COMPENSATORY EPISTASIS EXPLORED BY MOLECULAR DYNAMICS SIMULATIONS

Catarina Serrano[1,2,3], Carla S.S. Teixeira[4], David N. Cooper[5], João Carneiro[6], Monica Lopes-Marques[1,2,3], Peter D. Stenson[5], António Amorim[1,2,3], Maria J. Prata[1,2,3], Sérgio F. Sousa[4], Luísa Azevedo[1,2,3]

[1] i3S- Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Population Genetics and Evolution Group, Rua Alfredo Allen 208, 4200-135 Porto, Portugal

[2] IPATIMUP-Institute of Molecular Pathology and Immunology, University of Porto, Rua Júlio Amaral de Carvalho 45, 4200-135 Porto, Portugal

[3] Department of Biology, Faculty of Sciences, University of Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

[4] UCIBIO/REQUIMTE, BioSIM - Departamento de Biomedicina, Faculdade de Medicina da Universidade do Porto, Porto, Portugal

[5] Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff CF14 4XN, UK

[6] CIIMAR, Interdisciplinary Centre of Marine and Environmental Research, Terminal de Cruzeiros do Porto de Leixões, Av. General Norton de Matos, s/n, 4450-208, Matosinhos, Portugal

Sérgio F. Sousa and Luísa Azevedo contributed equally to this manuscript.

To whom correspondence should be addressed:
- Sérgio F. Sousa (sergiofsousa@med.up.pt)
- Luísa Azevedo (lazevedo@ipatimup.pt)

ABSTRACT

A non-negligible proportion of human pathogenic variants are known to be present as wild-type in at least some non-human mammalian species. The standard explanation for this finding is that molecular mechanisms of compensatory epistasis can alleviate the mutations' otherwise pathogenic effects. Examples of compensated variants have been described in the literature but the interacting residue(s) postulated to play a compensatory role have rarely been ascertained. In this study, the examination of five human X-chromosomally-encoded proteins (FIX, GLA, HPRT1, NDP and OTC) allowed us to identify several candidate compensated variants. Strong evidence for a compensated/compensatory pair of amino acids in the coagulation FIXa protein (involving residues 270 and 271) was found in a variety of mammalian species. Both amino acid residues are located within the 60-loop, spatially close to the 39-loop that performs a key role in coagulation serine proteases. To understand the nature of the underlying interactions, molecular dynamics simulations were performed. The predicted conformational change in the 39-loop consequent to the Glu270Lys substitution (associated with hemophilia B) appears to impair the protein's interaction with its substrate but, importantly, such steric hindrance is largely mitigated in those proteins that carry the compensatory residue (Pro271) at the neighboring amino acid position.

Keywords: Evolutionary conservation; pathogenic variants; molecular interactions; molecular dynamics; compensatory epistasis

INTRODUCTION

It is almost a truism that the evolutionary conservation of any given amino acid residue in a particular protein reflects its relevance to that protein's function. Thus, if a specific amino acid position is evolutionarily highly conserved, this is indicative of enduring functionality for the position and any substitution in humans is predicted to be associated with pathogenicity (de Beer et al. 2013; Kumar et al. 2009; Subramanian and Kumar 2006). The corollary to this is that as the evolutionary conservation of the site in which the amino acid substitution occurs decreases, the lower is the severity of the resulting clinical phenotype (Miller and Kumar 2001; Wacey et al. 1994).

This notwithstanding, the genetic context in which any given pathogenic variant occurs may be responsible for exacerbating its deleterious effect or, conversely, for ameliorating its pathogenicity through an epistatic mechanism of compensation between variants (Genin et al. 2008; Jordan et al. 2015; van Leeuwen et al. 2016). Such a pairwise compensatory interaction is exemplified by the occurrence of deleterious variants in humans that in non-human species constitute wild-type residues (Azevedo et al. 2009; Azevedo et al. 2016; Mouse Genome Sequencing Consortium 2002; Ferrer-Costa et al. 2007; Kondrashov et al. 2002; Marin et al. 2019; Rhesus Macaque Genome Sequencing and Analysis Consortium 2007). If a deleterious variant in a human disease-associated protein does not have the same detrimental effect on an orthologous protein in other evolutionarily related species, the toleration of the variant in non-human genomes can be explained, at least in some cases, by antagonistic epistatic interactions, in which the effect of a deleterious allele is ameliorated by the co-occurrence of another variant that acts so as to introduce a compensatory amino acid residue into the same protein or an interacting partner protein (Jordan et al. 2015; Kondrashov et al. 2002; Suriano et al. 2007). Such sequence variants have been termed 'compensated pathogenic deviations' (CPDs) (Kondrashov et al. 2002). Numerous descriptions of CPDs have emerged through comparative analyses of genomes from multiple mammalian species (Chimpanzee Sequencing and Analysis Consortium 2005; Rhesus Macaque Genome Sequencing and Analysis Consortium 2007; Xue et al. 2015; Zhang et al. 2011; Zhang et al. 2010). Further examples of CPDs were identified among human inherited pathogenic variants, and it has been estimated that at least 3.7% of human pathogenic missense variants correspond to CPDs (Azevedo et al. 2016). Despite the increasing number of CPDs reported in the literature, only rarely has the compensatory site been identified (Azevedo et al. 2009;

Jordan et al. 2015; Kondrashov et al. 2002) and even more rarely has experimental validation been obtained (e.g.Jordan et al. 2015; Suriano et al. 2007).

Since in almost all mammals the X-chromosome is present as a single copy in males, the process of identifying and validating CPDs can be simplified by examining X-linked genes. Therefore, taking advantage of this natural model system, in which the action of purifying selection is expected to be fully exposed in hemizygous individuals, we opted to study X-chromosome-encoded disease proteins. Five X-chromosome-encoded proteins were specifically targeted viz. coagulation factor IX (FIX), alpha-galactosidase A (GLA), hypoxanthine-guanine phosphoribosyltransferase (HPRT1), norrin (NDP) and ornithine transcarbamylase (OTC).

