


# The Structural Determinants of Intra-Protein Compensatory Substitutions

Shilpi Chaurasia<sup>1,†</sup> and Julien Y. Dutheil <sup>1,2,\*</sup>

<sup>1</sup>RG Molecular Systems Evolution, Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, August-Thienemann-Straße 2, 24306 Plön, Germany

<sup>2</sup>Institute of Evolution Sciences of Montpellier (ISEM), CNRS, University of Montpellier, IRD, EPHE, 34095 Montpellier, France

\*Corresponding author: E-mail: [dutheil@evolbio.mpg.de](mailto:dutheil@evolbio.mpg.de).

†Present address: Excelra Knowledge Solutions Pvt Ltd, Hyderabad, India.

Associate editor: Rebekah Rogers

## Abstract

Compensatory substitutions happen when one mutation is advantageously selected because it restores the loss of fitness induced by a previous deleterious mutation. How frequent such mutations occur in evolution and what is the structural and functional context permitting their emergence remain open questions. We built an atlas of intra-protein compensatory substitutions using a phylogenetic approach and a dataset of 1,630 bacterial protein families for which high-quality sequence alignments and experimentally derived protein structures were available. We identified more than 51,000 positions coevolving by the mean of predicted compensatory mutations. Using the evolutionary and structural properties of the analyzed positions, we demonstrate that compensatory mutations are scarce (typically only a few in the protein history) but widespread (the majority of proteins experienced at least one). Typical coevolving residues are evolving slowly, are located in the protein core outside secondary structure motifs, and are more often in contact than expected by chance, even after accounting for their evolutionary rate and solvent exposure. An exception to this general scheme is residues coevolving for charge compensation, which are evolving faster than noncoevolving sites, in contradiction with predictions from simple coevolutionary models, but similar to stem pairs in RNA. While sites with a significant pattern of coevolution by compensatory mutations are rare, the comparative analysis of hundreds of structures ultimately permits a better understanding of the link between the three-dimensional structure of a protein and its fitness landscape.

**Key words:** molecular coevolution, compensatory mutations, epistasis, protein structure, evolutionary rate, substitution mapping.

## Introduction

The function of a biological molecule depends on its structure. It results that the structural characteristics of biomolecules impact the fitness effect of mutations in the genes that encode them and, therefore, determine their fate. The impact of structure on the process of molecular evolution has been extensively documented, both in RNA and proteins (Chen et al. 1999; Liberles et al. 2012; Moutinho et al. 2019). The structure and function of macromolecules, however, stem from the complex interactions—rather than the sum of the properties—of the underlying residues. As a consequence, the fitness effect of a mutation at a given position may depend on the state of the interacting residues, inducing a nonindependent evolution, or coevolution (Starr and Thornton 2016). The interest of studying the signature of coevolution in molecular sequences has long been recognized, as detecting coevolving positions has the potential to point at functionally and structurally important interactions (de Juan et al. 2013).

Detecting coevolving positions is a statistically complex task, as the evolutionary process is not directly observable. Only its result is, in the form of extant sequences. The study of molecular coevolution is, therefore, grounded in phylogenetic comparative analysis: the shared history of species induces correlations in the sampled sequences that need to be disentangled from the functional correlations between sites (Pollock and Taylor 1997; Atchley et al. 2000; Dimmic et al. 2005). Furthermore, the geometry of interactions may vary in time and sequence space, not necessarily involving the same set of residues at different time points during the evolutionary history of the molecule. A large corpus of methods has been developed in order to address (some of) these issues, some using explicit evolutionary modeling of coevolution (e.g., Pollock et al. 1999; Dib et al. 2014; Behdenna et al. 2016), others relying on increasingly large data sets and advanced data mining procedures (Weigt et al. 2009; Jones et al. 2012; Wang et al. 2017; Li et al. 2019) to assess the patterns of site covariation. Furthermore, several case studies have provided a detailed understanding of the structural mechanisms of particular

© The Author(s) 2022. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

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

Open Access

compensatory mutations (Ivankov et al. 2014; Storz 2018). The distribution of coevolving positions in proteins, however, and the underlying molecular mechanisms of coevolution remain largely unknown. A reason underlying this state-of-the-art is that model-based approaches are computationally intensive, preventing large scale comparisons (Pollock et al. 1999; Dimmic et al. 2005), and typically produce small numbers of candidate coevolving positions with strong statistical support (Tufféry and Darlu 2000; Dutheil and Galtier 2007; Dunn et al. 2008). Conversely, data mining approaches are able to predict the interaction network of residues in a molecule with good accuracy, providing that a very large number of sequences are available (Tetchner et al. 2014), thus preventing the use of such methods on a large variety of gene families and protein structures: only few gene families will match the necessary sample size. Furthermore, these methods are typically calibrated to detect physically interacting residues, which may or may not be coevolving, excluding potentially coevolving residues not in direct physical contact (Di Lena et al. 2012).

There is an apparent discrepancy between evolutionary methods, which predict relatively few coevolving positions, and pattern-based methods, which successfully predict a large proportion of residues that are physically in contact. The first point is explained by a theoretical argument (Ivankov et al. 2014; Talavera et al. 2015): for compensatory mutations to occur, the first mutation must be sufficiently deleterious for the second mutation to be advantageous and invade the population. If the first mutation is too deleterious; however, it will be removed quickly from the population and will not be compensated. As a result, compensatory mutations are predicted to be relatively rare, and coevolving sites should evolve rather slowly. Talavera et al. (2015) further argue that covariation methods primarily infer slowly evolving sites, which tend to be located at the core of proteins and are, therefore, more likely to be in close proximity. Paradoxically, in RNA molecules, interacting pairs within double-stranded helices have long been recognized as exhibiting a clear pattern of coevolution resulting from Watson–Crick interactions and show an accelerated rate of evolution compared with single-stranded regions (Smit et al. 2007). The frequency of compensatory mutations in proteins, their distribution with regard to structural properties, and their relation to the evolutionary rate of coevolving sites remain to be established (Starr and Thornton 2016).

Here, we aim at generating an atlas of protein coevolution. We gathered a large dataset of sequence alignments of protein families for which at least one protein structure has been experimentally determined. We use a phylogenetic method that exhibits positions undergoing substitutions on the same branches of the phylogeny (Dutheil et al. 2005; Dutheil and Galtier 2007). We further restrict our analysis to the case of coevolution by compensatory mutations, where the first mutation at a given position has a negative fitness effect, for instance, because it results in a less stable protein structure, which is compensated by a second mutation at another position in the

protein. In order to predict compensatory mutations, we need a proxy for the fitness effect of single mutations. Considering that the fitness impact of a mutation depends on the biochemical properties of the encoded amino acids, we can predict this effect by quantifying the change in such properties (Grantham 1974). Neher (1994) first introduced this idea, developing a method looking at positive and negative correlations of biochemical properties between positions of sequence alignments. Dutheil and Galtier (2007) introduced a measure of phylogenetic compensation (referred to as the “compensation index”), which assesses the conservation of a biochemical property in a group of sites, given their individual variation. More specifically, considering  $n$  sites and a phylogeny with  $m$  branches, we note  $x_{ij}^P$  the change of a biochemical property  $P$  for the site  $i$  on branch  $j$ . We further note  $X_i^P = (x_{ij}^P)_{1 \leq j \leq m}$  the vector of branch-specific changes for the site  $i$ . The compensation index  $C(G)$  for a group of sites  $G$  is then computed as

$$C(G) = 1 - \frac{|\sum X_i|}{\sum |X_i|}, \quad (1)$$

where  $|X|$  denotes the norm of  $X$  vector. When the changes at a given set of positions tend to be in opposite directions,  $|\sum X_i|$  tends towards 0 and  $C$  towards 1. Conversely, if the changes are in the same direction,  $|\sum X_i|$  equals  $\sum |X_i|$  and  $C$  is equal to 0. The CoMap method uses a model-based substitution mapping procedure to infer the changes at each site on each branch of the phylogeny (Tataru and Hobolth 2011), combined with a clustering approach to detect candidate coevolving groups, which are then tested using simulations under the null hypothesis of independently evolving positions (Dutheil and Galtier 2007). While these simulations assume site independence, they conserve other aspects of the evolutionary process, such as the phylogenetic relationship of sequences, the probabilities of individual amino acid substitutions, and the site-specific rate of substitutions, allowing a precise evaluation of the coevolution signal while accounting for potentially confounding factors. Applying this method to hundreds of protein families and statistically analyzing hundreds of thousands of residues with structural annotations, we determine the drivers of protein intra-molecular coevolution.

## Results

We gathered a large data set of protein sequence families from complete genome sequences and for which structural information was available for at least one representative sequence. We initially analyzed gene families from the Archaea, Eukaryotes, and Bacteria separately. However, we later restricted our analyses to bacterial families due to the small number of sequences and structures retained in the two other domains. We developed a stringent pipeline controlling for alignment uncertainty and

sequence redundancy (see Material and Methods). Our curated data set contains 1,630 protein families, with a number of sequences per family ranging from 100 to 400 (supplementary table S1, Supplementary Material online). A maximum-likelihood phylogeny was reconstructed for each family and used as input of the coevolution detection method. In order to assess the compensating nature of substitutions, we considered several biochemical properties. Following previous studies, we considered the volume, polarity, and charge of amino acid residues (Neher 1994; Pollock et al. 1999; Tufféry and Darlu 2000). The AAIndex database (Kawashima et al. 2008), however, contains more than 500 nonindependent indices. Using clustering techniques, Saha et al. (2012) have shown that these redundant properties are clustered into eight groups. We included the eight indices corresponding to the centers of these clusters (hereby referred to as “synthetic indices”), in order to provide an objective and comprehensive measure of amino acid biochemical properties. These indices are defined as follow (Saha et al. 2012): electric properties (I1), hydrophobicity (I2), alpha and turn propensities (I3), physicochemical properties (I4), residue propensity (I5), composition (I6), beta propensity (I7), and intrinsic propensities (I8). We ran the CoMap coevolution detection method to detect nonoverlapping groups of coevolving positions, with a size ranging from 2 to 10, for the 11 biochemical properties. As a result, each site was annotated as coevolving for a given property if it belonged to a significant group. Structural properties, such as secondary structure motif and solvent exposure, as well as evolutionary rates, were recorded for each analyzed site (see Materials and Methods).

