




Article

Structural and Computational Characterization of Disease-Related Mutations Involved in Protein-Protein Interfaces

Dàmaris Navío ^{1,†} , Mireia Rosell ^{1,†}, Josu Aguirre ², Xavier de la Cruz ^{2,3} and Juan Fernández-Recio ^{1,4,5,*}

¹ Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain; damarisnavio@gmail.com (D.N.); mireia.rosell@bsc.es (M.R.)

² Vall d'Hebron Institute of Research (VHIR), Universitat Autònoma de Barcelona, 08035 Barcelona; Spain; josu.aguirre@vhir.org (J.A.); xavier.delacruz@vhir.org (X.d.l.C.)

³ Institutió Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain

⁴ Institut de Biologia Molecular de Barcelona (IBMB), Consejo Superior de Investigaciones Científicas (CSIC), 08028 Barcelona, Spain

⁵ Instituto de Ciencias de la Vid y del Vino (ICVV), CSIC—Universidad de La Rioja—Gobierno de La Rioja, 26007 Logroño, Spain

* Correspondence: juan.fernandezrecio@icvv.es

† These authors contributed equally to this work.

Received: 1 March 2019; Accepted: 27 March 2019; Published: 29 March 2019



Abstract: One of the known potential effects of disease-causing amino acid substitutions in proteins is to modulate protein-protein interactions (PPIs). To interpret such variants at the molecular level and to obtain useful information for prediction purposes, it is important to determine whether they are located at protein-protein interfaces, which are composed of two main regions, core and rim, with different evolutionary conservation and physicochemical properties. Here we have performed a structural, energetics and computational analysis of interactions between proteins hosting mutations related to diseases detected in newborn screening. Interface residues were classified as core or rim, showing that the core residues contribute the most to the binding free energy of the PPI. Disease-causing variants are more likely to occur at the interface core region rather than at the interface rim ($p < 0.0001$). In contrast, neutral variants are more often found at the interface rim or at the non-interacting surface rather than at the interface core region. We also found that arginine, tryptophan, and tyrosine are over-represented among mutated residues leading to disease. These results can enhance our understanding of disease at molecular level and thus contribute towards personalized medicine by helping clinicians to provide adequate diagnosis and treatments.

Keywords: protein-protein interactions; single amino acid variants; structural bioinformatics; computational docking; interface prediction

1. Introduction

Large-scale sequencing initiatives such as the 1000 Genomes Project [1] and the Exome Sequencing Project together with the important drop experienced by next-generation sequencing (NGS) costs [2] have provided a significant source of genomic information for an increasing number of healthy individuals and patient populations. In this context, understanding the molecular-level impact of genetic variants and its relationship to disease would further contribute to bridging the gap between genotypes and phenotypes and thus improve methods of prevention, diagnosis, and treatment of pathological conditions [3].

The predominant source of genetic variations detected in human population comes from single nucleotide variants (SNVs), which imply one single base substitution in the genome. Among SNVs, those that result in amino acid substitutions at the protein sequence level might be associated with genetic diseases since they can alter protein stability, interfere with protein-protein interaction properties [4–6], eliminate catalytic activity, affect protein folding [7], or lead to aggregation [8]. It has been estimated that around 58% of the ~13,000 exonic SNVs carried per person lead to single amino acid variants (SAVs) [9].

Therefore, studying the effects of protein sequence variants on molecular function is crucial, but experimental methods are costly, time-consuming and challenging, making it infeasible to analyze a large number of amino acid substitutions. Hence, computational tools that rely on conservation-related attributes reflecting structural and functional relevance, as well as on protein structure and stability-related properties following the relationship between structure and function, are used to estimate the phenotypic effect of these variants. Some examples of such reported methods are SIFT [10], CADD [11], PolyPhen-2 [12], PON-P2 [13] or PMut [14]. However, pathogenicity predictors do not accomplish the requirements of clinical applications for standalone tools since they show success rates of only around 80% on average, with prediction rates dramatically lower for specific diseases [15,16]. It is evident that current predictors are not able to capture all the possible effects of mutations at the molecular level. For this reason, a more detailed description of these variants, including information such as their potential involvement in protein-protein interactions (PPIs), would help to improve the prediction of their pathogenic character, providing a more accurate representation of the association between genetic variants and their phenotype by complementing general predictive methods.

Recent studies show that mutations in protein-protein interfaces are over-represented among disease-causing mutations [17–19]. While common variants from healthy individuals rarely affect interactions, almost two-thirds of disease-associated mutations perturb PPIs. Half of these pathogenic variants are ‘edgetic’ mutations, which impair only a subset of interactions while leaving most others unperturbed [20]. Consequently, within the context of PPI networks, knowledge about the molecular mechanisms by which genetic variants affect interaction networks can elucidate how mutations on the same gene might cause different phenotypes [6].

Regarding disease-causing mutations at PPI interfaces, they can induce geometrical and physicochemical changes at interaction sites that may affect interface stability, interface conformation dynamics through disruption or stabilization of specific conformational states, and the direct interactions between partner protomers [19]. Thus, the structural location of PPI interface mutations is important concerning pathogenicity. It has been demonstrated that disease-causing protein sequence variants are preferentially located in the solvent-inaccessible interface zones (‘core’), as opposed to the interface regions that remain partially solvent accessible (‘rim’) and are enriched in polymorphisms in the same way as the non-interacting surface. Moreover, energetic hot-spot residues, which play a crucial role in the free binding energy of the complex, tend to be enriched in disease-causing mutations regardless of the interface location [18]. All these findings highlight the importance of understanding the effects of protein sequence variants in protein structure to grasp the genotype to phenotype relationships.

In this study, we have characterized protein-protein interactions involving 58 proteins with pathogenic mutations related to diseases detected in newborn screening. Interpretation of mutational data in this set of proteins is of major clinical interest regarding the possibility of large-scale gene sequencing to detect disorders in newborn testing. We used the experimentally solved structures of the protein complexes when available, and when not, the protein-protein interface was predicted by an ab-initio docking approach. The distribution of disease-causing and neutral variants across the different interface regions, as well as the substitution susceptibility of distinct amino acids, was discussed.

2. Results

2.1. Structural Characterization of Proteins and Interactions in Diseases Detected in Newborn Screening

A total of 58 proteins with pathogenic mutations involved in diseases detected in newborn screening were analyzed (Table S1). As many as 56 of these proteins had structural information in Interactome3D, from which 42 had more than 80% structural coverage (Table S2). Only 16 of these proteins were monomers; 35 were homo-oligomers, and 5 hetero-oligomers. There were experimental structures for 62% of these 42 proteins, while for the remaining 38% of proteins, the Interactome3D database provided a homology-based model.

Among the 58 analyzed proteins, 50 of them had interaction data available in Interactome3D database (although three of them form only self-interactions). From the total of 389 PPIs found in such databases, only a small fraction (<12%) had available complex 3D structure. All these PPIs involved a total of 351 protein partners, the majority of which (75%) had available 3D structure. Among the partners with known 3D structure, 37% of them had good structural coverage (>80% of its sequence) in a single PDB file, while 40% of them have their structure split in separated PDB files.

