Potential aggregation hot spots in recombinant human keratinocyte growth factor; a computational study

Mansoureh Shahbazi Dastjerdeh
Pasteur Institute of Iran
Mohammad Ali Shokrgozar
Pasteur Institute of Iran
Hamzeh Rahimi (rahimi.h1981@gmail.com)
Pasteur Institute of Iran
Majid Golkar
Pasteur Institute of Iran

Research article

Keywords: In silico, Aggregation-prone regions, Aggregation hot spots, Sequence-based aggregation prediction, Structure-based aggregation prediction, Spatial Aggregation Propensity (SAP), Aggregation, Unfolding, Recombinant human keratinocyte growth factor

Posted Date: November 2nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-98614/v1

Version of Record: A version of this preprint was published on April 10th, 2021. See the published version at https://doi.org/10.1080/07391102.2021.1908912.

Abstract

Background: Recombinant human keratinocyte growth factor is a highly aggregation-prone therapeutic protein. The high aggregation liability of rhKGF is manifested by loss of the monomeric form of the protein and accumulation of the aggregated species even at moderate temperatures. Here, we analyzed rhKGF for its vulnerability towards aggregation by detection of aggregation-prone regions (APRs) using several sequence-based computational tools including TANGO, SolubiS, ZipperDB, AGGRESCAN, Zyggregator, Camsol, PASTA, SALSA, WALTZ, SODA, Amylpred, AMYPDB, and structure-based tools including Aggrescan3D and molecular dynamics-based spatial aggregation propensity (SAP) algorithm.

Results: The sequence-based prediction of APRs in rhKGF indicated that they are mainly located at positions 10-30, 40-60, 61-66, 88-120, and 130-140 which are rich in β-branched aliphatic, hydrophobic, aromatic and Glutamine/Aspargine (Q/N) residues. Mapping on the rhKGF tertiary structure revealed that most of these residues including F16-R25, I43, E45, R47-I56, F61, Y62, N66, L88-E91, E108-F110, A112, N114, T131, and H133-T140 are surface-exposed in the natively folded protein which can promote aggregation without major unfolding event or the conformational change may occur in the oligomers composed of natively folded monomers. The other regions are buried in the native state and their contribution to non-native aggregation is mediated by a preceding unfolding event in the monomeric state of the protein. The structure-based prediction of APRs using SAP tool limited the number of identified APRs to the dynamically-exposed hydrophobic residues including V12, A50, V51, L88, I89, L90, I118, L135, and I139 mediating the native-state aggregation.

Conclusion: Our analysis of APRs in rhKGF identified the regions determining the intrinsic aggregation propensity in both folded (native) and unfolded state of the protein. These regions are the candidate positions for engineering the rhKGF sequence to reduce its aggregation tendency.

1. Background

The biotherapeutics generally undergo a complex series of processing steps including production, harvest, purification, refolding, freeze-thaw, viral clearance, drying, filtration, filling, nebulization, packaging, and shipping to obtain the final product. All of these procedures can impact product stability via different stresses including high concentration, temperature and pH extremes, different ionic strengths, agitation, shear stresses, light, and air-water as well as solid-liquid interfaces. These stresses cause the protein molecules to undergo various physicochemical degradation mechanisms such as oxidation, fragmentation, deamidation, surface adsorption, denaturation, and aggregation [1].

Protein-based pharmaceuticals are typically active as correctly folded monomers composed of one or more protein chains [2]. The most prevalent, yet the least understood degradation pathway, protein aggregation, is a biological phenomenon in which the misfolded proteins accumulate and clump together in an irreversible form [1, 3]. Protein aggregation can reduce the protein shelf-life, *in vivo* half-life, protein binding affinity, elicit immune responses, and hamper the efficacy and safety of biotherapeutics [1, 4].

Protein aggregation is considered as a major challenge in the development and production of protein-based biotherapeutics, because of its potential to result in a product unfit for release and cause technical and economic problems [5, 6]. As such, predicting and mitigating the aggregation of biotherapeutics is of great importance from a pharmaceutical perspective [1]. Understanding of the aggregation mechanisms obtained via computational and experimental discoveries has led to the creation of robust algorithms for prediction of “aggregation-prone regions” (APRs), i.e., the sequence and structural features of proteins facilitating protein aggregation [7]. The APR prediction tools are of great interest to biopharmaceutical research and development, and can be potentially employed in the rational design of therapeutic candidates exhibiting less aggregation propensity, more stability, and solubility while maintaining potency and specificity [3].

One of the most (studied) aggregation-prone biotherapeutics with pharmaceutical applications is the recombinant human keratinocyte growth factor (rhKGF). KGF, also known as FGF-7, is a member of the fibroblast growth factor (FGF) family with the proven therapeutic properties in wound healing [8, 9]. As a potent paracrine-acting epithelial-specific mitogen, KGF acts exclusively through the FGFR2-IIIB spliced variant of FGFR2 expressed by the epithelial cells and is able to protect these cells from various injuries like chemotherapeutic agents used in cancer therapy. A truncated form of KGF known as Palifermin, hereafter referred to as rhKGF is used as the only FDA-approved medication for the treatment of chemotherapy- and radiotherapy-induced oral mucositis [8]. It has shown the ability to reduce the severity, incidence, and duration of oral mucositis in cancer patients [10]. However, rhKGF is a relatively unstable protein denaturing at ~ 37 °C (around the body temperature) which is the lower end of temperature stability range for most of the globular proteins [11]. Low conformational stability of rhKGF is attributed to the repulsion of the positively charged residues in the clusters forming the heparin-binding sites. Following the unfolding event, the high intrinsic aggregation liability of rhKGF which is attributed to the presence of aggregation-promoting hotspots leads to the formation of irreversible aggregation products including particles and precipitates [10].

According to the Lumry and Eyring equation, aggregate formation can be suppressed either through stabilizing the native state or reducing aggregation rate [9]. Various stabilizers including heparin, sulfated polysaccharides like dextran sulfate, anionic polymers, osmolytes like N, N'-dimethylglycine, trehalose, and sucrose, and high concentrations of salts including sodium phosphate, ammonium sulfate, sodium citrate, and sodium chloride were successfully used to reduce rhKGF aggregation tendency via increasing its denaturation temperature and conformational stability. The stabilizing effect of these factors on the rhKGF may be due to the neutralization of the positive charges [12]. The solid-phase PEGylating of rhKGF also significantly improved the stability without affecting its structure [11]. One of the major studies on KGF stability was performed by Amgen Company to generate more stable, yet active analogs of KGF. The most stable analog representing the optimal balance between the stability and activity was the truncated form of KGF lacking the 23 N-terminal amino acids marketed as Palifermin [10]. However, the high aggregation tendency, low storage stability and short plasma half-life of rhKGF has remained a challenge limiting its pharmaceutical applications [13, 14]. To the best of our knowledge, there is no previous study addressing the intrinsic aggregation propensity of rhKGF.

In the current work, we have taken a different approach and attempted to find the aggregation-prone regions in rhKGF by computational methods. We explored the intrinsic aggregation propensity of rhKGF using different sequence-based APR detection tools including TANGO [15], SolubiS [16], ZipperDB [17], AGGRESCAN [18], Zyggregator [19], Camsol [20], PASTA [21], SALSA [22], WALTZ [23], and SODA [24], consensus methods including Amylpred [25], and AMYPDB [26] and structure-based tools including Aggrescan3D [27] and molecular dynamics-based SAP tool [28]. The capability of these tools in the prediction of APRs has previously been shown both for proteins causing protein conformational diseases such as light-chain amyloidosis, prion disease, Alzheimer, and for aggregation-prone therapeutic proteins. These tools were successfully applied for the development of new drug candidates with less aggregation propensities [1, 28].

Our results identified the candidate positions for the rational design of rhKGF to reduce its vulnerability towards aggregation while maintaining (or even increasing) its potency and specificity. Developing the less aggregation-prone variants of rhKGF might reduce the need for stabilizers used in the drug formulation, allow preparation of the drug in the ready-to-use prefilled syringes, increase the drug shelf-life or storage stability, and increase the *in vivo* half-life which can result in the reduction of the required therapeutic doses, adverse effects and treatment costs. This approach might also be applicable to reduce aggregation propensity of the other aggregation-prone therapeutic proteins. Our findings of the aggregation hot spots in rhKGF may be useful for the other FGF family members.

2. Methods

The rhKGF sequence and structure were analyzed using most of the available APR detection tools applying different algorithms to prevent biases from the training sets, parameterization, and the specific characteristics of any given method. A schematic diagram of the study was presented in Fig. 1.

2.1. Retrieval of the rhKGF sequence

To identify the aggregation hotspots in the rhKGF using the sequence-based APR prediction tools, its amino acid sequence was retrieved from the DrugBank database (the accession number DB00039) and the Uniprot ID of P21781-1 except for deletion of 23 N-terminal residues .

2.2. Homology modeling of rhKGF

Since the three-dimensional structure of rhKGF has not yet been determined by experimental techniques, homology modelling (template-based modelling) was performed using GalaxyTBM server to predict its 3D-structure [29]. The server used 1QQL_A as the template which belongs to the Rattus Norvegicus FGF7 crystal structure. The quality of the regenerated models was analyzed using MolProbity structure-validation server (http://molprobity.biochem.duke.edu/) based on the protein geometry properties including the percentage of poor and favored rotamers, Ramachandran favored and outliers, Cβ deviations, bad bonds, bad angles and cis-prolines. The best model was refined using GalaxyRefine2 web server. All-atom contact analysis besides covalent-geometry and torsion-angle criteria were analyzed

using MolProbity for all of them. The less MolProbity score, the highest quality model is achieved. The model validation was also performed using PROCHECK and Verify3D (https://servicesn.mbi.ucla.edu).

2.3. Sequence-based detection of APRs in rhKGF

The amino acid sequence of the rhKGF was analyzed using different sequence-based APR detection tools whose basis of prediction and methodologies are described in the following sections.

2.3.1. TANGO

TANGO applies a statistical mechanics based algorithm for the prediction of protein aggregation nucleating regions as well as the effect of mutations and environmental conditions on the aggregation propensity of these regions [1]. TANGO calculates the cross-beta aggregation in peptides and denatured proteins and considers the propensity of different competing structural states including β-turn, alpha-helix, and β-sheet [1]. For the prediction of each conformation propensity, a partition function is used and cross-beta aggregation is predicted based on the assumption that the core regions involved in the aggregation process are fully buried and satisfy their hydrogen-bond potential. TANGO considers the protein as several overlapping peptides and calculates their aggregation tendency. The environmental conditions like stability, pH, ionic strength, protein concentration and concentration of denaturant TFE are also considered in the TANGO aggregation score. To analyze the rhKGF sequence with TANGO, a text file containing rhKGF sequence as twenty six overlapping peptides (known as the Sup) with the length of ten residues, each one containing five common residues with the next sup was prepared with the default pH, temperature and ionic strength conditions. For the data interpretation, any residue with an aggregation score above 5% over 5–6 residue is a potential aggregation-prone region (APR).

2.3.2. SolubiS

SolubiS is a combined method for identification of APRs with high aggregation propensity and low thermodynamic stability using 3-D structure of the protein. It uses TANGO algorithm and FoldX empirical force field for prediction of β-aggregation-prone regions and protein stability, respectively. SolubiS makes distinction between the APRs that are thermodynamically protected from triggering aggregation by folding (structural APRs) and those occurring in the aggregation-competent conformations that form without major unfolding event (critical APRs). The latter can trigger protein aggregation under near-native conditions [16, 30]. The 3-D structure of rhKGF was submitted to SolubiS web-based tool available at http://solubis.switchlab.org/ and stretch-plots representing the APRs (TANGO summed score) and their contribution to thermodynamic stability (FoldX summed ΔG of each APR), besides mutant aggregation and stability spectrum (MASS) plot showing the changes in aggregation propensity and thermodynamic stability upon mutation of APRs to different gate keepers were created.

2.3.3. AGGRESCAN

AGGRESCAN is a server for the identification of aggregation nucleating segments or “aggregation hot spots” in the polypeptides based on an aggregation-propensity scale obtained from the experimentally

validated hotspots in a group of both natively unfolded and pathogenic proteins like Aβ42, synuclein, prion, amylin and etc. For each protein sequence, AGGRESCAN calculates and presents:

1. An aggregation-propensity value for each amino acid (a^3v), a^3v windows average (a^4v) and a graphical view of the protein aggregation profile (AP)
2. The area of the aggregation profile (Hotspot Area: HSA) above the hotspot threshold (HST) in a given “hot spot” and the graphical representation of the peak area
3. Putative aggregation “hot spots” which are defined as the regions with five or more residues with an a^4v larger than the HST.
4. HSA divided by the number of residues in the entry sequence, known as “Normalized Hotspot Area” (NHSA) per residue [18]

To make APR prediction using AGGRESCAN, the rhKGF sequence was submitted to the AGGRESCAN database available at http://bioinf.uab.es/aggrescan/ and all of the above parameters were calculated.

