Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses

David Moi[1,2,#], Charles Bernard[1,2], Martin Steinegger[3,4,5], Yannis Nevers[1,2], Mauricio Langleib[6,7], Christophe Dessimoz[1,2,#]

[1]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland

[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

[3]School of Biological Sciences, Seoul National University, Seoul, South Korea

[4]Artificial Intelligence Institute, Seoul National University, Seoul, South Korea

[5]Institute of Molecular Biology and Genetics, Seoul National University, Seoul, South Korea

[6]Unidad de Bioinformática, Institut Pasteur de Montevideo, Montevideo, Uruguay

[7]Unidad de Genómica Evolutiva, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay

[#]Correspondence and requests for materials should be addressed to D.M. or C.D.

Abstract

Recent advances in AI-based protein structure modeling have yielded remarkable progress in predicting protein structures. Since structures are constrained by their biological function, their geometry tends to evolve more slowly than the underlying amino acids sequences. This feature of structures could in principle be used to reconstruct phylogenetic trees over longer evolutionary timescales than sequence-based approaches, but until now a reliable structure-based tree building method has been elusive. Here, we demonstrate that the use of structure-based

phylogenies can outperform sequence-based ones not only for distantly related proteins but also, remarkably, for more closely related ones. This is achieved by inferring trees from protein structures using a local structural alphabet, an approach robust to conformational changes that confound traditional structural distance measures. As an illustration, we used structures to decipher the evolutionary diversification of a particularly challenging family: the fast-evolving RRNPPA quorum sensing receptors enabling gram-positive bacteria, plasmids and bacteriophages to communicate and coordinate key behaviors such as sporulation, virulence, antibiotic resistance, conjugation or phage lysis/lysogeny decision. The advent of high-accuracy structural phylogenetics enables myriad of applications across biology, such as uncovering deeper evolutionary relationships, elucidating unknown protein functions, or refining the design of bioengineered molecules.

Introduction

Since Darwin, phylogenetic trees have depicted evolutionary relationships among organisms, viruses, genes, and other evolving entities, enabling an understanding of shared ancestry and tracing the events that led to the observable extant diversity. Trees based on molecular data are typically reconstructed from nucleotide or amino-acid sequences, by aligning homologous sequences and inferring the tree topology and branch lengths under a model of character substitution[1–3]. However, over long evolutionary time scales, multiple substitutions occurring at the same site cause uncertainty in alignment and tree building. The problem is particularly acute when dealing with fast evolving sequences, such as viral or immune-related ones, or when attempting to resolve distant relationships, such as at the origins of animals[4–6] or beyond.

In contrast, the fold of proteins is often conserved well past sequence signal saturation. Furthermore, because 3D structure determines function, protein structures have long been studied to gain insight into their biological role within the cell whether it be catalyzing reactions, interacting with other proteins to form complexes or regulating the expression of genes among a myriad of other functions.

Until recently, protein structures had to be obtained through labor intensive crystallography, with modeling efforts often falling short of the level of accuracy required to describe a fold for the many tasks crystal structures were used for. Due to these limitations, structural biology and phylogenetics have developed as largely separate disciplines and each field has created models describing evolutionary or molecular phenomena suited to the availability of computational power and experimental data.

Now, the widespread availability of accurate structural models[7,8] opens up the prospect of reconstructing trees from structures. However, there are pitfalls to avoid in order to derive evolutionary distances between homologous protein structures. Geometric distances between rigid body representations of structures, such as root mean square deviation (RMSD) distance or template modeling (TM) score[9], are confounded by spatial variations caused by conformational changes[10,11]. More local structural similarity measures have been proposed in the context of protein classification[10], but due to the relative paucity of available structures until recently, little is known about the accuracy of structure-based phylogenetic reconstruction beyond a few isolated case studies[12,13].

Here, we report on a comprehensive evaluation of phylogenetic trees reconstructed from the structures of thousands of protein families across the tree of life, using multiple kinds of distance measures. We built trees from structural divergence measures obtained using Foldseek[14], which outputs scores from rigid body alignment, local superposition-free alignment and structural alphabet based sequence alignments. The performance of these measures has been previously assessed on the task of detecting whether folds are homologous and belong to the same family[14–16], but have never been benchmarked with regards to how well they perform as evolutionary distances. Remarkably, we found that the structural alphabet-based measure outperforms phylogenies from sequence alone even at relatively short evolutionary distances. To demonstrate the capabilities of structural phylogenetics, we employ our methodology, released as open-source software named *Foldtree*, to resolve the difficult phylogeny of a fast-evolving protein family of high relevance: the RRNPPA (Rap, Rgg, NprR, PlcR, PrgX and AimR) receptors of

communication peptides. These proteins allow gram-positive bacteria, their plasmids and their viruses to assess their population density and regulate key biological processes accordingly. These communication systems have been shown to regulate virulence, biofilm formation, sporulation, competence, solventogenesis, antibiotic resistance or antimicrobial production in bacteria[17–21], conjugation in conjugative elements, lysis/lysogeny decision in bacteriophages[22] and host manipulation by mobile genetic elements (MGEs)[19,23]. Accordingly, the RRNPPA family has a substantial impact on human societies as it connects to the virulence and transmissibility of pathogenic bacteria and the spread of antimicrobial resistance genes through horizontal gene transfers. We analyze and discuss the parsimonious characteristics of the phylogeny of this family, highlighting the contrasts with the sequence-based tree.

Results

Structural trees outperform sequence based trees at both short and long evolutionary divergence times

To find a structural distance metric with high informative phylogenetic signal, we investigated the use of local superposition-free comparison (local distance difference test; LDDT[16]), rigid body alignment (TM score[9]) and a distance derived from similarity over a structural alphabet (Fident)[14]. These measures were used to compute distance trees using neighbor joining, after being aligned in an all-vs-all comparison using the Foldseek structural alphabet (*Methods*).

Assessing the accuracy of trees reconstructed from empirical data is notoriously difficult. We used two complementary indicators. The first one, taxonomic congruence score (TCS) (*Methods* **and Supplementary Figures 1-2**), assesses the congruence of reconstructed protein trees with the known taxonomy[24]. Among several potential tree topologies reconstructed from the same set of input proteins, the better topologies can be expected to have higher TCS on average.

For trees reconstructed from closely related protein families using standard sequence alignments, both local structure LDDT and global structure TM measures

showed poorer taxonomic congruence than sequence-based trees on average (**Figure 1a)**. By contrast, trees derived from the Fident distance (henceforth referred to as the *Foldtree* measure) outperformed those based on sequence. The difference was even larger if we excluded families for which the Alphafold2-inferred structures are of low confidence (**Figure 1b**). This trend was observed consistently across various protein family subsets, taken from clades with different divergence levels (**Supplementary Figure 8).** We also experimented with statistical corrections and other parameter variations, but they did not lead to further improvements (**Supplementary Figures 4-9).**

We then assessed the Foldtree measure's performance against sequence-based trees over larger evolutionary distances, using structure-informed homologous families from the CATH database[25]. This database classifies proteins hierarchically, grouping them based on **C**lass, **A**rchitecture, **T**opology and **H**omology of experimentally determined protein structures. We examined both proteins from the same homology set as well as proteins within the same topology sets (*Methods*). Efforts were made to correct structures with discontinuities or other defects before treebuilding (*Methods*) since these adversely affect structural comparisons. With this more divergent CATH dataset, structure-based methods performed better overall. Foldtree outperformed the sequence-based method even more (**Figure 1c**). Results for LDDT versus sequence flipped in favour of LDDT, while results for the global TM measure remained inferior to sequence (**Supplementary Figure 9**).