With few exceptions, eutherian X-chromosomes are evolutionarily highly conserved both in terms of gene content and function, with only a few genes known to escape dosage compensation and monoallelic expression. We therefore operated under the assumption that if an allele that was deleterious in humans appeared to be fixed in a non-human species, it was very likely to co-occur in the latter with compensatory partner amino acid residue(s), otherwise it would have been long since removed from the population under the action of purifying selection. This study represents a concerted attempt to explore the molecular basis of the epistatic interactions between two residues in mammalian genomes. To address the issue, we employed protein modeling and molecular dynamics simulations, key tools for the study of intra-molecular amino acid interactions (Castellana et al. 2021; O'Rourke et al. 2016).

METHODS

Proteins and disease-causing variants

Missense mutation density (i.e. the number of pathogenic missense variants per protein sequence length in amino acids) was determined from data retrieved from the Human Gene Mutation Database (HGMD) (Stenson et al. 2020). The five proteins encoded by X-linked genes with the highest mutation density and with high resolution crystal structures available from the Protein Data Bank (Berman et al. 2000) were selected for this study: FIX, GLA, HPRT1, NDP and OTC. These proteins are associated, respectively, with the following Mendelian conditions: hemophilia B (OMIM 300746), Fabry disease (OMIM 300644), Lesch-Nyhan syndrome (OMIM 308000), Norrie disease

(OMIM 300658) and ornithine transcarbamylase deficiency (OMIM 300461). A graphical representation of the analyses performed are shown in Fig. 1 and are detailed below.

Multiple sequence alignment and evolutionary analysis

DNA coding sequences (CDS) from placental mammals were retrieved from Ensembl Genome Browser v.97 and v.100 (Cunningham et al. 2019; Yates et al. 2020). The corresponding protein accession numbers are given in Supplementary Table S1. Using an in-house Python script, sequences with more than 1% unknown nucleotides in a DNA sequence and/or with a size difference of at least 25% compared to the human reference sequence were removed from the dataset to ensure valid comparisons between *bona fide* orthologous sequences and to avoid the analysis of less well characterized sequences. A codon-based multiple sequence alignment (MSA) was built using PRANK (Loytynoja and Goldman 2010) and the resulting alignments were manually inspected.

CPD variants dataset and their putative compensatory partners

A total of 628, 428, 259, 170 and 76 disease-causing missense variants for FIX, GLA, OTC, HPRT1 and NDP respectively, were retrieved from the HGMD (Stenson et al. 2020). Protein residues in the mammalian orthologues that harbored the human deleterious variants as wild-type amino acids were classified as potential CPDs. In order to maximize the probability of correctly identifying the corresponding compensatory partners, only cases in which the CPD was shared by at least six mammalian species were considered, under the assumption that the larger the number of species harboring a given CPD, the higher the probability that they would share a recent common ancestor and, therefore, the higher the probability that all the species in which the CPD occurs will share the same compensatory partner (Supplementary Table S2).

Three-dimensional (3D) analysis was performed on the crystal structure of the serine protease catalytic domain of coagulation FIXa (the active enzymatic form of FIX) (6MV4 (Vadivel et al. 2019)) and a dimeric structure of α-galactosidase (3s5z (Guce et al. 2011)) available from the Protein Data Bank (Berman et al. 2000). For each CPD shared by at least six species and located within the crystal structure of the corresponding protein (6MV4 and 3S5Z for FIXa and GLA, respectively), putative compensatory

variants were identified by protein sequence comparisons. A variant was considered to be a putative compensatory partner for a CPD when it co-occurred in all species in which the CPD was present and the amino acid residue differed from the wild-type found in humans. The molecular distance between the CPD and its putative compensatory variant was determined with PyMOL software (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC) (Supplementary Table S3).

Human coagulation FIXa model set-up and parameterization

A molecular model for the serine protease domain of wild-type human coagulation FIXa was prepared from 6MV4 (Vadivel et al. 2019), an X-ray crystallographic structure with a resolution of 1.37 Å. The protonation states of all the amino acid residues were predicted using PropKA version 3.0 at pH 7.0 (Olsson et al. 2011). The system was prepared using the AMBER18 software package and LEAP, using the ff14SB force field (Maier et al. 2015). Charges on the system were neutralized through the addition of counter-ions (Na^+) and the system was placed in a TIP3P water box with a minimum distance of 12 Å between the protein-surface and the side of the box, using the LEAP module of AMBER. For the mutants, the initial wild-type molecule was modeled with the mutagenesis feature in PyMOL 1.7.2.1 using the Dunbrack rotamer libraries available (Shapovalov and Dunbrack 2011). Selection of the initial amino-acid conformation of each mutant was based on the dominant conformer predicted, excluding clashes with other amino acid residues in the protein. This computational mutagenesis protocol has been previously used successfully in the study of other enzymes (Ferreira et al. 2017). The mutant systems were subject to the same solvation and neutralization protocol described for the wild-type protein. The modelled FIXa systems are listed in Table 1.

Molecular Dynamics simulations

All systems were subjected to four consecutive energy minimization stages to remove clashes prior to the molecular dynamics (MD) simulation. In these four stages, the minimization procedure was applied to the following atoms of the system: 1 - water molecules (2500 steps); 2 - hydrogen atoms (2500 steps); 3 - side chains of the amino acid residues (2500 steps); 4 - full system (10,000 steps). The energy minimized systems were then subjected to a two-stage MD equilibration procedure: in the first stage (50 ps),

the systems were gradually heated to 310.15 K using a Langevin thermostat at constant volume (NVT ensemble); in the second stage (50 ps), the density of the systems was further equilibrated at 310.15 K.

Finally, MD production simulation runs were performed for 100 ns for the wild-type FIXa protein and for the three mutants listed in Table 1. These were performed with an NPT ensemble at constant temperature (310.15 K, Langevin thermostat) and pressure (1 bar, Berendsen barostat), with periodic boundary conditions, with an integration time of 2.0 fs using the SHAKE algorithm to constrain all covalent bonds involving hydrogen atoms. A 10 Å cutoff for nonbonded interactions was used during the entire molecular simulation procedure. Coordinates were saved at each 10 ps. This procedure (energy minimization to MD production simulation) was repeated twice in order to obtain two MD replicates for each FIXa variant listed in Table 1. Final trajectories were analyzed in terms of backbone Root-mean-square deviation (RMSD), root-mean-square fluctuation (RMSF), cluster analysis, hydrogen bonds formed, solvent accessible surface areas (SASAs) and dynamic cross-correlation maps (DCCM). The DCCMs were obtained with the Bio3D R package (Grant et al. 2006).

