Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company’s public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
A proteomics-MM/PBSA dual approach for the analysis of SARS-CoV-2 main protease substrate peptide specificity

Gloria Gallo a, Uilla Barcick a, Camila Coelho a, Murilo Salardani a, Maurício F. Camacho a, Daniela Cajado-Carvalho b, Flávio V. Loures a, Solange M.T. Serrano b, Leon Hardy c, André Zelanis a, Martin Würtele a,∗

a Department of Science and Technology, Federal University of São Paulo, São José dos Campos, Brazil
b Laboratory of Applied Toxicology, Center of Toxins, Immune-Response and Cell Signaling (CoTICS), Butantan Institute, São Paulo, Brazil
c Department of Physics, University of South Florida, Tampa, United States

ARTICLE INFO

Keywords:
Proteomics
TAILS
Free energy calculations
MM/PBSA
COVID-19
Main protease

ABSTRACT

The main protease M pro of SARS-CoV-2 is a well-studied major drug target. Additionally, it has been linked to this virus’ pathogenicity, possibly through off-target effects. It is also an interesting diagnostic target. To obtain more data on possible substrates as well as to assess the enzyme’s primary specificity a two-step approach was introduced. First, Terminal Amine Isobaric Labeling of Substrates (TAILS) was employed to identify novel M pro cleavage sites in a mouse lung proteome library. In a second step, using a structural homology model, the MM/PBSA variant MM/GBSA (Molecular Mechanics Poisson-Boltzmann/Generalized Born Surface Area) free binding energy calculations were carried out to determine relevant interacting amino acids. As a result, 58 unique cleavage sites were detected, including six that displayed glutamine at the P1 position. Furthermore, modeling results indicated that M pro has a far higher potential promiscuity towards substrates than expected. The combination of proteomics and MM/PBSA modeling analysis can thus be useful for elucidating the specificity of M pro and thus open novel perspectives for the development of future peptidomimetic drugs against COVID-19, as well as diagnostic tools.

1. Introduction

SARS-CoV-2 spread worldwide rapidly, initiating the COVID-19 pandemic. Although several effective vaccines and small molecule inhibitors were rapidly developed, COVID-19 remains of great concern to global public health [1,2]. Due to several issues, including the emergence of new variants of SARS-CoV-2, global efforts for drug discovery to address the disease continue being of uttermost importance.

SARS-CoV-2 is an enveloped (+) single-stranded RNA coronavirus. Its genome size is about 29.9 kb in length. About two thirds of SARS-CoV-2 genome is occupied by orf1ab that encodes the non-structural proteins, while the remaining region next to the 3′ end encodes the structural proteins [3]. Orf1ab is translated into two polyproteins. They are processed by the virus’s main protease M pro (also called 3CL pro) and a second papain-like protease (PL pro) [4]. The structure of M pro from SARS-CoV-2, a protein with 96% sequence identity to M pro from SARS-CoV, was recently solved [5,6].

Because of its essential activity in the viral replication machinery, M pro, a cysteine protease, is an excellent target for the rational development of SARS-CoV-2 medications [7–11]. Additionally, it is considered an important marker for diagnostic purposes [12,13]. Its off-target activity is especially intriguing, as it could be linked to SARS-CoV-2 pathogenicity [14]. For all these reasons, we here have opted for a dualistic approach. First, we employed Terminal Amine Isobaric Labeling of Substrates (TAILS; [15]) for the identification of M pro cleavage sites in a complex proteome (mouse lung proteome, as described

Abbreviations: M pro, Main protease; TAILS, Terminal Amine Isobaric Labeling of Substrates; MM/PBSA, Molecular Mechanics Poisson-Boltzmann; MM/GBSA, Molecular Mechanics Generalized Born Surface Area; PL pro, second papain-like protease; MD, molecular dynamics; SASA, solvent accessible surface area; TEV, Tobacco Etch Virus protease; GuHCl, Guanidine hydrochloride; DTT, dithiothreitol; IAA, Iodoacetamide; 4FDA, 4-formylbenzene−1,3-disulfonate acid; ACN, acetonitrile; SCX, strong cation exchange; LC−MS/MS, nanoflow liquid chromatography/tandem mass spectrometry; CID, Collisional Induced Dissociation; AGC, Automatic Gain Control; FDR, False Discovery Rate; PME, Particle Mesh Ewald; SEM, Standard Error of the Mean.

∗ Correspondence to: Department of Science and Technology, Federal University of São Paulo, Talim 330, São José dos Campos, 12231-280, Brazil.

