THE EUROPEAN PHYSICAL JOURNAL D

Regular Article – Molecular Physics and Chemical Physics

The same, but different, but still the same: structural and dynamical differences of neutrophil elastase and cathepsin G

Fabian Schuhmann1, Xiangyin Tan2, Luca Gerhards1, Heloisa N. Bordallo2, and Ilia A. Solov’yov1,3,4,a

1 Department of Physics, Carl von Ossietzky Universität Oldenburg, Carl-von-Ossietzky-Str. 9-11, 26129 Oldenburg, Germany
2 Niels Bohr Institute, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark
3 Research Centre for Neurosensory Science, Carl von Ossietzky Universität Oldenburg, Carl-von-Ossietzky-Str. 9-11, 26129 Oldenburg, Germany
4 Center for Nanoscale Dynamics (CENAD), Carl von Ossietzky Universität Oldenburg Institut für Physik, Ammerländer Heerstr. 114-118, 26129 Oldenburg, Germany

Received 25 April 2022 / Accepted 2 July 2022 / Published online 20 July 2022
© The Author(s) 2022

Abstract. Although the general mechanism for serine protease catalysis is well established, some questions still remain. For instance, the two enzymes, neutrophil elastase and cathepsin G, have a lot of structural resemblances. However, elastase degrades virulence factors, while cathepsin G does not. This paper studies both enzymes computationally to probe for their conformational differences. In the process, a methodology is established to not only quantify similarities between the protein trajectories describing proteins’ temporal evolution but also account for a varying number of amino acid residues comprising each structure. Our results indicate slight differences in the behavior of the active sites of neutrophil elastase and cathepsin G in the solvent. These subtle changes could indicate differences in the general behavior responsible for the different specificity of the two enzymes.

Introduction

*Shigella* are Gram-negative rod-shaped bacteria, which are typically 0.3–1 µm in diameter and 1–6 µm in length, appearing singly, in pairs, and in chains, being closely related to *E. coli*, that invade the colonic mucosa [1]. *Shigella* are present wherever humans exist, and the direct spread from one person to another is a major route of transmission, which is facilitated by the very low inoculum of organisms required for infection [2]. An individual with dysentery and a presumed *Shigella* infection should be treated with an appropriate antimicrobial agent when first seen [2]. However, the increasing prevalence of multiple drug-resistant strains has made the selection of an appropriate agent increasingly difficult [3], making *Shigella* the second leading cause of diarrhoeal mortality in 2016 [4].

The inflammatory response initiated by *Shigella* is characterized by the recruitment of neutrophils, a part of a family of granulocytes that play an essential role in innate immunity. Neutrophil granules house critical enzymes that initially damage the colonic mucosa and exacerbate the infection but, ultimately, also control the infection by trapping *Shigella* in a phagosome where it is killed efficiently. Perhaps the best studied among these are the azurophilic granules [7], which contain high levels of bactericidal enzymes including the chymotrypsin-like serine proteases neutrophil elastase (NE) and cathepsin G (CG) [8–11], which use the Ser195/His57/Asp102 catalytic triad mechanism [12], where serine is the nucleophile, histidine is the general base and acid, and the aspartate helps orient the histidine residue and neutralize the charge that develops on the histidine during the transition states. This arrangement of amino acids in the active site presumably allows a nucleophilic attack by Ser195 on the carbonyl carbon (C = O) of the substrate at which the hydrolysis cleavage occurs, thus setting off the catalysis process [12].

The following convention for the sections of the two proteins has been adopted from an earlier study [10].

---

123
The specific S1 pocket in individual chymotrypsin-like serine proteases is crucial in defining their substrate specificity. In NE, the S1 pocket is hemispherical and hydrophobic because of the presence of valine, phenylalanine, alanine, and the disulfide bridge Cys182-Cys201, see Fig. 1. In contrast, the S1 pocket in CG reveals a significantly stronger electrostatic character due to a negatively charged glutamic acid (S1’ sub-site) at the far end. The S2’ sub-site, which is located in the S2 pocket of the protein (Fig. 1C, cyan dotted lines) of NE and CG, is formed by leucine (Leu106, Leu105) residues (Fig. 1A–B) and the flat side of the imidazole ring of His57 from the catalytic triad, which is rather hydrophobic. X-ray structural analysis showed that CG displays two positively charged residues (Arg41 and Arg143) in close vicinity to the active S2’ sub-site that could contribute to the specificity of the protease [13]. Visualization of the sequence alignment of NE and CG and a graphical illustration of the spatial arrangement of the aforementioned sites is shown in Fig. 1.

The structure-function analysis has established several regions in NE as essential for cleaving Shigella virulence proteins [13,14]. For example, it was hypothesized that Shigella virulence factor specificity belongs to a distinct region close to NE’s active site. In order to thoroughly determine the reasons for the different functional specificity in NE and CG and to find the subtle indicators of differences between these two proteins, atomistic simulations can be employed. For this purpose, we suggest and utilize a novel protocol to probe the differences in dynamics of the two protein structures, explicitly dealing with structures with a different number of amino acids residues.