RESULTS AND DISCUSSION

Identification and evolutionary conservation of CPDs in non-human placental mammals

For each protein (FIX, GLA, HPRT1, NDP and OTC), an MSA alignment derived from placental mammalian sequences was analysed, and manually annotated with disease-causing variants. Initially, homologous mammalian amino acid positions harboring a human pathogenic allele as the wild-type in any of the species studied, were considered to be potential CPDs. The complete set of the potential CPDs identified is given in Fig. 2a and Supplementary Table S2. The proportion of pathogenic missense variants assumed to be CPDs varied between 0.5% (HPRT1) and 4.2 % (GLA) in the set of proteins examined (Fig. 2b), which is in accord with previous findings (Azevedo et al. 2016). In some cases, the CPDs were detected in only one or very few species, whereas in other cases the CPDs were spread over a large number of species. For example, the FIX pathogenic variants Glu142Lys, Val257Ile and Glu323Lys were present in about one half

of the mammalian species examined whereas Arg75Gln and Thr84Ala were each found in only one species (Fig. 2a and Supplementary Table S2).

Identification of putative compensatory amino acid partners

Under the hypothesis that our candidate CPDs are tolerated in the genomes of non-human mammalian species due to compensatory interactions with other amino acids, we next attempted to identify the putative compensatory partners so that the CPD/compensatory pairs could be studied by MD. The strategy adopted was to seek amino acid differences between the human protein and its orthologues harboring the CPDs, under the assumption that any difference between the human and non-human sequences would represent a potential compensatory amino acid position. If the analysis was based on comparisons made with one or a very few species, the amino acid differences with respect to the human protein would be too numerous to reach meaningful conclusions. Therefore, to increase the prospect of finding *bona fide* compensatory sites among those common to a group of species while being distinct from the human sequence, we adopted a conservative approach whereby the results were filtered so that we considered only CPDs present in at least six mammalian species. A total of seven candidate CPDs (Arg3His, Cys18Phe, Glu142Lys, Phe224Leu, Val257Ile, Glu270Lys and Glu323Lys) in FIX, two (Met72Ile and Arg363Cys) in GLA and one (Thr125Met) in OTC, matched this criterion. Because the OTC Thr125Met has been previously studied by enzyme activity assays (Suriano et al. 2007), it was not explored further by MD simulations.

Since the available 3D structures for FIXa and GLA do not encompass the entirety of these proteins, only three candidate CPDs in FIXa (Val257Ile, Glu270Lys and Glu323Lys) and two in GLA (Met72Ile and Arg363Cys) were contained within the retrieved 3D structures. The list of all shared amino acid differences between the human and other mammalian FIXa and GLA proteins for these five CPDs are shown in Supplementary Table S3. In four out of the five putative CPDs, the number of differences between the human and non-human proteins was too high to identify the compensatory site(s) with any degree of certainty, according to the aforementioned criteria. However, in the case of Glu270Lys in FIXa, two putative compensatory residues were disclosed: Pro271 and Ile368. As far as Ile368 is concerned, the 3D distance to the mutated amino acid at residue 270 (24 Å, Supplementary Table S3) renders it unlikely to be the compensatory partner. By contrast, the immediate proximity of Pro271 reinforces our

argument that Pro271 could be the sought after compensatory site. We therefore retained the FIXa 270/271 pair to be further investigated by MD simulations. Thus, from our initial list of 44 possible CPDs, we ended up with a single high-confidence CPD/compensatory pair suitable for MD simulations. This protracted process demonstrates how challenging the identification of a *bona fide* compensatory partner for a given CPD can be, a process which can be hampered by the extensive background genetic variability that exists between orthologous protein sequences.

Despite the juxtaposition of Lys270 and Pro271 in the FIXa 3D structure, the question nevertheless remained as to whether they constituted a genuine CPD/compensatory pair. We next recruited MD simulations as a tool to establish whether the putative compensatory variant might serve to stabilize the protein that would otherwise be destabilized by the pathogenic variant. To this end, MD simulations were performed and compared i) between sequences containing Glu270 or Lys270, to assess whether the differences between both simulations could provide clues as to the changes in function and likely pathogenicity associated with Lys270, and ii) between sequences containing different combinations of Glu270Lys and Thr271Pro to establish if the variant Thr271Pro could act as a suppressor of the pathogenic effect of Glu270Lys.

Analyses of the impact of amino acid variants at positions 270 and 271 in the catalytic domain of FIXa

To investigate the possible interactions between Glu270Lys and Thr271Pro, four FIXa structures were prepared containing the four variants described in Table 1 (wild-type, Glu270Lys, Glu270Lys+Thr271Pro and Thr271Pro); and the RMSD of the alpha carbon atoms (Cα) of all residues from the FIXa catalytic domain were calculated along the 100 ns of the MD simulations.

The analysis of the graphical representation of the RMSD of all structures (Fig. 3) shows that the protein containing the Glu270Lys deleterious variant has a distinctive profile when compared with the other three FIXa structures. In line with the data from both the graphical RMSD representation and the 2D RMSD plots, the calculated RMSD mean and respective standard deviation of the RMSD values reveal that the FIXa Glu270Lys protein has a higher mean RMSD and a higher standard deviation when compared with the other three protein variants (Supplementary Table S4). The addition of the compensatory variant (Glu270Lys + Thr271Pro) reduced the RMSD of the protein

to a value closer to that of the wild-type protein. The RMSD profiles also indicated that FIXa proteins stabilize at 40 ns of simulation, with the exception of the FIXa Glu270Lys protein which shows an irregular pattern with a higher standard deviation. Analysis of equilibrium properties were based on the last 60 ns of the MD simulations. These results are in agreement with the analysis of the RMSD of the Cα atoms from the residues 265 to 274 of the loop containing the mutated residues (60-loop) (Supplementary Table S4). The analysis of the overall solvent accessible surface area (SASA) of the different proteins (Supplementary Table S4) did not reveal any meaningful differences.

