Machine learning aided top-down proteomics on a microfluidic platform

Yuewen Zhang[1,†,‡], Maya A. Wright[1,‡], Kadi L. Saar[1,2,‡], Pavankumar Challa[1,‡], Alexey S. Morgunov[1,3,‡], Quentin A. E. Peter[1], Sean Devenish[3], Christopher M. Dobson[1,¶], Tuomas P.J. Knowles[1,2,*]

[1]**Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK**
[2]**Cavendish Laboratory, Department of Physics, University of Cambridge, J J Thomson Ave, Cambridge CB3 0HE, UK**
[3]**Fluidic Analytics Ltd., Cambridge, UK**
[†] **Current address: Department of Chemistry, Lanzhou University, Lanzhou, Gansu, 730000, P. R. China**
[‡] **These authors contributed equally**
[¶] **Deceased September 8, 2019**

* tpjk2@cam.ac.uk

Abstract

The ability to determine the identity of specific proteins is a critical challenge in many areas of cellular and molecular biology, and in medical diagnostics. Here, we present a microfluidic protein characterisation strategy that within a few minutes generates a three-dimensional fingerprint of a protein sample indicative of its amino acid composition and size and, thereby, creates a unique signature for the protein. By acquiring such multidimensional fingerprints for a set of ten proteins and using machine learning approaches to classify the fingerprints, we demonstrate that this strategy allows proteins to be classified at a high accuracy, even though classification using a single dimension is not possible. Moreover, we show that the acquired fingerprints correlate with the amino acid content of the samples, which makes it is possible to identify proteins directly from their sequence without requiring any prior knowledge about the fingerprints. These findings suggest that such a multidimensional profiling strategy can lead to the development of novel method for protein identification in a microfluidic format.

Introduction

The diverse nature of proteins and their central role in a multitude of biological processes [1–3] necessitates a requirement for highly specific and sensitive approaches for protein detection and analysis. Indeed, protein detection and characterisation approaches have been of fundamental importance for a range of biological and medical research fields and have provided valuable information for better understanding the onset of a multitude of diseases, including various forms of cancer and neurodegenerative disorders. [4–10] In particular, at the centre of the discovery of novel protein-based disease biomarkers lies the ability to identify proteins. [11–16] In this context, protein microarrays are currently one of the most widely used techniques. By providing a high spatial density array of
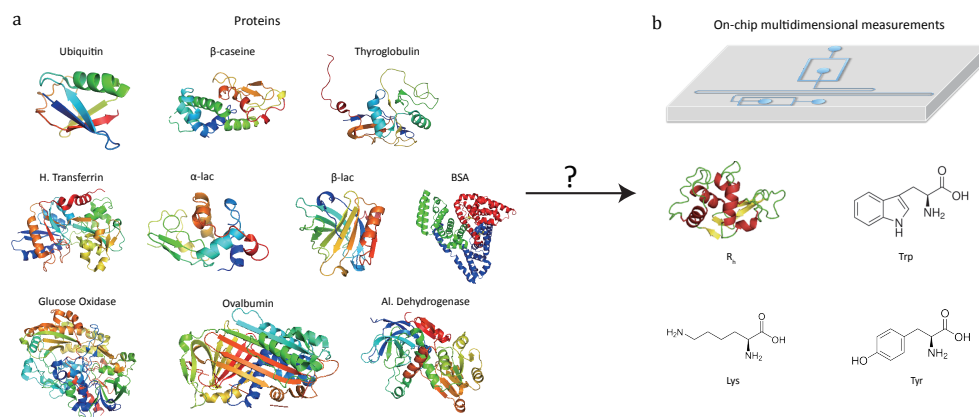
Figure 1. Microfluidic top down identification of proteins. **(a)** The proteins and **(b)** the microfluidic device used in this study. The device allows obtaining multi-dimensional fingerprints of protein samples that include information about their tryptophan, tyrosine and lysine content as well as the hydrodynamic radius, R_h.

solid-phase supported affinity reagents, such as antibodies, protein microarrays allow proteins of interest to be selectively captured and subsequently detected through the introduction of a second, frequently fluorescently labelled affinity reagent. [17,18] As such, protein microarray based approaches usually require access to two distinct antibodies each targeting a different epitope of a single protein and, similarly to any other affinity reagent mediated system, their performance is sensitive to potential undesired cross-reactivity events. On a fundamental level, such an affinity-reagent mediated strategy is inherently limited to detecting known targets for which a suitable affinity reagent was consciously included in the library and does not allow for the detection and the discovery of hitherto unknown markers.

