Structural Bioinformatics Survey on Disease-inducing Missense Mutations

Pietro Bongini[1], Simone Gardini[2], Monica Bianchini[1], Ottavia Spiga[1], and Neri Niccolai[1]

[1]University of Siena
[2]GenomeUp

August 18, 2020

Abstract

Understanding the molecular mechanisms that correlate pathologies with missense mutations is of critical importance for disease risk estimations and for devising personalized therapies. Thus, we have performed a bioinformatic survey of ClinVar, a database of human genomic variations, to find signals that can account for missense mutation pathogenicity. Arginine resulted as the most frequently replaced amino acids both in benign and pathogenic mutations. By adding the structural dimension to this investigation to increase its resolution, we found that arginine mutations occurring at the protein-DNA interface increase pathogenicity 6.5 times with respect to benign variants. Glycine is the second amino acid among all the pathological missense mutations. Necessarily replaced by larger amino acids, glycine replacements perturb the structural stability of proteins and, therefore, their functions, being mostly located in buried protein moieties. Arginine and glycine appear as representative of missense mutations causing respectively changes in interaction processes and in protein structural features, the two main molecular mechanisms of genome-induced pathologies.

1. Introduction

Molecular dialog among proteins, nucleic acids and small molecules is the essence of Life at an atomic resolution and, hence, understanding the basis of such dialog represents a step forward for genetic medicine [1]. Nowadays, genotype-to-phenotype associations in human diseases can be efficiently explored by accessing to mutation databases, such as HGMD [2], SwissVar [3], COSMIC [4], HuVarBase [5], HUMSAVAR [6] and ClinVar [7], where information on pathogenic mutations are collected. The fact that missense mutations constitute the most common sequence alteration in Mendelian disorders [8] offers a good starting point to understand the mechanisms of disease appearance due to amino acid variations in mutated proteins. Pathogenicity can arise from missense mutations whenever mutated proteins alter their structural stability and, consequently, their function. The large array of examples of this kind has driven the implementation of several algorithms to predict functional damages due to amino acid replacements [9-13]. An additional way to explain the pathological effect of some missense mutations takes into account the protein interactome dimension [14]. In the interactome, missense mutated proteins can be considered as the network nodes, being responsible for altered biochemical or biophysical properties that represent the network edges. Thus, specific modifications of the interaction pattern due to protein mutations define an edgotype, which has been proposed as a way of monitoring the effects that link genotypes to phenotypes [15,16]. The abundance of structures that are currently available in the Protein Data Bank (PDB) [17] allows a detailed view of protein interactome, particularly in light of tools such as PISA (Protein Structure, Interface and Assembly) provided by EBI (the European Bioinformatics Institute) [18]. PISA, indeed, is a database of pre-calculated results for the whole PDB archive for retrieving information on structural and chemical properties of macromolecular surfaces and interfaces. In the present report, we have performed a structural bioinformatic analysis of human mutation databases by using PISA database to obtain information on mechanisms of pathogenicity

of missense mutations at atomic resolution. By comparing benign and pathological missense mutations we tried to improve our understanding of specific roles of single amino acids in biological processes.

2. Materials and Methods

Due to the weekly updating, all results reported in this investigation refer to the database content that was declared on June 10, 2020. PDB files containing proteins were 161,599 out of the total 165,117 ones. PISA considered 137,793 protein-containing files. ClinVar listed 789,266 items.

2.1 Selection of human missense mutations.

By choosing from the default filter list in ClinVar web page*missense* as *molecular consequence* , we have made a preliminary selection of 308,326 entries. Then, we have refined our selection by applying the *clinical significance* filter for benign and pathogenic missense mutations obtaining respectively 25,579 and 22,153 items entries. From pathogenic missense mutations, we excluded 558 entries that were related to more than one mutations, often unlikely to occur, being associated with double or triple nucleotide changes between replaced amino acids. Thus, we have considered 21,595 ClinVar items with single missense mutations, which represent the content of our pathogenic missense mutation (PMM) data set. In the case of benign missense mutations, instead, all 25,579 were related to single mutations, being all these entries controlled with at least one star in the ClinVar nomenclature. Thus, all the 25,579 benign missense mutation (BMM) are single mutations entries and form the BMM data set.