E-mail address: martin.wurtele@unifesp.com (M. Würtele).

https://doi.org/10.1016/j.peptides.2022.170814
Received 8 March 2022; Received in revised form 15 May 2022; Accepted 16 May 2022
Available online 26 May 2022
0196-9781/© 2022 Elsevier Inc. All rights reserved.
Then, as a second step, molecular modelling techniques via MM/PBSA free energy calculations were carried out to determine the amino acids that are most relevant for the interaction with protein substrates.

TAILS is an N-terminomic analytical approach that enables the profiling of the substrate repertoire of a given protease (degradome) in complex biological mixtures, therefore allowing subsite mapping based in the use of mass spectrometry and bioinformatics for the identification of (new) N-termini generated by active proteases. Since the protease is incubated with native substrates in complex biological mixtures such as cell lysates, long-range interactions (e.g. the interaction of protease exosites with their substrates) are potentially allowed under assay conditions. Such a feature is the main advantage of TAILS over approaches involving the incubation of protease with peptide libraries, in which mainly short-range interactions are allowed. Briefly, using dimethyl labeling, prime-side peptides (i.e. those that were generated by active proteases in the sample) are labeled at their N-termini, therefore indicating the cleavage site. After the addition of a protease such as trypsin, peptides bearing free N-termini are produced. This latter set of peptides is then sulfo-modified and pulled out from the sample after strong cation exchange chromatography. The resulting peptide set (generated by active proteases in the sample) is submitted to mass spectrometry followed by bioinformatic analyses for protease specificity profiling.

The MM/PBSA and its variant MM/GBSA (Molecular Mechanics Poisson-Boltzmann Surface Area and Molecular Mechanics Generalized Born Surface Area) are popular methods used to determine the free binding energy between a protein and a ligand [16–19]. Essentially, these methods evaluate the free binding energy of the ligand from MD simulation snapshots using energy terms of the MD force field, but approximating the polar energy terms with the Poisson-Boltzmann (PB) equation, estimating the non-polar energy terms using the solvent-accessible surface area (SASA) method and roughly approximating the entropic terms using normal-mode analysis. They have proven to be reliable methods for the computation of binding energy leading to numerous applications in small-molecule ligand binding and drug design [20–22], protein-peptide interactions [23–26], protein-protein interactions [27,28] and protein design [29,30]. Although their accuracy in some systems is in the order of 10 kJ/mol, both computational improvements in the calculation of individual terms of the binding energy as well as carrying out many short MD simulations [31,32] has helped obtain more accurate estimations of binding energies up to the 1 kJ/mol range. As a whole, the accuracy of MM/PBSA methods ranks in between more simple semi-empirical molecular docking calculations and more computational demanding alchemical perturbation (AP) methods [33].

2. Material and methods

2.1. Protein expression and purification

SARS-CoV-2 M\textsuperscript{pro} (GenBank entry MT358641.1) was synthesized (GenScript, USA) and cloned into the pET21a plasmid to create an expression construct with an N-terminal cleavable His-tag for recombinant expression in E.coli BL21(DE3), purification by affinity chromatography and size exclusion chromatography after cleavage of the His-tag using TEV protease, as previously described [34].