The possibility to detect the presence of hitherto unknown targets and perform explorative screening arises when affinity-reagent free protein analysis approaches are used. In this context, various forms of mass-spectrometry have been widely used for protein identification for many decades due to their high sensitivity, resolution, accuracy and dynamic range. [19, 20] In a typical experiment, fragments of proteins are formed and separated through approaches, such as liquid chromatography before their injection to a mass-spectrometer. [21–24] While top-down identification has allowed characterising a number of different protein species, its application becomes challenging in the limit of high molecular weight and low solubility species. Due to these limitations, less than 10% of mammalian proteome can be accessed through these techniques. [25] For the analysis of higher molecular weight species, bottom-up sequencing approaches have been developed, which usually involve proteolysis of a complex mixture of proteins followed by a chromatographic separation of the peptides prior to their sequencing through tandem mass spectrometry (MS/MS). Whether the analysis is performed in a top-down or bottom-up manner, mass-spectrometry generally requires extensive sample preparation steps, often resulting in significant losses, and long experimental analysis time. Moreover, the presence of less abundant species is usually masked by more abundant ones, which prevents it effective use for detecting targets that are present at low concentrations, as is the case for biomarkers during the onset and early stages of diseases. Last but not least, its operation in gas-phase, has made it challenging to extend the analysis to protein complexes that are held together through transient interactions.

Recently, approaches that would enable overcoming the drawbacks of mass spectrom-

etry have been demonstrated and proposed. In particular, Swaminathan et al. [26] have demonstrated the possibility of immobilising peptides onto a glass slide and measuring their fluorescence through total internal reflection microscopy in consecutive cycles of Edman degradation after selectively labelling lysine and cystine residues. While demonstrating the first steps towards the feasibility of single molecule peptide fluorosequencing, [27] the approach involves a number of consecutive Edman steps, setting a limit on the speed at which the analysis can be performed.

To open up the possibility of minute-scale liquid-phase protein identification, here, we devised and demonstrated a microfluidic platform that permits the identification of protein samples on a single chip by relying on obtaining its characteristic multidimensional physicochemical signature. Specifically, by using a multi-wavelength detection system, we obtained readouts describing the tryptophan (Trp), tyrosine (Tyr) and lysine (Lys) content of the protein sample together with an estimate for their hydrodynamic radius (Figure 1). By obtaining such multidimensional signatures for a total of ten proteins and using machine learning approaches for identifying the origin of a set of validation proteins, we showed that such a strategy can be used for identifying proteins with a high confidence. The characterisation and identification process is performed on unlabelled protein samples and on a minute timescale.

Materials and Methods

Preparation of protein samples and the labelling solution

Bovine serum albumin (BSA), β-lactoglobulin (b-lac), glucose oxidase, α-lactalbumin (a-lac), ovalbumin, human transferrin, thyroglobulin (thglb), β-casein and ubiquitin (ubiq) were obtained from Sigma-Aldrich, and alcohol dehydrogenase (alc. dehydr) from Alfa Aesar. All the proteins were dissolved in 25 mM phosphate buffer at pH 8.0 to a micromolar concentration range. The solution used for labelling the lysine residues (Figure 2a) included 12 mM o-phthaldialdehyde (OPA), 18 mM β-mercaptoethanol (BME) and 4% wt/vol sodium dodecyl sulfate (SDS) in 200 mM carbonate buffer at pH 10.5.

UV-LED microscope

The schematic of the optical layout is shown in Figure 2b. The sample was excited using either a 280 nm LED (Thorlabs M280L3, UK) or a 365 nm LED (Thorlabs M365L2, UK) light source with a flip mirror used to switch between the two sources. The light from either of the LEDs was passed through an aspherical lens of focal length 20 mm to get a collimated output beam. The beam was passed through a dichroic filter cube, which consisted of an excitation filter (Semrock FF01-280/20-25) and a dichroic mirror (Semrock FF310-Di01-25x36). The light reflected by the dichroic mirror was then focussed onto the sample flowing in a microfluidic chip by an infinity corrected UV objective lens (Thorlabs LMU-10X-UVB, UK) of numerical aperture NA = 0.25. The emitted fluorescent light from the sample was collected through the same objective and an emission filter (Semrock FF01-357/44-25 for a charactersitic tryptophan, FF01-302/10-25 for a charactersitic tyrosine and FF01-452/45-25 for a charactersitic lysine signal) with an air-spaced achromatic doublet lens of focal length 20 mm (Thorlabs ACA254-200-UV) focussing it onto the camera (Rolera EMC2). All the optics used in the set-up were made out of fused silica for high transmission in the UV region. [28, 29]

