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

Mansoureh Shahbazi Dastjerdeh, Mohammad Ali Shokrgozar, Hamzeh Rahimi and Majid Golkar

Department of Molecular Medicine, Biotechnology Research Center, Pasteur Institute of Iran, Tehran, Iran; National Cell Bank, Pasteur Institute of Iran, Tehran, Iran; Department of Parasitology, Pasteur Institute of Iran, Tehran, Iran

Communicated by Ramaswamy H. Sarma

ABSTRACT

The recombinant human keratinocyte growth factor (rhKGF) is a highly aggregation-prone therapeutic protein. The high aggregation liability of rhKGF is manifested by loss of the monomeric state, and accumulation of the aggregated species even at moderate temperatures. Here, we analyzed the rhKGF for its vulnerability toward aggregation by detection of aggregation-prone regions (APRs) using several sequence-based computational tools including TANGO, ZipperDB, AGGREGSCAN, Zygggregator, Camsol, PASTA, SALSA, WALTZ, SODA, Amylpred, AMYPD8, and structure-based tools including Solubis, CamSol structurally corrected, Aggrescan3D and spatial aggregation propensity (SAP) algorithm. 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. Mapping on the rhKGF structure revealed that most of these residues including F16-R25, I43, E45, R47-I56, F61, Y62, N66, L88-E91, E108-E110, A112, N114, T131, and H133-T140 are surface-exposed in the native state which can promote aggregation without major unfolding event, or the conformational change may occur in the oligomers. The other regions are buried in the native state and their contribution to non-native aggregation is mediated by a preceding unfolding event. The structure-based prediction of APRs using the 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. Our analysis of APRs in rhKGF identified the regions determining the intrinsic aggregation propensity of the rhKGF which are the candidate positions for engineering the rhKGF to reduce its aggregation tendency.

ABBREVIATIONS: rhKGF: Recombinant Human Keratinocyte Growth Factor; APR: Aggregation-Prone Region; MD: Molecular Dynamics; SAP: Spatial Aggregation Propensity

1. Background

Keratinocyte growth factor (KGF), also known as FGF-7, is a member of the fibroblast growth factor (FGF) family with the proven therapeutic properties in wound healing (Dastjerdeh et al., 2019; Rajan et al., 2010). 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 (Huang, 2012). 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 (Dastjerdeh et al., 2019). It has shown the ability to reduce the severity, incidence, and duration of oral mucositis in cancer patients (Hsu et al., 2006). However, KGF is a fairly unstable and highly aggregation-prone therapeutic protein denaturing at ~37°C (around the body temperature) which is the lower end of temperature stability range for most of the globular proteins (Huang, 2012). 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 (Hsu et al., 2006).

According to the Lumry and Eyring equation, aggregate formation can be suppressed either through stabilizing the native state or reducing aggregation rate (Rajan et al., 2010). 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 (Chen et al., 1994). The solid-phase PEGylating of rhKGF also significantly improved the stability without affecting its structure (Huang, 2012). 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 (Hsu et al., 2006). However, the high aggregation tendency, low storage stability and short plasma half-life of rhKGF has remained a challenge limiting its pharmaceutical applications (Poorebrahim et al., 2017; Zahra et al., 2016). Protein aggregation can also reduce the protein binding affinity, elicit immune responses, and hamper the safety and efficacy of rhKGF as other biotherapeutics (Buck, 2012; Wang & Roberts, 2010). As such, predicting and mitigating the aggregation of rhKGF is of great importance from a pharmaceutical perspective. 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 (Lee et al., 2013). 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 (Rousseau et al., 2006), ZipperDB (Sawaya et al., 2007), AGGRESCAN (Conchillo-Sole et al., 2007), Zyggregator (Tartaglia & Vendruscolo, 2008), CamSolintrinsic (Sormanni et al., 2015), PASTA (Walsh et al., 2014), SALSA (Zibae et al., 2007), WALTZ (Oliveberg, 2010), and SODA (Paladin et al., 2017), consensus methods including Amylpred (Tsolis, 2013), and AMYPDB (Pawlicki et al., 2008) and structure-based tools including SolubiS (van der Kant et al., 2017), CamSol Structurally Corrected (Sormanni et al., 2015), Aggrescan3D (Zambrano et al., 2015) and molecular dynamics-based SAP tool (Chennamsetty et al., 2010). 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 (Buck, 2012; Chennamsetty et al., 2010).