2.2. N-terminomic analysis - terminal amine isotopic labeling of substrates (TAILS)

N-terminomic analysis was carried out using the Terminal Amine Isotopic Labeling of Substrates (TAILS) protocol [15] with modifications [35]. A lung lysate from C57BL/6 male mice (n = 2) was obtained after submitting the tissue samples to 3 cycles of 16,000 rpm of homogenization in a homogenizer (Bio-Gen PRO200 homogenizer PRO Scientific, Oxford CT, USA), on ice bath using a lysis buffer (50 mM HEPES 150 mM NaCl 1 mM EDTA, pH 7.5). The tissue lysate was cleared by centrifugation (14,000 x g, 10 min, 4 °C) and the protein content of the supernatant was measured using the Bradford method. Five hundred micrograms of proteins were incubated for 12 h with 10 μg of recombinant M\textsuperscript{pro} or as the control with 50 mM HEPES buffer alone (pH 7.5) at 37 °C. Guanidine hydrochloride (GuHCl) was added to the mixture to a final concentration of 3 M, followed by the addition of diithiothreitol (DTT) to a final concentration of 5 mM. The mixture was incubated at 37 °C for 60 min. Iodoacetamide (IAA) was then added to a final concentration of 15 mM and the samples were incubated for 30 min at room temperature, in the dark. To quench the excess of IAA, DTT was added to a final concentration of 15 mM. N-termini were differentially labeled via stable-isotope reductive dimethylation with either light (for controls) or heavy (for M\textsuperscript{pro} experiments) formaldehyde solutions. Protein samples from each experimental condition were incubated overnight at 37 °C with either light or heavy sodium cyanoborohydride (NaBH\textsubscript{4}CN, light, or NaBD\textsubscript{4}CN, heavy) to a final concentration of 20 mM followed by the addition of formaldehyde \textsuperscript{12}C\textsubscript{2}H\textsubscript{4}O (light) or \textsuperscript{13}CD\textsubscript{2}O (heavy) to a final concentration of 40 mM, resulting in mass differences of + 28.031300 Da and + 36.075670 Da for the light and heavy-labeled samples, respectively. The reaction was terminated by adding 1 M Tris (pH 6.8; to a final concentration of 100 mM) to each sample and the mixture was incubated for 2 h at 37 °C. Samples were then combined at a 1:1 ratio. Clean-up of samples was performed by the addition of ice-cold acetone (8 volumes) and methanol (1 vol), followed by the incubation of samples for 3 h at – 80 °C. After centrifugation at 14,000 g for 10 min, 4 °C, protein pellets were washed twice with one volume of ice-cold methanol and then resubilbilized with 100 mM NaOH solution (final concentration of 2.5 mM), followed by the addition of HEPES buffer, pH 7.5, to a final concentration of 25 mM in 500 μL of reaction solution. Trypsin (Proteomics grade, Sigma, USA) was added at a 1:100 ratio (enzyme/substrate) and the mixture was incubated at 37 °C for 18 h. N-terminal peptides were enriched by negative selection using sodium 4-formylbenzene-1,3-disulfonate as described [35]. The peptide mixture was subjected to a three-step disulfomodification procedure [35]. Briefly, 4-formylbenzene – 1,3-disulfonate acid (4FDA) was added to the peptide mixture solution to a final concentration of 20 mM, followed by the addition of NaBH\textsubscript{4}CN 20 mM (final concentration). The reaction was incubated at room temperature for 1 h. Two subsequent steps of incubation with 40 mM of 4FDA and 40 mM of NaBH\textsubscript{4}CN, and 60 mM of 4FDA and 60 mM of NaBH\textsubscript{4}CN were performed for 1 h and 18 h, respectively. The reaction was terminated by adding 1 M Tris (pH 6.8) to a final concentration of 100 mM, for 1 h at room temperature. After desalting using C18 cartridges (3 M Empore™, SPE Extraction disks, USA) peptide samples were dried in a SpeedVac, resuspended in 300 μL of 1% acetic acid and 15% acetonitrile (ACN) and subjected to strong cation exchange (SCX) using SCX StageTips [36]. While disulfomodified (trypptic) peptides were mainly eluted in the flowthrough, N-terminal blocked peptides (either by chemical dimethylation or naturally acetylated), corresponding to natural or neo-N-terminal peptides, were eluted in two fractions collected after the addition of 200 mM and 500 mM of NaCl in 1% acetic acid 15% ACN, respectively. Finally, fractions were desalted using C18 StageTips, dried in a SpeedVac and redissolved in 50 μL of 0.1% formic acid prior to nanoflow liquid chromatography/tandem mass spectrometry (LC–MS/MS) analysis.

2.3. Mass spectrometric analysis

An aliquot (5 μL) of the desalted peptide mixture was injected into a trap column packed with C18 (100 μm i.d. × 2 cm) for desalting with 100% solvent A (0.1% formic acid). Peptides were then eluted onto an analytical column (75 μm i.d. × 100 mm) packed in house with Aqua® C-18 5 μm beads (Phenomenex, USA). Nanoflow liquid chromatography was performed on an Easy nanoLC system (Thermo Fisher Scientific, USA) coupled to an LTQ-Orbitrap Velos mass spectrometer (Thermo
Fisher Scientific, USA). Peptides were loaded onto the column with solvent A (0.1% formic acid) and eluted with a 90 min linear gradient. Database searches using the MS data.