To evaluate the influence of the Glu270Lys, Glu270Lys + Thr271Pro and Thr271Pro mutations in the dynamic behavior of the FIXa protein, dynamic cross-correlation maps (DCCMs) of Cα atoms of all residues from the catalytic domain were calculated for the four variants. Comparison of the four resulting DCCMs (Fig. 4) clearly shows a distinct pattern for the FIXa with the Glu270Lys variant, with an increased level of correlated and anti-correlated motions, particularly in the ranges between residues 257 and 277 (which includes the 60-loop) and between residues 377 and 477.

At this stage, both RMSD and DCCM analyses clearly indicated that the structural impact of the human deleterious variant Glu270Lys is particularly evident in the 60-loop but that such an impact is ameliorated by the inclusion of the putative compensatory partner Thr271Pro.

Analyses of the flexibility of the FIXa catalytic domain

To identify the amino acid residues associated with the differences in behavior observed between the four FIXa protein sequences in the previous analyses, a Root-Mean-Square Fluctuation (RMSF) calculation of the Cα atoms of all residues from the catalytic domain was performed. RMSF measures the relative positional variability (i.e. flexibility) of the different amino acid positions along the protein backbone, in other words how much a particular residue fluctuates during a simulation, thereby allowing a distinction to be made between highly and poorly mobile regions of a protein. The analysis of the four different profiles demonstrated that the residues which stand out in the overall profile are Lys247 (from the 39-loop), plus residues Arg384 and Lys387. The flexibility of these three residues is clearly increased in the Glu270Lys protein and reduced in the Glu270+Thr271 protein, becoming closer to that of the wild-type protein (Fig. 5a).

The analysis of the RMSF values from the residues that comprise the active site (Fig. 5b; active site) further demonstrates that the flexibility of residue Ser411 (which contributes to the catalytic triad His221-Asp269-Ser365) is markedly higher in the wild type FIXa. All three mutant FIXa proteins present similar RMSF values, even though a small difference is evident in the double mutant FIXa Glu270Lys + Thr271Pro, whose RMSF is slightly higher and closer to the wild-type. The flexibility of the 39-loop (Fig b: 39-loop) is significantly affected by the introduction of the Glu270Lys variant. The RMSF values of this loop are clearly higher for the FIXa Glu270Lys protein than for the remaining three proteins. The introduction of a putative compensatory Thr271Pro variant serves to revert this increase in flexibility to values closer to the wild-type, whereas the single mutant Thr271Pro does not appear to affect the 39-loop.

Analysis of the 60-loop (Fig. 5b; 60-loop), where the Glu270Lys and Thr271Pro variants are located (yellow bars in Fig. 5b), reveals some small differences between the four RMSF profiles, although the overall behavior of the loop in terms of RMSF is similar between the four FIXa proteins.

The flexibility of the region between residues 375 to 393 (Fig. 5b, last group) is profoundly disturbed by the introduction of the Glu270Lys variant. Residues 384 and 387 from the FIXa Glu270Lys protein clearly stand out from the RMSF profile. The introduction of the compensatory Thr271Pro variant however reverses this effect and lowers the flexibility of residues 385 to 390 to values significantly lower than those obtained for the wild-type FIXa protein. The Thr271Pro replacement on its own does not appear to have such a marked effect in this region, although it increases the flexibility of residues 385 and 386 to RMSF values very close to those obtained for the FIXa Glu270Lys mutant.

Taken together, the results obtained from the RMSF analyses are in agreement with data from both the RMSD and DCCM analyses in which the structural disturbance caused by the FIXa disease-associated variant Glu270Lys was less dramatic in relation to the wild-type when the Thr271Pro was added to the background, thereby reinforcing the compensatory effect that a proline at position 271 has when it co-occurs in *cis* with a lysine at position 270. We performed these analyses in order to explain the structural differences between the Glu270Lys mutant and the wild-type using a cluster analysis as described in the next section.

Cluster analysis of the 39-loop and residues 375 to 393

In order to obtain the representative structures of the FIXa variants containing different conformations for the 39-loop and residues 375 to 393, a cluster analysis was performed of the protein conformations adopted along the MD for the four variants, based on the variability at these regions. The MD structures of each FIXa protein were grouped into two clusters according to the RMSD for these regions. Table 2 summarizes the proportion of structures along each MD simulation that comprise each cluster. The wild-type, Glu270Lys+Thr271Pro and Thr271Pro proteins adopt a set of conformations distributed between two main clusters: a dominant cluster, that represents between 72.6% (WT) and 74.7% (Thr271Pro) of the conformations, and an alternative cluster, representing between 27.4% (WT) and 25.3% (Thr271Pro) of the conformations observed in the MD simulations. By contrast, Glu270Lys remains almost locked in a single dominant conformation during 98.5% of the simulation.

To obtain further insights into the major structural differences between the wild-type and mutant FIXa proteins, representative structures of the dominant cluster of each FIXa protein were structurally aligned. This alignment showed that Lys247 (39-loop) from the Glu270Lys + Thr271Pro structure adopts a rather distinctive conformation when compared with the other three versions of the protein. In addition, residues Arg384 and Lys387 also assume distinctive conformations in the Glu270Lys and Glu270Lys + Thr271Pro proteins when compared both with each other and with the other two proteins (Fig. 6). Finally, the analysis of the 2D RMSD plots for the 100 ns of MD simulations of the 39-loop and residues 375 to 393 from the different FIXa variants shows that the Glu270Lys protein has the highest RMSD variation (1-10 Å) during the last 70 ns of the MD simulation whereas the Glu270Lys+Thr271Pro protein has the lowest range of RMSD variation (1-5 Å) during the 100 ns of simulation (Supplementary Figure S1).
Both the mean RMSD and SASA from the 39-loop and residues 375-393 are consistent with the RMSF values that suggested relevant conformational changes in these particular locations between the different FIXa proteins (Supplementary Table S5).