Our results identified the candidate positions for the rational design of rhKGF to reduce its vulnerability toward 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 Figure 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 D800039) 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 modeling (template-based modeling) was performed using GalaxyTBM server to predict its 3D-structure (Ko et al., 2012). 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 was 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 (Buck, 2012). 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 (Buck, 2012). 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. AGGRESCAN
AGGRESCAN is a server for the identification of aggregation nucleating segments or "aggregation hot spots" in the poly-peptides 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. AGGRESCAN is also able to detect the mutation effects on protein aggregation tendency as a function of protein sequence composition. For each protein sequence, AGGRESCAN calculates and presents:

1. An aggregation-propensity value for each amino acid (a³v), a³v windows average (a⁴v) 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 a⁴v larger than the HST.
4. HSA divided by the number of residues in the entry sequence, known as “Normalized Hotspot Area” (NHSA) per residue (Conchillo-Solé et al., 2007)

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.

2.3.3. ZipperDB
ZipperDB predicts protein aggregation by evaluation of fibril-forming propensity of the protein segments using the 3D-profile method (Kuhlman & Baker, 2000). 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 (Nelson et al., 2005). 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 (Sawaya et al., 2007). 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)

2.3.4. Zyggregator
Zyggregator is also a sequence-based method which calculates different propensities including the local stability of the monomeric state (In Pi), formation of β-rich oligomers (Zi^tox), and the formation of fibrillar aggregates (Zi^agg) to predict the intrinsic amyloid aggregation propensity of the protein. The Zi^agg score 0 is related to the aggregation propensity equal

Figure 1. A schematic diagram describing the method applied for prediction of aggregation hotspots in rhKGF using different APR detection tools applying different algorithms.
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 (Tartaglia & Vendruscolo, 2008).

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.

2.3.5. CamSol
CamSol consists of both a sequence-, and a structure-based algorithm 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 available at (http://www-vendruscolo.ch.cam.ac.uk/camsolmethod.html) to analyze its intrinsic solubility.

2.3.6. 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 (Oliveberg, 2010). 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 \( \beta \)-structure forming propensity, and a position-specific pseudoenergy matrix determined from the crystal structure of Sup35 GNNQQNY peptide which forms the cross-\( \beta \) spine of amyloid fibrils to estimate the relative energies using FoldX (Ahmed & Kajava, 2013; Meric et al., 2017). 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 (Meric et al., 2017).

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.

2.3.7. 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 (Walsh et al., 2014). PASTA makes aggregation predictions based on the assumption that mechanisms governing the native \( \beta \)-sheet formation, are also responsible for the formation of \( \beta \)-sheets in amyloid aggregates (Meric et al., 2017). The peptides with the energy level below a specified threshold are defined as aggregation hotspots. To analyze the rhKGF, its amino acid sequence was submitted to PASTA server available at http://protein.bio.unipd.it/pasta2/.

2.3.8. 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 (\( \alpha \)-helix and \( \beta \)-strand) propensity to predict changes of protein solubility based on the physicochemical properties (Paladin et al., 2017). 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 (Paladin et al., 2017). 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.

2.3.9. Amylpred
Amylpred utilizes a consensus of various strategies that have been found or explicitly developed to foresee the features related to the formation of amyloid fibrils. 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 (Tsolis et al., 2013). AMYLPRED makes a 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 at least two out of five methods (Tsolis et al., 2013). To make a consensus prediction of APRs in rhKGF, the AmylPred server was used which makes
2.3.10. 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 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 (Pawlicki et al., 2008).

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 (Pawlicki et al., 2008). 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. 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 (van der Kant, 2019; van der Kant et al., 2017). 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.

2.4.2. CamSol structurally corrected
The second algorithm of CamSol, known as CamSol structurally corrected, applies a PDB file as input, 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 (Sormanni et al., 2015). The rhKGF PDB prepared by homology modeling was used as the input file for CamSol structural analysis (http://www-vendruscolo.ch.cam.ac.uk/camsol-method.html).

2.4.3. AGGRESCAN3D (A3D)
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. In the A3D static mode calculation, the input structure is minimized with FoldX force-field, and then the AGGRESCAN intrinsic aggregation propensity score of each amino acid is modified based on its structural context. This procedure provides a structurally corrected aggregation score for each residue. In the dynamic mode of A3D, the input structure is first minimized by FoldX and applied as the starting structure for simulation of structural flexibility using CABS-flex protocol which is a fast and effective alternative to classic MD simulation. Finally, the A3D static mode calculation is performed for all CABS-flex generated models, and the model with the highest A3D score is presented in output as the most aggregation-prone structural variant of protein (Zambrano et al., 2015). AGGRESCAN3D server, available at http://biocomp.chem.uw.edu.pl/A3D/, was used to identify APRs in the rhKGF structure at the radius of 5 Å.

