Biosensors and Bioelectronics 180 (2021) 113108

Contents lists available at ScienceDirect

Biosensors and Bioelectronics

journal homepage: http://www.elsevier.com/locate/bios

Short communication

Single-aminoacid discrimination in proteins with homogeneous nanopore sensors and neural networks

David Rodriguez-Larrea

Biofisika Institute (CSIC, UPV/EHU) and Department of Biochemistry and Molecular Biology (UPV/EHU), Barrio Sarriena S/n, Leioa, 48940, Spain

ARTICLE INFO

Keywords:
Single-molecule
Nanopore
Neural network
Protein sequencing

ABSTRACT

A technology capable of sequencing individual protein molecules would revolutionize our understanding of biological processes. Nanopore technology can analyze single heteropolymer molecules such as DNA by measuring the ionic current flowing through a single nanometer hole made in an electrically insulating membrane. This current is sensitive to the monomer sequence. However, proteins are remarkably complex and identifying a single residue change in a protein remains a challenge. In this work, I show that simple neural networks can be trained to recognize protein mutants. Although these networks are quickly and efficiently trained, their ability to generalize in an independent experiment is poor. Using a thermal annealing protocol on the nanopore sample, and examining many mutants with the same nanopore sensor are measures aimed at reducing training data variability which produce an increase in the generalizability of the trained neural network. Using this approach, we obtain a 100% correct assignment among 9 mutants in >50% of the experiments. Interestingly, the neural network performance, compared to a random guess, improves as more mutants are included in the dataset for discrimination. Engineered nanopores prepared with high homogeneity coupled with state-of-the-art analysis of the ionic current signals may enable single-molecule protein sequencing.

1. Introduction

The number of protein species produced by a genome vastly exceeds the number of genes (Ponomarenko et al., 2016; Smith and Kelleher, 2013). Variable promoter usage, alternative splicing of pre-mRNA transcripts, and alternative translation initiation produce a diversity of protein isoforms which, in addition, may suffer one or multiple post-translational modifications. Each unique combination of these variables is known as a 'proteoform' (Smith and Kelleher, 2013), and there may be millions of them. An additional level of complexity is the concentration at which each proteoform is found (Ghaemmaghami et al., 2003), notably this variable ultimately determines the phenotypic outcome. To date, the proteome size and composition remain largely unknown (Aebersold and Mann, 2016).

We lack appropriate methods to analyze the enormous complexity of the proteome. State-of-the-art mass spectrometry is by far the most potent tool available (Aebersold and Mann, 2016). Bottom-up proteomics can identify thousands of proteins in a complex mixture, but by analyzing peptide fragments, they produce a puzzle that cannot be univocally solved (Schaffer et al., 2019). Top-down proteomics can distinguish between closely related proteoforms (Donnelly et al., 2019). However, since this method requires the purification/separation of the sample constituents, it is hardly quantitative, not suitable for low copy number proteins, and the analysis of proteins >50 kDa remains a challenge (Schaffer et al., 2019). There is a need for methods capable of single-molecule protein sequencing in complex mixtures in order to determine the proteome size and composition.

Nanopore technology is a single-molecule technique that allows sequencing of individual DNA molecules (Check Hayden, 2015; Cockroft et al., 2008; Maglia et al., 2010) and has been proposed for single-molecule protein sequencing (Restrepo-Pérez et al., 2018). It works by measuring the ionic current flowing through a single, nanometer-wide pore made in an electrically insulating membrane (Maglia et al., 2010). Molecules being threaded through the nanopore occupy a significant portion of its lumen and cause a decrease in the ionic current. A heteropolymer threading through the nanopore modulates the net current depending on which monomer of the polymer is lying within the pore. Despite the sensitivity of the technique, proteins pose an additional challenge if compared to DNA. They are made of at least 20 different and sometimes very similar amino acids (e.g., Ile and Leu, Ser and Thr, Ala and Ser), which in addition may undergo post-translational modifications. Amino acids are also smaller than

E-mail address: david.rodriguezl@ehu.es.

https://doi.org/10.1016/j.bios.2021.113108
Received 28 October 2020; Received in revised form 8 February 2021; Accepted 19 February 2021
Available online 24 February 2021

