Direct Detection of Twenty Amino Acids and Discrimination of Pathological Peptides with Functionalized Nanopore Postprint

Authors: Zhang, Ming, Tang, Chao, Wang, Zichun, Chen, Shanchuan, Xu, Mengying, Li, Kaiju, Sun, Ke, Zhao, Changjian, Wang, Yu, Dai, Lunzhi, Lu, Guangwen, Shi, Hubing, Chen, Lu, Geng, Jia, Chen, Lu, Geng, Jia

Date: 2024-03-05T00:00:00+00:00

Abstract

Single-molecule discrimination among amino acids is crucial to the realization of next-generation protein sequencing. Owing to the heterogeneous charge and subtle volume difference of underivatized amino acids, it remains a challenge for single-molecule techniques to recognize each of them. Here, we report the direct detection of twenty proteinogenic amino acids using a copper(II)-functionalized MspA nanopore. The binding sites for copper(II) ion are constructed by introducing histidine mutation (N91H) to M2MspA protein. With copper ion binding to histidine residues, amino acids can reversibly coordinate the copper-histidine complex, generating well-defined current signals. Using this strategy, all twenty amino acids can be detected. Assisted by a machine learning algorithm, we can identify 100% of signals with 70.2% accuracy or 60% of signals with 93.4% accuracy in the validation set. In successively addition experiment, each amino acid in a mixture of 10 amino acids can be identified precisely. Furthermore, we use carboxypeptidase A1 to partly release the C-terminal amino acids of peptides with different lengths (9, 10 and 22 residues). The hydrolysates of peptides can be identified and distinguished. These results demonstrate the feasibility of this system for amino acids detection and peptide identification, shedding new lights on the development of single-molecule protein sequencing.

Full Text

Preamble

Direct Detection of Twenty Amino Acids and Discrimination of Pathological Peptides with Functionalized Nanopores

Ming ZHANG[1]#, Chao TANG[2]#, Zichun WANG[1]#, Shanchuan CHEN[1]#, Mengying XU[3], Kaiju LI[1], Ke SUN[1], Changjian ZHAO[1], Yu WANG[1], Lunzhi DAI[4], Guangwen LU[5], Hubing SHI[6], Lu CHEN[3]* & Jia GENG[1]*

[1]Department of Laboratory Medicine, State Key Laboratory of Biotherapy and Cancer Center, Med-X Center for Manufacturing, West China Hospital, Sichuan University, Chengdu, 610041, China
[2]Key Laboratory of Birth Defects and Related Diseases of Women and Children of MOE, Department of Laboratory Medicine, State Key Laboratory of Biotherapy, West China Second University Hospital, Sichuan University, Chengdu, 610041, China
[3]Department of Laboratory Medicine, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu, 610041, China
[4]National Clinical Research Center for Geriatrics and Department of General Practice, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu 610041, China
[5]West China Hospital Emergency Department (WCHED), State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu, 610041, China
[6]Laboratory of Tumor Targeted and Immune Therapy, Clinical Research Center for Breast, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University and Collaborative Innovation Center, Chengdu, 610041, China

#These authors contributed equally to this work.

*Correspondence to: geng.jia@scu.edu.cn or luchen@scu.edu.cn

Abstract

Single-molecule discrimination among amino acids is crucial for realizing next-generation protein sequencing. Owing to the heterogeneous charge and subtle volume differences of underivatized amino acids, recognizing each of them remains a challenge for single-molecule techniques. Here, we report the direct detection of twenty proteinogenic amino acids using a copper(II)-functionalized MspA nanopore. Binding sites for copper(II) ions were constructed by introducing a histidine mutation (N91H) to the M2MspA protein. With copper ions bound to histidine residues, amino acids can reversibly coordinate to the copper-histidine complex, generating well-defined current signals. Using this strategy, all twenty amino acids can be detected. Assisted by a machine learning algorithm, we can identify 100% of signals with 70.2% accuracy or 60% of signals with 93.4% accuracy in the validation set. In successive addition experiments, each amino acid in a mixture of ten amino acids can be precisely identified. Furthermore, we used carboxypeptidase A1 to partially release the C-terminal amino acids of peptides of different lengths (9, 10, and 22 residues). The hydrolysates of these peptides can be identified and distinguished. These results demonstrate the feasibility of this system for amino acid detection and peptide identification, shedding new light on the development of single-molecule protein

sequencing.