### Substitution Mapping Enables the Detection of Compensatory Mutations

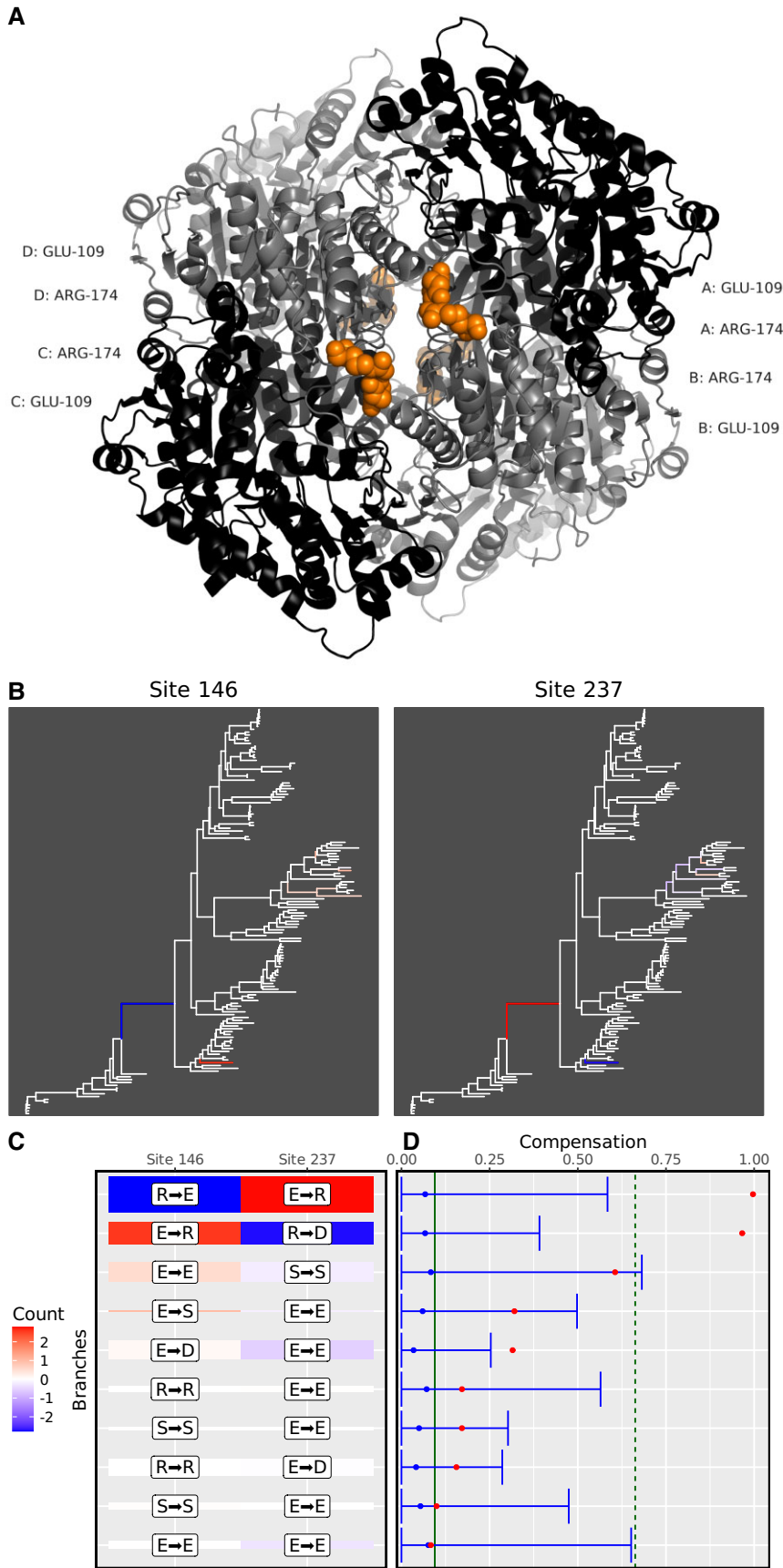
We detail two candidate groups to illustrate the nature of the positions predicted as coevolving in our data set. The first group involves a pair of distant sites at positions 146 and 237 in the sequence alignment of protein family HOG000218359. The three-dimensional structure used as a representative is the Menaquinone biosynthesis MenD protein of *Bacillus subtilis* (PDB ID:2X7J). The protein is a tetramer (Dawson et al. 2010), and the A chain was used as a reference. The coevolving positions correspond to position Glu-109 and Arg-174 in the *B. subtilis* sequence (fig. 1A). The positions are significantly coevolving for charge compensation ( $P$ -value =  $4.940333 \times 10^{-6}$ , significant after correction for multiple testing with a false discovery rate [FDR] of 1%). While the two sites show low substitution rates, they undergo two significant changes in two branches of the tree (fig. 1B). These changes show a perfect compensation signal, from a positively charged residue to a negatively charged residue at one position and from a negatively charged residue to a positively charged residue at the other position (fig. 1B). This signal is illustrated on the “compensogram” in fig. 1D (see

Materials and Methods): the compensation is perfect ( $C = 1$ ) for the first branch and almost perfect for the second branch ( $C > 0.9$ ).

The second example group involves three positions from the HOG000227724 family, whose representative structure is the A-chain of the dTDP-6-deoxy-D-xylo-4-hexulose 3,5 epimerase (RmlC) of *Salmonella typhimurium* (PDB ID: 1DZR, fig. 2A) (Giraud et al. 2000). The coevolving positions include the two close residues Glu-16 and Phe-20, as well as Asn-150. The three residues are located at the surface of the protein but are not in contact. They were detected as coevolving for volume compensation ( $P$ -value = 0.0007818069), with several branches showing substitutions leading to a change in residue volume (fig. 2B). At least six branches show a significant signal of compensation (fig. 2C and D). The most significant branch shows a mutation from isoleucine (Grantham’s volume: 111, large) to alanine (Grantham’s volume: 31, small,  $-72\%$  volume change) at site 63, compensated by two substitutions at site 59 and 208: glutamine (Grantham’s volume: 85, medium) to arginine (Grantham’s volume: 124, large,  $+46\%$  volume change) and glutamic acid (Grantham’s volume: 83, medium) to isoleucine (Grantham’s volume: 111, large,  $+34\%$  volume change). The total volume of the three sites changes from 279 to 266, which only represents a  $-5\%$  volume change, illustrating the compensatory nature of the substitutions and the high compensation index ( $C > 0.9$ ). These two examples illustrate the substitution patterns of the sites detected as coevolving. In particular, the coevolution signal is here defined in a phylogenetic context and can be traced back to individual substitutions.

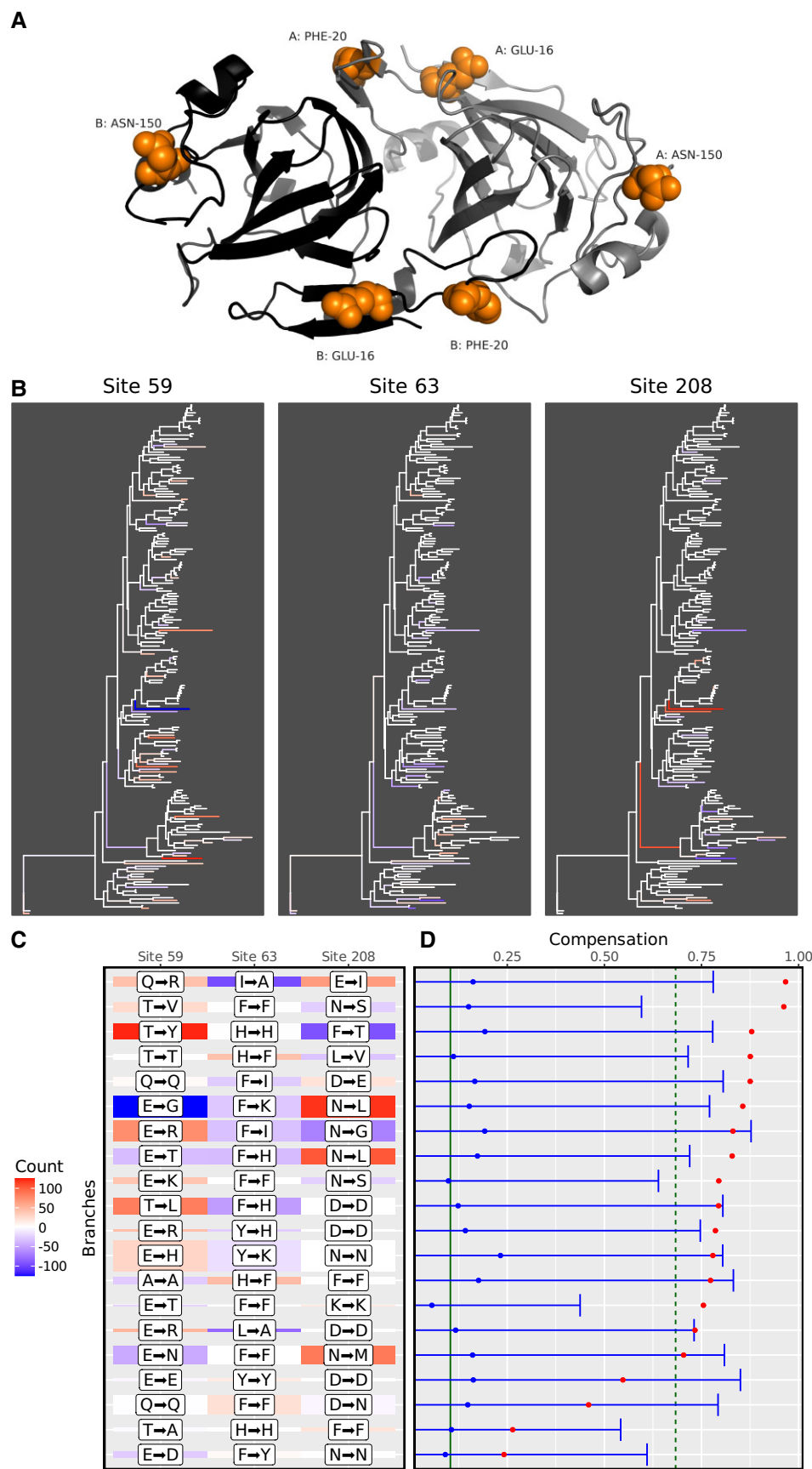
### Coevolution is Scarce Within, but Widespread Among Protein Families