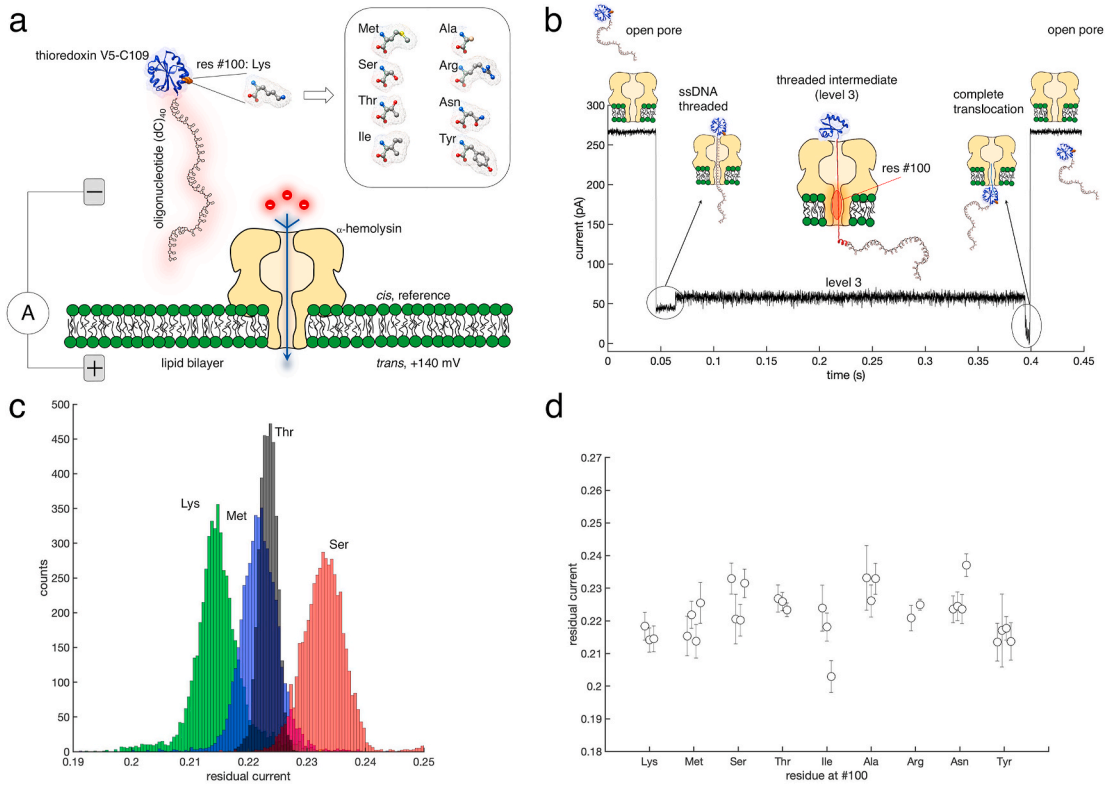

Fig. 1. Single-molecule analysis of thioredoxin with nanopores. a) Diagram showing the different parts of the experimental set-up: a 40mer oligonucleotide is covalently attached to the C-terminus of thioredoxin. *Inbox* amino acids explored at position #100. In response to the applied electric field, negative ions flow through a lipid embedded hemolysin pore towards the cathode. b) The ionic current signal characteristic of a protein-oligonucleotide complex is captured, pulled, unfolded, and translocated through the nanopore at +140 mV, 2 M KCl at both sides of the membrane. This study aims to use the ionic current observed during the intermediate threaded state (level 3) to identify the residue at position #100. c) Histogram of the residual currents of 10 ms long segments derived from level 3 when the amino acid at position #100 was Lys (green, number of thioredoxin molecules analyzed = 74), Met (blue, n = 87), Thr (black, n = 4) and Ser (red, n = 46). Each mutant was analyzed in an independent experiment (i.e., analyzed with a different nanopore sensor molecule). d) Mean (circles) and standard distribution (error bars) obtained from fitting the residual currents (Supplementary Fig 1-a) gaussian distribution. The experimental variability observed among independent replicas, in addition to the width of the observed residual current distribution, prevents accurate discrimination of the residue occupying position #100 based on the residual current. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

nucleotides, meaning that more amino acid residues than nucleotides can simultaneously fit in the nanopore, and therefore the number of different sequences is larger. In addition, each residue contributing marginally to the ionic current signal. Furthermore, proteins are usually folded and, in that state, cannot enter the nanopore (Nivala et al., 2013; Rodriguez-Larrea and Bayley, 2013). If linearized in order to enter the nanopore, their non-homogeneous charges and propensity to fold may cause a complex dynamic behaviour (Feng et al., 2020; Nanopore et al., 2014; Rodriguez-Larrea and Bayley, 2014; Rosen et al., 2020). Finally, it is unclear whether the ionic current signal arising from single protein molecules translocating through a nanopore can be useful for single-molecule protein sequencing. In this line, machine-learning algorithms hold enormous potential for classifying ionic current signals. For example, it has been used to identify amino acids (Zhao et al., 2014), sequence DNA (McIntyre et al., 2019) or to decode DNA encoded nanostructures (Misiunas et al., 2018) from their ionic current signature, but to date it has never been shown whether it can aid in single-molecule protein sequencing with nanopores.

Here I use a 4 layers deep, 10 units each, neural network (Bengio et al., 2016; Lecun et al., 2015; Rumelhart et al., 1986) (NN) for the classification of ionic current signals produced by 9 different single-residue mutants of the model protein thioredoxin as they translocate a solitary nanopore inserted in a lipid membrane. I found that the classifier is trained faster to higher accuracies with the ionic current data presented in the frequency domain. Nonetheless, the trained network failed at classifying data obtained in an independent experiment. This lack of generalization can be overcome working with nanopore samples homogenized with a thermal annealing protocol and collecting data for several mutants on the same nanopore sensor. The NN trained with new batch of data correctly identified the mutant in 54% of independent experiments (n = 28). Finally, I found that the accuracy, relative to a random guess, both improves with the number of mutants in the dataset and the number of independent experiments used to build it. Therefore, identifying a single amino acid substitution in the context of a complex protein molecule is feasible and suggest that NNs have enough power to allow single-molecule protein sequencing with nanopores.

2. Materials and methods

2.1. Substrate and nanopore sensor preparation

The substrate was a A22P–I23V–C32S–C35S–P68A-C109 (V5–C109) mutant of thioredoxin 2 from *E. coli* with a poly $(dC)_{40}$ oligonucleotide covalently linked at the C-terminal cysteine (Rodriguez-Larrea and Bayley, 2013). The protein was expressed in BL21 (DE3) cells, purified and modified following published protocols ((Celaya and Rodriguez-Larrea, 2021; Rodriguez-Larrea and Bayley, 2013), see Supplementary Information for details).