2.3.4. ZipperDB

ZipperDB predicts protein aggregation by evaluation of fibril-forming propensity of the protein segments using the 3D-profile method. [31]. ZipperDB contains segments with high fibril-forming propensity obtained from the analysis of more than 20,000 protein sequences that can form the “steric zipper” structure consist of two self-complementary beta-sheets creating the spine of an amyloid fibril [32]. Each proline-free hexapeptide from a protein sequence is threaded onto the NNQQNY structure and its energetic fit is calculated by RosettaDesign program. To avoid the problems with disulfide bonds, the Cysteine residues are replaced with Serine during modeling. Based on the experimental amyloid-like fibrils, the energy-threshold of -23 kcal/mol was determined. The sequences with the energy level equal to or below the − 23 kcal/mol are defined as segments with high fibril-forming propensity [17]. To analyze the rhKGF (*Palifermin*) with the ZipperDB, its sequence was submitted to ZipperDB web interface (https://services.mbi.ucla.edu/zipperdb/intro) and a graphical representation of its fibril-formation propensity besides the following data were provided for each peptide segment (The preferences presented below are related to the fibril-forming tendency):

1. Rosetta energy of one layer composed of two beta-strands (lower energies are preferred)
2. Shape complementarity of the steric zipper interface (range 0–1, higher values are preferred)
3. Area of Interface: The solvent-accessible surface area (SASA) at the steric zipper interface per layer (higher values are preferred)
4. The contact area between two sheets (the larger area is preferred)
5. The SASA of the five-layer structure (lower values are preferred)
6. The composite score obtained from the combination of Rosetta energy, shape complementarity and area of the interface (lower values are preferred)

2.3.5. Zyggregator

Zyggregator is also a sequence-based method which calculates different propensities including the local stability of the monomeric state (ln *P*i), formation of β–rich oligomers (Z_i^{tox}), and the formation of fibrillar aggregates (Z_i^{agg}) to predict the intrinsic amyloid aggregation propensity of the protein. The Z_i^{agg} score 0 is related to the aggregation propensity equal to that of a random sequence at position *i*, and when it is one standard deviation more aggregation-prone the score is 1. The Z_i^{agg} score 1 or more is related to the aggregation-prone residues [19].

To analyze the rhKGF with zyggregator, its sequence was submitted to the zyggregator database available at http://www-mvsoftware.ch.cam.ac.uk/index.php/zyggregator and the above parameters were calculated for it. Then a graphical representation based on the residues Z_i^{agg} score was prepared.

2.3.6. CamSol

CamSol consists of two algorithms for the calculation of protein solubility. The first algorithm, known as CamSolintrinsic, is a sequence-based method for the prediction of intrinsic solubility of the protein unfolded state. The rhKGF sequence was used in CamSolintrinsic to analyze its intrinsic solubility. The second one is CamSol structurally corrected which applies a PDB file as input and provides structural correction to the intrinsic solubility and calculates the protein solubility based on the amino acid proximity in the structure and regarding their solvent exposure. The rhKGF PDB prepared by homology modeling was used as the input file for CamSol structural analysis (http://www-vendruscolo.ch.cam.ac.uk/camsolmethod.html).

2.3.7. WALTZ

WALTZ is a computer algorithm for the amyloid prediction in proteins which is capable of distinguishing the true amyloids from the amorphous aggregates. WALTZ is also able to identify hard-to-predict amyloid aggregates like those involved in prion disease [23]. WALTZ has a three-component scoring function: a position-specific scoring matrix (PSSM) for hexapeptides to identify amyloid-forming regions using the AmylHex dataset which is supplemented with a large number of experimentally determined amyloidogenic and non-amyloidogenic sequences, the physicochemical properties of amino acids including hydrophobicity and β-structure forming propensity, and a position-specific pseudoenergy matrix determined from the crystal structure of Sup35 GNNQQNY peptide which forms the cross-β spine of amyloid fibrills to estimate the relative energies using FoldX [33, 34].

The amyloid-forming peptides can be predicted with each of the three existing thresholds in WALTZ database including “Best overall performance”, “High specificity”, and “High sensitivity” or with the customized threshold defined by the user. To analyze the rhKGF using WALTZ, its amino acid sequence in FASTA format was submitted to the online WALTZ web server (http://waltz.switchlab.org/) and APR prediction was made with different thresholds.

2.3.8. PASTA

PASTA is a sequence-based server to identify the propensity of different protein sequences for amyloid aggregation, the effect of mutations on the aggregation rate, the aggregation hotspots establishing the pathological conditions in amyloidosis, and the portions stabilizing the cross-beta core of fibrillar aggregates. PASTA also provides information about the protein intrinsic disorder and secondary structure which complement the aggregation prediction [21]. To analyze the rhKGF, its amino acid sequence was submitted to PASTA server available at http://protein.bio.unipd.it/pasta2/.

2.3.9. SODA

SODA can estimate the protein solubility changes based on the protein physico-chemical properties like sequence-based aggregation propensity, intrinsic disorder, plus hydrophobicity and the secondary structure preferences. SODA applies PASTA aggregation propensity, ESpritz intrinsic disorder scores, Kyte and Doolittle hydrophobicity profile and FESS secondary structure (α-helix and β-strand) propensity to predict changes of protein solubility based on the physicochemical properties [24]. Furthermore, SODA can predict the mutations increasing or decreasing the solubility of the protein at all of the positions and also the effect of mutations created by the user on the protein solubility [24]. The SODA input should be in the FASTA or PDB format and the output is the graphical solubility profile of the wild type protein and the effects of variations on protein solubility. The server is available at http://protein.bio.unipd.it/soda/. The SODA server was used to determine the low solubility regions of the rhKGF.

2.3.10. AMYLPRED and AMYLPRED2

AMYLPRED2 is an improved version of AmylPred web tool which utilizes a consensus of various strategies that have been found or explicitly developed to foresee the features related to the formation of amyloid fibrils. The consensus of these strategies is defined as the hit overlap of at least half of the selected methods [35]. The methods employed by the AMYLPRED2 server to make the consensus predictions are AGGRESCAN, AmyloidMutants, Amyloidogenic Pattern, Average Packing Density, Beta-strand contiguity, Hexapeptide Conformational Energy, NetCSSP, Pafig, SecStr (Possible Conformational Switches), TANGO, and WALTZ. Since the AmylPred2 server was not accessible for a long time, to make a consensus prediction of APRs in rhKGF, the AmylPred server was used which makes consensus prediction of five methods including Average Packing Density, Possible Conformational Switches, Amyloidogenic Pattern, TANGO, and Hexapeptide Conformational Energy and is available at http://aias.biol.uoa.gr/AMYLPRED/. In the Amylpred, the hits are defined as the overlap of two out of five methods.

2.3.11. AMYPDB

AMYPDB as a comprehensive amyloid protein database provides a pattern discovery and analysis tool enabling us to detect regions of significance in amyloid formation in any protein query. AMYPDB is also able to make a consensus prediction of APRs using different APR detection tools including Aggrescan, PASTA, SALSA, Pafig, Fold-amyloid, and TANGO [26]. AMYPDB server, available at http://amypdb.genouest.org/, was also used to make a consensus prediction of APRs in rhKGF.

2.4. Structure-based detection of APRs in rhKGF

2.4.1. Aggrescan3D

Aggrescan3D is a structure-based APR prediction approach developed to detect surface-exposed spatially-adjacent APRs involved in the aggregation of natively folded proteins in two static and dynamic states based on both amino acids position, and structure. Aggrescan3D server available at http://biocomp.chem.uw.edu.pl/A3D/ was used to identify APRs in the rhKGF at the radius of 5 Å.

2.4.2. Spatial Aggregation Propensity (SAP)

The hydrophobic residues generally participate in the formation of the hydrophobic core of the proteins, but they will form aggregation hotspots when they cluster together on the protein surfaces. The spatial aggregation propensity algorithm, briefly known as “SAP”, predicts protein structural regions which can form aggregates by making hydrophobic patches on the protein surface. Utilizing molecular dynamics (MD) simulation, SAP can also simulate the size change of the patches under physiological conditions. SAP identifies dynamically exposed hydrophobicity of the certain patches on the surface of the protein.

The SAP calculation procedure yields SAP scores for each atom, average SAP score for each residue and a SAP_mapped PDB file in which the residues are color-scaled based on the measured SAP values. A SAP value is calculated and assigned to each residue based on the sum of the effective hydrophobicity (*Φeff*) of the neighboring residues with a radius R. The effective hydrophobicity of each residue depends on both its intrinsic hydrophobicity and surface exposure (solvent accessible area; SAA). The aggregation propensity of each residue is calculated from the SAP values for each atom, SAP_{atom} [1, 7]:

$$SAP_{atom\ i} = \sum_{\substack{\text{Simulation} \\ \text{average}}} \sum_{\substack{\text{Residues with at least one side chain atom within } R \text{ from atom } i}} \left(\frac{\substack{\text{SAA of side chain} \\ \text{atoms within radius } R}}{\substack{\text{SAA of side chain atoms} \\ \text{of fully exposed residue}}} \times Hydrophobicity_i \right)$$

The positive SAP scores indicate the aggregation-prone regions. Whereas measuring the SAP scores at low resolution or high SAP radius (> 10 Å) can find the large aggregation prone patches and does not provide a detailed view of these patches, the SAP values at high resolution or low radius (5 Å) can identify the sites to be mutated to mitigate protein aggregation. When the SAP values are measured at 5 Å, the residues with the scores above 0.15 are considered as hits [5].

2.4.2.1. Molecular dynamics simulation of rhKGF

Molecular dynamics (MD) simulation is a powerful tool for the characterization of the atomic-level behavior of the biomolecules like the protein motions and conformational changes in different physiological states [5]. The MD simulation was conducted with the 3-D structure of rhKGF obtained from

homology modeling process using GROMACS software package (version 5.1.4). The rhKGF structure was solvated in a cubic box containing the transferable intermolecular potential with 3 points (TIP3P) water molecules with 10 Å distance between the protein and the edges of the solvation box. The solvated system was neutralized by adding Na^+ and Cl^- ions. Then, the system was energy minimized in 1000 steps of steepest descent, and equilibrated in the canonical (NVT) and isothermal–isobaric (NPT) ensembles with the leapfrog algorithm and an integration time step of 0.002 picosecond (ps). The production run was performed for 100 ns.

2.4.2.2. Identification of the rhKGF dynamic SAP scores

To measure the dynamic SAP scores of the rhKGF residues, the MD trajectory was converted into a pdb file containing 100 pdb(s) generated with the time interval of 1 ns (1000 ps) using GROMACS software. Then, the SAP values were computed at 5 Å and 10 Å for all 100 pdb files obtained from the MD trajectory based on the SAP calculation procedure developed by *Dr. Naresh Chennamsetty* using the CHARMM simulation program. Finally, the average SAP score of each residue was computed over the 100 ns MD production run. The SAP scores were also measured for the rhKGF 3D structure obtained from homology modelling in the static state.

3. Results

3.1. Retrieval of rhKGF sequence and structure:

The sequence for rhKGF (Palifermin) in the DrugBank was 140 amino acids in length without the first methionine, consequently all of the analyses were performed with this sequence. In homology modeling using Galaxy web server, the refined model with the MolProbity score of 0.54 was chosen as the best one. MolProbity score is a log-weighted combination of clash score, percentage of Ramachandran outliers, and bad rotamers that gives a number reflecting the crystallographic resolution that those values are expected.

3.2. Sequence-based detection of APRs in rhKGF:

Analysis of rhKGF using TANGO identified three segments as potential APRs with the aggregation scores above the threshold (5%) including residues 48–56: TVAVGIVAI, 63–67: FYLAM, and 109–114: MFVALN which can participate in cross-β aggregation (Fig. 2).

The stretch-plot of rhKGF created by SolubiS representing the aggregation propensity and local stability of APRs showed that from the three APRs identified by TANGO, the regions 48–56, and 63–67 have positive summed ΔG and were defined as critical APRs and the region 109–114 with negative SolubiS score was defined as structural APR (Fig. 2).

AGGRESCAN recognized three aggregation hotspots in rhKGF including residues 43–58, 61–66, and 108–113 (Fig. 3).

