns in protein structure data. Bioinformatics. 2016;32(5):792–794.

94. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577–2637. doi:10.1002/bip.360221211.

Supporting information

S1 Fig. Binding mode definition. Binding modes M1 and M2 are defined based on the complexed ligand: ligands that bind to the adenosine phosphate moiety (highlighted in red, only in contact when adenosine phosphate is part of the ligand) of the binding site (M1), no ligands or ligands that bind exclusively to the aminoacyl part (green) of the binding site (M2).

S2 Fig. Core-interaction patterns. Both aaRS classes contain highly conserved patterns, responsible for proper binding of the adenosine phosphate part of the ligand. Class I aaRS share a highly conserved set of backbone hydrogen interactions with the ligand: the backbone brackets. Class II active sites contain a pattern of two arginine residues grasping the adenosine phosphate ligand: the arginine tweezers. Interactions

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

were calculated with PLIP [50] and are represented with colored (dashed) lines: hydrogen bonds (solid, blue), π-stacking interactions (dashed, green), π-cation interactions (dashed, orange), salt bridges (dashed, yellow), metal complexes (dashed, purple), and hydrophobic contacts (dashed grey). (**A**) Class I backbone brackets motif and interactions with the ligand Tryptophanyl-5'AMP as observed in TrpRS structure PDB:1r6u chain A. (**B**) Class II arginine tweezers motif and interactions with the ligand Lysyl-5'AMP as observed in LysRS structure PDB:1e1t chain A.

S3 Fig. Secondary structure of backbone brackets adjacent residues. WebLogo [49] representation of secondary structure elements around the backbone brackets residues (274 and 1361) annotated by DSSP [94]: helices (blues), strands (red), and unordered (black). Unassigned states are represented by the character 'C'. The height of each character corresponds to the relative frequency.

S4 Fig. Secondary structure of arginine tweezers adjacent residues. WebLogo [49] representation of secondary structure elements around the arginine tweezers residues (698 and 1786) annotated by DSSP [94]: helices (blues), strands (red), and unordered (black). Unassigned states are represented by the letter 'C'. The height of each character corresponds to the relative frequency.

S5 Fig. Distributions of alpha carbon distances for backbone brackets and arginine tweezers. Distributions of alpha carbon distances for class I backbone brackets motif and class II arginine tweezers motif in adenosine phosphate bound (M1) and unbound state (M2). The alpha carbon distance of the backbone brackets differs significantly between the two states (Mann-Whitney U $p<0.01$).

S6 Fig. Distributions of side chain angles for backbone brackets and arginine tweezers. Distributions of side chain angle θ for class I backbone brackets motif and class II arginine tweezers motif in adenosine phosphate bound (M1) and unbound state (M2). The side chain angles of the arginine tweezers differs differs significantly between the two states (Mann-Whitney U $p<0.01$).

S7 Fig. Alignments of backbone brackets and arginine tweezers. Structural backbone-only alignments computed with Fit3D [93] of relevant binding site residue motifs derived from M1 and M2 representative structures in respect to their binding modes for aaRS class I and class II. (**A**) The class I backbone brackets motif aligned in respect to binding modes. A high side chain variance (gray line representation) is evident. However, backbone orientations are highly conserved to realize consistent hydrogen bond interaction with the adenosine phosphate part of the ligand. (**B**) The class II arginine tweezers motif aligned in respect to binding modes. Low side chain variance can be observed if an adenosine phosphate ligand is bound (M1), whereas the absence of an adenosine phosphate ligand (M2) allows an increased degree of freedom for side chain movement. Averaged backbone and side chain RMSD values after all-vs-all superimposition are shown in S1 Table.

S8 Fig. Pairwise sequence and structure similarity. Structure and sequence similarity for pairs of cluster representative chains for aaRS class I (**A**) and II (**B**). Depicted is the sequence similarity (% identity) after a global Needleman-Wunsch [91] alignment of both structures against the structure similarity determined by TMAlign [48]. For class I (class II) 95% of all pairs exhibit <33% (29%) sequence identity and <0.85 (0.84) TM score. The 95% quantile borders are depicted as red dashed lines.