The nanopore sensor was hetameric α-hemolysin (α-HL) from *S. aureus*. A plasmid with the gene encoding the monomer was expressed in BL21 (DE3) cells, purified and heptamerized according published protocols ((Celaya and Rodriguez-Larrea, 2021), see Supplementary

Information for details).

2.2. Single-molecule measurements

Briefly, I built a 1,2-diphytanoyl-sn-glycero-3-phosphocholine (DPhPC, Avanti Polar Lipids) membrane that separated two compartments, each with an Ag/AgCl electrode connected to the headstage of an Axopatch 200b amplifier (Molecular Devices) connected to a PC through a digitizer (either Digidata 1440A or 1550A, Molecular Devices). Each compartment was filled with 2 M KCl, 10 mM Hepes, pH 7.4. The α-HL sample was added to the *cis* (ground) side and following the first insertion the *cis* compartment was perfused with fresh buffer to avoid further insertions. Then the substrate sample was added and analyzed at +140 mV (see Supplementary Information for details).

2.3. Data processing and neural network

Raw data was collected with Clampex 10.3 and further analyzed with MatLab (the software codes can be found in the Supplementary Information). Briefly, the ionic current signals corresponding to the translocation of single substrate molecules were identified, selected and divided in 10-ms segments. The NNs were built in Python using Keras (Chollet, n.d.) with TensorFlow (Martín Abadi et al., 2015) as backend. The fully connected network consisted in an input layer, 4 inner fully connected layers with 10 units each and 'relu' activation functions, and the last layer had 9 units and a 'softmax' activation function. Training was carried out on an iMac Pro 2018 and a MacBook Pro 2019. (More details and full codes can be found in the Supplementary Information).

3. Results and discussion

3.1. Discrimination of thioredoxin mutations with a neural network

The single molecule sensor is a solitary α-HL nanopore inserted in a DPhPC membrane. This system is robust and has been widely employed in the characterization of single DNA molecules (Stoddart et al., 2009). With it the four nucleobases can be discerned and sequenced (Cockroft et al., 2008). Its crystallographic structure shows a heptameric transmembrane pore with a β-barrel stem spanning the membrane decorated at the *cis* side with a protruding cap (Fig. 1a (Jones et al., 1996)). Further, the α-HL nanopore has been used for recording single protein molecules of thioredoxin unfolding and translocating the nanopore (Rodriguez-Larrea and Bayley, 2013) and it has been shown that the ionic current signal is modulated by phosphorylation of residue #100 (Rosen et al., 2014).

The co-translocational unfolding of thioredoxin produces a characteristic ionic current signal with 4 distinguishable ionic current levels (Fig. 1b). The third level is caused by an unfolding intermediate with the C-terminal part unfolded and threading through the nanopore, including residue #100, and the N-terminal part above the nanopore detector (Feng et al., 2020; Rodriguez-Larrea and Bayley, 2013; Rosen et al., 2020) (Fig. 1b). To explore the protein sequencing capabilities of nanopores I produced nine thioredoxin variants, all at position #100 (K, M, S, T, R, Y, A, I, or N) to keep the background sequence constant, and focused the analysis on the ionic current signals of level 3.

Discrimination between nanopore signals caused by different molecular entities is frequently based on the residual current (I_{res}), i.e., how much of the open pore current remains when the molecule is inside the detector (Manrao et al., 2012; Stoddart et al., 2009). I analyzed 9 different mutants and repeated each experiment in at least three different α-HL nanopore molecules. In some cases, mutants may be distinguished based on their I_{res} (Fig. 1c). However, I_{res} shows experimental variability (Fig. 1d), which prevents correct residue assignment based on this parameter alone as previously reported for oligonucleotides (Stoddart et al., 2009). Nonetheless, the ionic current signal contains additional information. More recently, the ionic current signal produced by a protein molecule inside a wide nanopore of 15–30 nm in diameter, has been analyzed beyond the I_{res} to provide parameters such as charge, dipole moment, or even shape (Houghtaling et al., 2019; Yusko et al., 2017). Indeed, the ionic current signal produced by a protein threading the nanopore may be seen as a fingerprint. Despite recent advances, for narrower nanopores (Wilson et al., 2019), it is not yet possible to predict the ionic current from the atomic-scale geometry of the nanopores and amino acids (Javidpour et al, 2008, 2009). Nonetheless, deep NNs do not require prior knowledge of this link and could potentially use properties found in the data to assign an ionic current signal to particular mutants translocating the nanopore (Lin et al., 2017).

To test this, I chose a fully connected, deep feedforward, NN comprised of an input layer, 4 hidden layers with 10 neurons each, and an output layer with 9 neurons (one for each mutant, see Supplementary Information). This is a fairly simple deep NN which performs well in tasks such as digit recognition (>90% accuracy on the MINST database, data not shown). Data were processed as follows (Supplementary Fig. 2): i) Data were acquired with a 100 kHz low-pass Bessel filter and sampled at 250 kHz or 500 kHz -and later downsampled to 250 kHz. ii) For each single-molecule translocating the nanopore I extracted the data corresponding both to level 3 (I_{lvl3}) as well as the open pore (I_{lvl1}) right before the molecule entered; iii) each reading I_{lvl3} was split into overlapping 10-ms segments (S) and normalized:

$$NormS = \frac{\left(S - \bar{I}_{lvl3}\right)}{\sigma_{I_{lvl1}}} + \frac{\bar{I}_{lvl3}}{\bar{I}_{lvl1}} \quad (1)$$

where *NormS* is the normalized segment, $\bar{I}_{lvl3}$is the mean of I_{lvl3}, $\sigma_{I_{lvl1}}$ is the standard deviation of I_{lvl1}, and $\bar{I}_{lvl1}$ is the mean of I_{lvl1}. The motivation to work with overlapping segments was two-fold: firstly, it augments dataset size; and secondly, it allows each experiment to contribute equally to the final dataset. The degree of overlapping was chosen cautiously to avoid overloading the computer memory. Each mutant was analyzed in at least 3 different experiments of different nanopore molecules. The final dataset contained ~5000 normalized segments of 10-ms each, for each experiment (n = 32), which required ~2.5 gigabytes of memory.

Before training the NN, each column (feature) of the dataset was Z-score normalized, the data (rows) shuffled, and the NN fit to 90% of the dataset, while the remaining 10% was used as a validation set. Training stopped if no improvement was observed in the validation-loss after 10 epochs. At this point, the accuracy in segment assignment was 48% in the training set and 44% in the validation set. Both values are higher than random assignment (~11.1%), which suggests the NN can learn to classify the ionic current signals.

As an alternative, I explored how the NN performed with data presented in the frequency domain, an approach that performs better in deep learning audio signals (Hertel et al., 2016). The power spectral density (PSD) of each normalized segment gives a 2046-feature vector, each feature represents the amplitude of a 60 Hz frequency range. Before proceeding to Z-score normalization, each feature was scaled:

$$SS = \ln(PSD + 0.001) \quad (2)$$

where *SS* is the scaled signal. Then I Z-score normalized each column (feature) of the dataset, shuffled the data, and fit 90% of the data to the NN while the remaining 10% of the data was used as a validation set. Training stopped when the validation loss did not improve for 10 consecutive epochs. When this happened, the NN correctly assigned 84% of the examples in the training set and 86% in the validation set. Longer training (training ceased after 50 consecutive epochs without improvement), further improves the accuracy (>90%). This shows that a simple NN can accurately classify the signals arising from 9 different mutants particularly well if presented in the frequency domain.