Introduction

Amino acids are the building blocks of proteins and raw materials for biosynthesis, playing fundamental roles in various physiological and pathophysiological processes such as epigenetic regulation and tumor metabolism[1–4]. Therefore, detecting and identifying amino acids with higher spatiotemporal resolution is crucial and has recently aroused great interest among researchers, particularly in the field of single-molecule protein sequencing[5–8]. Due to alternative RNA splicing and post-translational modifications, the resulting proteoforms are highly complex and contain deeper-level information that cannot be accessed directly from the transcriptome[9]. In addition, there is no existing method to amplify proteins similar to DNA amplification. Consequently, it is difficult for mass spectrometry-based methods to identify low-abundance proteins from proteomes[10],[11]. To address these problems, single-molecule methods that can distinguish the twenty proteinogenic amino acids must be developed for protein sequencing.

For fluorophore-based methods, specific amino acids like cysteine and lysine can be selectively modified with fluorescent molecules. Then, by sequentially degrading peptides using Edman chemistry or direct imaging using single-molecule FRET, the relative positions of labeled amino acids can be deduced from the fluorescent signals[12–14]. Additionally, fluorophore-labeled N-terminal amino acid recognizers have been engineered to bind specific amino acids reversibly[15],[16]. The repetitive signals of the same amino acid can greatly improve identification accuracy[17]. Although these methods have high throughput and good reliability, it is difficult for chemists to label all twenty amino acids. Label-free methods such as tunneling current measurement[18],[19] and molecular junctions[20] enable rapid and precise detection of up to twelve amino acids. Given that the nanopore technique has demonstrated its superiority in single-molecule DNA sequencing, it has also been considered an ideal candidate for amino acid detection and protein sequencing[5],[21],[22]. Studies have shown that peptides with different properties can be directly detected and distinguished, including molecular weight[23],[24], length[25],[26], post-translational modifications[27],[28], and single-amino acid mutations[29]. To further analyze the amino acid sequence of a peptide, the translocation of the peptide must be well controlled to generate sequence-dependent signals. Protein unfoldase ClpX was used to unfold and drive proteins through a nanopore, enabling different segments of proteins to be discerned[30]. Moreover, the ratcheting motion of DNA-peptide conjugates through nanopores was achieved using DNA helicase or polymerase, generating clear sequence-dependent signals[31–33]. Single-file translocation of linearized proteins can be facilitated by engineering electroosmotic force[34],[35]. Unfortunately, it is challenging to deconvolute signals contributed by 5–6 amino acids, as twenty types of amino acids generate more combinations than four types of nucleotides. Therefore, analysis of single amino acid translocation events could

provide valuable information. Underivatized amino acids can be detected using copper ion-modified α-hemolysin and MoS_2 nanopores[36],[37]. Furthermore, taking advantage of pore structure, the aerolysin nanopore is capable of distinguishing thirteen out of twenty amino acids using only a polyarginine carrier[38]. Biosensors that can discriminate all twenty proteinogenic amino acids directly with high sensitivity and specificity at the single-molecule level are required for biosensing and protein sequencing.

Here, we report the direct identification of twenty proteinogenic amino acids using a copper(II)-functionalized MspA nanopore. Benefiting from its conical pore geometry, MspA nanopore has proven to be an ideal choice for sensing ions and small molecules[39],[40]. We constructed binding sites for copper ions by introducing histidine mutations in the restriction region of the pore lumen. With copper ions bound to histidine residues, reversible coordination between amino acids and the copper-histidine complex could generate well-defined current signals. Next, a machine learning-based classifier was trained for identifying the twenty amino acids. Furthermore, three types of peptides (α-Bag Cell Peptide (1-9), adrenocorticotropic hormone (ACTH, 18-39), and Angiotensin I) were partially hydrolyzed from the C-terminus using carboxypeptidase A1. The composition of their hydrolysates could be identified accurately. These results suggest the great potential of this system for amino acid detection and peptide identification, paving the way for next-generation protein sequencing.

Results

Sensing of Twenty Amino Acids Using Copper(II)-MspA

In a typical experiment, amino acids and copper(II) chloride are added to the cis (grounded) and trans chambers, respectively (Fig. 1a [Figure 1: see original paper]). The binding sites for copper ions are located at the constriction region of the MspA nanopore (Fig. 1b). For each of the eight subunits, the 91st asparagine is substituted by histidine to create a copper-binding structure similar to a histidine brace motif[41]. We suppose that one 90th asparagine residue and two adjacent 91st histidine residues could reversibly coordinate a single copper ion and then an amino acid (Fig. 1c). The corresponding three binding states could be observed as stepwise current changes (Fig. 1d). According to this supposition, there are at least four binding sites for copper ions, and the reversible binding of multiple copper ions was also observed[42]. However, such stochastic binding events interfered with precise assay of subsequent amino acid binding. To keep the current baseline at a constant level (I_0), excess copper ions at a final concentration of 200 M were added to saturate the binding sites during most of the measuring time (ca. 87.8 $\pm$ 3.1%) (Supplementary Table 1).