### 2.5. Homology modelling

Galo et al. (2015) suggested the TAILS approach targets prime-side peptides, non-prime side positions (i.e. P5 to P1) were bioinformatically inferred from peptide identifications, after database searches using the MS data.

| Cleavage site (P5-P5+)/| Identified peptide (P1-‘P’)/| Gene name | UniProt Accession | Protein name |
|-----------------------|-----------------------------|-----------|-------------------|--------------|
| GAEIQ/DGRFN           | DGBNLKVKQQR                | FIP1      | Q9DB24.4          | Pre-mRNA 3’-end-processing factor FIP1 |
| GETEQ/KRRK            | KRRKKKAKKRK                | CA131     | Q8CI4             | Uncharacterized protein C1orf131 homolog |
| SPSQG/SQDTO           | SGDTQFQKLQKLR              | GOGA4     | Q91VWS            | Golgin subfamily A member 4 |
| QLSNQQ/LFVGG          | LFVGGSSQKGGFFGCR           | CTP5C     | Q0V177            | Contactin-associated protein like 5–3 |
| RISSQQ/EDDDQ          | EDDQEFFEDR                 | FBX43     | Q8CD62            | F-box only protein 42 |
| SSSQVQ/IDPPL          | IDDPLLQVRQRTKQR            | CO8A1     | Q2UY11            | Collagen alpha 1(XXVIII) chain |

* Since TAILS approach targets prime-side peptides, non-prime side positions (i.e. P5 to P1) were bioinformatically inferred from peptide identifications, after database searches using the MS data.

### 2.6. Molecular dynamics (MD)

MD simulations were carried out using the AMBER simulation software package [41]. The model was prepared with tleap using the ff99 AMBER force field, neutralized with Na+ ions and solvated with the TIP3P water model in a cuboid integration box with 12 Å solvent margi.ns. After structure minimization, a short heating run and a density equilibration run, several parallel 10 ns MD production runs were conducted with AMBER. Production runs were carried out with the GPU-version of pmemd at 300 K in time steps of 2 fs using a constant pressure with isotropic position scaling. Particle Mesh Ewald (PME) model of periodic boundary conditions with an 8 Å classical non-bonded cut-off, a Langevin thermostat and the SHAKE algorithm to constrain the bonds of all hydrogen atoms.

### 2.7. MM/GBSA

Essentially, a molecular dynamics (MD) simulations of protein-peptide complexes were carried out, and from single snapshots of the MD-trajectory, the binding free energy of the peptides to the protein $\Delta G_{\text{bind}}$ was calculated from protein-peptide complex MD simulations as an averaged difference between the free energies of the complex state ($G_{\text{bind}}$) and unbound states of the protein ($G_P$) and the peptide ligand ($G_L$):

$$\Delta G_{\text{bind}} = \langle G_{\text{bind}} \rangle - \langle G_P \rangle - \langle G_L \rangle$$

The states were evaluated using following approximation:

$$G = E_{\text{bonded}} + E_T + E_{\text{el}} + G_{\text{polar}} + G_{\text{empirical}} - TS$$

In the second equation, $E_{\text{bonded}}$ is the bonded, $E_T$ is the electrostatic and $E_{\text{el}}$ is the Van der Waals energy component from the MM force field. $G_{\text{polar}}$ is the polar solvation energy component. $G_{\text{polar}}$ is approximated using the generalized Born (GB) model approximation for solving the Poisson-Boltzmann (PB) equation. $G_{\text{empirical}}$ is the non-polar solvation energy contribution and is estimated using a linear approximation including the solvent accessible surface area (SASA). Finally, TS is the entropic term, which was approximated using normal mode analysis [41–43]. In detail, the MM/GBSA calculations were carried with the AMBER MMPBSA.py script [41] using the “OBC” model (gb=2) [44] together with mbond2 atomic radii. At least 20 independent 10 ns production runs were carried out for each peptide. For each run, 25 snapshots for the Δ$\text{H}_{\text{GBSA}}$ and 5 snapshots for the TΔS calculations were evaluated. After averaging results, error values were calculated as SEM (Standard Error of the Mean) from the independent runs.

### 2.8. Ethics statement and animal model

Animal experiments were performed in accordance with the Brazilian Federal Law 11,794 establishing procedures for the scientific use of animals in accordance with the principles outlined by the Brazilian
College of Animal Experimentation (CONCEA), and the State Law establishing the Animal Protection Code of the State of São Paulo. All protocols adopted in this study were approved by the local ethics committee for animal experiments from the Federal University of São Paulo (Permit Number: 9552180122). All efforts were made to minimize suffering. Animals (C57BL/6 mouse) were sacrificed by intraperitoneal injection of ketamine and xylazine.

3. Results and discussion

3.1. Mpro specificity screening using TAILS

The N-terminomic analysis of the proteolytic activity of Mpro upon a mouse lung proteome library resulted in the identification of 58 unique cleavage sites. Thus, 58 peptides, which were absent in the control condition, were identified as heavy dimethylated-labeled (singleton) peptides (Table S1 and Table S2). Interestingly, out of the 58 unique cleavage sites, six contained Glutamine (Q) at the P1 position (Table 1). The cleaved proteins include contactin-associated protein like 5 (i.e., other than Mpro), which is possibly involved in damage repair processes [46] . These proteins might be bonicide substrates of Mpro. However, the physiological implications of their cleavage must be further validated. Thus, we chose to focus on structural aspects of Mpro subsite specificity.