To delve deeper into the reasons for these performance differences, we applied a gradient-boosted decision tree regressor[26] on features derived from the input structures and taxonomic lineages of the input protein sets, aiming to predict the TCS difference (**Supp Methods Table 1**). We found that features measuring the confidence of the AlphaFold structure prediction (predicted LDDT or pLDDT) emerged as significant factors in the analysis (**Supplementary Figure 3**). This suggests that advancements in structural prediction might further benefit structural trees in the future.

To validate our findings using an entirely different indicator of tree quality, we assessed the "ultrametricity" of trees—how uniform a tree's root-to-tip lengths are for all its tips, akin to following a molecular clock. Although strict adherence to a molecular clock is unlikely in general, it is reasonable to assume that distance measures resulting in more ultrametric trees on average (i.e., with reduced root-to-tip variance, see *Methods*) are more accurate[27]. We found that in the sequence-based family dataset, Foldtree trees had by far the lowest root-to-tip variance of all approaches (**Figure 1d**). The difference was so pronounced that it is evident in visual comparison of tree shapes for several randomly chosen families (**Figure 1e**). Foldtree performed the best of all metrics and sequence-based trees the worst.

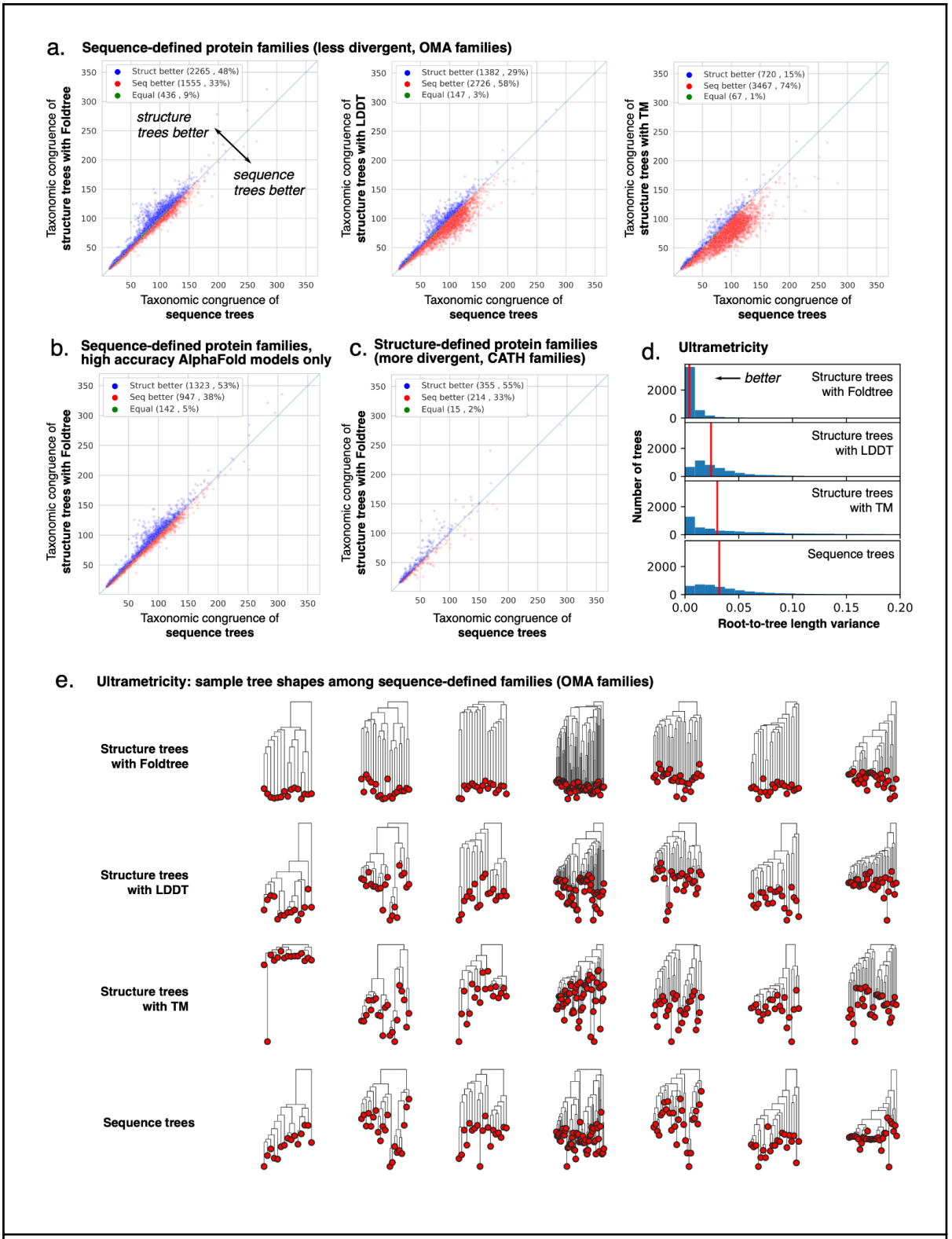

Figure 1 a) Trees using the Foldtree metric exhibit higher taxonomic congruence than sequence trees on average (protein families defined from sequences); by contrast, structure trees from LDDT and TM underperform sequence trees; **b)** After filtering the input dataset for structural quality (families with average pLDDT structure scores > 40) , the proportion of Foldtree trees which have a greater normalized congruence than sequence-based trees increased from 48% to 53%; **c**) the Foldtree metric on the CATH dataset of structurally defined families using experimental structures

outperforms sequence trees to an even greater proportion; **d)** The variance of normalized root-to-tip distances were compiled for all trees within the OMA dataset for all tree structural tree methods and sequence trees. Foldtree has a lower variance than other methods. The median of each distribution is shown with a vertical red line. Distributions are truncated to values between 0 and 0.2; **e)** A random sample of trees is shown where each column is from from equivalent protein input sets and each row of trees is derived using a distinct tree building method.

Both of the orthogonal metrics of ultrametricity and species tree discordance indicate that Foldtree produces trees with desirable characteristics that are ideal for constructing phylogenies with sets of highly divergent homologs.

