Application of a new in silico strategy to evidencing the role of missense mutations properties in determining Hemophilia A

Mariana Meireles[1], Lara Stelmach[2], Eliane Bandinelli[2], and Gustavo Fioravanti Vieira[3]

[1]Universidade Federal do Rio Grande do Sul Instituto de Biociencias
[2]Universidade Federal do Rio Grande do Sul
[3]Centro Universitario La Salle

November 7, 2020

Abstract

Hemophilia A (HA) consists of a genetic X-linked blood disorder. It is caused by a diversity of F8 gene mutations, with missense type being the most prevalent. Amino acid substitutions may impact physicochemical properties of the protein, providing an abundant scenario for investigation. This work evaluates 71 substitutions contributing to distinct patients' phenotypes (mild, moderate, and severe), in terms of physicochemical alterations (PA - electrostatic potential, hydrophobicity, surface solvent-accessible/excluded areas, disulfide disruptions, and substitutions indexes), through an in silico strategy. PA information extracted from models fuels a hierarchical clustering analysis (HCA - independently and in combination) in an attempt to connect mutations and patients' phenotypes. The combined use of PA over the analysis of single features seems to better reflect the impact of substitutions in severity degree, apparently in a domain-dependent way. Besides, a principal component analysis (PCA) identified prominent properties impacting clustering results for each domain. Electrostatic potential has a greater contribution to A3 than C1 domain clustering, probably by A3 involvement in FVIII activation, for example. The conjugated use of HCA and PCA is a powerful tool to assess if and what kind of structural features are involved in FVIII protein functionality impact and HA disease severity.

Introduction

The hemophilia A (HA) condition is the most prevalent bleeding X-linked disorder and affects about 1: 5.000 male live births (Hoyer 1994). The condition expresses due to mutations in the F8 gene that encodes the coagulation Factor VIII (FVIII), a crucial protein in the coagulation process (Gitschier et al. 1984). The mature factor VIII protein is composed of homologous internal regions classified into three type domains: A, B, and C (Eaton and Vehar 1986, Vehar et al. 1984). The proteolytic cleavage by thrombin in several Arginine residues leads to the removal of B domain and plasmatic secretion (Lenting, van Mourik and Mertens 1998, Ngo et al. 2008). According to the International Society of Thrombosis and Haemostasis, the HA is classified based on the plasmatic level of FVIII as mild (5-40%), moderate (1-5%), and severe (<1%)(White et al. 2001).

The disease manifestations involve a broad clinical heterogeneity with a range of mutations considered as disorder cause. According to the European Association of Hemophilia and Allied Disorders, 66,2% of F8 variations are point mutations, with missense ones the most prevalent of them (Venkateswarlu 2014). Amino acids substitutions consist of a challenge in terms of impact evaluation, once these mutations have a more distributed spectrum that includes since beneficial to lethal variants. There are a lot of aspects involving amino acid substitution that could impact the FVIII functionality, and consequently, disease severity and its classification(into mild, moderate, and severe): amino acid characteristics themselves, their location, bond disruptions, and modifications that affect protein conformation and affinity with coagulation cascade partners.

A recent study by our group shows a singular investigation approach to associate hemophilia B disorder causality using amino acids and protein chain properties ((Meireles et al. 2019). That pipeline takes into account some physicochemical features from the Factor IX molecular 3D structure.

To a better HA disease comprehension and the cited application improvement, the present study makes an effort to apply and refine the method, evaluating how the analyzed aspects contribute to genotype-phenotype correlation in HA disease. In this context, physicochemical properties (electrostatic patterns (EP), hydrophilic/hydrophobic changes, hydrogen and disulfide bond alterations, solvent surface accessibility, and combined indexes of substitutions impact) were verified and evaluated. A combination of these elements was analyzed to respond to the following queries: (a) How the missense substitutions affect the protein properties? (b) Is this new pipeline approach useful to a better understanding of the physicochemical features involved in the disorder causality and severity?

Methods:

To a better comprehension of the pipeline, **Figure 1** summarizes it in a scheme illustrating the applied methodological process. For all the effects, the F8 sequence accession number used for the genome is NG_-011403.1 (F8) obtained on NCBI, as the transcript NM_000132.4 (F8). The protein annotated and reviewed accession code at UniProt knowledgebase (Swiss-Prot) is P00451. All the variants nomenclatures were verified using the Name checker service from Mutalyzer (https://mutalyzer.nl/name-checker). The verification results are available at Supplemental material.

Mutations choice:

The amino acid substitutions were chosen based on previous studies, taking into account mutations evidenced in our country, by our research group (Gorziza et al. 2013, Rosset et al. 2014) and also other references sources, including the European Association for Haemophilia and Allied Disorders (EAHAD) Factor VIII mutation database (accessible at f8-db.eahad.org). It leads to a total of 71 different mutations selected from the five mature peptide domains (A1, A2, A3, C1, and C2). Information concerning these substitutions, including their domain location and references, are available in **Table 1** . Aiming to distinguish the effect of amino acid position where the mutation appears and the residue substituted, we select many mutations occurring at the same protein position. Regarding the selected mutations, 27 are found in patients with mild, ten with moderate, and 34 with severe HA phenotype.

Modeling approach

In intent to verify the impact caused by missense mutations in the structural protein level, we proceed with an analysis based on generated tridimensional models. These models were examined and have their physicochemical properties assessed.

Generating models:

The FASTA protein sequence from human FVIII was recovered from the annotated and reviewed UniProt code P00451, and for each mutation, it was manually edited. The obtained sequences for all substitutions, and also the wild type (WT) sequence, were applied in a modeling approach using Phyre-2 software (**P** rotein **H** omology/analogy**R** ecognition **E** ngine V 2.0-(Kelley et al. 2015)). This tool used through its expert "one-to-one threading job" mode, enables the use of the sequences obtained combined with the Protein Data Bank (Rose et al. 2016) templates (PDB id: 2R7E) to produce structural models regarding the mutated and wild types.