The primary specificity of Mpro for residues containing a Glutamine at the P1 position was also observed in recent reports on Mpro cleavage specificity [14,47] and is in accordance with the primary specificity observed for SARS-CoV-2 and SARS-CoV main proteases [5,6,48]. Recent N-terminomics reports on Mpro specificity used a protease to proteome ratio ranging from 1:20 to 1:50. This might explain why we identified peptides spanning the canonical Mpro preference of Q at P1. Moreover, for the cleavage sites containing Q at P1 position, we did not verify the recurrent presence of L at P2, as recently reported [14, 45]. On the other hand, most cleavage sites (~90%) varied in relation to the residues at the scissile bond (P1–P1’ positions), and, notably, they deviated from the canonical presence of Q at P1 position in the set of identified peptides. In this context, we cannot rule out the activation of other proteases of mouse lung tissue upon incubation with Mpro therefore resulting in peptides spanning distinct cleavage sites (i.e. other than Q at P1). Thus, we focused our structural analysis on peptides in which the preference of Q at P1 was observed (Table 1).

3.2. MM/GBSA binding analysis

In order to estimate the free binding energy of the different detected Mpro binding peptides, MM/GBSA analysis was performed. To obtain reproducible data several 10 ns MD production runs were carried out.

| Peptide      | ΔHGBSA (kcal/mol) | ΔS (kcal/mol) | ΔG (kcal/mol) |
|--------------|-------------------|---------------|---------------|
| GAIEQDRFRN   | -64.5 + 0.9       | -37.9 + 1.2   | -26.6 + 1.3   |
| GETEQQRBRK   | -74.3 + 1.1       | -47.5 + 1.0   | -26.8 + 1.6   |
| KRSSQEDDOQ   | -73.7 + 1.2       | -45.3 + 1.1   | -28.4 + 1.3   |
| QLNNQTVFVGG  | -78.6 + 1.0       | -42.7 + 1.6   | -35.9 + 1.7   |
| SSSVQGDDTQ   | -56.2 + 1.1       | -39.9 + 1.5   | -16.3 + 1.6   |
| SSSVQIDPPL   | -62.2 + 0.7       | -37.7 + 1.1   | -24.5 + 1.4   |

3.3. Binding energy decomposition analysis

To further discern the importance of the amino acids to the free
energy of binding, binding energy decomposition analysis was carried out with the AMBER MMPBSA.py script. The obtained results can be considered a measure of the interaction energy of each amino acid with the rest of the system. They are shown in Fig. 2. In this figure, the importance of the P1 and P2 residues, for peptide recognition by Mpro become apparent. To complement this analysis, in Fig. 3, the different MM/GBSA energy components, for both the peptides and the protein are highlighted. As shown in Table 2, due to the compensation effect of electrostatic interaction energy by polar solvent interaction energies, in the employed MM/GBSA model, the binding is determined more by Van der Waals interactions, not electrostatic interactions.

### Table 3

**MM/GBSA energy components.** Different energy components from the MM/GBSA analysis of the binding of the peptides displaying the canonical Gln at P1 position to SARS-CoV-2 Mpro are given in kcal/mol.