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

S9 Fig. Origin organisms of aaRS class I and class II structures in the dataset. The organisms of origin for aaRS class I (**A**) and class II (**B**) structures in the dataset. The inner circles correspond to the superkingdom of the organism. The outer circle depicts the partition into specific species (combining different strains). Sections representing eukaryotic species are colored in violet, bacteria are colored in green, archaea are colored in orange and vira are colored in gray. Species, that are origin of less than two percent of the structures are condensed to the "other" cluster for each superkingdom. All superkingdoms are represented in both datasets. Class I contains more bacterial structures than class II, but fewer originating from eukaryotes or archaea. Interestingly, class I also contains one viral structure. Despite the diverse origins of the structures the conserved interaction patterns can be observed.

S1 Table. Backbone and side chain RMSD of backbone brackets and arginine tweezers after superimposition. Averaged backbone and side chain RMSD values after all-vs-all superimposition are shown in this table.

Table 1. Averaged backbone RMSD values of backbone brackets and arginine tweezers after superimposition.

motif	binding mode	observations	backbone RMSD [Å]
backbone brackets	M1	28	0.32
backbone brackets	M2	59	0.34
arginine tweezers	M1	39	0.24
arginine tweezers	M2	47	0.28

M1 contains an adenosine phosphate ligand, no adenosine phosphate ligand in M2.

S1 Appendix. Dataset preparation. All selected protein chains from the PDB carry one of the following protein family annotations, according to Pfam [90]: PF00133, PF00152, PF00579, PF00587, PF00749, PF00750, PF01406, PF01409, PF01411, PF02091, PF02403, PF02912, PF03485, PF09334. Additionally, structures annotated with an EC number indicating tRNA-ligation activity were selected: 6.1.1.1 (TyrRS), 6.1.1.2 (TrpRS), 6.1.1.3 (ThrRS), 6.1.1.4 (LeuRS), 6.1.1.5 (IleRS), 6.1.1.6 (LysRS), 6.1.1.7 (AlaRS), 6.1.1.9 (ValRS), 6.1.1.10 (MetRS), 6.1.1.11 (SerRS), 6.1.1.14 (GlyRS), 6.1.1.15 (ProRS), 6.1.1.16 (CysRS), 6.1.1.17 (GluRS), 6.1.1.18 (GlnRS), 6.1.1.19 (ArgRS), 6.1.1.20 (PheRS), 6.1.1.21 (HisRS), 6.1.1.22 (AsnRS), 6.1.1.23 (AspRS), 6.1.1.26 (PylRS).

For each of the resulting structures, the existence of a catalytic domain was checked manually and only the chains containing a domain with confirmed catalytic activity were retained. If there were ligands present in the structure, these ligands were annotated manually to avoid errors in the assignment of ligands to their catalytic chain. This procedure resulted in a high quality dataset of 972 individual aaRS chains containing a catalytic domain.

S2 Appendix. Selection of representative entries. In order to avoid redundancy, representatives were defined for each sequence cluster with >95% sequence identity and discriminated between three types: cluster representatives, representatives that contain an adenosine phosphate ligand (M1), and representatives that do not contain an adenosine phosphate ligand (M2).

The selection criteria for these categories were defined as follows:

- cluster representative:

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1. protein must be wild type (if wild type exists)
2. best resolution
3. longest sequence coverage

- representatives with an adenosine-relevant ligand
  1. chain must contain an adenosine phosphate ligand
  2. this ligand must be standard (adenosine phosphate or close derivate)
  3. no experimentally validated inhibitor ligand in the binding site
  4. protein must be wild type (if wild type exists)
  5. best resolution
- representatives without an adenosine phosphate ligand
  1. chain must not contain an adenosine phosphate ligand
  2. binding site must not contain an inhibitor ligand
  3. protein must be wild type (if wild type exists)
  4. best resolution

For ties in the selection, structures were sorted naturally ascending according to their PDB identifier and chain identifier and the first was chosen.

S1 File. Dataset as table. Summary table of all aaRS protein chains used for the analysis, including PDB identifier, chain identifier, superkingdom, taxonomy identifier, and ligand information (if any).

S2 File. Dataset as JSON file. Machine-readable JSON version of the dataset. Additionally enriched with protein sequence, sequence cluster identifier, and representative types for each dataset entry.

S3 File. Class I sequences in FASTA format. Protein sequences of class I aaRS structures used to construct the structure-guided MSA in FASTA format.

S4 File. Class II sequences in FASTA format. Protein sequences of class II aaRS structures used to construct the structure-guided MSA in FASTA format.

S5 File. Class I multiple sequence alignment. Structure-guided MSA of class I sequences in FASTA format.

S6 File. Class II multiple sequence alignment. Structure-guided MSA of class II sequences in FASTA format.

