Structural effects driven by rare point mutations in amylin hormone, the type II diabetes associated peptide

Wendy S. Mendes[1,2], Octavio L. Franco[1,2,3], Sergio A. Alencar[1] and William F. Porto[1,4*]

[1] Programa de Pós-Graduação em Ciências Genômicas e Biotecnologia, Universidade Católica de Brasília, Brasília-DF, Brazil.
[2] Centro de Análises Proteômicas e Bioquímicas, Pós-Graduação em Ciências Genômicas e Biotecnologia, Universidade Católica de Brasília, Brasília-DF, Brazil.
[3] S-Inova Biotech, Pós-Graduação em Biotecnologia, Universidade Católica Dom Bosco, Campo Grande, MS, Brazil.
[4] Porto Reports, Brasília-DF, Brazil. http://www.portoreports.com

* Corresponding author: williamfp7@yahoo.com.br (WFP)

Abstract

Amylin is a 37 amino acid peptide hormone co-secreted with insulin, which participates in the glucose homeostasis. This hormone is able to aggregate in a β-sheet conformation and deposit in islet amyloids, a hallmark in type II diabetes. Since amylin is a gene encoded hormone, this peptide has variants caused by single nucleotide polymorphisms (SNPs) that can impact its functions. The effect of these variants can be evaluated by means of computational tools, such as molecular dynamics simulations. Here, we analyzed the structural effects caused by S20G and G33R point mutations, which, according to the 1,000 Genomes Project, have frequency in East Asian and Europe populations, respectively. By means of the aggrescan software and molecular dynamics, we found that both mutations have aggregation potential and cause changes in the monomeric forms when compared with wild-type amylin. The S20G mutation has been described as pathological related, and G33R mutation may have a deleterious effect. Furthermore, comparative analyses with pramlintide, an amylin drug analogue, allowed inferring the second α-helix maintenance may be related to the aggregation potential. Our data could help to better understand the mutations influence in the wild-type amylin sequence, being a starting point for the evaluation and characterization of other variations. The data presented here also propitiate new therapy opportunities, whether creating more effective drugs to diabetes or implementing a specific treatment to the patients with these mutations. Hence, these findings could improve the lifestyle to patients with type II diabetes.

Keywords: Molecular dynamics; Islet amyloid; S20G mutation; Pramlintide.

1 Introduction

Amylin is a hormone co-secreted with insulin in a 1:100 rate by the pancreatic beta cells [1]–[3], which participates in the glucose homeostasis causing a number of effects, including reduce in food intake, nutrient delivery delay, muscle glycogen synthesis inhibition and lipolytic-like effects [1]. Encoded from the *IAPP* gene, Amylin has a precursor and a pro-form, being the precursor synthetized with 89 amino acid residues [3]. This precursor loses the signal peptide, resulting in the 67 residues pro-form, which is processed to generate the 37 amino acid amylin mature form [4] (Fig. 1A). This peptide is part of the calcitonin gene peptide superfamily [5] and is encoded, in humans, on chromosome 12p12.1 [3].

This hormone is often related to the development or aggravation of type II diabetes [6]. Amylin was first isolated from the islet amyloid in patients with non-insulin-dependent diabetes mellitus [3], and thus, is also known as islet amyloid polypeptide. The islet amyloid is a denomination to the deposition of amylin fibrils, an aggregate polypeptide state with several β-sheet conformations [1], [7]. This amyloid form is related to cytotoxicity. However, its mechanisms are still not fully understood [8]. Even though the islet amyloid is a hallmark in type II diabetes, there are some hypothesis the oligomeric states could be the toxic form in amylin, and the process of deposition is an immune response to prevent damage [1], [9].

Comparisons involving amylin from different species and even with other peptides from the calcitonin superfamily enabled a better understanding of some regions in this molecule [2], [10]. Analyzing the conservation between these sequences and the aggregation response, the region harboring the amino acids Ser^{20}-Ser^{29} was suggested as the main responsible for the aggregation function [2], [8]. However, mutations within and out of this region are capable to impact the amyloid properties [8]. The only

substitution described with amyloid effect *in vivo* is the S20G in the amylin mature peptide [11]. Since the amidated C-Terminal and the Cys^2-Cys^7 disulphide bond existence and maintenance are essential to amylin bioactivity [12], the hormone activity is also susceptible to other mutations.

The study of mismatches in homologous sequences enables the development of synthetic peptides with specific properties. In this context, the pramlintide drug was developed based on non-amyloidogenic rat amylin sequence, which diverges in 7 residues from the human mature amylin sequence [10]. This amylin analogue drug has three mutations (A25P, S28P and S29P) and aggregates less than human amylin. Pramlintide also mimics some of the amylin functions, such as gastric emptying delay, being used as pharmacological treatment to some type II diabetics patients [13].

Nowadays, computational tools have been used for variants effects evaluation [14] with focus on mutations with frequency in population, described in projects such as the 1,000 Genomes Project [15]. The prediction tools are able to predict some deleterious behavior by means of different strategies that can be classified in four groups, sequence homology based, supervised learning based, structure based and consensus based [16]. Although these tools have limitations described in benchmarking analyses [17]–[20], they still can be used as filter or indicative in SNP effect studies [16]. Then, to improve the accuracy of the predictions, other techniques, such as molecular dynamics, can be used [21]. This approach has been used by our group to study other mutations related to peptide hormones, such as guanylin [22], uroguanylin [23], lymphoguanylin [24] and growth hormone receptor [25]. Here we show that the rare mutations in mature amylin caused by SNPs are located in conserved position, maintaining the aggregation potential and a similar structure conformation to the wild-

type after an unfolding dynamics simulation. Despite that, we also observed the variants cause changes in the monomeric forms when compared to the wild-type.

2 Results

2.1 Amylin mutations from European or Asian populations are in highly conserved positions

The amylin hormone is coded by its gene as a precursor protein and undergoes several processing steps until the release of the 37 amino acid mature peptide (Fig. 1A), therefore, this work was focused on mutations in the mature chain. According to the 1,000 Genomes data, the S20G mutation had frequency in the East Asian population (0.011) and was reported in three Chinese and one Japanese subpopulation while the G33R mutation had frequency in the European population (0.001), reported in British subpopulation. The positions where the mutations occur were classified as conserved according to Consurf's amino acids evolutionary conservation evaluation (Fig. 1A and Fig. S1). According to the pairwise F_{ST}, both mutations presented no genetic differentiation in their respective populations (Table S1).

