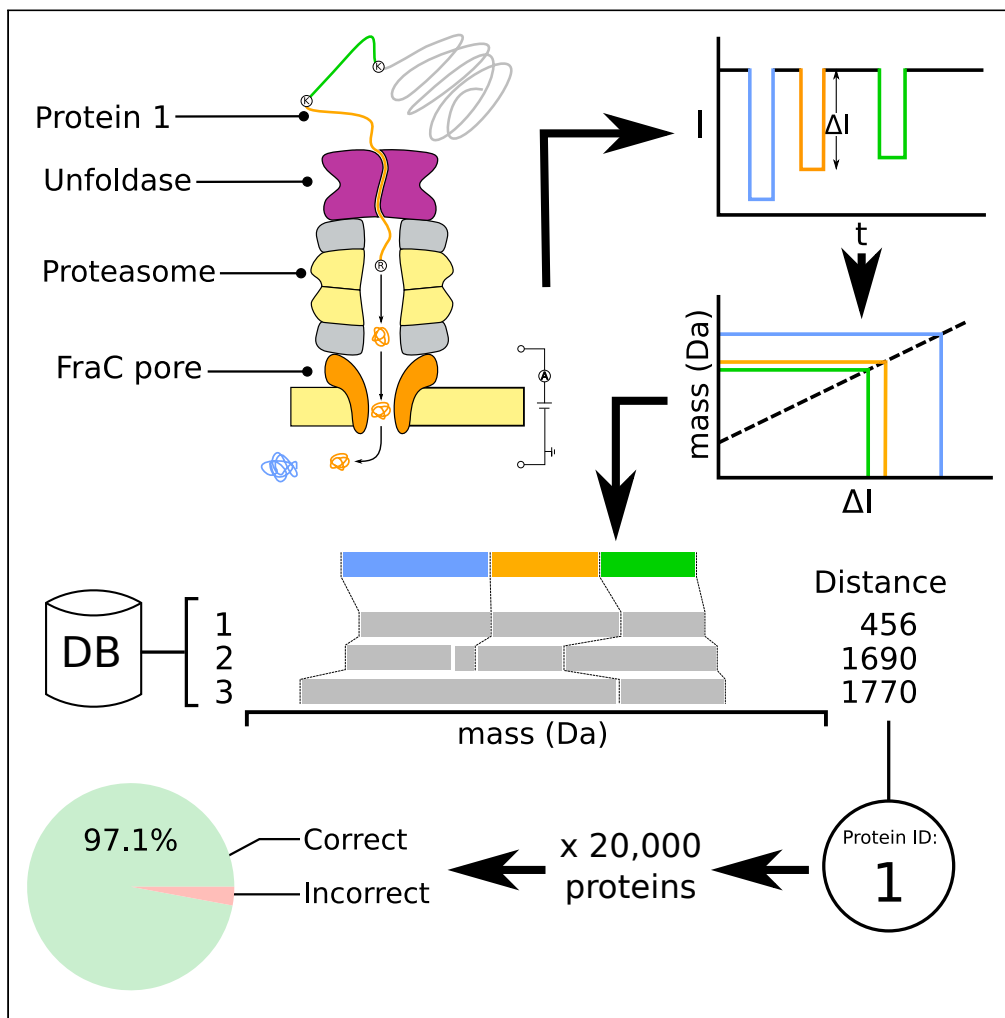


Article

*In silico* assessment of a novel single-molecule protein fingerprinting method employing fragmentation and nanopore detection



Carlos de Lannoy,  
Florian Leonardus  
Rudolfus Lucas,  
Giovanni Maglia,  
Dick de Ridder

carlos.delannoy@wur.nl

Highlights

A proteasome-nanopore construct can cleave a protein and interrogate the fragments

In simulations, this process provides sufficient information to identify proteins

An alignment-based classifier attains 97% accuracy on the entire human proteome

## Article

# *In silico* assessment of a novel single-molecule protein fingerprinting method employing fragmentation and nanopore detection

Carlos de Lannoy,<sup>1,3,\*</sup> Florian Leonardus Rudolfus Lucas,<sup>2</sup> Giovanni Maglia,<sup>2</sup> and Dick de Ridder<sup>1</sup>**SUMMARY**

The identification of proteins at the single-molecule level would open exciting new venues in biological research and disease diagnostics. Previously, we proposed a nanopore-based method for protein identification called chop-n-drop fingerprinting, in which the fragmentation pattern induced and measured by a proteasome-nanopore construct is used to identify single proteins. In the simulation study presented here, we show that 97.1% of human proteome constituents are uniquely identified under close to ideal measuring circumstances, using a simple alignment-based classification method. We show that our method is robust against experimental error, as 69.4% can still be identified if the resolution is twice as low as currently attainable, and 10% of proteasome restriction sites and protein fragments are randomly ignored. Based on these results and our experimental proof of concept, we argue that chop-n-drop fingerprinting has the potential to make cost-effective single-molecule protein identification feasible in the near future.

**INTRODUCTION**

Over the past decades, mass spectrometry (MS) has allowed for ground-breaking discoveries in proteomics, enabling such impressive feats as the definition of a human protein atlas (Pontén et al., 2008) and large-scale screening for protein disease biomarkers (Keshishian et al. 2017). However, not all protein-related research questions may be addressed by MS. Examples are found in the nascent field of single-cell proteomics which, following the example of single-cell transcriptomics, is expected to give unprecedented insight into cell functioning and pathology (Vistain and Tay, 2021). While MS has already made strides in this field by enabling the detection of proteins present at thousands of copies per cell (Specht et al., 2021), some important and clinically relevant proteins such as signaling molecules and transcription factors are expected to be present in the range of dozens of copies (Righetti and Boschetti, 2013). The development of novel single-molecule protein identification methods is therefore necessary to unlock the true potential of single-cell proteomics.