The generated models comprise the domains of circulating activated FVIII (A1, A2, A3, C1, and C2), which are included in the crystal structure. After all, the analysis comprises a total of 76 structures: 17 models for A1 (16 mutated and the wild type (WT)), 18 for A2 (17 mutated and the wild one), 25 for A3 (24 mutated), seven for C1 (six mutated), and nine for C2 (eight mutated). All these chains models were submitted for quality evaluation using ModFold (Maghrabi and McGuffin 2017, McGuffin, Buenavista and Roche 2013)and PDBsum (Laskowski et al. 1997).

Physicochemical alterations:

To investigate alterations caused by amino acid substitutions, the wild and the mutated models were applied in the Delphi web server tool (Sarkar et al. 2013, Smith et al. 2012). This tool permits the electrostatic potential (EP) files calculations and turns possible a visual inspection of EP distribution through its input in the Chimera interface (Pettersen et al. 2004). Improving the EP evaluation, we also perform a clusterization exclusively based on that property, using the webPIPSA software (Protein Interaction Property Similarity Analysis)(Richter et al. 2008). This analysis considered the domains independently, with their corresponding mutated and wild structures, obtaining five clustering schemes (A1, A2, A3, C1, C2). The clusters group models based on the EP similarity distance measures. The use of both procedures allow a comprehensive relation of EP modifications, with a complete analysis that identifies not only the electrostatic distance between structures (clusterization in webPIPSA) but also the detection of charge surface alterations of each model (Delphi calculations applied in Chimera interface).

The Chimera interface resources also contribute to assessing other features: hydrophobicity changes, hydrogen bonds and cysteine bonds disruptions, solvent-accessible surface area (areaSAS), and solvent-excluded surface area (areaSES) modifications. The values of each attribute were extracted from each model individually and compared to the wild one.

The following scores for physicochemical distances concerning amino acid replacements were calculated by the following matrices: Grantham's distance (Grantham 1974), Sneath´s index (Sneath 1966), Epstein´s coefficient of distance (Epstein 1967), Miyata´s distance (Miyata, Miyazawa and Yasunaga 1979), and Experimental Exchangeability (Yampolsky and Stoltzfus 2005).

All the properties cited above were applied in a combined manner as an input for a Hierarchical Clustering Analysis (HCA), objecting to improve the formed EP clusters of each analyzed domains. The HCA analysis became viable through the R Studio platform, using the pvclust package(Suzuki and Shimodaira 2006).The output of this program returns clusters and branches with associated probability values using bootstrap resampling techniques: approximately unbiased (AU), obtained with multiscale bootstrap resampling, and bootstrap probability (BP), calculated by the ordinary bootstrap resampling technique. Generally, the AU p-value is a better estimator of the reliability of the clusters obtained and was the selected statistical metric in our work.

Evaluating the principal components involved in mutations variance

All the considered features have their role in the substitutions effects, but how each one of them contributes to the final clustering? To solved this puzzle, we performed a Multivariate analysis using the FactoMineR package (FactoShiny) in the R cross-platform (Lê, Josse and Husson 2008). This package allowed a Principal Components Analysis (PCA), which indicated the most contributive variable in a complex scenario, as the variance presented by the mutations properties spectrum. Principal components correspond to the directions where there is the more variance, the orientation where the data is more spread out. The substitutions were considered for each evaluated domain in separate, reflecting that the domain functions are diverse, and their roles contributes to determines the impact degree of a physicochemical change. We also considered all mutations together, aiming to identify the principal components (PC) of entire data and comparing its findings to the specific domains PCs.

Results and discussion

The electrostatic potential influence in the disorder determination

Electrostatic potential (EP) consists of a fundamental property that can be used to infer interaction with other coagulation cascade partners, as the FVIII link with Factor IX generating the tenase complex that leads to Factor X activation, an essential step in the coagulation process. Changes in surface EP caused by missense mutations put their respective models away from normal phenotype. The epograms, **(Figures 2, S1, S2, S3, S4)** obtained based on EP similarities for each FVIII domain analyzed, illustrates the distances between clusters (grouping models with similar EP patterns). The visual inspection of EP calculations (using

Delphi web server) applied at these models confirmed that (It occurs especially for HCAs where the members present more distant EP scales, as in **Figure S3** and **S4** .). If taken into account the total of substitutions included in this research, about 69 percent of them (49/71) fall separated from the cluster containing WS (wild structure) based on electrostatic similarities, indicating the crucial role of EP in protein alterations. The EP influences about mutations vary according to the domain and substitutions.

Considering the A1 domain epogram and heatmap (**Figure 2 and S5**), five of 16 mutated structures are electrostatically similar to WS. They represent different patient phenotypes: two mild (p.Val64Met, p.Val285Gly), one moderate (p.Cys267Tyr), and two severe (p.Phe70Ser, p.Cys348Tyr). At the A2 domain, 11/17 structures (corresponding to multiple phenotypes) grouped with the wild one with identical EP patterns, except for p.Asp544Glu, showing a small distance from WS (which indicates a little alteration in the EP distribution) that was not enough to separate them, as can be observed at **Figure S1** . On the other hand, every structure of the A3 domain showed some differences that incurred in distances from WS, even those grouped with it (**Figure S2)** . The same occurs to the C2, having only the p.Pro2224Arg (severe phenotype) clustered with WS but also presenting EP divergences between the structures (**Figure S4**). The C1 domain has only the p.Cys2188Tyr (severe) in the WS cluster, which even with the critical phenotype, mutated structure have no EP change in comparison to wild model (**Figure S3**).

While we understand the importance of the EP in protein interaction, evidenced in some of the above-described clusters, many members clustered altogether with structures presenting discordant phenotypes. In some cases, structures from mutations involved in severe disease grouped with wild type structure. To disentangle this intricate network of proteins relations, other amino acid properties aspects should be considered for a better understanding of HA and other disorders.

Amino acid substitutions alter important properties to protein function

Trying to identify another amino acid feature that influences severity determination, the hydrophobicity values were calculated for each one of the mutated and WS models (using Chimera interface resources). The residues affinity or repulse to water consist of a molecular characteristic capable of modifies the residues solvent exposure, leading to protein structural changes and affecting its function and dynamics.