The motivation for the study primarily comes from the fact that although NE and CG sequences are only about 32% similar sequence-wise [5], their X-ray crystal structures are remarkably similar [10,11]. The sequence alignment is shown in Fig. 1 as well as in Appendix 9. Despite that, solely NE is a key host defense protein responsible for controlling Shigella, as well as Salmonella and Yersinia, the causative agents of typhoid fever and the plague [15]. Why does CG lack the NE specificity for bacterial virulence factors? Our results suggest that understanding the role of neutrophils in disease mechanisms might be facilitated by computational methodologies.
Methodology

Molecular dynamics simulation and stability analysis

The known crystal structure of boar elastase (NE) (pdbID: 1ela) [11] and human cathepsin G (CG) (pdbID: 1cgh) [10] were simulated employing all-atom molecular dynamics (MD) approach. Molecular dynamics (MD) simulations were conducted using the NAMD [16,17] software; the interatomic interactions were described using the CHARMM36 force field with CMAP corrections [18–22]. The simulations were set up employing the online-platform VIKING [23]. The same simulation parameters were used for the NE and CG structures. The structures were solvated in a water box with 20 Å padding distance around the protein structure. The system was neutralized and ionized using NaCl to reach a concentration of 0.2 mol/l, representing the typical salt concentration inside a cell. The temperature of the system was set to 315 K, and a pressure of 1 bar was chosen. Following the 10,000 energy minimization steps, equilibration simulations were performed. Equilibrations were carried out for 21 ns at a timestep of 1 fs with varying constraints: In the first nanosecond, everything but water and ions was restrained, while the next 10 ns had a restrained protein backbone, allowing the amino acid side chains to reach a favorable conformation. The last 10 ns of the equilibration protocol was done without any restraints. Following equilibration, the production simulations were carried out for 400 ns with rigid hydrogen bonds in the NVT (constant number of particles, volume, and temperature) ensemble.

The simulated structures were probed for their stability using the radius of gyration, \( R_g \), which was determined for every timestep for the two structures, respectively, by calculating

\[
R_g = \sqrt{\frac{1}{K} \sum_{k=1}^{K} ||\mathbf{r}_i - \mathbf{r}_c||^2},
\]

with \( K \) denoting the number of atoms in the protein, \( \mathbf{r}_i \) being the location of the \( i \)th atom and \( \mathbf{r}_c \) the geometric center of the protein.

Small angle neutron scattering

To substantiate the results of MD simulations, small-angle neutron scattering (SANS) was performed to compare the shape of NE and CG dissolved in \( D_2O \).

The SANS allows testing of NE and CG conformations without inhibitors, while crystal structures were obtained only with inhibitors [10,11]. Additionally, the molecular weight of 25.98 kDa and 26.58 kDa for CG and NE, respectively, limits the possibility of using NMR, making SANS a favorable method in this study. Furthermore, although very high protein concentrations were necessary to obtain a reasonable signal-to-noise ratio, enzymatic assay results showed that both proteins were still active even after 72 hours of exposure to the neutron beam.

Experiments were carried out using V4 at the Helmholtz-Zentrum Berlin für Materialien und Energie (Berlin, Germany) in standard geometry at 25 °C [24]. The samples were placed in quartz cells of 1 mm thickness and filled with the buffer as immersion liquid. In all cases, circular cadmium diaphragms of 14 mm diameter defined the volume exposed to the incident beam. The samples were measured with a wavelength of 6 Å at three different sample-detector distances of 0.98 m, 4 m, and 12 m, covering a momentum transfer range between 0.04 nm\(^{-1}\) and 3.5 nm\(^{-1}\).

NE and CG were purchased from Elastin Products Company, Owensville, Missouri, USA. Their enzymatic activity was determined per the manufacturer’s protocols. For this experiment, NE and CG were both dissolved in 100 mM Tris D\(_2\)O buffer at a concentration of 10 mg/mL.

Theory of dynamical trajectory comparison

Alignment of two trajectories describing protein dynamics, followed by their comparison, utilized the so-called Kullback–Leibler divergence [25]. The essential idea of the methodology employed is described below.

General aspects

Qualitative comparison of two trajectories describing the dynamics of two proteins requires the structural alignment of both structures that evolve in time. The alignment is meant to diminish the effect of global movements and rotations of the evolving structures in order to focus on internal rearrangements within the protein structures. A common strategy to align two protein structures is the Kabsch algorithm [26], which rotates and translates the structures to achieve the smallest relative root mean square deviation (RMSD) value for the two structures. The Kabsch algorithm requires the same number of residues to be present in the two studied structures, which is not applicable for the NE and CG structures; the NE structure is comprised of 240 residues, while the CG structure contains 224 residues. A multi-step alignment scheme was therefore employed.

At first, a single frame alignment using the MultiSeq tool in VMD [27,28] was performed, and 134 pairs of residues from both structures were manually selected following a visual inspection. Next, the Kabsch alignment was used to line up the protein structures according to the selected 134 reference residues. The procedure followed earlier studies [29,30], in which the NE and CG structures were aligned iteratively to their own average structure for all MD timesteps until the changes became insignificant. The aligned trajectory of CG was then aligned to the aligned trajectory of NE for all time steps.
The Kabsch-aligned structures were then used in a less parameter-dependent approach, which permits aligning two protein structures independently of a manual choice of an equal number of residues, utilizing a two-step algorithm based on the Fréchet distance as proposed by Jiang et al. [31]. The Fréchet distance analyzes the position of the backbone atoms in a protein structure and interprets them as an ordered path in the three-dimensional space. While traversing the two paths of the protein backbone, one should then search for the shortest possible connection between the two paths. From these shortest connections, the Fréchet distance is defined as the length of the longest connection. The Fréchet distance has the advantage that the two discrete ordered sequences of data points do not have to be of the same length for a successful Fréchet distance calculation, thus permitting a comparison of the 224 residues of CG directly to the 240 residues of NE. The concept of the Fréchet distance calculation is illustrated in Fig. 2. The orange cylinders show the shortest connections that atoms from one structure can have to the atoms in the other structure. The purple connection marks the longest possible connection from all the possible pairs of atoms. The length of the purple connection is then the result of the Fréchet distance calculation.