In the search for single-molecule alternatives to MS, two main venues are currently being explored. On the one hand, conceptual methods utilizing the read-out of fluorescent dyes attached to a subset of residue types have shown promising results (Ohayon et al., 2019; Swaminathan et al., 2018; Yao et al., 2015). However, methods using fluorescence-based readout strategies require efficient and specific labeling of residues. Optimizing labeling strategies is non-trivial (e.g. Abello et al., 2007; Nanda and Lorsch, 2014), and less-than-perfect labeling may decrease accuracy; thus, a label-free method would be preferred.

On the other hand, unlabeled proteins may be analyzed using a nanopore, over which an electrical potential is applied; as a protein is passing through the pore, changes in electrical resistance may give information on the protein's properties (Hu et al., 2021). Proteins may be analyzed in their folded states (Houghtaling et al., 2019; Huang et al., 2017; Ouldali et al., 2020; Piguet et al., 2018; Sutherland et al., 2004) which is relatively straightforward but does not provide sufficient information to discriminate between similarly shaped and charged proteins. Furthermore, no single pore aperture size is suitable for proteome-wide analysis due to the wide variety of protein sizes found in nature (Brocchieri and Karlin, 2005). Alternatively, proteins may be unfolded and threaded single file through the pore using a molecular

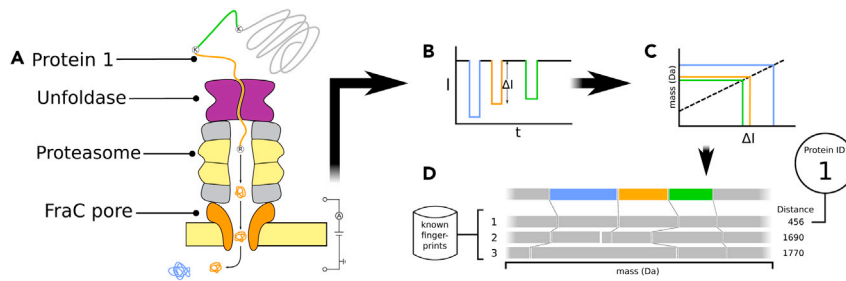
<sup>1</sup>Bioinformatics Group, Wageningen University, 6708PB Wageningen, The Netherlands

<sup>2</sup>Groningen Biomolecular Sciences & Biotechnology Institute, University of Groningen, 9747AG Groningen, The Netherlands

<sup>3</sup>Lead contact

\*Correspondence: carlos.delannoy@wur.nl  
<https://doi.org/10.1016/j.isci.2021.103202>





**Figure 1. Schematic overview of the chop-n-drop fingerprinting method**

(A) A protein is unfolded by an unfoldase and fragmented by a proteasome directly introduced above a nanopore. The protease is engineered to lyse proteins at particular residues.  
 (B) As the fragments pass the pore, a change in electrical current through the pore is measured.  
 (C) The molecular weights of the fragments are estimated from the magnitudes of the current changes.  
 (D) Finally, the produced sequence of fragment weights is aligned to database fingerprints of known proteins to identify the protein.

motor (Brinkerhoff et al., 2021; Nivala et al., 2014). This approach allows for finer interrogation of the residue sequence and may analyze proteins of any size using a single pore aperture size.

In prior work, we showed that engineered complexes of heptameric nanopores and proteasomes can be readily assembled without loss of proteasome activity or electrical conductance of the pore (Zhang et al., 2020). Furthermore, we have shown that residual current through FraC pores correlates well with the molecular weight of passing protein fragments in the 500–1600 Da range (Huang et al., 2019). Presumably, this is because weight is correlated to properties that directly influence residual current, such as fragment size, shape, and charge (Di Muccio et al., 2019; Houghtaling et al., 2019; Hu et al., 2018; Waduge et al., 2017; Wilson et al., 2019; Zernia et al., 2020). We thus proposed that proteasome-nanopore constructs can be used to identify proteins, in a conceptual method dubbed chop-n-drop fingerprinting (Zhang et al., 2020). An unknown protein can be processed terminal-to-terminal by the construct, cleaving it at proteasome target sites, after which the molecular weight of sequentially released fragments can be estimated based on the residual electrical current as they pass through the nanopore. The sequence of measured fragment weights can then serve as a characteristic signature—a fingerprint—of the protein. Once proven, this fingerprinting method can easily be implemented in a highly parallel fashion by adapting existing hardware that was developed for nucleic acid sequencing. Compared to both MS and existing fluorescence-based measurement equipment, this hardware is inexpensive and has a small benchtop footprint, thus opening up opportunities for field diagnosis and in-house analysis for even small laboratories. It is as of yet however unclear whether chop-n-drop fingerprints are sufficiently characteristic to identify a single protein in highly complex mixtures.

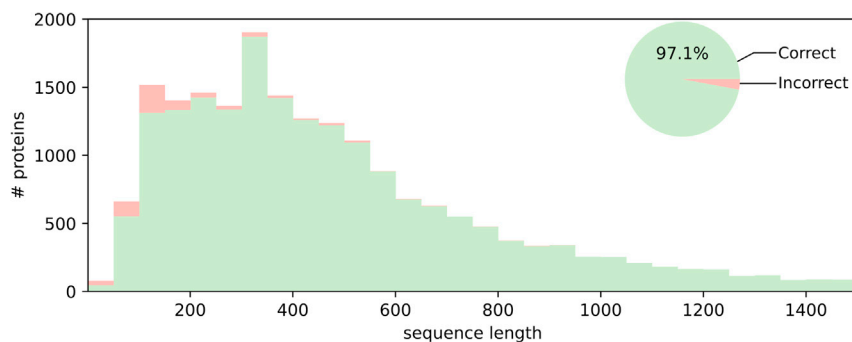
Here, we present a computational analysis of the chop-n-drop method, in which we show that simulated fingerprints of all proteins in the UniProt human proteome can be accurately classified using a simple alignment-based method. Considering these and previously published experimental results, we argue that chop-n-drop fingerprinting is a promising concept for cost-effective single-molecule protein identification.