In-depth analysis of the structural changes resulting from the Glu270Lys replacement

Analysis of the tertiary structure of the different proteins (Fig. 6) showed that residues Arg384 and Lys387 are too distant from the 60-loop to be able to interact directly with the loop and therefore with amino acid positions 270 and 271. Although the results from the RMSF, RMSD and SASA analyses clearly show that these residues undergo a conformational change in the Glu270Lys mutant protein that is not observed in the other proteins, it is not possible to directly correlate these changes with the Glu270Lys replacement. Nevertheless, we cannot rule out that the observed alterations in this region could have resulted from the influence of Lys270 via long-range electrostatic interactions or from the rearrangement of the network of hydrogen bonds.

On the other hand, analysis of Lys247 (part of the 39-loop, Fig. 5) revealed that it is spatially close to Glu270 and Thr271 (60-loop) and may therefore be influenced by the mutated residues. More specifically, inspection of the 3D structure of FIXa harboring Glu270Lys shows that the conformational change observed at Lys247 results from the formation of a stable hydrogen bond between the positively charged ε-amino group of Lys247 and the oxygen from the peptide bond of Cys268. Analysis of the distance between the nitrogen atom of the ε-amino group of Lys247 and the oxygen atom of the Cys268 confirms that these residues are closer together in the FIXa Glu270Lys protein than in the other three 3D structures during the last 60 ns of the simulation. This difference is particularly markedin the last 20 ns of simulation, as shown in Supplementary Table S6. Given this result, we then attempted to clarify how Lys247 (from the 39-loop) is affected by residues 270 and 271 of the FIXa protein.

Closer inspection of the spatial arrangement of the moiety surrounding residue Cys268 clearly indicates that a bulky and charged amino acid such as a lysine requires free access in order to bend and become hydrogen bonded to the oxygen of Cys268. In the Glu270Lys + Thr271Pro and Thr271Pro mutant proteins, the protruding side chain of Pro271 appears to prevent access of the lysine to the Cys268 oxygen (Fig. 7a). In the case of the wild-type and Glu270Lys proteins, the 271 residue is the same but, the proximity of a negatively charged Glu or a positively charged Lys affects its RMSF and hence its conformation. The RMSF of Thr271 is 2.65 Å for the wild-type and 2.02 Å for the Glu270Lys protein. The higher RMSF for the wild-type protein is consequent to the hydrogen bond interactions that the hydroxyl group of Thr271 may establish with Glu270 and which are absent in the Glu270Lys mutant.

In order to evaluate how the absence of a nearby glutamate residue influences the conformational freedom of Thr271 in the wild-type and Glu270Lys mutant, the potential

hydrogen bonds established between the Thr271 hydroxyl oxygen (271@OG1) and the neighboring amino acid residues were identified and characterized. Two hydrogen bonds were identified (represented by green dotted lines in Fig. 7b), one of them between the Thr271@OG1 and the nitrogen from the peptide bond of Gly272 (272@N); the other between the Thr271@OG1 and the nitrogen from the peptide bond of Val273 (273@N). According to the data in Supplementary Table S6, both Thr271@OG1–272@N and Thr271@OG1–273@N distances and standard deviations are greater for the wild-type protein suggesting that in the FIXa Glu270Lys protein the side chain of Thr271 is stabilized in a conformation that allows Lys247 to interact with Cys268. The Thr271 from the wild-type protein has greater conformational freedom which precludes access of Lys247 to the Cys268 oxygen.

Functional implications of replacements at residues 270 and 271 of the FIX protein

At a functional level, it is well established that the 39-loop has a vital role in all serine proteases of coagulation (Yang and Rezaie 2013), accounting for the restriction of substrate and inhibitor specificity. We therefore examined whether the human Glu270Lys deleterious variant might interfere with these functions. The alignment of the representative structure from the dominant cluster in the Glu270Lys simulation with the high-resolution structure of the Michaelis complex between pentasaccharide-activated antithrombin (AT) and human FIXa (PDB ID 3KCG) (Johnson et al. 2010) revealed that the conformational change of the 39-loop in the Glu270Lys mutant protein is likely to impair its interaction with the reactive center loop (RCL) from AT (Fig. 8). Since AT is a suicide-substrate inhibitor of FIXa (Law et al. 2006), its binding mode is expected to be similar to the binding mode of the natural substrate of FIXa - coagulation factor X. The alignment of catalytic domains from the compensated protein and from the 3KCG crystallographic structure showed that the presence of a proline at position 271 serves to reverse the conformational change observed in the Glu270Lys in the 39-loop, thereby strengthening the evidence supporting a structural compensatory effect for Pro271 when *in cis* with Lys270. The FIX protein with compensatory amino acid Pro271 on a wild-type background behaved similarly to the wild-type protein with Thr271, and it may therefore be assumed that Pro271 *per se* should not significantly affect FIXa activity.

Evolutionary history of the FIX Glu270Lys CPD and its compensatory partner

Having obtained strong evidence through MD simulations that positions 270 and 271 harbor respectively a CPD and its *cis*-compensatory partner, we sought to complete our analysis by integrating an evolutionary perspective into the alternative combinations involving the Glu270Lys-Thr271Pro pair during mammalian evolution. To this end, we constructed a phylogenetic tree of FIXa with sequences encompassing our amino acid residues of interest at positions 270 and 271.

The ancestral Pro271 appears to have mutated to a threonine in the primate lineage that preceded the divergence of Old World and New World monkeys approximately 43 MYA (Fig. 9). Mammalian FIX homologs harboring Pro271 have, over evolutionary time, also accepted residues with very different biochemical properties at position 270 such as Lys and His (basic), Glu (acidic), Val (nonpolar), and Gln and Asn (polar). However, once threonine became fixed at position 271, only two different residues with similar biochemical properties (both acidic) were subsequently found at position 270, namely aspartate and glutamate. This concurs with the considerable flexibility conferred by Pro271 in terms of its toleration of an evolutionary replacement at position 270. It should also be noted that residue Lys270 emerged recurrently (on at least four occasions) during mammalian evolution, but always on a background where residue Pro271 is present, suggesting that the two residues have indeed been subject to compensatory interaction.