Our coevolution analysis encompasses 366,794 sites, among which 51,661 (14%) were found to be coevolving for at least one biochemical property (fig. 3). While most sites were coevolving for a single biochemical property, the number of sites found to be coevolving by at least two methods was significantly greater than expected by chance (permutation test,  $P$ -value  $< 1 \times 10^{-4}$ ). This significant overlap may be explained by some sites being more likely to be detected as coevolving than others, either because of heterogeneous statistical power across sites, or their underlying functional and structural properties. We note that the eight synthetic indices detected fewer coevolving sites than the classical properties volume, polarity, and charge. This result suggests that all biochemical properties are not as likely to inducing coevolution. Interestingly, the two properties for which a coevolutionary scenario is perhaps most intuitive (volume, big-to-small compensated by small-to-big mutations, and charge, positive-to-negative compensated by negative-to-positive mutations) were the ones leading to the largest number of detected positions. The coevolving sites predicted

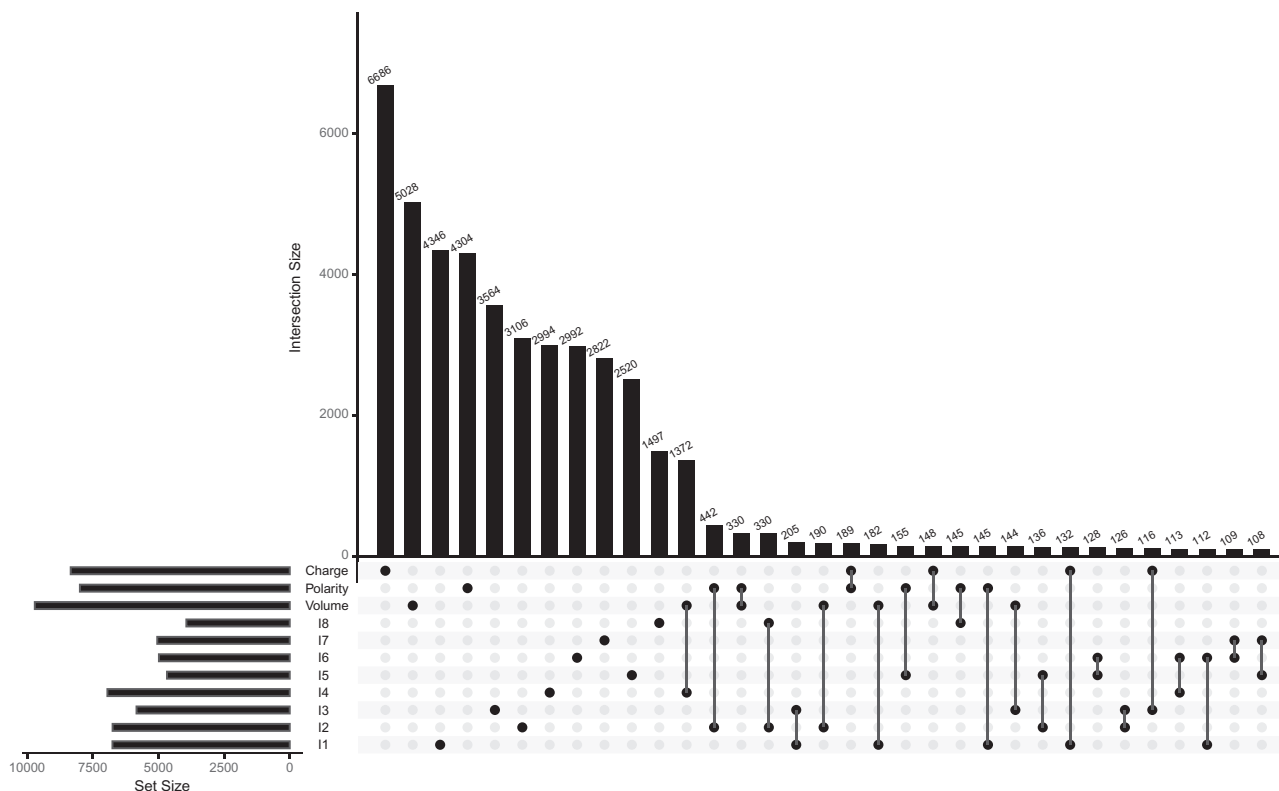


**FIG. 1.** Example 1, positions Glu-109 and Arg-174 of the Menaquinone biosynthesis MenD protein. (A) Three-dimensional structure of the Menaquinone biosynthesis MenD tetramer. Residues at coevolving positions are shown in spacefill and colored in orange, for each monomer. (B) Estimated amounts of charge change are plotted on the branches of the phylogeny for the two sites. Negative changes (i.e., charge reduction) are plotted with a blue gradient, while positive changes (i.e., charge increase) are colored with a red gradient. (C) Heatmap showing the changes for the top 10 branches, ranked according to the amount of charge compensation. Colors as in (B). The height of the tile is proportional to the total branch length. The marginal maximum likelihood ancestral state reconstruction is indicated within white labels but was not used for the compensation calculations, as these were integrated over all possible ancestral states, weighted by their respective likelihood values. (D) Compensogram of the two sites for the top 10 branches, as defined in (B). Red dots indicate the charge compensation index for each branch. Blue dots and error bars indicate the mean and 95% confidence interval of the site-permutation test, for each branch. The green vertical solid and dash lines represent the mean and 95% upper bound of the branch permutation test (see Materials and Methods).





**FIG. 2.** Example 2, positions Glu-16, Phe-20, and Asn-150 of the dTDP-6-deoxy-D-xylo-4-hexulose 3,5 epimerase (RmlC) protein. (A) Three-dimensional structure of RmlC dimer. Residues at coevolving positions are shown in full and colored in orange, for each monomer. (B) Changes of volume are plotted on the branches of the phylogeny for the three sites. Negative changes (i.e., volume reduction) are indicated in blue, while positive changes (i.e., volume increase) are colored in red. (C) Heatmap showing the changes for the top 20 branches, ranked according to the amount of volume compensation. Colors as in (B). The height of the tile is proportional to the total branch length. The marginal maximum likelihood ancestral state reconstruction is indicated with white labels but was not used for the compensation calculations, as these were integrated over all possible ancestral states, weighted by their respective likelihood values. (D) Compensogram of the top 20 branches, as defined in (B). Red dots indicate the charge compensation index for each branch. Blue dots and error bars indicate the mean and 95% confidence interval of the site-permutation test for each branch. The green vertical solid and dash lines represent the mean and 95% upper bound of the branch permutation test (see Materials and Methods).



**Fig. 3.** Upset diagram shows the overlap of sites detected as coevolving for compensation of various biochemical properties. The numbers of sites detected as coevolving (set size) are shown on the bottom left barplot. The numbers of sites detected by one or several methods (intersection size) are shown as the top barplot, sorted by decreasing size.

with these properties encompassed more than a third (34%) of all predicted positions.

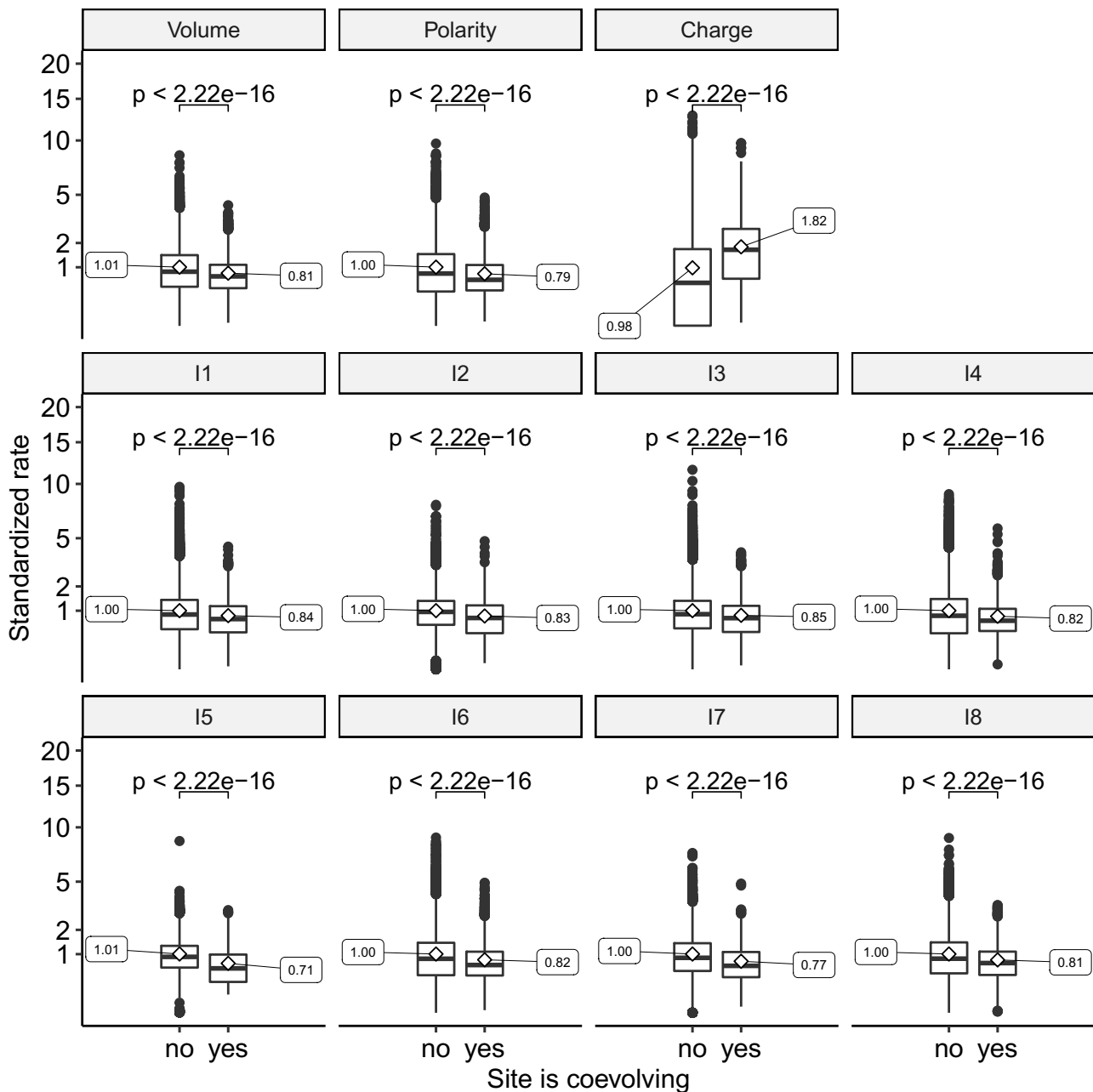
The number of detected coevolving positions in a protein family linearly increases with the size of the number of analyzable sites when this number is below circa 230 amino acids but decreases when the number of sites exceeds this threshold ([supplementary fig. S1A, Supplementary Material](#) online). Two opposite trends can explain this relation. Larger proteins have a larger number of residues interactions and offer more possibilities for coevolution. Conversely, they are also more conserved, meaning that mutations occurring in these proteins tend to be more deleterious and removed from the population, leaving less chance for compensatory mutations to occur. In agreement with this hypothesis, we find a negative correlation between protein length and the tree diameter of each family (Kendall's tau =  $-0.27$ ,  $P$ -value  $< 2.2 \times 10^{-16}$ , [supplementary fig. S1B, Supplementary Material](#) online), and a positive correlation between the tree diameter and the proportion of coevolving sites per family (Kendall's tau =  $0.17$ ,  $P$ -value  $< 2.2 \times 10^{-16}$ , [supplementary fig. S1C, Supplementary Material](#) online).