2.4.4. Spatial aggregation propensity (SAP)
The hydrophobic residues generally participate in the formation of the hydrophobic core of the proteins, increasing the protein stability 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 when they cluster together on the protein surfaces. 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 (Ψeff) of the neighboring residues

\[
\text{SAP}_{\text{atom } i} = \sum_{\text{Simulation average}} \sum_{\text{Residues with at least one side chain atom within } R \text{ from atom } i} \left( \frac{SAA \text{ of side chain atoms within radius } R}{\text{SAA of side chain atoms of fully exposed residue}} \right) \times \text{Hydrophobicity}_{j}
\]
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$_{atom}$ (Buck, 2012; Lee et al., 2013):

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 (Courtois et al., 2016).

2.4.4.1. 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 (Courtois et al., 2016). 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.

2.4.4.2. 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.

2.5. Integration of prediction methods

We integrated the result of APR prediction in rhKGF by different sequence- and structure-based tools on a graph to determine the consensus APRs. Furthermore, to distinguish the surface-exposed APRs from the buried ones mediating aggregation of rhKGF through native and non-native pathways, respectively, the consensus APRs were mapped onto the 3D-structure of rhKGF using PYMOL software version 2.4.1.

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

Amyloid fibrils, as highly ordered fibrillar structures, are composed of the polypeptide chains arranged in a cross-β-sheet conformation. Many predictive algorithms have been developed to predict β-aggregation, and amyloid fibril-forming propensity of peptides and proteins (Garvey et al., 2013). Here, we tried to integrate most of the available predictive algorithms to predict the regions determining the cross-β aggregation and amyloid-forming propensity of rhKGF.

TANGO, as one such statistical algorithm, incorporates not only the peptide or protein sequence, but also the physico-chemical parameters including pH, temperature, and ionic strength to calculate the cross-β aggregation propensity of proteins (Garvey et al., 2013). 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 (Figure 2). The highest TANGO β-aggregation score belongs to the TVAVGIVAI peptide.

To distinguish the buried and exposed APRs predicted by TANGO, the rhKGF was analyzed using SolubiS server. 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 predicted as critical (surface-exposed) APRs mediating native-state aggregation and the region 109–114 with negative SolubiS score was defined as structural (buried) APR which participates in non-native aggregation (Figure 2).

According to the AGGRESCAN a4v numerical data (not shown) and hot spot area (Figure 3) prepared for the rhKGF, three aggregation hotspots have been recognized in the rhKGF including residues 43–58, 61–66, and 108–113. AGGRESCAN predicts these regions 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.

According to the 3D profile approach implemented by ZipperDB and by threading the sequence of rhKGF on the crystal structure of NNQNY peptide of the sup35 prion protein, the residues G29, T48-V54, G58, Y64, H93, T96, A98, E108, F110, and Q129 with the energy below −23 Kcal/mol were identified as amyloid-prone regions in rhKGF (Figure 4).
These regions can form the cross-beta spine of amyloid-like fibrils in rhKGF. Most of the APRs predicted by ZipperDB were common with those identified by TANGO and AGGRESCAN except for G29, H93, T96, A98, E108 and Q129.

The possibility of the fibril or protofibril structure formation and amyloid aggregation propensity of rhKGF was also predicted with Zyggregator. Zyggregator identified residues Y2, R18, T19, I43, V49-V54, I56, Y64, L65, D82, Y94-Y97, S99,
A100, T103-N105, A112, and T140 with the Ziagg score of one or higher as aggregation-prone residues (Figure 5). 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 (Chiti et al., 2002; Meric et al., 2017). Zyggregator also evaluates the flanking residues (“gatekeepers”) of a sliding window regarding the presence of charged residues which may decrease aggregation.

The intrinsic solubility profile of the rhKGF prepared by CamSol identified the residues R18, T19, Q20, Y22, A66, V49-V54, I56, F110-A112, F134-A138, and T140 as poorly soluble (aggregation-prone) amino acids in rhKGF (Figure 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 (Figure 7). These are the solvent-exposed poorly soluble residues of rhKGF which have been determined through structural correction to the CamSol intrinsic scores.

The result of amyloid aggregates prediction in rhKGF using WALTZ algorithm are presented in Table 1 and Figure 8. Most of the amyloidogenic regions predicted by WALTZ were common with the other APR prediction tools. The least conserved

| Positions | Sequence | Average score per residue |
|-----------|----------|---------------------------|
| 18-22     | RTQWY    | 94.98                     |
| 39-47     | NNYNMEIR | 94.98                     |
| 52-57     | GIVAIX    | 96.98                     |
| 93-98     | HINTYXA   | 96.32                     |

Figure 5. Zyggregator scores ($Z_{agg}$) 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 solvent-exposed 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.
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.