S7 File. Backbone brackets residue mapping. Mapping of the backbone brackets class I motif to sequence positions in origin structures.

S8 File. Backbone brackets failed mapping. List of structures where the mapping of the backbone brackets motif was not possible.

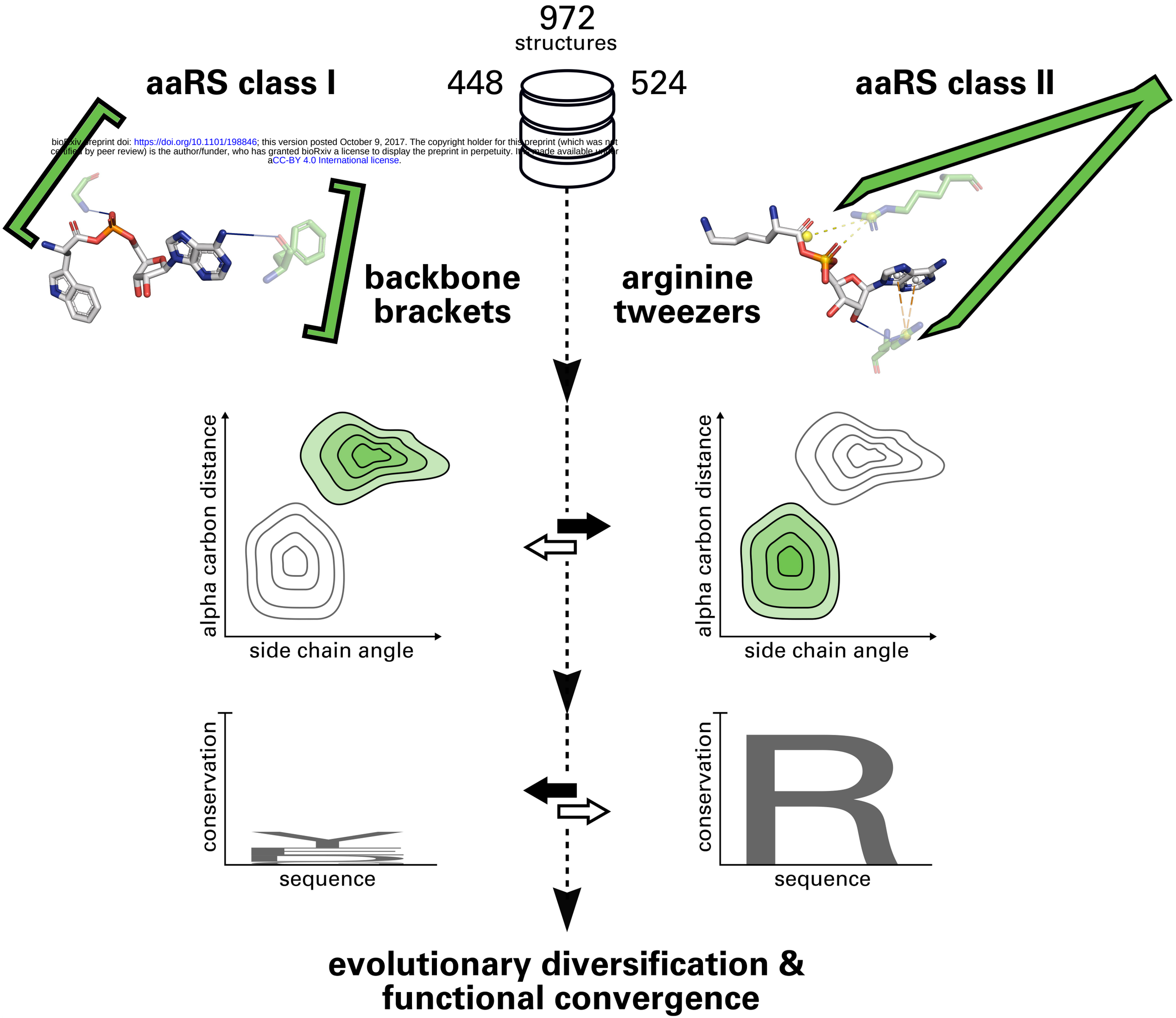
S9 File. Arginine tweezers residue mapping. Mapping of the arginine tweezers class II motif to sequence positions in origin structures.

bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

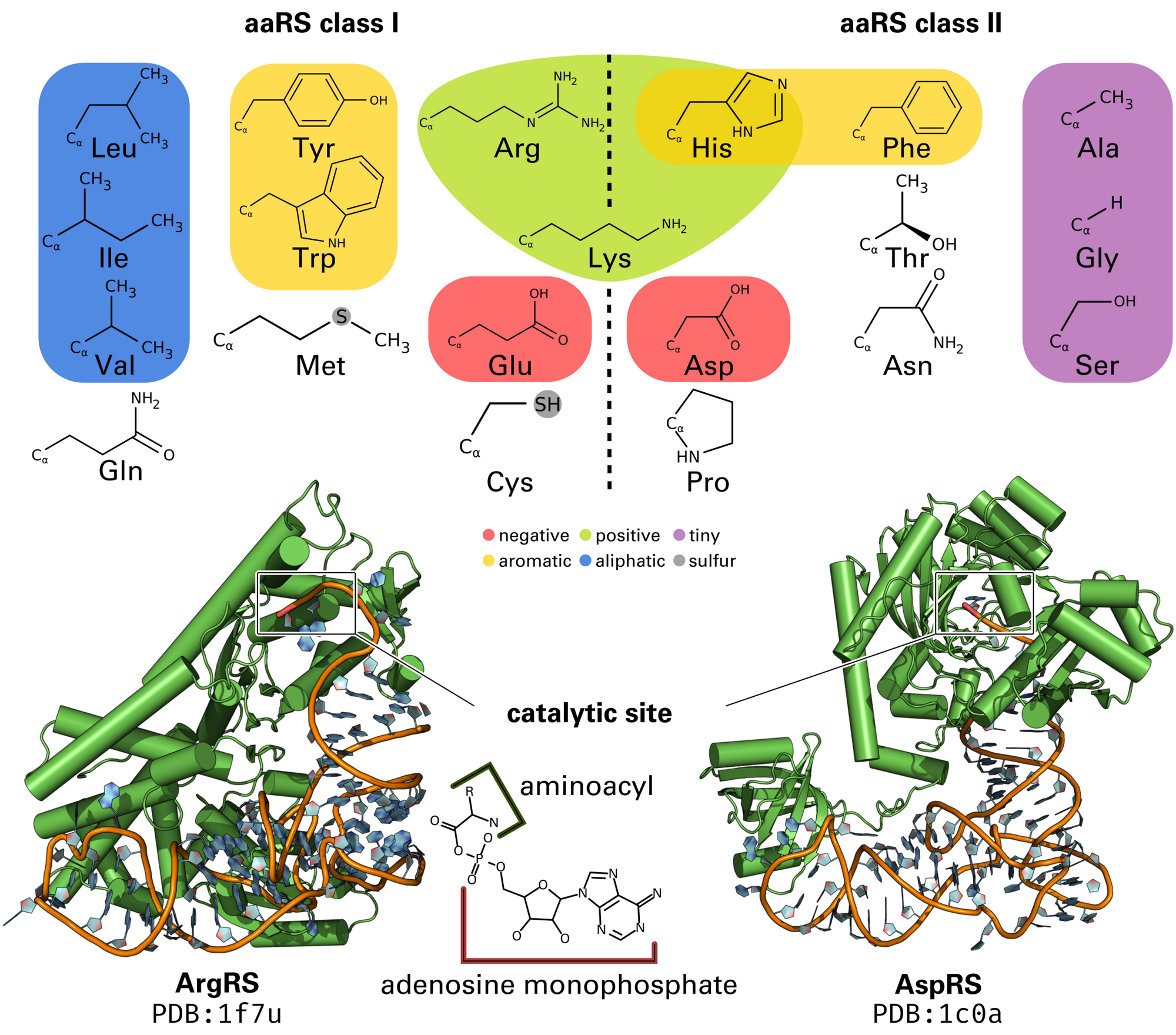
S10 File. Arginine tweezers failed mapping. List of structures where the mapping of the arginine tweezers motif was not possible.