## RESULTS

### Simulation and classification method

To estimate the performance of the chop-n-drop fingerprinting method on a highly complex protein identification task, we developed a simulation pipeline mimicking the experimental procedure, including several sources of biological and technical noise that we expect to encounter (Figure 1).

In essence, the chop-n-drop fingerprint of a protein only consists of a sequence of weights, which are deduced from pore current blockades caused by sequentially cleaved-off fragments passing through the nanopore. The simulation of this process follows a straightforward two-step process. First, akin to the proteasome cleaving a protein into fragments, we divide a given protein sequence into sub-sequences by splitting it at the proteasome's target sites. We assume here that we can force it to exhibit only



**Figure 2. Simulated fingerprint identification accuracy assuming low-noise conditions**

Cumulative histogram of correct and incorrect classifications of simulated chop-n-drop protein fingerprints for all human proteome constituents, assuming low noise parameters; resolution  $r = 5$  Da, capture rate  $C = 0.99$  and proteasome cleaving efficiency  $e_p = 0.99$ . Numbers shown are distributed over sequence length (bars) and relative to the total number of proteins (pie chart). Alignment examples and sequence identity distribution for erroneous alignments are shown in [Figures S3](#) and [S4](#), respectively. Results assuming charge-dependent fragment capture are shown in [Figure S5A](#).

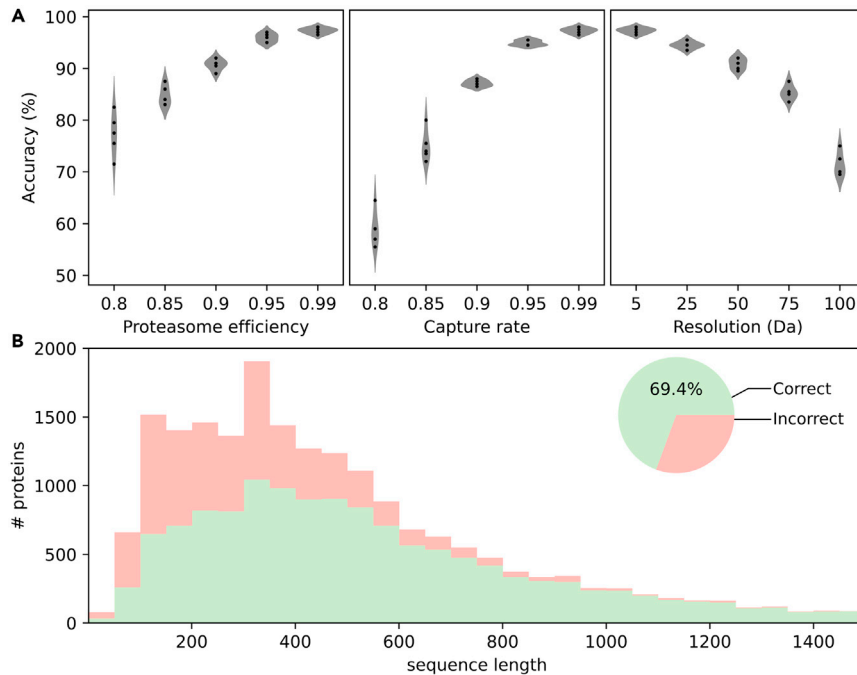
trypsin-like behavior by mutating proteasome subunits exhibiting chymotrypsin-like and caspase-like activities, while leaving its subunits exhibiting trypsin-like activity intact ([Collins et al., 2010](#); [Li et al., 2004](#)). To account for the fact that the proteasome will likely fail to cleave at a fraction of the target sites, we only cleave each target site with a certain probability, which we refer to as the proteasome efficiency ( $e_p$ ).

Subsequently, we mimic the passing of fragments through a heptameric FraC pore, the readout of the current blockade, and the estimation of the fragment weight by simply translating the sub-sequences into corresponding fragment weights. Although weights can be calculated from sequences with high accuracy, experimental measurements may be less accurate and marked by a given resolution ( $r$ ), the smallest detectable weight difference. In experimental setups, this parameter is dependent on pore and measuring equipment properties. The smallest weight difference we have detected with FraC so far is 4Da ([Figure S1](#)); thus, in simulations, we consider  $r$ -values above 4 Da attainable. To account for resolution in simulation, Gaussian noise is added to fragment weights, where the standard deviation of the noise is related to  $r$  (see [STAR Methods](#)). Fragments weighing less than 500 Da are removed, as they typically escape detection of heptameric FraC nanopores ([Huang et al., 2019](#)). Furthermore, as it has not been shown that the relation between weights above 1.6 kDa and current blockades remains monotonic ([Huang et al., 2019](#)), all fragment weights larger than this value are reduced to 1.6 kDa. Lastly, although we expect the seal between proteasome and pore to be extremely tight based on molecular dynamics simulations ([Zhang et al., 2020](#)), fragments may fail to enter the pore after cleavage. We account for this by only retaining each fragment with a certain probability, which we refer to as the capture rate ( $C$ ). Although  $C$  is likely dependent on the size and charge of individual fragments, the relationship between these factors is unclear; thus, we assume  $C$  to be constant. The resulting sequence of fragment weights returned by this process constitutes the fingerprint for a protein.

We used fingerprints generated using our pipeline to develop a classification method, which assigns a protein identity to a given fingerprint. We follow an alignment-based approach, where a query fingerprint is aligned to a database of previously generated fingerprints, using a custom dynamic programming implementation ([Figure S2](#)). The database fingerprint that is most similar to the query fingerprint is assumed to have come from the same protein.