Protein-protein interfaces were divided into core and rim residues (see Methods). For a given protein, residues can be defined as surface, interface core or interface rim depending on the considered partner. Figure 1 shows an example of a protein in which there is an available structure for the majority of its interactions, and the residues have been annotated according to such interaction data. Interface patches in one protein can be the same for some partners and different for others. This is important, since protein sequence variants in these regions could disrupt only a subset of interactions, possibly leading to 'edgetic' effects [20].

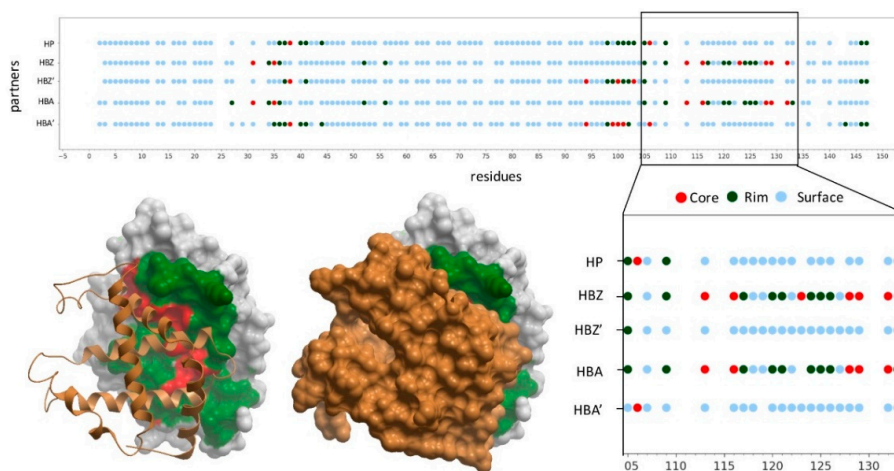


Figure 1. Structural characterization of hemoglobin subunit beta (HBB) interactions. The graphic represents the binding interface of HBB with different partners for which there is an available complex structure: haptoglobin (HP, hemoglobin subunit zeta (HBZ) and hemoglobin subunit alpha (HBA). As HBB interacts with both HBA and HBZ forming a heterotetramer, two different interfaces are formed with each of the HBB subunit (HBB-HBA and HBB-HBA', or HBB-HBZ and HBB-HBZ'). The graphic represents as dots the non-interacting surface residues (in blue), the interface rim residues (in dark-green) and the interface core ones (in red). The complex structure between HBB (white skin) and HBZ (gold ribbon or skin) is represented, with HBB interface rim residues (in green) and interface core ones (in red).

In all the annotated PPIs involving the 58 analyzed proteins, there were a total of 11,199 residues, of which 6019 were found to be buried in at least one structure (Table 1). Of the remaining non-buried (surface) residues, 2062 were located at the interface with at least one protein partner, and of these, 1146 residues were found at the interface core at least in one complex.

Table 1. Distribution of residues along the different protein regions and odds ratio for disease-causing and neutral variants.

<i>Disease-causing SAVs</i>									
Region	All Residues ¹	Observed ²	Expected ³	O/E ⁴	Regions	OR ⁵	95% C.I.	<i>p</i> -Value	Adjusted <i>p</i> -Value
Buried	6019	1842	1.548.96	1.19	Buried versus Surface	2.05	1.83–2.28	<0.00001	<0.00001
Surface	3118	552	802.40	0.69	Core versus Buried	0.94	0.82–1.09	0.441	1
Rim	916	151	235.73	0.64	Core versus Rim	2.11	1.69–2.64	<0.00001	<0.00001
Core	1146	337	294.92	1.14	Core versus Surface	1.94	1.65–2.27	<0.00001	<0.00001
Total	11199	2882			Rim versus Surface	0.92	0.75–1.12	0.428	1
					Rim versus Buried	0.45	0.37–0.54	<0.00001	<0.00001
					Interface versus Surface	1.44	1.25–1.54	<0.00001	<0.00001
<i>Neutral SAVs</i>									
Region	All Residues ¹	Observed ²	Expected ³	O/E ⁴	Regions	OR ⁵	95% C.I.	<i>p</i> -Value	Adjusted <i>p</i> -Value
Buried	6019	524	834.14	0.63	Buried versus Surface	0.29	0.25–0.33	<0.00001	<0.00001
Surface	3118	767	432.10	1.78	Core versus Buried	0.82	0.63–1.04	0.105	0.738
Rim	916	178	126.94	1.40	Core versus Rim	0.32	0.24–0.43	<0.00001	<0.00001
Core	1146	83	158.82	0.52	Core versus Surface	0.24	0.19–0.30	<0.00001	<0.00001
Total	11199	1552			Rim versus Surface	0.74	0.61–0.89	0.001187	0.008209
					Rim versus Buried	2.53	2.08–3.06	<0.00001	<0.00001
					Interface versus Surface	0.44	0.38–0.52	<0.00001	<0.00001

¹ Total number of residues in each protein region. ² Observed number of residues involving pathogenic (or neutral) variants in each protein region. ³ Expected number of residues involving pathogenic (or neutral) variants in each protein region, according to a random distribution based on the total number of residues. ⁴ Ratio of observed to expected residues involving pathogenic (or neutral) variants in each protein region. ⁵ Odds ratio for different possibilities is calculated with a 95% confidence interval and a *p*-value for a two-tailed Fisher's exact test. This *p*-value is adjusted using Bonferroni correction. A *p*-value < 0.05 is considered indicative of statistical significance.

2.2. Residues Energetically Relevant for the Interaction Are More Likely to Be at the Interface Core

The energetic contribution to protein complex stability was not uniform across the interface, and for instance, interface residues that were estimated to be energetically relevant for the interaction (i.e., those with residue binding energy < -2 a.u., as calculated by pyDock) tended to be located at the interface core region more often than expected by random (Table 2). This was consistent with previous studies reporting that core amino acids contribute significantly more than rim amino acids to the binding free energy of the complex [18,21].

Table 2. Distribution of all interface residues and those energetically relevant for the interaction.

Interface Region	All Residues ¹	Observed Low-Energy Residues ²	Expected Low-Energy Residues ³	O/E ⁴
Rim	916	201	298.08	0.67
Core	1146	470	372.92	1.26
Total	2062	671		

¹ Total number of residues in the set of PPIs analyzed here in each interface region (core and rim). ² Residues with binding energy < -2 a.u., as calculated by pyDock, in each interface region. ³ Expected number of low binding energy residues in each interface region according to a random distribution based on the total number of residues.

⁴ Ratio of observed to expected residues.

2.3. Pathogenic and Neutral Variants Are Differentially Distributed in Protein-Protein Interfaces

A total of 2882 disease-causing mutations and 1552 neutral variants were mapped onto the 3D structures of the protein-protein interactions involving the 58 genes analyzed here (Table 1). Around 47% of all variants occurred in solvent-accessible residues, which included non-interacting regions (surface) and interacting ones (interface).