Nonetheless, it does not imply that the NN generalizes well, i.e.,

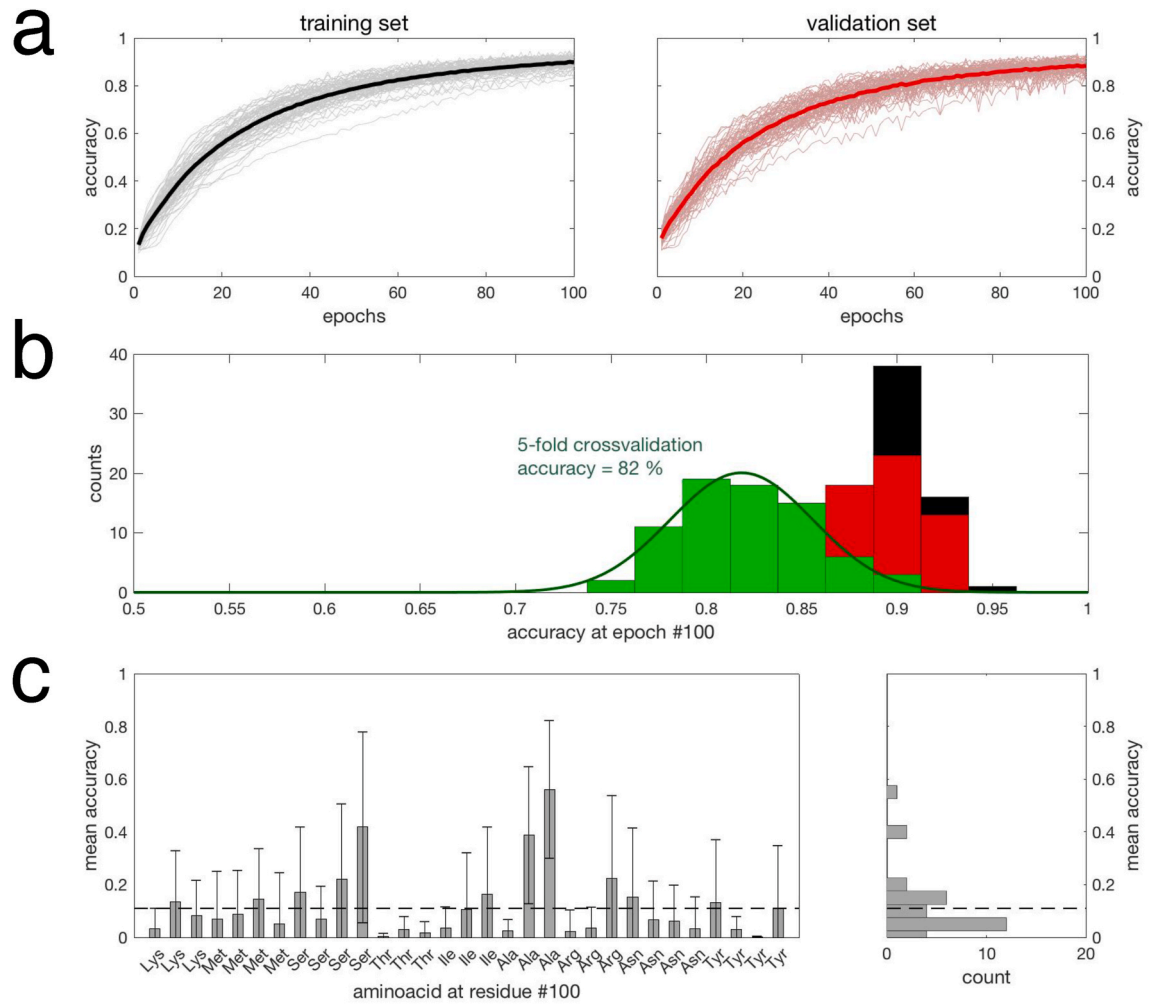

Fig. 2. The NN fits the data but lacks generalization. a) The data is split into 5 blocks. 4 blocks were used for training the NN (90% of the data. left panel, black) and validation (10% of the data, right panel, red). Each block was analyzed 15 times for 100 epochs (optimization cycles), and the averages of all training and validation results at each epoch are shown in dark black and red, respectively. b) Histogram showing the accuracies after 100 epochs on the training set (black), validation set (red), and the cross-validation set (green). c) Bar chart (left) showing the mean accuracies and standard deviation in the classification when all the data coming from a single-pore was used to test the generalization of the fitted NN (rather than using 20% of its data as cross-validation). The correct amino acid is indicated at the bottom. On the right, histogram of the observed accuracies. The dashed line in both panels indicates the estimated accuracy from a random guess (11%). When the NN generalization is tested with data coming from a nanopore that does not contribute to the training set, the resulting model lacks generalization power. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

accurately classifies data that is not included in the training and validation sets. To evaluate the generalization, I performed a 5-fold cross-validation as follows (Bengio et al., 2016): first, I divided the data collected in each nanopore into 5 non-overlapping blocks, I used one of them as a cross-validation set, and the data from the other 4 blocks as training (80%) and validation sets (20%). Then I trained the NN for 100 epochs and evaluated its performance on the cross-validation set (Fig. 2a and b). This process was repeated 4 more times, using a different block as a cross-validation set each time. Again, the training and validation sets had an accuracy of >90%. And the 5-fold cross-validation accuracy was 82%.

This would be an excellent result, but it is misleading because it does not consider the experimental variability. To estimate the model generalization, the classification accuracy was evaluated on data collected in a nanopore not used in NN training (i.e., an independent experiment). I built a training set with data collected from all the nanopores except for one (n = 31) and used the data from that nanopore as a validation set. During training I monitored the validation loss and stopped if no improvement was observed over 10 consecutive epochs. In 30 cases out of 32, the validation set data was classified with ~11% accuracy, similar to what would be expected using a random classifier (Fig. 2c). In summary, optimized NNs accurately classify ionic current signals arising from single protein molecules translocating through nanopores. Nonetheless, the differences between independent experiments lead to optimized NNs that lack any generalization ability. The challenge is how to accurately classify data collected in independent experiments.

3.2. *Obtaining neural networks that generalize well*

I hypothesized that the NN would fail to generalize well because differences between experiments have a major impact on the ionic current signal. As a consequence, the NN would learn the experiment rather than differences between mutants. One source of variability between experiments is the fact that not all the nanopores are the same (Stoddart et al., 2009). While steadily improving (Gilbert et al., 2017), synthetic, solid-state nanopores encounter this variability due to lack of control during the manufacturing process. Instead biological nanopores were considered more reproducible because they are dependent on protein folding. Nonetheless, there is also heterogeneity in the ionic