### Simulations under low-noise conditions

We ran our simulation pipeline and classification method on all sequences in the UniProt human proteome ( $n = 20,395$ ). Under close to ideal simulated noise parameters ( $e_p = 0.99$ ,  $r = 5.0$ Da,  $C = 0.99$ ), we find that our alignment-based approach retrieves the correct identity for 97.1% of fingerprints ([Figure 2](#)). Inspection of made alignments shows that our algorithm correctly handles missing and fused fragments ([Figure S3A](#)). Then, 77% of misclassifications occur for shorter proteins, under 250 residues in length. Of misclassified fingerprints, 42% shows more than 80% amino acid sequence identity to the protein as which it was wrongly identified, indicating that the resolution of 5Da assumed here is insufficient to consistently separate such



**Figure 3. Simulated fingerprint identification accuracy under noisy conditions**

(A) Fingerprint classification accuracy over a range of noise parameter values; resolution (left), capture rate (mid), and proteasome efficiency (right). For each case the unvaried noise parameters are set to low-noise values (capture rate  $C = 0.99$ , resolution  $r = 5.0$  Da, and proteasome efficiency  $e_p = 0.99$ ). Five replicates were generated for each parameter combination.

(B) Cumulative histogram of correct and incorrect classifications of simulated chop-n-drop protein fingerprints for all human proteome constituents, assuming more realistic noise parameters;  $r = 10$  Da,  $C = 0.90$  and  $e_p = 0.90$ . Numbers shown are distributed over sequence length (bars) and relative to the total number of proteins (pie chart). Results assuming charge-dependent fragment capture are shown in [Figure S5B](#).

similar entities ([Figure S4](#)). Upon inspection of these cases, we find that many misclassifications were in fact mix-ups between paralogous sequences. The remaining misclassifications are caused by chance alignments with different fingerprints ([Figure S3B](#)). This is expected to occur more often if a protein is shorter, as it will generally produce a fingerprint of fewer elements, which is less likely to yield a unique pattern.

### Simulations under high-noise conditions

We subsequently probed how resistant chop-n-drop fingerprinting is to higher levels of experimental noise, by varying one noise parameter at a time while keeping all others near their low-noise values ( $e_p = 0.99$ ,  $r = 5.0$ Da,  $C = 0.99$ ). To keep computations tractable, these simulations were run on a random subset of 200 query sequences, while the reference database still contained all UniProt sequences so that the classification task was no less challenging. In each case, we find that accuracy deteriorates gracefully with parameter values ([Figure 3A](#)). Interestingly, we still attain an accuracy of 90.9% at a resolution of 50 Da, which is worse than the 44 Da resolution we reported previously ([Huang et al., 2019](#)) and more than tenfold worse than the current best resolution of 4Da, as reported in this work ([Figure S1](#)). Similarly, we find that a lower proteasome efficiency or capture rate of 90% still results in 90.7% and 87.1% accuracy on average, respectively. We then repeated the simulation on the entire dataset with all noise parameters at high-noise values ( $e_p = 0.90$ ,  $r = 10.0$ Da,  $C = 0.90$ ). Even under these circumstances, we find that 69.4% of proteins are correctly classified ([Figure 3B](#)). Here too, it should be noted that most incorrectly classified proteins were of lower sequence length.

Finally, we investigated the effect of fragment charge on accuracy. In FraC pores, electro-osmotic flow (EOF) is sufficiently strong to overcome an opposing electrophoretic force (EF) to some degree so that even negatively charged fragments are pulled through the pore toward an anode at the trans side.

However, fragments carrying larger negative charges increase the EF so that the EOF can no longer cancel it out, and thus, such fragments cannot enter the pore. We investigated the effect of the omission of fragments that are too negatively charged ( $\leq -1e$ ) at FraC's operating pH levels (pH = 4.0) (Huang et al., 2019) and found that the effect on accuracy is negligible in both low-noise and high-noise scenarios (97.1% and 69.3%, respectively, Figure S5).

## DISCUSSION

Single-molecule (SM) protein fingerprinting holds great promise to revolutionize biological research and diagnostics (Restrepo-Pérez et al., 2018). We have previously proposed that this may be accomplished using a proteasome-nanopore construct, which cleaves a target protein into fragments and subsequently reads out the fragment weights (Zhang et al., 2020). Here, we present simulation results indicating that the produced sequence of fragment weights contains sufficient information to identify a protein.

The hypothetical construct investigated in our simulations consists of a heptameric dihedral FraC pore, which we show to be well suited to detect differences between fragment weights at high resolution and a proteasome exhibiting trypsin-like cleaving activity. In prior work, we have shown that constructs of artificial heptameric beta-barrel PA-pores and proteasomes can be built without loss of function of either component. Given the structural similarity, we are confident that our hypothetical construct can be built in a similar way, by replacing PA-pore monomers with FraC monomers and by making use of engineered proteasomes (Collins et al., 2010; Li et al., 2004).

In the presented simulations, we included sources of noise that may hamper fingerprint measurements in practice. We assumed that the proteasome may not cleave each target site that weight measurements may be inaccurate up to a given weight resolution and that not all cleaved-off fragments may be caught in the nanopore. Assuming higher noise parameter settings—a fragment capture rate and proteome efficiency of 90%, with a measurement resolution of 10Da—for each of these noise sources, we find that overall accuracy remains sufficiently high at 69.4%. As accuracy increases with protein length, we find that chop-n-drop fingerprinting should be particularly suitable to identify larger proteins.

Over the past years, the obstacles on the road toward SM protein fingerprinting have been attacked vigorously from multiple angles, with several groups showing promising initial results and proofs of concept. While each proposed method has shown particular strengths, we argue that chop-n-drop combines several properties not found together in other methods. First, unlike fluorescence-based methods (Ohayon et al., 2019; Swaminathan et al., 2018; Yao et al., 2015), it does not require the implementation of any labeling chemistries as properties of the target protein are read out directly, thus evading issues with erroneous labeling and simplifying sample preparation. As a trade-off, fluorescence-based methods are more sensitive to differences between proteoforms as long as the difference involves the position or presence of a targeted residue type. As we show here that even at high resolution our method misclassifies proteins with high sequence similarity to other entries, it is likely that differences between highly similar proteoforms may also remain unnoticed.