Except for p.Asp544Glu (in the A2 domain), every studied substitution presents modifications in hydrophobicity when compared to the wild structure (WS), as can be observed in **Figure 3** (the A1 domain) and **Supplementary Figures 6,7,8,9** (A2, A3, C1, and C2, respectively). Some alterations are more evident due to its capability to shift the local hydrophobicity from negative (more hydrophilic residues) to positive or vice versa, as can be noticed in the values of p.Cys547 and p.Ile494 substitutions from A1 domain (**Figure 3**).Others are blander and thus hard to percept. Interestingly, even substitutions with no electrostatic modifications (in comparison to WS) presented alterations when considering hydrophobicity. These observations highlight this feature role in missense mutation effect determination, notwithstanding recognizing that the hydrophobic change dimension is not linearly related to phenotype.

Regarding the other aspects investigated, disulfide-bonds disruptions occur in 12 of the 71 substitutions of the present study. Considering the domains, the A1 shows two disulfide disruptions, corresponding to approximately 12,5 percent (2/16 – p.Cys267Tyr, p.Cys348Tyr) of domain mutations, A2 presents five (approx. 29,5%, 5/17 - p.Cys547Arg, p.Cys547Tyr, p.Cys547Trp, p.Cys649Arg, p.Cys649Ser). The C1 and C2 domains present respectively two (28,6%, 2/7- p.Cys2040Tyr, p.Cys2188Tyr) and three (37,5%, 3/8- p.Cys2193Gly, p.Cys2345Ser, p.Cys2345Tyr) cysteine-cysteine breakage. The A3 domain shows none alteration in this property. This molecular ligation disruption seems to be a good predictor of severity, since it is involved in domains maintenance, establishment, and sustention. No wonder, none of the mutations presenting this property change exhibit a mild phenotype

(Table 1).

The solvent-accessible surface area (areaSAS) and solvent excluded surface area (areaSES) properties are modified in the majority of the mutations when compared to wild structure, except for p.Ala2080Thr at the C1 domain. It occurs because of the high level of influences that interferes with molecular surface areas, turning the residues more or less buriedor apparent. The access to these characteristics using the Chimera interface returns unique values for every amino acid residue in each model, making it feasible the comparison among structures to verify how divergent they are concerning WS.

After individual evaluation of the cited features, the extracted values of differences between mutated and wild structures (for each considered characteristic) are compiled in the HCA analysis, considering that all properties investigated affect the FVIII functioning since alters solvent interaction and structural conformation. Considering the physicochemical aspects evidenced so far, every mutated model shows alterations in one or more than one of the cited attributes, evidencing the importance of a combined analysis.

Use of multiple amino acid replacement indexes improves the results

The replacement distance indexes consist of a mathematical calculation to estimate values for the difference between amino acids considering their particularities. This approach is very used in evolutionary genetics. Each index examines different residues properties, making its estimations values. It does not seem correct to consider any of them separately, which could increase the estimated weight bias of the physicochemical property effect at the expense of others. Resuming, the factors considered by each index are not identical (Dagan, Talmor and Graur 2002), so the combined use of them seems to be effective. Therefore, when individually observed and compared to patient phenotypes, these index values do not explain the severity of mutations described, nor singly or in combination (**Figure S10** shows an HCA for A1 domain mutations using only the substitutions matrices data). Despite that, the substitution matrices demonstrate to improve the refinement of the HCA approach, as an additional step.

The application of multiple physicochemical features in an HCA has no mutations grouped with WS type.

When applying the HCA methodology based on the combination of the different physical and chemical aspects above cited, it was possible to verify that the clusters have a quite improvement if compared to those based only on the EP pattern. As a result, none of the domains showed mutant structures grouped with the normal phenotype (WS) (**Figures 5, S11, S12, S13, and S14**).

Figure 5 shows the HCA for the A1 domain. Beyond the WS and mutated structures are correctly separated, we also can notice a greater homogeneity of members presenting the same phenotypes in each formed clusters. There is only one case grouping a model corresponding to a severe phenotype together with a mild one (p.Lys67Glu and p.Glu162Lys, respectively). Interestingly, this case involves substitutions concerning the same residues exchanging. The same advances in clusters resolution, separating similar phenotypes occur for the remaining domains (**Figures S11, S12, S13, and S14**). It brings a good sign being an indicative of the considered physicochemical features relevance as phenotypic determining aspects in HA (and probably other coagulation disorders).

A Principal component multivariate analysis contributes to elucidate the features role in specific domains.

Principal component analysis (PCA) is useful to clarify the directions of the data where is more variance. This linear transformation fits our information in a coordinate system, where related mutations contribute with similar properties values as a component. At the A1 domain, the first principal component contributes with 34,63 percent of the total variance, while the second one with 22,52 percent **(Figure 5A and Table2)** . Together the first two PCs explain 57%of the variance in a ten-dimensional data, indicating its essential role. To a better understanding of how the elements influence the determinations in coordinate and contributions terms, the five first dimensions results (explaining almost 91% of the variance) are in **Table 2** . (**Table S1** presents the coordinate and contributions for elements in domains A2, A3, C1 and C2).

In a two dimensional graphic (**Figure 5B**), the eigenvalues depiction reproduces the eigenvectors force, and the distribution of the mutations in this coordinate system tends to group in conformity with characterized elements. In the PCA graph of individuals, the eight more contributive mutations distribution is represented according to variables effects. The p.Cys267Tyr and p.Cys348Tyr mutations are highly influenced by the eigenvector "C-C disruption"; another group (p.Lys67Glu, p.Glu162Lys, and p.His228Arg) are positioned due to other features (Granthams, Exchangeability, areaSES); the remain three mutations (p.Leu117Arg, p.Leu217Arg, and p.Glu30Val) are characterized by high values for Epstein coefficient, hydrophobicity, and Miyata. It corroborates with the qualitative and quantitative observed hydrophobicity features for the mutations p.Leu117Arg, p.Leu217Arg, and p.Glu30Val, (**Figure 3**) which have a more prominent change than all others evaluated substitutions in the same domain.