All twenty amino acids can be detected and produce clear signals with high reproducibility (Fig. 1e). Commonly, blockade ($|I_1\text{-}I_0|/I_0$) and dwell time (Δt) are analyzed to characterize the signals. The signal blockades for each type of amino acid exhibit a unimodal distribution (Fig. 2a [Figure 2: see original

paper]). Most of them can be well distinguished from each other. The mean blockade and its standard deviation for each amino acid were calculated from the mean value of the Gaussian fit. The blockades show a good positive correlation with the volume of amino acids (Fig. 2b). When amino acids with charged side groups and proline (P) are excluded, the coefficient of determination of linear fitting reaches up to 0.92 (Supplementary Fig. 1), indicating that the generation process of current blockade for these amino acids obeys the classical volume exclusion model. While for amino acids with charged side groups, the volume exclusion model is no longer applicable[43]. The blockade of aspartic acid, glutamic acid, and histidine (D, E, and H) is larger than expected, which could be attributed to possible interactions between their side chains and the copper-histidine complex. For lysine and arginine (K and R), electrical repulsion between their amine groups and copper ions was expected to lower the current blockade. Due to the strong interaction between copper ions and the sulfhydryl group of cysteine (C), the binding of copper ions to histidine residues could be extremely unstable (Supplementary Fig. 2). The fluctuation of open pore current made it difficult to determine the baseline current I_0 and also shortened its duration. Therefore, few signals of cysteine were extracted, causing a high standard deviation of the mean blockade.

In terms of signal frequency, there are remarkable differences among different amino acids (Fig. 2c, e). Among them, the signal frequency of P is the lowest because of its unique structure, which can hardly interact with copper ions. The mean signal frequency of polar amino acids is significantly higher than that of nonpolar amino acids. For amino acids with charged side chains, signal frequencies of K and R are significantly lower than those of negatively charged D and E. One reason may be that K and R can be driven away from the nanopore by electrophoretic force. The isoelectric point of H (7.59) is quite close to 7.5, and the interaction between H and copper ions is very strong, so its signal frequency is even higher than that of D and E. Therefore, both the charge distribution of amino acids and their interaction with copper ions contribute to signal frequency. To demonstrate that the α-amine group and α-carboxyl group of amino acids are essential for coordination, we synthesized two dipeptides, EF and γEF (peptide bond is formed using the γ-carboxyl group of glutamate). The translocation of dipeptide γEF generated distinguishable signals with larger blockade ($37.8 \pm 0.6\%$ vs $15.3 \pm 0.2\%$) and longer dwell time (2.521 ± 0.149 ms vs 0.554 ± 0.014 ms) than dipeptide EF (Supplementary Fig. 3 [Figure 3: see original paper]). Since all these amino acids bind to a copper ion with their α-amine group and α-carboxyl group, the mean dwell times of signals are in the same order of magnitude, ranging from 1.87 ms (G) to 6.86 ms (W) (Fig. 2d). Additionally, acetylated leucine and amidated leucine did not generate characteristic signals, providing direct evidence for our chemical model (Supplementary Fig. 4 [Figure 4: see original paper]).

Amino Acid Identification by Machine Learning

Multiple amino acids can bind to the nanopore simultaneously, resulting in superimposed multi-level signals (Supplementary Fig. 2). These signals cannot be identified simply by the two parameters of blockade and dwell time. To improve the accuracy of amino acid identification, we trained a machine learning-based classifier (Fig. 3a).

We first normalized the classified amino acid signals by dividing the current amplitude of signals by their baseline current (I_0). The distribution of current density of each normalized signal was divided into 1000 equally sized intervals from 0 to 1. Features X0001-X1000 were then calculated from the data of their corresponding intervals, representing the density of data points within each interval. The mean blockade, dwell time, standard deviation, and features X0001-X1000 of normalized signals were used as input features to train the classifier using a machine learning algorithm. Next, we assessed the performance of six different classification algorithms. Results showed that the random forest (RF) algorithm performed best, and the corresponding receiver operating characteristic (ROC) analysis revealed an area under the curve (AUC) of 0.9708. Features ranging from X150 to X178 have larger importance values compared with others (Fig. 3b). These features were generated from signal points within the range of level 1 blockade of amino acids, indicating that the portion of the signal generated from the single amino acid binding event remains the most important for signal identification.

To minimize the influence of background signals and further improve the accuracy of amino acid identification, we filtered signals by applying cutoff values to dwell time. Signals with dwell times lower than 1 ms, 2 ms, 3 ms, 4 ms, 5 ms, and 6 ms were filtered in RF1-6 models, respectively. We found that with increasing cutoff values, the performance of models improved despite the reduction in the number of available signals (Supplementary Fig. 6a [Figure 6: see original paper], b, c). The results indicate that signals with higher dwell time are more likely to be correctly identified. The AUCs of ROC in the test set and validation set increased from 0.9837 and 0.9708 to 0.9936 and 0.9838, respectively (Supplementary Fig. 6d, e). In addition, the sensitivity, specificity, precision, recall, and F1 score of random forest classification models improved significantly (Supplementary Fig. 7a [Figure 7: see original paper]), and the prediction probability of corrected signals is also enriched around 1 (Supplementary Fig. 7b).