The PASTA aggregation and disorder profile shows the aggregation probability versus disorder probability at residue k in self-aggregation mode. The pairing free energy of residue k with residue m is presented as free energy (k, m) in self-alignment (Trovato et al., 2006). The PASTA aggregation profile provides a normalized per-residue aggregation probability as h (k). This parameter is calculated from the h (k) equation (Equation (5)) (Trovato et al., 2006) and shows that which regions of the protein have more aggregation probability as h (k). This parameter is calculated which the k and m letters represent the first and second amino acids of a pair, respectively (Raja & Kinne, 2020). The PASTA aggregation profile provides a normalized per-residue aggregation probability as h (k). This parameter is calculated from the h (k) equation (Equation (5)) (Trovato et al., 2006) and shows that which regions of the protein have more aggregation tendency. The h^2 (k, m), calculated from the h (k) equation (Equation (5)) (Trovato et al., 2006) demonstrates the result of Spatial Aggregation Propensity (SAP) calculation for rhKGF. The residue SAP scores and the SAP mapped PDB files showed that almost half of the rhKGF residues have positive SAP scores forming the hydrophobic exposed patches on the surface of rhKGF which make it prone to aggregate in the aqueous media. The SAP mapped PDB files were visualized and presented by the Visual Molecular Dynamics (VMD) software (version 1.9.3). The color intensity indicates the degree of hydrophobicity or hydrophilicity. The graph of SAP values was prepared by the GraphPad Prism software (version 8.3.0). 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. The static SAP calculation defined the same residues as those identified by dynamic calculation as high SAP residues, except for the residue Y2 which had a static SAP score of 0.15, but an average dynamic SAP score of 0.14.

3.4. Integration of results

Integration of the APR prediction in rhKGF made with different tools indicated that the high intrinsic aggregation propensity of rhKGF is mainly mediated by the APRs located at positions 10–30, 40–60, 61–66, 88–120, and 130–140 (Figure 15(A)). Mapping the computationally predicted APRs found by the sequence-based tools on the tertiary structure of the rhKGF revealed 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.

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 (Chen et al., 1994) and other FGFs (Benington et al., 2020) 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

**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.
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 h to more than 20 days at 37°C (Benington et al., 2020). 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 (Hsu et al., 2006). 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 (van der Kant et al., 2017). 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 (Meric et al., 2017).

A number of computational methods have been developed to identify the sequence aggregation-prone regions in the amyloidogenic proteins. Some of these algorithms like TANGO, AGGRESCAN, AGGRESCAN3D, CamSol, and PAGE focus on prediction of β-aggregation prone regions, but not necessarily the amyloid ones. Some of the other algorithms like PASTA, SALSA, Waltz, Zyggregator, 3D profile, Pafig, Fold-amyloid, NetCSSP, AmyloidMutants, Amyloidogenic Pattern, and SecStr have focused on both prediction of β-aggregation and amyloid-fibril prone regions. Here, we tried to analyze the rhKGF with most of the publicly 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.

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 (Nelson et al., 2005). 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 β2-microglobulin (Siepen et al., 2005).
Figure 12. The results of consensus APR prediction in rhKGF by AMYPDB. The three major APRs identified by different tools include RVRR (11–14), NIMEIRTVAVGIVA (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 (Pallfermin). (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) Table of residues with the average dynamic SAP scores higher than 0.15 at $R = 5$ Å, and the corresponding SAP scores. (C) The plot of average dynamic SAP values of the rhKGF residues at $R = 5$ Å. The residues with SAP scores greater than 0.15 are marked.

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 $\alpha$-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, 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, J43, E45, R47-I56, F61, Y62, N66, L88-E91, E108-F110, A112, N114, T131, and H133-T140.
Moreover, Dobson’s group showed that fibril formation is the general properties of proteins under stress conditions (Chiti et al., 2002). The rhKGF is also a β-rich protein motivating us to explore if the cross β-motif conformation promotes its aggregation. To examine the β-sheet aggregation propensity of rhKGF, it was analyzed using different algorithms of which TANGO, AGGRESCAN, AGGRESCAN3D, and CamSol have been dedicated to identification of cross β-motifs in proteins. TANGO calculations revealed that rhKGF contains three motifs with the aggregation scores above the threshold. The highest aggregation score belonged to the region TVAVGIVAI (48–56) implying that this peptide has the highest aggregation propensity in rhKGF. The existence of one such aggregation-prone motifs can be sufficient for cross-β motif-derived aggregation, but it does not necessarily guarantee that the protein aggregation proceeds by this mechanism (Wang et al., 2009).