Performing an HCPC (Hierarchical clustering for Principal Components) analysis with Euclidean metrics (p value of 0.05) and six clusters based on only the five first dimensions (**Figure 5C)** , we obtain a graphical clusterization with few differences concerning the previous HCA (**Figure 4**). The use of only five more contributive features did not improve the cluster's severity, pointing out to the importance of using a broad approach including the whole set of available variables**(Figure S11).** It is still more challenging considering that the five principal components differ according to domains, and the properties contributions to dimensions also differ. When analyzing the remaining FVIII domains, it is possible to notice distinct forces and contributions of components according to variables. It occurs due to the particular mutations profile and the specific domain's function that leads to differential influences in features. It tends to be more or less affected in a context dependent way (for example, substitutions considered for A3 domain leads to different coordinates variables distribution in comparison to C1 (**Figure S15B and S15C**). The A3 domain have an important function in FVIII activation (considering the proximity and participation at B domain cleavage) and stabilization by its linkage with Von Willebrand Factor (vWF) (Bloem et al. 2013). Thus, the detected electrostatic alterations seem to have a major influence in the first two Principal Components in A3 domain, due to the affinity requirement with the region and interactor above cited. The influence of the electrostatic pattern in the A3 domain contributes to explain the similar results found in electrostatic potential clustering**(Figure S2)** and HCA (**Figure S12**). On the other hand, hydrophobicity and disulfide-bond disruptions have a higher impact for the C1 domain, due to the inclusion of cysteine mutations in residues 2040 and 2188, both part of an intra-domain bond. As expected, the results of PCA considering all included mutations for every FVIII active domain (**Figure S16)** also differ from specific domains analysis.

Final conclusions

The physicochemical features changes generated by missense mutations influence the disease determination in a complex role. None of the included aspects could themselves, individually, explain the disorder severity. However, the combined use of this set of characteristics applied in a Hierarchical Clustering Analysis (HCA) proved to be an effective method to discriminate missense mutations according to severity since the amino acid substitutions have multiple interferences at the protein level. Every feature included has its contribution in the relative distance inference of mutations physicochemical profile concerning the wild type, indicating their importance to explain the current disorder. The applied approach takes into account essential physicochemical properties, which has the potential to be used in machine learning methods to predict how new mutations contribute to disrupt FVIII protein function, even known that they do not have a straight effect.

Hemophilia A has the status of a monogenic disorder with complete penetrance. The phenotypes classification adopted in this study reflects the plasmatic protein levels, and it is important to emphasize that the degree of available information varied according to accessed databank and references. Also important to notice, the clinical investigations reveal that habits and environment influence the bleeding tendency (the clinical phenotype, which not always reflect plasmatic factor levels) and inhibitors' development to treatment (factor reposition) (Peyvandi, Garagiola and Seregni 2013, Oldenburg et al. 2018). Our study takes into account the missense mutations genetic component and their protein effect behind the disorder, relating them to patients' phenotype in terms of plasmatic FVIII level.

The identification of which variables have their effect on the mutations impact brings an essential issue about predictions: they should consider domain specificities. As observed, a mutational landscape sometimes leads to a more variable severity according to domain functionality. Considering that, prediction method must contemplate the domains in separate and its influence in global protein analysis. Although the amino acid substitutions properties (and also their penalty score matrices) constitute a valid source of information and initial approach investigation, the impact caused by mutations is not fully estimated using only this information.

The implementation of all investigated features brings progress to the phenotype's severity understanding. Nonetheless, the PCA analysis helped to identify the most relevant variables influencing the clustering of chosen mutations in a FVIII domain specific. The variations spectrum tolerated by each protein region is unique, and consequently, the effect in disorder determination too.

Acknowledgments

Financing for this work was provided by Conselho Nacional de Desenvolvimento Científico e Tecnológico and Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (Apoio a Núcleos de Excelência Program). The authors declare that they do not have any conflict of interest.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

References

Bloem, E., H. Meems, M. van den Biggelaar, K. Mertens & A. B. Meijer (2013) A3 domain region 1803-1818 contributes to the stability of activated factor VIII and includes a binding site for activated factor IX. *J Biol Chem,* 288**,** 26105-11.

Dagan, T., Y. Talmor & D. Graur (2002) Ratios of radical to conservative amino acid replacement are affected by mutational and compositional factors and may not be indicative of positive Darwinian selection. *Mol Biol Evol,* 19**,** 1022-5.

Eaton, D. L. & G. A. Vehar (1986) Factor VIII structure and proteolytic processing. *Prog Hemost Thromb,* 8**,** 47-70.

Epstein, C. J. (1967) Non-randomness of amino-acid changes in the evolution of homologous proteins. *Nature,* 215**,** 355-9.

Gitschier, J., W. I. Wood, T. M. Goralka, K. L. Wion, E. Y. Chen, D. H. Eaton, G. A. Vehar, D. J. Capon & R. M. Lawn (1984) Characterization of the human factor VIII gene. *Nature,* 312**,** 326-30.

Gorziza, R. P., I. A. Vieira, D. B. Kappel, C. Rosset, M. Sinigaglia, L. B. Leiria, F. M. Salzano & E. Bandinelli (2013) Genetic changes in severe haemophilia A: new contribution to the aetiology of a complex disease. *Blood Coagul Fibrinolysis,* 24**,** 164-9.

Grantham, R. (1974) Amino acid difference formula to help explain protein evolution. *Science,* 185**,** 862-4.

Hoyer, L. W. (1994) Hemophilia A. *N Engl J Med,* 330**,**38-47.

Kelley, L. A., S. Mezulis, C. M. Yates, M. N. Wass & M. J. Sternberg (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc,* 10**,** 845-58.

Laskowski, R. A., E. G. Hutchinson, A. D. Michie, A. C. Wallace, M. L. Jones & J. M. Thornton (1997) PDBsum: a Web-based database of summaries and analyses of all PDB structures. *Trends Biochem Sci,*22**,** 488-90.

Lenting, P. J., J. A. van Mourik & K. Mertens (1998) The life cycle of coagulation factor VIII in view of its structure and function.*Blood,* 92**,** 3983-96.