The fibril-forming propensity analysis of the rhKGF using the 3D profile method showed that the residues G29, T48-V54, G58, Y64, H93, T96, A98, E108, F110, and Q129 have Rosetta energy below − 23 Kcal/mol, and therefore are identified as amyloid-prone regions (Fig. 4).

Zyggregator APR prediction results identified Y2, R18, T19, I43, V49-V54, I56, Y64, L65, D82, Y94-Y97, S99, A100, T103-N105, A112, and T140 with the Z_i^{agg} score of one or higher as aggregation-prone residues in rhKGF (Fig. 5).

The intrinsic solubility profile of the rhKGF prepared by CamSol identified the residues R18, T19, Q20, Y22, V49-V54, I56, A66, F110-A112, F134-A138, and T140 as poorly soluble amino acids in rhKGF (Fig. 6). According to the Camsol structurally-corrected solubility profile, the residues V51, G52, and T140 have structurally-corrected solubility scores of -1.97, -1.05 and − 1.43, respectively (Fig. 7).

Table 1
The amyloidogenic regions identified in *Palifermin* using the best overall performance threshold of WALTZ.

Best Overall Performance		
Positions	**Sequence**	**Average score per residue**
18–22	RTQWY	94.98
39–47	NNYNIMEIR	94.98
52–57	GIVAIK	96.98
93–98	HYNTYA	96.32

According to PASTA aggregation and disorder profile, aggregation free energy profile and aggregation-pairing matrix (Fig. 9), the portion 40–60, owing to the lowest aggregation free energy, is the most amyloid- stabilizing region responsible for rhKGF amyloid-forming propensity.

According to the SODA solubility profile prepared for the rhKGF (Fig. 10), the lowest solubility regions are located at positions 40–60, 87–89, and 108–112 (Fig. 10).

The result of consensus APR prediction in rhKGF using Amylpred server is presented in Fig. 11. The prediction is based on the overlap of at least two out of five methods. The consensus aggregation hits identified by Amylpred in rhKGF are the regions 15–25, 43–56, 63–67, 88–91, 109–114, and 133–139.

According to AMYPDB APR prediction results presented in Fig. 12, the concensus APRs in rhKGF are the regions RVRR (11–14), NIMEIRTVAVGIVAIK (42–57), and MFVA (109–112).

3.3. Structure-based detection of APRs in rhKGF:

Structure-based prediction of APRs using Aggrescan3D identified residues Y2, M5, I10, V51, V59, Y74, L88, I89, L90, and I139 as spatially-adjacent APRs involved in the aggregation of rhKGF (Fig. 13).

4. Discussion

As previously mentioned, high aggregation tendency of rhKGF, manifested by loss of the monomeric state and accumulation of the aggregated species at moderate temperatures, has made it a fairly unstable protein with limited pharmaceutical applications. Optimization of the formulation/storage condition e.g. using different stabilizers (heparin), osmolytes (like trehalose and sucrose) and salts (like ammonium sulfate, and sodium chloride) applied for KGF [12] and other FGFs [36] may be implemented for in vitro applications, but in vivo applications are limited to the naturally occurring solution environment. For the latter, aggregation management can be obtained by designing aggregation-resistant variants. For example, stable mutants of FGF2 (bFGF) were designed by sequence alignment with FGF1 stabilized mutants. Furthermore, the structural free energy analysis to enhance conformational stability of FGF2 have resulted in nine point mutants with increased in vitro functional half-life from 10 hours to more than 20 days at 37 °C [36]. Different groups of mutations including deletion, cysteine substitution, charge substitution, and combined mutations were also assessed on native KGF and the D23 analog with ~ 2 fold increase in half-life to 1.2 days with optimal activity was approved by FDA for clinical applications [10]. Increasing the protein conformational stability protects the protein from unfolding and the APRs buried in the hydrophobic core of the protein from triggering aggregation and therefore reduces the protein aggregation tendency [16]. Characterization of the sequence/structural APRs in proteins may help mitigate aggregation in biotherapeutics via engineering the surface-exposed APRs (either in the native or unfolded state) and selection of drug candidates with reduced aggregation tendency [33].

A number of computational methods have been developed to identify the sequence aggregation-prone regions in the amyloidogenic proteins. Here, we tried to analyze the rhKGF with most of the available tools applying different algorithms to prevent biases from the training sets, parametrization, and the specific characteristics of any given method. This approach helped us to make a prediction of APRs in rhKGF with higher confidence. We also measured the dynamic exposure of the hydrophobic patches in the rhKGF using SAP tool to identify structural APRs mediating native-state aggregation.

TANGO aggregation propensity scores are based on the beta-sheet formation. Therefore, the regions predicted as APR by TANGO can participate in cross-β aggregation of rhKGF [1, 37, 38]. Previous studies recognized several protein conformations as the seed of protein aggregates at the molecular level. The most studied conformational seed is the cross β-motif observed in the amyloid-like fibrils. The cross β-motifs composed of intermolecular β-sheets form the spine of amyloids aggregation as highly ordered aggregates [32]. Siepen et al proved that the exposed side β-strands may also give rise to amyloids by forming the cross β-motifs as seen in transthyretin and β_2-microglobulin [39]. Moreover, Dobson's group showed that fibril formation is the general properties of proteins under stress conditions [40]. The rhKGF is also a β-rich protein motivating us to explore if the cross β-motif conformation promotes its aggregation. Detection of these motifs in rhKGF showed that its aggregation may proceed by cross- β aggregation. The existence of one aggregation-prone motif can be sufficient for cross β-motif-derived aggregation, though it does not necessarily guarantee that the protein aggregation proceeds by this mechanism. Generally, the sequence motifs involved in protein aggregation are rich in hydrophobic (like

Gly, Ala, Val, Leu, Ile, and etc) or aromatic (Phe, Tyr, Trp) residues. The sequence motifs known to be involved in the aggregation of prion or amyloidal proteins have also contained Aspargine (N) / Glutamine (Q) residues. However, charged residues (Arg, Lys, His, Asp and Glu) are rarely observed in such regions [41].

The common sequence-based APR prediction tools identify the APRs in the primary sequence and consider the protein in the unfolded state. SolubiS combines the TANGO APR prediction results with FoldX free energy estimation to make distinction between the structural and critical APRs. Among the APRs detected by TANGO, the regions 48–56, and 63–67 positive summed ΔG were characterized as critical APRs which are the able to mediate aggregation under native conditions. The least stable APR with the highest TANGO score is the region 48–56 which is also involved in the receptor binding and it can be problematic when the protein is not engaged in the protein-receptor interaction. Previous studies showed that mutation of the critical APRs can reduce the overall aggregation tendency and increased amount of soluble protein produced in mammalian cells [16]. The region 109–114 has the lowest SolubiS score due to its high contribution to local thermodynamic stability and therefore is considered as a structural APR which is naturally protected from triggering aggregation by folding. Increasing the conformational stability can protect such APRs from exposure and contact with the similar sequences on the neighboring molecules and prevents non-native aggregation [16, 33].

AGGRESCAN recognized three aggregation hotspots in rhKGF including residues 43–58, 61–66, and 108–113. The APRs predicted by AGGRESCAN were highly similar to those predicted by TANGO. The similarity was expected as both tools use the same approach in the characterization of APRs i.e., predicting β-aggregation prone regions, but not necessarily amyloid ones. Based on the AGGRESCAN algorithm these regions are sequentially contiguous regions which mediate protein aggregation after exposure in at least partially unfolded state of the protein, and the aggregation propensity of these sequences is determined by their amino acid composition. AGGRESCAN predicts APRs in the input polypeptide sequence by identification of the protein fragments which were experimentally proved to be involved in the aggregation of disease-linked proteins. AGGRESCAN is also able to detect the mutation effects on protein aggregation tendency as a function of protein sequence composition [18].

According to the 3D profile approach implemented by ZipperDB and by threading the sequence of rhKGF on the crystal structure of NNQQNY peptide of the Saccharomyces cerevisiae sup35 prion protein, the regions identified by ZipperDB method, can form the cross-beta spine of amyloid-like fibrils [42]. Most of the APRs predicted by ZipperDB are common with those identified by TANGO and AGGRESCAN except for G29, H93, T96, A98, E108 and Q129. ZipperDB is unique regarding its prediction approach since it makes use of the structural information to find the probability of fibril formation for a particular sequence.

Most of the APRs detected by Zyggregator were common with the other APR detection tools, and the most conserved aggregation-prone motif is VAVGIV (V49-V54). The Zyggregator method predicts the possibility of the fibril or protofibril structure formation and amyloid aggregation propensity based on the combination of physico-chemical properties of the amino acid content of the protein including

hydrophobicity, α-helical and β-sheet propensity, hydrophilic/hydrophobic patterns and net charge of the polypeptide. Zyggregator also evaluates the flanking residues (“gatekeepers”) of a sliding window regarding the presence of charged residues which may decrease aggregation. Like TANGO, Zyggregator also considers the effect of physicochemical conditions including pH, temperature, ionic strength, and concentration of trifluorethanol on aggregation. Zyggregator is also able to predict the local instabilities of the structured proteins resulting in aggregation using the CamP program [33, 40].

The intrinsic solubility profile of the rhKGF prepared by CamSol identified the poorly soluble and aggregation-prone residues in rhKGF. The most conserved aggregation-prone motifs between CamSol and other APR detection tools are VAVGIV (V49-V54) and FVA (F110-A112). Camsol structurally-corrected solubility score which is based on the proximity of the amino acids in the 3D structure and their solvent exposure, identified three residues including V51, G52, and T140. These residues were identified as poorly soluble amino acids which are the most important candidate positions for designing variants of rhKGF with enhanced solubility. CamSol, as a protein solubility predictor, is also able to provide predictions of protein aggregation propensity, as opposed to the other algorithms which may only focus on amyloid formation. CamSol employs the variables similar to those implemented in Zyggregator with the difference in the parameters and definition of the gatekeeping effect.

Most of the amyloidogenic regions predicted by WALTZ were common with the other APR prediction tools. The “RTQWY” region was also predicted as APR by CamSol, Zyggregator, AmylPred, SALSA, PAFIG, and Fold-Amyloid. Most of the residues in the “NNYNIMEIR” motif were also observed in APRs defined by TANGO, AGGRESCAN, AMYLPRED, PASTA, SALSA, PAFIG, and Fold-Amyloid. The “GIVAIK” motif is also a common APR predicted by most of the APR detection tools like AGGRESCAN, ZipperDB, Zyggregator, CamSol, AmylPred, PASTA, SALSA, PAFIG, Fold-Amyloid, AMYPDB, and some of these residues were also categorized as high SAP-valued residues. The motif “MFVALN” was also a common APR defined by TANGO, AGGRESCAN, CamSol, AmylPred, SALA, PAFIG, Fold-Amyloid, and AMYPDB. The least conserved APR defined by WALTZ is “HYNTYA” which has only a few common residues with those detected by ZipperDB, Zyggregator and SALSA, but not with the other tools. WALTZ algorithm have been trained by a large set of experimentally validated amyloid-forming peptides with eighty percent of the data provided has been validated with technologies like circular dichroism, electron microscopy and infrared spectroscopy, Fourier-transform infrared spectroscopy (FTIR) and Thioflavin-T binding assays. It is very important for a computerized tool that its data have been validated by the experimental methods. No other existing amyloid prediction tools has such a high validation rate [43].

According to PASTA prediction results, the portion 40–60, owing to the lowest aggregation free energy, is the most amyloid- stabilizing region responsible for rhKGF amyloid-forming propensity. Furthermore, this area has the highest pairing probability which is responsible for protein self-association. These results are also consistent with the previous APR prediction results provided by other tools such as TANGO, Zyggregator, ZipperDB, CamSol, SALSA, WALTZ, AmylPred, AMYPDB, and Fold-Amyloid. Some of the residues in this area including A50 and V51 have also been identified as hydrophobic exposed residues with the SAP scores over the threshold (0.15) at 5 Å. PASTA makes aggregation predictions based on the

assumption that mechanisms governing the native β-sheet formation, are also responsible for the formation of β-sheets in amyloid aggregates [33]. The peptides with the energy level below a specified threshold are defined as aggregation hotspots.

According to the SODA solubility profile, the lowest solubility regions of rhKGF are located at positions 40–60, 87–89, and 108–112 which are mainly consistent with the APRs identified with the other tools. These regions can be considered as the candidate positions to create the mutants with reduced aggregation propensity. SODA predicts protein solubility based on disorder and aggregation which both of them are affected by the physicochemical properties of the amino acids including hydrophobicity, secondary structure preferences, and charge. As the highly dynamical disordered regions of protein can increase its aggreagation propensity, the intrinsic disorder is a major determinant of protein aggregation [44].