The classification accuracy of the test set ranged from 75.64% to 87.66% with 100% signal recovery. When a prediction probability cutoff (>0.7) was used to filter out unclear results, the average classification accuracy increased up to 95.64% with 63.85% signal recovery (Supplementary Fig. 6f). In the validation set, the accuracy ranged from 70.2% to 83.2% with 100% signal recovery. When only signals with a prediction probability greater than 0.7 were considered, the average classification accuracy increased up to 93.4% with 60% signal recovery

(Fig. 3c, Supplementary Fig. 7b). These results indicate that there is no overfitting in these models. In model RF4, the accuracy reached up to 95% with all amino acids showing good differentiation except C, N, and T (Fig. 3d).

Furthermore, in successive addition experiments, the results showed that each amino acid in a mixture of ten proteinogenic amino acids and S-carboxymethyl-L-cysteine can be precisely identified (Supplementary Fig. 5 [Figure 5: see original paper]).

Discrimination of Peptides by Recognizing Their Hydrolysates

As it is challenging to sequence a linear peptide directly, detecting individual amino acids cleaved from a peptide may offer an alternative approach. We next tested the feasibility of this system for identifying peptide hydrolysates. Carboxypeptidase A1 was used to sequentially release single amino acids from the C-terminus of peptides.

The hydrolysis reaction stops when any of three amino acids (K, R, and P) becomes the first amino acid at the C-terminus (Fig. 4a). This approach not only avoids detecting K, R, and P but also produces limited types of amino acids, reducing the complexity of amino acid identification. The results showed that peptide hydrolysis could be monitored in real time after mixing carboxypeptidase A1 with peptide EAFNL directly in the cis chamber (Fig. 4b). To make the hydrolysis reaction more complete, three types of peptides (α-Bag Cell Peptide (1-9), ACTH (18-39), and Angiotensin I) were hydrolyzed respectively with a higher concentration of carboxypeptidase A1 at 37 °C, then added to the cis chamber. Their hydrolysates were detected and identified, which were mostly consistent with the theoretical amino acid products (Fig. 4c). For all three types of peptides, the percentage of leucine signals is lower than other theoretical products (Fig. 4d), similar to the trend of signal frequency when amino acids were detected separately (Fig. 2c). For ACTH (18-39) and Angiotensin I, some hydrolysates were identified incorrectly, as F and L were identified as other amino acids with similar blockades. Although the hydrolysates of these peptides can be distinguished, the accuracy for identifying amino acid mixtures still needs improvement.

Discussion

In summary, this study enables the direct detection of twenty proteinogenic amino acids using a copper ion-functionalized MspA nanopore. The interaction between the α-amine group and α-carboxyl group of amino acids and the copper-nanopore complex is considered key to generating current blockade. Consequently, unnatural amino acids are also expected to be detectable. A shortcoming of this method is the presence of multiple binding sites for copper ions and amino acids due to the sequence homology of eight MspA subunits, which results in superimposed event signals (Supplementary Fig. 2). Therefore, MspA with two adjacent N91H-mutant subunits is required for binding of a single

copper ion and amino acid, which helps reduce signal complexity and improve sensing accuracy[44],[45]. In addition, we noticed that the maximum blockade of amino acids only accounted for ~1/4 of the open pore current, which could be attributed to the large pore diameter and short translocation duration of amino acids. Therefore, a narrower pore with stronger interaction with amino acids should be rationally designed[46],[47]. It is observed that besides hydrophobic volume, the charge of amino acids has considerable influence on current blockade. This needs to be further investigated combined with theoretical calculation and may provide a valuable reference for polypeptide sequencing[43],[48]. Compared with analysis using blockade and dwell time as parameters, the machine learning algorithm developed here allows analysis of all data points of one signal, improving accuracy for amino acid identification. We also demonstrated that C-terminal amino acids of peptides could be partially released by carboxypeptidase A1 and identified using this system. As single-molecule protein sequencing would revolutionize proteomics research and has valuable practical applications such as identifying neoantigens, this method needs to be further developed toward single-molecule resolution. It may be beneficial to construct a peptidase-nanopore conjugation or even a nanopore with peptidase activity[49],[50], which could enable hydrolysis and sequencing of a single peptide.