Foldtree reveals the evolutionary diversification of RRNPPA communication systems

To illustrate the potential of structural phylogenies, we reconstructed the intricate evolutionary history of the RRNPPA family of intracellular quorum sensing receptors in gram-positive Bacillota bacteria, their conjugative elements and temperate bacteriophages[17,21,28]. These receptors, vital for microbial communication and decision-making, are paired with a small secreted communication peptide that accumulates extracellularly as the encoding population replicates. Once a quorum of cells, plasmids or viruses is met, communication peptides get frequently internalized within cells and binds to the tetratricopeptide repeats (TPRs) of cognate intracellular receptors, leading to gene or protein activation or inhibition, facilitating a coordinated response beneficial for a dense population. The density-dependent regulations of RRNPPA systems control behaviors like bacterial virulence, biofilm formation, sporulation, competence, conjugation and bacteriophage lysis/lysogeny decisions[17–21]. Although these receptors were identified in the early 1990s[29,30], their evolutionary history is unclear due to frequent mutations and transfers, making sequence comparisons challenging[28,31,32]. This is reflected by the nomenclature of the family: RRNPPA is an acronym for Rap, Rgg, NprR, PlcR, PrgX and AimR, which were historically described as six different families of intracellular receptors, and of which only structural comparisons allowed to establish the actual consensus on their common evolutionary origin[28,33,34]. Recently, a pioneer work combining

structural comparisons among folds and sequence-based phylogenetics have provided insights among some of these families[28], but a comprehensive reconstruction of the evolutionary history of this family that includes all described subfamilies[19] remains elusive.

The Foldtree structure-based phylogeny illuminates key evolutionary features of the diversification of RRNPPA communication systems that could not be resolved based on sequences (**Figure 2**). The evolutionary trajectory it implies is more parsimonious in terms of subfamily classification, taxonomy, functions, and protein architectures than a phylogeny obtained with a state-of-the-art sequence-based method (details in Supplementary Figure 9). In particular, the structure-based phylogeny implies that folds composed of 9 tetratricopeptide repeats (TPRs) and folds composed of 5 TPRs emerged only once while the sequence-based tree implies a less plausible scenario of convergent evolution of two clades toward 5-TPR protein architectures.

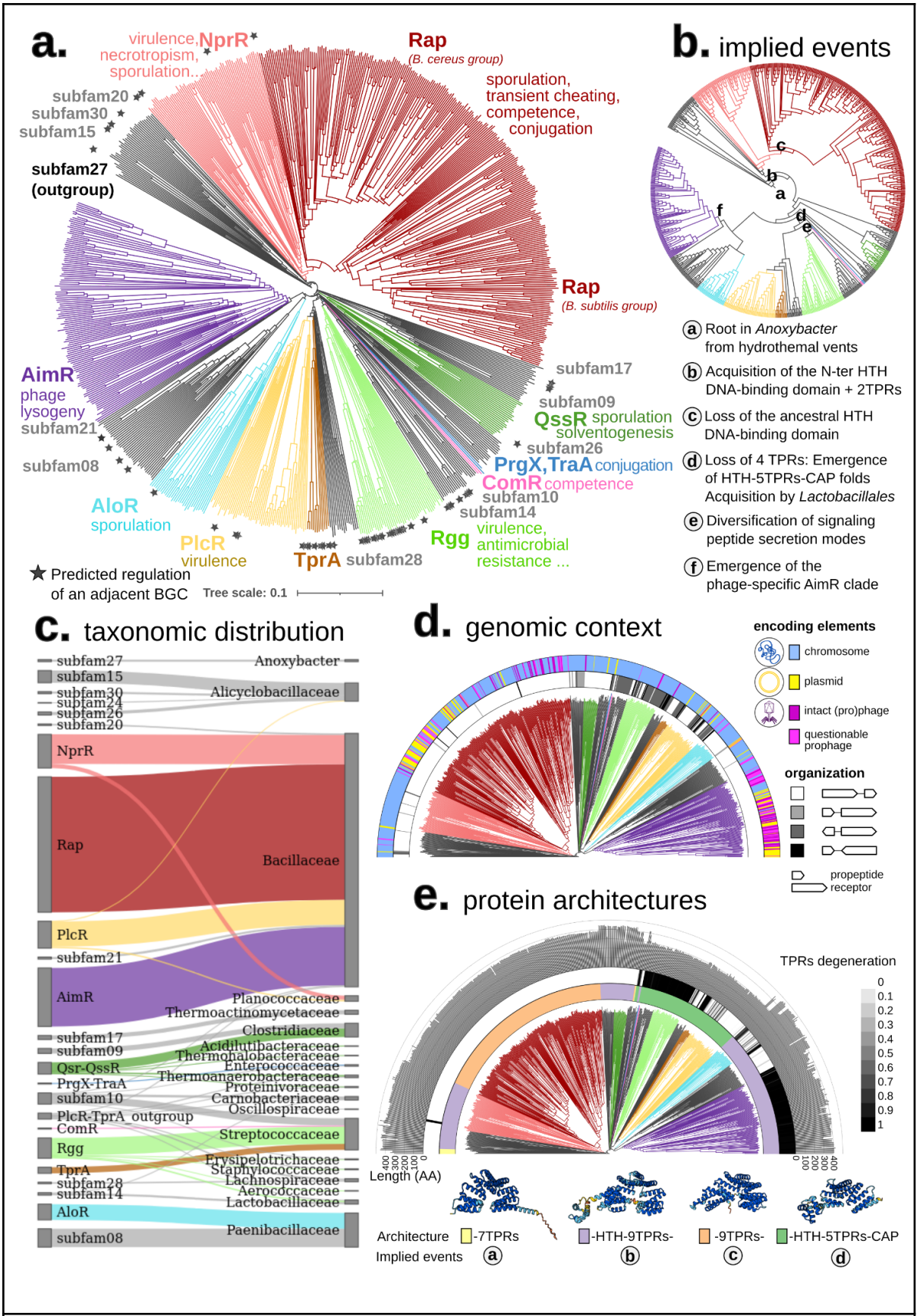

Figure 2. Phylogeny of cytosolic receptors from the RRNPPA family paired with a communication propeptide. **a**) Functional diversity of the RRNPPA family. The MAD root separates paralogs of *Anoxybacter fermentans* with a singular architecture from the other canonical RRNPPA systems.

Subfamilies with experimental validation of at least one member are highlighted in color. Other subfamilies correspond to high-confidence candidate subfamilies detected with RRNPP_detector in [19]. Biological processes experimentally shown to be regulated in a density-dependent manner by a QS system are displayed for each validated subfamily. Subfamilies in gray correspond to novel, high-confidence candidate RRNPPA subfamilies from[19]. A star mapped to a leaf indicates a predicted regulation of an adjacent biosynthetic gene cluster by the corresponding QS system. **b)** Main implied events of the tree, with normalized branch length for visualization purposes (the events that are implied from alternate roots are shown in Supplementary Figure 10). **c)** Distribution and prevalence of the different members of each RRNPPA subfamily into the different taxonomic families. **d)** Genomic orientation and encoding element of the receptor - adjacent propeptide pairs. **e)** The first colorstrip indicates the domain architecture of each receptor. A representative fold for each domain architecture is displayed in the legend (AlphaFold models of subfamily 27, NprR, Rap and PlcR, respectively) with an indication of the implied events from panel a) at the origin of each fold/architecture. The second colorstrip gives the degeneration score of TPR sequences of each receptor (given as 1 - TprPred_likelihood, as in[33]). The histogram shows the length (in amino-acids) of each receptor.