Microfluidic device fabrication

The microfluidic devices were cast using polydimethylsiloxane (PDMS) (Sylgard 184 kit, Dow Corning, USA) from a silicon wafer master imprinted with 50 μm high structures based using standard singlelayer soft-lithography techniques. [30] The precise height of the photoresist structures on different locations across the master mould were measured by a profilometer (DektakXT, UK) to correct the analysis for any variations in structure height across the master. Carbon black nanopowder (Sigma-Aldrich, UK) was added to the PMDS to minimise undesired autofluorescence from the PDMS devices under UV illumination during the measurements. The devices were bonded to a quartz slide (Alfa Aesar, 76.2x25.4x1.0 mm, UK) using plasma treatment (Electronic Diener Femto plasma bonder; 15 seconds at 40% of the full power). The PDMS-glass microfluidic devices were then exposed to an additional extended plasma treatment step (500 seconds at 80% of the full power) to render channel surfaces more hydrophilic with the inlets and outlets blocked with water-filled gel-loading tips immediately after the exposure to maintain their hydrophilic character.

Device operation

To obtain a multidimensional signature for a sample, the channels of the microfluidic were first filled from the common outlet using a glass syringe (Hamilton, 500 μL, UK), 27 gauge needle (Neolus Terumo, 25 gauge, 0.5 x 16 mm, UK), and polyethene tubing (Scientific Laboratory Supplies, inner diameter 0.38 mm, outer diameter 1.09 mm, UK). Gel loading tips filled with the relevant solutions were then inserted into the device inlets (Figure 2a). The fluid flow through of the solutions into the microfluidic channels was controlled using neMESYS syringe pumps (Cetoni GmbH, Germany) that was set to withdraw the solutions at a total flow rate of 200 μL h^{-1}. As described previously, [29] in order to increase the accuracy of the diffusional sizing process, the gel loading tip in the sample inlet was first filled with the auxiliary buffer and a background image of the diffusional sizing area recorded. This micrograph was later used for subtracting the static background arising from the autofluorescence of the PDMS device. The gel loading tip in the sample inlet was then carefully exchanged to a tip including the protein sample with care taken not to introduce any air bubbles in the process. For both images, an exposure time of 500 ms was used.

Finally, in order to account for any potential fluctuations in the power output of the LEDs, the intensities of standard calibration solutions (10 μM L-Tryptophan and 10 μM 4-methylumbelliferone both in 400 mM potassium borate buffer at pH 9.7) were recorded in a channel adjacent to the top-down identification device itself. The obtained characteristic tryptophan and tyrosine fluorescence values were then normalised by the former of this calibration readings and the lysine value by the latter of the two calibration readings.