2.2 Selection of structural files of proteins bearing pathogenic and benign missense mutations.

To bring the analysis to the structural level we have queried VarMap [19], to associate each of our mutations to a protein structure file. Structures were available in the PDB only for a fraction of the mutations. Furthermore, as PISA database contains only 84% of all protein PDB entries, we had to take into account 1,253 and 1,580 protein structures to locate respectively 5,641 PMMs and 3,018 BMMs. The surface accessibility of replaced amino acids has been calculated with POPS software [20]. PyMOL [21] was used for molecular graphics and to generate mutant structures by using its Mutagenesis routine; we used the Optimize plugin for energy minimization of mutants [22].

2.3 Profiling of amino acid occurrences for pathogenic and benign missense mutations. We have implemented a series of Python scripts, downloadable from https://github.com [23] to collect all amino acid replacements that occur in BMM and PMM data sets.

3. Results

3.1 Mapping amino acid replacements in human missense mutations .

Structural Bioinformatics procedures to find signals that can help to understand the mechanisms of benign and pathogenic mutagenesis and the role of single amino acids in this process require suitable data sets. Hence, we have selected ClinVar [24] as the basis for our investigations, due to the massive variety of weekly updated information on clinically relevant mutations that this databank offers. As of June 10, 2020, ClinVar reports 789,266 mutations, which can be directly filtered to obtain 308,326 missense mutation items by applying the*Molecular consequence* options offered by the ClinVar web home page. All amino acid replacements found in the latter missense mutation dataset are reported in Table 1, apart from those entries which did not give either the natural amino acid or the replacing one. It is apparent that several replacements are not allowed, like 597 and 14 missense mutations that would imply changes between amino acids with codons differing respectively for two or three nucleotides. Genome sequencing errors should be mostly responsible for these findings, but it will not be considered further. Furthermore, the fact that 13 self-mutations are also included in Table 1 suggests that some additional control is needed, such as the *At least one star* from *Review status*among ClinVar filtering options.

Thus, we got a final number of 25,579 BMMs and 21,595 PMMs, which we normalized, despite their close similarity, for a direct comparison of amino acid replacements in the two datasets. Afterward, to correlate expected and experimentally obtained mutation distributions, we have compared values of the scoring matrix

for amino acid substitutions with matrix elements obtained from the difference between BMM and PMM, see Table 2. Among the large variety of PFAM and BLOSUM substitution matrices, we have chosen BLOSUM62 [25], which is the default option for several sequence analysis procedures such as BLAST [26], see Table 2.

Three interesting features emerge from Table 2: a) only replacements among amino acids having codons with one nucleotide change are observed; b) in both datasets the most frequent missense mutations involve arginine; c) as expected, in the case of BMM the overall amount of normalized mutations are 3,829 (38,29%) favorable 3,925 (39,25%) less favorable and 2,246 (22,46%) unfavorable variants; in the case of PMM, instead, an opposite distribution is observed with 2,015 (20,15%) favorable, 3,362 (33,62%) less favorable and 4,623 (46,23%) unfavorable variants.

3.2 Mapping amino acid replacements in benign and pathogenic missense mutations .

Different missense mutation profiles were obtained for BMMs and PMMs. Apart from the large predominance of arginine in both sets of data, alanine is the second most frequent mutated amino acid among BMM dataset and glycine behaves similarly among PMMs, as summarized in Fig. 1.