The minimal ancestor deviation (MAD) method placed the root right next to receptors encoded by *Anoxybacter fermentans* DY22613, a piezophilic and thermophilic endospore-forming bacterium from the *Clostridia* class isolated from a deep-sea hydrothermal vent. These proteins exhibit a unique domain architecture lacking the DNA-binding HTH domain and harboring 7 TPRs (Table S1). Their singular architecture, and the proximity to the MAD root lead us to infer Anoxybacter's receptors as the outgroup of all other RRNPPA systems (**Figure 2a-b**). This suggests that the early history of canonical RRNPPA systems could have been linked to extremophile endospore-forming Bacillota and may have started with a gain of a N-terminal HTH DNA binding domain, enabling to coupling quorum sensing with transcriptional regulation (**Figure 2e**). We considered alternative rooting scenarios (**Supplementary Figure 11)** but only the MAD rooting implies a unique origin of receptors with non-degenerated TPR sequences that predates the last common ancestor of each clade of receptors with degenerated TPRs (**Figure 2e**), in line with Declerck *et al.'s* conjecture[33].

The widespread distribution of sporulation-regulation on the tree (**Figure 2a**) suggests that the early history of the 9 TPRs group may have been linked to the regulation of the costly differentiation into a resistant endospore in extremophile spore-forming taxa from the *Clostridia* (*Biomaibacter acetigenes, Sulfobacillus thermotolerans, Thermoanaerobacter italicus*) and *Bacilli* (*Alicyclobacillaceae* and

Thermoactinomycetaceae families) classes (**Table S1**). Consistently, the NprR subfamily is suggested to have diversified first in extremophile spore-forming *Bacillaceae (Psychrobacilli, Halobacilli, Anoxybacilli* etc.*)* and *Planococcaceae* (*Sporosarcina, Planococcus antarticus, halotolerans, glaciei* etc.) (**Table S1**). The Rap clade, exclusively found in *Bacillus* and *Alkalihalobacillus* genera, is nested within NprR, and is inferred to have diverged from the same ancestral gene as that of NprR receptors found in *Halobacilli*, *Geobacilli*, *Virgibacilli*, *Oceanobacilli* and *Bacilli* from the *Bacillus cereus* group (**Table S1**). This indicates that the absence of the N-terminal HTH domain observed in Rap receptors originates from a loss of the ancestral domain (**Figure 2e**), as previously reported by Felipe-Ruiz et al[28]. However, many Rap receptors have retained the ability to regulate sporulation, but only through protein inhibition of the Spo0F-P and ComA regulators, rather than through transcriptional regulation[35]. The Rap clade is characterized by a wide occurrence in MGEs, consistent with the high rate of horizontal gene transfers described for this subfamily[32]. The MGE distribution in the Rap clade is polyphyletic, suggesting frequent exchanges of these communication systems between the host genome, phages and conjugative elements (**Figure 2d**). The QssR validated clade is specific to solventogenic *Clostridiaceae* (**Figure 2a-c**) while its sister clade (subfamily 09) is specific to pathogenic *Clostridium* such as *C. perfringens* and *C. botulinum*, which may indicate a novel link between quorum sensing and pathogenesis in these taxa of medical relevance that may warrant further investigation.

AloR and AimR members appear to be the most diverged representatives of the HTH-9TPRs architectural organization. Consistently, their TPR sequences harbor signs of degeneration, which is especially true in the AimR clade, consistent with its specificity to *Bacillus* phages, since viruses evolve at higher evolutionary rates (**Figure 2f**). The AimR receptors supporting phage-phage communication are adjacent to non-viral communication systems from subfamily 21, found in the chromosome of *Alkalihalobacillus clausii* or *lehensis*. For the first time, the structural phylogeny reveals that the AimR-subfamily 21 clade is evolutionary close from *Paenibacillaceae* receptors from the AloR subfamily prevalent in the *Paenibacillus*

genus and the candidate subfamily 08 predominantly found in the *Brevibacillus* genus (**Figure 2**). Subfamily 08, AloR and AimR are suggested to form a monophyletic group with a presumable *Paenibacillaceae* ancestry. This is supported by systems from *Paenibacillus xylanexedens* and *Brevibacillus formosus* position in the outgroup, close to the QssR subfamily (**Figure 2a, Table S1**). Remarkably, the cognate communication peptides of AimR receptors from the *Bacillus cereus* group are highly similar to that of subfamily 08, with the presence of the DPG amino-acid motif in the C-terminal (**Table S1**). Our results therefore suggest that a QSS similar to the ancestor of the AloR-subfam08 clade was co-opted by a temperate phage to regulate the lysis/lysogeny decision. This successful functional association has spread in *Bacillus* phages and led to the AimR clade. The numerous phage- and prophage-encoded systems from the subfamily 08 support this hypothesis[19].

The proteins composed of 5 TPRs are suggested to have emerged from the loss of 4 TPRs in the C-terminus, drastically shortening their length (**Figure 2b, Figure 2e**), although other evolutionary scenarios that do not imply such loss exist as well[28] (**Supplementary Figure 10**). The 5 TPRs group is divided in two sister clades: one with a wide taxonomic range composed of PlcR, TprA and their outgroup (**Figure 2a** and **Figure 2c**), the second including PrgX, TraA, ComR and Rgg validated subfamilies, specific to non-spore forming *Lactobactillales*. The emergence of the 5TPRs clade is associated with fundamental functional shifts. First, receptor-propeptide orientations are highly diversified compared to the HTH-9TPRs group (**Figure 9d**). These heterogeneous orientations correlate with functional changes as receptors divergently transcribed from their propeptide tend to repress target genes while co-directional receptors tend to activate them[36]. Second, the diversification of the PrgX-ComR-Rgg clade was accompanied with an important diversification of propeptide secretion modes: their cognate propeptides are exported through the alternative PptAB translocon rather than through the SEC translocon[17,28] and it has even been shown that a paralog of Rgg in *S. pyogenes* is paired with a functional leaderless communication peptide that lacks a signal sequence for an export system, highlighting that another secretory process of

communication peptides emerged in the clade[37]. Last, the biological processes controlled by these communication systems are not linked to cellular dormancy or viral latency, but rather to the production of virulence factors and antimicrobials[21]. This is mirrored by the substantial number of syntenic biosynthetic gene clusters (BGCs) predicted to be regulated by TprA and Rgg members (**Figure 2a**)[19]. Consistently, the primary role of members of the HTH-5TPRs clade may be to assess the threshold population density at which a collective production of biomolecules starts to be ecologically impactful and becomes the most evolutionary advantageous strategy, with a few exceptions such as the regulation of competence by ComR or conjugation by PrgX.