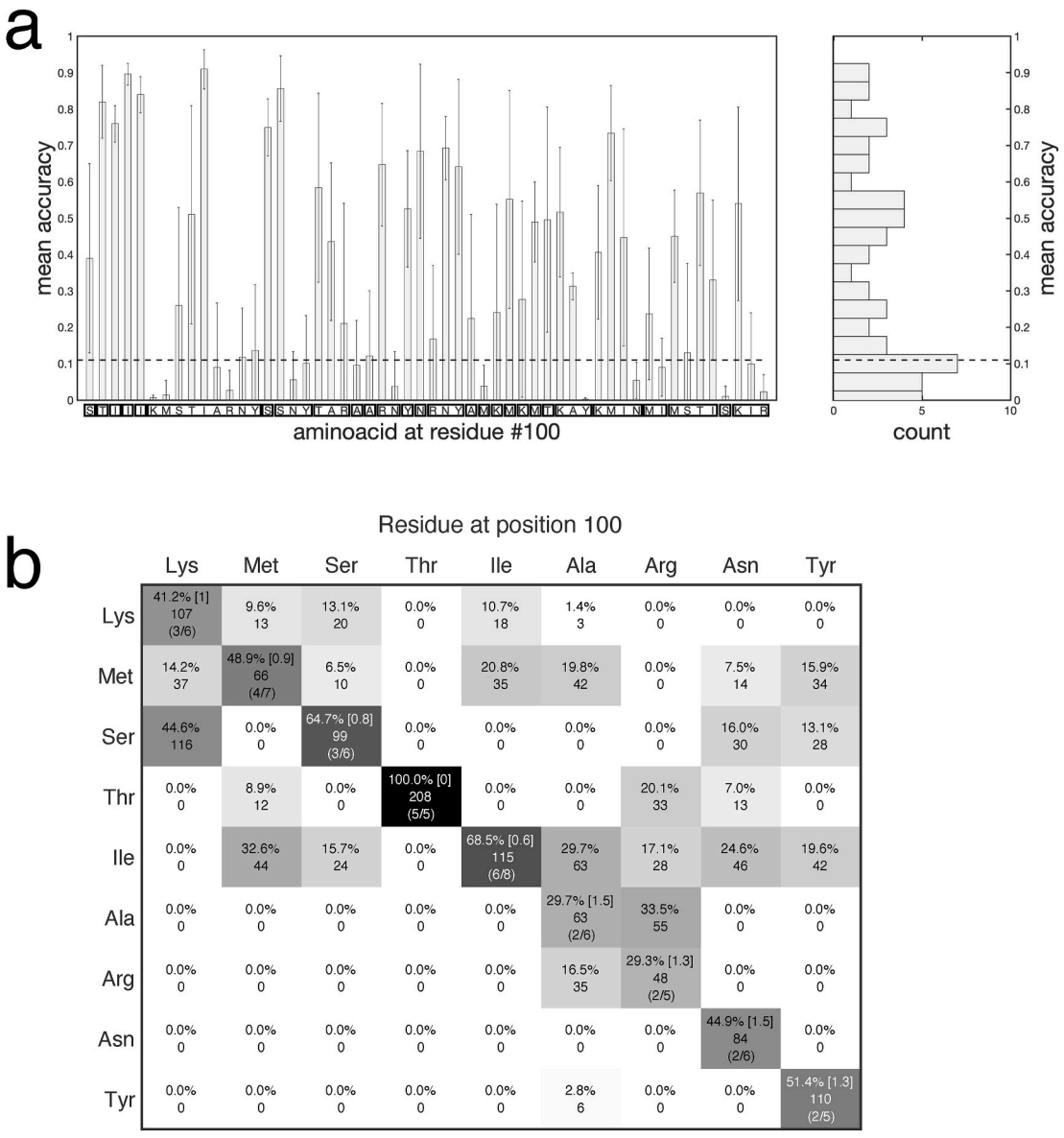

Fig. 3. Improved NN generalization. a) Left: bar chart showing the mean accuracies and standard deviation in the classification accuracy of segments coming from a nanopore that is not used in NN training. The amino acid at position #100 is below in one-letter code. Different mutants analyzed with the same nanopore are boxed. Right: histogram of the mean accuracies obtained in the left panel. The dashed line in both panels indicates the estimated accuracy from a random guess (11%). b) Confusion matrix showing the performance of the NN in predicting the thioredoxin mutant that caused a translocation signal. On diagonal cells, the percentage indicates the fraction of correctly assigned molecules x 100, in brackets the coefficient of variation; below the absolute number of molecules correctly assigned and in parenthesis the fraction of pores which showed 100% accuracy in the assignment.

current signals obtained on independent experiments (Stoddart et al., 2009), which is frequently omitted in the literature. I aimed to reduce this source of experimental variability by treating the heptameric α-hemolysin nanopore sample with a thermal treatment akin to the thermal annealing ramps used for DNA origami (Gilbert et al., 2017). The sample was heated to 70 °C and cooled to room temperature in 10 °C, 30 minute-steps, and stored at 4 °C no longer than 1 month. I expected this would allow most of the α-hemolysin molecules to reach a common free energy minimum.

Simultaneously, I examined more than one thioredoxin mutant on each α-HL nanopore. This would compel the NN to focus on features that allow distinguishing of one mutant from another rather than one experiment from another. Upon examining one mutant, I perfused the *cis* chamber with 20 vol of fresh buffer, waited for 5 mins and perfused with a further 20 vol of fresh buffer. Before adding a new mutant sample, I verified that no translocations were observed for 1 min. If no events were detected, a new mutant sample was added. Following this I examined at least 5 times each of the 9 different mutants in a total of 28 independent experiments.

For data analysis, I proceeded as before: the data collected for each mutant on each nanopore was split into 10-ms overlapping segments, each referenced to the open pore ionic current following equation (1) (~5000 segments/mutant/pore). The PSD was scaled with equation (2), and Z-score normalized. The training set was built with the same number of examples for each mutant using data from 27 nanopores, and the validation set using data from only one mutant obtained in the left nanopore. Optimization of the NN proceeded until no improvement in the validation loss was observed for 10 epochs. On average, the NN correctly classified 36 ± 28% [coefficient of variation, cv = 0.77] of the 10-ms segments. The average classification accuracy for different mutations ranged from 55 ± 35% [cv = 0.63] in the case of Isoleucine to 17 ± 15% [cv = 0.88] for Alanine, suggesting that some residues may be

either easier or harder to identify than others (Fig. 3a). Overall, the results improve if the whole signal arising from the translocation of a protein molecule is considered rather than single 10 ms segments. Each translocation signal may contain tens or hundreds of these segments. The averaged prediction of them increases the classification accuracy at the single-molecule level to 100% in 54% of the experiments (n = 28, Fig. 3b).

This improvement was obtained by both increasing the homogeneity of the nanopore sample and by analyzing several mutants with the same nanopore molecule. To estimate the effect of the homogenization, I prepared data subsets on which each mutant was collected on 3 different nanopores, and each nanopore only contributed data for one mutant. This dataset can be built in 140 different ways; I randomly picked 20 of them, and trained the NN as described before. On average, 10-ms segments were assigned with 26 ± 26% [cv = 1] accuracy (ranging from 70 ± 27% [cv = 0.38] for Ile to 10 ± 7% [cv = 0.7] for Ser).