The prevalence of arginine among all the missense mutations of our two data sets has been already observed [27] and ascribed mainly to the frequent presence of the 5'CpG dinucleotide along the DNA genomic sequence. CG moiety in genomic DNA has been observed to be prone to TG or CA mutations, due to the deamination of 5' methyl-cytosine (28). Arginine, indeed, has four nucleotide triplets, out of its six codons, that include the CG dinucleotide. Thus, CG/TG and CG/CA mutations yield R/C, H, Q, W that are the most abundant replacements, see Tables 1 and 2. Alanine, proline, serine and threonine have also one CG dinucleotide in their codons accounting, at least in part, for the amino acid occurrence profile in BMM dataset. The pathological effects of glycine replacements cannot be discussed on the basis of DNA sequences, and structural analysis of collected data is needed, vide infra.

3.3 Structural analysis of benign and pathogenic missense mutations .

From PISA database [18] we have derived the topology of 3,018 BMMs and 5,641 PMMs, even though the number of BMMs is larger than PMMs in ClinVar, underlining that structural biologists are predominantly concerned on proteins involved in diseases. For all the structurally characterized missense mutations, we have analyzed also the solvent accessible surface areas by using POPS [20], labeling as surface-exposed amino acids those having an exposed area higher than 20%. Table 3 summarizes our results for arginine and glycine, the two amino acids that are most involved in pathological mutations. We have reported also the topological distributions obtained for all the other eighteen amino acids, which can be considered as average reference values. From a preliminary overlook of Table 3, it is apparent that, among PMMs, arginine is more abundant than glycine and all the other amino acids both in PISA-defined interfaces and in protein surfaces. Furthermore, well above the PMM average, arginine is frequently located in protein-DNA interfaces.

4. Discussion

Almost one third of PMMs reported in Table 2, involves R (18.4%) and G (14.5%) with the remaining variants almost uniformly distributed among the other amino acids, see Fig. 1b. The peculiar characteristics of arginine and glycine have been invoked to explain this finding [27], as the former is prone to be replaced, despite its six different protecting codons, due to the high C-T and G-A transition probability observed for 5'-CpG dinucleotides [28]. However, data shown in Fig. 1a indicate that arginine is still very frequently encountered in BMMs (15.1%), while glycine occurrence reaches only a rather average value (6.5%), suggesting that different mechanisms contribute to their pathogenicity.

A comparative topological analysis of BMMs and PMMs, summarized in Table 3, clearly indicates that arginine mutations increase pathogenicity whenever they occur at PISA-defined interfaces and, particularly, at protein-DNA interfaces. In the protein-DNA interface, indeed, arginine is more than six times more frequent in PMM than BMM. This feature is in total agreement with the very critical role that this amino acid has in the interaction with nucleic acids [29]. Moreover, arginine PMM tends not to stay in buried protein moieties or in protein-protein interfaces, whereas glycine PMMs exhibit the opposite trend. It

is interesting to note that the latter glycyl mutations are well above the average more frequently found in protein-ligand interfaces, in agreement with the suggested role of this amino acid to stabilize concave moieties of the protein surface [30]. Prevalent localization of pathological glycine mutations indicates that its replacement with amino acids bearing larger side chains causes structural stress and, hence, functional changes in mutated proteins.

The fact that among BMMs there are also three cases of arginine substitutions at the protein-DNA interface, seems to contradict the relevance of this amino acid in the latter interface. Hence, we have manually checked the structural features of these three BMMs that are associated with two transcription regulators, ZFP568 [31] and DUX4 [32], structurally resolved in PDB ID: 5V3J and PDB ID: 5ZFZ respectively. We have used the two PDB structures to generate the R98/Q mutant structure in ZFP568, and R411/Q and R599/H mutant structures of DUX4. Fig. 2 shows how R/Q and R/H replacements can maintain protein-DNA binding with the glutamyl amide group and with the histidyl imidazole group. It is important to note that in both cases the original arginine duty was not to keep these two proteins tightly bound to DNA, as it would be needed in the case of histones, being transcription regulators rather mobile proteins along DNA trails.