Discussion

As early as 1975, Eventoff and Rossmann employed the number of structurally dissimilar residues between pairs of proteins to infer phylogenetic relationships by means of a distance method[38]. This approach has been revisited to infer deep phylogenetic trees and networks using different combinations of dissimilarity measures (*e.g.,* RMSD, Q_{score}, Z-score) and inference algorithms[12,39–43]. Conformational sampling has been proposed to assess tree confidence when using this approach[11]. Some models have been developed that mathematically describe the molecular clock in structural evolution[44] or integrate sequence data with structural information to inform the likelihood of certain substitutions[45]. Other studies have modeled structural evolution as a diffusion process in order to infer evolutionary distances[46], or incorporating it into a joint sequence-structure model to infer multiple alignments and trees by means of bayesian phylogenetic analysis[47,48]. To date, the quality of structure-based phylogenetics, especially compared to conventional sequence-based phylogenetics, has remained largely unknown, limiting its use to niche applications.

The extensive empirical assessment reported here, using two orthogonal indicators of tree quality, demonstrates the high potential of structure-based phylogenetics. The taxonomic congruence score (TCS) measures agreement with

the established classification. Individual gene trees can be expected to deviate substantially from the underlying species tree due to gene duplication, lateral transfer, incomplete lineage sorting, or other phenomena. However, the evolutionary history of the underlying species will still be reflected in many parts of the tree—which is quantified by the TCS. All else being equal, tree inference approaches which tend to result in higher TCS over many protein families can be expected to be more accurate. On this metric, we obtained the best trees using Foldtree, which is based on Foldseek's structural alphabet, and an alignment procedure combining structural and sequence information. Furthermore, after filtering lower quality structures out of the tree building process, tree quality improved further when compared to sequence-based trees (**Figure 1.b**), indicating that higher confidence models with accurate structural information provide better phylogenetic signal.

When considering the ultrametricity through the root-to-tip variances of the trees, the Foldtree trees adhered more closely to a molecular clock than other structural or sequence trees. We acknowledge that in and of itself, adherence to a molecular clock is only a weak indicator of tree accuracy. Nevertheless, considering the clear, consistent differences obtained, and the agreement with the TCS criterion, the ultrametricity appears to reflect meaningful performance difference among the tree inference methods.

Folds evolve at a slower rate than the underlying sequence mutations[49,50]. Structural distances are therefore less likely to saturate over time, making it possible to recover the correct topology deeper in the tree with greater certainty. This could be observed in our results on the distant, structurally defined CATH families. Interestingly, however, Foldtree distinguished itself even at divergence times when homology is identifiable using sequence to sequence comparison. It is thus both fine grained enough to account for small differences between input proteins at shorter divergence times, overcoming the often mentioned shortcoming of structural phylogenetics, and more robust than sequence comparison at longer evolutionary distances.

As the projection of each residue onto a structural character is locally influenced by its neighboring residues rather than global steric changes, Foldseek's representations of 3D structures are well suited to capture phylogenetic signals when comparing homologous proteins. In contrast, global structural similarity measures are confounded by conformational fluctuations which involve steric changes that are much larger in magnitude than the local changes observed between functionally constrained residues during evolution. Moreover, since Foldseek represents 3D structures as strings, the computational speed-ups and techniques associated with string comparisons implemented in MMseqs[51] can be applied to structural homology searches and comparisons making the Foldtree pipeline extremely fast and efficient.

Viral evolution, quickly evolving extracellular proteins and protein families with histories stretching back to the first self replicating cells are among the many cases that can be revisited with these new techniques. In our first study of a family using Foldtree, we present just one such case, with the fast evolving RRNPPA family of cytosolic communication receptors encoded by Firmicutes bacteria, their conjugative elements and their viruses. The phylogeny reconstructed by Foldtree includes, for the first time, all described RRNPPA subfamilies[19]. Remarkably, despite their significant divergence, the underlying diversifying history is parsimonious in terms of taxonomy, functions, and protein architectures (**Supplementary Figure 10**). The MAD rooting method flags a previously undescribed candidate outgroup with a singular architecture of 7 TPRs and no DNA-binding domain in *Anoxybacter fermentans*, which supports Declerck et al. speculation that the ancestral receptor at the origin of the RRNPPA clade lacked the DNA-binding domain, and that the latter was gained subsequently in the evolutionary history of the family. Declerck *et al.* also speculated that the level of TPR degeneracy in receptors is a marker of divergence from the last common ancestor of the family[33]. In this respect, root to tips lengths are remarkably uniform throughout the entire RRNPPA structural tree with slight differences being meaningful, as the longest branches correspond to receptors with degenerated TPR sequences (**Figure 2e**). Last, this rooting implies that receptors with non-degenerated TPRs sequences emerged only once, and

systematically involves a late emergence of clades with degenerated TPRs as a derived state of an ancestor harboring non-degenerated TPRs (**Figure 2e**). Although rooting is easier when a tree is more clock-like, there remains uncertainty regarding the precise placement of the root. Our interpretation of MAD rooting and domain architecture led us to infer an origin of the RRNPPA family linked to the regulation of sporulation in extreme environments, implying also that 9 TPRs folds predate 5 TPRs folds. Yet, alternative rootings of the structural phylogeny cannot be ruled out, with a root either within the HTH-5TPRs group as in[28] or within the AloR-AimR-subfamily08 group (hypotheses displayed in **Supplementary Figure 11**). Additional, yet-to-be-discovered members of RRNPPA homologs could help resolve the root with higher confidence.

Recently the fold universe has been revealed using AlphaFold on the entirety of the sequences in UniProt and the ESM model[8] on the sequences in MGNIFY[52] to reach a total of nearly one billion structures. The UniProt structures inferred by AlphaFold have recently been systematically organized into sequence- and structure-based clusters, shedding light on novel fold families and their possible functions[14,53]. In future work it may be desirable to add an evolutionary layer of information to this exploration of the fold space using structural phylogenetics to further refine our understanding of how this extant diversity of folds emerged.

In conclusion, this work shows the potential of structural methods as a powerful tool for inferring evolutionary relationships among proteins. For relatively close proteins, structured-based tree inference rivals sequence-based inference, and the choice of approach should be tailored to the specific question at hand and the available data. For more distant proteins, structural phylogenetics opens new inroads into studying evolution beyond the “twilight” zone[54]. We believe that there remains much room for improvement in refining phylogenetic methods using the tertiary representation of proteins and hope that this work serves as a starting point for further exploration of deep phylogenies in this new era of AI-generated protein structures.

Methods

No statistical methods were used to predetermine sample size.