AMYLPRED and AMYPDB are also referred as metapredictors due to their analogy to the metasearch engines. The rational in favor of consensus strategies is that different strategies consider different contributors in the estimation of protein aggregation propensity. Since it is still unknown that which contributors are most important, and various algorithms weigh these contributors differently, using metapredictors in which their members complement each other may enhance prediction accuracy. It is also more likely that applying consensus methods with the algorithms trained with different data sets may result in less false positive results [25, 33]. The APRs identified by Amylpred and AMYPDB as two consensus methods comprising different algorithms may be considered as representative of the result of other sequence-based APR prediction tools.

Amylpred has been useful to identify the amyloid-forming regions involved in the development of conformational diseases known as amyloidosis like Alzheimer's, Parkinson's, type II diabetes, and prion disease. Amylpred also enables us to identify the properties of protein folding/misfolding and to control the aggregation/solubility of biopharmaceuticals in biotechnology industry [35].

Like Amylpred, AMYPDB is a consensus method to predict APRs in proteins except for some differences in their members. AMYPDB as a comprehensive amyloid protein database has been dedicated to the accumulation of the sequence and structure of the amyloidogenic protein families associated with several pathologic conditions such as Alzheimer's disease, prion disease, type II diabetes mellitus, Parkinson, Huntington, and Creutzfeldt-Jakob. AMYPDB also provides the amyloid-related sequence signatures and patterns matching each amyloid protein family and provides a pattern discovery and analysis tool enabling us to detect regions of significance in amyloid formation in any protein query [26].

Aggrescan3D exhibited higher accuracy compared to the sequence-based algorithms in forecasting the aggregation hot spots of the globular proteins [27]. Prediction of APRs with Aggrescan3D as the optimized version of Aggrescan which is able to overcome the limitations of sequence-based tools, identified the spatially-adjacent APRs involved in the aggregation of rhKGF from the native state.

The Dynamic SAP calculation has identified nine residues with the SAP scores above 0.15 at R = 5 Å on the surface of the rhKGF which are considered as potential APRs and are the candidate positions for mutations against aggregation. Except for the residues which are involved in heparin or receptor binding, and essential for KGF activity, the other residues with the SAP scores greater than 0.15 are suitable positions which can be mutated to the hydrophilic or less hydrophobic residues in order to reduce rhKGF aggregation tendency while maintaining its biological activity.

The hydrophobic residues are usually buried inside the core of the protein increasing the protein stability. Exposure of the hydrophobic residues under the normal dynamic fluctuations may combine them with the surface residues and lead to the formation of larger surface-exposed hydrophobic patches [28]. The SAP algorithm have been designed to identify these dynamically exposed hydrophobic patches on the surface of the protein which their interactions with the other hydrophobic patches on the neighboring molecules in the native state of the protein lead to the accumulation and aggregation of the protein molecules [1, 7]. SAP considers both dynamic exposure and spatial proximity of each residue in the protein tertiary structure and is also applicable for large biotherapeutics like antibodies. In the SAP algorithm like Aggrescan3D, the unfolding event is not considered as a prerequisite of aggregation, and APRs are identified in the native state of the protein [28].As the hydrophobic interactions have been shown to be the predominant mechanism of aggregation, identification of the dynamically exposed hydrophobic patches of the protein with the SAP tool, and mutation of the high SAP residues; indicators of aggregation hot spots, to the residues with the less SAP scores can reduce protein aggregation tendency [39].

For a given aggregation-prone region to promote aggregation, it should have a high intrinsic aggregation propensity, be surface-exposed or become exposed after conformational transition like complete/partial unfolding or misfolding events to facilitate intermolecular interactions. Therefore, the presence of sequence/structural APRs may be essential but is not sufficient for protein aggregation [28].

The sequence-based APR prediction tools discussed above, are only capable of distinguishing sequentially contiguous residues as APR in small proteins. These tools generally assume that an unfolding event is required to expose and make the generally hydrophobic hot spots with a high β-sheet forming propensity available for contacting the analogous sequences on the neighboring molecules. The detected APRs are solvent exposed either natively or after an unfolding event. This allows stabilizing contacts which usually tend to form interprotein β-sheet structures converting the aggregates to net irreversible structures [33]. Because these tools use primary sequence to predict APRs, they are not able to distinguish non-contiguous surface-exposed APRs from the regions which are buried inside the natively-folded globular proteins [27]. Non-native aggregation as a prevalent challenge in biopharmaceutical industry is defined as the process of forming aggregates composed of fully/partially unfolded or misfolded protein monomers. Hence, one proposed solution to reduce protein aggregation is to enhance the conformational stability of the protein by increasing the free energy of unfolding (ΔGunf) [33]. Using SolubiS and structure-based APR detection tools enabled us to make distinction between APRs mediating native and non-native aggregation. Aggrescan3D (A3D) and SAP, as two structure-based APR prediction tools, are capable of detecting non-contiguous surface-exposed APRs in the large natively folded globular

proteins like monoclonal antibodies, and therefore recognizing APRs involved in the native-state aggregation [27, 28].

5. Conclusion

The aggregation propensity of rhKGF was analyzed using both sequence and structure-based APR predictors. The APRs identified in rhKGF by different computational tools have been shown in Fig. 15a. Accordingly, the high intrinsic aggregation propensity of rhKGF is mainly mediated by the amino acids located at positions 10–30, 40–60, 61–66, 88–120, and 130–140. The aggregation-prone motifs found in the rhKGF were rich in the hydrophobic residues including Gly, Ala, Val, Leu, and Ile, and aromatic residues like Phe, Tyr and Trp which have high β-sheet forming propensity. Some motifs also contained Asn (N) and Gln (Q) which have been demonstrated to be involved in the formation of amyloid aggregates. Charged residues have been rarely observed in these motifs. The APRs detected by the sequence-based APR detection tools might be buried inside the protein and only become exposed after an unfolding event or be surface exposed in the native state. Infact, mapping these APRs on the native structure of the rhKGF revealed that most of them are solvent-exposed in the natively folded protein including F16-R25, I43, E45, R47-I56, F61, Y62, N66, L88-E91, E108-F110, A112, N114, T131, and H133-T140 (Fig. 15b). These regions may promote aggregation of the rhKGF without the occurance of a significant unfolding event preceding the protein aggregation or the conformational change may occur within the oligomers composed of the folded monomers (aggregation-competent conformations) [33]. Some of these regions contribute toward receptor/heparin binding. As a result, mutation of these residues may disrupt the KGF biological activity [45, 46]. This is a significant challenge to design variants of biotherapeutics with low vulnerability towards aggregation and unchanged biological activity. The other regions are buried in the native state and their contribution to non-native aggregation is mediated by a preceding unfolding event in the monomeric state of the protein. Increasing the protein conformational stability can prevent these regions from triggering aggregation. Limiting the rate of aggregation can also be achieved via mutation of the APRs to less aggregation-prone residues. The structure-based predictions made with the Aggrescan3D and SAP limited the number of identified APRs to the dynamically-exposed hydrophobic residues including V12, A50, V51, L88, I89, L90, I118, L135, and I139 mediating the native-state aggregation. Except for the functional residues, the others can be substituted with either hydrophilic or even less hydrophobic residues to design variants with reduced aggregation propensity. Furthermore, the structure-based rational design (e.g. mutation of high SAP-valued residues which are not involved in receptor/heparin binding) as a part of QbD approach may also help reducing the total aggregation propensity without compromising the biological activity.

Abbreviations

rhKGF
Recombinant Human Keratinocyte Growth Factor
APR

Aggregation-Prone Region
MD
Molecular Dynamics
SAP
Spatial Aggregation Propensity

Declarations

- Ethics approval and consent to participate: Not Applicable.

- Consent for publication: Not Applicable.

- Availability of data and materials: All data generated or analysed during this study are included in this published article.

- Competing Interests: The authors declare that they have no competing interests.

- Funding: This research was financially supported by a grant from Pasteur Institute of Iran (for the Ph.D. thesis of Mansoureh Shahbazi Dastjerdeh, grant number. BP-9368). The molecular dynamics simulations and data analysis were performed by the computer servers provided by the Pasteur Institute of Iran.

- Authors' Contributions: Design of study was performed by Mansoureh Shahbazi Dastjerdeh, and Dr. Hamzeh Rahimi. The computational studies and data analysis was performed by Mansoureh Shahbazi Dastjerdeh under the supervision of Dr. Hamzeh Rahimi, and Dr. Majid Golkar. Manuscript writing was mainly performed by Mansoureh Shahbazi Dastjerdeh and it was revised by Dr. Majid Golkar, Hamzeh Rahimi, and Dr. Mohammad Ali Shokrgozar.

- Acknowledgements: We are deeply grateful to Pasteur Institute of Iran due to their support. We are also deeply thankful to Dr. Naresh Chennamsetty due to his technical guidance for SAP calculation.

References

1. Buck PM, et al., *Computational methods to predict therapeutic protein aggregation*, in *Therapeutic Proteins*. 2012, Springer. p. 425–451.
2. Roberts C. J.J.C.o.i.b. Protein aggregation its impact on product quality. 2014;30:211–7.
3. Kumar S, Wang X, and S.K.J.A.o.t.p. Singh, *Identification and impact of aggregation-prone regions in proteins and therapeutic monoclonal antibodies.* 2010: p. 103–118.
4. Wang W, Roberts CJ, *Aggregation of therapeutic proteins*. Vol. 100. 2010: Wiley Online Library.
5. Courtois F, et al. *Rational design of therapeutic mAbs against aggregation through protein engineering and incorporation of glycosylation motifs applied to bevacizumab*. in *MAbs*. 2016. Taylor & Francis.

6. Fink ALJF, design. *Protein aggregation: folding aggregates, inclusion bodies and amyloid*. 1998. 3(1): p. R9-R23.
7. Lee CC, Perchiacca JM. and P.M.J.T.i.b. Tessier. Toward aggregation-resistant antibodies by design. 2013;31(11):612–20.
8. Dastjerdeh MS, et al., *In silico analysis of different signal peptides for the secretory production of recombinant human keratinocyte growth factor in Escherichia coli*. 2019.
9. Rajan RS, Li T. and T.J.A.o.T.P. Arakawa, *Case studies involving protein aggregation.* 2010: p. 367–401.
10. Hsu E, et al., *Enhanced stability of recombinant keratinocyte growth factor by mutagenesis*. 2006. 19(4): p. 147–153.
11. Huang Z, et al., *A novel solid-phase site-specific PEGylation enhances the in vitro and in vivo biostabilty of recombinant human keratinocyte growth factor 1*. 2012. 7(5).
12. Chen BL, et al., *Strategies to suppress aggregation of recombinant keratinocyte growth factor during liquid formulation development*. 1994. 83(12): p. 1657–1661.
13. Kalhor HRJJoC, Research M, *Expression of the Full-length Human Recombinant Keratinocyte Growth Factor in Pichia pastoris.* 2016. **8**(1): p. 1–7.
14. Poorebrahim M, et al. *In* silico enhancement of the stability activity of keratinocyte growth factor. 2017;418:111–21.
15. Rousseau F, Schymkowitz J. and L.J.C.o.i.s.b. Serrano. Protein aggregation amyloidosis: confusion of the kinds? 2006;16(1):118–26.
16. van der Kant R, et al. Prediction reduction of the aggregation of monoclonal antibodies. 2017;429(8):1244–61.
17. Sawaya MR, et al., *Atomic structures of amyloid cross-β spines reveal varied steric zippers.* 2007. **447**(7143): p. 453.
18. Conchillo-Solé O, et al., *AGGRESCAN: a server for the prediction and evaluation of" hot spots" of aggregation in polypeptides.* 2007. **8**(1): p. 65.
19. Tartaglia GG, Vendruscolo MJCSR. The Zyggregator method for predicting protein aggregation propensities. 2008;37(7):1395–401.
20. Sormanni P, Aprile FA. and M.J.J.o.m.b. Vendruscolo. The CamSol method of rational design of protein mutants with enhanced solubility. 2015;427(2):478–90.
21. Walsh I, et al., *PASTA 2.0: an improved server for protein aggregation prediction*. 2014. 42(W1): p. W301-W307.
22. Zibaee S, et al., *A simple algorithm locates β-strands in the amyloid fibril core of α-synuclein, Aβ, and tau using the amino acid sequence alone*. 2007. 16(5): p. 906–918.
23. Oliveberg MJNm. *Waltz, an exciting new move in amyloid prediction*. 2010. 7(3): p. 187.
24. Paladin L, Piovesan D, and S.C.J.N.a.r. Tosatto, *SODA: prediction of protein solubility from disorder and aggregation propensity.* 2017. **45**(W1): p. W236-W240.