Different methods based on the readout of folded proteins by electrical current blockage of a nanopore have been proposed as well (Huang et al., 2017; Piguet et al., 2018; Yusko et al., 2017). These were unable to analyze a wide range of protein sizes however; as the pore lumen needs to be of an appropriate volume for the analysis of a given protein size, a single nanopore is not able to detect minute differences in both small and large proteins. Here, this problem is mitigated by the fragmentation step.

Most importantly, however, the hardware required to implement chop-n-drop fingerprinting in a highly parallelized setting can be readily borrowed from commercial platforms for DNA sequencing using nanopores, which are inexpensive and have already been miniaturized to a handheld format. As such, we envision that our method could soon fill a niche that no other method currently can: that of small-scale, in-house single-molecule protein identification.

In conclusion, we provide evidence that chop-n-drop fingerprints can provide sufficient information to identify proteins in complex samples and present a suitable alignment-based classification method. Upon optimization of the fingerprinting procedure, we envision that our method may see practical implementation in the near future.

### Limitations of the study

Our simulation builds on the assumption that fragment weight is correlated to the residual current measured while the fragment passes the nanopore. Indeed, we have previously shown that this is the case for fragments weighing between 500 and 1600 Da (Huang et al., 2019). However, it should be noted that rather than the fragment's weight, its volume, shape, charge, hydrophobicity, and interactions with the pore interior directly influence residual current (Di Muccio et al., 2019; Houghtaling et al., 2019; Hu et al., 2018; Waduge et al., 2017; Wilson et al., 2019; Zernia et al., 2020). As these properties are more difficult to model and considering how they apparently correlate to weight sufficiently well to in turn correlate weight and residual current (Huang et al., 2019), we consider using weight in simulated fingerprints justifiable. Once the experimental methodology has been further developed and protein fingerprints can be measured more routinely, we can define the relation between these properties and the residual current in more detail to predict fingerprints in a more robust manner.

The existence of different proteoforms, which was not accounted for in this simulation, presents both an opportunity and a challenge to chop-n-drop fingerprinting. Through alternative splicing and post-translational modification (PTM), multiple proteoforms with different functions may be generated from the same gene (Aebersold et al., 2018). Depending on the spliceoform or the PTM types present, different proteoforms may generate distinct fingerprints. This allows their individual identification at SM resolution, which is an important potential application of SM analysis, but also adds tens of thousands of potential fingerprint patterns, which further complicates the task of fingerprint classification. A solution may be to fractionate samples prior to chop-n-drop analysis, after which each fraction may be analyzed using a dedicated classifier which only considers the proteoforms that could be present in a given fraction.

### STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
  - Lead contact
  - Materials availability
  - Data and code availability
- **METHOD DETAILS**
  - Fragaceatoxin C purification
  - His6-tagged fragaceatoxin C sequence
  - Electrophysiological recordings
  - Electrophysiological recording analysis
  - *In silico* fingerprint generation
  - Simulation and classification

### SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.isci.2021.103202>.

### ACKNOWLEDGMENTS

This work was supported by the Vrije Programma of the Foundation for Fundamental Research on Matter under grant number 16SMPS05.

### AUTHOR CONTRIBUTIONS

C.L. developed and wrote the code for simulations, D.R. supervised simulation development, F.L.R.L. designed and executed practical experiments and performed experimental data analysis, G.M. conceived the analysis method and supervised practical experiments, C.L. wrote the manuscript, and all authors provided feedback on the manuscript.

### DECLARATION OF INTERESTS

G.M. holds a patent for biopolymer sensing and sequencing based on FRAC Actinoporin (patent number US20190292235A1) and has filed a patent for single-molecule recognition by chop-n-drop.