The evidence for coevolution within proteins was globally scarce since only 16% of the analyzed positions in a protein were involved in a coevolving group (median of all protein families). Coevolution, however, was found to be a general mechanism among proteins: 98% of the protein families that we analyzed had at least one coevolving group. In the following, we combined all positions from all

protein families and unraveled the factors determining the occurrence of coevolution.

### Coevolving Positions are Evolving Slowly Unless They are Coevolving Because of Charge Compensation

We assessed the evolutionary rate of coevolving positions and tested whether the detected positions were more conserved, as predicted by the coevolution model of [Talavera et al. \(2015\)](#). An alternative hypothesis is that fast-evolving positions are more likely to be detected as coevolving because of the increased statistical power stemming from the larger number of underlying mutations ([Dutheil 2012](#)). For each biochemical property, we compared the evolutionary rate (see Materials and Methods) for sites predicted to be coevolving or not. For all properties, coevolving positions had a lower evolutionary rate, except the charge property for which coevolving sites evolved faster than noncoevolving sites ([fig. 4](#)). A particularity of the charge property is its discrete nature: each amino acid is characterized by one of three possible states—positively charged, negatively charged, or neutral—while other properties are continuous. To test whether this difference in property nature could affect the capacity to detect coevolving sites, we discretized the volume and polarity indices in three categories (see Materials and Methods) and rerun the coevolution detection analysis. We found that positions coevolving for volume and polarity had a lower rate of evolution, even



**Fig. 4.** Evolutionary rates of coevolving sites. Box-whisker plots (first and third quartile, medians) of standardized rates for coevolving versus noncoevolving sites for each biochemical property. Mean values of each distribution are indicated as side labels.  $P$ -values of corresponding Wilcoxon signed-rank tests are indicated on top of each comparison.

when considering only three categories, suggesting that the discrete nature of the property is not responsible for the inferred faster evolution of coevolving charged residues (supplementary fig. S2, Supplementary Material online). We next aimed at assessing which structural characteristic impacts the propensity to coevolve, accounting for the underlying evolutionary rate of the sites.

#### There is Comparatively Less Coevolution Within Secondary Structure Motifs

We used mixed generalized linear models to assess the impact of structural properties on the propensity of sites to

coevolve. Each biochemical property was analyzed separately, and each site was considered an independent data point. The protein family was treated as a random effect, allowing sites within the same protein to share the same error distribution, while different protein families may have distinct ones. Whether each site was found to be coevolving was used as a binary response variable. Structural properties of each site were recorded, including its relative solvent accessibility (RSA), its secondary structure motif (one of  $\alpha$ -helix, 3–10 helix,  $\pi$ -helix, strand,  $\beta$ -bridge, turn, bend, intrinsically disordered or unknown), and the evolutionary rate of the site. Whilst several explanatory variables are potentially intrinsically correlated, all variance inflation

factors (VIFs) were found to be close to 1.0 and much lower than 5, an empirical threshold for identifying collinearity issues (James et al. 2013). The highest values were observed for the RSA variable, for which they are above 2.0 (supplementary table S2, Supplementary Material online).

The evolutionary rate of the sites was found to be the most significant variable (table 1). Consistent with the results above, the effect was significantly negative for all properties except charge, where it was significantly positive. Solvent exposure was generally found to have a negative impact on coevolution, meaning that coevolving residues tend to be buried. Indices 1 and 3 are an exception, though, as exposed residues have an increased probability of experiencing mutations that compensate for electric properties or alpha and turn propensities. Conversely, secondary structure was found to have little impact on the occurrence of compensatory mutations. The most substantial effect was observed for  $\beta$ -strands and  $\alpha$ -helices, where it was consistently negative: sites in helices and strands are less likely to experience compensatory mutations. We further note a significant positive interaction between solvent exposure and strand for properties “polarity”, “charge”, and index 1, indicating that for these amino acid properties, exposed residues in strands have an increased probability of coevolving. The property for which secondary structure has the strongest effect was index 3, “alpha and turn propensities”, showing a significant positive effect in turn and bend motifs, as well as in disordered regions. This index reflects the propensity of amino acids to be found in helices (high index values) versus turns (low index values). Bends, turns and disordered regions prefer amino acids with low values for index 3. A potential coevolution scenario would involve that a mutation of a preferred amino acid type toward a nonpreferred type may be compensated by a mutation at another position within the same region, involving a nonpreferred type toward a preferred amino acid type. We also report a significant negative interaction with solvent exposure in these three types of regions, suggesting that the chances for coevolution are higher for buried residues.

In the following, we more specifically studied the role of secondary and tertiary structure in the occurrence of compensatory mutations. First, we assessed whether the proximity of residues in the three-dimensional structure impacted the occurrence of coevolution. Second, we tested whether coevolving residues located in a secondary structure motif tended to be located in the same element (helix, strand, or sheet). To test these hypotheses, we developed a permutation test controlling for evolutionary rate and solvent exposure.

**Coevolving Residues are More Likely to be in Contact, Independently of Their Exposure or Evolutionary Rate** Residues at the core of proteins (low RSA) tend to be more conserved and more connected (supplementary fig. S3, Supplementary Material online). In order to test whether coevolving residues are more often in contact than

randomly selected sites, we need to compare with random positions of similar rates or solvent exposure. For that purpose, we developed a Monte-Carlo algorithm that samples groups of sites conditioned on a third variable (see Materials and Methods). As expected, we showed that sampling residues without conditioning led to a bias: randomly selected sites had a higher rate and exposure on average than the groups of sites detected as coevolving by all methods but the one accounting for the charge property, for which the opposite pattern was observed. This bias is successfully removed by using conditional sampling (supplementary fig. S4, Supplementary Material online). These results align with the observation that coevolving sites evolve more slowly and are buried, while sites detected as coevolving for charge evolve faster and are exposed.

Using the conditional sampling algorithm, we tested whether sites detected as coevolving were closer to each other in the three-dimensional structure than random sites. We computed the average pairwise three-dimensional distance between the alpha-carbons of the residues within each detected group, which we averaged over all groups. We then computed the same statistic over 1,000 random sets of groups with the same sizes and similar rates or RSA. Only 0.35% of all site groups were excluded from the randomization test due to the lack of sites with a similar rate within the same protein. Conversely, 31% had to be discarded when conditioning on RSA. We found that the observed statistic is significantly lower than in the case of random sets for all methods (fig. 5A, all  $P$ -values  $< 0.001$ ), suggesting that coevolving residues are in closer proximity than expected by chance. To further assess whether this proximity is explained by residues being in physical contact, we investigated the connectivity graph of the residues in each group. As an approximation, we considered two residues to be in contact if their  $\alpha$ -carbon distance was shorter than 8 Å and computed the proximity graph of the residues, where an edge connects two residues if they are in contact. To measure the number of contacts, we counted the number of graphs for the group: this number is equal to 1 if all residues are in contact, either directly or indirectly. If no residue is in contact with any other, the number of graphs equals the number of residues. We further standardized this measure so that it is comprised between 0 and 1 to be comparable between groups of different sizes (see Materials and Methods). We showed that this statistic is lower than expected by chance when sampling groups of sites with a similar rate or RSA (Fig. 5B, all  $P$ -values  $< 0.001$ ). Therefore, we conclude that coevolving residues are more likely to be in contact in the three-dimensional structure. This effect is not due to their slow evolutionary rate and low solvent exposure.

Using a similar approach, we further investigated whether coevolving residues in a given secondary structure motif were located in the same element. We generated two sub-datasets, only containing sites in helices or strands, respectively. We then computed the connectivity graphs in each case and considered two residues in contact if they were



**Table 1.** Impact of Structural Properties on the Probability of Sites to Coevolve. For each biochemical property, a generalized linear model with mixed effects was fitted. Whether a site was found coevolving according to the biochemical property was set as a binary response variable, and the evolutionary rate and structural properties were set as explanatory variables. The protein family was set as a random factor.