Regarding the disease-causing variants, 36% of them (1040) occurred in solvent-accessible residues, of which 488 were found at the interface with at least one partner (with 337 of them at the core region in at least one complex). The odds of being located at the interface rather than at a non-interacting surface was 1.44 higher for pathogenic variants compared to the rest of residues (OR 1.44, 95% CI 1.25–1.54, $p < 0.0001$), which is consistent with previous studies [4,17,18,21]. More specifically, the odds of being located at the interface core region rather than rim was 2.11 higher for disease-causing variants compared to the rest of the residues, similar to the odds of being located at the interface core rather than at non-interacting protein surface (1.94). On the other side, there was no significant difference between the location of disease-causing variants at the interface rim region and the non-interacting protein surface (Table 1). These results show clearly the different impact of interface core and rim mutations in human disease.

Regarding the neutral variants, 66% of them occurred in solvent-accessible residues, of which 261 were found at the interface region (being 83 of them at the core, and 178 at the rim), and 767 at the non-interacting protein surface. Contrarily to disease-causing variants, the odds of being located at the interface rather than at a non-interacting protein surface was smaller for neutral variants compared to the rest of residues (OR 0.44, 95% CI 0.38–0.52, $p < 0.00001$). Moreover, for these neutral variants, the odds of being located at the interface core rather than the rim or the non-interacting surface was 0.32 and 0.24, respectively. As in the case of disease-causing variants, there was no significant difference between the location of neutral variants at the interface rim region and the non-interface protein surface.

The division of the interface into core and rim regions showed that the core was enriched in disease-causing variants, while the rim was enriched in neutral variants. As in the case of the non-interacting protein surface, amino acid changes in the rim region can be easily accommodated without significant distortions in the overall fold of the protein and without affecting the PPIs. This was consistent with their lower evolutionary conservation and higher side-chain flexibility [22]. Figure 2 shows the distribution of neutral and pathogenic SAVs in a case example (HBB, in which there is available structure for the majority of its interactions). There were five PPIs annotated in Interactome3D for HBB. Four of these PPIs had an available structure (or reliable model): HBB-HP, HBB-HBA1, HBB-HBZ and

HBB-HBB. As can be seen, the proportion of pathogenic variants located in the interface rim and the non-interacting surface was much lower than the proportion of neutral variants in the same regions.

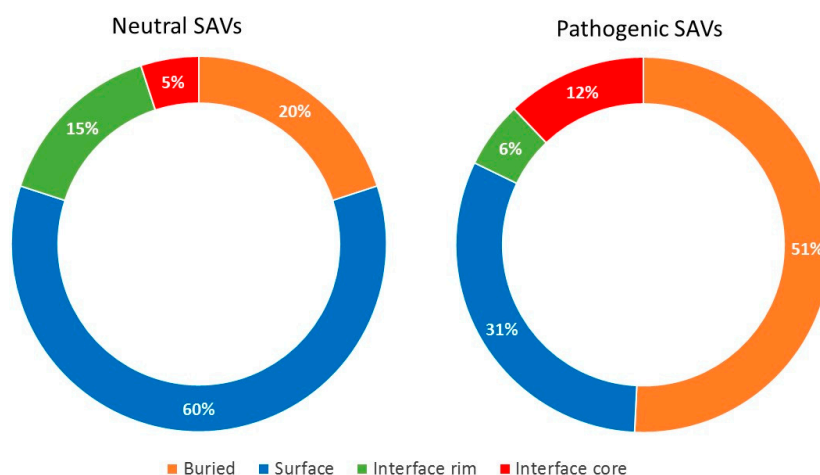


Figure 2. Structural characterization of residues in HBB affected by neutral or pathogenic variants. The graphics show the percentage of HBB residues affected by either neutral or pathogenic variants, as a function of their location in the available protein-protein complex structures.

2.4. Amino Acid Substitution Susceptibility in the Interface Is Larger in Pathogenic Variants

The amino acids mutability susceptibility was analyzed to determine whether it could be relevant for the molecular characterization of disease-causing variants. Arginine (R) was the most mutated residue in both neutral and pathogenic variants in protein-protein interfaces. This high mutability can be explained by the fact that four out of the six codons for R include CpG dinucleotides, which tend to mutate at rates 10–15 times higher than other dinucleotides in the DNA [23]. Arginine (R), tryptophan (W) and tyrosine (Y) were significantly over-represented among mutated residues leading to disease (Figure 3), which is coherent with previous findings [24].

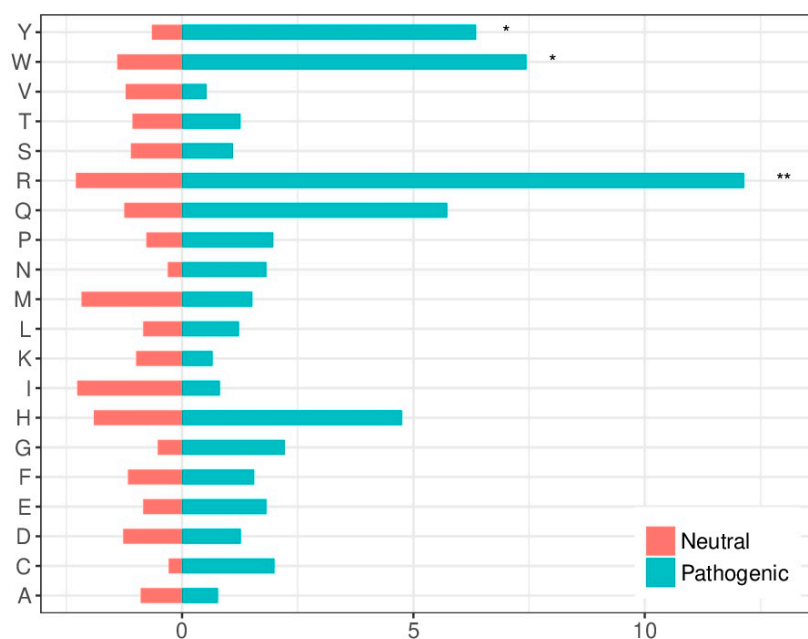


Figure 3. Amino acid substitution susceptibility to disease-causing or neutral variants within protein interfaces. The bars show the percentage of residues of a given type that are found mutated in disease-causing variants (in blue) or neutral variants (in red). Asterisks show statistical significance according to a “N-1” χ^2 test (* $0.01 < p < 0.05$; ** $0.001 < p < 0.01$).

2.5. Docking-Based Interface Prediction for Further Characterization of Protein Sequence Variants: A Case Study

Many of the proteins analyzed here are involved in protein-protein interactions for which there is no available complex structure. In these cases, we could apply docking simulations to identify potential interface residues, using a pyDock approach and NIP interface prediction module. First of all, in order to apply docking simulations, we needed to check whether we had a complete structure or a reasonable model for the interacting proteins. Available databases such as Interactome3D provide this information. However, there are several issues to consider here. For example, in many cases either a model or an experimental version of the overall structure of the target protein is available, but with incomplete structural coverage. Or, it may also happen that the overall structure is split between different PDB files, and then we would need to infer the global structure from these different parts, a non-trivial problem. In Table S2 we identified those proteins that have >80% structural coverage in a single PDB file. This global structure would be suitable for docking. If this global structure is not available, because of incomplete coverage or because it is split between different PDB files, we could still use for docking the individual domains that have >80% structural coverage, and then we would try to rebuild the whole protein in the context of the docking models. Since this is out of the scope of automatic docking, we have not used these cases here for docking. In addition to the previous issues, it is important to identify the oligomeric state of the interacting proteins, so that we use for docking the correct assembly form. Table S2 provides such information.