The two-step algorithm described by Jiang et al. [31] was used to align all backbone atoms in the protein structures. First, both protein structures have been transposed, such that their geometric center of mass coincides with the origin point at (0, 0, 0). The following two steps were repeated for each MD simulation snapshot. **In the first step**, the first atom (the N atom of the first residue) and the last atom (the C atom of the last residue) of the three-dimensional protein backbone, together with the origin point, form a triangle. Considering such triangles for the NE and CG structures, one performs rotations in a way that the triangles become coplanar and their heights align, as illustrated in Fig. 3. The coplanar triangles guarantee a sensible initial orientation of the two protein backbones relative to each other for comparing the Fréchet distances. The first and the last residues of the compared structures are always considered when traversing the backbone to compute the distance. Therefore, both residues have a great influence on the resulting distance which is initially minimized by the triangle construction. **In the second step** of the alignment procedure, one structure (CG, reference) is kept rigid while the other (NE, alignee) is rotated. Two atoms from the alignee structure are chosen at random. The rotation of the alignee structure is then attempted by a small angle around the axis defined by the two atoms. The rotation is performed if the overall Fréchet distance between the reference and the alignee structure decreases and is rejected otherwise. The rotation procedure is repeated for a fixed number of iterations.

Note that the alignment algorithm is not deterministic, and it is possible that a successive rotation would decrease the Fréchet distance between the two structures without reaching the best relative alignment if the random rotation axis was not chosen optimally. The algorithm always follows the decrease in the Fréchet distance at each iteration, and therefore, it might follow a potentially wrong rotation direction. In order to decrease the chance of a non-optimal fit, the Fréchet alignment was repeated ten times from the initial coplanar triangle configuration. Finally, the alignment with the lowest Fréchet distance was chosen for each simulation snapshot.

In this study, the rotation angle was set to 0.0006 degrees, and 500 iterations were performed to find a better alignment of the two studied structures. The rather rotation small angle was assumed to correct the alignment established with the Kabsch algorithm in the previous alignment step. In each rotation, the average maximum displacement of a single residue was just 0.00014 Å. Alternative strategies that involved multiple rotation angles as well as a scheme, that would alter the rotation angle depending on the improvements achieved in prior steps were also considered. For the size of neutrophil elastase and cathepsin G, such methods, however, did not show any improvement in the merit of the alignment procedure.

**Residues coupling: generating relative displacements**

Every amino acid residue can be described through its three backbone atoms for the alignment procedure. However, considering all backbone atoms is computationally demanding and largely unnecessary, as the description of every amino acid residue can be further
reduced to only the C$_\alpha$ atoms. Their central location in each residue makes the C$_\alpha$ atoms a suitable anchor point for the description of residue motion.

In contrast to the structural alignment, where all atoms of interest for a single simulation snapshot were considered, now every C$_\alpha$ atom for all simulation snapshots is taken into account. To perform comparative measurements, one needs to establish pairs of residues between both structures for which the differences in the position are to be analyzed. These pairs are necessary in order to compare each fragment of the proteins with sufficient precision. Since the two studied structures do not have the same number of residues, the problem can be addressed through the dynamic time warping algorithm [32], which creates a $224 \times 240$ penalty matrix from the locations of the C$_\alpha$ atoms in CG and NE protein structures. The penalties are assigned by the dynamic time warping algorithm explained in greater detail in Appendix A. In the matrix, every row corresponds to an amino acid residue in CG, while every column corresponds to a residue in NE.

Each element in the penalty matrix corresponds to a pair of residues, assigning one residue from each structure. The procedure of finding the suitable residue pairs is therefore equivalent to establishing a route through the matrix, which passes every row and every column at least once. For example, for a square penalty matrix, i.e., describing two protein structures with the same number of residues, the route could follow the matrix’ diagonal. Since the studied proteins (NE and CG) have a different number of residues, a more elaborate route through the penalty matrix has to be determined. A fragment of the route for the first nine residues (residue numbers 16 to 24) of the CG and NE structures is shown in Fig. 4D. Assume that the residue 23 in the CG and NE structures should be used for comparison. The repetition of the process where the couplings are established from the penalty matrix calculation. A single residue can be matched with multiple residues in the other structure. For instance, residue Ile17 in CG is paired with both Val17 and Gly18 in the NE. The purple circle illustrates the pair in Fig. 4 with both pairs linked in yellow. The pairs of atoms that belong to the different proteins can be used to quantitatively compare the dynamic evolution of the two structures.

A description of the dynamic time warping algorithm with further details on how the penalty matrix is populated and how the backtracking algorithm is performed is given in Appendix A.