Variable	Volume	Polarity	Charge	I1	I2	I3	I4	I5	I6	I7	I8
RE(Intercept)	1.17 (***)	1.23 (***)	1.36 (***)	1.68 (***)	1.51 (***)	1.69 (***)	1.82 (***)	2.18 (***)	2.09 (***)	2.34 (***)	2.58 (***)
(Intercept)	-3.45 (***)	-3.80 (***)	-4.94 (***)	-3.63 (***)	-4.09 (***)	-4.24 (***)	-4.25 (***)	-5.18 (***)	-4.74 (***)	-5.05 (***)	-5.67 (***)
Np <sup>a</sup>	-0.35 (***)	-0.38 (***)	0.52 (***)	-1.44 (***)	-0.34 (***)	-0.71 (***)	-0.34 (***)	-0.34 (***)	-0.61 (***)	-0.34 (***)	-0.31 (***)
RSA <sup>b</sup>	-0.70 (***)	-0.33 (*)	-0.32 (*)	0.40 (**)	-0.34 (*)	0.34 (*)	-0.51 (***)	-0.28	-0.54 (**)	-0.69 (***)	-0.44 (*)
Secondary structure											
α-helix	-0.17 (***)	-0.17 (**)	0.01	-0.11 (*)	-0.15 (**)	-0.17 (**)	-0.08	0.00	0.00	-0.24 (***)	-0.24 (**)
3–10 helix	0.17 (*)	0.24 (*)	-0.03	0.19 (.)	-0.09	0.2 (.)	0.07	0.13	0.18	-0.20	-0.20
π-helix	-0.02	0.26	0.25	-0.30	0.24	0.52 (*)	0.14	0.48 (.)	0.04	0.09	0.05
Strand	-0.21 (***)	-0.26 (***)	-0.31 (***)	-0.17 (**)	-0.17 (**)	-0.26 (***)	-0.13 (*)	-0.05	-0.09	-0.11	-0.18 (*)
β-bridge	-0.24	-0.23	-0.15	0.03	-0.09	-0.13	-0.23	0.00	-0.04	0.04	-0.07
Turn	0.16 (*)	0.14 (.)	0.16 (.)	0.06	-0.06	0.18 (*)	0.20 (*)	-0.03	0.12	-0.09	-0.06
Bend	0.12 (.)	0.11	0.24 (*)	0.08	-0.06	0.24 (**)	0.14 (.)	0.14	0.04	-0.23 (*)	-0.03
Disordered	-0.22 (*)	0.09	0.25 (*)	0.12	-0.08	0.40 (***)	-0.02	-0.03	0.06	-0.13	-0.11
RSA:secondary structure											
α-helix	-0.22	0.11	0.14	0.28	-0.17	-0.45 (*)	-0.52 (**)	-0.38 (.)	0.15	0.01	-0.17
3–10 helix	-0.37	-0.54 (.)	0.10	-0.20	0.20	-0.71 (*)	-0.32	-0.08	0.24	0.74 (*)	0.45
π-helix	-0.29	-0.16	0.15	1.77 (*)	-0.20	-2.19 (.)	-0.59	-1.26	0.41	0.43	0.48
Strand	0.13	0.68 (***)	1.22 (***)	0.8 (***)	0.29	0.06	-0.06	0.37	0.38	0.35	0.23
β-bridge	0.47	0.91	0.03	0.86	-0.08	0.39	-0.51	0.39	-0.46	-0.22	0.25
Turn	-0.23	-0.53 (*)	-0.39 (.)	-0.44 (*)	-0.35	-0.65 (**)	-0.53 (*)	-0.19	-0.12	-0.06	-0.15
Bend	-0.05	-0.14	-0.31	-0.09	0.11	-0.72 (**)	-0.29	-0.39	0.07	0.24	-0.13
Disordered	0.59 (*)	-0.11	-0.23	-0.12	0.13	-0.65 (*)	0.04	0.11	0.08	0.52 (.)	0.03

Significance levels for P-values: (.)  $P < 0.10$ , (\*)  $P < 0.05$ , (\*\*)  $P < 0.01$ , and (\*\*\*)  $P < 0.001$ .

<sup>a</sup>Standardized norm of the weighted substitution vector (evolutionary rate).

<sup>b</sup>Relative solvent accessibility (solvent exposure).

in the same helix, strand, or sheet. By sampling among helix or strand sites within each protein family, we computed the expected distribution of these statistics. We also compared the results with randomization where sites were constrained to have similar evolutionary rate or RSA. We report consistent results between these analyses (fig. 6): coevolving sites located within helices are significantly more often in the same helix than expected by chance (fig. 6A).

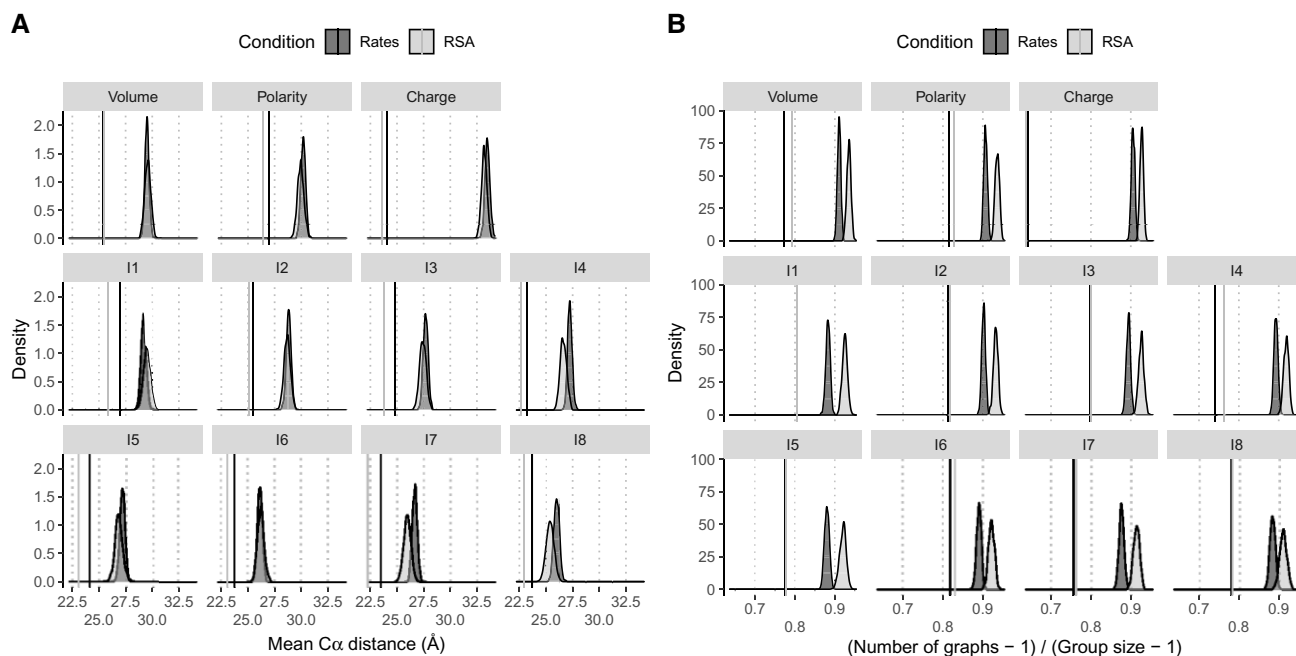
Conversely, coevolving sites within strands are not more often in the same one (fig. 6B). They are also generally not more often in the same sheet, except sites coevolving for charge (fig. 6C). These results allow us to refine our previous conclusions: while there are fewer compensatory mutations in secondary structure motifs, this level of organization creates evolutionary constraints susceptible to induce coevolution. This is particularly clear for helices, while only charge compensation could be evidenced in the case of β-sheets.

## Discussion

We generated an atlas of protein compensatory mutations, applying a phylogenetic method on a curated set of protein alignments for which a representative three-dimensional structure was available. We employed a substitution mapping procedure, inferring the position of all amino acid substitutions that occurred at every site of the alignment and every branch of the underlying phylogeny. We used indices of amino acid biochemical properties to build a proxy for the fitness effect of amino acid substitutions and assess their compensatory nature (Dutheil and Galtier 2007). The statistical procedure used to assess the significance of the coevolution signal accounts for the phylogenetic correlation and the site-specific rate of evolution, addressing some of the common pitfalls in predicting site associations (de Juan et al. 2013).

As alignment errors can induce a false signal of coevolution (Dickson et al. 2010), our analysis pipeline included a stringent quality control and filtering out of missing data and ambiguously aligned positions. While ensuring that the detected coevolving groups are not artifacts of the alignment process, this conservative approach may bias our site sample toward slowly evolving positions. To account for this possibility, the underlying statistics (linear models, randomization) were carried on the set of sites included in the coevolution analysis only. Another putative source of error is the phylogeny reconstruction. We note that the alignment filtering procedure also acts positively on the phylogeny inference (Md Mukarram Hossain et al. 2015), and that the sequence selection procedure to remove highly similar sequences also warrants a minimum phylogenetic signal. Despite these precautions, local branching uncertainty may persist in some of the inferred phylogenies. Because our analyses integrate over hundreds of protein families, we expect such residual uncertainty to act as a source of anisotropic noise, and not to create a statistical bias. Possible additional complementary strategies would involve computing branch support values and collapsing of the underlying nodes, or generating consensus sets over coevolution predictions obtained by distinct but equally likely phylogenies.

We found the number of positions with a significant coevolution signal to be generally low (14% of sites are part of a coevolution group), and coevolving positions were generally evolving slowly, showing a higher-than-average level of conservation (the median evolutionary rate rank per family and per method over the full dataset was 4.5 for the coevolving positions, while it was 125 for the noncoevolving ones). These results indicate that compensatory mutations are rare, as predicted by theoretical models (Talavera et al. 2015). Most of the families that we tested, however, exhibited some coevolving positions, showing