To estimate the contribution to increased accuracy of analyzing more than one mutant on the same nanopore, I constructed the training dataset with data from 7 nanopores so that each mutant was analyzed on 3 different nanopores (7 was the minimal number of nanopores needed to build the dataset with the available data). Training the NN with this dataset led to 24 ± 19% [cv = 0.79] average accuracy in signal classification. Mutations such as Thr were correctly assigned in 60 ± 23% [cv = 0.38] of the cases while others such as Ala did not improve on random assignment (7 ± 12% [cv = 1.71]). Importantly, I did not observe any correlation in the accuracies obtained with each approach (r = −0.053, p-value = 0.89, Supplementary Fig. 4). Therefore, it is likely that both homogenization of the nanopore sample and examining several mutants with the same nanopore detector contribute to the observed improvement in mutant identification with NN.

Next, I examined the effect of collecting data on multiple nanopores. Ile mutation was studied on 8 different nanopores, and data collected in experiment #4 was correctly assigned to Ile in 90 ± 5% [cv = 0.05] of the cases. I trained the NN providing Ile examples for the training process coming from a range of 1 up to 7 different nanopores (all possible combinations of them were considered) and evaluated the accuracy in predicting the data of experiment #4. The accuracy in the assignment increased monotonically from 40 ± 40% [cv = 1] to 90 ± 5% [cv = 0.05] as data from more pores were included in the training set. As expected the more nanopore sensors that are used to build the training data set, the higher the prediction accuracy.

Finally, I asked how the number of mutations in the dataset influences the classification performance of the NN. I generated 42 random ways in which each of the 9 mutants were sequentially added to a data subset. Upon each addition, the NN was trained and evaluated in each independent experiment within the data subset (repeated 10 times, and the results averaged). The observed trend was that the more mutations that were present in the data subset used to train the NN, the lower the accuracy in the classification (Supplementary Fig. 5). Nonetheless, the accuracy of the NN relative to a random guess increases as more mutations are added to the data subset (Supplementary Fig. 5), suggesting that this simple NN would perform even better in a dataset with more mutants to discriminate between.

4. Conclusions

In summary, I have shown that NNs can learn from the ionic currents produced by single protein molecules translocating a nanopore to discern single residue mutations. Generalization of the trained NN improves as more nanopore sensors are used to build the training set, and by reducing the heterogeneity in the data set. Interestingly, the accuracy relative to a random guess improves as the number of different mutations in the dataset increases. This suggests that NNs would perform better with larger datasets, and therefore are a promising tool in single-molecule protein sequencing.

The system here described could be used to distinguish and quantify closely related forms of a protein -such as single point mutations or post-translational modifications (Rosen et al., 2014)- because the number of possible outcomes is limited to a few tens. Instead, true sequencing would require further developments similar to those used for single-molecule DNA sequencing, namely i) the use of a molecular motor that produces a step-by-step, processive, translocation through the nanopore (Cockroft et al., 2008), and ii) the use of nanopores with narrow sensing transmembrane regions (Ayub et al., 2015; Faller et al., 2004; Manrao et al., 2012). Both introductions would result in smaller peptide stretches inside the nanopore and, therefore, in a reduction in the number of signals that must be discerned. This would allow to predict from the ionic current signal the most likely sequence of a molecule translocating the nanopore.

CRediT authorship contribution statement

David Rodriguez-Larrea: This work was fully authored by.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

I thank Andrina Chambers for critical reading of the manuscript and helpful suggestions. I thank G. Celaya for the preparation of hemolysin monomers and erythrocyte membranes. DRL work was supported by grants BIO2017-88946-R, BFU2016-81754-ERC from MINECO (FEDER funds), IT1201-19 from the Basque Government, and a Ramon y Cajal fellowship (RYC-203-12799). I also thank Fundacion Biofisika Bizkaia for their support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bios.2021.113108.

References

Aebersold, R., Mann, M., 2016. Nature.
Ayub, M., Stoddart, D., Bayley, H., 2015. ACS Nano 9, 7895–7903.
Bengio, Y., Courville, A., Goodfellow, I., 2016. MIT Press 5, 118–120.
Celaya, G., Rodriguez-Larrea, D., 2021. In: Methods in Molecular Biology. Humana Press Inc., pp. 135–144
Check Hayden, E., 2015. Nature.
Abadi, Martín, Agarwal, Ashish, Barham, Paul, Brevdo, Eugene, Chen, Zhifeng, Citro, Craig, Corrado, Greg S., Davis, Andy, Dean, Jeffrey, Devin, Matthieu, Ghemawat, Sanjay, Goodfellow, Ian, Harp, Andrew, Irving, Geoffrey, Isard, Michael, Jozefowicz, Rafal, Jia, Yangqing, Kaiser, Lukasz, Kudlur, Manjunath, Levenberg, Josh, Mané, Dan, Schuster, Mike, Monga, Rajat, Moore, Sherry, Murray, Derek, Olah, Chris, Shlens, Jonathon, Steiner, Benoit, Sutskever, Ilya, Talwar, Kunal, Tucker, Paul, Vanhoucke, Vincent, Vasudevan, Vijay, Viégas, Fernanda, Vinyals, Oriol, Warden, Pete, Wattenberg, Martin, Wicke, Martin, Yu, Yuan, Zheng, Xiaoqiang, 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from. tensorflow.org.
Chollet, F., 2015. Keras. https://keras.io.
Cockroft, S.L., Chu, J., Amorin, M., Ghadiri, M.R., 2008. J. Am. Chem. Soc. 130, 818–820.
Donnelly, D.P., Rawlins, C.M., DeHart, C.J., Fornelli, L., Schachner, L.F., Lin, Z., Lippens, J.L., Aluri, K.C., Sarin, R., Chen, B., Lantz, C., Jung, W., Johnson, K.R., Koller, A., Wolff, J.J., Campuzano, I.D.G., Auclair, J.R., Ivanov, A.R., Whitelegge, J. P., Paša-Tolić, L., Chamot-Rooke, J., Danis, P.O., Smith, L.M., Tsybin, Y.O., Loo, J.A., Ge, Y., Kelleher, N.L., Agar, J.N., 2019. Nat. Methods 16, 587–594.
Faller, M., Niederweis, M., Schulz, G.E., 2004. Science (80-) 303, 1189–1192.
Feng, J., Martin-Baniandres, P., Booth, M.J., Veggiani, G., Howarth, M., Bayley, H., Rodriguez-Larrea, D., 2020. Commun. Biol. 3.
Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., Weissman, J.S., 2003. Nature 425, 737–741.
Gilbert, S.M., Dunn, G., Azizi, A., Pham, T., Shevitski, B., Dimitrov, E., Liu, S., Aloni, S., Zettl, A., 2017. Sci. Rep. 7, 1–7.
Hertel, L., Phan, H., Mertins, A., 2016. arXiv::1603.05824v1.