25. Tsolis AC, et al., *A consensus method for the prediction of 'aggregation-prone'peptides in globular proteins.* 2013. **8**(1).
26. Pawlicki S, Le A, Béchec. and C.J.B.b. Delamarche, *AMYPdb: a database dedicated to amyloid precursor proteins*. 2008. 9(1): p. 273.
27. Zambrano R, et al., *AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures*. 2015. 43(W1): p. W306-W313.
28. Chennamsetty N, et al. Prediction of aggregation prone regions of therapeutic proteins. 2010;114(19):6614–24.
29. Ko J, Park H. and C.J.B.b. Seok, *GalaxyTBM: template-based modeling by building a reliable core and refining unreliable local regions*. 2012. 13(1): p. 198.
30. van der Kant R, et al., *Solubis: Optimizing protein solubility by minimal point mutations*, in *Protein Misfolding Diseases*. 2019, Springer. p. 317–333.
31. Kuhlman B. and D.J.P.o.t.N.A.o.S. Baker, *Native protein sequences are close to optimal for their structures*. 2000. 97(19): p. 10383–10388.
32. Nelson R, et al., *Structure of the cross-β spine of amyloid-like fibrils*. 2005. 435(7043): p. 773.
33. Meric G, et al., *Driving forces for nonnative protein aggregation and approaches to predict aggregation-prone regions*. 2017. 8: p. 139–159.
34. Ahmed AB, Kajava AVJFl, *Breaking the amyloidogenicity code: methods to predict amyloids from amino acid sequence.* 2013. **587**(8): p. 1089–1095.
35. Tsolis AC, et al., *A consensus method for the prediction of 'aggregation-prone'peptides in globular proteins.* 2013. **8**(1): p. e54175.
36. Benington L, et al. Fibroblast Growth Factor 2—A Review of Stabilisation Approaches for Clinical Applications. 2020;12(6):508.
37. Fernandez-Escamilla A-M, et al., *Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins*. 2004. 22(10): p. 1302.
38. Linding R, et al., *A comparative study of the relationship between protein structure and β-aggregation in globular and intrinsically disordered proteins*. 2004. 342(1): p. 345–353.
39. Siepen JA, Radford SE. and D.R.J.P.s. Westhead, *β Edge strands in protein structure prediction and aggregation*. 2003. 12(10): p. 2348–2359.
40. Chiti F, et al., *Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases*. 2002. 99(suppl 4): p. 16419–16426.
41. Wang X, et al. *Potential aggregation prone regions in biotherapeutics: a survey of commercial monoclonal antibodies*. in *MAbs*. 2009. Taylor & Francis.
42. Thompson MJ, et al., *The 3D profile method for identifying fibril-forming segments of proteins*. 2006. 103(11): p. 4074–4078.
43. Maurer-Stroh S, et al., *Exploring the sequence determinants of amyloid structure using position-specific scoring matrices.* 2010. **7**(3): p. 237.

44. Chiti F, et al., *Rationalization of the effects of mutations on peptide andprotein aggregation rates*. 2003. 424(6950): p. 805–808.
45. Li Y, et al., *Heparin binding preference and structures in the fibroblast growth factor family parallel their evolutionary diversification*. 2016. 6(3): p. 150275.
46. Xu R, et al., *Diversification of the structural determinants of fibroblast growth factor-heparin interactions implications for binding specificity*. 2012. 287(47): p. 40061–40073.

Figures

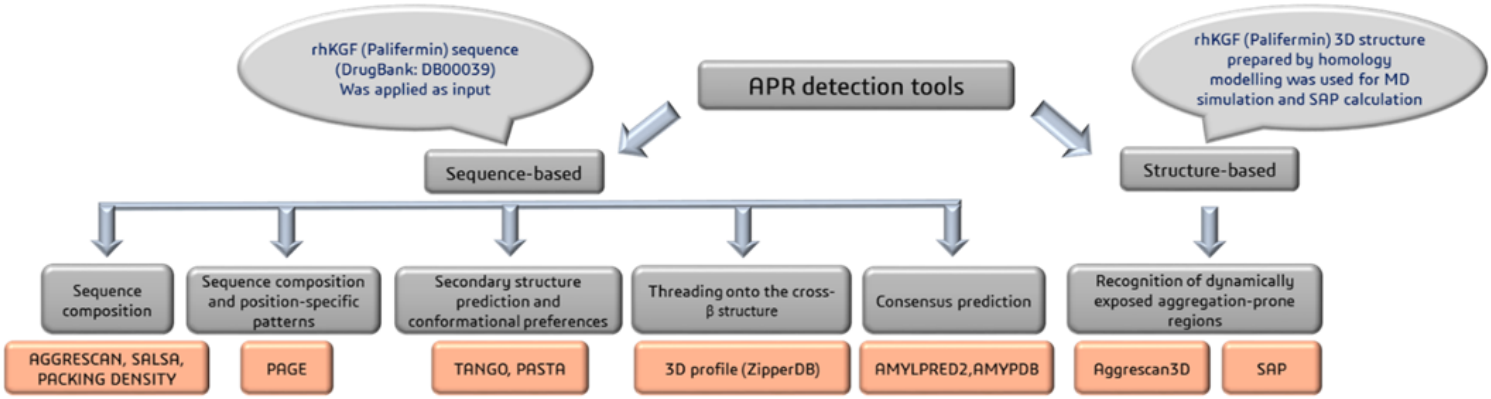


Figure 1

A schematic diagram describing the method applied for prediction of aggregation hotspots in rhKGF using different APR detection tools applying different algorithms.

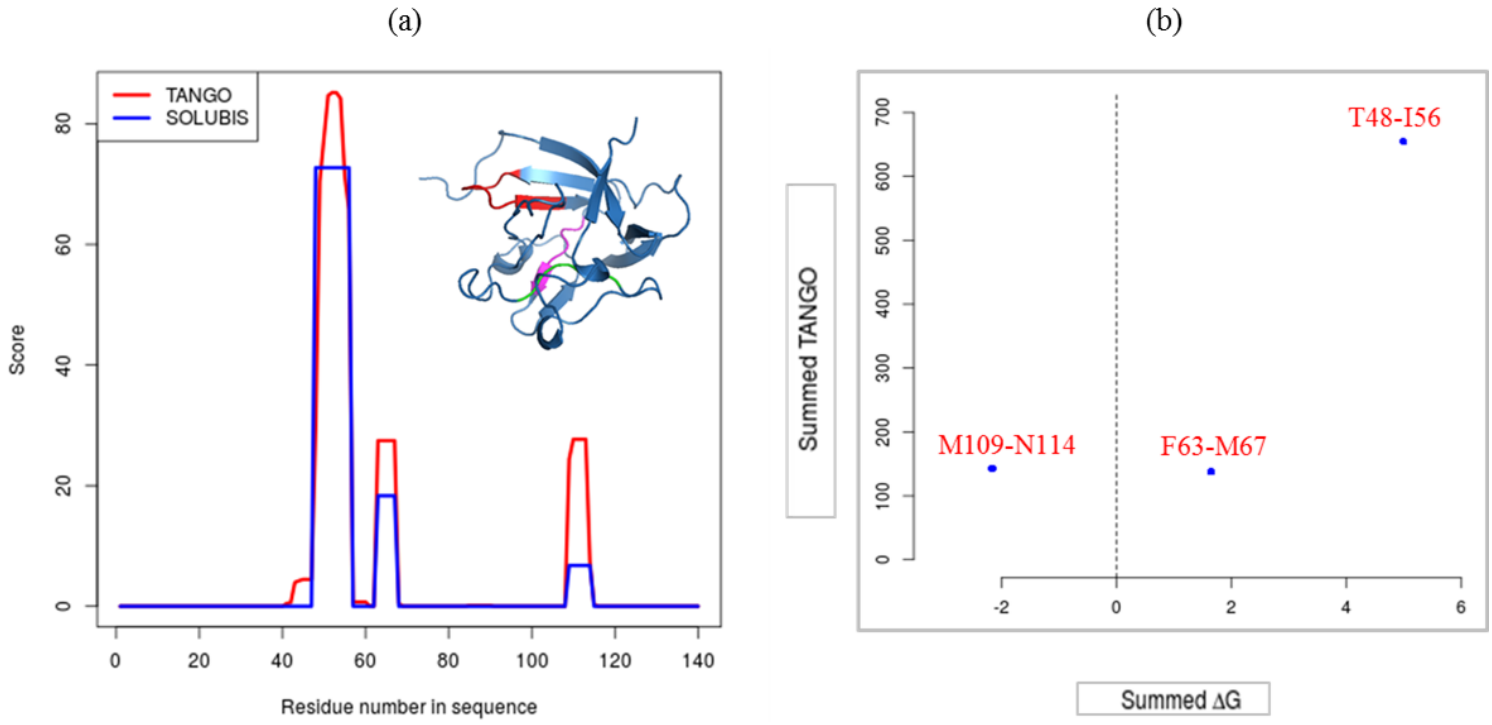


Figure 2

a) Schematic representation of the APRs predicted by TANGO in rhKGF. The TANGO aggregation propensity and the SolubiS score (normalized based on the thermodynamic stability of APRs) were plotted versus the protein sequence. Three regions were identified as APR by TANGO, including residues 48-56: TVAVGIVAI, 63-67: FYLAM, and 109-113: MFVAL, and highlighted in the schematic 3D structure with red, pink and green colors, respectively. b) Stretch-plot of aggregation propensity and thermodynamic stability of APRs in rhKGF. The APRs located at top right of the plot are problematic under native conditions (high aggregation tendency, low thermodynamic stability). The ideal situation is in the bottom left. These APRs are protected from triggering aggregation by folding.

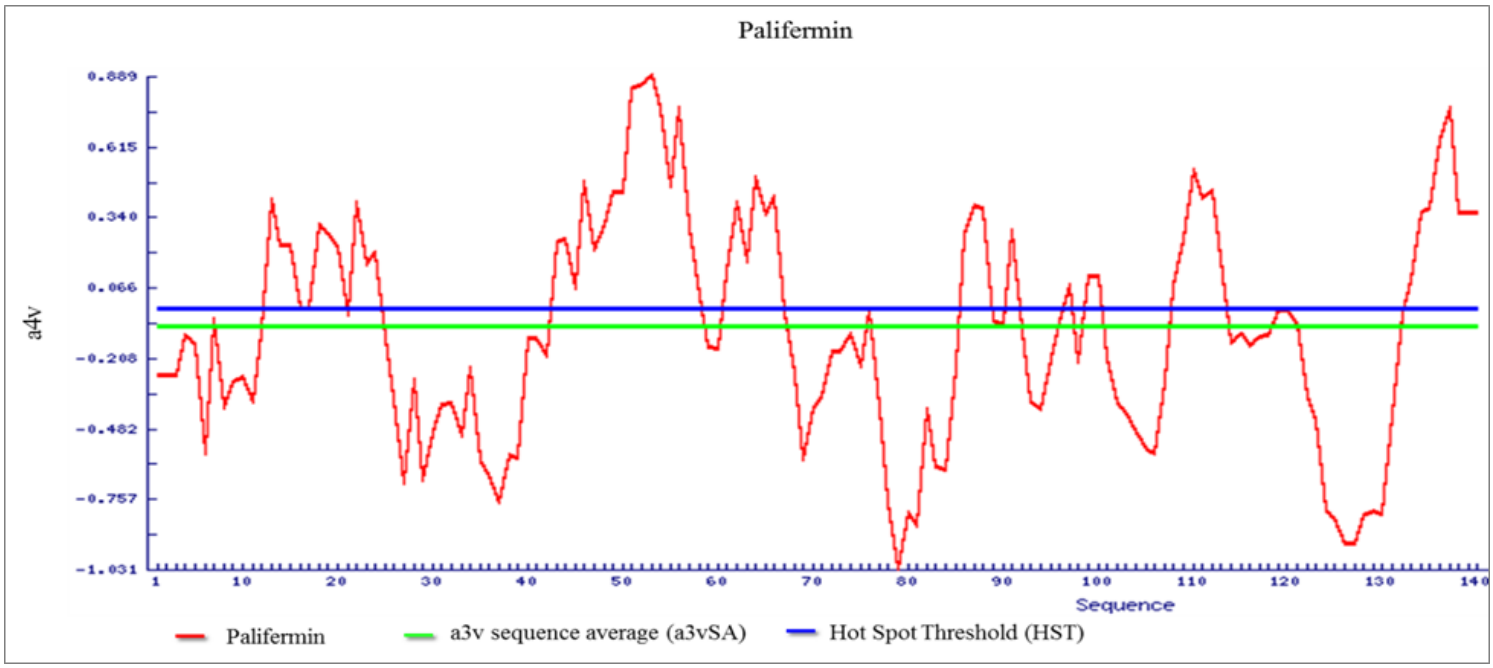


Figure 3

Aggrescan graphic profile of rhKGF. a4v (red), a3vSA (green line) and HST (blue) as function of amino acid sequence. Three aggregation hotspots were identified by Aggrescan in rhKGF including residues 43-58, 61-66, and 108-113.