Lê, S., J. Josse & F. Husson (2008) FactoMineR: An R Package for Multivariate Analysis. *2008,* 25**,** 18.

Maghrabi, A. H. A. & L. J. McGuffin (2017) ModFOLD6: an accurate web server for the global and local quality estimation of 3D protein models.*Nucleic Acids Res,* 45**,** W416-W421.

McGuffin, L. J., M. T. Buenavista & D. B. Roche (2013) The ModFOLD4 server for the quality assessment of 3D protein models. *Nucleic Acids Res,* 41**,** W368-72.

Meireles, M. R., M. A. S. Bragatte, E. Bandinelli, F. M. Salzano & G. F. Vieira (2019) A new in silico approach to investigate molecular aspects of factor IX missense causative mutations and their impact on the hemophilia B severity. *Hum Mutat,* 40**,** 706-715.

Miyata, T., S. Miyazawa & T. Yasunaga (1979) Two types of amino acid substitutions in protein evolution. *J Mol Evol,* 12**,**219-36.

Ngo, J. C., M. Huang, D. A. Roth, B. C. Furie & B. Furie (2008) Crystal structure of human factor VIII: implications for the formation of the factor IXa-factor VIIIa complex. *Structure,* 16**,** 597-606.

Oldenburg, J., G. Young, E. Santagostino & C. Escuriola Ettingshausen (2018) The importance of inhibitor eradication in clinically complicated hemophilia A patients. *Expert Rev Hematol,* 11**,** 857-862.

Pettersen, E. F., T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng & T. E. Ferrin (2004) UCSF Chimera–a visualization system for exploratory research and analysis. *J Comput Chem,* 25**,** 1605-12.

Peyvandi, F., I. Garagiola & S. Seregni (2013) Future of coagulation factor replacement therapy. *J Thromb Haemost,* 11 Suppl 1**,** 84-98.

Richter, S., A. Wenzel, M. Stein, R. R. Gabdoulline & R. C. Wade (2008) webPIPSA: a web server for the comparison of protein interaction properties. *Nucleic Acids Res,* 36**,** W276-80.

Rose, P. W., A. Prlić, A. Altunkaya, C. Bi, A. R. Bradley, C. H. Christie, L. D. Costanzo, J. M. Duarte, S. Dutta, Z. Feng, R. K. Green, D. S. Goodsell, B. Hudson, T. Kalro, R. Lowe, E. Peisach, C. Randle, A. S. Rose, C. Shao, Y. P. Tao, Y. Valasatava, M. Voigt, J. D. Westbrook, J. Woo, H. Yang, J. Y. Young, C. Zardecki, H. M. Berman & S. K. Burley (2016) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res* .

Rosset, C., R. P. Gorziza, M. R. Botton, F. M. Salzano & E. Bandinelli (2014) Factor VIII mutations and inhibitor formation in a southern Brazilian population. *Blood Coagul Fibrinolysis,* 25**,**125-7.

Sarkar, S., S. Witham, J. Zhang, M. Zhenirovskyy, W. Rocchia & E. Alexov (2013) DelPhi Web Server: A comprehensive online suite for electrostatic calculations of biological macromolecules and their complexes. *Commun Comput Phys,* 13**,** 269-284.

Smith, N., S. Witham, S. Sarkar, J. Zhang, L. Li, C. Li & E. Alexov (2012) DelPhi web server v2: incorporating atomic-style geometrical figures into the computational protocol. *Bioinformatics,*28**,** 1655-7.

Sneath, P. H. (1966) Relations between chemical structure and biological activity in peptides. *J Theor Biol,* 12**,** 157-95.

Suzuki, R. & H. Shimodaira (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics,*22**,** 1540-2.

Vehar, G. A., B. Keyt, D. Eaton, H. Rodriguez, D. P. O'Brien, F. Rotblat, H. Oppermann, R. Keck, W. I. Wood, R. N. Harkins, E. G. Tuddenham, R. M. Lawn & D. J. Capon (1984) Structure of human factor VIII. *Nature,* 312**,** 337-42.

Venkateswarlu, D. (2014) Structural insights into the interaction of blood coagulation co-factor VIIIa with factor IXa: a computational protein-protein docking and molecular dynamics refinement study.*Biochem Biophys Res Commun,* 452**,** 408-14.

White, G. C., F. Rosendaal, L. M. Aledort, J. M. Lusher, C. Rothschild, J. Ingerslev & F. V. a. F. I. Subcommittee (2001) Definitions in hemophilia. Recommendation of the scientific subcommittee on factor

VIII and factor IX of the scientific and standardization committee of the International Society on Thrombosis and Haemostasis. *Thromb Haemost,* 85**,** 560.

Yampolsky, L. Y. & A. Stoltzfus (2005) The exchangeability of amino acids in proteins. *Genetics,* 170**,** 1459-72.

Supplementary Figures Captions:

Figure S1. The Electrostatic similarities and patterns of structural models in the A2 domain. The clusters generation using webPIPSA tool groups models in terms of electrostatic potential (EP) similarities (left) based on the "skin" pattern. Next to the epogram, on the right, the respective EP surface distribution was obtained with the Delphi web server (imputed in Chimera interface) for mutated clusters representants models and the wild structure (WS). Information about the severity phenotype and others are specified by color in the box. Cluster 5 groups the WS together with eleven mutated structures, with no distance in relation to ten of them, which represent distinct severity (mild and severe), indicating that other properties are probably intervening in the phenotypes determination.

Figure S2. The Electrostatic similarities and patterns of structural models in the A3 domain. The clusters generation using webPIPSA tool groups models in terms of electrostatic potential (EP) similarities (left) based on the "skin" pattern. Next to the epogram, on the right, the respective EP surface distribution was obtained with the Delphi web server (imputed in Chimera interface) for mutated clusters representants models and the wild structure (WS). Information about the severity phenotype and others are specified by color in the box. Cluster 8 (up to down) groups the WS together with five mutated structures, which represent distinct severity (mild, moderate, and severe), indicating that other properties are probably intervening in the phenotypes determination. The WS shows no electrostatic distance from two mutated structures (Tyr1805Phe, Tyr1811Cys).