Methods

Protein Preparation

M2MspA-N91H mutant was expressed and purified as described previously[42],[51]. Briefly, a gene of M2MspA with the 91st histidine mutation for each of eight subunits was cloned into a pET28b vector. The plasmid was heat-shock transformed into E. coli BL21 (DE3) competent cells. The cells were cultured in LB medium containing kanamycin (50 g/mL). When the OD_{600} reached 0.6–0.8, 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) was added. Afterward, cells were incubated at 15 °C for 12 h with shaking at 220 rpm. The cells were harvested by centrifugation at 4000 rpm, 4 °C for 15 min and then resuspended. Cell disruption was performed by sonication using an ultrasonic cell disruption device. The supernatant was retained, and the target protein was further purified using an anion exchange column (Q-Sepharose) and size exclusion column (Superdex 200 16/90).

Amino Acid Detection and Peptide Hydrolysis

Electrophysiology experiments were performed using a classical vertical lipid bilayer setup (Warner Instruments). A pair of Ag/AgCl electrodes were placed in the trans and cis (grounded) sides of the chamber, which was filled with 1 mL of electrolyte solution (1 M KCl, 10 mM MOPS, pH 7.5). The planar lipid bilayer membrane was formed on a 150 m-diameter aperture by painting a thin film of 1,2-diphytanoyl-sn-glycero-3-phosphocholine (DPhPC) (Avanti Polar Lipids). A voltage of +300 mV was applied to induce nanopore insertion after adding the MspA protein (final concentration of 60–90 ng/mL) to the cis chamber. After

a single nanopore insertion, $CuCl_2$ solution was added to the trans chamber to a final concentration of 200 M (20 M in peptide hydrolysis experiments). High-purity L-amino acids (>98%) were dissolved in Milli-Q water away from light immediately before use. To collect more signals, amino acids were added to the cis chamber to a high final concentration of 100 M (except 5 M, 200 M, and 2 M for H, P, and C, respectively). For peptide hydrolysis, the peptide was dissolved in Milli-Q water to a final concentration of 2 mM. 8 L of peptide solution was mixed with 2 L of 3.3 U carboxypeptidase A1 and reacted at 37 °C for 15 min, then the mixture was added to the cis chamber. For real-time monitoring of peptide hydrolysis, peptide N' -EAFNL-C' was added to the cis chamber to a final concentration of 20 M, followed by the addition of 10 L of 16.7 U carboxypeptidase A1.

Electrophysiology Recording

Single-channel current recordings were amplified using an Axopatch 200B amplifier (Molecular Devices) and filtered with a built-in four-pole low-pass Bessel filter at 2 kHz. Data were digitized by a Digidata 1550B converter (Molecular Devices) at a sampling rate of 100 kHz. All electrophysiology experiments were performed at room temperature (23 ± 2 °C).

Signal Extraction for Amino Acid Translocation Events

First, to reduce noise in raw current recordings, we calculated the optimal changepoints of current according to the mean and variance and polished the current recording using the mean current of each segmented time range according to the identified changepoints. Then we extracted translocation events from the polished signal based on a minimum blockade threshold value (0.1) against the baseline current. For all extracted events, we calculated the blockade, dwell time, and standard deviation of signal current (SD). Additionally, to better describe the characteristics of each signal, we uniformly extracted the density values of 1000 points from the density curve of the normalized current of each signal as characteristic values.

Raw Signal Filtering Based on Similarity with Background Noise

For the raw signals of each independent experiment, we randomly selected the same number of noise signals from the corresponding blank control to calculate the Euclidean distance matrix. The eigenvalue of Euclidean distance is the predicted value of the machine learning model.

Classification Model Training

We developed a machine learning (ML) algorithm to automatically predict the corresponding amino acid from the signal of a translocation event. The strategy was to utilize an algorithm to "learn" from classified training data and build an optimum classification model to recognize unknown events. To train the model,

the blockade, dwell time, SD value, and estimated density values of the normalized signal were calculated using R program to form a feature matrix (Fig. 3). For each amino acid, we randomly selected one experimental dataset as the independent validation set, then randomly selected 80% of all remaining signals as the training dataset and 20% as the test set. Model training was performed using the R package caret. A set of classifiers including random forest (RF), naive Bayes (NB), K nearest neighbors (KNN), Bagged CART, AdaBoost.M1 (AdaBoost), and neural networks (NNet) were tested. To prevent model overfitting, 10-fold cross-validation was performed for each model. Considering that the dwell time of the signal reflects signal quality, and signals with dwell time lower than 1 ms may be generated from spontaneous gating of the nanopore, signals were then filtered in each model to improve accuracy. Signals with dwell times lower than 1 ms, 2 ms, 3 ms, 4 ms, 5 ms, and 6 ms were filtered in models RF1, RF2, RF3, RF4, RF5, and RF6, respectively. Finally, the trained model was used to predict unclassified events.

Data Availability

The datasets generated and/or analyzed in this study are available within the source data. All data supporting the findings of this study are available from the corresponding authors upon reasonable request.