In summary, we have provided evidence showing that is possible to identify CPDs and their compensatory partners by means of a combination of fine-scale computational and physicochemical methods. In the case of the CPD/compensatory pair identified, characterized and validated in this study, the lack of toleration of a FIX Glu270Lys substitution in humans (as evidenced by the hemophilia B phenotype) appears to be due to the replacement of the ancestral Pro271 by a threonine prior to the radiation of the primate lineage. Such epistatic interactions between amino acid residues are likely to have shaped the acceptability of certain amino acid residues on an evolutionary timescale. One consequence of this is that novel variants identified in human disease-associated proteins will sometimes be located at less evolutionarily conserved sites and may, therefore, escape detection by conventional methods of pathogenicity assessment that rely heavily upon measures of evolutionary conservation (Azevedo et al. 2016). In other words, a low level of evolutionary conservation at a particular amino acid residue in a given protein should not automatically be interpreted as a predictor of negligible pathogenicity when

that residue is substituted. To circumvent this issue, variant prioritization should take into account the epistatic interactions between amino acid residues (e.g. Kim et al. 2019), in order to potentiate the identification of pathogenic mutations that occur at less well conserved amino acid positions.

DECLARATIONS

FUNDING

This work was supported by Fundo Europeu de Desenvolvimento Regional (FEDER) through the COMPETE 2020 - Operacional Programme for Competitiveness and Internationalization, Portugal 2020 and by Foundation for Science and Technology (FCT) [POCI-01-0145-FEDER-007274, POCI-01-0145-FEDER-29723]; Foundation for Science and Technology [UIDB/04423/2020, UIDP/04423/2020, UIDP/04378/2020, UIDB/04378/2020, SFRH/BD/137925/2018 to C.S.]; Infraestrutura Nacional de Computação Distribuída (INCD) [01/SAICT/2016 number 022153, CPCA/A00/7140/2020, CPCA/A00/7145/2020] through funds from Foundation for Science and Technology and Fundo Europeu de Desenvolvimento Regional; Foundation for Science and Technology [under the transitional rule of Decree Law 57/2016, amended by Law 57/2017 to J.C.]; Foundation for Science and Technology [2020.01423.CEECIND to S.S.].

CONFLICT OF INTEREST STATEMENT

The authors are unaware of any conflict of interest.

AVAILABILITY OF DATA

Sequences used in this study were retrieved from the Ensembl Genome Browser (https://www.ensembl.org) and are shown in Supplementary Table S1. Human disease-causing missense variants were retrieved from HGMD (http://www.hgmd.cf.ac.uk). The crystal structure of FIXa (6MV4) and GLA (3S5Z) were retrieved from the Protein Data Bank (PDB) (https://www.rcsb.org).

CODE AVAILABILITY

Not applicable

REFERENCES

Azevedo L, Carneiro J, van Asch B, Moleirinho A, Pereira F, Amorim A (2009) Epistatic interactions modulate the evolution of mammalian mitochondrial respiratory complex components. BMC Genomics 10:266. https://doi:10.1186/1471-2164-10-266

Azevedo L, Mort M, Costa AC, Silva RM, Quelhas D, Amorim A, Cooper DN (2016) Improving the *in silico* assessment of pathogenicity for compensated variants. Eur. J. Hum. Genet. 25:2-7. https://doi:10.1038/ejhg.2016.129

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res. 28:235-242. https://doi:10.1093/nar/28.1.235

Castellana S, Biagini T, Petrizzelli F, Parca L, Panzironi N, Caputo V, Vescovi AL, Carella M, Mazza T (2021) MitImpact 3: modeling the residue interaction network of the Respiratory Chain subunits. Nucleic Acids Res. 49:D1282-D1288. https://doi:10.1093/nar/gkaa1032

Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69-87. https://doi:10.1038/nature04072

Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-562. https://doi:10.1038/nature01262

Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J, Billis K, Boddu S, Cummins C, Davidson C, Dodiya KJ, Gall A, Giron CG, Gil L, Grego T, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, Kay M, Laird MR, Lavidas I, Liu Z, Loveland JE, Marugan JC, Maurel T, McMahon AC, Moore B, Morales J, Mudge JM, Nuhn M, Ogeh D, Parker A, Parton A, Patricio M, Abdul Salam AI, Schmitt BM, Schuilenburg H, Sheppard D, Sparrow H, Stapleton E, Szuba M, Taylor K, Threadgold G, Thormann A, Vullo A, Walts B, Winterbottom A, Zadissa A, Chakiachvili M, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Yates AD, Zerbino DR, Flicek P (2019) Ensembl 2019. Nucleic Acids Res. 47:D745-D751. https://doi:10.1093/nar/gky1113

de Beer TAP, Laskowski RA, Parks SL, Sipos B, Goldman N, Thornton JM (2013) Amino acid changes in disease-associated variants differ radically from variants