2.2 European rare mutation is predicted as deleterious

The G33R mutation was predicted as deleterious in all five supervised learning method-based tools, and in four out of five sequence-based tools, while only one tool of the consensus-based group had a deleterious prediction (Table 1). In contrast to the S20G mutation with only three deleterious prediction results, two of them in tools part of the sequence-based group, and the other in a tool based in supervised learning method (Table 1). The consensus-based prediction for S20G mutation resulted in none of these tools with a deleterious prediction (Table 1). As well as to the structure-based

tools that evaluate stability effect, three out of five predicted G33R as a stabilizing mutation, while only one out of five had the same prediction to the S20G (Table 1).

2.3 Aggregation and secondary structure are maintained in the variants

Aggrescan analysis showed two aggregation regions in the wild-type sequence, which were maintained in both mutations. While the pramlintide, known to aggregates less than human amylin, loses the second region (Fig. 1C).

The resolved native amylin structure is composed by two α-helices (Fig. 1B). Structurally, the selected mutations, S20G and G33R, are located in the second α-helix and in the C-terminal loop, respectively (Fig. 1B). The validated models (Table S2) for the variants and the pramlintide used the native amylin structure as template. Despite the changes suffered after 1μs in an unfolding dynamics simulation, a similar structure was kept by the variants. On the other hand, pramlintide was capable of maintaining only the N-Terminal α-helix.

The molecular dynamics simulations were performed in water, while the NMR structure used as template (PDB ID: 2L86) was resolved in SDS micelles [26]. Hence, we can observe the unfolding behavior in molecular dynamics, with backbone's RMSD reaching values higher than 4 Å, indicating several conformation changes in all models (Fig. 2A). Regarding the variants, G33R mutation reached the highest radius of gyration (Rg) and solvent accessible surface area (SASA) values between all in the ensemble in the first half of the simulation period (Fig. 2B and 2C). Despite S20G mutation kept a similar structure to wild-type, its Rg and SASA values were similar to those of pramlintide (Fig. 2B and 2C).

On the principal component analysis (Fig. 2D), two very distinct clusters were observed in the G33R mutation, where a long α-helix, in the distinct cluster was noted at

the same time period of the Rg and SASA peak (Fig. 2B and 2C). Despite the same secondary structures were kept in the wild-type and the variants, they have differences in the length of their α-helices and the variants eigenvectors projection clusters were mainly overlapped with the pramlintide structure (Fig. S2).

3 Discussion

Amylin is a peptide hormone that participates in glucose homeostasis and, due to amyloidosis phenomenon, has a key role in type II diabetes pathophysiology [27], [28]. As a matter of fact, the amyloidogenic characteristic is extensively discussed [1], [7], [29], [30], and is understood as essential to determine normal or pathological peptide behavior [28].

Because amylin is a gene encoded hormone, its amino acid sequence could be modified by SNPs on the *IAPP* gene. Many strategies to SNP discovery have been described, but none of them is flawless [31]. Using SNPs with reported frequency means the variants were found in population and they could already affect people. Thus, the understanding of the impact that mutations with frequency cause is an immediate need propitiating create or adapt therapeutic intervention to avoid and remediate the possible issues in these affected population.

The rare variants selected here were caused by SNPs with global frequencies lower than 1%. The evolutionary amino acid maintenance represents the amino acid is probably needed in function or structure in the peptide and the S20G and G33R mutations were found in positions classified as evolutionary conserved (Fig. S1). As matter of fact, amino acid residues that perform specific functions, such as catalytic and interactions are often more conserved [32], [33]. Although this suggests conserved

mutations have higher propensity to functional or structural effects, the relation between pathogenicity and conservation must be carefully analyzed [34].

To contribute to the pathogenicity determination we analyze the SNP effect prediction, but this one, specifically, must be carefully analyzed because benchmarking analyses have shown these predictors lose accuracy to specific proteins [17]–[20]. Due to the known prediction limitations, other analyses were performed, such as aggregation prediction and molecular dynamics. Regarding the predictions the G33R mutation has highly deleterious prediction (Table 1) and no effects reported in the literature yet. Whereas S20G mutation is mostly neutral predicted (Table 1), however, its deleterious effect has been shown [11], [35], [36], and therefore, the predictions for S20G mutation could be false negatives.

Other interesting analysis regards the prediction tools that evaluate effects in molecule stability (Table 1). We can observe G33R was mainly predicted as stabilizer while the S20G results were often not stabilizer. The amyloidosis process is dynamic, reversible and directed to fibrils formation, therefore, molecules more stable usually hinder to reverse the process with the oligomeric formation that is believed to be the amylin toxic form [9].

The S20G mutation is described as involved in type II diabetes and has increased aggregation capacity [11], [35]–[37], being demonstrated this capacity is due to structural changes and the existence of three subunits in a fibril compared to wild-type fibrils with only two subunits [7]. Interestingly, Asians have higher type II diabetes rates [38], which might be related to the S20G frequency in the East Asian population. On the other hand, G33R mutation has a unique structural form caused by a repulsion of

between the C-terminal amide group and the Arg^{33} (Fig. S2). This structural form could impact the peptide function.

Point mutations are able to affect the aggregation capacity, *e.g.*, the pramlintide drug has three variants, A25P, S28P and S29P, near the C-terminal and a reduced amyloidogenicity. Because of that and also due to amyloidosis capacity being important to amylin evaluation, we used the pramlintide drug sequence as a control in the analyses with the wild-type sequence and the mutations (Fig. 1A). Our data shows that for both S20G and G33R mutations, the aggregation potential is maintained (Fig. 1C), besides these mutations cause structural behavior changes in the monomeric forms when compared with the wild-type (Fig. 2). Such data supports what has been described for S20G mutation, and also suggest a pathological potential for the G33R mutation. Thus, the aggregation potential is maintained in G33R (Fig. 1C) and therefore, this variant could be involved in the type II diabetes pathophysiology increasing the amyloid deposits formation.

Interestingly, the second region of aggregation and the C-Terminal α-helix present in both variants and in the wild-type is absent in the pramlintide sequence. Therefore, we can infer the existence of this secondary structure is probably related to the aggregation capacity. Thus, there could be a relation between aggregation potential and α-helix maintenance (Fig. 1C and Fig. S2), which could act as a classifier to amyloidogenic potential, with possibility to be used as an amylin deleterious mutation indicative.