Houghtaling, J., Ying, C., Eggenberger, O.M., Fennouri, A., Nandivada, S., Acharjee, M., Li, J., Hall, A.R., Mayer, M., 2019. ACS Nano 13, 5231–5242.
Javidpour, L., Tabar, M.R.R., Sahimi, M., 2009. J. Chem. Phys. 130, 85105.
Javidpour, L., Tabar, M.R.R., Sahimi, M., 2008. J. Chem. Phys. 128, 115105.
Jones, M.J., Murray, A.W., Chem, J.B., Chen, M., Latinis, K.M., Song, L., Hobaugh, M.R., Shustak, C., Cheley, S., Bayley, H., Gouauxt, J.E., 1996. Science 80–, 274.
Lecun, Y., Bengio, Y., Hinton, G., 2015. Nature.
Lin, H.W., Tegmark, M., Rolnick, D., 2017. J. Stat. Phys. 168, 1223–1247.
Maglia, G., Heron, A.J., Stoddart, D., Japrung, D., Bayley, H., 2010. Methods in Enzymology. Elsevier Inc.
Manrao, E.A., Derrington, I.M., Laszlo, A.H., Langford, K.W., Hopper, M.K., Gillgren, N., Pavlenok, M., Niederweis, M., Gundlach, J.H., 2012. Nat. Biotechnol. 30, 349–353.
McIntyre, A.B.R., Alexander, N., Grigorev, K., Bezdan, D., Sichtig, H., Chiu, C.Y., Mason, C.E., 2019. Nat. Commun. 10, 1–11.
Misiunas, K., Ermann, N., Keyser, U.F., 2018. Nano Lett. 18, 4040–4045.
Nanopore, U.U., Nivala, J., Mulroney, L., Li, G., Schreiber, J., Akeson, M., 2014. ACS Nano 12365–12375.
Nivala, J., Marks, D.B., Akeson, M., 2013. Nat. Biotechnol. 31, 247–250.
Ponomarenko, E.A., Poverennaya, E.V., Ilgisonis, E.V., Pyatnitskiy, M.A., Kopylov, A.T., Zgoda, V.G., Lisitsa, A.V., Archakov, A.I., 2016. Int. J. Anal. Chem., 7436849
Restrepo-Pérez, L., Joo, C., Dekker, C., 2018. Nat. Nanotechnol. 13, 786–796.
Rodriguez-Larrea, D., Bayley, H., 2014. Nat. Commun. 5, 1–7.
Rodriguez-Larrea, D., Bayley, H., 2013. Nat. Nanotechnol. 8, 288–295.
Rosen, C.B., Bayley, H., Rodriguez-Larrea, D., 2020. Commun. Biol. 3, 160.
Rosen, C.B., Rodriguez-Larrea, D., Bayley, H., 2014. Nat. Biotechnol. 32, 179–181.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Nature 323, 533–536.
Schaffer, L.V., Millikin, R.J., Miller, R.M., Anderson, L.C., Fellers, R.T., Ge, Y., Kelleher, N.L., LeDuc, R.D., Liu, X., Payne, S.H., Sun, L., Thomas, P.M., Tucholski, T., Wang, Z., Wu, S., Wu, Z., Yu, D., Shortreed, M.R., Smith, L.M., 2019. Proteomics.
Smith, L.M., Kelleher, N.L., 2013. Nat. Methods 10, 186–187.
Stoddart, D., Heron, A.J., Mikhailova, E., Maglia, G., Bayley, H., 2009. Proc. Natl. Acad. Sci. Unit. States Am. 106, 7702–7707.
Wilson, J., Sarthak, K., Si, W., Gao, L., Aksimentiev, A., 2019. ACS Sens. 4, 634–644.
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., Li, J., Kalonia, D.S., Mayer, M., 2017. Nat. Nanotechnol. 12, 360–367.
Zhao, Y., Ashcroft, B., Zhang, P., Liu, H., Sen, S., Song, W., Im, J., Gyarfas, B., Manna, S., Biswas, S., Borges, C., Lindsay, S., 2014. Nat. Nanotechnol. 9, 466–473.