observed in the 1000 Genomes Project dataset. PLoS Comput. Biol. 9:e1003382. https://doi:10.1371/journal.pcbi.1003382
Ferreira P, Sousa SF, Fernandes PA, Ramos MJ (2017) Improving the catalytic power of the DszD enzyme for the biodesulfurization of crude oil and derivatives. Chemistry 23:17231-17241. https://doi:10.1002/chem.201704057
Ferrer-Costa C, Orozco M, de la Cruz X (2007) Characterization of compensated mutations in terms of structural and physico-chemical properties. J. Mol. Biol. 365:249-256. https://doi:10.1016/j.jmb.2006.09.053
Grant, B.J., Rodrigues, A.P., ElSawy, K.M., McCammon, J.A., Caves, L.S. (2006) Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics, 22, 2695–2696. https://doi.org/10.1093/bioinformatics/btl461
Genin E, Feingold J, Clerget-Darpoux F (2008) Identifying modifier genes of monogenic disease: strategies and difficulties. Hum. Genet. 124:357-368. https://doi:10.1007/s00439-008-0560-2
Guce AI, Clark NE, Rogich JJ, Garman SC (2011) The molecular basis of pharmacological chaperoning in human alpha-galactosidase. Chem. Biol. 18:1521-1526. https://doi:10.1016/j.chembiol.2011.10.012
Johnson DJ, Langdown J, Huntington JA (2010) Molecular basis of factor IXa recognition by heparin-activated antithrombin revealed by a 1.7-Å structure of the ternary complex. Proc. Natl. Acad. Sci. USA 107:645-650. https://doi:10.1073/pnas.0910144107
Jordan DM, Frangakis SG, Golzio C, Cassa CA, Kurtzberg J, Task Force for Neonatal G, Davis EE, Sunyaev SR, Katsanis N (2015) Identification of *cis*-suppression of human disease mutations by comparative genomics. Nature 524:225-229. https://doi:10.1038/nature14497
Kim D, Han SK, Lee K, Kim I, Kong J, Kim S (2019) Evolutionary coupling analysis identifies the impact of disease-associated variants at less-conserved sites. Nucleic Acids Res. 47:e94-e94. https://doi:10.1093/nar/gkz536
Kondrashov AS, Sunyaev S, Kondrashov FA (2002) Dobzhansky-Muller incompatibilities in protein evolution. Proc. Natl. Acad. Sci. USA 99:14878-14883. https://doi:10.1073/pnas.232565499
Kumar S, Stecher G, Suleski M, Hedges SB (2017) TimeTree: a resource for timelines, timetrees, and divergence times. Mol. Biol. Evol. 34:1812-1819. https://doi:10.1093/molbev/msx116
Kumar S, Suleski MP, Markov GJ, Lawrence S, Marco A, Filipski AJ (2009) Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations. Genome Res. 19:1562-1569. https://doi:10.1101/gr.091991.109
Law RH, Zhang Q, McGowan S, Buckle AM, Silverman GA, Wong W, Rosado CJ, Langendorf CG, Pike RN, Bird PI, Whisstock JC (2006) An overview of the serpin superfamily. Genome Biol. 7:216. doi: 10.1186/gb-2006-7-5-216
Loytynoja A, Goldman N (2010) webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. BMC Bioinformatics 11:579. https://doi:10.1186/1471-2105-11-579
Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C (2015) ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. J. Chem. Theory Comput. 11:3696-3713. https://doi:10.1021/acs.jctc.5b00255
Marin O, Aguirre J, de la Cruz X (2019) Compensated pathogenic variants in coagulation factors VIII and IX present complex mapping between molecular

impact and hemophilia severity. Sci. Rep. 9:9538. https://doi:10.1038/s41598-019-45916-3
Miller MP, Kumar S (2001) Understanding human disease mutations through the use of interspecific genetic variation. Hum. Mol. Genet. 10:2319-2328. https://doi:10.1093/hmg/10.21.2319
O'Rourke KF, Gorman SD, Boehr DD (2016) Biophysical and computational methods to analyze amino acid interaction networks in proteins. Comput. Struct. Biotechnol. J. 14:245-251. https://doi:10.1016/j.csbj.2016.06.002
Olsson MH, Sondergaard CR, Rostkowski M, Jensen JH (2011) PROPKA3: consistent treatment of internal and surface residues in empirical pKa predictions. J. Chem. Theory Comput. 7:525-537. https://doi:10.1021/ct100578z
Rhesus Macaque Genome Sequencing and Analysis Consortium, Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, Batzer MA, Bustamante CD, Eichler EE, Hahn MW, Hardison RC, Makova KD, Miller W, Milosavljevic A, Palermo RE, Siepel A, Sikela JM, Attaway T, Bell S, Bernard KE, Buhay CJ, Chandrabose MN, Dao M, Davis C, Delehaunty KD, Ding Y, Dinh HH, Dugan-Rocha S, Fulton LA, Gabisi RA, Garner TT, Godfrey J, Hawes AC, Hernandez J, Hines S, Holder M, Hume J, Jhangiani SN, Joshi V, Khan ZM, Kirkness EF, Cree A, Fowler RG, Lee S, Lewis LR, Li Z, Liu YS, Moore SM, Muzny D, Nazareth LV, Ngo DN, Okwuonu GO, Pai G, Parker D, Paul HA, Pfannkoch C, Pohl CS, Rogers YH, Ruiz SJ, Sabo A, Santibanez J, Schneider BW, Smith SM, Sodergren E, Svatek AF, Utterback TR, Vattathil S, Warren W, White CS, Chinwalla AT, Feng Y, Halpern AL, Hillier LW, Huang X, Minx P, Nelson JO, Pepin KH, Qin X, Sutton GG, Venter E, Walenz BP, Wallis JW, Worley KC, Yang SP, Jones SM, Marra MA, Rocchi M, Schein JE, Baertsch R, Clarke L, Csuros M, Glasscock J, Harris RA, Havlak P, et al. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222-234. https://doi:10.1126/science.1139247
Shapovalov MV, Dunbrack RL, Jr. (2011) A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure 19:844-858. https://doi:10.1016/j.str.2011.03.019
Stenson PD, Mort M, Ball EV, Chapman M, Evans K, Azevedo L, Hayden M, Heywood S, Millar DS, Phillips AD, Cooper DN (2020) The Human Gene Mutation Database (HGMD((R))): optimizing its use in a clinical diagnostic or research setting. Hum. Genet. 139:1197-1207. https://doi:10.1007/s00439-020-02199-3
Subramanian S, Kumar S (2006) Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. BMC Genomics 7:306. https://doi:10.1186/1471-2164-7-306
Suriano G, Azevedo L, Novais M, Boscolo B, Seruca R, Amorim A, Ghibaudi EM (2007) *In vitro* demonstration of intra-locus compensation using the ornithine transcarbamylase protein as model. Hum. Mol. Genet. 16:2209-2214. https://doi:10.1093/hmg/ddm172
Vadivel K, Schreuder HA, Liesum A, Schmidt AE, Goldsmith G, Bajaj SP (2019) Sodium-site in serine protease domain of human coagulation factor IXa: evidence from the crystal structure and molecular dynamics simulations study. J. Thromb. Haemost. 17:574-584. https://doi:10.1111/jth.14401
van Leeuwen J, Pons C, Mellor JC, Yamaguchi TN, Friesen H, Koschwanez J, Usaj MM, Pechlaner M, Takar M, Usaj M, VanderSluis B, Andrusiak K, Bansal P,

Baryshnikova A, Boone CE, Cao J, Cote A, Gebbia M, Horecka G, Horecka I, Kuzmin E, Legro N, Liang W, van Lieshout N, McNee M, San Luis BJ, Shaeri F, Shuteriqi E, Sun S, Yang L, Youn JY, Yuen M, Costanzo M, Gingras AC, Aloy P, Oostenbrink C, Murray A, Graham TR, Myers CL, Andrews BJ, Roth FP, Boone C (2016) Exploring genetic suppression interactions on a global scale. Science 354: aag0839. https://doi:10.1126/science.aag0839