The established pairs of atoms are influenced by the dynamics of the proteins during the course of the simulation, and the pairs established for one timestep might not be the best for another. Therefore, the coupled pairs between CG and NE were formed on the basis of every timestep of the performed MD simulations. The similarity between both protein structures was then calculated for each coupling, and the cumulative similarity result is achieved as the arithmetic mean computed for all similarities for all determined residue couplings. The repetition of the process where the couplings are obtained from different snapshots is introduced to cir-
Fig. 4 A An exemplary extract from the $224 \times 240$ penalty matrix calculated by the dynamic time warping algorithm. The penalty values (in nm$^2$) are shown for each entry in the matrix, as well as the corresponding residue pairs (which are described by their $C_{\alpha}$ atoms) for the NE and CG structures. The pairs of atoms to be coupled for the similarity measurement can be found by backtracking through the penalty matrix, which yields the indicated route (red). The route through the matrix can either go up, left, or diagonally up left, always following the direction with the smallest element as highlighted by the cyan square. Within the square, one starts with the pair $(23, 23)$ and continues to the smallest penalty value within the square, which is 0.73, corresponding to the residue pair $(22, 22)$. Note, however, that a single residue can be coupled with more than one residue in the other structure, e.g., residue number 17 in CG is coupled with residue numbers 17 and 18 in NE; see the purple circle. B The backbone of NE (blue) and CG (red) is shown for the first residues of the protein structures. The yellow connections show the pairs between the $C_{\alpha}$ atoms of CG and NE. The atoms between the linked $C_{\alpha}$ are the nitrogens and carbons that form the backbone of the protein structures. An example of a double paired connection is indicated by the purple circle.

cumvent the need of choosing a specific reference frame for the calculation of the penalty matrix couplings.

The Kullback–Leibler divergence

After the proteins have been structurally aligned and the to-be-compared pairs have been established, the comparison of the two structures can commence. In general, the Kullback–Leibler divergence (KLD) \cite{25} permits quantifying a difference between two arbitrary normalized distributions. Let us illustrate how the KLD can be used for the studied protein structures. For Ile17CG, the distribution of its $x$ coordinates would then be denoted as $D(Ile17CG, x)$. Consider now two residues, e.g., Ile17CG in CG, and Val17NE in NE. Based on the distributions introduced above, one defines $KLD(Ile17CG, Val17NE)$ as

$$KLD(Ile17CG, Val17NE) = \sum_{k \in \{x, y, z\}} \sum_{i=0}^{N-1} D(Ile17CG, k) \cdot \ln \left( \frac{D(Ile17CG, k)}{D(Val17NE, k)} \right),$$

(3)

where $D(\cdot, k)_i$ denotes the $i$th discrete entry in the distribution of the $k$th coordinate \cite{33}. In general, the KLD is not symmetric, i.e.,

$$KLD(Ile17CG, Val17NE) \neq KLD(Val17NE, Ile17CG).$$

(4)

The measure can, however, be made symmetric by considering the arithmetic mean of KLD computed in two
directions. Thus, the similarity $S$ between a Ile$^{17}_{\text{CG}}$ and Val$^{17}_{\text{NE}}$ residues is defined as

$$S(Ile^{17}_{\text{CG}}, Val^{17}_{\text{NE}}) = \frac{KLD(Ile^{17}_{\text{CG}}, Val^{17}_{\text{NE}}) + KLD(Val^{17}_{\text{NE}}, Ile^{17}_{\text{CG}})}{2}.$$  \hspace{1cm} (5)

As a final step, similarity for every residue has to be evaluated according to the number of its partners. For instance, one notices that Ile$^{17}_{\text{CG}}$ is actually paired with both Val$^{17}_{\text{NE}}$ and Gly$^{18}_{\text{NE}}$. On the other hand, Val$^{17}_{\text{NE}}$ is only forming the single pair to Ile$^{17}_{\text{CG}}$ as illustrated in Fig. 4. The final similarity value for Ile$^{17}_{\text{CG}}$ is therefore calculated as

$$\hat{S}(Ile^{17}_{\text{CG}}) = \frac{S(Ile^{17}_{\text{CG}}, Val^{17}_{\text{NE}}) + S(Ile^{17}_{\text{CG}}, Gly^{18}_{\text{NE}})}{2},$$

while the similarity for Val$^{17}_{\text{NE}}$ would be defined as

$$\hat{S}(Val^{17}_{\text{NE}}) = S(Ile^{17}_{\text{CG}}, Val^{17}_{\text{NE}}).$$ \hspace{1cm} (6)

In general, if one assumes a residue $a$ in NE to be paired with residues $b_1, \ldots, b_l$ in CG, then the similarity for the residue $a$ is defined as

$$\hat{S}(a) = \frac{1}{l} \sum_{i=1}^{l} S(a, b_i).$$ \hspace{1cm} (8)

The similarity of a residue in CG is calculated analogously by considering all the pairs to residues in NE.

Intuitively, one would assume that NE is as different to CG as CG is different to NE and therefore, the overall similarity between the two structures should not depend on whether CG is compared to NE or NE is compared to CG. The symmetry is, however, violated since it is possible to have residues in one protein that couple to several residues in the other protein (see Eq. (6) and (7)). Therefore, similarities in both directions have to be considered for analyses, i.e., CG→NE and NE→CG.

By repeating the KLD procedure for all residue pairs, their similarity can be calculated and analyzed. In summary, the KLD quantifies the difference between two residues in the two studied proteins over the time span of the whole simulation trajectory and yield a single number. The smaller the resulting number is, the more similar the two residues are expected to be.

---

**Results and discussion**

**Radius of gyration and comparison with small-angle neutron scattering data**

The radius of gyration ($R_g$) values were calculated for the two simulated NE and CG structures, and their respective time evolution is visualized in Fig. 5. The average radius of gyration is $17.1 \pm 0.1$ Å for CG and $16.9 \pm 0.1$ Å for NE. The values fluctuate narrowly around the average, which indicates stable structures. A difference of 0.2 Å is considered neglectable in terms of $R_g$ [34], once more demonstrating the overall similarity in structure between the NE and CG proteins.