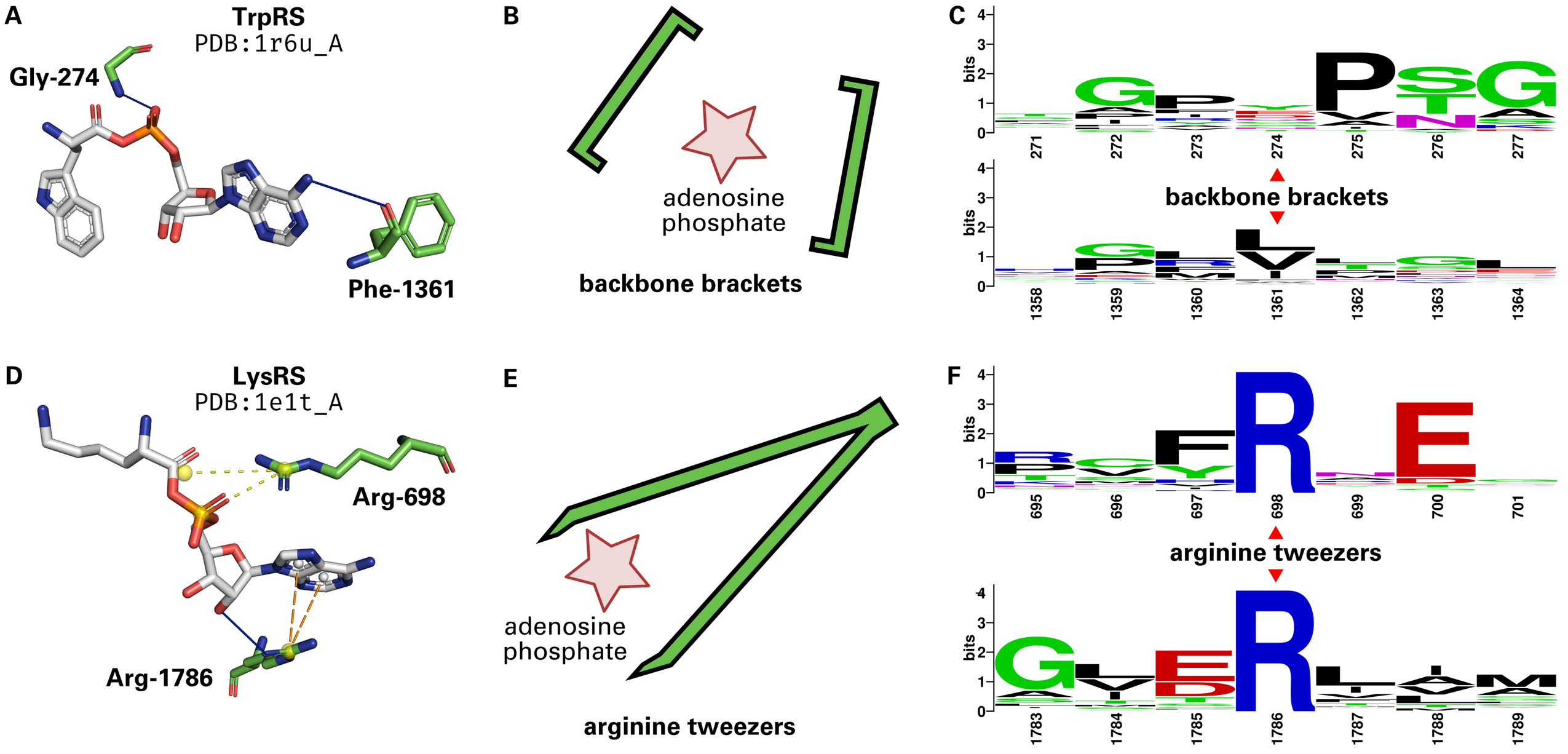
S11 File. Archive containing class I renumbered structures. All structures of class I aaRS with residues renumbered according to the MSA.