Given all the above considerations, we have selected one example case, HADHA, involved in 13 different protein-protein interactions for which there is no complex structure available. There are 116 neutral and 31 pathogenic mutations described in HADHA. Figure 4a shows the location of these mutations in HADHA structure. However, they cannot be located at any protein-protein interface due to the lack of this structural information. We explored whether docking-based estimation of interface residues could help to further characterize such mutations. In six of these interactions, interacting partners have sufficient structural coverage (i.e., >80%) for docking. We used pyDock to run docking in these cases, and based on that, we estimated the interface rim and core residues from the NIP calculations. Figure 4b–g shows the predicted interface core and rim residues for each of these interactions, which can be used to visually check whether any of the known variants in HADHA are located at the predicted interfaces. Table 3 shows in detail the neutral and pathogenic SAVs in HADHA that are located at the different predicted interfaces. Disease-related mutations R399*, Y740* and V412L, involved in mitochondrial trifunctional protein deficiency, are found in all predicted interfaces. Pathological mutations with more specific effects are Q358K, involved in haemolysis, elevated liver enzymes, and low platelets, and found at the predicted interface with Q14134 (Figure 4d), or R610K, involved in long-chain 3-hydroxyacyl-CoA dehydrogenase deficiency, and found at the predicted interface with Q14134 and Q99714 (Figure 4d,e).

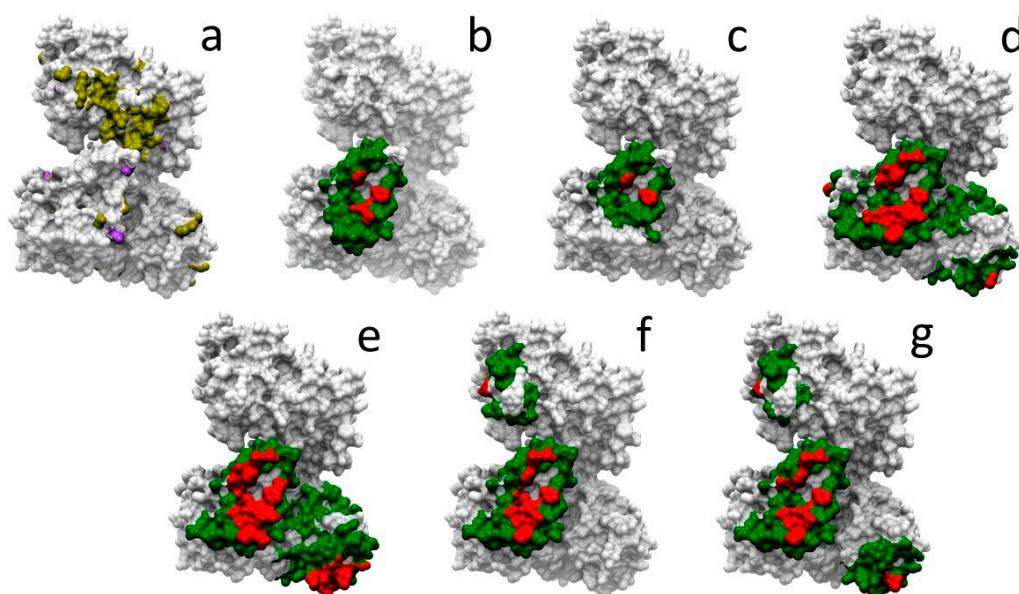


Figure 4. Docking-based characterization of HADHA mutations related to protein interactions. (a) Neutral (yellow) and disease-related (purple) mutations mapped on HADHA structure. In panels (b–g), docking-based predicted interface core (red) and rim (green) residues in HADHA for the interaction with the following partners: (b) O95166, (c) P60520, (d) Q14164, (e) Q99714, (f) Q9GZQ8, and (g) Q9H0R8.

Table 3. Docking-based characterization of HADHA mutations.

UniProt ¹ (Partner)	Neutral Mutations		Pathogenic Mutations ²	
	Core ³	Rim ³	Core ³	Rim ³
O95166	D398G	A396G, K406R	-	R399*, V412L
P60520	-	D398G, K406R	-	R399*, V412L
Q14164	D398G	A396G, K406R, K519R, A596V, S654N, K734Q	-	<i>Q358K</i> , R399*, V412L, R610G, Y740*
Q99714	D398G, S654N	A396G, K406R, A596V, R645S, R645N, L661I, K734Q	-	R399*, V412L, R610G, Y740*
Q9GZQ8	D398G	V52I, V526I, N142S, L221I, E223T, I237M, A396G, K406R	-	R399*, V412L
Q9H0R8	D398G	N142S, L221I, E223T, I237M, A396G, K406R, S654N, L661I	-	R399*, V412L

¹ UniProt code of the corresponding interacting partner. ² The symbol "*" indicates stop codon. Mutations R399*, Y740* and V412L (in bold) are associated with mitochondrial trifunctional protein deficiency. Mutation *Q358K* (in italics) is associated to haemolysis, elevated liver enzymes, and low platelets. Mutation R610G is associated to long-chain 3-hydroxyacyl-CoA dehydrogenase deficiency. ³ Interface core and rim estimated from the docking calculations.

3. Discussion

To better understand the functional influence of genetic variants at the protein level, structural characterization of single amino acid variants and their interactions is one of the basic steps. In this regard, several structural databases of protein interaction data can be used, such as Interactome3D [25]. However, a major limitation is the low availability of 3D structures for protein-protein complexes. Consequently, a large fraction of protein sequence variants cannot be precisely located at protein interfaces. For this reason, using docking models to estimate whether variants can be involved in PPIs may be auspicious. In this regard, a potential problem for the application of docking at a large-scale is that most of the available interaction databases essentially provide sets of binary interactions (i.e., protein-protein interacting pairs), while for this type of experiment we would need more detailed data, such as the identification of the contacting domains, or the oligomeric state of the interacting proteins. We have collected all this information for the proteins and interactions analyzed here (Table S2). This can be valuable information in order to run docking simulations in the most

realistic conditions. To test this in a real example, we chose HADHA interactions, in which interacting partners had available 3D structure with >80% structural coverage. We applied protein-protein docking using the available structures of the interacting partners and their biological units in order to predict the binding residues for these interactions.

Mutations in the same gene can affect different phenotypic traits (pleiotropy). In this context, the number of interactions and interfaces in a protein is key to understand pleiotropic effects in disease genes. Recent studies show that SAVs located at distinct protein-protein interfaces of the same protein are prone to produce different disease phenotypes [20,26,27]. Moreover, it has been demonstrated that one-third of the SAVs produce an ‘edgetic’ effect, by impairing only a subset of the interactions [20].