| Peptide     | MM/GBSA Term | Ligand Energy | Ligand SEM | Receptor Energy | Receptor SEM | Total Energy | Total SEM |
|-------------|---------------|----------------|-------------|-----------------|--------------|--------------|-----------|
| GAEEIQDGRFN | ΔHGBSA        | -25.0          | 1.3         | -39.8           | 0.6          | -64.8        | 1.9       |
|             | ΔGnsolv       | -6.80          | 0.12        | -4.83           | 0.04         | -11.63       | 0.16      |
|             | ΔGpol         | 38.2           | 4.3         | 21.4            | 1.2          | 59.7         | 5.5       |
|             | ΔEel          | -11.7          | 4.7         | -11.7           | 1.4          | -23.4        | 6.1       |
| GETEQKRIRK  | ΔHGBSA        | -23.3          | 1.3         | -51.8           | 0.6          | -75.1        | 1.9       |
|             | ΔGnsolv       | -7.36          | 0.12        | -5.20           | 0.03         | -12.57       | 0.15      |
|             | ΔGpol         | 160.8          | 5.4         | 130.2           | 1.4          | 290.9        | 6.8       |
|             | ΔEel          | -129.1         | 5.6         | -129.1          | 1.7          | -258.2       | 7.3       |
| KRSQEDDDQE  | ΔHGBSA        | -47.6          | 0.9         | -47.6           | 0.3          | -95.3        | 1.2       |
|             | ΔGnsolv       | -7.29          | 0.11        | -5.25           | 0.02         | -12.55       | 0.13      |
|             | ΔGpol         | 61.0           | 6.3         | 30.7            | 3.5          | 91.7         | 9.9       |
|             | ΔEel          | -26.6          | 7.1         | -26.6           | 4.2          | -53.2        | 11.2      |
|             | ΔEel          | -49.6          | 1.1         | -49.6           | 0.3          | -99.3        | 1.5       |
| QLNSQLFVG   | ΔHGBSA        | -35.4          | 1.4         | -43.0           | 0.5          | -78.4        | 2.0       |
|             | ΔGnsolv       | -7.16          | 0.13        | -4.95           | 0.03         | -12.12       | 0.16      |
|             | ΔGpol         | 90.2           | 2.5         | 80.5            | 1.3          | 170.7        | 3.8       |
|             | ΔEel          | -70.2          | 3.0         | -70.2           | 1.6          | -140.3       | 4.6       |
|             | ΔEel          | -48.4          | 1.0         | -48.4           | 0.3          | -96.7        | 1.5       |
| SSPGQSGD    | ΔHGBSA        | -19.1          | 1.4         | -37.0           | 0.6          | -56.1        | 2.0       |
|             | ΔGnsolv       | -6.05          | 0.13        | -3.99           | 0.03         | -10.03       | 0.16      |
|             | ΔGpol         | 53.1           | 3.0         | 33.3            | 1.3          | 86.4         | 4.3       |
|             | ΔEel          | -29.3          | 3.5         | -29.3           | 1.6          | -58.6        | 5.1       |
|             | ΔEel          | -36.9          | 1.1         | -36.9           | 0.4          | -73.8        | 1.5       |
| SSVDIPPL    | ΔHGBSA        | -26.2          | 1.2         | -36.5           | 0.4          | -62.7        | 1.6       |
|             | ΔGnsolv       | -6.28          | 0.10        | -4.51           | 0.02         | -10.79       | 0.12      |
|             | ΔGpol         | 44.9           | 2.7         | 32.7            | 1.1          | 77.6         | 3.8       |
|             | ΔEel          | -21.7          | 3.1         | -21.7           | 1.2          | -43.4        | 4.3       |
|             | ΔEel          | -43.0          | 0.9         | -43.0           | 0.4          | -86.1        | 1.2       |

**Fig. 2.** Per-residue binding energy decomposition of peptides. A) GAEIQDGRFN. B) GETEQKRIRK. C) KRSQEDDDQ. D) QLNSQLFVG. E) SSPGQSGD. F) SSVDIPPL. Using MM/GBSA analysis, the interaction energy of each amino acid with the rest of the system was estimated. The importance of the residues at P1 and P2 positions become apparent. Variations at e.g. P4 and P1’ positions could also explain improved binding of the corresponding peptides. Values are given in kcal/mol.
P2 Ile and P1 Gln side chains and to a lesser degree with the P4’ Phe and P4 Ala side chains (Fig. 2A).

Compared to this peptide, the GETEQKRIRK peptide shows a similar binding affinity (∆G = -26.8 kcal/mol) in the modeling studies. In this case, the per-residue MM/GBSA energy decomposition analysis indicates that the less favorable substitutions Ile/Glu (at P2), Ala/Glu (P4) and Phe/Arg (P4’) are mostly countered by several more favorable substitutions (e.g. Glu/Thr at P3, Asp/Lys at P1’ and Gly/Arg at P2’, Fig. 2A and B). Analyzing the ligand MM/GBSA energy components, improvement of the electrostatics for binding (∆Eel) is reversed by less favorable solvent free energies (∆GSolv, Table 3). On the side of the protein, electrostatics are more favorable in comparison with the GAEIQDGRFN peptide. However, the resulting more favorable MM/GBSA binding energy term (∆HGBSA) is weakened by a more unfavorable ∆TS value (Table 3). Entropy evaluations with MM/PBSA or MM/GBSA methods must be interpreted carefully, due to obvious methodological limitations [16]. In the current analysis, most peptides showed however a similar conformational entropy (-T∆S) value of about 41.8 kcal/mol. Because this -T∆S value is similar, the task of ranking the binding can be thus in most cases simplified in first approximation to comparison of the other binding energy values. The somewhat atypical -T∆S value of 47.5 kcal/mol for the GETEQKRIRK peptide is possibly the result of its increased number of large, positively charged residues, which will induce more vibrational and rotational movements, including side-chain rotamer states [49,50]. These states are stabilized upon binding, leading to a reduction of conformational entropy. This effect is expected to be very significant when these residue side-chains bind to the interface of a protein. This is certainly the case of the all the three interacting residues of the KRIRK sequence, except the last lysine, which presumably remains mostly solvent exposed after peptide binding.