Using TANGO APR prediction algorithm, SolubiS is also able to define the motifs involved in cross-β aggregation of proteins. 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 (van der Kant et al., 2017). 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 (Meric et al., 2017; van der Kant et al., 2017).

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 the amyloid ones (Conchillo-Solé et al., 2007). AGGRESCAN3D as the optimized version of AGGRESCAN which is capable of distinguishing the buried from surface-exposed APRs, is also a predictor of β-sheet aggregation. AGGRESCAN3D identified the spatially-adjacent APRs involved in the aggregation of rhKGF from the native state by structural correction applied to the AGGRESCAN score. AGGRESCAN3D has exhibited higher accuracy compared to the sequence-based algorithms in forecasting the aggregation hot spots of the globular proteins (Zambrano et al., 2015).

CamSol is also able to provide predictions of protein β-aggregation propensity. However, unlike the other algorithms which may only focus on amyloid formation, CamSol also provides predictions of protein solubility (Sormanni et al., 2015). The intrinsic solubility profile of the rhKGF prepared by CamSol identified the poorly soluble and aggregation-prone residues in rhKGF. The results of APR prediction in rhKGF using camsol are largely consistent with TANGO and AGGRESCAN and the most conserved regions are VAVGIV (V49-V54) and FVA (F110-A112). These residues are the most important candidate positions for designing variants of rhKGF with enhanced solubility. Like AGGRESCAN3D, CamSol structurally-corrected solubility algorithm applies the structural correction to the CamSol intrinsic scores to identify the solvent-exposed poorly soluble amino acids. CamSol employs the variables similar to those implemented in Zyggregator with the difference in the parameters and definition of the gatekeeping effect (Sormanni et al., 2015).

ZipperDB, Zyggregator, WALTZ, PASTA, SODA, AMYLPRED, and AMYPDB are servers which have been developed to predict both cross-β aggregation and amyloid-forming propensity of peptides and proteins (Siepen et al., 2003). Among these tools, 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 (Sawaya et al., 2007). Zyggregator 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 (Meric et al., 2017). WALTZ has the highest accuracy, since its data have been validated by the experimental methods. No other existing amyloid prediction tools has such a high validation rate (Maurer-Stroh et al., 2010). The unique feature of PASTA and SODA is considering the intrinsic disorder of the protein as a major determinant of protein aggregation. Because the highly dynamical disordered regions of protein can increase its aggregation propensity (Chiti, 2003).

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 (Meric et al., 2017; Tsolis, 2013). The APRs identified by AMYLPRED and AMYPDB may be considered as representative of the result of other sequence-based APR prediction tools. Despite the differences in the algorithm and parameters used in different tools, it can be concluded that the regions 10–30, 40–60, 61–66, 88–120, and 130–140 play the most important roles in amyloid-forming propensity of rhKGF. These regions can be considered as the candidate positions to create the mutants with reduced aggregation propensity.
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 (Chennamsetty et al., 2010). 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 \( \beta \)-sheet forming propensity available for contacting the analogous sequences on the neighboring molecules. The APRs which are exposed after an unfolding event mediate protein aggregation through non-native pathways (Meric et al., 2017). Hence, one proposed solution to reduce protein aggregation is to enhance the conformational stability of the protein by increasing the free energy of unfolding (\( \Delta G_{\text{unf}} \)) (Meric et al., 2017). Using structure-based APR detection tools like SolubiS, Structurally corrected version of CamSol, Aggrescan3D, and SAP enabled us to make distinction between APRs mediating native (aggregation proceeding from the native monomeric state of the protein) and non-native aggregation (aggregation after partial/complete unfolding event). Aggrescan3D (A3D) and SAP are capable of detecting noncontiguous surface-exposed APRs in the large natively-folded globular proteins like monoclonal antibodies, and therefore recognizing APRs involved in the native-state aggregation (Meric et al., 2017; Nelson et al., 2005).

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. The high sap residues are considered as potential APRs and are the candidate positions for mutations against aggregation. There was no significant difference between the result of SAP calculation in the static, and dynamic states. According to the SAP algorithm, these are the hydrophobic residues which may expose under the normal dynamic fluctuations (Chennamsetty et al., 2010). These fluctuations may combine them with the other surface-exposed hydrophobic residues and lead to the formation of larger surface-exposed hydrophobic patches. The interaction of the hydrophobic patches on the neighboring molecules in the native state of the protein leads to the accumulation and aggregation of the protein molecules (Buck, 2012; Lee et al., 2013). 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 (Chennamsetty et al., 2010). 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 the protein aggregation tendency (Wang et al., 2009).