In this line, structural analysis of the case example hemoglobin subunit beta (HBB) showed that the same variant could affect the interaction with different partner proteins if their interface patches are the same, and different variants could perturb different partner proteins if these have distinct interface patches. Figure 5 shows some of the pathogenic mutations found in HBB as well as the interaction they impair. For instance, F123S only affects the interaction between HBB and hemoglobin subunit zeta (HBZ); E27A perturbs the interaction between HBB and hemoglobin subunit alpha (HBA); E44Q hampers the interactions between HBB and both HBA and haptoglobin (HP); C113R affects the interactions of HBB with HBA and HBZ, same as R105W, which also hinders the interaction between HBB and HP.

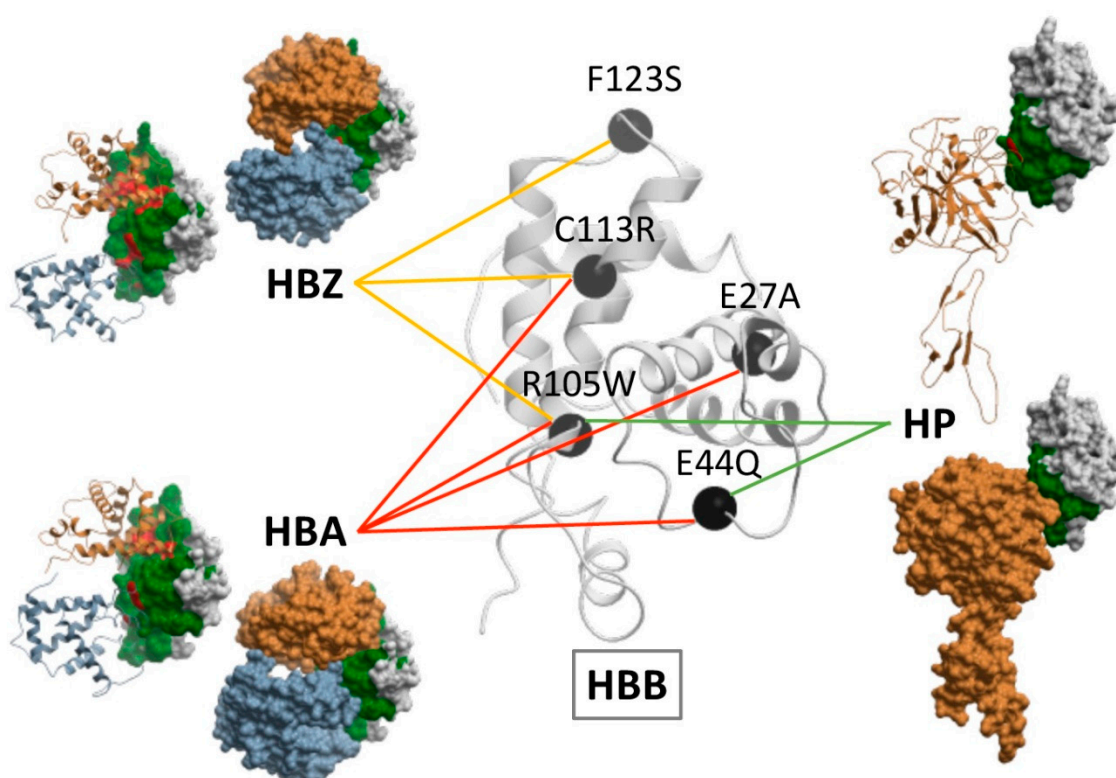


Figure 5. Estimation of the effect of disease-causing variants in the HBB interaction network based on experimentally solved complex structures. HBB is represented in white ribbon, with selected pathogenic variants in black, linked to the protein-protein interactions in whose interface they are located. The structures of such interactions are represented, showing HBB in white skin with interface rim residues (in green) and interface core residues (in red), and HBB partners (HBZ, HBA, HP) either as ribbon or skin in two different views.

The extensive analysis of protein interface residues shown here, combining complex structures and docking predictions, demonstrate that pathogenic variants are more likely to be located at the interface rather than at the non-interacting surface. More precisely, we found that they are more

probable to occur at the interface core region rather than at the rim, in agreement with previous studies [4,17,18,28]. On the contrary, neutral variants occur significantly more often in the interface rim as well as in the non-interacting surface, as compared with the interface core region. Furthermore, the residues that contribute the most to the binding free energy of the protein-protein complex (hot-spots) are more likely to be located at the interface core. This is in line with previous studies [18,21,29], which revealed that hot-spot residues are not equally distributed among interface regions, but they tend to be clustered within the interface core. Thus, this core region is critical for the stabilization of PPIs; this is reflected in the fact that core residues show a higher level of conservation and coevolution among homologous proteins as compared to those in the rim [23,30]. This energetical relevance of the core region also explains why protein sequence variants are not as likely to be tolerated there as in the interface rim or the non-interacting surface [18]. We found that arginine, tryptophan and tyrosine are over-represented among disease-causing mutated residues. This was consistent with previous studies reporting that the most frequent hot-spot residues are tryptophan (21%), arginine (13.3%) and tyrosine (12.3%) [31–33]. Indeed, arginine mutations in interface core residues are not likely to be tolerated and tend to have a profound effect in phenotype [31,32].

The present study has some limitations, such as the low availability of 3D protein structures or the global consideration in the analysis of both transient and permanent PPIs, which are known to show very different mechanistic, structural and energetic properties. Despite these limitations, this study shows that the structural characterization of PPIs and the analysis of the location of pathogenic and neutral variants, together with the identification of the interface residues that are more prone to be mutated and lead to disease, can provide novel information on disease-causing variants. This can be useful in order to characterize protein sequence variants in future studies, interpret them at the molecular level, improve the accuracy of pathogenicity predictors on new mutations, and help to advance toward precision medicine by helping clinicians to provide adequate diagnosis and treatments.

Further studies with more docking simulations will need to be undertaken. For partners with 3D structures split in different PDBs, a template could be used to model the missing amino acids and join the distinct protein fragments in a correct global 3D structure. Moreover, if these PDBs contain at least one complete domain, docking simulations could be done at the domain level. Homology models could be generated for those proteins without available 3D structure, so that docking can be run afterwards in order to find a possible protein-protein interface. This would help to achieve a better understanding of disease at molecular level since more PPIs could be characterized and more disease-causing and neutral variants could be mapped on the structural models.

4. Materials and Methods

4.1. Protein Interaction and Mutational Data

Human PPI data and structural information for both protein complexes and their individual components were retrieved from Interactome3D [25]. In this database, human protein-protein interaction data is generated by integrating information available from nine major public PPI databases: Intact [34], MINT [35], DIP [36], MPIDB [37], MatrixDb [38], InnateDb [39], BioGRID [40], BIND [41] and HPRD [42]. The Interactome3D database also provided the experimentally solved structures of protein-protein complexes, when they were available in the Protein Data Bank (PDB) [43].