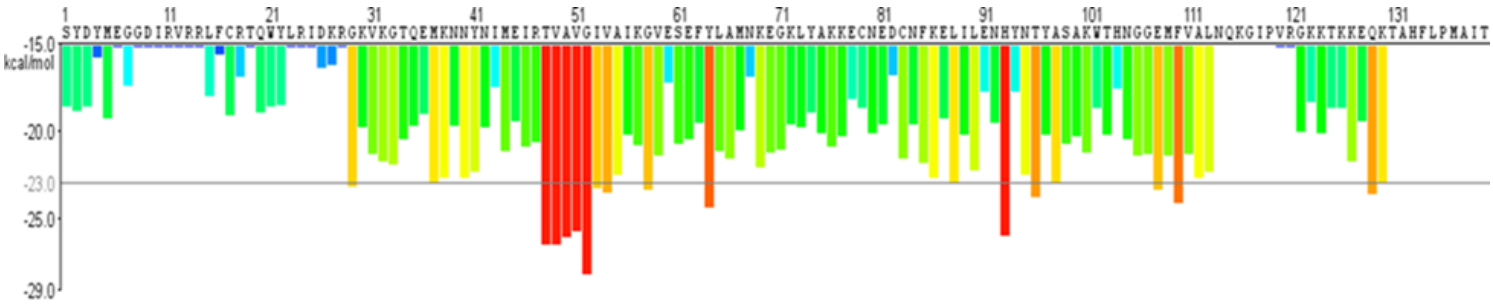


Figure 4

The rhKGF fibril-forming propensity analysis by ZipperDB based on 3D-profile method. Each histogram bar represents a hexapeptide starting at the indicated position which was colored based on the Rosetta energy. The orange-red segments with the energy level below the -23 Kcal/mol threshold were predicted to

form fibril. Residues G29, T48-V54, G58, Y64, H93, T96, A98, E108, F110, and Q129 have Rosetta energy below -23 Kcal/mol which also showing lowest composite scores.

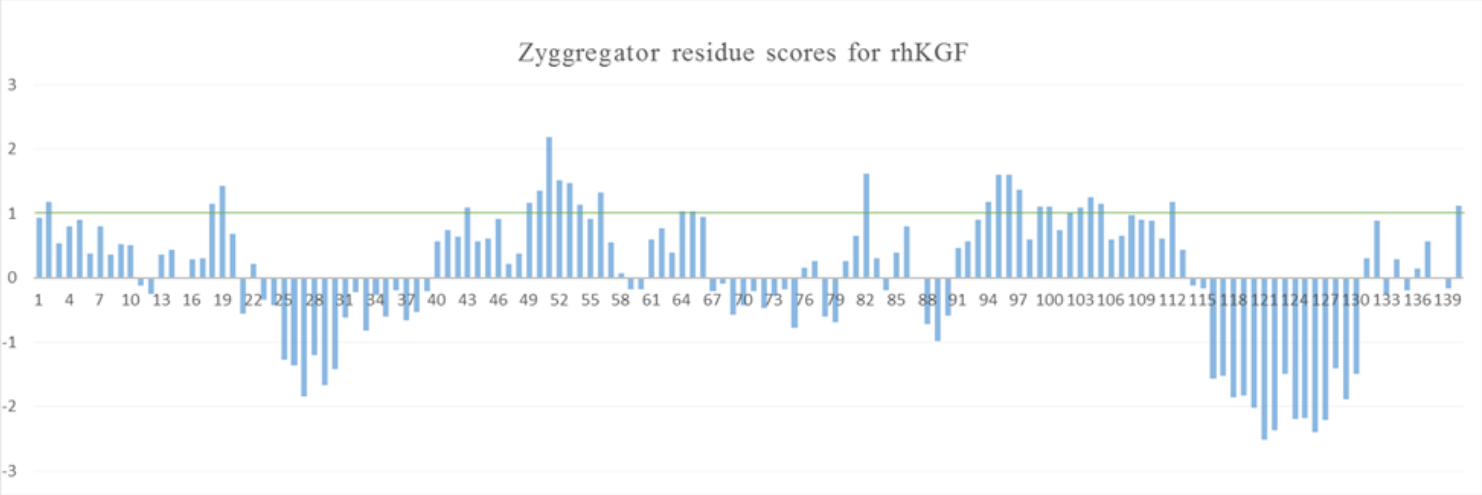


Figure 5

Zyggregator scores (Ziagg) for rhKGF. The green-colored line shows the threshold for aggregation tendency. Residues with scores above the threshold are predicted to participate in the formation of fibrillar aggregates. These residues are Y2, R18, T19, I43, V49-V54, I56, Y64, L65, D82, Y94-Y97, S99, A100, T103-N105, A112, and T140.

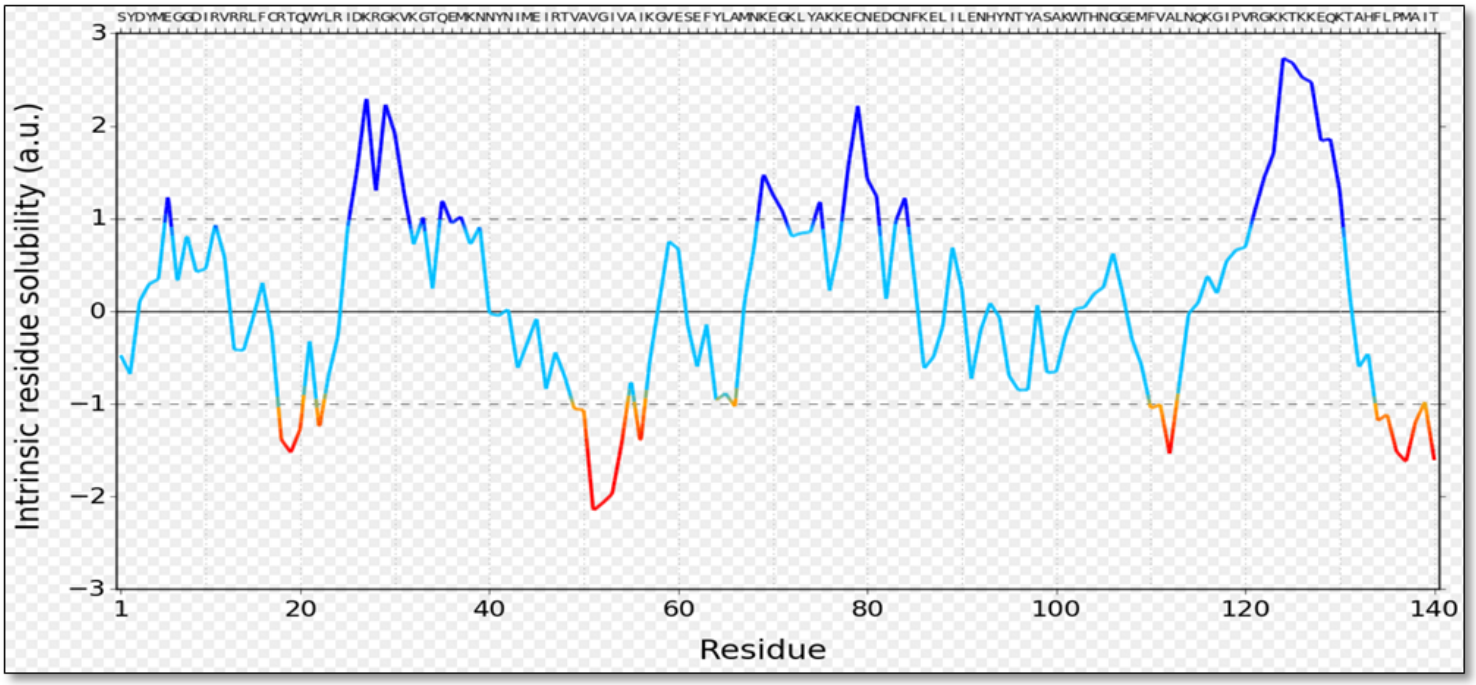


Figure 6

The intrinsic solubility profile of rhKGF created by CamSol method based on the amino acid sequence. The regions with the scores more than +1 are highly soluble (dark blue) and the regions with the solubility

scores below -1 (shown in orange) are identified as poorly soluble ones. The amino acids with the score below -1 in rhKGF are R18, T19, Q20, Y22, A66, V49-V54, I56, F110-A112, F134-A138, and T140.

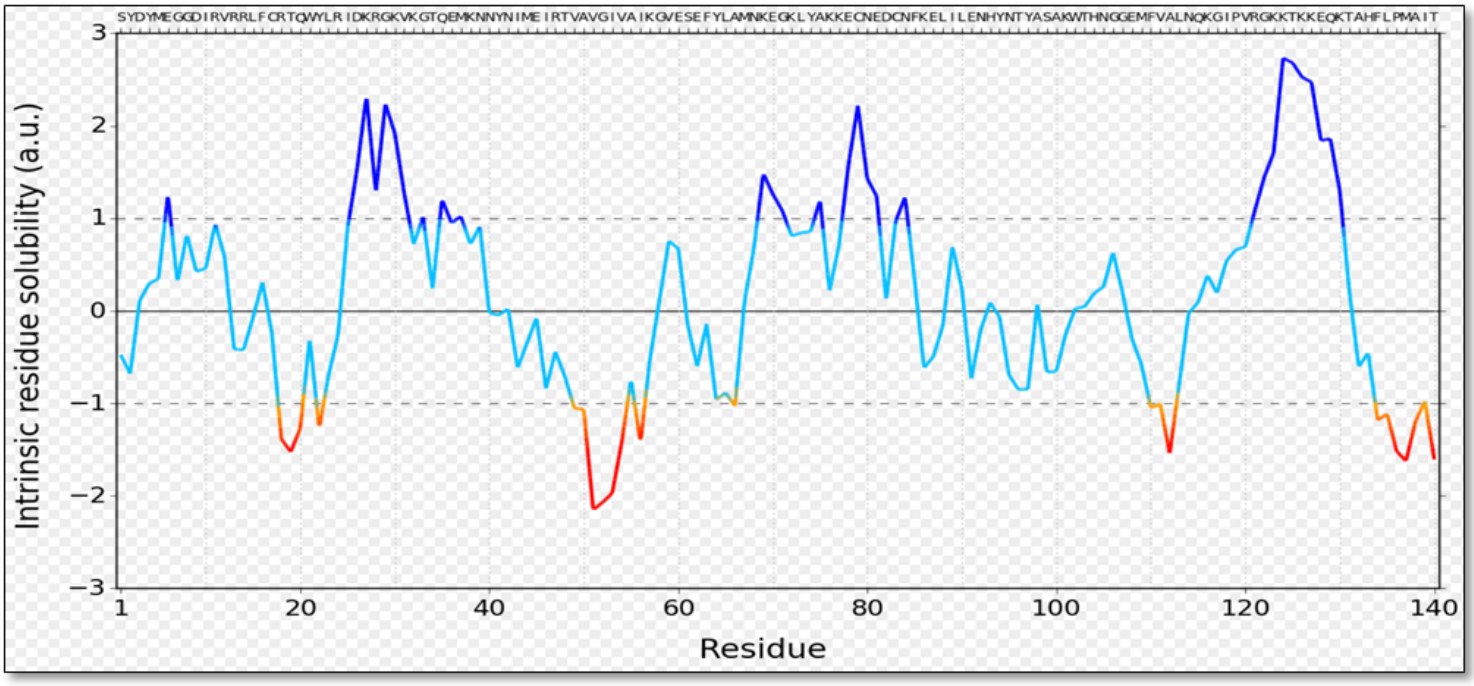


Figure 7

The poorly soluble amino acids in rhKGF with the CamSol structurally-corrected solubility scores below -1 are shown with the black arrows and marked with the yellow color. These residues are the candidate positions for mutations to enhance solubility. The structurally-corrected solubility scores are calculated from the intrinsic solubility scores based on the proximity of the amino acids in the 3D structure and their solvent exposure.