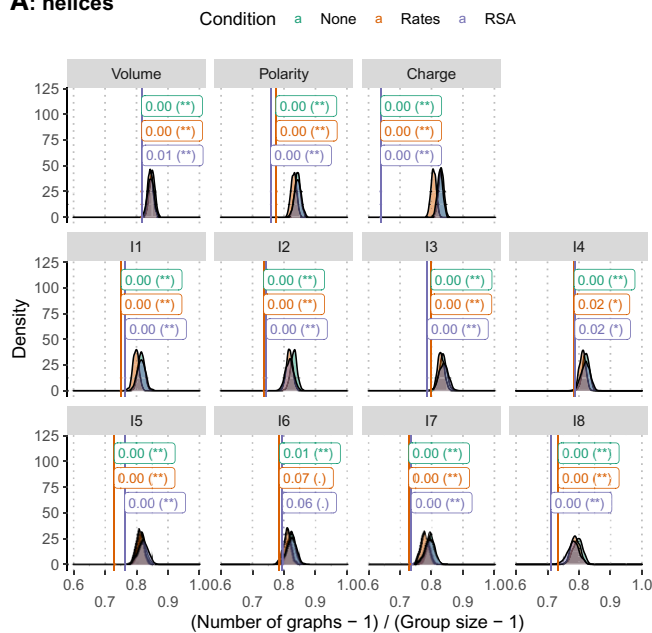
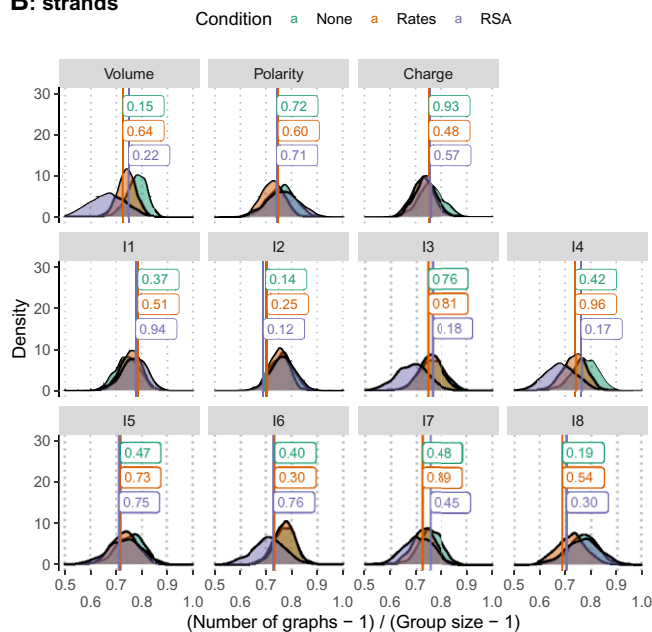
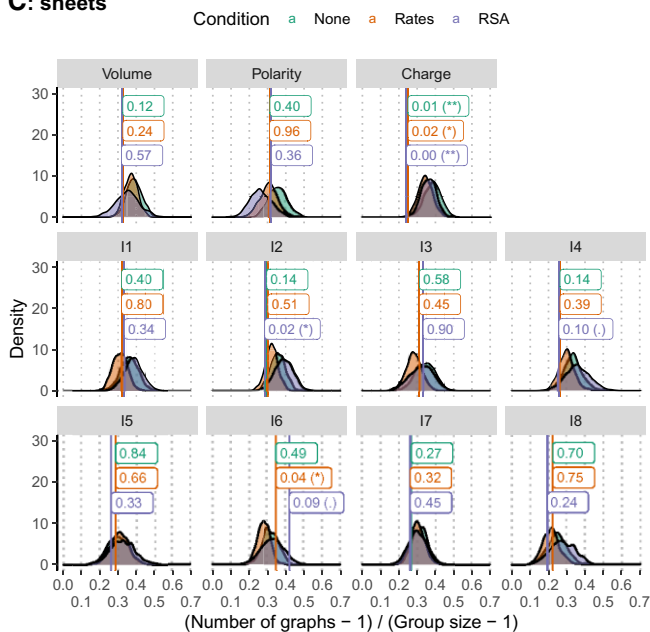
**FIG. 5.** Three-dimensional proximity of coevolving sites. Average C $\alpha$  distance (A) and relative number of contact groups (B) of all groups detected to be coevolving by compensation for each biochemical property. Densities are computed from 1,000 samplings, conditioned over the evolutionary rate or the solvent exposure (see Material and Methods). Only groups for which sites with a similar rate or RSA values could be sampled were included. Observed values are indicated as vertical bars and are all significantly lower than expected by chance, with a  $P$ -value  $< 1/1001$ .

that while rare, compensatory mutations are widespread among proteins with distinct structural and functional properties. From a genetic perspective, the mechanism of compensatory mutations implies that the negative fitness effect of a first mutation is compensated by the positive fitness effect of a second mutation. The fitness effect of the two mutations, therefore, depends on the order in which they occur: the same mutation may have a negative effect if it appears first or positive if it happens second, compensating for a preceding deleterious mutation. Compensatory mutations constitute a particular case of nonadditive fitness, termed reciprocal sign epistasis, which induces valleys of low fitness in fitness landscapes (see Whitlock et al. (1995) for a review, and Poelwijk et al. (2011)). Our empirical results suggest that reciprocal sign epistasis is a general feature of protein-coding gene fitness landscapes, independent of the function of the encoded proteins.

Our analyses confirm a strong link between protein structure and coevolution of residues (de Juan et al. 2013). In particular, our results highlight the importance of tertiary structure in shaping the fitness landscape of proteins: coevolving residues tend to be located outside secondary structure motifs, at the hydrophobic core of proteins, and generally in close proximity. Importantly, we show that the link between structural properties and patterns of sites coevolution is not a byproduct of the structural constraints shaping site-specific evolutionary rates (Talavera et al. 2015): coevolving residues are more likely to be physically in contact than random sites with an equal rate or solvent exposure. While failing to account

for evolutionary rates may lead to sites being falsely labeled as coevolving, in particular when large amounts of such sites are predicted for a given protein, this fact does not imply that compensatory mutations do not occasionally occur and cannot be detected.

The patterns of coevolving positions with respect to site properties appeared to be consistent across the range of biochemical properties that we used as a fitness proxy. This robustness extends to the properties selected from a multi-dimensional analysis of more than 500 properties available, ensuring that our conclusions are not biased toward any prior assumption on which biochemical property might be relevant. A notable exception is the pattern of sites detected as coevolving for charge compensation, which have a higher substitution rate than noncoevolving sites. Coevolution for charge compensation mirrors the pattern of coevolution in RNA, where Watson–Crick pairs in double-stranded helices evolve at a faster rate than single-stranded regions and pairs interacting via non-Watson–Crick interactions (Dutheil et al. 2010). It is particularly intriguing that distinct biochemical constraints lead to opposite patterns of evolutionary rates, suggesting that the frequency at which fitness valleys are crossed in a fitness landscape cannot be fully predicted by simple models with fixed selection coefficients. A key to the understanding of these dynamics may lie in the accounting of selective constraints acting on intermediate organizational levels. This seems to be the case in RNA, where selective forces may act at the level of double-stranded stems (Dutheil et al. 2010). Here, however, we show that secondary structure motifs comparatively seem to play little role in the occurrence of coevolving positions in proteins, where intermediate levels of

**A: helices****B: strands****C: sheets**

**FIG. 6.** Are coevolving positions located within the same secondary structure element? (A) Helices, (B) strands, and (C) sheets. The x-axis measures the relative number of secondary structure elements: a value of 0 indicates that all sites within a coevolving group are in the same element, while a value of 1 is obtained when each site is in a distinct element. Densities are computed from 1,000 samplings, conditioned over the secondary structure element alone (None), secondary structure and evolutionary rate (Rate), or secondary structure and solvent exposure (RSA, see Material and Methods). Only groups for which sites with a similar rate or RSA values could be sampled were included. Observed values are indicated as vertical bars together with the corresponding *P*-values.

organization pertinent to molecular coevolution are only starting to be characterized (Halabi et al. 2009).

## Materials and Methods

### Sequence Retrieval

Protein sequences were retrieved from the HOGENOM database (release 06) (Penel et al. 2009), which contains families of homologous protein-coding genes from

completely sequenced genomes of Bacteria, Archaea and Eukaryotes. We sampled the sequences in all three domains of life and found that only Bacteria had enough sequences to provide sufficient signal for coevolution detection. Thus, we targeted members of the Bacteria domain for this study. Using the “query\_win” retrieval program (Gouy and Delmotte 2008), 2,047 families were selected for which at least one bacterial sequence had at least one experimentally solved structure available. All the Archaea and Eukaryotic

sequences were removed from the selected families. For each family, species having more than one sequence were discarded to avoid the comparison of paralogous sequences. To ensure the comparison of families with similar statistical power, we discarded families with less than 100 sequences.

We developed a pipeline to process each protein family, with the goal to minimize the amount of missing data and maximize the alignment reliability. Phylogenetic inference relies on alignment quality, but alignment filtering depends on the knowledge of the phylogenetic tree. To solve this conundrum, the pipeline uses an iterative approach where fast alignment and phylogeny building tools are used first to obtain a filtered set of sequences, and more accurate but computationally demanding methods are used in a second step:

- 1) *Approximate alignment and phylogeny*: a fast multiple sequence alignment for each family was performed using Clustal Omega (version 1.2.4) (Sievers et al. 2011). For each alignment, a first phylogenetic tree was built with FastTree (version 2.1.11) (Price et al. 2010).
- 2) *Minimization of missing data*: the phylogenetic tree obtained at step 1 was used to remove sequences from the alignment in order to maximize the number of sites with sufficient coverage using the “bppAlnOptim” program (version 1.1.0) (Dutheil and Figuet 2015). Unresolved characters were considered together with gaps in coverage calculations, and only sites without gaps were kept.
- 3) *High-quality alignment*: filtered families were realigned using both Muscle (version 3.8) (Edgar 2004) and Clustal Omega (Sievers et al. 2011). We built a consensus alignment by computing sum-of-pairs scores (SPS) for each alignment site and masking sites with an SPS lower than 80% using the “bppAlnScore” program from the Bio++ Program Suite (version 2.4.0) (Guéguen et al. 2013). These high-quality consensus alignments were used in the following analyses.
- 4) *Phylogenetic sampling*: we performed a two-step phylogenetic sampling, first to remove highly similar sequences, which do not carry a biological signal, and second to limit the maximum number of sequences in each family in order to reduce computational time. The “bppPhySamp” program from the Bio++ program suite was used to ensure that sequences in each alignment were at least 1% different from the other sequences and that there is a maximum of 400 sequences per alignment. Families with less than 100 sequences after sampling were further discarded, leaving a total of 1,684 families.
- 5) *Accurate phylogeny reconstruction*: the resulting masked alignments were used to build a maximum likelihood phylogenetic tree, using the PhyML program (version 3) (Guindon and Gascuel 2003; Guindon et al. 2010). The LG substitution model was shown to provide an improved fit on a large

set of alignments and was, therefore, selected for the analysis of all our protein families (Le and Gascuel 2008). We further used a discrete gamma distribution of substitution rates with four categories (Yang 1994) as a compromise between model-fit and computational efficiency. PhyML 3 implements several tree search algorithms. We selected the option that retains the best tree from the nearest neighbor interchange and subtree pruning regrafting topology estimation algorithms (Felsenstein 2003). The inferred phylogenies were used for the identification of coevolving sites. Unless explicitly stated otherwise, all programs were run with their default settings.