The existence of SNPs in a sequence can be related to composition and structural effects [39]. The non-synonymous mutations, for instance, represent an amino acid change with several potential effects in structure, interactions, and conformations, even

though with different disease severity [40]. That way, a structural study can be a valuable tool to analyze the effects the mutations may have on protein function or physiology. Usually, this type of study is limited by the need of a resolved structure. However, due to recent advances in computational three-dimensional structure prediction field, such limitation might be overcome. The program AlphaFold2, an accurate method to structure prediction based on deep-leaning method, has improved the structural predictions, with some results being indistinguishable from structures determined by experimental methods [41]. Nevertheless, the AlphaFold2 limitations are still not clear, including cases of very dynamic structures, such as amylin, which changes upon oligomerization (recently, cryo-em experiments demonstrated changes in the oligomeric conformation caused by the S20G mutation [7]). Even with a possible limitation, the AlphaFold2 breakthrough could represent an immensurable improvement in the future studies.

4 Conclusions

Mature human amylin sequence has only two missense mutations with reported frequency in the 1,000 Genomes Project, the S20G and the G33R mutations. There are several studies trying to describe amylin cytotoxicity and, others trying to evaluate whether these mutations influence such cytotoxicity or even whether they play a pathological role in other forms. In that way, structural studies can combine both fields and try to identify the relation between the variation and the deleterious effect, as well to propitiate a better understand in the peptide behavior.

In this view, here, structural analysis demonstrated both mutations probably cause changes in the monomeric forms in comparison to the wild-type. Furthermore, a unique structure form in the G33R mutation was also observed, and its biological effects are still unknown. Data presented here also shows aggregation potential was

maintained in structures with S20G and G33R mutations and that this capacity is probably related to the C-Terminal α-helix maintenance. This work allows a better understanding of the structural effects caused by *IAPP* SNPs and could improve future SNPs evaluation. Therefore, a better comprehension on how mutations influence the wild-type amylin sequence allows new therapy opportunities, including creating more effective drugs to type II diabetes or implementing a specific treatment to the patients with these mutations. Our findings can also help and be used to continue the evaluation and characterization of other variations.

5 Material and Methods

5.1 Databases and SNPs selection

The SNPs were searched in the dbSNP database (build 153) using the Variation Viewer navigator from the NCBI (https://www.ncbi.nlm.nih.gov/variation/view/) [42]. The missense variant filter was applied and only the SNPs located in the mature sequence with frequency in the 1,000 Genomes Project phase 3 (https://www.internationalgenome.org/), reference assembly GRCh38 [15] were selected. The amylin sequence and the protein structure files of the mature peptide (PDB ID: 2L86) were obtained from the RCSB Protein Data Bank [26], [43].

5.2 SNP frequency

The SNP frequency data of *IAPP* gene were obtained from the 1,000 Genomes Project (phase 3) browser [15]. The browser provides the frequencies of all SNPs identified in the genomes of 2,504 individuals from 26 populations obtained through a combination of low-coverage (2–6×) whole genome sequence data, targeted deep (50–100×) exome sequencing, and dense SNP genotype data. The 26 populations studied were grouped by the predominant component of ancestry into five super-populations: African (AFR; 661 individuals), Admixed American (AMR; 347 individuals), East Asian (EAS; 504 individuals), South Asian (489 individuals), and European (EUR; 503 individuals). The fixation index (F_{ST}) was measured according to the model developed by Wright [44].

5.3 Functional analyzes of variants

The SNPs were submitted to functional impact prediction using 20 computational tools, divided into 4 different groups, as describe below.

5.3.1 Sequence homology-based methods

These methods determine whether a variation has functional effect or not by means of homologous sequence multiple alignments and sequence conservation. The principle in Sorting Intolerant From Tolerant (SIFT) [45], Provean [46], Mutation Assessor [47], Panther [48] and FATHMM [49] predictors. We used the SIFT sequence option in the single protein tools; and the PROVEAN protein tool in the PROVEAN web server, the remaining tools were used in default settings.

5.3.2 Supervised learning methods (SLM)

Tools such as SNAP [50], SuSPect [51], MutPred [52], PolyPhen-2 [53] and PhD-SNP [54] use an algorithm to learn patters of conservation and/or physicochemical properties. In SuSpect, we analyzed the mutation residues position in the precursor sequence using the Mutate an entire protein option; the other tools were performed using the default settings.

5.3.3 Structure-based methods

This group analyzes the SNPs effects by means of structural information, for instance solvent accessibility and free energy variation. We used SDM [55], Fold-X [56], PopMuSiC [57], HOTMuSiC [58] and SNPMusic [59] predictors in this category. Default settings were used in the first two tools and in the three last, the manual approach to select the specific SNPs. In SDM the negative and positive values represent a destabilizing and stabilizing prediction, respectively [55]. In FoldX, positive results are usually indicative of problems with the structure [56]. In PoPMuSiC and HoTMuSiC, negative and positive values, respectively, means stabilizing mutation. And in SNPMuSiC, positive and negative scores correspond to a deleterious and neutral prediction.

5.3.4 Consensus-based methods

The predictors in this category combine methods and tools to create a SNP effect classifier. In this group, the following tools were used: Condel [60], Meta-SNP [61], SNPeffect [62], TransFIC [63], [64] and PredictSNP [65]. The last one used the precursor FASTA sequence as an input and the tools nsSNPAnalyzer and PANTHER were included in the consensus. Default settings were used in the remaining predictors.

5.4 Evolutionary conservation analysis

The ConSurf server is a tool that, based on the phylogenetic relations between homologous sequences, can estimate the evolutionary conservation of amino acid positions in a protein molecule [66]. Using the precursor sequence in FASTA format obtained from the National Center of Biotechnology Information (NP_000406.1), ConSurf, in ConSeq mode, carried out a search for close homologous sequences using CSI-BLAST (3 iterations and 0.0001 e-value cutoff) against the UNIREF-90 protein database [67], [68]. The parameters were set as follows, maximum number of homologs as 150 and a 35 minimal to 95 maximal percentage ID between sequence. Multiple sequence alignment and calculation methods were used as default (MAFFT-L-INS-I and Bayesian). The sequences were then clustered and highly similar sequences removed using CD-HIT [69]. The empirical Bayesian algorithm was used to compute the position-specific conservation scores [70].