Human pathogenic mutations were compiled by pooling variants obtained from UniProt [44], selecting SAVs labelled as “Disease” in the downloadable file humsavar.txt (therefore not including “Polymorphism” or “Unclassified” variants), as well as from the Human Gene Mutation Database (HGMD) [45], containing both missense and nonsense variants (Table S1). For neutral variants (Table S1), we used the homology-based model described in Riera et al. [15,46,47], where variants were obtained from a multiple sequence alignment (MSA) for each protein family and corresponded to mismatches between the human protein and its close homologs (more than 95% sequence identity with respect to the human protein sequence).

4.2. Interacting Proteins Analysis

As a further analysis, protein structures involved in each interaction were characterized in more detail regarding sequence identity, structural coverage, domains, and biological assembly (Table S2). Sequence identity and structural coverage were calculated using the UniProt canonical sequence as a reference. Missing loops were not considered in the structural coverage calculations (as opposed to the structural coverage value given by Interactome3D, which includes the missing loops). To identify the protein domains, HMMER3 [48] was used to search against Pfam database [49], based on the canonical sequence. Based on the structural coverage, PPIs could be defined depending on whether the interacting proteins had: (i) global structural coverage greater than 80% in a single PDB file, (ii) global structural coverage < 80% and at least one domain with more than 80% structural coverage, and (iii) global or domain structural coverage < 80% (Table S2).

4.3. Experimental Protein-Protein Interfaces

Protein-protein complex structures, when available, were retrieved from PDB based on Interactome3D information. Protein-protein interfaces were defined in a similar way as previously described [50]. Prior to the interface calculation, the sequence and numbering of the PDB structures were extracted and aligned with the corresponding canonical sequence fetched from UniProt database to ensure a correct residue numbering.

Residues were defined as buried if they had relative Accessible Surface Area in the uncomplexed structure ($rASA_u$) < 0.1, or surface if they had $rASA_u \geq 0.1$. Surface residues were classified as interface residues when the difference in rASA between the uncomplexed and complexed form ($rASA_u - rASA_c$) was > 0, or non-interface surface otherwise. Interface residues were further divided into core and rim. Core was formed by interface residues that were buried in the complex ($rASA_c < 0.1$), and rim was formed by interface residues that remained exposed in the complex ($rASA_c > 0.1$). The value rASA was computed as the ratio between the Accessible Surface Area (ASA) of a given residue, and the ASA of the corresponding residue type in the extended conformation of the Gly-X-Gly peptide. All (ASA) calculations were done with ICM-Browser (<http://www.molsoft.com>).

4.4. Predicted Protein-Protein Interfaces

For selected protein-protein interactions without available protein complex structure, we applied a computational procedure to estimate the interface residues. For this, the uncomplexed structures were retrieved from PDB, considering the oligomeric state as defined in the biological unit in the PDB. In this work, ab initio protein-protein docking was used to model the PPI when both proteins forming the complex had more than 80% structural coverage.

First, the sequence and numbering of the PDB structures were extracted and aligned with the corresponding canonical sequence fetched from UniProt database, to ensure a correct residue numbering. Then, docking simulations were run with the Fast Fourier Transform (FFT)-based program FTDock 2.0 [51], and the resulting 10,000 rigid-body orientations were rescored by pyDock scoring function, which includes electrostatics, desolvation energy, and a limited van der Waals contribution [52].

From the resulting docking poses, a normalized interface propensity (NIP) was obtained per residue with the built-in *patch* module in pyDock, implementing the pyDockNIP algorithm [53]. A normalized interface propensity (NIP) value of one indicates that the corresponding residue is involved in all predicted interfaces of the 100 lowest energy docking solutions, while a value of zero means that it appears as expected by random. On the other hand, a negative NIP value implies that the residue appears at the low-energy docking interfaces less often than expected by random [53]. Usually, residues with $NIP \geq 0.2$ are considered as hot-spot residues when using FTDock but given the large size of the proteins analyzed here, we used a cutoff of $NIP \geq 0.1$ to define the predicted hot-spot residues. These constituted the predicted interface core residues. Then, predicted interface

rim residues were built by surface residues located within 10 Å distance from the predicted hot-spot (core) residues [28].

4.5. Energetic Characterization of Protein-Protein Interfaces

The energetic characterization of protein-protein interfaces was performed with the pyDock *bindEy* and *resEnergy* modules. The *bindEy* module computes the total binding energy for a given protein-protein interaction, based on the complex structure or a model. The *resEnergy* module calculates the contribution of each individual protein residue to the binding energy for a given protein-protein complex structure.

4.6. Statistical Analysis

The statistical analyses were performed using version 3.4.4 of the R statistical package [54]. The probability of observing a protein sequence variant in the protein region i is calculated as shown in Equation (1), where n_i is the number of variants observed in the protein region i , and N_i is the total number of residues in that region. The likelihood of a variant to be in region i rather than in region j in the protein was expressed then in terms of odds ratio (OR_{ij}) (Equation (2)). The χ^2 test was used to compare the observed number of variants in each region with the expected one if variants were distributed according to the number of residues in the different regions. A two-tailed p -value < 0.05 indicated statistical significance of the preference for variants to be in one region over another. Bonferroni correction was used to adjust p -value for multiple comparisons.

$$x_i = \frac{n_i}{N_i} \quad (1)$$

$$OR_{ij} = \frac{x_i/(1-x_i)}{x_j/(1-x_j)} \quad (2)$$

The amino acid substitution susceptibility to disease-causing variants or neutral ones at protein interfaces was calculated, and a two-tailed p -value < 0.05 implied statistical significance according to an “ $N-1$ ” χ^2 test.

Supplementary Materials: Supplementary Materials can be found at <http://www.mdpi.com/1422-0067/20/7/1583/s1>.

Author Contributions: Methodology, D.N., M.R. and J.A.; software, D.N. and M.R.; formal analysis, D.N. and M.R.; data curation, J.A.; writing—original draft preparation, D.N.; writing—review and editing, M.R., J.A., X.C. and J.F.-R.; supervision, X.C. and J.F.-R.; funding acquisition, X.C. and J.F.-R.