Figure S3. The Electrostatic similarities and patterns of structural models in the C1 domain. The clusters generation using webPIPSA tool groups models in terms of electrostatic potential (EP) similarities (left) based on the "skin" pattern. Next to the epogram, on the right, the respective EP surface distribution was obtained with the Delphi web server (imputed in Chimera interface) for mutated clusters representants models and the wild structure (WS). Information about the severity phenotype and others are specified by color in the box. Cluster 3 (up to down) groups the WS together with two mutated structures, which have moderate (no distance in relation to WS) , and severe phenotypes, indicating that other properties are probably intervening in the phenotypes determination.

Figure S4. The Electrostatic similarities and patterns of structural models in the C2 domain. The clusters generation using webPIPSA tool groups models in terms of electrostatic potential (EP) similarities (left) based on the "skin" pattern. Next to the epogram, on the right, the respective EP surface distribution was obtained with the Delphi web server (imputed in Chimera interface) for mutated clusters representants models and the wild structure (WS). Information about the severity phenotype and others are specified by color in the box. The clusterization includes WS and a severe mutation at the same cluster, even with some distance between them. Other properties are probably intervening in the phenotypes determination.

Figure S5. Heatmap of the electrostatic similarities matrice calculations for A1 domain. The tridimensional structures pair-to-pair comparison of models and wild type. The upper left side of the picture indicates the color key and density plots. The epograms consists of a distinct way to visualize identical information. WS: wild (normal) structure.

Figure S6. Visualization of position and values attributed to hydrophobicity of the mutated and wild structure residues in A2 domain. The amino acid residues locations in the protein chain can be visualized in the Wild Structural model (WS) in the center. The boxes depicted changes in terms of atomic conformation and hydrophobicity from the mutated residues (Mut) in comparison to wild type (WS). The hydrophobicity scale values vary from –4.5 (more hydrophilic residues) to 4.5 (more hydrophobic

ones). The maroon pigment represents more hydrophobic residues (those with positive values), while white and cyan colors represent neutral and hydrophilic (negative values), respectively.

Figure S7. Visualization of position and values attributed to hydrophobicity of the mutated and wild structure residues in A3 domain. The amino acid residues locations in the protein chain can be visualized in the Wild Structural model (WS) in the center. The boxes depicted changes in terms of atomic conformation and hydrophobicity from the mutated residues (Mut) in comparison to wild type (WS). The hydrophobicity scale values vary from –4.5 (more hydrophilic residues) to 4.5 (more hydrophobic ones). The maroon pigment represents more hydrophobic residues (those with positive values), while white and cyan colors represent neutral and hydrophilic (negative values), respectively.

Figure S8. Visualization of position and values attributed to hydrophobicity of the mutated and wild structure residues in C1 domain. The amino acid residues locations in the protein chain can be visualized in the Wild Structural model (WS) in the center. The boxes depicted changes in terms of atomic conformation and hydrophobicity from the mutated residues (Mut) in comparison to wild type (WS). The hydrophobicity scale values vary from –4.5 (more hydrophilic residues) to 4.5 (more hydrophobic ones). The maroon pigment represents more hydrophobic residues (those with positive values), while white and cyan colors represent neutral and hydrophilic (negative values), respectively.

Figure S9. Visualization of position and values attributed to hydrophobicity of the mutated and wild structure residues in C2 domain. The amino acid residues locations in the protein chain can be visualized in the Wild Structural model (WS) in the center. The boxes depicted changes in terms of atomic conformation and hydrophobicity from the mutated residues (Mut) in comparison to wild type (WS). The hydrophobicity scale values vary from –4.5 (more hydrophilic residues) to 4.5 (more hydrophobic ones). The maroon pigment represents more hydrophobic residues (those with positive values), while white and cyan colors represent neutral and hydrophilic (negative values), respectively.

Figure S10. Substitution matrices data of A1 domain mutations compiled in a Hierarchical Clustering Analysis (HCA). The clusters are only based on mutational values obtained from Sneath's index, Miyata's distance, Exchangeability, and Granthams score. The analysis conductions use correlation distance and average clustering method. The calculation of Approximately unbiased (AU) values is based on 10,000 bootstraps (BP) replications. The clusters formed shows a minor resolution in terms of phenotypes and disorder evaluation if compared to Figure 4.

Figure S11. Hierarchical clustering analysis (HCA) for the A2 domain, involving the physicochemical properties values obtained from mutated structures against the wild type. The analysis conductions use correlation distance and average clustering method. The calculation of Approximately unbiased (AU) values is based on 10,000 bootstraps (BP) replications. The clusterization considered the substitutions and the wild-type structures values for the following aspects: electrostatic potential, hydrophobicity, disulfide bond disruption, area SAS, area SES. Also included in the analysis the substitutions matrices and indexes: Sneath's index, Miyata's distance, Exchangeability, and Granthams score. The phenotype severity of the condition is colored as the legend: green, yellow, and red for the mild, moderate, and severe phenotypes, respectively. The WS legend contains an inside-circle star. Clusters are formed with an AU larger than 95% percent ($P<0,0001$).

Figure S12. Hierarchical clustering analysis (HCA) for the A3 domain, involving the physicochemical properties values obtained from mutated structures against the wild type. The analysis conductions use correlation distance and average clustering method. The calculation of Approximately unbiased (AU) values is based on 10,000 bootstraps (BP) replications. The clusterization considered the substitutions and the wild-type structures values for the following aspects: electrostatic potential, hydrophobicity, disulfide bond disruption, area SAS, area SES. Also included in the analysis the substitutions matrices and indexes: Sneath's index, Miyata's distance, Exchangeability, and Granthams score. The phenotype severity of the condition is colored as the legend: green, yellow, and red for the mild, moderate, and severe phenotypes, respectively. The WS legend contains an inside-circle star. Clusters are formed with an

AU larger than 95% percent (P<0,0001).