Additionally, the stability of the protein structures was assessed based on analysis of the temporal evolution of the RMSD for the proteins’ backbone as well as the potential energy of each protein structure without the solvent contribution. The small fluctuation in the RMSD as well as the normal distributed energies indicate equilibrated and stable structures. The RMSD and energy data are shown in Appendix Fig. 10.

The small-angle neutron scattering (SANS) data were used to compare the shapes of NE and CG. The variation of the absolute scattering intensity $I$ with the momentum transfer rate $q$ for the samples in a 10 mg/ml solution is shown in Fig. 6.

Differences between NE and CG are observed primarily for $q \leq 0.02 \text{Å}^{-1}$. This region of the scattering curve is sensitive to internal molecular effects, describing protein–protein interactions. Three factors were considered to be important for the proteins’ affinity to form oligomers. NE has a molecular charge of $4^- | e^- |$, while the total electrostatic charge of CG is $23^- | e^- |$; both are positively charged, and therefore, repulsive interaction is expected between several proteins.
proteins, however, also possess a strong dipole moment, which was computed as the average value over the entire MD simulation for both cases; CG exhibits a higher dipole moment value (305 Debye) than NE (238 Debye). Lastly, the solvent-accessible surface area (SASA) was calculated for all hydrophobic residues in both proteins and averaged over the course of the MD simulations. NE exhibits a greater average SASA value of 2990 Å², while CG has a value of 2128 Å². NE might therefore bind more favorably than CG through hydrophobic interactions. Additionally, the electrostatic potential was mapped onto the protein structures as visualized in Appendix Fig. 11. The figure shows that the backbones of the two proteins are folded remarkably similar, while the electrostatic potential around the proteins is clearly the electrostatic potential of CG and NE at the Ser195 residue decreases and Asp102 are located in a “trough,” i.e., residues at which the two structures are more similar, the similarity of CG and NE at the Ser195 residue decreases (as seen through a corresponding peak in Fig. 7). In the active state, Ser195 is expected to undergo a slight conformational change to stabilize the catalytic process [39,40]. As NE and CG were simulated inactively, the slight difference between the position of Ser195 could explain the difference in the functionality of the proteins and might therefore facilitate or hinder the nucleophilic attack that initiates the catalysis process.

As indicated by the gray bars in Fig. 7, the large variance observed for the three loops decisive for individual specificity of the NE and CG enzymes highlights the clear relation between structure and function in the NE and CG proteins. However, no large variance (greater than the 0.9 quantile) was observed in the region of the three key residues of the active site, namely His57, Asp102, and Ser195, as indicated by the labeled vertical black lines. On the other hand, while both His57 and Asp102 are located in a trough, i.e., residues at which the two structures are more similar, the similarity of CG and NE at the Ser195 residue decreases (as seen through a corresponding peak in Fig. 7). In the active state, Ser195 is expected to undergo a slight conformational change to stabilize the catalytic process [39,40]. As NE and CG were simulated inactively, the slight difference between the position of Ser195 could explain the difference in the functionality of the proteins and might therefore facilitate or hinder the nucleophilic attack that initiates the catalysis process.

Lastly, Fig. 7 shows a dashed line for CG (red) and NE (blue), which marks the Kullback–Leibler divergence for CG and NE for the crystal structure only. If one were to see differences in the crystal structure, it could indicate a predisposition to a conformational change. However, the differences between the crystal...
structures are not only visually small but also based on the similarity measure. Only by scaling the difference by a factor of 100, they become detectable in Fig. 7.

**Conclusion**

Neutrophil serine proteases are essential for the function of the neutrophils, and although they are relatively well-characterized, several important unanswered questions remain. In particular, even though the role of NE and CG in antibacterial defense is relatively well-established, their extended specificities have not yet been fully determined.

One main question comes from the fact that although X-ray crystallography reveals that the structures of the inhibited proteins are very similar, they have different biophysical specificity; only NE specifically destroys *Shigella* virulence factors [8]. Given that the NE and CG molecules are flexible and soft, small changes within the molecular structure as well as slight differences in electrostatic potentials, dipole moments and charges may play a significant role for their functionality. The similarity in the crystal structures of NE and CG could stem from the introduction of inhibitors, which are necessary for a successful determination of the structure; the specific inhibitors for NE and CG mimic the peptide substrates and block the enzymatic activity of NE and CG. To test the hypothesis that the conformations of the two studied proteins, while not inhibited and in solution, are in reality different, an experiment was performed comparing NE and CG using small-angle neutron scattering (SANS). Due to the natural aggregation of the NE system, the information that can be extracted from the SANS data becomes difficult to be evaluated. To overcome such difficulties, computational methodologies were used to relate the structure and function of NE and CG. The basic idea is that subtle conformational changes might not be easily detected through (even advanced) experimental approaches. To achieve this goal, we have developed a matching method that uses a dynamic time warping algorithm enabling a pairwise evaluation of the similarity of the protein dynamics trajectory based on the symmetrized mean Kullback–Leibler divergence. Furthermore, it was demonstrated that the experimentally determined radius of gyration, $R_g$, obtained for CG is close to the value calculated from the MD simulations, implying that the value calculated for NE has also been established reliably and that the NE and CG structures are expected to be similar in solution. The MD simulation allowed the identification of subtle changes in the dynamics of the two proteins. These subtle differences between NE and CG in the solvent could be responsible for the difference in proteins’ specificity. Additionally, it has been established that the electrostatic nature of the active sites varies significantly between NE and CG, facilitating the difference in their catalytic behavior [14].
In conclusion, we argue that elaborate analyses based on MD simulations provide a detailed molecular understanding of the two studied protein systems and that not only electrostatic, but also conformational differences can be established, even if the crystal structures are very similar. Of course, the MD analyses have to be backed by experimental data, e.g., SANS data, and vice versa. The MD studies can suggest, for instance, possible mutations to be tested to avoid unnecessary costly experiments.