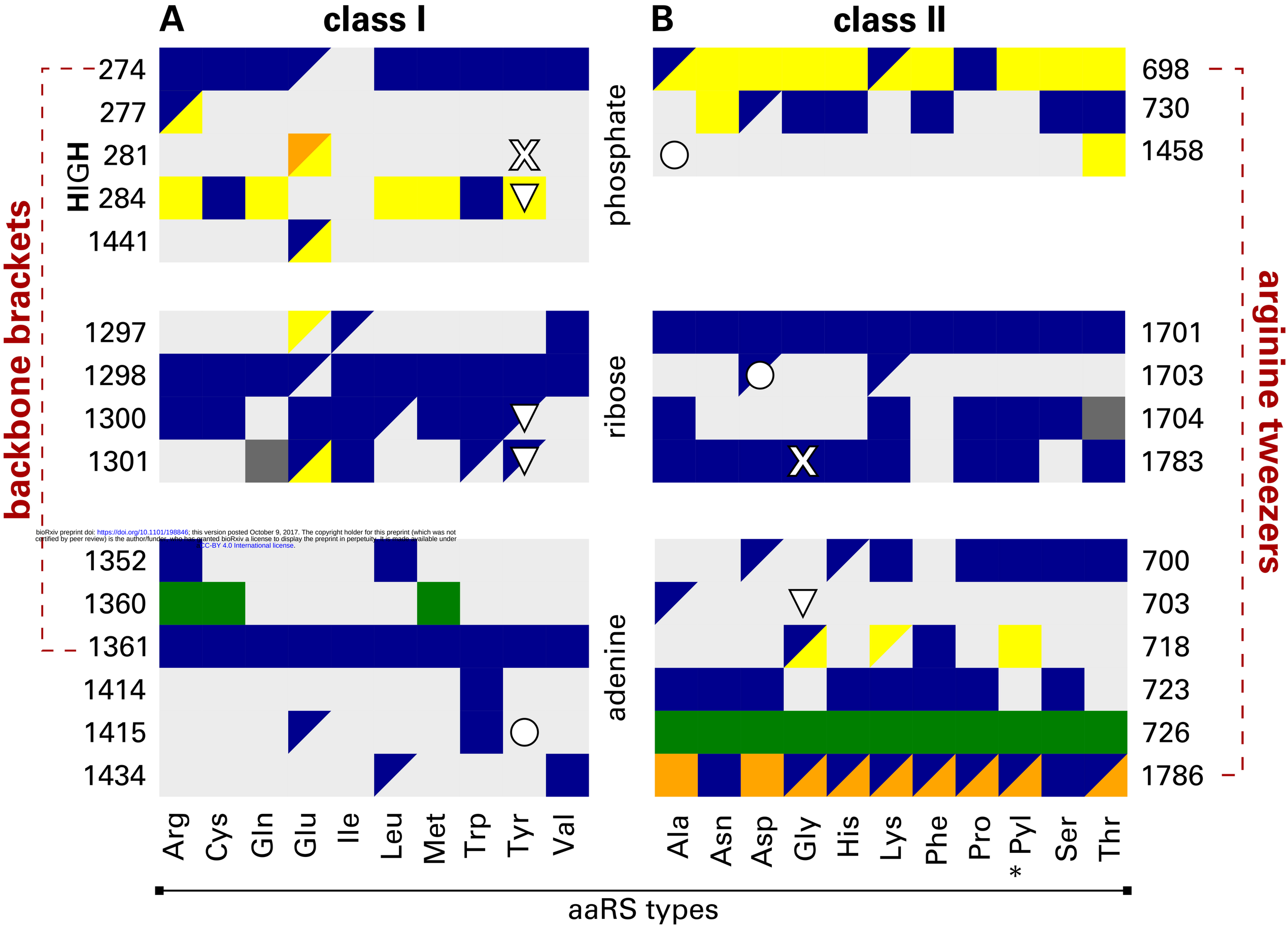
S12 File. Archive containing class II renumbered structures. All structures of class II aaRS with residues renumbered according to the MSA.



bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.


aaRS class I
aaRS class II
Leu
Ile
Val
Tyr
Trp
Arg
Lys
His
Phe
Thr
Ala
Gly
Ser
Met
Glu
Asp
Asn
Gln
Cys
Pro
negative
positive
tiny
aromatic
aliphatic
sulfur
catalytic site
aminoacyl
adenosine monophosphate
ArgRS
PDB:1f7u
AspRS
PDB:1c0a


A
TrpRS
PDB:1r6u_A
Gly-274
Phe-1361
B
adenosine
phosphate
backbone brackets
C
bits
271
272
273
274
275
276
277
backbone brackets
1358
1359
1360
1361
1362
1363
1364
D
LysRS
PDB:1e1t_A
Arg-698
Arg-1786
E
adenosine
phosphate
arginine tweezers
F
695
696
697
698
699
700
701
arginine tweezers
1783
1784
1785
1786
1787
1788
1789