OMA HOG selection for large scale benchmark

The OMA set of protein families consists of "root hierarchical orthologous groups" (root HOGs) which are derived from all-vs-all sequence comparisons[55]. The quest for orthologs benchmarking dataset[56] consists of 78 proteomes.The 2020 release of this dataset was used as input into the OMA orthology prediction pipeline[55] (version 2.4.1). A random selection of at most 500 orthologous groups with at least 10 proteins were compiled for each group of HOGs that were inferred to have emerged in different ancestral taxa (Bacteria, Bilateria, Chordata, Dikarya, Eukaryota, Eumetazoa, Euteleostomi, Fungi, LUCA, Opisthokonta and Tetrapoda). The UniProt identifiers of the proteins within each group were used as input to the Foldtree pipeline.

CATH family selection for large scale benchmark

CATH structural superfamilies are constructed using structural comparisons and classification[25]. Each level of classification designates a different resolution of structural similarity. These are delineated as Class, Architecture, Topology and Homology. We chose to investigate tree quality using input sets within the same homology classification as well as sets within the same topology. We selected a random subsample of at most 250 proteins (or the number of proteins within the family if there were less) from each family for 635 CATH families and 500 CAT families. The Topology-based dataset is designated as CAT and the Homology-based dataset is designated as CATH. Each CAT or CATH family contains the PDB identifiers and chains of the structures they correspond to.

The PDB files were programmatically obtained from the PDB database. 3D structures of monomers corresponding to the chain identified in the CATH classification for each fold were extracted from PDB crystal structures using Biopython. PDBfixer from the OpenMM[57] package was used to fix crystal structures with discontinuities, non-standard residues or missing atoms before tree building since these adversely affect structural comparisons.

Structure tree construction

Sets of homologous structures were downloaded from the AFDB or PDB and prepared according to the OMA and CATH dataset sections above. Foldseek[14] is then used to perform an all vs all comparison of the structures.

Structural distances between all pairs are compiled into a distance matrix which is used as input to quicktree[58] to create minimum evolution trees. These trees are then rooted using the MAD method[59]. Foldseek (Version: 30fdcac78217579fa25d59bc271bd4f3767d3ebb) has two alignment modes where character based structural alignments are performed and are scored using the 3Di substitution matrix or a combination of 3Di and amino-acid substitution matrices. A third mode, using TMalign to perform the initial alignment was not used. It is then possible to output the fraction of identical amino acids from the 3Di and amino acid based alignment (Fident), the LDDT (locally derived using Foldseek's implementation) score and the TM score (normalized by alignment length). This results in a total of 6 structural comparison methods. We then either directly used the raw score or applied a correction to the scores to transform them to the distance matrices so that pairwise distances would be linearly proportional to time (Supplementary methods). This resulted in a total of 12 possible structure trees for each set of input proteins. To compile these results, Foldseek was used with alignment type 0 and alignment type 2 flags in two separate runs with the '--exhaustive-search' flag. The output was formatted to include lddt and alntmscore columns. The pipeline of comparing structure- and sequence-based trees is outlined in **Supplementary Figure 1**.

Before starting the all vs all comparison of the structures we also implemented an optional filtering step to remove poor AlphaFold models with low pLDDT values. If the user activates this option, the pipeline removes structures (and the corresponding sequences) with an average pLDDT score below 40, before establishing the final protein set and running structure and sequence tree building pipelines. We performed similar benchmarking experiments on filtered and unfiltered

versions of the OMA dataset to observe the effect of including only high quality models in the analysis.

Sequence based tree construction

Sets of sequences and their taxonomic lineage information were downloaded using the UniProt API. Clustal Omega (version 1.2.4)[60] or Muscle5 (version 5.0)[61] was then used to generate a multiple sequence alignment on default parameters. This alignment was then used with either FastTree(version 2.1)[62] on default parameters or IQ-TREE (version 1.6.12 using the flags LG+I) to generate a phylogenetic tree. Finally, this tree was rooted using the MAD (version 1775932) method on default parameters.

Taxonomic congruence metric for phylogenetic trees

Taxonomic lineages were retrieved for each sequence and structure of each protein family via the UniProt API. It is assumed that the vast majority of genes will follow an evolutionary trajectory that mirrors the species tree with occasional loss or duplication events. The original development and justification for this score to measure tree quality in an unbiased way can be found in the following work [24]. In this version of the metric we reward longer lineage sets towards the root by calculating a score for each leaf from the root to the tip.

The agreement of the tree with the established taxonomy (from UniProt) can be calculated recursively in a bottom up fashion when traversing the tree using equation 1. Leaves of trees were labeled with sets representing the taxonomic lineages of each sequence before calculating taxonomic congruence.

$$C(tree) = \sum^{Leaves} C(leaf)$$

$$C(x) = \begin{cases} |\mathrm{s(x)}| \text{ if x is root} \\ |s(x)| + |s(x.ancestor)| & \text{if x is an internal node} \end{cases}$$

where

$$s(x) = \begin{cases} L(x), \text{ if x is a leaf} \\ s(x.Left) \bigcap s(x.Right)) & \text{if x is an internal node} \end{cases}$$

Equation 1- taxonomic congruence metric. This score is used to measure the agreement of binary tree topologies with the known species tree. s(x) denotes the set of lineages found in the tree node x. C(x) denotes the congruence score of node x based on its two child nodes. L(x) denotes the labels of leaves. The total score of a tree is defined as the sum of the leaf scores. The code to calculate this metric is available on the git repository.

Both structure and sequence trees were rooted using the MAD method to make TCS comparisons between the methods equivalent. To compare large collections of trees with varying input set sizes, we normalized the congruence scores of trees by the number of the proteins in the tree.

Ultrametricity quantification

Ultrametricity[63] describes the consistency of tip to root lengths of a given phylogenetic tree. If a tree building approach has an accurate molecular clock on all branches, the amount of inferred evolutionary time elapsed between the root and all of the extant species should be equivalent and proportional to real time. This would imply that the sums of branch length along a lineage from the root to any tip of the tree should be equivalent since the amount of clock time elapsed from the common ancestor until the sequencing of species in the present day is the same.

$$E(rootdist) = \sum_{i=1}^{nleaves} dist(l_i, root)/nleaves$$

$$S_{norm}(rootdist) = \sum_{i=1}^{nleaves} (dist(l_i, root)/E(rootdist) - 1)^2/(nleaves - 1)$$

Equation 2- To derive a unified metric for ultrametricity that could easily be applied to the trees generated by different methods, we normalized the branch lengths to center the distribution of root to tip lengths at 1. We then measured the variance of these normalized root to tip lengths. E(.) represents the average root to tip length for a given tree. S_{norm}(.) represents the variance of these normalized root to tip distances. dist(l_i,root) denotes the length of the tip (l_i) to root.

To describe the ultrametricity of the different methods of structural tree derivation, we measured the length of root-to-tip distances of a given tree (equation 2). We then normalized this collection of distances by their mean and calculated their variance. We compiled this variance measurement for collections of trees with corresponding input protein sets for all methods used to derive trees and compared their distributions. **Supplementary Figure 2** shows a visual representation of how this score is calculated.