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. 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. Therefore, these regions may promote aggregation of the rhKGF without the occurrence 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) (Meric et al., 2017). The other regions are buried in the native state and contribute to non-native aggregation by a preceding unfolding event in the monomeric state of the protein. Increasing the protein conformational stability can prevent these regions from triggering aggregation. The structure-based predictions 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. The FGFR mutational analysis has indicated that some of these regions including M37-K57, which has been predicted as APR by most of the sequence-based APR detection tools, is important for the interaction of all FGFs with FGFR receptors (FGFR) with the RTVAV motif (residues 47–51) being specific for FGFR-FGFR interaction [44]. Furthermore, some of the APRs identified in rhKGF are also located in the heparin-binding sites including R11, R13, R14, Q20, K27, K101, and R121. Mutation of the residues involved in receptor/heparin-binding may disrupt the KGF biological activity (Chiti, 2003; Maurer-Stroh et al., 2010). Except for these regions, the other residues can be substituted with either hydrophilic or even less hydrophobic residues to design biologically active analogs of rhKGF with reduced aggregation propensity.

5. Conclusion

The aggregation propensity of rhKGF was analyzed using both sequence and structure-based APR predictors. 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 \( \beta \)-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. Using sequence- and structure-based APR prediction tools enabled us to distinguish buried and exposed APRs mediating non-native and native aggregation of rhKGF, respectively. As most of the predicted APRs in rhKGF, even those identified by the sequence-based tools are solvent-exposed in the native state of the protein, it can be concluded that the non-
covalent aggregation of rhKGF mainly occurs through the native pathway. Limiting the rate of aggregation can be achieved via increasing conformational stability or mutating the APRs to the less aggregation-prone residues. Some of these regions contribute toward receptor/heparin binding. This is a significant challenge to design variants of biotherapeutics with low vulnerability toward aggregation and unchanged biological activity. However, except for the functional residues, the others can be substituted with either hydrophilic or even less hydrophobic residues to design active analogs of rhKGF 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.

**Acknowledgments**

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.

**Disclosure statement**

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 PhD 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 MShD, and HR. The computational studies and data analysis was performed by MShD under the supervision of HR, and MG. Manuscript writing was mainly performed by MShD and it was revised by MG, HR, and MASH.

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Availability of data and materials**

All data generated or analyzed during this study are included in this published article.

**References**

Ahmed, A. B., & Kajava, A. V. (2013). Breaking the amyloidogenicity code: Methods to predict amyloids from amino acid sequence. FEBS Letters, 587(8), 1089–1095. https://doi.org/10.1016/j.febslet.2012.12.006

Bennigton, L., Rajan, G., Locher, C., & Lim, L. Y. (2020). Fibroblast growth factor 2—a review of stabilisation approaches for clinical applications. Pharmaceuticals, 12(6), 508. https://doi.org/10.3390/pharmaceuticals12060508

Buck, P. M. (2012). Computational methods to predict therapeutic protein aggregation. In Voynov Vladimir & Caravella Justin A. (Eds.), Therapeutic proteins (pp. 425–451). Springer.

Chen, B. L., Arakawa, T., Hsu, E., Narhi, L. O., Tressel, T. J., & Chien, S. L. (1994). Strategies to suppress aggregation of recombinant keratinocyte growth factor during liquid formulation development. Journal of Pharmaceutical Sciences, 83(12), 1657–1661. https://doi.org/10.1021/jps2600831204

Chennamsetty, N., Vovov, V., Kayser, V., Helk, B., & Trout, B. L. (2010). Prediction of aggregation prone regions of therapeutic proteins. The Journal of Physical Chemistry B, 114(19), 6614–6624. https://doi.org/10.1021/jp911706q

Chiti, F. (2003). Rationalization of the effects of mutations on peptide and protein aggregation rates. Nature, 424(6950), 805–808.

Chiti, F., Calamai, M., Taddei, N., Stefani, M., Ramponi, G., & Dobson, C. M. (2002). Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases. Proceedings of the National Academy of Sciences, 99(Suppl. 4), 16419–16426. https://doi.org/10.1073/pnas.212527999

Conchillo-Solé, O., de Groot, N. S., Avilés, F. X., Vendrell, J., Daura, X., & Ventura, S. (2007). AGGRESCAN: A server for the prediction and evaluation of “hot spots” of aggregation in polypeptides. BMC Bioinformatics, 8(1), 65. https://doi.org/10.1186/1471-2105-8-65

Courtois, F., Agrawal, N. J., Lauer, T. M., & Trout, B. L. (2016). Rational design of therapeutic mAbs against aggregation through protein engineering and incorporation of glycosylation motifs applied to bev-acizumab. mAbs, 8(1), 99–112. https://doi.org/10.1080/19420862.2015.1112477