Acknowledgements We acknowledge Arturo Zychlinsky (Max Planck Institute for Infection Biology) and Molly Ingersoll (Institute Pasteur) for the fruitful discussions. The authors would like to thank the Danish Councils for Independent Research, the Volkswagen Foundation (Lichtenberg Professorship to IAS), the DFG, German Research Foundation, (GRK1885—Molecular Basis of Sensory Biology and SFB 1372—Magnetoreception and Navigation in Vertebrates), and the Ministry for science and culture of Lower Saxony (Simulations meet experiments on the nanoscale: Opening up the quantum world to artificial intelligence (SMART)). Computational resources for the simulations were provided by the CARL Cluster at the Carl-von-Ossietzky University Oldenburg, which is supported by the DFG and the ministry for science and culture of Lower Saxony. The work was also supported by the North-German Supercomputing Alliance (HLRN). We thank the Helmholtz-Zentrum Berlin für Materialien und Energie (former Hahn-Meitner Institut) for the allocation of neutron beamtime.

Author contributions

FS contributed to conceptualization, methodology, software, formal analysis, writing—original draft, visualization. XT contributed to investigation, writing—review & editing, visualization. LG contributed to validation, writing—review & editing. HNB contributed to conceptualization, investigation, resources, data curation, writing—original draft, supervision. IAS contributed to resources, data curation, writing—review & editing, supervision, funding acquisition.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability Statement This manuscript has no associated data or the data will not be deposited. [Authors’ comment: All data generated during this study are contained in this published article.]

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Graphical Software Protein renderings were done in VMD, and plots were generated employing the Matplotlib package in Python and Origin. Postproduction was done in Adobe Photoshop.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need
Appendix A. Dynamic time warping

The Dynamic Time Warping algorithm was originally designed to handle speech recognition problems, as the similarity of two phrases needed to be analyzed without obscuring the order [32]. Here, it was generalized and applied to study the dynamics of proteins.

Assume two ordered sequences of coordinate points (i.e., atomistic locations) with \( n \) and \( m \) elements, respectively. The dynamic time warping algorithm between two ordered sequences \( A = \{a_1, \ldots, a_n\} \) and \( B = \{b_1, \ldots, b_m\} \) creates an \( n \times m \) penalty matrix \( P \) in three steps; here, \( P(i,j) \) denotes the entry in the \( i \)th row and the \( j \)th column of \( P \).

First, one sets \( P(1,1) = \|a_1 - b_1\|^2 \) with \( \|\cdot\| \) being the Euclidean norm. As a second, the first row and first column are traversed, such that

\[
P(i,1) = P(i-1,1) + \|a_i - b_1\|^2
\]

and

\[
P(1,j) = P(1,j-1) + \|a_1 - b_j\|^2,
\]

respectively. Having computed the elements of the first row and the first column of the matrix \( P \), the remaining entries are determined by setting

\[
P(i,j) = \|a_i - b_j\|^2 + \min(P(i-1,j), P(i,j-1), P(i-1,j-1)).
\]

In other words, the dynamic time warping algorithm is designed to calculate a penalty value that depends on the distance between the points that build the corresponding pair, built from an element in \( A \) and \( B \). The penalty for the pair \((a,b)\) is calculated while considering the penalties of previous steps that needed to be traversed in the ordered lists to reach the pair \((a,b)\). The result for the dynamic time warping distance between the \( A \) and \( B \) sequences of coordinates corresponds to the \( P(n,m) \) element in the penalty matrix. A smaller value of \( P(n,m) \) indicates a higher similarity of the two ordered \( A \) and \( B \) sequences. The primary interest in the case of the studied proteins is not the dynamic time warping distance but the indices of the paired residues (couplings) generated in the process, which can be backtracked considering the whole penalty matrix \( P \).

A protein-specific introduction of the backtracking is visualized in Fig. 4. A route through the penalty matrix is defined as a list of tuples containing the row and column number of an element in the penalty matrix. The route must include the tuples \((n,m)\) as well as \((1,1)\). Additionally, each row number and each column number must appear at least once within the tuples comprising the route.

The route is constructed iteratively by consecutively adding the next steps. Starting at \( P(n,m) \), the tuple \((n,m)\) is added to the route. From there, one considers the elements in the penalty matrix \( P(n-1,m) \), \( P(n,m-1) \), and \( P(n-1,m-1) \), choosing the element with the smallest value and adding the corresponding tuple to the route. The procedure is repeated from the new element of the penalty matrix analogously. For a general \( P(i,j) \) with \( i \leq n, j \leq m \), one sequentially selects one of the \( P(i-1,j) \), \( P(i,j-1) \), or \( P(i-1,j-1) \) elements at every step of the algorithm until the \( P(1,1) \) element is reached and the tuple \((1,1)\) is added to the route. This backtracking procedure, also visualized as an example in Fig. 4, yields the route through the matrix.