5.5 Aggregation hot spot areas prediction

Aggrescan is based on an aggregation-propensity scale for natural amino acids and is able to predict aggregation-prone segments in protein sequences [71]. The sequence in FASTA format were used as input, for each the aggrescan return the aggregation-propensity value for residue and aggregation hot spots from the peptide aggregation profile.

5.6 Molecular modelling

One hundred molecular models for each ensembles (the two variants and the pramlintide drug) were constructed by comparative molecular modelling by means of MODELLER 9.16 [72] using the structure of amylin (PDB ID 2l86) as template. Three-dimensional models were constructed using automodel and environ classes from MODELLER with default settings. The final models were select according to the discrete optimized protein energy score (DOPE score) which indicates the most probable structures [73]. The best models were evaluated through PROSA II [74], QMEAN [75] and PROCHECK [76] softwares. PROSA II is a tool applicated to evaluate protein structures by means of a score based in a comparison with all known protein structures that indicates the fold quality. QMEAN uses a scoring function that combines six structural descriptors to estimate the absolute quality of a protein structure. PROCHECK checks the stereochemical quality of a protein structure by means of Ramachandran plot, in which models with good quality have more than 90% of residues in the most favored and additional allowed regions. The models structures were visualized in PyMOL (http:// www.pymol.org).

5.7 Molecular dynamics simulation

The molecular dynamics simulation of the wild-type, the variants and the pramlintide drug were performed in the GROMACS 4.6 [77] package using the CHARMM force field [78], [79]. The simulations were done in a saline environment (0.2 M NaCl) where the structures were immersed in water cubic boxes with a 1.2 nm distance between the protein and the edge of the box. Geometry of water molecules was constrained by using the SETTLE algorithm [80]. All atom bond lengths were linked by using the LINCS algorithm [81]. Electrostatic corrections were made by Particle Mesh Ewald algorithm [82], with a threshold of 1.4 nm to minimize the computational time.

The same cut-off radius was also used for van der Waals interactions. The steepest descent algorithm was applied to minimize system energy for 50,000 steps. The system with temperature normalized to 310 K and pressure normalized to 1 bar were simulated for 1 µs.

5.8 Analyses of molecular dynamics trajectories

Molecular dynamics simulations were analyzed by means of the backbone root mean square deviation (RMSD), radius of gyration (Rg) and solvent accessible surface area (SASA) using the g_rms, g_gyrate and g_sas built in functions of the GROMACS package. Except the RMSD that used the GROMACS group 4 (backbone) to analyses, the functions used the GROMACS group 1 (protein). The principal component analysis (PCA) was used to analyze and visualize the protein motion in the simulation. For that, the protein movement is extract and split in different components. For visualization, the two largest class of motion (the 1st and the 2nd eigenvectors) were plot against each other. The PCA, or essential dynamics analysis, were done using the g_covar and g_anaeig functions.

6 Acknowledgements

This work was supported by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico); CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), FAPDF (Fundação de Amparo a Pesquisa do Distrito Federal) and FUNDECT (Fundação de Apoio ao Desenvolvimento do Ensino, Ciência e Tecnologia do Estado de Mato Grosso do Sul).

7 References

[1] X.-X. Zhang, Y.-H. Pan, Y.-M. Huang, and H.-L. Zhao, “Neuroendocrine hormone amylin in diabetes,” *World J. Diabetes*, vol. 7, no. 9, 2016, doi: 10.4239/wjd.v7.i9.189.

[2] R. A. Pittner *et al.*, “Molecular physiology of amylin,” *J. Cell. Biochem.*, vol. 55, no. S1994A, 1994, doi: 10.1002/jcb.240550004.

[3] G. J. S. Cooper, A. J. Day, A. C. Willis, A. N. Roberts, K. B. M. Reid, and B. Leighton, “Amylin and the amylin gene: structure, function and relationship to islet amyloid and to diabetes mellitus,” *BBA - Molecular Cell Research*, vol. 1014, no. 3. 1989, doi: 10.1016/0167-4889(89)90220-6.

[4] R. Akter *et al.*, “Islet Amyloid Polypeptide: Structure, Function, and Pathophysiology,” *J. Diabetes Res.*, vol. 2016, 2016, doi: 10.1155/2016/2798269.

[5] S. J. Wimalawansa, “Amylin, calcitonin gene-related peptide, calcitonin, and adrenomedullin: A peptide superfamily,” *Critical Reviews in Neurobiology*, vol. 11, no. 2–3. 1997, doi: 10.1615/CritRevNeurobiol.v11.i2-3.40.

[6] P. Westermark, C. Wernstedt, E. Wilander, and K. Sletten, “A novel peptide in the calcitonin gene related peptide family as an amyloid fibril protein in the endocrine pancreas,” *Biochem. Biophys. Res. Commun.*, vol. 140, no. 3, 1986, doi: 10.1016/0006-291X(86)90708-4.

[7] R. Gallardo *et al.*, “Fibril structures of diabetes-related amylin variants reveal a basis for surface-templated assembly,” *Nat. Struct. Mol. Biol.*, vol. 27, no. 11, 2020, doi: 10.1038/s41594-020-0496-3.

[8] P. Cao *et al.*, “Islet amyloid: From fundamental biophysics to mechanisms of

cytotoxicity," *FEBS Letters*, vol. 587, no. 8. 2013, doi: 10.1016/j.febslet.2013.01.046.

[9] K. E. Marshall, R. Marchante, W. F. Xue, and L. C. Serpell, "The relationship between amyloid structure and cytotoxicity," *Prion*, vol. 8, no. 2, 2014, doi: 10.4161/pri.28860.

[10] W. M. Berhanu and U. H. E. Hansmann, "Inter-species cross-seeding: Stability and assembly of rat - Human amylin aggregates," *PLoS One*, vol. 9, no. 5, 2014, doi: 10.1371/journal.pone.0097051.

[11] S. Sakagashira *et al.*, "S20G mutant amylin exhibits increased in vitro amyloidogenicity and increased intracellular cytotoxicity compared to wild-type amylin," *Am. J. Pathol.*, vol. 157, no. 6, 2000, doi: 10.1016/S0002-9440(10)64848-1.

[12] R. L. Bower and D. L. Hay, "Amylin structure-function relationships and receptor pharmacology: Implications for amylin mimetic drug development," *British Journal of Pharmacology*, vol. 173, no. 12. 2016, doi: 10.1111/bph.13496.