### Identification of Coevolving Sites

Coevolving positions within each protein family were predicted using the CoMap package (version 1.5.2) (Dutheil and Galtier 2007). The built-in algorithm was then used to correct for multiple testing and estimate the false discovery rate (FDR). In this study, 10,000 simulations per tested group were performed, and the groups with an adjusted  $P$ -value  $< 1\%$  were considered as coevolving. Compensation was used as a measure of coevolution that assesses the compensatory nature of cosubstitutions based on a given physicochemical property. The indexes used as weights were retrieved from the AAindex database (Kawashima et al. 2008). Three commonly studied properties: Volume (as defined by Grantham, AAindex id: GRAR740103); Polarity (as defined by Grantham, AAindex id: GRAR740102); Charge (as defined by Klein, AAindex id: KLEP840101) were tested. In addition, we also tested eight nonoverlapping properties as described by Saha et al. (2012). Saha et al. have categorized all published 544 amino acid indexes of the AAindex database into eight clusters, and these eight categories were used in this study: Electric property; Hydrophobicity; Alpha and turn propensities; Physicochemical properties; Residue propensity; Composition; and Beta propensity and Intrinsic propensities. Because the charge property has three discrete states (positively charged, negatively charged, and neutral), we further assess the impact of discrete versus continuous properties on the detection of coevolution by discretizing the volume and polarity properties. Residues G (volume (Grantham 1974): 3), A (31), S (32), and P (32.5) were considered as small, residues D (54), C (55), N (56), T (61), E (83), Q (85), and V (84) as medium, and residues H (96), M (105), I (111), L (111), K (119), R (124), F (132), Y (136), and W (170) as large. Residues L (polarity (Grantham 1974): 4.9), I (5.2), F (5.2), W (5.4), C (5.5), M (5.7), V (5.9), and Y (6.2) were classified as hydrophobic, residues P (8.0), A (8.1), T (8.6), G (9.0), and S (9.2) as intermediate, and residues H (10.4), Q (10.5), R (10.5), K (11.3), N (11.6), E (12.3), and D (13.0) as polar.

Out of 1,684 families, 24 families were discarded because likelihoods could not be computed because of numerical underflow. Numerical saturation, unavailability of secondary



structure or solvent accessibility information (see below) resulted in a final total of 1,630 analyzable protein families (supplementary table S1, Supplementary Material online). The diameter of each phylogenetic tree (maximum distance between any two leaves) as well as the median of sequence lengths for each family were computed using the “ape” (Paradis et al. 2004) and “seqinr” (Charif et al. 2005) packages for R, and recorded. The relationship between the number of detected sites and the alignment length was assessed using a broken regression model, as implemented in the “lm.br” package for R (Adams 2017).

### Post hoc Tests of Branch-Specific Compensatory Substitutions

We developed two statistical *post hoc* tests in order to better characterize the signal of coevolution for a group of candidate coevolving sites. Following the notations from the introduction, these tests compute the compensation statistic for a group of sites  $G$ , at a specific branch  $j$ . We note as  $x_{ij}^P$  the change of a biochemical property  $P$  for site  $i$  on branch  $j$ . The compensation index  $C_j(G)$  for a group of sites  $G$  on branch  $j$  is then computed as

$$C_j(G) = 1 - \frac{\left| \sum x_{ij}^P \right|}{\sum |x_{ij}^P|}.$$

This measure is identical to the compensation index defined in the introduction for the full tree but applied to a single branch. We introduce two permutation tests to assess the significance of the compensation statistics. The first test randomizes the branches for each site, conditioning on the total amount of change at each site. The second test randomizes the sites for each branch, conditioning on the branch length. In both tests, the conservation statistic is computed for each permutation and its distribution compared to the observed value.

### Evolutionary Rates

We measured the rate of molecular evolution according to a given biochemical property using the norm of the weighted substitution vector for each site (Dutheil and Galtier 2007). For a given site in the alignment, the number of substitutions that occurred on each branch of the phylogenetic tree was inferred by probabilistic substitution mapping. Each substitution type from an amino acid  $A$  into an amino acid  $B$  was then weighted by the absolute difference in the amino acid property for  $A$  and  $B$ .

### Structural Information

The PDB ids of the structures associated with each family were extracted from the HOGENOM database and the corresponding PDB files were retrieved from the Protein Data Bank (PDB) (Berman et al. 2000). For each protein family, all pairwise alignments of each homologous sequence with each subunit of each matching PDB structure were generated, and the PDB subunit leading to the highest alignment

score was used as a reference structure for the protein family. The corresponding alignment was used to translate alignment positions into positions in the PDB structure. All scripts used for analyses were written in Python version 3 using the Biopython modules (Hamelryck and Manderick 2003). The reference PDB structures were then subsequently used to extract structural properties: secondary structure and RSA were obtained using the DSSP program (Kabsch and Sander 1983), run via the Biopython DSSP module. The distance of each residue to the solvent-accessible surface was computed via the MSMS program (Sanner et al. 1996), accessed via the “Biopython.ResidueDepth” module. Both the residue depth and the  $C\alpha$  depths were computed. We further computed the number of residues in contact with each residue using  $C\alpha$  distance thresholds of 5, 8, and 10 Å. Because several of these properties may be intrinsically correlated, we examined their relationships using principal component analysis and nonparametric pairwise correlation (Spearman’s rank correlation). We found that the structural variables were highly correlated (supplementary fig. S3, Supplementary Material online): residue depths and numbers of contacts were highly positively correlated while highly negatively correlated with RSA. In summary, the more exposed a residue is, the closer it is to the surface of the protein and the less contacts it has with other residues. Because all these variables carry the same biological information, we only used RSA in the statistical analyses.

Intrinsic disorder measures for all residues were obtained using the DisEMBL version 1.4 program (Linding et al. 2003), using default parameter values as suggested by the program and after porting the original Python script to Python 3.0. We considered a site to be in a disordered region if its holoop index was greater than 0.1204 and the site was not predicted by DSSP to be in a secondary structure motif. We considered sites labeled as disordered as not having any secondary structure, and we combined the results of DSSP and DisEMBL into a single discrete variable with states “no structure”, “alpha helix”, “3–10 helix”, “pi helix”, “strand”, “beta bridge”, “turn”, “bend”, and “disordered”.

In order to test whether the coevolving sites of a group are in proximity in the three-dimensional structure, we measured the distance between the two alpha carbon atoms of the residues. For groups with more than two sites, the average pairwise distance was computed. As an alternative measure of proximity, we considered residues with a distance  $\leq 8$  Å to be in contact. We then computed the number of subgroups of sites in contact within a group: if all residues are in contact with each other, then the number of subgroups is one. If each residue is distant from all others, then the number of subgroups is equal to the size of the group. We further standardized this measure by removing 1 and further dividing by the group size—1, so that it is comprised between 0 (all residues in contact) and 1 (all residues apart). We note  $N_{\text{sub}}$  the resulting statistic, averaged over all candidate groups. Testing of residues proximity was performed using a dedicated Monte Carlo procedure. For each detected group in each protein family, a group of identical size was sampled among analyzed positions of the same protein family. In order to further

enforce that sampled positions have evolutionary rates similar to the ones of the observed group, we performed a conditional sampling. For each site in each group, called the “focus” site, we sampled among the subset of sites with a rate no more than  $x\%$  different from the rate of the focus site,  $x$  being an adjustable similarity threshold. We further ensured that sites were not sampled more than one time in each group, although a site in a protein was allowed to be sampled multiple times between distinct groups. Because the distribution of rates is typically skewed, we further restricted the sampled set of sites so that it contained as many values below and above the rate of the focus site. When less than five sites with similar rates were available, the group was not further considered in the randomization test. We first generated 100 replicates without conditioning, with a 20% similarity threshold and with a 10% threshold for comparison. We found that the 10% threshold removed any significant effect of the evolutionary rate ([supplementary fig. S4, Supplementary Material](#) online). We then generated 1,000 replicates per predicted group with the 10% threshold and computed  $P$ -values as  $(N + 1)/1001$ , where  $N$  is the number of replicates with a measure greater or equal (or lower or equal, depending on the test) to the observed value in the data. A similar procedure was conducted using RSA instead of the evolutionary rate.

To test whether coevolving residues were located within the same secondary motif (helix, strand, or sheet), we applied a similar randomization procedure after discarding all sites not located within secondary structures. In order to annotate sites with their secondary structure labels (helix and strand numbers), mmCIF files were used, and annotations were parsed using scripts developed with the “BioPython.PDB” package ([Hamelryck and Manderick 2003](#)). We computed the  $N_{\text{sub}}$  statistic (see above), considering whether sites are in the same motif or not. Sites in helices and strands were analyzed separately, and sampling was done in each case without further conditioning or by conditioning on evolutionary rates or RSA, with a 10% similarity threshold and with 1,000 replicates. Analyses were run on a Linux workstation using the “GNU parallel” software ([Tange 2011](#)).

### Statistical Analyses

All statistical analyses were conducted with the R statistical software ([R Core Team 2020](#)) using the “ggplot2” ([Wickham 2016](#)), “cowplot” ([Wilke 2020](#)), and “ggpubr” ([Kassambara 2020](#)) packages for results visualization. Methods overlap ([fig. 3](#)) was plotted using the “UpSetR” package ([Gehlenborg 2019](#)). To assess whether the number of sites detected as coevolving by two or more methods was significant, we defined the statistic  $S$  as the number of sites detected by at least two methods. We randomized the positions detected as coevolving independently for each method and computed  $S$  for the pseudo dataset. We repeated the procedure 10,000 times and computed a  $P$ -value as  $(x + 1)/10001$ , where  $x$  is the number of cases where  $S$  in the randomized data sets is at least equal to the value of  $S$  on the nonrandomized data.

We fitted generalized linear models for all analyzed sites independently for each type of biochemical weight used in the coevolution prediction. Whether a site was predicted as part of a coevolving group was used as a binary response variable, and the evolutionary rate of the site, its relative solvent exposure (RSA) in the protein structure, the secondary structure motif and the interaction between RSA and secondary structure were used as putative explanatory variables. As the evolutionary rate measure depends on the phylogenetic tree, we standardized them by dividing the measure by the respective family average. To account for correlated errors of sites within protein families, we further added the protein family as a random factor, resulting in a generalized linear model with mixed effects (GLMM). GLMM analyses were conducted in the R statistical software ([R Core Team 2020](#)).

The “lme4” package for R ([Bates et al. 2015](#)) failed to estimate parameters for our models. We further tried different estimation procedures: (1) penalized quasi-likelihood, using the “glmmPQL” function from the “MASS” package ([Venables and Ripley 2002](#)), and (2) Laplace approximation, (3) adaptive Gaussian quadrature, and (4) sequential reduction with the “glmm” function from the “glmsr” package ([Ogden 2019](#)). All methods converged to the same model parameters, but  $P$ -values could only be obtained with the “glmmPQL” method after increasing the default number of iterations. In one case (index 8), convergence was not reached after 100 iterations. We report the results from the Laplace method but used the “glmmPQL” fitted models to compute VIFs using the “vif” function from the “car” package ([Fox and Weisberg 2019](#)), which is not compatible with the “glmsr” package.