Funding: This research was funded by the EU European Regional Development Fund (ERDF) through the Program Interreg V-A Spain-France-Andorra (POCTEFA), by the CSIC (intramural grant number 201720I031), and by the Spanish Ministry of Economy and Competitiveness (grants BIO2016-79930-R and SAF2016-80255-R). M.R. is recipient of an FPI fellowship from the Severo Ochoa program.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

NGS	Next-generation sequencing
SNV	Single nucleotide variant
SAV	Single amino acid variant
PPI	Protein-protein interaction
HBB	Hemoglobin subunit beta
HBZ	Hemoglobin subunit zeta
HBA	Hemoglobin subunit alpha
HP	Haptoglobin
PDB	Protein Data Bank
MSA	Multiple Sequence Alignment
ASA	Accessible Surface Area
NIP	Normalized Interface Propensity
HADHA	Hydroxyacyl-CoA dehydrogenase trifunctional multienzyme complex subunit α

References

1. The 1000 Genomes Project Consortium; Boerwinkle, E.; Doddapaneni, H.; Han, Y.; Korchina, V.; Lee, S.; Zhu, Y.; Chang, Y.; Feng, Q.; Fang, X.; et al. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74.
2. Muir, P.; Li, S.; Lou, S.; Wang, D.; Spakowicz, D.J.; Salichos, L.; Zhang, J.; Weinstock, G.M.; Isaacs, F.; Rozowsky, J.; et al. The real cost of sequencing: Scaling computation to keep pace with data generation. *Genome Biol.* **2016**, *17*, 53. [[CrossRef](#)]
3. Xue, Y.; Ankala, A.; Wilcox, W.R.; Hegde, M.R. Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: Single-gene, gene panel, or exome/genome sequencing. *Genet. Med.* **2014**, *17*, 444–451. [[CrossRef](#)] [[PubMed](#)]
4. David, A.; Razali, R.; Wass, M.N.; Sternberg, M.J. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Hum. Mutat.* **2011**, *33*, 359–363. [[CrossRef](#)]
5. Gao, M.; Zhou, H.; Skolnick, J. Insights into disease-associated mutations in the human proteome through protein structural analysis. *Structure* **2015**, *23*, 1362–1369. [[CrossRef](#)] [[PubMed](#)]
6. Jubb, H.C.; Pandurangan, A.P.; Turner, M.A.; Ochoa-Montano, B.; Blundell, T.L.; Ascher, D.B. Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. *Prog. Biophys. Mol. Boil.* **2017**, *128*, 3–13. [[CrossRef](#)]
7. Gregersen, N.; Bross, P.; Vang, S.; Christensen, J.H. Protein Misfolding and Human Disease. *Annu. Rev. Genom. Hum. Genet.* **2006**, *7*, 103–124. [[CrossRef](#)]
8. Aguzzi, A.; O'Connor, T. Protein aggregation diseases: Pathogenicity and therapeutic perspectives. *Nat. Rev. Drug Discov.* **2010**, *9*, 237–248. [[CrossRef](#)]
9. Tennessen, J.A.; Bigham, A.W.; O'Connor, T.D.; Fu, W.; Kenny, E.E.; Gravel, S.; McGee, S.; Do, R.; Liu, X.; Jun, G.; et al. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science* **2012**, *337*, 64–69. [[CrossRef](#)]
10. Vaser, R.; Adusumalli, S.; Leng, S.N.; Sikic, M.; Ng, P.C. SIFT missense predictions for genomes. *Nat. Protoc.* **2015**, *11*, 1–9. [[CrossRef](#)] [[PubMed](#)]
11. Kircher, M.; Witten, D.M.; Jain, P.; O'Roak, B.J.; Cooper, G.M.; Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **2014**, *46*, 310–315. [[CrossRef](#)]
12. Adzhubei, I.A.; Schmidt, S.; Peshkin, L.; Ramensky, V.E.; Gerasimova, A.; Bork, P.; Kondrashov, A.S.; Sunyaev, S.R. A method and server for predicting damaging missense mutations. *Nat. Chem. Boil.* **2010**, *7*, 248–249. [[CrossRef](#)]
13. Niroula, A.; Urolagin, S.; Vihinen, M. PON-P2: Prediction method for fast and reliable identification of harmful variants. *PLoS ONE* **2015**, *10*, e0117380. [[CrossRef](#)]
14. López-Ferrando, V.; Gazzo, A.; De La Cruz, X.; Orozco, M.; Gelpí, J.L. PMut: A web-based tool for the annotation of pathological variants on proteins, 2017 update. *Acids Res.* **2017**, *45*, W222–W228. [[CrossRef](#)] [[PubMed](#)]

15. Riera, C.; Lois, S.; De La Cruz, X. Prediction of pathological mutations in proteins: The challenge of integrating sequence conservation and structure stability principles. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *4*, 249–268. [[CrossRef](#)]
16. Sunyaev, S.R. Inferring causality and functional significance of human coding DNA variants. *Hum. Mol. Genet.* **2012**, *21*, R10–R17. [[CrossRef](#)] [[PubMed](#)]
17. Yates, C.M.; Sternberg, M.J. The Effects of Non-Synonymous Single Nucleotide Polymorphisms (nsSNPs) on Protein–Protein Interactions. *J. Mol. Boil.* **2013**, *425*, 3949–3963. [[CrossRef](#)] [[PubMed](#)]
18. David, A.; Sternberg, M.J. The Contribution of Missense Mutations in Core and Rim Residues of Protein–Protein Interfaces to Human Disease. *J. Mol. Boil.* **2015**, *427*, 2886–2898. [[CrossRef](#)]
19. Kucukkal, T.G.; Petukh, M.; Li, L.; Alexov, E. Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. *Curr. Opin. Struct. Boil.* **2015**, *32*, 18–24. [[CrossRef](#)]
20. Sahni, N.; Yi, S.; Taipale, M.; Bass, J.I.F.; Coulombe-Huntington, J.; Yang, F.; Peng, J.; Weile, J.; Karras, G.I.; Wang, Y.; et al. Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders. *Cell* **2015**, *161*, 647–660. [[CrossRef](#)]
21. Agius, R.; Torchala, M.; Moal, I.H.; Fernandez-Recio, J.; Bates, P.A. Characterizing Changes in the Rate of Protein–Protein Dissociation upon Interface Mutation Using Hotspot Energy and Organization. *PLoS Comput. Boil.* **2013**, *9*, e1003216. [[CrossRef](#)] [[PubMed](#)]
22. Guharoy, M.; Chakrabarti, P. Conservation and relative importance of residues across protein–protein interfaces. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15447–15452. [[CrossRef](#)] [[PubMed](#)]
23. Sipos, B.; Goldman, N.; Laskowski, R.A.; Parks, S.L.; De Beer, T.A.P.; Thornton, J.M. Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. *PLoS Comput. Boil.* **2013**, *9*, e1003382.
24. Thusberg, J.; Olatubosun, A.; Vihinen, M. Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.* **2011**, *32*, 358–368. [[CrossRef](#)] [[PubMed](#)]
25. Mosca, R.; Ceol, A.; Aloy, P. Interactome3D: Adding structural details to protein networks. *Nat. Chem. Boil.* **2012**, *10*, 47–53. [[CrossRef](#)]
26. Wang, X.; Wei, X.; Thijssen, B.; Das, J.; Lipkin, S.M.; Yu, H. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* **2012**, *30*, 159–164. [[CrossRef](#)]
27. Kammenga, J.E. The background puzzle: How identical mutations in the same gene lead to different disease symptoms. *FEBS J.* **2017**, *284*, 3362–3373. [[CrossRef](#)]
28. Barradas-Bautista, D.; Fernández-Recio, J. Docking-based modeling of protein–protein interfaces for extensive structural and functional characterization of missense mutations. *PLoS ONE* **2017**, *12*, e0183643. [[CrossRef](#)]
29. Keskin, O.; Ma, B.; Nussinov, R. Hot Regions in Protein–Protein Interactions: The Organization and Contribution of Structurally Conserved Hot Spot Residues. *J. Mol. Boil.* **2005**, *345*, 1281–1294. [[CrossRef](#)]
30. Teppa, E.; Zea, D.J.; Marino-Buslje, C.; Marino-Buslje, C.; Marino-Buslje, C. Protein–protein interactions leave evolutionary footprints: High molecular coevolution at the core of interfaces. *Protein Sci.* **2017**, *26*, 2438–2444. [[CrossRef](#)]
31. A Bogan, A.; Thorn, K.S. Anatomy of hot spots in protein interfaces. *J. Mol. Boil.* **1998**, *280*, 1–9. [[CrossRef](#)] [[PubMed](#)]
32. Moreira, I.S.; Fernandes, P.A.; Ramos, M.J. Hot spots—A review of the protein–protein interface determinant amino-acid residues. *Proteins Struct. Funct. Bioinform.* **2007**, *68*, 803–812. [[CrossRef](#)]
33. Morrow, J.K.; Zhang, S. Computational Prediction of Protein Hot Spot Residues. *Curr. Drug Metab.* **2012**, *18*, 1255–1265. [[CrossRef](#)]
34. Orchard, S.; Ammari, M.; Aranda, B.; Breuza, L.; Briganti, L.; Broackes-Carter, F.; Campbell, N.H.; Chavali, G.; Chen, C.; Del-Toro, N.; et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Acids Res.* **2013**, *42*, D358–D363. [[CrossRef](#)]
35. Licata, L.; Briganti, L.; Peluso, D.; Perfetto, L.; Iannuccelli, M.; Galeota, E.; Sacco, F.; Palma, A.; Santonico, E.; Castagnoli, L.; et al. MINT, the molecular interaction database: 2012 update. *Acids Res.* **2011**, *40*, D857–D861. [[CrossRef](#)]
36. Salwinski, L.; Miller, C.S.; Smith, A.J.; Pettit, F.K.; Bowie, J.U.; Eisenberg, D. The Database of Interacting Proteins: 2004 update. *Acids Res.* **2004**, *32*, D449–D451. [[CrossRef](#)] [[PubMed](#)]
37. Goll, J.; Wu, H.; Uetz, P.; Rajagopala, S.V.; Shiau, S.C.; Lamb, B.T. MPIDB: The microbial protein interaction database. *Bioinformatics* **2008**, *24*, 1743–1744. [[CrossRef](#)] [[PubMed](#)]