[13] M. Samsom, L. A. Szarka, M. Camilleri, A. Vella, A. R. Zinsmeister, and R. A. Rizza, "Pramlintide, an amylin analog, selectively delays gastric emptying: Potential role of vagal inhibition," *Am. J. Physiol. - Gastrointest. Liver Physiol.*, vol. 278, no. 6 41-6, 2000, doi: 10.1152/ajpgi.2000.278.6.g946.

[14] Z. Zhang, M. A. Miteva, L. Wang, and E. Alexov, "Analyzing effects of naturally occurring missense mutations," *Computational and Mathematical Methods in Medicine*, vol. 2012. 2012, doi: 10.1155/2012/805827.

[15] D. M. Altshuler *et al.*, “An integrated map of genetic variation from 1,092 human genomes,” *Nature*, vol. 491, no. 7422, 2012, doi: 10.1038/nature11632.

[16] A. S. Pires, W. F. Porto, O. L. Franco, and S. A. Alencar, “In silico analyses of deleterious missense SNPs of human apolipoprotein E3,” *Sci. Rep.*, vol. 7, no. 1, 2017, doi: 10.1038/s41598-017-01737-w.

[17] K. C. do V. Ferreira, L. F. Fialho, O. L. Franco, S. A. de Alencar, and W. F. Porto, “Benchmarking analysis of deleterious SNP prediction tools on CYP2D6 enzyme,” *Chem. Biol. Drug Des.*, vol. 96, no. 3, 2020, doi: 10.1111/cbdd.13676.

[18] D. G. Grimm *et al.*, “The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity,” *Hum. Mutat.*, vol. 36, no. 5, 2015, doi: 10.1002/humu.22768.

[19] S. Hicks, D. A. Wheeler, S. E. Plon, and M. Kimmel, “Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed,” *Hum. Mutat.*, vol. 32, no. 6, 2011, doi: 10.1002/humu.21490.

[20] C. Rodrigues, A. Santos-Silva, E. Costa, and E. Bronze-da-Rocha, “Performance of In Silico Tools for the Evaluation of UGT1A1 Missense Variants,” *Hum. Mutat.*, vol. 36, no. 12, 2015, doi: 10.1002/humu.22903.

[21] A. Kumar and R. Purohit, “Use of Long Term Molecular Dynamics Simulation in Predicting Cancer Associated SNPs,” *PLoS Comput. Biol.*, vol. 10, no. 4, 2014, doi: 10.1371/journal.pcbi.1003318.

[22] W. F. Porto, O. L. Franco, and S. A. Alencar, “Computational analyses and prediction of guanylin deleterious SNPs,” *Peptides*, vol. 69, 2015, doi: 10.1016/j.peptides.2015.04.013.

[23] A. C. S. Marcolino, W. F. Porto, Á. S. Pires, O. L. Franco, and S. A. Alencar, "Structural impact analysis of missense SNPs present in the uroguanylin gene by long-term molecular dynamics simulations," *J. Theor. Biol.*, vol. 410, 2016, doi: 10.1016/j.jtbi.2016.09.008.

[24] Á. S. Pires, W. F. Porto, P. O. Castro, O. L. Franco, and S. A. Alencar, "Theoretical structural characterization of lymphoguanylin: A potential candidate for the development of drugs to treat gastrointestinal disorders," *J. Theor. Biol.*, vol. 419, 2017, doi: 10.1016/j.jtbi.2017.02.016.

[25] W. F. Porto *et al.*, "Computational Investigation of Growth Hormone Receptor Trp169Arg Heterozygous Mutation in a Child With Short Stature," *J. Cell. Biochem.*, vol. 118, no. 12, 2017, doi: 10.1002/jcb.26144.

[26] R. P. R. Nanga, J. R. Brender, S. Vivekanandan, and A. Ramamoorthy, "Structure and membrane orientation of IAPP in its natively amidated form at physiological pH in a membrane environment," *Biochim. Biophys. Acta - Biomembr.*, vol. 1808, no. 10, 2011, doi: 10.1016/j.bbamem.2011.06.012.

[27] S. C. Woods, T. A. Lutz, N. Geary, and W. Langhans, "Pancreatic signals controlling food intake; insulin, glucagon and amylin," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 361, no. 1471. 2006, doi: 10.1098/rstb.2006.1858.

[28] L. Haataja, T. Gurlo, C. J. Huang, and P. C. Butler, "Islet amyloid in type 2 diabetes, and the toxic oligomer hypothesis," *Endocrine Reviews*, vol. 29, no. 3. 2008, doi: 10.1210/er.2007-0037.

[29] C. Goldsbury *et al.*, "Amyloid fibril formation from full-length and fragments of amylin," *J. Struct. Biol.*, vol. 130, no. 2–3, 2000, doi: 10.1006/jsbi.2000.4268.

[30] D. C. Rodriguez Camargo *et al.*, "The redox environment triggers conformational changes and aggregation of hIAPP in Type II Diabetes," *Sci. Rep.*, vol. 7, 2017, doi: 10.1038/srep44041.

[31] A. Vignal, D. Milan, M. SanCristobal, and A. Eggen, "A review on SNP and other types of molecular markers and their use in animal genetics," *Genetics Selection Evolution*, vol. 34, no. 3. 2002, doi: 10.1051/gse:2002009.

[32] R. A. Studer, B. H. Dessailly, and C. A. Orengo, "Residue mutations and their impact on protein structure and function: Detecting beneficial and pathogenic changes," *Biochemical Journal*, vol. 449, no. 3. 2013, doi: 10.1042/BJ20121221.

[33] S. Gong, C. L. Worth, G. R. J. Bickerton, S. Lee, D. Tanramluk, and T. L. Blundell, "Structural and functional restraints in the evolution of protein families and superfamilies," in *Biochemical Society Transactions*, 2009, vol. 37, no. 4, doi: 10.1042/BST0370727.

[34] H. Sun and G. Yu, "New insights into the pathogenicity of non-synonymous variants through multi-level analysis," *Sci. Rep.*, vol. 9, no. 1, 2019, doi: 10.1038/s41598-018-38189-9.

[35] S. C. Lee *et al.*, "The islet amyloid polypeptide (amylin) gene S20G mutation in Chinese subjects: Evidence for associations with type 2 diabetes and cholesterol levels," *Clin. Endocrinol. (Oxf).*, vol. 54, no. 4, 2001, doi: 10.1046/j.1365-2265.2001.01244.x.