Dastjerdeh, M. S., Marashiyan, M., Boroujeni, M. B., Golkar, M., Shokrgozar, M. A., & Rahimi, H. (2019). In silico analysis of different signal peptides for the secretory production of recombinant human keratinocyte growth factor in Escherichia coli. Computational Biology and Chemistry, 80, 225–233. https://doi.org/10.1016/j.compbiolchem.2019.03.003

Garvey, M., Meehan, S., Gras, S. L., Schirra, H. J., Craik, D. J., Van der Weerden, N. L., Anderson, M. A., Gerrard, J. A., & Carver, J. A. (2013). A radish seed antifungal peptide with a high amyloid fibril-forming propensity. Biochimica et Biophysica Acta, 1834(8), 1615–1623. https://doi.org/10.1016/j.bbabio.2013.04.030

Hsu, E., Osslund, T., Nybo, R., Chen, B.-L., Kenney, W. C., Morris, C. F., Arakawa, T., & Narhi, L. O. (2006). Enhanced stability of recombinant keratinocyte growth factor by mutagenesis. Protein Engineering, Design & Selection: PEDS, 19(4), 147–153. https://doi.org/10.1093/protein/gzq013

Huang, Z. (2012). A novel solid-phase site-specific PEGylation enhances the in vitro and in vivo biostability of recombinant human keratinocyte growth factor 1. PLoS one, 7(5), e36423.

Ko, J., Park, H., & Seok, C. (2012). GalaxyTBM: Template-based modeling by building a reliable core and refining unreliable local regions. BMC Bioinformatics, 13(1), 198. https://doi.org/10.1186/1471-2105-13-198

Kuhlman, B., & Baker, D. (2000). Native protein sequences are close to optimal for their structures. Proceedings of the National Academy of Sciences of the United States of America, 97(19), 10383–10388. https://doi.org/10.1073/pnas.97.19.10383

Lee, C. C., Perchiacca, J. M., & Tessier, P. (2013). Toward aggregation-resistant antibodies by design. Trends in Biotechnology, 31(11), 612–620. https://doi.org/10.1016/j.tibtech.2013.07.002

Maurer-Stroh, S., Debulpaepe, M., Kuenmerer, N., Lopez de la Paz, M., Martins, I. C., Reumers, J., Morris, K. L., Copland, A., Serpell, L., Serrano, L., Schymkowitz, J. W. H., & Rousseau, F. (2010). Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nature Methods, 7(3), 237–242. https://doi.org/10.1038/nmeth.1432

Meric, G., Robinson, A. S., & Roberts, C. J. (2017). Driving forces for non-native protein aggregation and approaches to predict aggregation-prone regions. Annual Review of Chemical and Biomolecular
Engineering, 8, 139–159. https://doi.org/10.1146/annurev-chembioeng-060816-101404

Nelson, R., Sawaya, M. R., Balbirnie, M., Madsen, A. Ø., Riekel, C., Grothe, R., & Eisenberg, D. (2005). Structure of the cross-beta spine of amyloid-like fibrils. Nature, 433(7043), 773–778. https://doi.org/10.1038/nature03680

Oliveberg, M. (2010). Waltz, an exciting new move in amyloid prediction. Nature Methods, 7(3), 187–188. https://doi.org/10.1038/nmeth0310-187

Paladin, L., Piovesan, D., & Tosatto, S. (2017). SODA: Prediction of protein solubility from disorder and aggregation propensity. Nucleic Acids Research, 45(W1), W236–W240. https://doi.org/10.1093/nar/gkw412

Pawlicki, S., Béche, A. L., & Delamarche, C. (2008). AMYPdb: A database dedicated to amyloid precursor proteins. BMC Bioinformatics, 9(1), 273. https://doi.org/10.1186/1471-2105-9-273

Poorebrahim, M., Sadeghi, S., Ghorbani, R., Asghari, M., Abazari, M. F., Rajan, R. S., Li, T., & Arakawa, T. (2010). Case studies involving protein structure prediction and aggregation. Protein Science: A Publication of the Protein Society, 19(9), 1367–1378. https://doi.org/10.1002/pro.51

Raj, M., & Kinne, R. (2020). Mechanistic insights into protein stability and self-aggregation in GLUT1 genetic variants causing GLUT1-deficiency syndrome. The Journal of membrane biology, 1–13.

Rajan, R. S., Li, T., & Arakawa, T. (2010). Case studies involving protein aggregation. In Wang Wei & J. Roberts Christopher (Eds.), Aggregation of therapeutic proteins (pp. 367–401). Wiley.