Code Availability

The experimental data were analyzed using R software, and all in-house developed codes and algorithms used in this study are available at https://github.com/LuChenLab/AANanopore.git.

Acknowledgments

This project was funded by the National Key Research and Development Program of China (Grant No. 2022YFB3205600), Science & Technology Department of Sichuan Province (Grant No. 2020YFS0579), and the 1 • 3 • 5 project for disciplines of excellence, West China Hospital, Sichuan University (Grant No. ZYYC20011 to J.G.).

Author Contributions

J.G., L.C., and M.Z. conceived the project. M.Z. and Z.C.W. performed the electrophysiology measurements for amino acid detection and peptide identification. C.T. wrote the signal processing programs and analyzed the data with assistance from M.Y.X., S.C.C., and Z.C.W. K.J.L. prepared the MspA protein. K.S., C.J.Z., Y.W., L.Z.D., G.W.L., and H.B.S. contributed to experimental design. J.G., L.C., M.Z., C.T., and Z.C.W. wrote the manuscript, and all other authors commented on it.

Competing Interests

Sichuan University has filed patent applications for methods described herein, with J.G., L.C., M.Z., C.T., Z.C.W., and S.C.C. listed as inventors.

References

1. Lieu, E. L., Nguyen, T., Rhyne, S. & Kim, J. Amino acids in cancer. *Exp. Mol. Med.* 52, 15–30 (2020).
2. Vettore, L., Westbrook, R. L. & Tennant, D. A. New aspects of amino acid metabolism in cancer. *Br. J. Cancer* 122, 150–156 (2020).
3. Thandapani, P. et al. Valine tRNA levels and availability regulate complex I assembly in leukaemia. *Nature* 601, 428–433 (2022).
4. Maddocks, O. D. K. et al. Modulating the therapeutic response of tumours to dietary serine and glycine starvation. *Nature* 544, 372–376 (2017).
5. Alfaro, J. A. et al. The emerging landscape of single-molecule protein sequencing technologies. *Nat. Methods* 18, 604–617 (2021).
6. Restrepo-Pérez, L., Joo, C. & Dekker, C. Paving the way to single-molecule protein sequencing. *Nat. Nanotechnol.* 13, 786–796 (2018).
7. Hu, Z. L., Huo, M. Z., Ying, Y. L. & Long, Y. T. Biological Nanopore Approach for Single-Molecule Protein Sequencing. *Angew. Chemie - Int. Ed.* 60, 14738–14749 (2021).
8. Cressiot, B., Bacri, L. & Pelta, J. The Promise of Nanopore Technology: Advances in the Discrimination of Protein Sequences and Chemical Modifications. *Small Methods* 4, 1–13 (2020).
9. Zhu, Y. et al. Nanodroplet processing platform for deep and quantitative proteome profiling of 10-100 mammalian cells. *Nat. Commun.* 9, 1–10 (2018).
10. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* 422, 198–207 (2003).
11. BAILEY, J. L. Proceedings of the Biochemical Society. *Biochem. J.* 52, i.2-xiii (1952).
12. Swaminathan, J. et al. Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. *Nat. Biotechnol.* 36, 1076–1091 (2018).
13. Van Ginkel, J. et al. Single-molecule peptide fingerprinting. *Proc. Natl. Acad. Sci. U. S. A.* 115, 3338–3343 (2018).
14. de Lannoy, C. V., Filius, M., van Wee, R., Joo, C. & de Ridder, D. Evaluation of FRET X for single-molecule protein fingerprinting. *iScience* 24, 103239 (2021).
15. Tullman, J., Callahan, N., Ellington, B., Kelman, Z. & Marino, J. P. Engineering ClpS for selective and enhanced N-terminal amino acid binding. *Appl. Microbiol. Biotechnol.* 103, 2621–2633 (2019).
16. Tullman, J., Marino, J. P. & Kelman, Z. Leveraging nature' s biomolecular designs in next-generation protein sequencing reagent development. *Appl. Microbiol. Biotechnol.* 104, 7261–7271 (2020).