RRNPPA phylogeny

The metadata of "strict" known and candidate RRNPPA QSSs described in the RRNPP_detector paper were fetched from TableS2 in the corresponding supplementary materials[19]. The predicted regulations by QSSs of adjacent BGCs were fetched from TableS5. The propeptide sequences were downloaded from the following Github repository: https://github.com/TeamAIRE/RRNPP_candidate_propeptides_exploration_dataset. The 11,939 receptors listed in TableS2 were downloaded from the NCBI Genbank database, and redundancy was removed by clustering at 95% identity with CD-HIT[64], yielding 1,418 protein clusters. The Genbank identifiers of the 11,939 receptors were used as queries in the UniProt Retrieve/ID mapping research engine (https://www.uniprot.org/id-mapping) to retrieve corresponding UniProt/AlphaFoldDB identifiers. 768 protein clusters successfully mapped to at least one UniProt/AlphaFoldDB identifier. The 768 predicted protein structures were downloaded and Foldseek was used to perform an all vs all comparison. Based on

our benchmarking results we used the Fident scores from a comparison using amino-acid and 3Di alphabet alignment scoring (alignment mode 1 in Foldseek). Since this family had undergone domain architecture modifications, we decided to extract the structural region between the first and last positions of each fold where 80% of all of the other structures in the set mapped. With these core structures we performed a second all vs all comparison. We again used the Fident scores (alignment mode 1) and no statistical correction to construct a distance matrix between the core structures. This matrix was then used with FastME[65] to create a distance based tree. The resulting tree was annotated with ITOL[66], using the metadata available in **Table S1**. To derive the sequence-based phylogeny, we built a multiple sequence alignment (MSA) of receptors, using mafft[67] with the parameters –maxiterate 1000 –localpair for high accuracy. The MSA was then trimmed with trimAl[68] under the -automated 1 mode optimized for maximum likelihood reconstruction. The trimmed alignment of 304 sites was given as input to IQ-TREE[2] to infer a maximum likelihood phylogenetic under the LG+G model with 1000 ultrafast bootstraps.

Acknowledgements

We thank the Dessimoz lab members for thoughtful discussions on the topic of structural evolution and their encouragement and input on this work. We especially thank Clement Train for his brilliant work on the tree visualization tool accompanying this work. We also gratefully acknowledge helpful suggestions by Pedro Beltrao.

The work was supported by SNSF grant 216623 to C.D.. M. L. is a recipient of a doctoral scholarship from Agencia Nacional de Investigación e Innovación (ANII), Uruguay.

Author contributions

David Moi designed and wrote the treebuilding pipeline and analysis pipelines, collected benchmarking data for CATH structural families, carried out large scale analysis for benchmarking, generated trees for protein families, and drafted the manuscript. Charles Bernard collected data relevant to the bacterial signaling case

study, analyzed and annotated the case study in light of the existing literature and wrote the corresponding sections of the paper. Martin Steinegger contributed advice and feedback on the structural distance measures evaluated in this paper. Yannis Nevers collected HOG benchmarking data and curated examples of protein families to test the pipeline. Mauricio Langlieb wrote the documentation and collected benchmarking data and curated examples of protein families. Christophe Dessimoz supervised the project and contributed to the conception of the study, the interpretation of results, and the manuscript writing.

Correspondence and requests for materials should be addressed to D.M.

Competing interests

The authors declare no competing interests.

Supplementary Information Guide

1. Supplementary data

The homologue list of RRNPPA sequences and their metadata is available in the RRNPPAlist.xls file. In the text it is referred to as **Table S1**.

2. Supplementary methods, results and discussion are found in the SI section pdf

Code and Data availability

All UniProt identifiers necessary to replicate the experimental results are available on Zenodo: https://doi.org/10.5281/zenodo.8346286

The Foldtree pipeline is available on github: https://github.com/DessimozLab/fold_tree

All metadata used to annotate the RRNPPA phylogeny are available in the supplementary data file or on the Zenodo archive.

References

1. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
2. Minh, B. Q., Trifinopoulos, J., Schrempf, D., Schmidt, H. A. & Lanfear, R. IQ-TREE version 2.0: tutorials and Manual Phylogenomic software by maximum likelihood. *URL http://www. iqtree. org* (2019).
3. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, e1006650 (2019).
4. Laumer, C. E. *et al.* Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proc. Biol. Sci.* **286**, 20190831 (2019).
5. Li, Y., Shen, X.-X., Evans, B., Dunn, C. W. & Rokas, A. Rooting the Animal Tree of Life. *Mol. Biol. Evol.* **38**, 4322–4333 (2021).
6. Schultz, D. T. *et al.* Ancient gene linkages support ctenophores as sister to other animals. *Nature* **618**, 110–117 (2023).
7. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
8. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
9. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
10. Le, Q., Pollastri, G. & Koehl, P. Structural alphabets for protein structure classification: a

comparison study. *J. Mol. Biol.* **387**, 431–450 (2009).

11. Malik, A. J., Poole, A. M. & Allison, J. R. Structural Phylogenetics with Confidence. *Mol. Biol. Evol.* **37**, 2711–2726 (2020).
12. Bujnicki, J. M. Phylogeny of the restriction endonuclease-like superfamily inferred from comparison of protein structures. *J. Mol. Evol.* **50**, 39–44 (2000).
13. Balaji, S. & Srinivasan, N. Use of a database of structural alignments and phylogenetic trees in investigating the relationship between sequence and structural variability among homologous proteins. *Protein Eng.* **14**, 219–226 (2001).
14. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01773-0.
15. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889–895 (2010).
16. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).
17. Neiditch, M. B., Capodagli, G. C., Prehna, G. & Federle, M. J. Genetic and Structural Analyses of RRNPP Intercellular Peptide Signaling of Gram-Positive Bacteria. *Annu. Rev. Genet.* **51**, 311–333 (2017).
18. Fleuchot, B. *et al.* Rgg proteins associated with internalized small hydrophobic peptides: a new quorum-sensing mechanism in streptococci. *Mol. Microbiol.* **80**, 1102–1119 (2011).
19. Bernard, C., Li, Y., Lopez, P. & Bapteste, E. Large-Scale Identification of Known and Novel RRNPP Quorum-Sensing Systems by RRNPP_Detector Captures Novel Features of Bacterial, Plasmidic, and Viral Coevolution. *Mol. Biol. Evol.* **40**, (2023).
20. Kotte, A.-K. *et al.* RRNPP-type quorum sensing affects solvent formation and

sporulation in Clostridium acetobutylicum. *Microbiology* **166**, 579–592 (2020).