Rousseau, F., Schymkowitz, J., & Serrano, L. J. (2006). Protein aggregation and amyloidosis: Confusion of the kinds? Current Opinion in Structural Biology, 16(1), 118–126. https://doi.org/10.1016/j.sbi.2006.01.011

Sawaya, M. R., Sambashivan, S., Nelson, R., Ivanova, M. I., Sievers, S. A., Apostol, M. I., Thompson, M. J., Balbirnie, M., Wilczius, J. J. W., McFarlane, H. T., Madsen, A. Ø., Riekel, C., & Eisenberg, D. (2007). Atomic structures of amyloid cross-beta spines reveal varied steric zippers. Nature, 447(7143), 453–457. https://doi.org/10.1038/nature05695

Siepen, J. A., Radford, S. E., & Westhead, D. (2003). Beta edge strands in protein structure prediction and aggregation. Protein Science, 12(10), 2348–2359. https://doi.org/10.1100/p.03234503

Sormanni, P., Aprile, F. A., & Vendruscolo, M. (2015). The CamSol method of rational design of protein mutants with enhanced solubility. Journal of Molecular Biology, 427(2), 478–490. https://doi.org/10.1016/j.jmb.2014.09.026

Tartaglia, G. G., & Vendruscolo, M. J. C. S. R. (2008). The zyggregator method for predicting protein aggregation propensities. Chemical Society Reviews, 37(7), 1395–1401. https://doi.org/10.1039/b706784b

Trovato, A., Chiti, F., Maritan, A., & Seno, F. (2006). Insight into the structure of amyloid fibrils from the analysis of globular proteins. PLoS Computational Biology, 2(12), e170. https://doi.org/10.1371/journal.pcbi.0020170

Tsolis, A. C. (2013). A consensus method for the prediction of ‘aggregation-prone’ peptides in globular proteins. PLoS one, 8(1), e54175.

Tsolis, A. C., Papandreou, N. C., Iconomidou, V. A., & Hamodrakis, S. J. (2013). A consensus method for the prediction of ‘aggregation-prone’ peptides in globular proteins. PLoS One, 8(1), e54175. https://doi.org/10.1371/journal.pone.0054175

van der Kant, R. (2019). Solubis: Optimizing protein solubility by minimal point mutations. In Protein misfolding diseases (pp. 317–333). Springer.

van der Kant, R., Karow-Zwick, A. R., Van Durme, J., Blech, M., Gallardo, R., Seeliger, D., Alifalq, K., Baatsen, P., Compenolle, G., Gils, A., Studts, J. M., Schulz, P., Gardel, P., Schymkowitz, J., & Rousseau, F. (2017). Prediction and reduction of the aggregation of monoclonal antibodies. Journal of Molecular Biology, 429(8), 1244–1261. https://doi.org/10.1016/j.jmb.2017.03.014

Walsh, I., Seno, F., Tosatto, S. C. E., & Trovato, A. (2014). PASTA 2.0: An improved server for protein aggregation prediction. Nucleic Acids Research, 42(Web Server issue), W301–W307. https://doi.org/10.1093/nar/gku399

Wang, W., & Roberts, C. J. (2010). Aggregation of therapeutic proteins. Vol. 100. Wiley Online Library.

Wang, X., Das, T. K., Singh, S. K., & Kumar, S. (2009). Potential aggregation prone regions in biotherapeutics: A survey of commercial monoclonal antibodies. mAbs, 1(3), 254–267. https://doi.org/10.4161/mabs.1.3.8035

Wang, X., Westhead, D., & Roberts, C. J. (2010). MABBanner: A web-based tool for rapid detection of aggregation-prone residues in proteins. Protein Science: A Publication of the Protein Society, 19(9), 1631–1644. https://doi.org/10.1002/pro.51

Zahrara, B., Reza, K. H., & Javad, M. S. (2016). Expression of the full-length human recombinant keratinocyte growth factor in Pichia pastoris. Journal of Cell and Molecular Research, 8(1), 1–7.

Zambrano, R., Jamroz, M., Szczasniuk, A., Pujols, J., Kmiecik, S., & Ventura, S. (2015). AGGRESCAN3D (A3D): Server for prediction of aggregation properties of protein structures. Nucleic Acids Research, 43(W1), W306–W313. https://doi.org/10.1093/nar/gkv359

Zibaee, S., Makin, O. S., Goedert, M., & Serpell, L. C. (2007). A simple algorithm locates beta-strands in the amyloid fibril core of alpha-synuclein, abeta, and tau using the amino acid sequence alone. Protein Science: A Publication of the Protein Society, 16(5), 906–918. https://doi.org/10.1110/ps.062624507