38. Launay, G.; Salza, R.; Multedo, D.; Thierry-Mieg, N.; Ricard-Blum, S. MatrixDB, the extracellular matrix interaction database: Updated content, a new navigator and expanded functionalities. *Acids Res.* **2014**, *43*, D321–D327. [[CrossRef](#)]
39. Breuer, K.; Chen, C.; Sribnaia, A.; Lo, R.; Foroushani, A.K.; Laird, M.R.; Winsor, G.L.; Hancock, R.E.W.; Brinkman, F.S.L.; Lynn, D.J. InnateDB: Systems biology of innate immunity and beyond—recent updates and continuing curation. *Acids Res.* **2012**, *41*, D1228–D1233. [[CrossRef](#)] [[PubMed](#)]
40. Chatr-Aryamontri, A.; Breitkreutz, B.-J.; Oughtred, R.; Boucher, L.; Heinicke, S.; Chen, D.; Stark, C.; Breitkreutz, A.; Kolas, N.; O'Donnell, L.; et al. The BioGRID interaction database: 2015 update. *Acids Res.* **2014**, *43*, D470–D478. [[CrossRef](#)]
41. Isserlin, R.; A El-Badrawi, R.; Bader, G.D. The Biomolecular Interaction Network Database in PSI-MI 2.5. *Database* **2011**, *2011*, baq037. [[CrossRef](#)]
42. Prasad, T.S.K.; Goel, R.; Kandasamy, K.; Keerthikumar, S.; Kumar, S.; Mathivanan, S.; Telikicherla, D.; Raju, R.; Shafreen, B.; Venugopal, A.; et al. Human Protein Reference Database—2009 update. *Acids Res.* **2008**, *37*, D767–D772. [[CrossRef](#)]
43. Berman, H.M.; Kleywegt, G.J.; Nakamura, H.; Markley, J.L. The Protein Data Bank archive as an open data resource. *J. Comput. Mol. Des.* **2014**, *28*, 1009–1014. [[CrossRef](#)]
44. Bateman, A. The UniProt Consortium UniProt: The universal protein knowledgebase. *Acids Res.* **2016**, *45*, D158–D169.
45. Stenson, P.D.; Ball, E.V.; Mort, M.; Phillips, A.D.; Shaw, K.; Cooper, D.N. The Human Gene Mutation Database (HGMD) and Its Exploitation in the Fields of Personalized Genomics and Molecular Evolution. *Curr. Protoc. Bioinform.* **2012**, *39*, 1–13.
46. Riera, C.; Lois, S.; Dominguez, C.; Fernandez-Cadenas, I.; Montaner, J.; Rodríguez-Sureda, V.; De La Cruz, X.; Fernández-Cadenas, I.; Rodríguez-Sureda, V. Molecular damage in Fabry disease: Characterization and prediction of alpha-galactosidase A pathological mutations. *Proteins: Struct. Funct. Bioinform.* **2014**, *83*, 91–104. [[CrossRef](#)] [[PubMed](#)]
47. Riera, C.; Padilla, N.; De La Cruz, X.; La Cruz, X. The Complementarity Between Protein-Specific and General Pathogenicity Predictors for Amino Acid Substitutions. *Hum. Mutat.* **2016**, *37*, 1013–1024. [[CrossRef](#)]
48. Mistry, J.; Finn, R.D.; Eddy, S.R.; Bateman, A.; Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Acids Res.* **2013**, *41*, e121. [[CrossRef](#)] [[PubMed](#)]
49. Finn, R.D.; Bateman, A.; Clements, J.; Coghill, P.; Eberhardt, R.Y.; Eddy, S.R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J.; et al. Pfam: The protein families database. *Acids Res.* **2013**, *42*, D222–D230. [[CrossRef](#)] [[PubMed](#)]
50. Levy, E.D. A Simple Definition of Structural Regions in Proteins and Its Use in Analyzing Interface Evolution. *J. Mol. Boil.* **2010**, *403*, 660–670. [[CrossRef](#)]
51. Jackson, R.M.; Gabb, H.; Sternberg, M.J. Rapid refinement of protein interfaces incorporating solvation: Application to the docking problem. *J. Mol. Boil.* **1998**, *276*, 265–285. [[CrossRef](#)] [[PubMed](#)]
52. Cheng, T.M.-K.; Blundell, T.L.; Fernandez-Recio, J.; Fernandez-Recio, J.; Fernández-Recio, J. pyDock: Electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins Struct. Funct. Bioinform.* **2007**, *68*, 503–515. [[CrossRef](#)] [[PubMed](#)]
53. Grosdidier, S.; Fernández-Recio, J. Identification of hot-spot residues in protein-protein interactions by computational docking. *BMC Bioinform.* **2008**, *9*, 447. [[CrossRef](#)] [[PubMed](#)]
54. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2014.