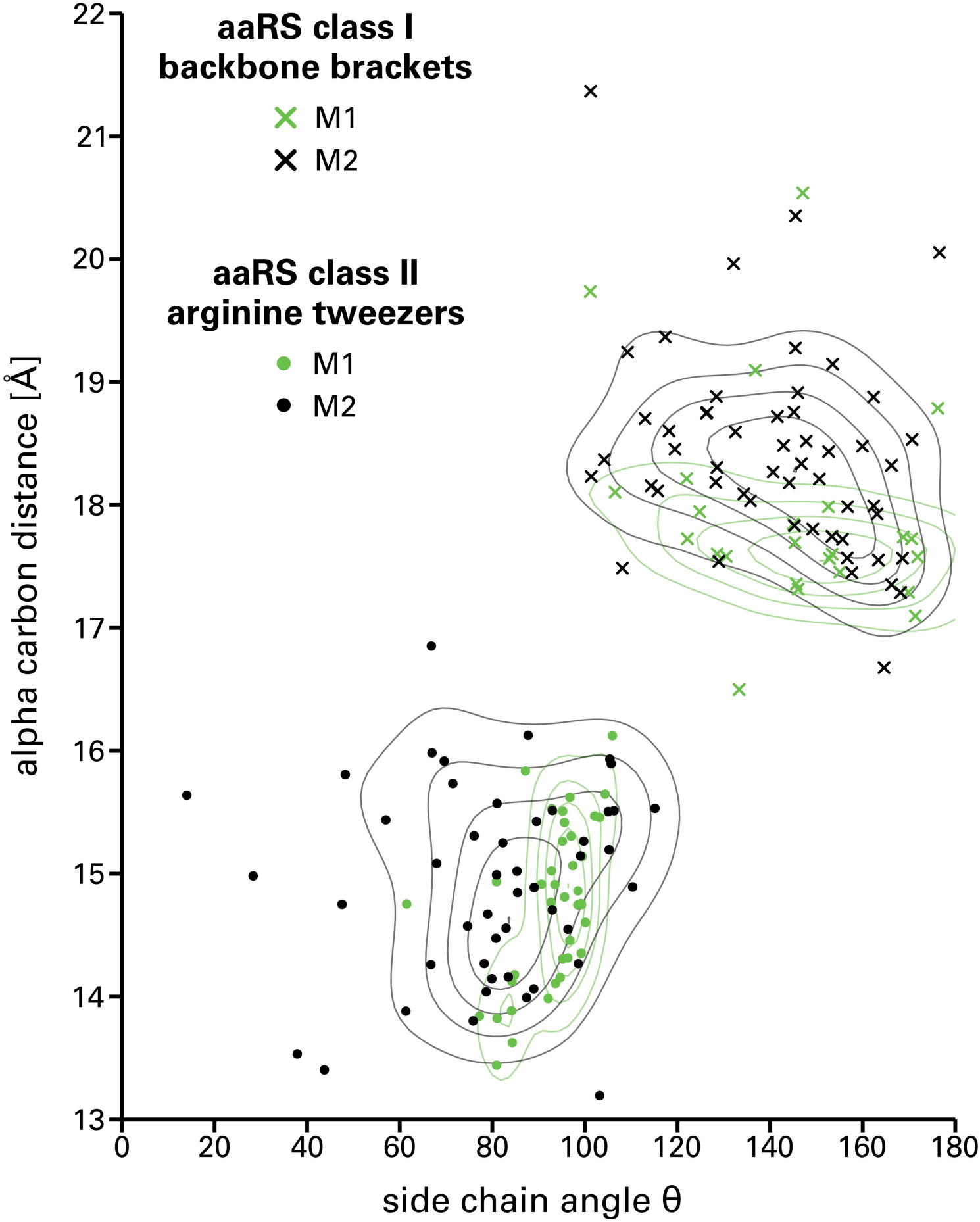
bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under CC-BY 4.0 International license.