17. Reed, B. D. et al. Real-time dynamic single-molecule protein sequencing on an integrated semiconductor device. *Science (80-.).* 378, 186–192 (2022).
18. Zhao, Y. et al. Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling. *Nat. Nanotechnol.* 9, 466–473 (2014).
19. Ohshiro, T. et al. Detection of post-translational modifications in single peptides using electron tunnelling currents. *Nat. Nanotechnol.* 9, 835–840 (2014).
20. Liu, Z. et al. A single-molecule electrical approach for amino acid detection and chirality recognition. *Sci. Adv.* 7, 1–10 (2021).
21. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nat. Biotechnol.* 34, 518–524 (2016).
22. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* 39, 1348–1365 (2021).
23. Lucas, F. L. R., Versloot, R. C. A., Yakovlieva, L., Walvoort, M. T. C. & Maglia, G. Protein identification by nanopore peptide profiling. *Nat. Commun.* 12, 1–9 (2021).
24. Afshar Bakshloo, M. et al. Nanopore-Based Protein Identification. *J. Am. Chem. Soc.* 144, 2716–2725 (2022).
25. Ji, Z., Kang, X., Wang, S. & Guo, P. Nano-channel of viral DNA packaging motor as single pore to differentiate peptides with single amino acid difference. *Biomaterials* 182, 227–233 (2018).
26. Piguet, F. et al. Identification of single amino acid differences in uniformly charged homopolymeric peptides with aerolysin nanopore. *Nat. Commun.* 9, 966 (2018).
27. Versloot, R. C. A. et al. Quantification of Protein Glycosylation Using Nanopores. *Nano Lett.* 22, 5357–5364 (2022).
28. Ensslen, T., Sarthak, K., Aksimentiev, A. & Behrends, J. C. Resolving Isomeric Posttranslational Modifications Using a Biological Nanopore as a Sensor of Molecular Shape. *J. Am. Chem. Soc.* 144, 16060–16068 (2022).
29. Huang, G., Voet, A. & Maglia, G. FraC nanopores with adjustable diameter identify the mass of opposite-charge peptides with 44 dalton resolution. *Nat. Commun.* 10, 835 (2019).
30. Nivala, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation through an α-hemolysin nanopore. *Nat. Biotechnol.* 31, 247–250 (2013).
31. Brinkerhoff, H., Kang, A. S. W., Liu, J., Aksimentiev, A. & Dekker, C. Multiple rereads of single proteins at single-amino acid resolution using nanopores. *Science (80-.).* 374, 1509–1513 (2021).
32. Yan, S. et al. Single Molecule Ratcheting Motion of Peptides in a Mycobacterium smegmatis Porin A (MspA) Nanopore. *Nano Lett.* 21, 6703–6710 (2021).
33. Nova, I. C. et al. Detection of phosphorylation post-translational modifications along single peptides with nanopores. *Nat. Biotechnol.* (2023)

doi:10.1038/s41587-023-01839-z.
34. Sauciuc, A., Morozzo della Rocca, B., Tadema, M. J., Chinappi, M. & Maglia, G. Translocation of linearized full-length proteins through an engineered nanopore under opposing electrophoretic force. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01954-x.
35. Yu, L. et al. Unidirectional single-file transport of full-length proteins through a nanopore. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-022-01598-3.
36. Boersma, A. J. & Bayley, H. Continuous stochastic detection of amino acid enantiomers with a protein nanopore. *Angew. Chemie - Int. Ed.* 51, 9606–9609 (2012).
37. Wang, F. et al. MoS_2 nanopore identifies single amino acids with sub-1 Dalton resolution. *Nat. Commun.* 14, 1–8 (2023).
38. Ouldali, H. et al. Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore. *Nat. Biotechnol.* 38, 176–181 (2020).
39. Cao, J. et al. Giant single molecule chemistry events observed from a tetrachloroaurate(III) embedded Mycobacterium smegmatis porin A nanopore. *Nat. Commun.* 10, 5668 (2019).
40. Wang, S. et al. Single molecule observation of hard-soft-acid-base (HSAB) interaction in engineered Mycobacterium smegmatis porin A (MspA) nanopores. *Chem. Sci.* 11, 879–887 (2020).
41. Chalkley, M. J., Mann, S. I. & DeGrado, W. F. De novo metalloprotein design. *Nat. Rev. Chem.* 6, 31–50 (2022).
42. Zhang, X. et al. Real-time sensing of neurotransmitters by functionalized nanopores embedded in a single live cell. *Mol. Biomed.* 2, 6 (2021).
43. Li, M.-Y. et al. Revisiting the Origin of Nanopore Current Blockage for Volume Difference Sensing at the Atomic Level. *JACS Au* 1, 967–976 (2021).
44. Wang, Y. et al. Identification of nucleoside monophosphates and their epigenetic modifications using an engineered nanopore. *Nat. Nanotechnol.* 17, 976–983 (2022).
45. Zhang, S. et al. A Nanopore Based Saccharide Sensor. *Angew. Chemie Int. Ed.* 61, e202203769 (2022).
46. Zhao, C. et al. High-fidelity biosensing of dNTPs and nucleic acids by controllable subnanometer channel PaMscS. *Biosens. Bioelectron.* 200, 113894 (2022).
47. Zhang, M., Chen, C., Zhang, Y. & Geng, J. Biological nanopores for sensing applications. *Proteins Struct. Funct. Bioinforma.* 90, 1786–1799 (2022).
48. Huo, M. Z., Li, M. Y., Ying, Y. L. & Long, Y. T. Is the volume exclusion model practicable for nanopore protein sequencing? *Anal. Chem.* 93, 11364–11369 (2021).
49. Zhang, S. et al. Bottom-up fabrication of a proteasome–nanopore that unravels and processes single proteins. *Nat. Chem.* 13, 1192–1199 (2021).
50. Sun, K. et al. Active DNA unwinding and transport by a membrane-adapted helicase nanopore. *Nat. Commun.* 10, 5083 (2019).