Thus, we have used the large array of items contained in the ClinVar database for generating maps of amino acid replacements, confirming that arginine and glycine are the most involved protein residues in missense mutations. As expected, by comparing BMMs and PMMs, we have also proved that amino acid similarity plays a significant role in determining pathogenicity. With the present Structural Bioinformatics approach, by using PISA as a protein interface analyzer, we have searched at an atomic resolution those features that are responsible for pathogenic mutations. Arginine and glycine, the most frequently involved in PMMs, resulted as representatives of two different mechanisms of pathogenicity. Arginine replacements, indeed, resulted to be pathogenic when they involve interaction processes and glycine substitutions can be deleterious whenever they can determine structural stresses in mutated proteins. In the edgotype view of missense mutation effects [14], arginine perturbs network edges and glycine modifies its nodes.

Structural characterization of PMMs can be expanded outside the current limits of the PISA database, by implementing algorithms that can work on reliably predicted structures for the advancement of genomic medicine.

Authorship contribution statement

Pietro Bongini : Data curation, script implementation, writing review. **Simone Gardini** : Conceptualization, methodology, writing review. **Monica Bianchini** : Conceptualization, methodology, data curation, writing review. **Ottavia Spiga** : Methodology, sequence analyses, writing review & editing. **Neri Niccolai** : Conceptualization, methodology, writing original draft, writing review & editing.

Conflict of Interest Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statement

Data sharing not applicable – no new data generated

References

1 Horton, R.H., Lucassen, A.M. (2019). Recent developments in genetic/genomic medicine. *Clin Sci (Lond)* . 133(5), 697-708. doi:10.1042/CS20180436

2 http://www.hgmd.cf.ac.uk

3 https://swissvar.expasy.org/cgi-bin/swissvar/home)

4 https://cancer.sanger.ac.uk/cosmic

5 https://www.iitm.ac.in/bioinfo/huvarbase

6 https://omictools.com/analytics/bioinformatics/database/humsavar

7 https://www.ncbi.nlm.nih.gov/clinvar

8 Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A., Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet.* 133(1), 1-9. doi:10.1007/s00439-013-1358-4

9 Kulshreshtha, S., Chaudhary, V., Goswami, G.K., Mathur, N. (2016). Computational approaches for predicting mutant protein stability.*J Comput Aided Mol Des.* 30(5), 401-412. doi:10.1007/s10822-016-9914-3

10 Capriotti, E., Fariselli, P., Casadio, R. (2004). A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics* . 20 Suppl 1:i63-i68. doi:10.1093/bioinformatics/bth928

11 Fariselli, P., Martelli, P.L., Savojardo, C., Casadio, R. (2015). INPS: predicting the impact of non-synonymous variations on protein stability from sequence. *Bioinformatics* . 31(17), 2816-2821. doi:10.1093/bioinformatics/btv291

12 Pires, D.E., Chen, J., Blundell, T.L., Ascher, D.B. (2016). In silico functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. *Sci Rep.* 6:19848. doi:10.1038/srep19848

13 Pandurangan, A.P., Ochoa-Montaño, B., Ascher, D.B., Blundell, T.L. (2017). SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res.* 45(1), 229-235. doi:10.1093/nar/gkx439

14 Sahni, N., Yi, S., Taipale, M., et al. (2015). Widespread macromolecular interaction perturbations in human genetic disorders.*Cell* . 161(3), 647-660. doi:10.1016/j.cell.2015.04.013

15 Zhong, Q., Simonis, N., Li, Q.R., et al. (2009). Edgetic perturbation models of human inherited disorders. *Mol Syst Biol* . 5, 321. doi:10.1038/msb.2009.80

16 Sahni, N., Yi, S., Zhong, Q., et al. (2013). Edgotype: a fundamental link between genotype and phenotype. *Curr Opin Genet Dev* .23(6), 649-657. doi:10.1016/j.gde.2013.11.002

17 Berman, H.M., Westbrook, J., Feng, Z., et al. (2000). The Protein Data Bank. *Nucleic Acids Res* . 28(1), 235-242. doi:10.1093/nar/28.1.235

18 Krissinel, E., Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J Mol Biol.* 372(3), 774-797. doi:10.1016/j.jmb.2007.05.022

19 Stephenson, J.D., Laskowski, R.A., Nightingale, A., Hurles, M.E., Thornton, J.M. (2019). VarMap: a web tool for mapping genomic coordinates to protein sequence and structure and retrieving protein structural annotations. *Bioinformatics.* 35(22), 4854–4856.