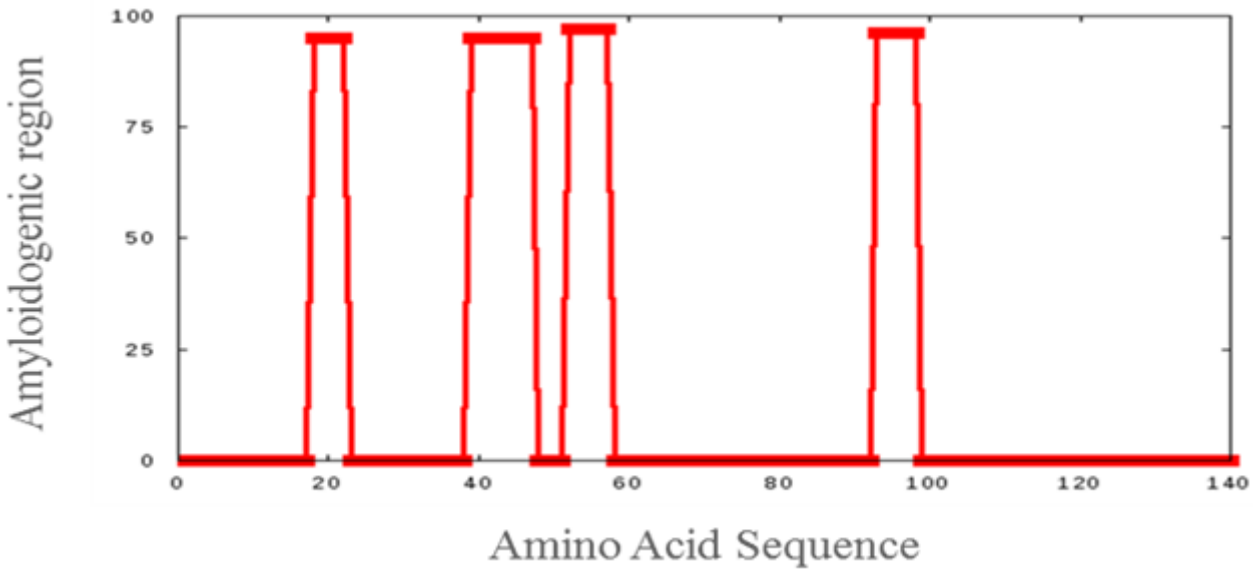


Figure 8

The amyloidogenic regions identified in Palifermin using WALTZ. The best overall performance threshold has identified four regions as APR in rhKGF including 18-22, 39-47, 52-57, and 93-98.

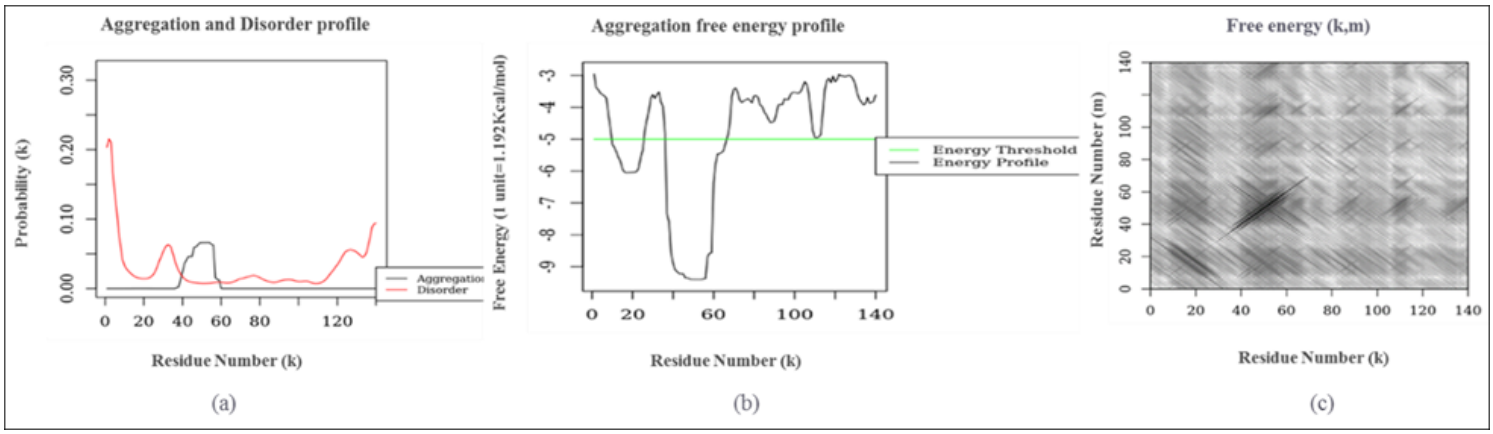


Figure 9

Prediction of amyloid structure aggregation in rhKGF primary structure using PASTA. According to aggregation and disorder profile (a), aggregation free energy profile (b), and aggregation-pairing matrix (c) the portion 40-60, having the lowest aggregation free energy, is the most amyloid stabilizing region responsible for rhKGF amyloid-forming propensity. The graph (c) shows the paring probability important for protein self-aggregation. The darkest area belongs to the region 40-60 and has the most paring probability.

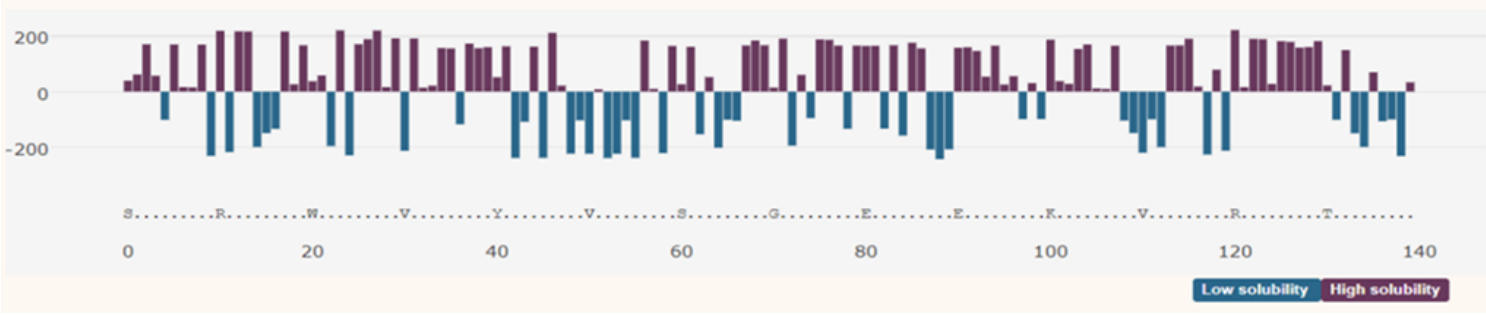


Figure 10

The graphical solubility profile of rhKGF prepared by SODA server. The low solubility regions (the aggregation hits) are shown in blue.

Figure 11

The consensus results of APR prediction with AMYLPRED server. The prediction is based on the overlap of two out of five methods. The hits (15-25, 43-56, 63-67, 88-91, 109-114, and 133-139) were presented with the * symbol.

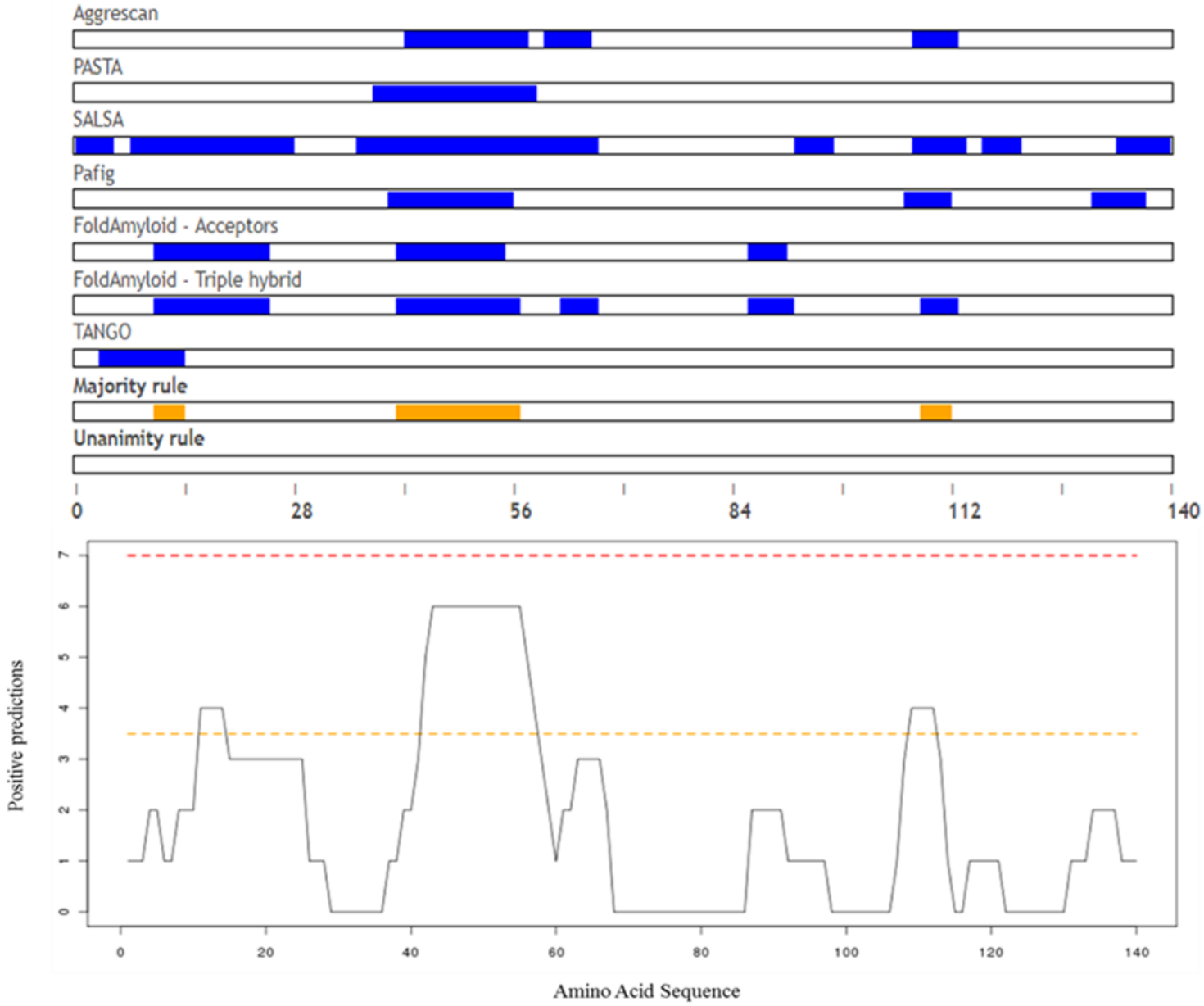


Figure 12

The results of consensus APR prediction in rhKGF by AMYPDB. The three major APRs identified by different tools include RVRR (11-14), NIMEIRTVAVGIVAIK (42-57), and MFVA (109-112).