Appendix B. Sequence alignment of NE and CG

See Fig. 9.

![Sequence alignment](image)

Fig. 9 Visualization of the sequence alignment of NE and CG. Sequence alignment mapped onto the individual protein structures of NE and CG. A white residue indicates an individual residue in both proteins (sequence similarity), while red shows a difference in residue in the particular location.

Appendix C. Stability of CG and NE

See Fig. 10 here.
Fig. 10 Root mean square deviation (RMSD) and energy distribution for CG and NE. Panel A shows the time evolution of the RMSD computed for the backbone of CG (red) and NE (blue) over the simulation duration. The low fluctuation indicates stable structures. Panel B shows the probability distribution of the potential energy of the proteins. The resulting normal distribution (black line) illustrates further that the proteins appear in a stable configuration.

Appendix D. Electrostatic surfaces of the protein structures

See Fig. 11 here.

Fig. 11 Electrostatic differences in CG and NE. Panels A and B show the CG and NE protein with their electrostatic potential mapped onto the surface of the two protein structures. The color scale ranges from the smallest potential (red) to the largest potential (blue). The active triad is shown as a reference to the orientation of the protein structures with His57 in cyan, Asp102 in black, and Ser195 in yellow. Panels C and D show the cartoon representation of the protein structures (CG in red, NE in blue) in the same orientation as the surface mappings to visualize the alignment of the structures in all four panels. Noticeably, close to the active triad, a difference can be detected. NE features a higher negative potential (Panel B) exhibits a higher negative charge. Overall, the intensity of charges is greater in NE compared to CG.

References

1. H. Bridle (ed.), Waterborne Pathogens: Detection Methods and Applications (Elsevier, Netherlands, 2014)
2. C.J. Kuehl, J.D. D’Gama, A.R. Warr, M.K. Waldor, V. Sperandio, An oral inoculation infant rabbit model for Shigella infection. mBio 11(1), 03105–03119 (2020). https://doi.org/10.1128/mBio.03105-19
3. R.B. Sack, M. Rahman, M. Yunus, E.H. Khan, Antimicrobial resistance in organisms causing diarrhoeal disease. Clinical Infect. Dis. 24(Supplement 1), 102–105 (1997)
4. I.A. Khalil, C.E. Troeger, B.F. Blacker, P.C. Rao, A. Brown, D.E. Atherly, T.G. Brewer, C.M. Engmann, E.R. Houpt, G. Kang, K.L. Kotloff, M.M. Levine, S.P. Luby, C.A. MacLennan, W.K. Pan, P.B. Pavlinac, J.A. Platts-Mills, F. Qadri, M.S. Riddle, E.T. Ryan, D.A. Shoulitz, A.D. Steele, J.L. Watson, J.W. Sanders, A.H. Mokdad, C.J.L. Murray, S.I. Hay, R.C. Reiner, Morbidity and mortality due to shigella and enterotoxigenic escherichia coli diarrhoea: the global burden of disease study 1990–2016. The Lancet. Infect. Dis. 18, 1229–1240 (2018)
5. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic Local Alignment Search Tool 2. J. Mol. Biol. 215, 403–410 (1990)
6. D.M. Blow, J.J. Birktoft, Douglas Mandell, Bennett’s Principles and Practice of Infectious Diseases, vol. 1–2 (Elsevier, Philadelphia, 2014)
7. J. Bennett, R. Dolin, Megson, Bennett’s Principles and Practice of Infectious Diseases, vol. 1–2 (Elsevier, Philadelphia, 2014)
8. Y. Weinrauch, D. Drujan, S.D. Shapiro, J. Weiss, A. Zychlinsky, Neutrophil elastase targets virulence factors of enterobacteria. Nature 417, 91–94 (2002)
9. E. Reeves, H. Lu, H. Jacobs, C. Messina, S. Bolsover, G. Gabella, E. Potma, A. Warley, J. Roes, A. Segal, Killing activity of neutrophils is mediated through activation of proteases by K+ flux. Nature 416, 291–297 (2002)
10. P. Hof, I. Mayr, R. Huber, E. Korzus, J. Potempa, J. Travis, J.C. Powers, W. Bode, The 1.8 Å crystal struc-
ture of human cathepsin G in complex with Suc-Val-Pro-Phe(P)-(OPh)2: A Janus-faced protease with two opposite specificities. EMBO J. 15, 5481–5491 (1996)

11. C. Mattos, B. Rasmussen, X. Ding, G.A. Petsko, D. Ringe, Analogous inhibitors of elastase do not always bind analogously. Struct. Biol. 1, 55–58 (1994)

12. L. Hedstrom, Serine Protease Mechanism and Specificity. Chem. Rev. 102, 4501–4523 (2002)

13. B. Korkmaz, M. Horwitz, D. Jenne, F. Gauthier, Neutrophil elastase, proteinase 3, and cathepsin G as therapeutic targets in human diseases. Pharmacol Rev. 62, 726–759 (2010)

14. P. Avorhoff, M. Kolbe, A. Zychlinsky, Y. Weinrauch, Single residue determines the specificity of neutrophil elastase for shigella virulence factors. J. Mol. Biol. 377, 1053–1066 (2008)

15. J. Conlan, Critical roles of neutrophils in host defense against experimental systemic infections of mice by Listeria monocytogenes, Salmonella typhimurium, and Yersinia enterocolitica. Infect Immun. 65, 630–635 (1997)