20 Cavallo, L., Kleinjung, J., Fraternali, F. (2003). POPS: A fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic Acids Res* . 31(13), 3364-3366. doi:10.1093/nar/gkg601

21 *https://pymol.org*

22 O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., Hutchison, G.R. (2011). Open Babel: An open chemical toolbox. *J Cheminform* . 3, 33. doi:10.1186/1758-2946-3-33

23*https://github.com/PietroMSB/ClinVarAnalyzer*

24 Landrum, M.J., Lee, J.M., Riley, G.R., et al. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype.*Nucleic Acids Res* . 42(Database issue), 980-985. doi:10.1093/nar/gkt1113

25 Henikoff, S., Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* . 89(22), 10915-10919. doi:10.1073/pnas.89.22.10915

26 Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol.* 215(3), 403-410. doi:10.1016/S0022-2836(05)80360-2

27 Vitkup, D., Sander, C., Church, G.M. (2003). The amino-acid mutational spectrum of human genetic disease. *Genome Biol.* 4(11), 72. doi:10.1186/gb-2003-4-11-r72

28 Antonarakis, S.E., Krawczak, M., Cooper, D.N. (2000). Disease-causing mutations in the human genome. *Eur J Pediatr* . 159 Suppl 3, 173-178. doi:10.1007/pl00014395

29 Gardini, S., Furini, S., Santucci, A., Niccolai, N. (2017). A structural bioinformatics investigation on protein-DNA complexes delineates their modes of interaction. *Mol Biosyst* . 13(5), 1010-1017. doi:10.1039/c7mb00071e

30 Bongini, P., Niccolai, N., Bianchini, M. (2019). Glycine-induced formation and druggability score prediction of protein surface pockets.*J Bioinform Comput Biol* . 17(5), 1950026. doi:10.1142/S0219720019500264

31 Patel, A., Yang, P., Tinkham, M., et al. (2018). DNA Conformation Induces Adaptable Binding by Tandem Zinc Finger Proteins. *Cell.*173(1), 221-233. doi:10.1016/j.cell.2018.02.058

32 Li, Y., Wu, B., Liu, H., et al. (2018). Structural basis for multiple gene regulation by human DUX4. *Biochem Biophys Res Commun* . 505(4), 1161-1167. doi:10.1016/j.bbrc.2018.10.056

Tables

Table 1

Amino acid distributions of ClinVar items only related to missense mutations

	A	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**	**M**	**N**	**P**	**Q**
A	0	2	1126	721	32	1677	1	22	17	50	7	21	1714	0
C	1	0	1	0	846	577	2	0	1	7	0	2	1	0
D	607	0	0	2395	10	2669	1535	9	1	2	0	5144	2	0
E	1046	3	2836	0	3	2301	2	5	7794	25	13	3	2	2047
F	5	669	0	1	1	1	2	412	0	2791	1	5	1	0
G	1759	943	3195	2696	4	0	3	5	3	10	0	6	2	1
H	0	7	500	1	1	0	1	2	0	497	0	470	654	1243
I	0	1	1	1	849	0	2	4	136	951	1742	749	0	1
K	9	0	1	2728	2	1	0	241	1	7	291	2164	1	1024
L	3	0	0	1	3550	4	356	840	2	0	753	1	4121	524
M	3	2	0	1	2	1	1	2387	521	1134	2	3	1	0
N	1	2	1651	1	1	2	763	600	1977	3	0	0	1	0
P	2067	1	1	0	8	1	786	2	2	7798	0	0	0	590
Q	0	0	0	1461	0	2	2151	0	982	550	1	0	1065	0
R	2	8057	1	2	2	2867	8765	247	1559	1992	166	6	1696	9871
S	800	1771	7	2	2268	1885	2	734	12	2897	1	2588	1841	0
T	3608	0	1	1	1	4	3	4299	656	3	4041	1050	969	1
V	2823	0	475	482	1050	1032	1	5858	2	3063	5212	1	3	1
W	0	721	0	0	4	244	0	0	3	205	1	0	1	4
Y	0	3376	396	1	483	0	1418	4	3	1	0	372	0	4
total	12734	15555	10192	10495	9117	13268	15794	15671	13672	21986	12231	12585	12075	15311