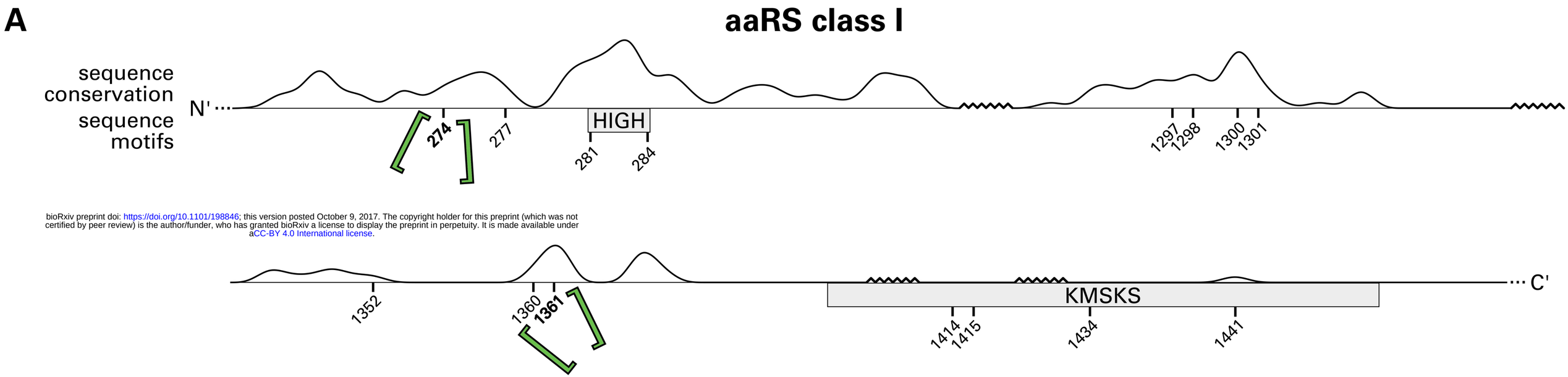
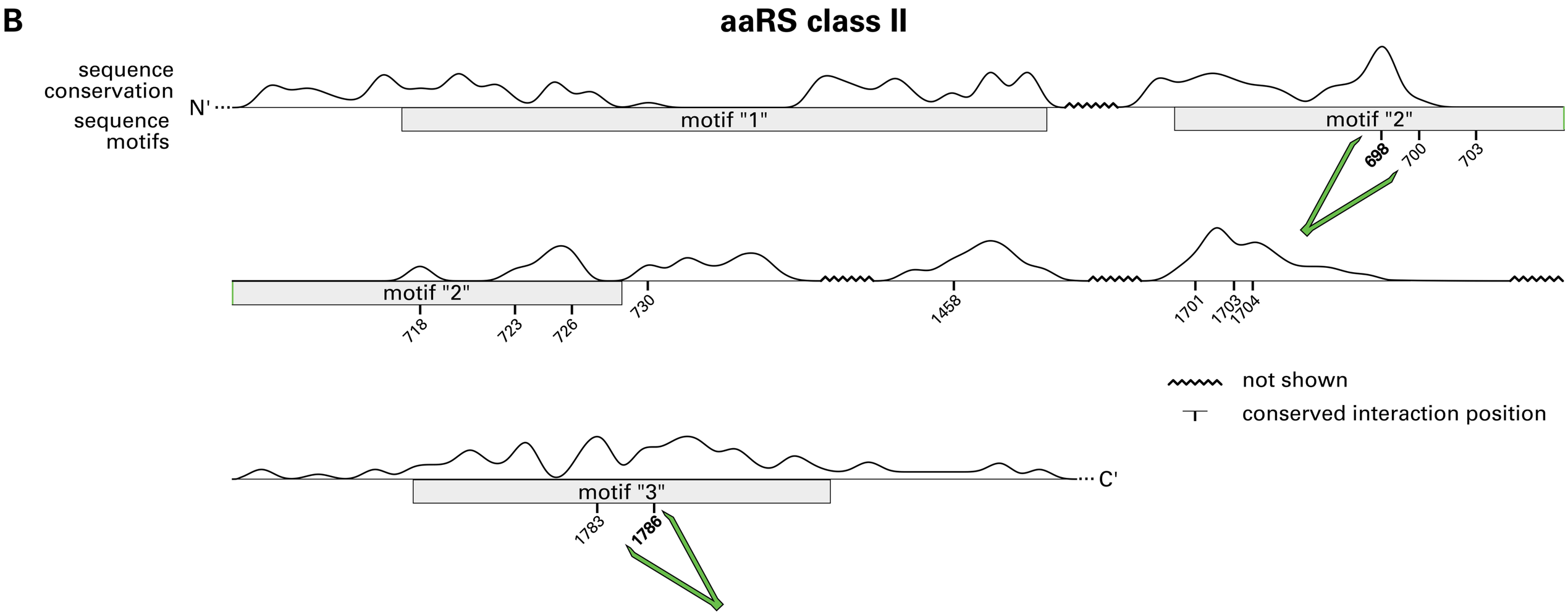
interaction types

- hydrophobic
- hydrogen bond
- salt bridge
- π-cation
- π-stacking
- no contact

mutations

- no effect
- decreased activity
- loss of activity


aaRS class I
backbone brackets
M1
M2
aaRS class II
arginine tweezers
M1
M2
alpha carbon distance [Å]
side chain angle θ
22
21
20
19
18
17
16
15
14
13
0
20
40
60
80
100
120
140
160
180



bioRxiv preprint doi: https://doi.org/10.1101/198846; this version posted October 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.