Figure S13. Hierarchical clustering analysis (HCA) for the C1 domain, involving the physicochemical properties values obtained from mutated structures against the wild type. The analysis conductions use correlation distance and average clustering method. The calculation of Approximately unbiased (AU) values is based on 10,000 bootstraps (BP) replications. The clusterization considered the substitutions and the wild-type structures values for the following aspects: electrostatic potential, hydrophobicity, disulfide bond disruption, area SAS, area SES. Also included in the analysis the substitutions matrices and indexes: Sneath's index, Miyata's distance, Exchangeability, and Granthams score. The phenotype severity of the condition is colored as the legend: green, yellow, and red for the mild, moderate, and severe phenotypes, respectively. The WS legend contains an inside-circle star. Clusters are formed with an AU larger than 95% percent (P<0,0001).

Figure S14. Hierarchical clustering analysis (HCA) for the C2 domain, involving the physicochemical properties values obtained from mutated structures against the wild type. The analysis conductions use correlation distance and average clustering method. The calculation of Approximately unbiased (AU) values is based on 10,000 bootstraps (BP) replications. The clusterization considered the substitutions and the wild-type structures values for the following aspects: electrostatic potential, hydrophobicity, disulfide bond disruption, area SAS, area SES. Also included in the analysis the substitutions matrices and indexes: Sneath's index, Miyata's distance, Exchangeability, and Granthams score. The phenotype severity of the condition is colored as the legend: green, yellow, and red for the mild, moderate, and severe phenotypes, respectively. The WS legend contains an inside-circle star. Clusters are formed with an AU larger than 95% percent (P<0,0001).

Figure S15. Multivariate Principal Component Analysis (PCA), including all evaluated features for the remained domains (A2, A3, C1, and C2). Two-dimensional graphics for variables and individuals, decomposition of total inertia graphic, and table (containing: the eigenvalues, the variance percentage, and the cumulative percentage of variance) depicted in A, B, C, and D corresponds to A2, A3, C1, and C2 domains, respectively. Colors for the Eigenvectors and variables are accord a scale (red depicts more contributive properties, and blue the less). The Individuals' distribution is accord to the features.

Figure S16. Multivariate Principal Component Analysis (PCA) comprehending the five mature domains of Factor VIII, with all the features evaluated in the study. In A - Two-dimensional graphics for variables and individuals. **B-** decomposition of total inertia graphic containing the percentage of variance in each one of the ten dimensions. On the right, a table showing: the eigenvalues, the variance percentage, and the cumulative percentage of variance.

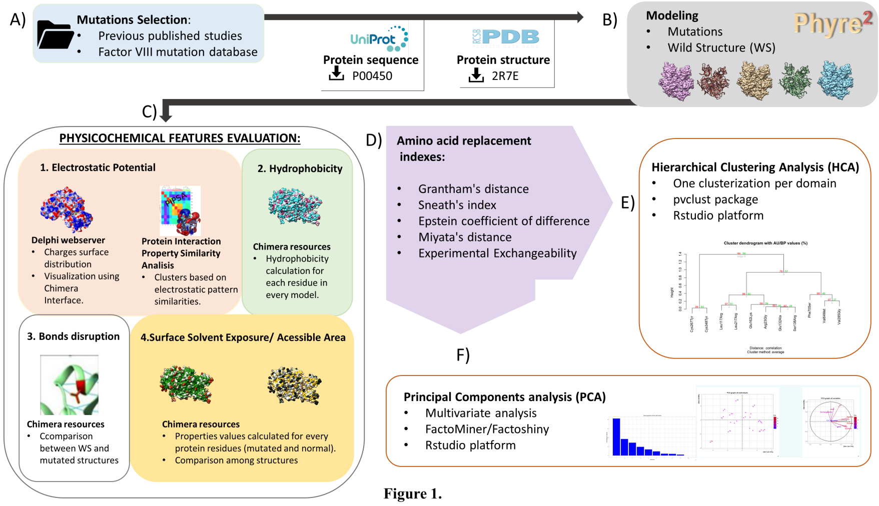

A)
Mutations Selection:
Previous published studies
Factor VIII mutation database
UniProt
Protein sequence
P00450
PDB
Protein structure
2R7E
B)
Modeling
Mutations
Wild Structure (WS)
Phyre2
C)
PHYSICOCHEMICAL FEATURES EVALUATION:
1. Electrostatic Potential
Delphi webserver
Charges surface distribution
Visualization using Chimera Interface.
PIPSA
Protein Interaction Property Similarity Analisis
Clusters based on electrostatic pattern similarities.
2. Hydrophobicity
Chimera resources
Hydrophobicity calculation for each residue in every model.
3. Bonds disruption
Chimera resources
Ccomparison between WS and mutated structures
4.Surface Solvent Exposure/ Acessible Area
Chimera resources
Properties values calculated for every protein residues (mutated and normal).
Comparison among structures
D)
Amino acid replacement indexes:
Grantham's distance
Sneath's index
Epstein coefficient of difference
Miyata's distance
Experimental Exchangeability
E)
Hierarchical Clustering Analysis (HCA)
One clusterization per domain
pvclust package
Rstudio platform
Cluster dendrogram with AU/BP values (%)
Height
Distance: correlation
Cluster method: average
F)
Principal Components analysis (PCA)
Multivariate analysis
FactoMiner/Factoshiny
Rstudio platform

Figure 1.

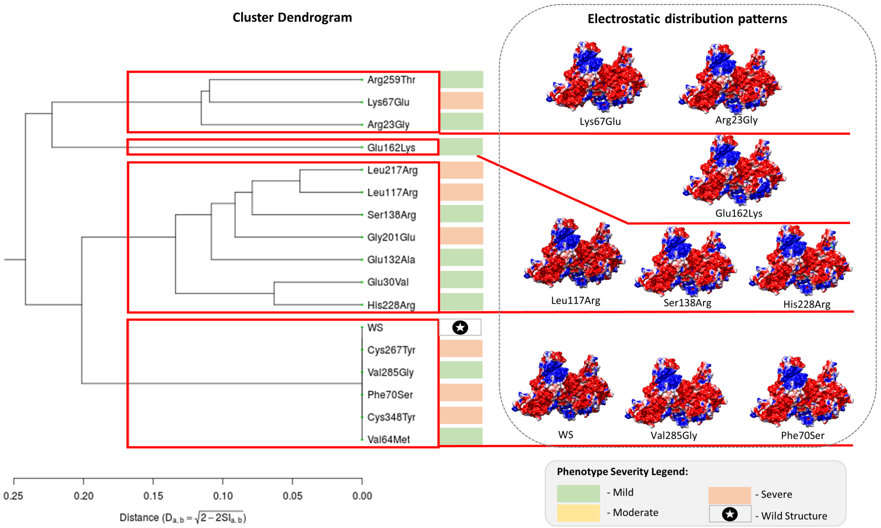


Figure 2.