### Supplementary Material

[Supplementary data](#) are available at *Molecular Biology and Evolution* online.

### Acknowledgments

The authors would like to thank Simon Penel for helping with the HOGENOM database. JYD acknowledges funding from the Max Planck Society.

### Data Availability

All sequence alignments, phylogenies, and coevolution detection results for all families are available at <https://doi.org/10.6084/m9.figshare.16586915>. The scripts necessary for producing the statistical analyses and figures of this manuscript are available at <https://gitlab.gwdg.de/molsysevol/coevolution/atlas-of-coevolution>.

### References

Adams M. 2017. lm.br: Linear Model with Breakpoint. Available from: <https://CRAN.R-project.org/package=lm.br>

- Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW. 2000. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol Biol Evol.* **17**:164–178.
- R Core Team. 2020. *R: a language and environment for statistical computing*. Vienna (Austria): R Foundation for Statistical Computing. Available from: <https://www.R-project.org/>
- Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. *J Stat Softw.* **67**:1–48.
- Behdenna A, Pothier J, Abby SS, Lambert A, Achaz G. 2016. Testing for independence between evolutionary processes. *Syst Biol.* **65**:812–823.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. *Nucl Acids Res.* **28**:235–242.
- Charif D, Thioulouse J, Lobry JR, Perrière G. 2005. Online synonymous codon usage analyses with the ade4 and seqinR packages. *Bioinformatics* **21**:545–547.
- Chen Y, Carlini DB, Baines JF, Parsch J, Braverman JM, Tanda S, Stephan W. 1999. RNA secondary structure and compensatory evolution. *Genes Genet Syst.* **74**:271–286.
- Dawson A, Chen M, Fyfe PK, Guo Z, Hunter WN. 2010. Structure and reactivity of *Bacillus subtilis* MenD catalyzing the first committed step in menaquinone biosynthesis. *J Mol Biol.* **401**:253–264.
- de Juan D, Pazos F, Valencia A. 2013. Emerging methods in protein co-evolution. *Nat Rev Genet.* **14**:249–261.
- Dib L, Silvestro D, Salamin N. 2014. Evolutionary footprint of co-evolving positions in genes. *Bioinformatics* **30**:1241–1249.
- Dickson RJ, Wahl LM, Fernandes AD, Gloor GB. 2010. Identifying and seeing beyond multiple sequence alignment errors using intra-molecular protein covariation. *PLoS ONE* **5**:e11082.
- Di Lena P, Nagata K, Baldi P. 2012. Deep architectures for protein contact map prediction. *Bioinformatics* **28**:2449–2457.
- Dimmic MW, Hubisz MJ, Bustamante CD, Nielsen R. 2005. Detecting coevolving amino acid sites using Bayesian mutational mapping. *Bioinformatics* **21**(Suppl 1):i126–i135.
- Dunn SD, Wahl LM, Gloor GB. 2008. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24**:333–340.
- Dutheil JY. 2012. Detecting coevolving positions in a molecule: why and how to account for phylogeny. *Brief Bioinform.* **13**:228–243.
- Dutheil JY, Figuet E. 2015. Optimization of sequence alignments according to the number of sequences vs. number of sites trade-off. *BMC Bioinform.* **16**:190.
- Dutheil J, Galtier N. 2007. Detecting groups of coevolving positions in a molecule: a clustering approach. *BMC Evol Biol.* **7**:242.
- Dutheil JY, Jossinet F, Westhof E. 2010. Base pairing constraints drive structural epistasis in ribosomal RNA sequences. *Mol Biol Evol.* **27**:1868–1876.
- Dutheil J, Pupko T, Jean-Marie A, Galtier N. 2005. A model-based approach for detecting coevolving positions in a molecule. *Mol Biol Evol.* **22**:1919–1928.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl Acids Res.* **32**:1792–1797.
- Felsenstein J. 2003. *Inferring phylogenies*. 2nd ed. Sinauer Associates, Inc.
- Fox J, Weisberg S. 2019. *An R companion to applied regression*. 3rd ed. Thousand Oaks (CA): Sage. Available from: <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>
- Gehlenborg N. 2019. UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets. Available from: <https://CRAN.R-project.org/package=UpSetR>
- Giraud MF, Leonard GA, Field RA, Berlind C, Naismith JH. 2000. RmlC, the third enzyme of dTDP-L-rhamnose pathway, is a new class of epimerase. *Nat Struct Biol.* **7**:398–402.
- Gouy M, Delmotte S. 2008. Remote access to ACNUC nucleotide and protein sequence databases at PBIL. *Biochimie* **90**:555–562.
- Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* **185**:862–864.
- Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, Bigot T, Fournier D, Pouyet F, Cahais V, et al. 2013. Bio++: efficient extensible libraries and tools for computational molecular evolution. *Mol. Biol. Evol.* **30**:1745–1750.
- Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* **59**:307–321.
- Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* **52**:696–704.
- Halabi N, Rivoire O, Leibler S, Ranganathan R. 2009. Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**:774–786.
- Hamelryck T, Manderick B. 2003. PDB file parser and structure class implemented in Python. *Bioinformatics* **19**:2308–2310.
- Ivankov DN, Finkelstein AV, Kondrashov FA. 2014. A structural perspective of compensatory evolution. *Curr Opin Struct Biol.* **26**:104–112.
- James G, Witten D, Hastie T, Tibshirani R. 2013. *An introduction to statistical learning: with applications in R*. 1st ed. 2013, Corr. 7th printing 2017 ed. New York: Springer.
- Jones DT, Buchan DWA, Cozzetto D, Pontil M. 2012. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**:184–190.
- Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**:2577–2637.
- Kassambara A. 2020. ggpubr: “ggplot2” Based Publication Ready Plots. Available from: <https://CRAN.R-project.org/package=ggpubr>
- Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. 2008. AAindex: amino acid index database, progress report 2008. *Nucl Acids Res.* **36**:D202–D205.
- Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol.* **25**:1307–1320.
- Li Y, Hu J, Zhang C, Yu D-J, Zhang Y. 2019. ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics* **35**:4647–4655.
- Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, Bornberg-Bauer E, Colwell LJ, de Koning APJ, Dokholyan NV, Echave J, et al. 2012. The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci.* **21**:769–785.
- Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. 2003. Protein disorder prediction: implications for structural proteomics. *Structure* **11**:1453–1459.
- Md Mukarram Hossain AS, Blackburne BP, Shah A, Whelan S. 2015. Evidence of statistical inconsistency of phylogenetic methods in the presence of multiple sequence alignment uncertainty. *Genome Biol Evol.* **7**:2102–2116.
- Moutinho AF, Trancoso FF, Dutheil JY. 2019. The impact of protein architecture on adaptive evolution. *Mol. Biol. Evol.* **36**:2013–2028.
- Neher E. 1994. How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci USA.* **91**:98–102.
- Ogden H. 2019. glmmsr: Fit a Generalized Linear Mixed Model. Available from: <https://CRAN.R-project.org/package=glmmsr>
- Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**:289–290.
- Penel S, Arigon A-M, Dufayard J-F, Sertier A-S, Daubin V, Duret L, Gouy M, Perrière G. 2009. Databases of homologous gene families for comparative genomics. *BMC Bioinform.* **10**(Suppl 6):S3.
- Poelwijk FJ, Tănase-Nicola S, Kiviet DJ, Tans SJ. 2011. Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *J Theor Biol.* **272**:141–144.
- Pollock DD, Taylor WR. 1997. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Eng.* **10**:647–657.
- Pollock DD, Taylor WR, Goldman N. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. *J Mol Biol.* **287**:187–198.

- Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**: e9490.
- Saha I, Maulik U, Bandyopadhyay S, Plewczynski D. 2012. Fuzzy clustering of physicochemical and biochemical properties of amino acids. *Amino Acids* **43**:583–594.
- Sanner MF, Olson AJ, Spehner JC. 1996. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **38**: 305–320.
- Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol Syst Biol*. **7**:539.
- Smit S, Widmann J, Knight R. 2007. Evolutionary rates vary among rRNA structural elements. *Nucl Acids Res*. **35**:3339–3354.
- Starr TN, Thornton JW. 2016. Epistasis in protein evolution. *Protein Sci*. **25**:1204–1218.
- Storz JF. 2018. Compensatory mutations and epistasis for protein function. *Curr Opin Struct Biol*. **50**:18–25.
- Talavera D, Lovell SC, Whelan S. 2015. Covariation is a poor measure of molecular coevolution. *Mol Biol Evol*. **32**: 2456–2468.
- Tange O. 2011. GNU parallel – the command-line power tool. *USEUNIX Mag*. **361**:42–47.
- Tataru P, Hobolth A. 2011. Comparison of methods for calculating conditional expectations of sufficient statistics for continuous time Markov chains. *BMC Bioinform*. **12**:465.
- Tetchner S, Kosciółek T, Jones DT. 2014. Opportunities and limitations in applying coevolution-derived contacts to protein structure prediction. *Bio-Algor Med-Syst*. **10**:243–254.
- Tufféry P, Darlu P. 2000. Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. *Mol. Biol. Evol*. **17**:1753–1759.
- Venables WN, Ripley BD. 2002. *Modern applied statistics with S*. New York: Springer. Available from: <http://link.springer.com/10.1007/978-0-387-21706-2>
- Wang S, Sun S, Li Z, Zhang R, Xu J. 2017. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput Biol*. **13**:e1005324.
- Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. 2009. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci USA*. **106**:67–72.
- Whitlock MC, Phillips PC, Moore FB-G, Tonsor SJ. 1995. Multiple fitness peaks and epistasis. *Ann Rev Ecol Syst*. **26**: 601–629.
- Wickham H. 2016. *ggplot2: elegant graphics for data analysis*. New York: Springer-Verlag. Available from: <https://ggplot2.tidyverse.org>
- Wilke CO. 2020. cowplot: Streamlined Plot Theme and Plot Annotations for “ggplot2”. Available from: <https://CRAN.R-project.org/package=cowplot>
- Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol*. **39**:306–314.