16. J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kalé, K. Schulten, Scalable molecular dynamics with NAMD. J. Comput. Chem. 26, 1781–1802 (2005)

17. J.C. Phillips, D.J. Hardy, J.D.C. Maia, J.E. Stone, J.V. Ribeiro, R.C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, R. McGreevy, M.C.R. Melo, B.K. Radak, R.D. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L.V. Kalé, K. Schulten, C. Chipot, E. Tajkhorshid, Scalable molecular dynamics on CPU and GPU architectures with NAMD. J. Chem. Phys. 153, 44130 (2020)

18. N. Foloppe, A.D. MacKerell, All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. J. Comput. Chem. 21, 86–104 (2000)

19. R.B. Best, X. Zhu, J. Shim, P.E.M. Lopes, J. Mittal, M. Feig, A.D. MacKerell, Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ1 and χ2 dihedral angles. J. Chem. Theory Comput. 8, 3257–3273 (2012)

20. K. Hart, N. Foloppe, C.M. Baker, E.J. Denning, L. Nelson, A.D. MacKerell, Optimization of the CHARMM additive force field for DNA: improved treatment of the B1/B2I conformational equilibrium. J. Chem. Theory Comput. 8, 348–362 (2012)

21. A.D. MacKerell, M. Feig, C.L. Brooks, Improved treatment of the protein backbone in empirical force fields. J. Am. Chem. Soc. 126, 698–699 (2004)

22. A.D. MacKerell, D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiwicz-Kuczera, D. Yin, M. Karplus, All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B 102, 3586–3616 (1998)

23. V. Korol, P. Husen, E. Sjulstok, C. Nielsen, I. Friis, A. Frederiksen, A.B. Salo, I.A. Solov’yov, Introducing VIKING: a novel online platform for multiscale modeling. ACS Omega 5, 1254–1260 (2020)

24. A. Zychlinsky, M. Ingersoll, H.N. Bordallo, Hauß, T., Heinemann, A.: BENSC Experimental Reports page 173. ISSN 0936 - 0891, Berlin Neutron Scattering Center Hahn-Meitner-Institut Berlin, Germany (2003)

25. S. Kullback, R.A. Leibler, On information and sufficiency. Ann. Math. Stat. 22(1), 79–86 (1951)

26. W. Kabsch, A discussion of the solution for the best rotation to relate two sets of vectors. Acta Crystallogr. Sect. A 34, 827–828 (1978)

27. E. Roberts, J. Eargle, D. Wright, Z. Luthey-Schulten, MultiSeq: Unifying sequence and structure data for evolutionary analysis. BMC Bioinform., 1–11 (2006)

28. W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics. J. Mol. Graph. 14, 33–38 (1996)

29. F. Schuhmann, D.R. Kattning, I.A. Solov’yov, Exploring post-activation conformational changes in pigeon cryptochrome 4. J. Phys. Chem. B 125, 9652–9659 (2021)

30. D.R. Kattning, C. Nielsen, I.A. Solov’yov, Molecular dynamics simulations disclose early stages of the photo-activation of cryptochrome 4. New J. Phys. 20, 083018 (2018)

31. M. Jiang, Y. Xu, B. Zhu, Protein structure-structure alignment with discrete Fréchet distance. J Bioinform Comput Biol 5, 131–141 (2007)

32. H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. 26 (1978)

33. P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. C.J. Larson, I. Carey, Y. Polat, E.W. Feng, J. Moore, VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A.P. Bardelli, A. Rothenberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C.N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D.A Nicholson, D.R. Hagen, D.V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G.A. Price, G.L. Ingold, G.E. Allen, G.R. Lee, H. Audren, I. Probst, J.P. Dietrich, J. Silterra, J.T. Webster, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J.L. Schönberger, J.V. de Miranda Cardoso, J. Reimer, J. Harrington, J.L.C. Rodríguez, J. Nunez-Iglesias, J. Kuczyński, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N.J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P.A. Brodtkorb, P. Lee, R.T. McGibbon, R. Feldbauer, S. Lewis, S. Tyghier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T.J. Pingel, T.P. Robitaille, T. Spura, T.R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y.O. Halchenko, Vázquez-Baeza, Y.: SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272 (2020)

34. D.N. Ivankov, N.S. Bogatyreva, M.Y. Lobanov, O.V. Galzitskaya, Coupling between properties of the protein shape and the rate of protein folding. PLoS One 4, 6476 (2009)
35. SasView. https://www.sasview.org/. Accessed: 2022-03-31
36. D. Svergun, Basics of X-ray scattering by solutions. EMBL Hamburg (2023)
37. A. Andreeva, D. Howorth, C. Chothia, E. Kulesha, A.G. Murzin, SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Res. 42(D1), 310–314 (2013)
38. A. Andreeva, E. Kulesha, J. Gough, A.G. Murzin, The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res. 48(D1), 376–382 (2019)
39. C.N. Fuhrmann, M.D. Daugherty, D.A. Agard, Sub-angstrom crystallography reveals that short ionic hydrogen bonds, and not a His-Asp low-barrier hydrogen bond, stabilize the transition state in serine protease catalysis. J. Am. Chem. Soc. 128, 9086–9102 (2006)
40. Y. Zhou, X. Zhang, Serine protease acylation proceeds with a subtle re-orientation of the histidine ring at the tetrahedral intermediate. Chem Commun 47, 1577–1579 (2012)