21. Perez-Pascual, D., Monnet, V. & Gardan, R. Bacterial Cell-Cell Communication in the Host via RRNPP Peptide-Binding Regulators. *Front. Microbiol.* **7**, 706 (2016).
22. Stokar-Avihail, A., Tal, N., Erez, Z., Lopatina, A. & Sorek, R. Widespread Utilization of Peptide Communication in Phages Infecting Soil and Pathogenic Bacteria. *Cell Host Microbe* **25**, 746–755.e5 (2019).
23. Cardoso, P. *et al.* Rap-Phr Systems from Plasmids pAW63 and pHT8-1 Act Together To Regulate Sporulation in the Bacillus thuringiensis Serovar kurstaki HD73 Strain. *Appl. Environ. Microbiol.* **86**, (2020).
24. Tan, G., Gil, M., Löytynoja, A. P., Goldman, N. & Dessimoz, C. Simple chained guide trees give poorer multiple sequence alignments than inferred trees in simulation and phylogenetic benchmarks. *Proceedings of the National Academy of Sciences of the United States of America* vol. 112 E99–100 (2015).
25. Knudsen, M. & Wiuf, C. The CATH database. *Hum. Genomics* **4**, 207–212 (2010).
26. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *aos* **29**, 1189–1232 (2001).
27. Bereg, S. & Zhang, Y. Phylogenetic networks based on the molecular clock hypothesis. in *Fifth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'05)* 320–323 (2005).
28. Felipe-Ruiz, A., Marina, A. & Rocha, E. P. C. Structural and Genomic Evolution of RRNPPA Systems and Their Pheromone Signaling. *MBio* **13**, e0251422 (2022).
29. Clewell, D. B. & Weaver, K. E. Sex pheromones and plasmid transfer in Enterococcus faecalis. *Plasmid* **21**, 175–184 (1989).
30. Rudner, D. Z., LeDeaux, J. R., Ireton, K. & Grossman, A. D. The spo0K locus of Bacillus subtilis is homologous to the oligopeptide permease locus and is required for

sporulation and competence. *J. Bacteriol.* **173**, 1388–1398 (1991).

31. Kalamara, M., Spacapan, M., Mandic-Mulec, I. & Stanley-Wall, N. R. Social behaviours by Bacillus subtilis: quorum sensing, kin discrimination and beyond. *Mol. Microbiol.* **110**, 863–878 (2018).
32. Even-Tov, E., Omer Bendori, S., Pollak, S. & Eldar, A. Transient Duplication-Dependent Divergence and Horizontal Transfer Underlie the Evolutionary Dynamics of Bacterial Cell-Cell Signaling. *PLoS Biol.* **14**, e2000330 (2016).
33. Declerck, N. *et al.* Structure of PlcR: Insights into virulence regulation and evolution of quorum sensing in Gram-positive bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 18490–18495 (2007).
34. Gallego Del Sol, F., Penadés, J. R. & Marina, A. Deciphering the Molecular Mechanism Underpinning Phage Arbitrium Communication Systems. *Mol. Cell* **74**, 59–72.e3 (2019).
35. Schultz, D., Wolynes, P. G., Ben Jacob, E. & Onuchic, J. N. Deciding fate in adverse times: sporulation and competence in Bacillus subtilis. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 21027–21034 (2009).
36. Monnet, V. & Gardan, R. Quorum-sensing regulators in Gram-positive bacteria: 'cherchez le peptide'. *Molecular microbiology* vol. 97 181–184 (2015).
37. Do, H. *et al.* Leaderless secreted peptide signaling molecule alters global gene expression and increases virulence of a human bacterial pathogen. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E8498–E8507 (2017).
38. Eventoff, W. & Rossmann, M. G. The evolution of dehydrogenases and kinases. *CRC Crit. Rev. Biochem.* **3**, 111–140 (1975).
39. Johnson, M. S., Sali, A. & Blundell, T. L. Phylogenetic relationships from three-dimensional protein structures. *Methods Enzymol.* **183**, 670–690 (1990).
40. Garau, G., Di Guilmi, A. M. & Hall, B. G. Structure-based phylogeny of the

metallo-beta-lactamases. *Antimicrob. Agents Chemother.* **49**, 2778–2784 (2005).

41. Lundin, D., Berggren, G., Logan, D. T. & Sjöberg, B.-M. The origin and evolution of ribonucleotide reduction. *Life* **5**, 604–636 (2015).
42. Moi, D. *et al.* Discovery of archaeal fusexins homologous to eukaryotic HAP2/GCS1 gamete fusion proteins. *Nat. Commun.* **13**, 3880 (2022).
43. Lakshmi, B., Mishra, M., Srinivasan, N. & Archunan, G. Structure-Based Phylogenetic Analysis of the Lipocalin Superfamily. *PLoS One* **10**, e0135507 (2015).
44. Pascual-García, A., Arenas, M. & Bastolla, U. The Molecular Clock in the Evolution of Protein Structures. *Syst. Biol.* **68**, 987–1002 (2019).
45. Arenas, M., Sánchez-Cobos, A. & Bastolla, U. Maximum-Likelihood Phylogenetic Inference with Selection on Protein Folding Stability. *Mol. Biol. Evol.* **32**, 2195–2207 (2015).
46. Grishin, N. V. Estimation of evolutionary distances from protein spatial structures. *J. Mol. Evol.* **45**, 359–369 (1997).
47. Challis, C. J. & Schmidler, S. C. A stochastic evolutionary model for protein structure alignment and phylogeny. *Mol. Biol. Evol.* **29**, 3575–3587 (2012).
48. Herman, J. L., Challis, C. J., Novák, Á., Hein, J. & Schmidler, S. C. Simultaneous Bayesian estimation of alignment and phylogeny under a joint model of protein sequence and structure. *Mol. Biol. Evol.* **31**, 2251–2266 (2014).
49. Illergård, K., Ardell, D. H. & Elofsson, A. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. *Proteins* **77**, 499–508 (2009).
50. Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826 (1986).
51. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for

the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

52. Richardson, L. *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).
53. Durairaj, J. *et al.* Uncovering new families and folds in the natural protein universe. *Nature* (2023) doi:10.1038/s41586-023-06622-3.
54. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94 (1999).
55. Altenhoff, A. M. *et al.* OMA standalone: orthology inference among public and custom genomes and transcriptomes. *Genome Res.* **29**, 1152–1163 (2019).
56. Altenhoff, A. M. *et al.* The Quest for Orthologs benchmark service and consensus calls in 2020. *Nucleic Acids Res.* **48**, W538–W545 (2020).
57. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13**, e1005659 (2017).
58. Howe, K., Bateman, A. & Durbin, R. QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics* **18**, 1546–1547 (2002).
59. Tria, F. D. K., Landan, G. & Dagan, T. Phylogenetic rooting using minimal ancestor deviation. *Nat Ecol Evol* **1**, 193 (2017).
60. Sievers, F. & Higgins, D. G. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.* **1079**, 105–116 (2014).
61. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
62. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
63. Moore, N. C. A. & Prosser, P. The Ultrametric Constraint and its Application to Phylogenetics. *arXiv [cs.AI]* (2014).
64. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of

protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

65. Lefort, V., Desper, R. & Gascuel, O. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. *Mol. Biol. Evol.* **32**, 2798–2800 (2015).
66. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
67. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
68. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* vol. 25 1972–1973 Preprint at https://doi.org/10.1093/bioinformatics/btp348 (2009).