Compared with GAEIQDGRFN, the KRSQQEDDQE peptide binding energy decomposition shows binding energy improvements due mostly to the Ala/Arg (at P4) that are however compensated by Ile/Gln(P2) and Phe/Gln (P4’) substitutions (Fig. 2C). Similar to GETEQKRIRK, the MM/GBSA energy components (Table 3) indicate that electrostatic improvements on the receptor side are neutralized by entropic disadvantages. As a whole the free energy of binding remains similar to the other described peptides (∆G = -28.4 kcal/mol).

The QLNSQLFVGG is the peptide with the most favorable ∆G value (-35.9 kcal/mol). Compared with GAEIQDGRFN, the most important favorable substitution is Ala to Leu at P4, which apparently fills the S4 subsite’s steric restraints better. The other improvements are at the P1’ (Asp to Leu), P2’ (Gly to Phe) and P3’ (Arg to Val) sites (Fig. 2D). On the other hand, QLNSQLFVGG has some less favorable substitutions at the P2 (Ile to Ser) and P4’ (Phe to Gly) positions.

The last two peptides considered in this study are SSPSQSGDTQ (∆G = -16.3 kcal/mol) and SSSVQIDPPL (∆G = -24.5 kcal/mol). In the case of the below average ∆HGBSA of SSPSQSGDTQ, MM/GBSA binding energy decomposition indicates most disadvantages in comparison with GAEIQDGRFN are due to an Ile/Ser substitution at P2 position and a Phe/Thr substitution at P4’ position (Fig. 2E). In the case of SSPSQIDPPL, which has a slightly less favorable binding enthalpy when compared with GAEIQDGRFN, MM/GBSA binding enthalpy indicates only slight improvements mainly through the P1’ Asp to Ile substitution in energy decomposition analysis (Fig. 2F). However, these small improvements are reversed by decrements in negative binding enthalpy in the protein part of the decomposition. These decrements are mostly related to both more unfavorable electrostatics (∆Eel and ∆EGSolv) and Van der Waals (∆Evdw) contacts being involved in the binding of the SSPSQIDPPL peptide (Table 3).

In conclusion, peptide binding to the Mpro active site does not seem to be driven mainly by electrostatics in any of the studied peptides. Furthermore, although rather specific, the site does show certain potential promiscuity towards binding variations of the peptides. From a structural point of view, the tight binding groove which binds to the N-terminal peptide shows correspondingly higher binding affinities in the MM/GBSA analysis than the more open C-terminal binding sites. As far as entropic calculations are reliable in the MM/GBSA method, the peptides GETEQKRIRK and KRQSQEDDQE, both containing several charged residues, show an above average unfavorable conformational entropy term.

3.4. Peptide design

The obtained results indicate that the ∆HGBSA energy term can be significantly improved by substitutions. These subtle substitutions allowed us, based on this limited dataset, to extrapolate a peptide with the sequence GRNIQVFVK. Using the MM/GBSA methodology, the following energetic values for this peptide were obtained for this peptide: ∆HGBSA = -97.7 ± /- 0.2 kcal/mol, ∆TS = 47.3 ± /- 0.2 kcal/mol, ∆G = -50.4 ± /- 0.2 kcal/mol. Peptide binding energy decomposition analysis confirms high energy of binding for the P4, P3, P2, P1,
P1’, P2’ and P4’ positions (Fig. S1A). Energy components from MM/GBSA indicates that strong electrostatic attraction is compensated by unfavorable polar solvation energy (Fig. S1C). The resulting high binding affinity indicates this approach can indeed be applied to obtain potent Mpro binding peptides.