Received: May 31, 2021

Revised: August 10, 2021

Accepted: September 28, 2021

Published: October 22, 2021

## REFERENCES

- Abello, N., Kerstjens, H.A., Postma, D.S., and Bischoff, R. (2007). Selective acylation of primary amines in peptides and proteins. *J. Proteome Res.* 6, 4770–4776.
- Aebersold, R., Agar, J.N., Amster, I.J., Baker, M.S., Bertozzi, C.R., Boja, E.S., Costello, C.E., Cravatt, B.F., Fenselau, C., Garcia, B.A., et al. (2018). How many human proteoforms are there? *Nat. Chem. Biol.* 14, 206.
- Bjellqvist, B., Hughes, G.J., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J.-C., Frutiger, S., and Hochstrasser, D. (1993). The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis* 14, 1023–1031.
- Brinkerhoff, H., Kang, A.C., Liu, J., Aksimentiev, A., and Dekker, C. (2021). Infinite re-reading of single proteins at single-amino-acid resolution using nanopore sequencing. *bioRxiv*. <https://doi.org/10.1101/2021.07.13.452225>.
- Brocchieri, L., and Karlin, S. (2005). Protein length in eukaryotic and prokaryotic proteomes. *Nucl. Acids Res.* 33, 3390–3400.
- Collins, G.A., Gomez, T.A., Deshaies, R.J., and Tansey, W.P. (2010). Combined chemical and genetic approach to inhibit proteolysis by the proteasome. *Yeast* 27, 965–974.
- Di Muccio, G., Rossini, A.E., Di Marino, D., Zollo, G., and Chinappi, M. (2019). Insights into protein sequencing with an  $\alpha$ -hemolysin nanopore by atomistic simulations. *Sci. Rep.* 9, 1–8.
- Houghtaling, J., Ying, C., Eggenberger, O.M., Fennouri, A., Nandivada, S., Acharjee, M., Li, J., Hall, A.R., and Mayer, M. (2019). Estimation of shape, volume, and dipole moment of individual proteins freely transiting a synthetic nanopore. *ACS Nano* 13, 5231–5242.
- Hu, Z.-L., Huo, M.-Z., Ying, Y.-L., and Long, Y.-T. (2021). Biological nanopore approach for single-molecule protein sequencing. *Angew. Chem.* 60, 14738–14749.
- Hu, R., Rodrigues, J.V., Waduge, P., Yamazaki, H., Cressiot, B., Chishti, Y., Makowski, L., Yu, D., Shakhnovich, E., Zhao, Q., et al. (2018). Differential enzyme flexibility probed using solid-state nanopores. *ACS Nano* 12, 4494–4502.
- Huang, G., Voet, A., and Maglia, G. (2019). FraC nanopores with adjustable diameter identify the mass of opposite-charge peptides with 44 dalton resolution. *Nat. Commun.* 10, 1–10.
- Huang, G., Willems, K., Soskine, M., Wloka, C., and Maglia, G. (2017). Electro-osmotic capture and ionic discrimination of peptide and protein biomarkers with FraC nanopores. *Nat. Commun.* 8, 1–11.
- Keshishian, H., Burgess, M.W., Specht, H., Wallace, L., Clauser, K.R., Gillette, M.A., and Carr, S.A. (2017). Quantitative, multiplexed workflow for deep analysis of human blood plasma and biomarker discovery by mass spectrometry. *Nat. Protoc.* 12, 1683.
- Li, Z., Arnaud, L., Rockwell, P., and Figueiredo-Pereira, M.E. (2004). A single amino acid substitution in a proteasome subunit triggers aggregation of ubiquitinated proteins in stressed neuronal cells. *J. Neurochem.* 90, 19–28.
- Lucas, F.L.R., Sarthak, K., Lenting, E.M., Coltan, D., van der Heide, N.J., Versloot, R.C.A., Aksimentiev, A., and Maglia, G. (2021). The manipulation of the internal hydrophobicity of FraC nanopores augments peptide capture and recognition. *ACS Nano* 15 (6), 9600–9613.
- Mutter, N.L., Huang, G., van der Heide, N.J., Lucas, F.L.R., Galenkamp, N.S., Maglia, G., and Wloka, C. (2021). Preparation of Fragaecotoxin C (FraC) Nanopores (Nanopore Technology), pp. 3–10.
- Nanda, J.S., and Lorsch, J.R. (2014). Labeling of a protein with fluorophores using maleimide derivitization. *Methods Enzymol.* 536, 79–86.
- Nivala, J., Mulroney, L., Li, G., Schreiber, J., and Akeson, M. (2014). Discrimination among protein variants using an unfoldase-coupled nanopore. *ACS Nano* 8, 12365–12375.
- Ohayon, S., Girsault, A., Nasser, M., Shen-Orr, S., and Meller, A. (2019). Simulation of single-protein nanopore sensing shows feasibility for whole-proteome identification. *PLoS Comput. Biol.* 15, e1007067.
- Ouldali, H., Sarthak, K., Ensslen, T., Piguet, F., Manivet, P., Pelta, J., Behrends, J.C., Aksimentiev, A., and Oukhaled, A. (2020). Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore. *Nat. Biotechnol.* 38, 176–181.
- Piguet, F., Ouldali, H., Pastoriza-Gallego, M., Manivet, P., Pelta, J., and Oukhaled, A. (2018). Identification of single amino acid differences in uniformly charged homopolymeric peptides with aerolysin nanopore. *Nat. Commun.* 9, 1–13.
- Pontén, F., Jirstrom, K., and Uhlen, M. (2008). The human protein atlas—a tool for pathology. *J. Pathol.* 216, 387–393.
- Restrepo-Pérez, L., Joo, C., and Dekker, C. (2018). Paving the way to single-molecule protein sequencing. *Nat. Nanotechnol.* 13, 786–796.
- Righetti, P.G., and Boschetti, E. (2013). Introducing low-abundance species in proteome analysis. In *Low-Abundance Proteome Discovery* (Elsevier), pp. 1–11.
- Specht, H., Emmott, E., Petelski, A.A., Huffman, R.G., Perlman, D.H., Serra, M., Kharchenko, P., Koller, A., and Slavov, N. (2021). Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biol.* 22, 1–27.
- Sutherland, T.C., Long, Y.-T., Stefureac, R.-I., Bediako-Amoa, I., Kraatz, H.-B., and Lee, J.S. (2004). Structure of peptides investigated by nanopore analysis. *Nano Lett.* 4, 1273–1277.
- Swaminathan, J., Boulgakov, A.A., Hernandez, E.T., Bardo, A.M., Bachman, J.L., Marotta, J., Johnson, A.M., Anslyn, E.V., and Marcotte, E.M. (2018). Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. *Nat. Biotechnol.* 36, 1076–1082.
- Vistain, L.F., and Tay, S. (2021). Single-cell proteomics. *Trends Biochem. Sci.* 46, 661–672.
- Waduge, P., Hu, R., Bandarkar, P., Yamazaki, H., Cressiot, B., Zhao, Q., Whitford, P.C., and Wanunu, M. (2017). Nanopore-based measurements of protein size, fluctuations, and conformational changes. *ACS Nano* 11, 5706–5716.
- Wilson, J., Sarthak, K., Si, W., Gao, L., and Aksimentiev, A. (2019). Rapid and accurate determination of nanopore ionic current using a steric exclusion model. *ACS Sens.* 4, 634–644.
- Yao, Y., Docter, M., Van Ginkel, J., de Ridder, D., and Joo, C. (2015). Single-molecule protein sequencing through fingerprinting: computational assessment. *Phys. Biol.* 12, 055003.
- Yusko, E.C., Bruhn, B.R., Eggenberger, O.M., Houghtaling, J., Rollings, R.C., Walsh, N.C., Nandivada, S., Pindrus, M., Hall, A.R., Sept, D., et al. (2017). Real-time shape approximation and fingerprinting of single proteins using a nanopore. *Nat. Nanotechnol.* 12, 360–367.
- Zernia, S., van der Heide, N.J., Galenkamp, N.S., Gouridis, G., and Maglia, G. (2020). Current blockades of proteins inside nanopores for real-time metabolome analysis. *ACS Nano* 14, 2296–2307.
- Zhang, S., Huang, G., Versloot, R., Herwig, B.M., de Souza, P.C.T., Marrink, S.-J., and Maglia, G. (2020). Bottom-up fabrication of a multi-component nanopore sensor that unfolds, processes and recognizes single proteins. *bioRxiv*. <https://doi.org/10.1101/2020.12.04.411884>.