Amino acid distributions of ClinVar items only related to missense mutations. Rows describe how each of the

natural amino acids has been replaced by column residues. Colors refer to the number of codon nucleotides involved in mutations: green, yellow and red indicate respectively one, two and three nucleotide changes.

Table 2

Replacing matrix of normalized elements of BMM (upper) and PMM (lower)

	A	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**
A	0	0	27	22	0	66	0	0	0	0	0	0	51	0	0	98	415	334	0
C	0	0	0	0	10	10	0	0	0	0	0	0	0	0	19	19	0	0	5
D	14	0	0	96	0	58	41	0	0	0	0	176	0	0	0	0	0	19	0
E	28	0	105	0	0	63	0	0	227	0	0	0	0	58	0	0	0	22	0
F	0	10	0	0	0	0	0	8	0	62	0	0	0	0	0	18	0	10	0
G	60	20	81	66	0	0	0	0	0	0	0	0	0	0	171	184	0	59	7
H	0	0	10	0	0	0	0	0	0	10	0	14	12	39	82	0	0	0	0
I	0	0	0	0	15	0	0	0	1	35	49	13	0	0	0	10	107	247	0
K	0	0	0	71	0	0	0	5	0	0	8	51	0	30	131	0	23	0	0
L	0	0	0	0	109	0	7	35	0	0	26	0	58	10	25	25	0	113	5
M	0	0	0	0	0	0	0	72	10	37	0	0	0	0	6	0	74	119	0
N	0	0	55	0	0	0	25	14	55	0	0	0	0	0	0	206	19	0	0
P	82	0	0	0	0	0	26	0	0	300	0	0	0	26	67	215	69	0	0
Q	0	0	0	50	0	0	73	0	28	20	0	0	26	0	107	0	0	0	0
R	0	241	0	0	0	71	360	3	60	44	4	0	32	435	0	44	17	0	190
S	35	53	0	0	49	83	0	19	0	118	0	118	61	0	54	0	72	0	5
T	170	0	0	0	0	0	0	149	17	0	215	29	25	0	14	83	0	0	0
V	110	0	4	6	21	19	0	361	0	104	227	0	0	0	0	0	0	0	0
W	0	9	0	0	0	2	0	0	0	1	0	0	0	0	19	2	0	0	0
Y	0	60	5	0	12	0	36	0	0	0	0	4	0	0	0	3	0	0	0
total	501	395	287	313	218	374	569	670	401	737	531	405	265	598	695	907	796	925	213

	A	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**
A	0	0	51	35	0	22	0	0	0	0	0	0	83	0	0	29	156	154	0
C	0	0	0	0	71	38	0	0	0	0	0	0	0	0	163	78	0	0	37
D	21	0	0	43	0	87	56	0	0	0	0	130	0	0	0	0	0	56	0
E	18	0	44	0	0	51	0	0	263	0	0	0	0	27	0	0	0	23	0
F	0	33	0	0	0	0	0	15	0	96	0	0	0	0	0	75	0	24	0
G	52	73	238	176	0	0	0	0	0	0	0	0	0	0	474	219	0	193	25
H	0	0	16	0	0	0	0	0	0	16	0	10	24	34	68	0	0	0	0
I	0	0	0	0	27	0	0	0	9	11	29	45	0	0	6	25	104	25	0
K	0	0	0	68	0	0	0	6	0	0	4	55	0	12	33	0	16	0	0
L	0	0	0	0	84	0	17	5	0	0	7	0	314	24	90	38	0	47	14
M	0	0	0	0	0	0	0	96	38	37	0	0	0	0	54	0	101	109	0
N	0	0	34	0	0	0	14	25	68	0	0	0	0	0	0	87	12	0	0
P	23	0	0	0	0	0	23	0	0	217	0	0	0	16	64	89	38	0	0
Q	0	0	0	19	0	0	37	0	19	10	0	0	55	0	52	0	0	0	0
R	0	387	0	0	0	115	293	6	31	106	4	0	128	346	0	64	20	0	327
S	6	29	0	0	72	23	0	18	0	88	0	35	79	0	60	0	12	0	10
T	51	0	0	0	0	0	0	94	22	0	103	19	44	0	29	12	0	0	0
V	45	0	35	26	37	46	0	29	0	62	114	0	0	0	0	0	0	0	0
W	0	50	0	0	0	19	0	0	0	14	0	0	0	0	87	20	0	0	0