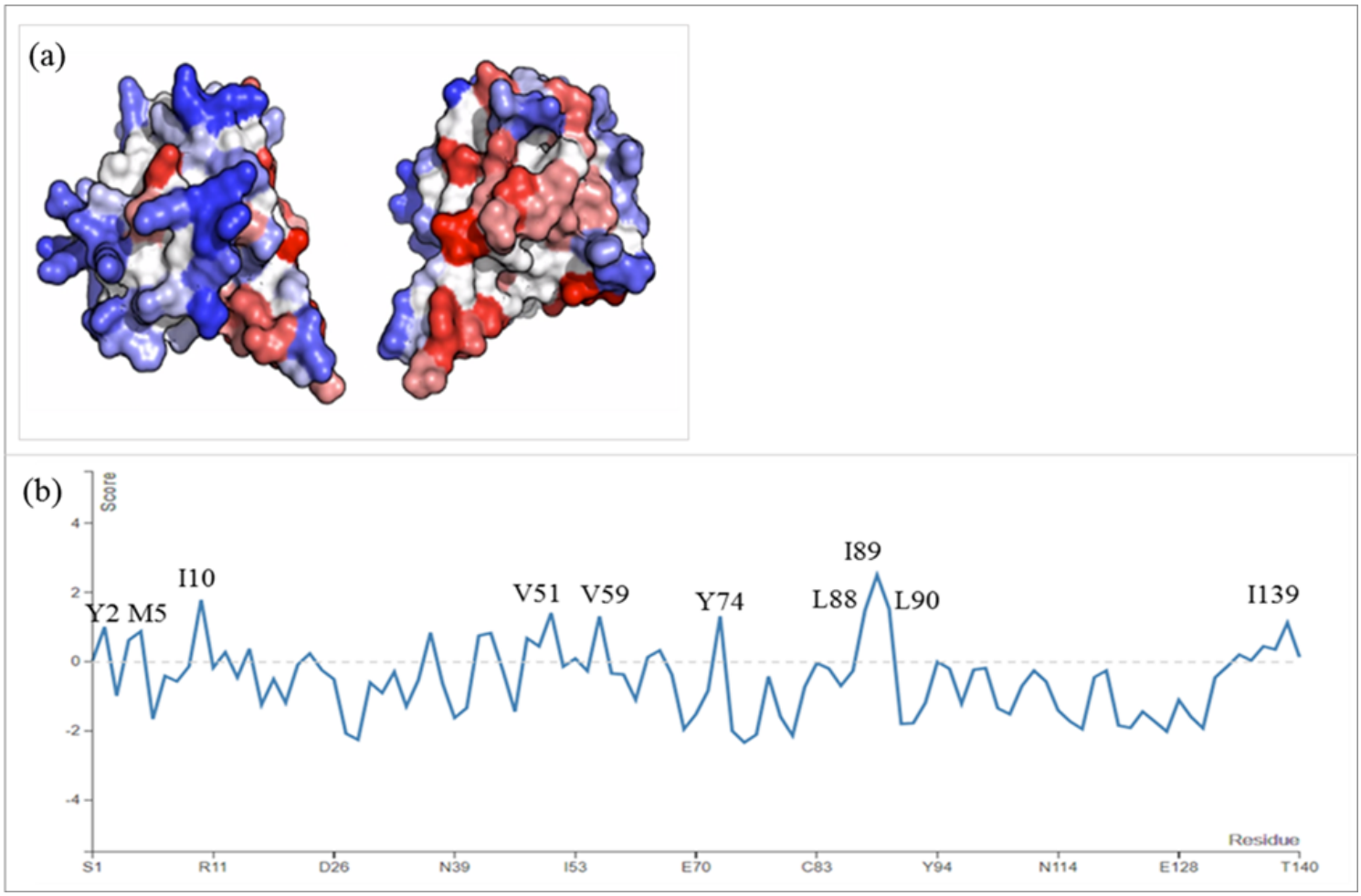


Figure 13

a) A3D structure-based analysis of the rhKGF aggregation propensity in the dynamic state. The protein surface has been colored according to the A3D score in gradient (from Red for high aggregation propensity to white for negligible effect on the aggregation propensity to blue for high solubility regions). b) The A3D profile of the rhKGF. The positive and negative scores are correlated to aggregation- and solubility-prone residues, respectively. The residues with the highest A3D score involved in the formation of the aggregation-prone patches of the rhKGF folded state have been marked.

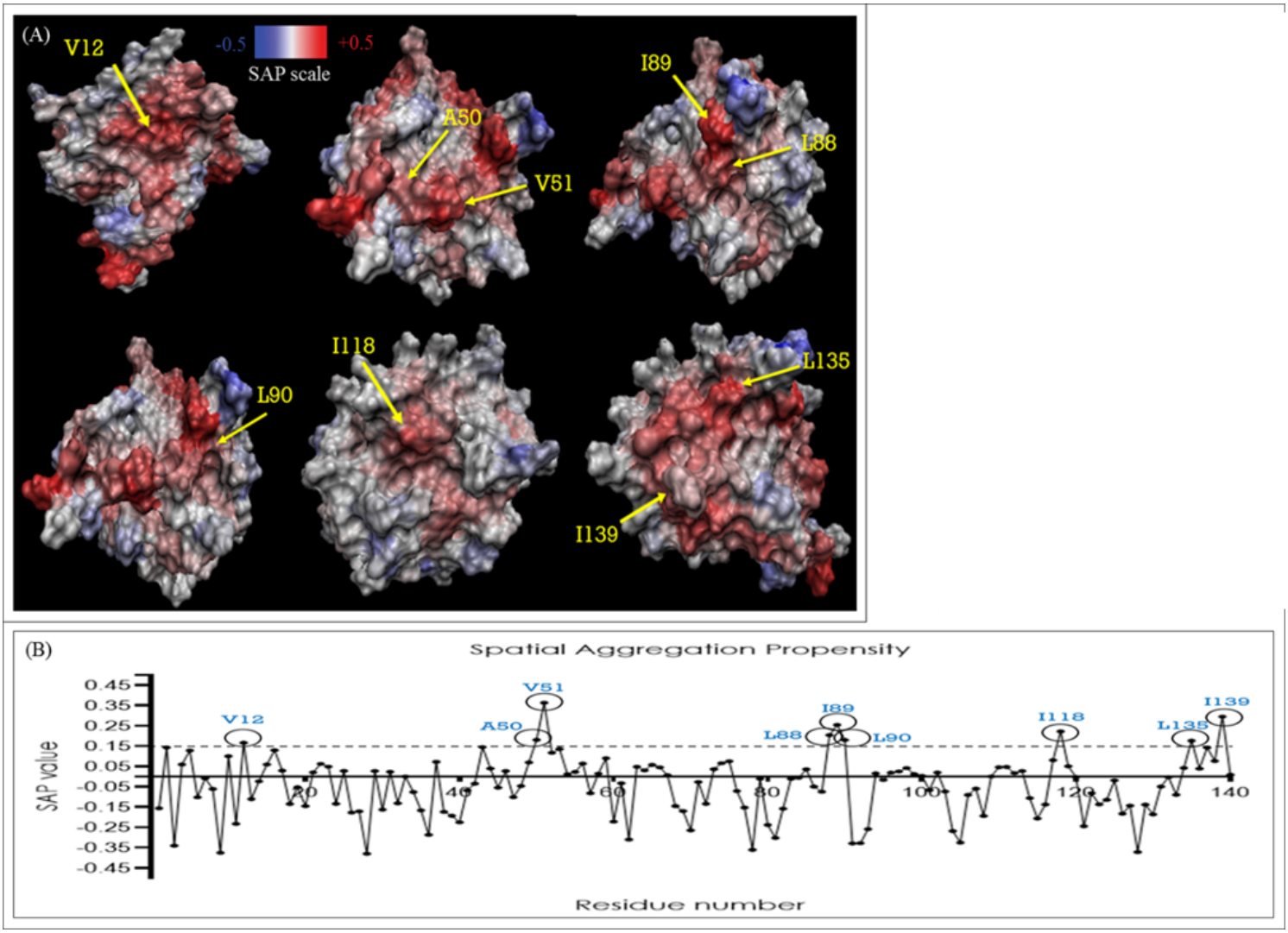
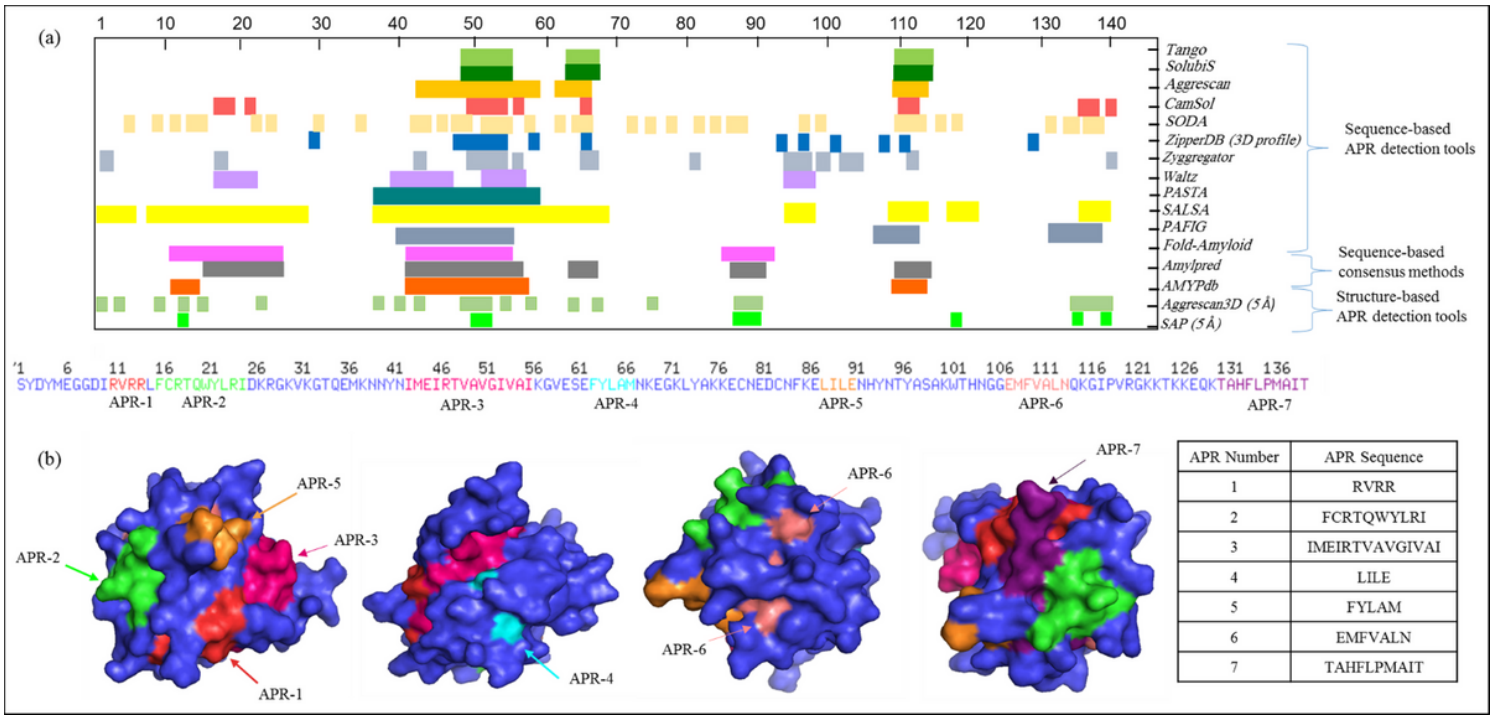


Figure 14

Spatial aggregation propensity for the rhKGF (Palifermin). A) The SAP values at R=5 Å are mapped onto the rhKGF structure. (Note: Due to the difference in residue SAP values and therefore color intensity in different SAP_mapped PDBs from dynamic SAP calculation, the SAP_mapped PDB obtained from the static SAP calculation was used to show the SAP values on the structure and it is the reason that the color intensity and SAP scores are not matched exactly.). The regions with the positive SAP scores (more hydrophobicity and aggregation propensity) are shown in red and the areas with the negative SAP values (less hydrophobicity and aggregation propensity) are represented in blue. The color intensity shows the degree of hydrophobicity or hydrophilicity. The residues with the SAP values above 0.15 at R=5 Å which are considered as the candidate positions for mutation are located. B) The average dynamic SAP values of the rhKGF residues at R=5 Å. The residues with the SAP scores greater than 0.15 are marked.



APR Number	APR Sequence
1	RVRR
2	FCRTQWYLRI
3	IMEIRTVAVGIVAI
4	LILE
5	FYLAM
6	EMFVALN
7	TAHFLPMAIT

Figure 15

(a) The aggregation propensity of rhKGF was analyzed using different sequence, structure-based and consensus APR predictors. The APRs in rhKGF have been indicated with different colors for different predictors. The predictions showed that the APRs in rhKGF are mainly located at positions 10-30, 40-60, 61-66, 88-120, and 130-140. Similar to APRs found in the other amyloidogenic proteins, rhKGF APRs are rich in β-branched aliphatic, hydrophobic, aromatic and in some regions Q/N residues. While the APRs identified by the sequence-based predictors might be buried or exposed in the folded protein, the structure-based predictions made with the Aggrescan3D and SAP limited the number of identified APRs to the dynamically-exposed hydrophobic residues including V12, A50, V51, L88, I89, L90, I118, L135, and I139 mediating the native-state aggregation. (b) The modeled crystal structure (space-filling model) of the rhKGF. Mapping the computationally predicted APRs found by the sequence-based tools on the tertiary structure of the rhKGF showed that most of the APRs are solvent-exposed in the natively folded protein including F16-R25, I43, E45, R47-I56, F61, Y62, N66, L88-E91, E108-F110, A112, N114, T131, and H133-T140.