## STAR★METHODS

### KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
<b>Bacterial and virus strains</b>		
electrocompetent <i>E. coli</i> ® EXPRESS BL21 (DE3) strain cells	Lucigen	Cat#60300
<b>Chemicals, peptides, and recombinant proteins</b>		
His6-tagged G13F Fragaceatoxin C (FraC)	Lucas et al., 2021	N/A
[Met5]-Enkephalin	Sigma Aldrich	M6638, CAS: 82362-17-2
[d-Ala2][d-Leu5]-Enkephalin	Sigma Aldrich	E7131, CAS: 94825-57-7
<b>Deposited data</b>		
Electrophysiological traces and simulated fingerprints	This paper	DOI: <a href="https://doi.org/10.5281/zenodo.5116022">10.5281/zenodo.5116022</a>
UniProt Human proteome UP000005640	UniProt	<a href="https://www.uniprot.org/proteomes/UP000005640">https://www.uniprot.org/proteomes/UP000005640</a>
<b>Software and algorithms</b>		
Simulation and analysis code	This paper	DOI: <a href="https://doi.org/10.5281/zenodo.5116022">10.5281/zenodo.5116022</a>

### RESOURCE AVAILABILITY

#### Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Carlos de Lannoy ([carlos.delannoy@wur.nl](mailto:carlos.delannoy@wur.nl)).

#### Materials availability

- The amino acid sequence for His6-tagged G13F Fragaceatoxin C (FraC) is included in the [method details](#) section.
- Plasmid pT7-SC1 containing Fragaceatoxin C (FraC) with a modification of glycine 13 to phenylalanine and a C-terminal His6-tag was obtained from an earlier study (Lucas et al., 2021).
- Other reagents are available without restrictions, from sources mentioned for each reagent.

#### Data and code availability

All generated data and analysis code was deposited on github, at: [https://github.com/cvdelannoy/chop\\_n\\_drop\\_simulation](https://github.com/cvdelannoy/chop_n_drop_simulation).

- Ion current measurement data have been deposited at Github and are publicly available as of the date of publication. DOIs are listed in the key resource table.
- All simulation and analysis code has been deposited at Github and is available as of the date of publication. DOIs are listed in the key resources table.
- For fingerprinting simulations we used all available sequences in the UniProt human proteome (UP000005640).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

### METHOD DETAILS

#### Fragaceatoxin C purification

Plasmid DNA containing the sequence for His6-tagged Fragaceatoxin C (FraC) with a phenylalanine modification on position G13 was transformed into electrocompetent *E. coli* ®EXPRESS BL21 (DE3) strain cells (Lucigen) for protein expression. Transformed cells were plated on Lysogeny broth (LB) plates containing 100 µg/mL ampicillin and incubated overnight at 37°C. A single colony was cultured in LB medium containing 100 µg/mL ampicillin and finally protein expression was induced using 0.5 mM

isopropyl- $\beta$ -D-thiogalactopyranoside (IPTG). Bacteria were harvested, treated with 0.2 units/mL DNase I and 20  $\mu$ g/ $\mu$ L lysozyme, and sonicated, after which the mixture was incubated with Ni-NTA beads. FraC monomers were washed from the beads using elution buffer containing 0.300M imidazole. Sphingomyelin (porcine, brain):1,2-Diphytanoyl-sn-glycero-3-phosphocholine (DPhPC) liposomes for FraC oligomerization were created by solubilizing equal weight parts sphingomyelin and DPhPC in a pentane-ethanol mixture and evaporating the solvent to create a lipid film. The film was then solubilized in 0.015M Tris-HCl dilution buffer (pH 7.5, 0.150M NaCl), briefly sonicated, frozen at  $-20^{\circ}\text{C}$  and thawed. Liposomes and FraC monomers were mixed at a mass ratio of 10:1 solubilized in 0.6% (w/v) N,N-Dimethyldodecylamine-N-oxide (LDAO) and diluted in 0.02% (w/v) Dodecyl- $\beta$ -D-maltoside (DDM). Finally oligomerized FraC was bound to Ni-NTA beads and eluted in oligo elution buffer (0.200M  $\text{Na}_2\text{EDTA}$ , pH 8.0, 0.02% (w/v) DDM). This protocol was previously described in (Mutter et al., 2021)

### His6-tagged fragaceatoxin C sequence

MASADVAGAVIDFAGLGFVLKTVLEALGNVKKRIAVGIDNESGKTWTAMNTYFRSGTSDIVLPHKVAHGKAL  
LYNGQKNRGPVATGVVGVIAYSMSDGNLAVLFSVPYDYNWYSNWWNVRVYKQKQRADQRMYEELYHRS  
PFRGDNGWHSRGLGYLKSFRGMNSSGHAILEIHVTKAGSAHHHHHHH.

### Electrophysiological recordings

For planar lipid bilayer electrophysiology measurements, a constant voltage of  $-70$  mV was applied over a single nanopore, and a buffer solution containing 1M KCl buffered to pH 3.8 using 50mM Citric acid titrated with bis-tris-propane was used. [Met5]-Enkephalin and [d-Ala2][d-Leu5]-Enkephalin (Sigma Aldrich) were added in 10 $\mu$ M concentration. Ionic currents were recorded using an Axopatch 200B amplifier coupled with a Digidata 1440a or Digidata 1550B A/D converter (Molecular Devices). All data was recorded using Clampex 10 (Molecular Devices) with a sampling frequency of 50 kHz and an analog Bessel filter of 10 kHz. Measurement procedures were first described in (Lucas et al., 2021).