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W
Y	0	165	27	0	6	0	55	0	0	0	0	20	0	0	0	24	0	0	0
total	218	739	448	370	299	405	497	295	454	662	263	317	731	462	1186	766	463	634	414

Table cells are colored according to BLOSUM62 matrix values: cyan, orange and magenta refer respectively to positive (favorable change), 0 to -1 (less favorable changes) and <-1 (unfavorable changes) values. Both rows and columns totals diverge from 10,000 due to figure rounding.

Table 3

Topology of ClinVar benign and pathological missense mutations.

PMM (BMM)	#G	#R	#others	%G	%R	%others
Total missense mutations from Clinvar	3,136 (1,666)	3,960 (3,855)	14,499 (20,058)			
Missense mutations from PISA database	680 (198)	1,134 (451)	3,827 (2,369)			
Missense mutations from PISA interfaces	171 (25)	419 (108)	1,082 (353)	25.1 (12.6)	36.9 (23.9)	28.3 (14.9)
Missense mutations exposed to protein surfaces	102 (98)	297 (230)	711 (1,124)	15.0 (49.5)	26.2 (51.0)	18.5 (47.4)
Missense mutations in protein cores	407 (69)	418 (113)	2,034 (892)	59.8 (34.8)	36.8 (25.0)	53.1 (37.6)
Missense mutations in protein-protein interfaces	109 (15)	228 (60)	706 (204)	**60.2 (53.6)**	**51.2 (52.2)**	**61.1 (55.4)**
Missense mutations in protein-DNA interfaces	2 (0)	75 (3)	61 (9)	**1.1 (0)**	**16.8 (2.6)**	**5.2 (2.4)**
Missense mutations in protein-RNA interfaces	0 (1)	2 (3)	2 (4)	**0 (3.5)**	**0.4 (2.6)**	**0.1 (1.0)**

PMM (BMM)	#G	#R	#others	%G	%R	%others
Missense mutations in protein-ligand interfaces	70 (12)	140 (49)	386 (151)	**38.7 (42.8)**	**31.4 (42.6)**	**33.4 (41.0)**

In parenthesis, data for benign missense mutations; percentages are calculated for mutations in PISA databases and, with bold figures, in PISA interfaces.

Captions to the figures

Table 1: Rows describe how each of the natural amino acids has been replaced by column residues. Colors refer to the number of codon nucleotides involved in mutations: green, yellow and red indicate respectively one, two and three nucleotide changes.

Table 2: Table cells are colored according to BLOSUM62 matrix values: cyan, orange and magenta refer respectively to positive (favorable change), 0 to -1 (less favorable changes) and <-1 (unfavorable changes) values. Both rows and columns totals diverge from 10,000 due to figure rounding.

Table 3: In parenthesis, data for benign missense mutations; percentages are calculated for mutations in PISA databases and, with bold figures, in PISA interfaces.

Figure 1: Profiles of amino acid occurrences in missense mutations reported in ClinVar database. Benign and pathological mutations are shown in a and b respectively.

Figure 2: Cartoon representations of two proteins bearing three ClinVar BMMs involving arginine. Mutants are obtained from wild type structures by using PyMOL mutation routine and its Optimize plugin.