Wacey AI, Krawczak M, Kakkar VV, Cooper DN (1994) Determinants of the factor IX mutational spectrum in haemophilia B: an analysis of missense mutations using a multi-domain molecular model of the activated protein. Hum. Genet. 94:594-608. https://doi:10.1007/BF00206951

Xue Y, Prado-Martinez J, Sudmant PH, Narasimhan V, Ayub Q, Szpak M, Frandsen P, Chen Y, Yngvadottir B, Cooper DN, de Manuel M, Hernandez-Rodriguez J, Lobon I, Siegismund HR, Pagani L, Quail MA, Hvilsom C, Mudakikwa A, Eichler EE, Cranfield MR, Marques-Bonet T, Tyler-Smith C, Scally A (2015) Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. Science 348:242-245. https://doi:10.1126/science.aaa3952

Yang L, Rezaie AR (2013) Residues of the 39-loop restrict the plasma inhibitor specificity of factor IXa. J. Biol. Chem. 288:12692-12698. https://doi:10.1074/jbc.M113.459347

Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, Boddu S, Marugan JC, Cummins C, Davidson C, Dodiya K, Fatima R, Gall A, Giron CG, Gil L, Grego T, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, Kay M, Lavidas I, Le T, Lemos D, Martinez JG, Maurel T, McDowall M, McMahon A, Mohanan S, Moore B, Nuhn M, Oheh DN, Parker A, Parton A, Patricio M, Sakthivel MP, Abdul Salam AI, Schmitt BM, Schuilenburg H, Sheppard D, Sycheva M, Szuba M, Taylor K, Thormann A, Threadgold G, Vullo A, Walts B, Winterbottom A, Zadissa A, Chakiachvili M, Flint B, Frankish A, Hunt SE, G II, Kostadima M, Langridge N, Loveland JE, Martin FJ, Morales J, Mudge JM, Muffato M, Perry E, Ruffier M, Trevanion SJ, Cunningham F, Howe KL, Zerbino DR, Flicek P (2020) Ensembl 2020. Nucleic Acids Res. 48:D682-D688. https://doi:10.1093/nar/gkz966

Zhang G, Pei Z, Ball EV, Mort M, Kehrer-Sawatzki H, Cooper DN (2011) Cross-comparison of the genome sequences from human, chimpanzee, Neanderthal and a Denisovan hominin identifies novel potentially compensated mutations. Hum. Genomics 5:453-484. https://doi:10.1186/1479-7364-5-5-453

Zhang G, Pei Z, Krawczak M, Ball EV, Mort M, Kehrer-Sawatzki H, Cooper DN (2010) Triangulation of the human, chimpanzee, and Neanderthal genome sequences identifies potentially compensated mutations. Hum. Mutat. 31:1286-1293. https://doi:10.1002/humu.21389

TABLES

Table 1. Protein models with FIXa variants simulated through molecular dynamics.

Protein	Position 270	Position 271
Wild-type	Glutamate	Threonine
Glu270Lys Mutant	Lysine	Threonine
Glu270Lys + Thr271Pro Mutant	Lysine	Proline
Thr271Pro Mutant	Glutamate	Proline

Table 2. Summary of the cluster analysis with the estimated proportion of time spent in each cluster during the last 60 ns of MD simulation.

	Wild-type	**Glu270Lys**	**Glu270Lys + Thr271Pro**	**Thr271Pro**
Dominant Cluster (%)	72.6	98.5	71.4	74.7
Alternative Cluster (%)	27.4	1.5	28.6	25.3

FIGURE LEGENDS

Fig. 1 Workflow of the strategy used in this work.

Fig. 2 a CPDs detected through the comparison of FIX, GLA, HPRT1, NDP and OTC in different mammalian species. Orange spheres represent the human deleterious missense variants that correspond to the wild-type amino acid residues in non-human mammalian species. Numbers below the orange spheres indicate the number of species in which the CPD was found. **b** Proportion of CPDs found in each protein.

Fig. 3 a Graphical representation of the RMSd of the alpha carbon atoms from the catalytic domain residues from the wild-type, Glu270Lys, Glu270Lys + Thr271Pro and Thr270Pro FIXa proteins along the 100 ns of MD simulation. **b** 2D RMSd plots for the MD simulations of the different FIXa variants. The simulation time is plotted on both the X and Y axes with the RMSd values represented as a gradation in color.

Fig. 4 DCCM analyses of Cα atoms of all amino acid residues from the FIXa catalytic domain.

Fig. 5 a Graphical representation of the RMSF of the Cα atom of each amino acid residue in the catalytic domain of wild-type, Glu270Lys, Glu270Lys + Thr271Pro and Thr270Pro proteins along the last 60 ns of MD simulation. **b** Graphical representation of the RMSF of all atoms from the residues of the active site, 39-loop, 60-loop and 375-393 of wild-type, Glu270Lys, Glu270Lys + Thr271Pro and Thr270Pro proteins along the last 60 ns of MD simulations.

Fig. 6 Visual Molecular Dynamics (VMD) representation of the alignment of the representative structure from the dominant cluster from each FIXa protein. The residues with the higher RMSF variation are represented in licorice.

Fig. 7 VMD representation of **a** the spatial arrangement of residues Glu/Lys270, Thr/Pro271, Cys268 and Lys247 in the four FIXa proteins. **b** Hydrogen bonds established by the 271@OG1 atom from Thr271 in the wild-type and Glu270Lys mutant proteins.

Fig. 8 VMD representation of the alignment between the catalytic domains of the representative structures of the dominant clusters taken from the MD simulations studies on the different FIXa protein variants and the catalytic domain from the crystallographic structure.

Fig. 9 Species tree of major mammalian lineages demonstrating the evolution of the residues homologous to the amino acid positions 270 and 271 of human FIX, constructed in TimeTree (Kumar et al. 2017). Values at tree nodes indicate the approximate time (in MYA) of divergence of each lineage.