### Electrophysiological recording analysis

Data was analyzed using Python 3.7 and is contained within a Jupyter notebook, available in the Github repository noted in the key resources table. Events from translocating peptides were characterized using a threshold search algorithm combined with a generalized flat-top normal distribution fit (Lucas et al., 2021). The resolution between the two peptides was determined using the difference between the peak centers and their standard deviation (Lucas et al., 2021).

### In silico fingerprint generation

Code for *in silico* fingerprint generation and classification was written in Python 3.8 (Python Software Foundation, [www.python.org](http://www.python.org)). We generate *in silico* chop-n-drop fingerprints by splitting protein sequences at protease target sites and calculating the weights of the resulting fragments from their sequences. We assume that fragments of a weight lower than 500 Da are undetectable, thus these fragments are removed from fingerprints. Fragments of a weight larger than 1.6 kDa are set to 1.6kDa, as prior investigations showed that the relationship between weight and current blockage is not monotonic above this weight (Huang et al., 2019). In simulations where fragments were selected based on charge, the fragment charge was calculated using biopython (v1.78), which employs a concentration ratio-based calculation (Bjellqvist et al., 1993).

Three parameters are set to represent different noise sources; capture rate  $C$ , proteasome efficiency  $e_p$  and resolution  $r$ . The capture rate denotes the fraction of fragments that enters the pore after lysis and is measured. In our simulations each fragment is retained with a probability of  $C$ . The proteasome efficiency denotes the fraction of target sites at which the proteasome cleaves. In simulations, each target site has a probability of  $e_p$  of being cleaved. Note that a failure to cleave will result in two fragments being fused together, after which they remain represented in the fingerprint as the sum of their weights. Finally, the resolution denotes the minimum difference in fragment weight that can still be detected by current blockage, expressed in Da. We adhere to an experimentally found minimum resolution of 4Da (Figure S1). In our simulations, the resolution is represented by the magnitude of Gaussian noise added to fingerprint weights. Specifically, we define the standard deviation of the distribution from which a noise value  $n$  is drawn such, that the probability of a fragment size measurement deviating  $r$  or less from its actual size is fifty percent:

$$P(n \leq r) = 0.5 \quad (\text{Equation 1})$$

The standard deviation enforcing this resolution  $r$ ,  $\sigma_r$ , can be found using the Z-score formula, in which the Z-score is calculated using the inverse cumulative distribution function of the standard normal distribution,  $\Phi^{-1}$ :

$$\sigma_r = \frac{-r}{\Phi^{-1}(0.5/2)} \quad (\text{Equation 2})$$

As resolution is expressed as a positive number and Equation 2 considers the lower tail of a distribution centered at 0, the resolution is multiplied by  $-1$ .

### Simulation and classification

We ran *in silico* digestions on all sequences in the UniProt human proteome (UP000005640). To compile a database of fingerprints with known identity, we first performed an *in silico* digestion under noiseless circumstances (i.e.  $C = 1.0$ ,  $e_p = 1.0$  and  $r = 0.0\text{Da}$ ). Then we ran several subsequent digestions for a range of values for  $C$ ,  $e_p$  and  $r$ . Fingerprints from these runs were classified by aligning them to database fingerprints obtained from noiseless digestions.

We gauge the similarity of query and database fingerprints by aligning them using a dynamic programming algorithm (Figure S2). The dynamic programming table is filled as follows:

$$S(i, j) = \min \begin{cases} |X_i - Y_j| + S(i-1, j-1) \\ |X_i + X_{i-1} - Y_j| + S(i-2, j-1) & i \leq |X|, j \leq |Y| \\ |X_i - Y_j| + S(i-2, j-1) + G \\ |X_i - Y_j| + S(i-1, j-2) + G \end{cases} \quad (\text{Equation 3})$$

with the following conditions for edge cases to ensure that the alignment is global:

$$\begin{aligned} S(0, 0) &= 0 \\ S(i, 0) &= \infty \quad \forall \quad 1 \leq i \leq N_X \\ S(0, j) &= \infty \quad \forall \quad 1 \leq j \leq N_Y \end{aligned} \quad (\text{Equation 4})$$

Here  $S(i, j)$  is the distance between query and database fingerprints  $X$  and  $Y$  respectively, up to fragments  $X_i$  and  $Y_j$  and  $G$  is a gap penalty.  $N_X$  and  $N_Y$  are the numbers of fragments in  $X$  and  $Y$  respectively. At each step in the alignment one of three actions may be taken. First, a single fragment of each fingerprint may be aligned, in which case the absolute difference of their weights is added to the total score. Second, two fragments of  $X$  may be aligned to one fragment of  $Y$ , corresponding to a missed proteasome target site. This action increases the score by the difference between the summed weight of the former and the single weight of the latter. Third, a gap may be introduced in either  $X$  or  $Y$  at the cost of a penalty. The gap penalty  $G$  is dependent on the resolution used during digestion:

$$G = (1.96 \cdot \sigma_r)^2 + L \quad (\text{Equation 5})$$

Here  $\sigma_r$  is the resolution-dependent standard deviation of Gaussian noise added to fragment sizes during *in silico* digestion (Equation 2) and  $L$  is the lower detection limit ( $L = 500\text{ Da}$ ). This means that introducing a gap is preferred over matching fragments if the difference between fragment weights exceeds the difference expected in 95 percent of correct matches. The addition of  $L$  is required to ensure that a match is still preferred if a normally undetected fragment (i.e. of which the weight is under  $L$ ) is fused to another fragment due to a missed proteasome target site.

A query fingerprint is classified by aligning it to all fingerprints in the database and assigning it the identity of the database fingerprint to which the distance is smallest.