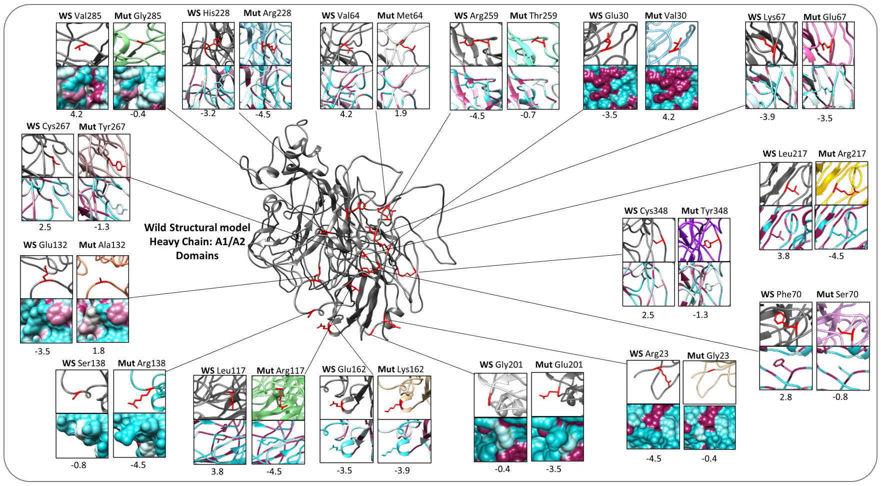


Figure 3

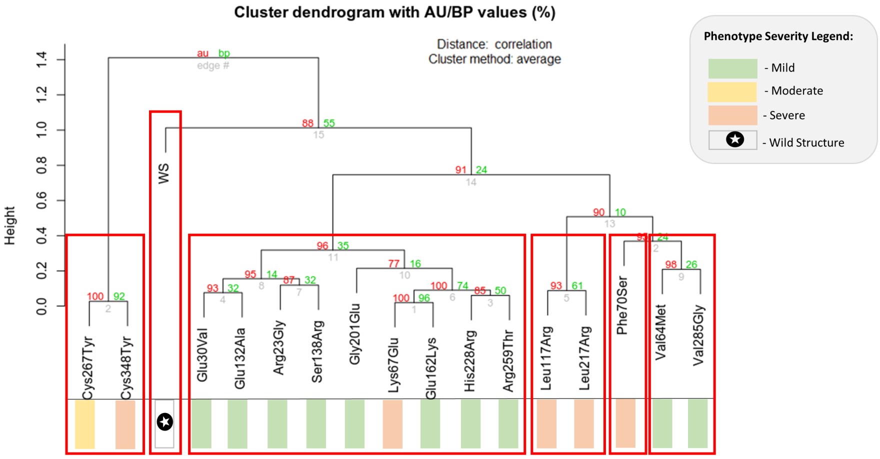
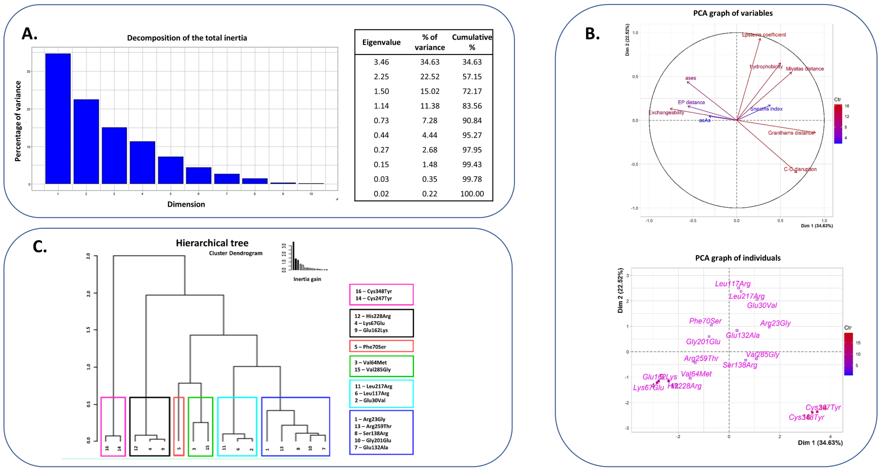


Figure 4.



Eigenvalue	% of variance	Cumulative %
3.46	34.63	34.63
2.25	22.52	57.15
1.50	15.02	72.17
1.14	11.38	83.56
0.73	7.28	90.84
0.44	4.44	95.27
0.27	2.68	97.95
0.15	1.48	99.43
0.03	0.35	99.78
0.02	0.22	100.00

16 – Cys348Tyr
14 – Cys247Tyr

12 – His228Arg
4 – Lys67Glu
9 – Glu162Lys

5 – Phe70Ser

3 – Val64Met
15 – Val285Gly

11 – Leu217Arg
6 – Leu117Arg
2 – Glu30Val

1 – Arg23Gly
13 – Arg259Thr
8 – Ser138Arg
10 – Gly201Glu
7 – Glu132Ala

Figure 5.

Hosted file

Table1.docx available at https://authorea.com/users/373906/articles/491521-application-of-a-new-in-silico-strategy-to-evidencing-the-role-of-missense-mutations-properties-in-determining-hemophilia-a

Hosted file

table_2.docx available at https://authorea.com/users/373906/articles/491521-application-of-a-new-in-silico-strategy-to-evidencing-the-role-of-missense-mutations-properties-in-determining-hemophilia-a