3.5. Non-canonical peptides

Additionally to the canonical peptides, MM/GBSA analysis was carried out with all the non-canonical i.e. peptides without a Q at P1) identified in TAILS analysis (Table S2 and Table S3). Interestingly, these peptides showed an average \( \Delta G \) of \( -27.3 ± 1.2 \) kcal/mol similar to the peptides bearing the canonical cleavage sites in the MM/GBSA (\( -26.4 ± 0.4 \) kcal/mol) analysis. The average per-residue energy decomposition for the canonical peptide amino acids is given in Fig. 4A and for the non-canonical peptides in Fig. 4B. This image confirms the importance of the amino acid residues at the P4, P2 and P1 positions. Similar results could be obtained for the canonical as non-canonical peptides. Finally, a decomposition matrix with average decomposition enthalpies (\( \Delta H_{\text{GBSA}} \)) was calculated for each of the 20 proteinogenic amino acids at each position (Table S4). Using this kind of analysis, an idealized peptide with the sequence PLIRKQLQYP could be extrapolated. MM/GBSA analysis led to following energetic values for this peptide with above average binding free energy: \( \Delta H_{\text{GBSA}} = -94.3 \) kcal/mol and \( \Delta G = -51.8 \) kcal/mol. Energy decomposition analysis indicates high energy of binding for the P4, P3, P2, P1, P2’ and P4’ positions (Fig. S1B) and the MM/GBSA energy components indicate again that binding is not determined by electrostatic interactions (Fig. S1D). Due to a slightly less favorable binding enthalpy (\( \Delta H_{\text{GBSA}} = -94.3 \) vs \(-97.7 \) kcal/mol) and entropy term (\( \Delta T\Delta S = 51.8 \) vs \( 47.3 \) kcal/mol), this peptide has a lower binding energy than the peptide suggested containing the canonical Gln at position P1. Coincidentally, the results in Table S4 were partially corroborated by experimental peptide library cleavage data [51] and the native viral sequences cleaved by Mpro [10].

In that e.g. our MM/GBSA studies indicate Leu in position P2 as one of the 3 top binders and Arg is the top binding amino acid in P3.

3.6. Final conclusions

Obtaining kinetical data from binding studies of substrates to proteases must be viewed with caution, as the binding of substrate is not usually the rate limiting step of the reaction. This can be exemplified in an extreme case by the well-studied peptidic protease inhibitors [52]. Another limitation of this study is related to the possibility that some of the detected peptides might have been generated by lung tissue proteases after incubation with Mpro. Notwithstanding, the modelling approach of peptides obtained by TAILS analysis described here can be of significant interest for both the design of novel diagnostic peptides as well as inhibitors of this prominent SARS-CoV-2 drug target. We exemplify this briefly by suggesting two model peptides based on MM/GBSA energy decomposition analysis: The high affinity GRNQIFVFK peptide and the non-canonical PLIRKQLQYP peptide, which binds somewhat weaker to Mpro. This approach can therefore be of use for the development of both potent inhibitors and selective substrates of Mpro. The modelling results obtained additionally indicate that the Mpro has a far higher potential promiscuity towards substrates than expected. Due to the methodological limitations stated above, this effect needs to be corroborated in further studies.

Funding

This project obtained funding from FAPESP grants 2011/50963–4 (MW), 2019/07282–8 (AZ) and 2013/07467–1 (CeTICS) as well as FINEP grants 04.11.0043.06 and 04.16.0054.02 (institutional).

CRediT authorship contribution statement

Conception of the work, AZ, MW; experimental procedures, GG, UB, CC, MS, MC, DC, FL; analysis of data, SS, LH, AZ, MW; writing of manuscript, GG, SS, LH, AZ, MW.

Conflict of interest

The authors declare no conflict of interest.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.peptides.2022.170814.

References

[1] A. Hijikata, C. Shionyu, S. Nakae, M. Shionyu, M. Ota, S. Kanaya, T. Shirai, Current status of structure-based drug repurposing against COVID-19 by targeting SARS-CoV-2 proteins, BIOPHYSICS 18 (2021) 226–240.
[2] A. Saxena, Drug targets for COVID-19 therapeutics: ongoing global efforts, J. Biosci. 45 (2020) 87.
[3] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Y. Bi, X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, et al., Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding, Lancet 395 (2020) 565–574.
[4] R. Hilgenfeld, From SARS to MERS: crystallographic studies on coronavirus proteases enable antiviral drug design, FEBS J. 281 (2014) 4085–4096.
[5] Z. Jin, X. Du, Y. Xu, Y. Deng, M. Liu, Y. Zhao, B. Zhang, X. Li, L. Zhang, C. Peng, Y. Duans, J. Yu, L. Wang, K. Yang, F. Liu, R. Jiang, X. Yang, T. You, X. Liu, X. Yang, et al., Structure of M pro from SARS-CoV-2 and discovery of its inhibitors, Nature (2020) 1–5.
[6] L. Zhang, D. Lin, X. Sun, U. Curth, C. Drosten, L. Sauerhering, S. Becker, K. Rox, R. Hilgenfeld, Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors, Science 368 (2020) 409–412.