[36] S. Seino, "S20G mutation of the amylin gene is associated with Type II diabetes in Japanese. Study Group of Comprehensive Analysis of Genetic Factors in Diabetes Mellitus," *Diabetologia*, vol. 44, no. 7, 2001.

[37] S. Sakagashira *et al.*, “Missense mutation of amylin gene (S20G) in Japanese NIDDM patients,” *Diabetes*, vol. 45, no. 9, 1996, doi: 10.2337/diab.45.9.1279.

[38] E. J. Rhee, “Diabetes in Asians,” *Endocrinology and Metabolism*, vol. 30, no. 3. 2015, doi: 10.3803/EnM.2015.30.3.263.

[39] J. Duan *et al.*, “Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor,” *Human Molecular Genetics*, vol. 12, no. 3. 2003, doi: 10.1093/hmg/ddg055.

[40] A. Ali, S. Ghosh, and A. Bagchi, “Structural study of the effects of mutations in proteins to identify the molecular basis of the loss of local structural fluidity leading to the onset of autoimmune diseases,” *Biochem. Biophys. Res. Commun.*, vol. 484, no. 1, 2017, doi: 10.1016/j.bbrc.2017.01.048.

[41] A. W. Senior *et al.*, “Improved protein structure prediction using potentials from deep learning,” *Nature*, vol. 577, no. 7792, 2020, doi: 10.1038/s41586-019-1923-7.

[42] S. T. Sherry *et al.*, “DbSNP: The NCBI database of genetic variation,” *Nucleic Acids Res.*, vol. 29, no. 1, 2001, doi: 10.1093/nar/29.1.308.

[43] F. C. Bernstein *et al.*, “The protein data bank: A computer-based archival file for macromolecular structures,” *J. Mol. Biol.*, vol. 112, no. 3, 1977, doi: 10.1016/S0022-2836(77)80200-3.

[44] S. WRIGHT, “The genetical structure of populations,” *Ann. Eugen.*, vol. 15, no. 4, 1951, doi: 10.2307/2407273.

[45] P. Kumar, S. Henikoff, and P. C. Ng, “Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm,” *Nat.*

Protoc., vol. 4, no. 7, 2009, doi: 10.1038/nprot.2009.86.

[46] Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, and A. P. Chan, "Predicting the Functional Effect of Amino Acid Substitutions and Indels," *PLoS One*, vol. 7, no. 10, 2012, doi: 10.1371/journal.pone.0046688.

[47] B. Reva, Y. Antipin, and C. Sander, "Predicting the functional impact of protein mutations: Application to cancer genomics," *Nucleic Acids Res.*, vol. 39, no. 17, 2011, doi: 10.1093/nar/gkr407.

[48] H. Mi *et al.*, "The PANTHER database of protein families, subfamilies, functions and pathways," *Nucleic Acids Res.*, vol. 33, no. DATABASE ISS., 2005, doi: 10.1093/nar/gki078.

[49] H. A. Shihab *et al.*, "Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models," *Hum. Mutat.*, vol. 34, no. 1, 2013, doi: 10.1002/humu.22225.

[50] Y. Bromberg, G. Yachdav, and B. Rost, "SNAP predicts effect of mutations on protein function," *Bioinformatics*, vol. 24, no. 20, 2008, doi: 10.1093/bioinformatics/btn435.

[51] C. M. Yates, I. Filippis, L. A. Kelley, and M. J. E. Sternberg, "SuSPect: Enhanced prediction of single amino acid variant (SAV) phenotype using network features," *J. Mol. Biol.*, vol. 426, no. 14, 2014, doi: 10.1016/j.jmb.2014.04.026.

[52] B. Li *et al.*, "Automated inference of molecular mechanisms of disease from amino acid substitutions," *Bioinformatics*, vol. 25, no. 21, 2009, doi: 10.1093/bioinformatics/btp528.

[53] I. A. Adzhubei *et al.*, "A method and server for predicting damaging missense mutations," *Nature Methods*, vol. 7, no. 4. 2010, doi: 10.1038/nmeth0410-248.

[54] E. Capriotti, R. Calabrese, and R. Casadio, "Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information," *Bioinformatics*, vol. 22, no. 22, 2006, doi: 10.1093/bioinformatics/btl423.

[55] A. P. Pandurangan, B. Ochoa-Montaño, D. B. Ascher, and T. L. Blundell, "SDM: A server for predicting effects of mutations on protein stability," *Nucleic Acids Res.*, vol. 45, no. W1, 2017, doi: 10.1093/nar/gkx439.

[56] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano, "The FoldX web server: An online force field," *Nucleic Acids Res.*, vol. 33, no. SUPPL. 2, 2005, doi: 10.1093/nar/gki387.

[57] Y. Dehouck, J. M. Kwasigroch, D. Gilis, and M. Rooman, "PoPMuSiC 2.1: A web server for the estimation of protein stability changes upon mutation and sequence optimality," *BMC Bioinformatics*, vol. 12, 2011, doi: 10.1186/1471-2105-12-151.

[58] F. Pucci, R. Bourgeas, and M. Rooman, "Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC," *Sci. Rep.*, vol. 6, 2016, doi: 10.1038/srep23257.

[59] F. Ancien, F. Pucci, M. Godfroid, and M. Rooman, "Prediction and interpretation of deleterious coding variants in terms of protein structural stability," *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, doi: 10.1038/s41598-018-22531-2.

[60] A. González-Pérez and N. López-Bigas, "Improving the assessment of the

outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel," *Am. J. Hum. Genet.*, vol. 88, no. 4, 2011, doi: 10.1016/j.ajhg.2011.03.004.

[61] E. Capriotti, R. B. Altman, and Y. Bromberg, "Collective judgment predicts disease-associated single nucleotide variants.," *BMC Genomics*, vol. 14 Suppl 3, 2013, doi: 10.1186/1471-2164-14-s3-s2.

[62] J. Reumers, J. Schymkowitz, J. Ferkinghoff-Borg, F. Stricher, L. Serrano, and F. Rousseau, "SNPeffect: A database mapping molecular phenotypic effects of human non-synonymous coding SNPs," *Nucleic Acids Research*, vol. 33, no. DATABASE ISS. 2005, doi: 10.1093/nar/gki086.