51. Butler, T. Z., Pavlenok, M., Derrington, I. M., Niederweis, M. & Gundlach, J. H. Single-molecule DNA detection with an engineered MspA protein nanopore. *Proc. Natl. Acad. Sci. U. S. A.* 105, 20647–20652 (2008).

Figure Captions

Figure 1. Experimental setup and principle of amino acid detection. (a) Schematic illustration of experimental setup. Amino acids and copper ions were added to cis and trans chambers respectively. A voltage of +50 mV was applied during measurement. The N91H mutant sites of eight subunits are highlighted in orange. (b) Bottom-view structure of M2MspA-N91H nanopore (predicted using SWISS-MODEL). The dotted box shows a binding site for copper ion. (c) Proposed sensing mechanism. Two adjacent histidine residues and one 90th asparagine residue first coordinate a copper ion. The α-amine and α-carboxyl group of amino acid then coordinate the histidine-copper complex. (d) A representative current trace showing the corresponding current change for three binding states in Fig. 1c. (e) A typical current signal of amino acid translocation event.

Figure 2. Characteristics of signals of twenty proteinogenic amino acids. (a) The distribution of relative abundance of the blockade of amino acid signals. (n = 4278 (E), 4211 (D), 650 (K), 193 (His1), 306 (His2), 7166 (F), 3934 (W), 2768 (Y), 3025 (I), 8004 (M), 3059 (R), 8131 (T), 8101 (S), 3750 (L), 857 (A), 1149 (G), 361 (P), 7873 (Q), 9634 (N), 2119 (V), 616 (C)). (b) Scatter plot of volume versus blockade of amino acids. For each amino acid, the blockade and its standard deviation were calculated using mean values of Gaussian fit from at least three independent experiments. Amino acids with charged side chains, nonpolar and polar amino acids are colored in green, purple and orange, respectively. Pearson correlation coefficient = 0.58, p = 0.0062. (c) Boxplot of signal frequency of amino acids. Each dot represents data from an independent experiment. (d) Mean dwell time of amino acid signals. (e) Boxplot of signal frequency of four categories of amino acids. The signal frequency of polar amino acids is significantly higher than nonpolar amino acids.

Figure 3. Amino acid identification assisted by machine learning algorithm. (a) Illustration of training process. First, classified level 1 (one amino acid binding) and level 2 (two same amino acids binding) signals of each type of amino acid were imported and normalized. Then the level 1 blockade, dwell time, and standard deviation were extracted. Additionally, 1000 data points were extracted from the current density of each signal (from 0 to 1 with an interval of 0.001), named Feature X0001-X1000. Performance of models were tested including random forest (RF), naive Bayes (NB), neural networks (NNet), K nearest neighbor (KNN), Bagged CART and AdaBoost.M1 (AdaBoost). Among these, random forest model was the best one with AUC of 0.990. A 10-fold cross-validation was used to prevent overfitting. (b) Feature importance generated from training of random forest model (RF) for L1 signals of all 20 amino acids. The upper y axis represents the corresponding blockade of each feature. Features

within the range of level 1 blockade of all amino acids have higher importance value. (c) The receiver operating characteristic curve (ROC) of random forest model for training, testing and independent validation dataset of L1 signals of all 20 amino acids. The area under curve (AUC) of different datasets are shown with colored text label. (d) Confusion matrix of amino acid identification using RF model. For amino acid Cys (C) and Lys (K), the percentage of signals was much lower than 100%, indicating they were identified incorrectly as other amino acids. For amino acid Asp (D) and Phe (F), the percentage of signals was much higher than 100%, indicating some other amino acids were incorrectly identified as these amino acids.

Figure 4. Peptide discrimination by identifying amino acid hydrolysates. (a) Schematic illustration of peptide hydrolysis using carboxypeptidase A1. Carboxypeptidase A1 releases amino acids (except R, K and P) from C-terminus of peptide. Then the released amino acids are detected and identified. (b) Current trace of real-time detection of hydrolysates from peptide EAFNL. (c) The distribution of relative abundance of the blockade of amino acid signals identified by machine learning algorithm from the hydrolysate of peptides. (d) The identified compositions of three peptides. Number in parentheses represents the count of recognized signals.

Note: Figure translations are in progress. See original paper for figures.