[63] R. Ben-Hamo and S. Efroni, "Gene expression and network-based analysis reveals a novel role for hsa-miR-9 and drug control over the p38 network in glioblastoma multiforme progression," *Genome Med.*, vol. 3, no. 11, 2011, doi: 10.1186/gm293.

[64] A. Gonzalez-Perez, J. Deu-Pons, and N. Lopez-Bigas, "Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation," *Genome Med.*, vol. 4, no. 11, 2012, doi: 10.1186/gm390.

[65] J. Bendl *et al.*, "PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations," *PLoS Comput. Biol.*, vol. 10, no. 1, 2014, doi: 10.1371/journal.pcbi.1003440.

[66] G. Celniker *et al.*, "ConSurf: Using evolutionary data to raise testable hypotheses about protein function," *Israel Journal of Chemistry*, vol. 53, no. 3–4. 2013, doi: 10.1002/ijch.201200096.

[67] C. Angermüller, A. Biegert, and J. Söding, “Discriminative modelling of context-specific amino acid substitution probabilities,” *Bioinformatics*, vol. 28, no. 24, 2012, doi: 10.1093/bioinformatics/bts622.

[68] B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu, “UniRef: Comprehensive and non-redundant UniProt reference clusters,” *Bioinformatics*, vol. 23, no. 10, 2007, doi: 10.1093/bioinformatics/btm098.

[69] W. Li and A. Godzik, “Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences,” *Bioinformatics*, vol. 22, no. 13, 2006, doi: 10.1093/bioinformatics/btl158.

[70] I. Mayrose, D. Graur, N. Ben-Tal, and T. Pupko, “Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior,” *Mol. Biol. Evol.*, vol. 21, no. 9, 2004, doi: 10.1093/molbev/msh194.

[71] O. Conchillo-Solé, N. S. de Groot, F. X. Avilés, J. Vendrell, X. Daura, and S. Ventura, “AGGRESCAN: A server for the prediction and evaluation of ‘hot spots’ of aggregation in polypeptides,” *BMC Bioinformatics*, vol. 8, 2007, doi: 10.1186/1471-2105-8-65.

[72] N. Eswar *et al.*, “Comparative Protein Structure Modeling Using MODELLER,” *Curr. Protoc. Protein Sci.*, vol. 50, no. 1, 2007, doi: 10.1002/0471140864.ps0209s50.

[73] M. Shen and A. Sali, “Statistical potential for assessment and prediction of protein structures,” *Protein Sci.*, vol. 15, no. 11, 2006, doi: 10.1110/ps.062416606.

[74] M. Wiederstein and M. J. Sippl, “ProSA-web: Interactive web service for the

recognition of errors in three-dimensional structures of proteins," *Nucleic Acids Res.*, vol. 35, no. SUPPL.2, 2007, doi: 10.1093/nar/gkm290.

[75] P. Benkert, M. Biasini, and T. Schwede, "Toward the estimation of the absolute quality of individual protein structure models," *Bioinformatics*, vol. 27, no. 3, 2011, doi: 10.1093/bioinformatics/btq662.

[76] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton, "PROCHECK: a program to check the stereochemical quality of protein structures," *J. Appl. Crystallogr.*, vol. 26, no. 2, 1993, doi: 10.1107/s0021889892009944.

[77] B. Hess, C. Kutzner, and D. Van Der Spoel, "GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation - Journal of Chemical Theory and Computation (ACS Publications)," *J. Chem. …*, 2008.

[78] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations," *J. Comput. Chem.*, vol. 4, no. 2, 1983, doi: 10.1002/jcc.540040211.

[79] B. R. Brooks *et al.*, "CHARMM: The biomolecular simulation program," *J. Comput. Chem.*, vol. 30, no. 10, 2009, doi: 10.1002/jcc.21287.

[80] S. Miyamoto and P. A. Kollman, "Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models," *J. Comput. Chem.*, vol. 13, no. 8, 1992, doi: 10.1002/jcc.540130805.

[81] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, "LINCS: A Linear Constraint Solver for molecular simulations," *J. Comput. Chem.*, vol. 18,

no. 12, 1997, doi: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

[82] T. Darden, D. York, and L. Pedersen, “Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems,” *J. Chem. Phys.*, vol. 98, no. 12, 1993, doi: 10.1063/1.464397.

8 Tables

Table 1. Prediction results of Amylin SNPs analyzed by 20 prediction tools classified in four different groups

		Sequence-based[b]					SLM-based[b]					Consensus-based[b]					Structure-based[c]				
SNP rs #	Amino acid change[a]	SIFT	PROVEAN	Mutation Assessor	Panther	FATHMM	SNAP	SuSPect	MutPred	PolyPhen-2	PhD-SNP	Condel	Meta-SNP	SNPeffect	PredictSNP	TransFIC	SDM	PoPMuSiC	HOTMuSiC	SNPMusic	Fold-X
rs1800203: A>G	S20G	N	D	N	D	N	D	N	N	N	N	N	N	N	N	N	0.16*	0.41	-0.75	-0.83	0.12
rs200376097: G>A	G33R	D	D	D	D	N	D	D	D	D	D	N	N	N	D	N	0.35*	-0.17*	1.27*	0.43	2.15

[a] Both SNPs were validated by frequency on 1,000 Genome Project.
[b] N: Neutral; D: Deleterious.
[c] These prediction tools present the results in the form of a score (represented in the table) that can be interpreted as stabilizer or destabilizer. The results interpreted as stabilizer are marked with an asterisk (*) after the number. The interpretation of whether stabilizer effect is deleterious or neutral is further discussed.

In the analyses was used the mature amylin peptide structure (PDB ID: 2L86).

9 Figures

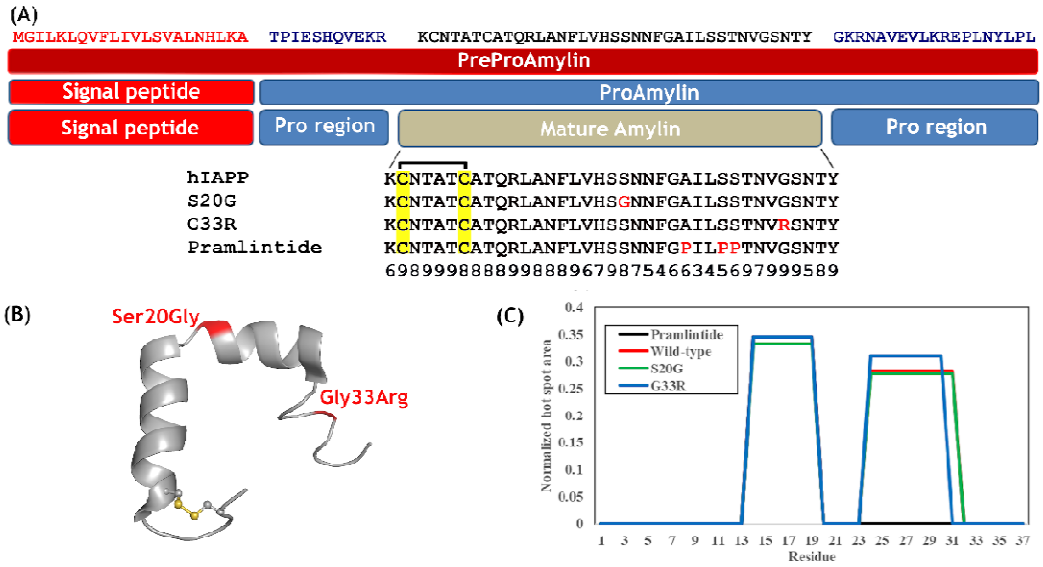

Fig. 1. Amylin and its variants. (A) Amylin processing and variants alignment. PreProAmylin is the amylin precursor with 89 amino acids. The precursor has a 22-residue signal peptide, which is cleaved to form the ProAmylin. This propeptide has two flanking regions the are removed by the prohormone convertases enzymes PC2 and PC3 in the processing to release the mature peptide. In the process, the peptidyl amidating mono-oxygenase complex acts catalyzing an amidation reaction on the C-terminal. The 37 amino acid residues mature amylin has a disulfide bond, formed by Cys^2 and Cys^7 (highlighted in yellow). The SNPs S20G and G33R are the only missense variants reported with frequency on the 1,000 genomes project. The amylin analogue drug, pramlintide, have the point mutations A25P, S28P e S29P derived from the rat amylin, to reduce the molecule aggregation capacity. The amino acids that differ from the wild-type human amylin in the alignment are highlighted in red. The numbers below the alignment represent an increasing scale of conservation from Consurf (see Fig S1). (B) Tridimensional structure of wild-type amylin. This molecule structure (PDB 2l86) present two alfa-helices in Thr^6-His^{18} and Ser^{20}-Ser^{29}. The disulfide bond is represented in ball and stick. The mutation residues in S20G and G33R are shown in red. (C) Hotspot area representation using aggrescan. This data indicates the wild-type has two distinct regions of aggregation in 13-20 and 23-32 residues, while the pramlintide drug, well known less amyloidogenic, have just the 13-20 residue region. Although the SNPs also have two regions with high hotspot area values, the second region from G33R only comprehend the residues 23-31.

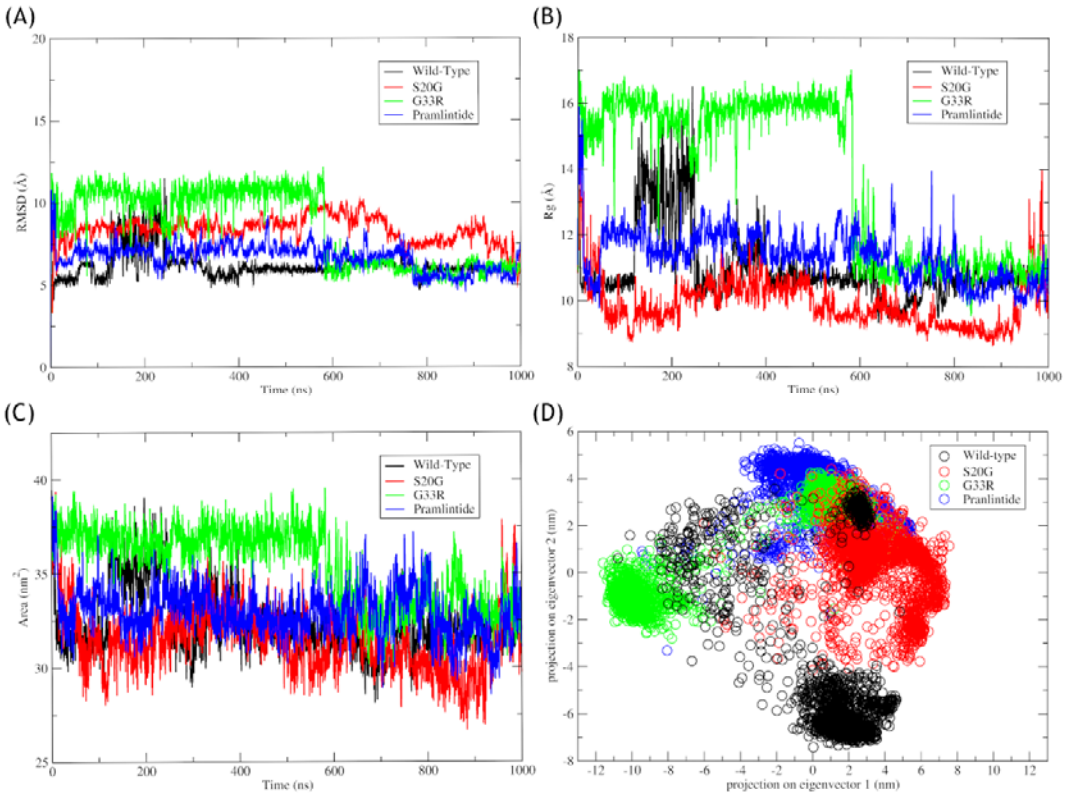

Fig. 2. Analyses of molecular dynamics trajectories. (A) The RMSD variation during the simulations. Due to values higher than 4Å, we can notice a structure change in all models. B) The radius of gyration (Rg) variation during the simulations. (C) The solvent accessible surface area (SASA) variation during the simulations. The Rg and the SASA have a similar behavior, showing the structure forms with higher expansion also have high solvent accessible surface area. (D) Projection of the motion of amylin in phase space along the first two principal eigenvectors. The variants and the drug pramlintide have a small overlap among the structures while most wild-type structures form a distant cluster (eigenvector 1: 2 and eigenvector 2: -6). The G33R left cluster (eigenvector 1: -10 and eigenvector 2: 1) represent an intermediate structure from the 0 – 600 ns dynamics period.