Results and Discussion

Microfluidic multidimensional protein characterisation strategy

To facilitate the acquisition of multidimensional physicochemical signatures of proteins directly in solution, we designed a microfluidic device that allowed simultaneously obtaining four characteristic parameters of an unlabelled protein sample. Specifically, after introducing a sample from its corresponding inlet (Figure 2a), first, the characteristic fluorescence intensities indicative of the tryptophan and tyrosine content of the sample
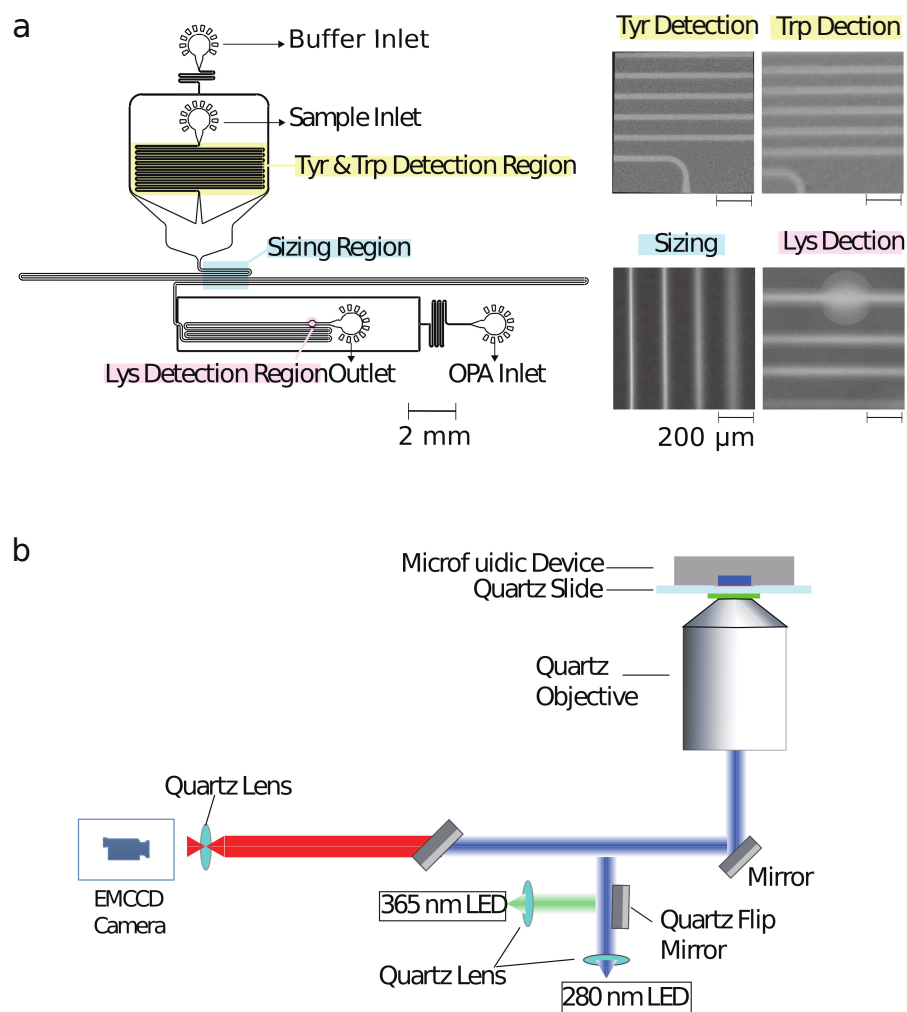
Figure 2. Microfluidic top-down identification strategy. **(a)** The microfluidic device used in this study allowed extracting a multidimensional characteristic signature of each analysed sample describing its (i) tryptophan (Trp) and (ii) tyrosine (Tyr) residues (yellow highlighted region), (iii) its hydrodynamic radius R_h obtained by monitoring the diffusion of the sample molecules into a co-flowing buffer (blue highlighted region) and (iv) its lysine (Lys) content (pink highlighted region). The scale bars on all insets are 200 μm. **(b)** Schematic representation of the home-built inverted fluorescence microscope used. The two light sources (280 nm and 365 nm) and emission filters can be switched readily to record the characteristic fluorescent signals.

were recorded in the yellow highlighted area by exciting the microfluidic chip with a 132
UV wavelength (280 nm) LED (Figure 2b) and collecting the emitted fluorescent light 133
using two distinct filters. The filters were chosen such that the collected light originated 134
either predominantly from its tryptophan or from its tyrosine residues (Materials and 135
Methods). 136

The protein sample was then surrounded by a co-flowing buffer in order to monitor 137
the lateral diffusion of the protein sample into an auxiliary carrier medium in space and 138
in time. Such a strategy has been previously shown to yield the diffusion coefficients 139
of protein samples. [31] In particular, the device we used in this study was designed 140
for the camera field of view (800 μm $\times$ 1000 μm) to include four distinct sections of 141
this channel (blue highlighted region), so that a single image could be used to extract 142
the diffusion coefficient as described earlier. [32] The channels were imaged using the 143
280 nm excitation LED in combination with the tryptophan filter as the signal from 144
latter residue was stronger than the signal from tyrosine residues. The diffusion profiles 145
on the micrographs were then fitted to simulated basis functions for particles of known 146
radii and each of the simulated profiles were compared to the measured profiles in order 147
to extract the hydrodynamic radius of each sample. [31–34] 148

Finally, downstream the sizing unit, an on-chip latent labelling strategy was used 149
to conjugate the lysine residues in each protein to o-phthaldialdehyde (OPA) dye 150
molecules [33, 35] (Materials and Methods). The characteristic fluorescence intensity 151
from the OPA labelled lysine residues was measured (pink highlighted region) by switching 152
the UV-LED light source to an LED light source with excitation at 365 nm wavelength 153
(Materials and Methods) at which unconjugated OPA molecules have been observed to 154
show only minimal background fluorescence. The dimensions of the labelling channel 155
were chosen such that the OPA dye and the protein sample would be able to mix for 156
over a 3 second long time period before the measurement was taken, a time scale that we 157
had previously shown allows quantitative insight into the abundance of lysine residues 158
in proteins. [33] 159

All in all, this strategy allowed us to obtain a four-dimensional signature for each 160
protein sample using a single microfluidic chip and a dual-wavelength excitation system. 161
One of the four parameters was later used for normalising the obtained fluorescent signals 162
to obtain a concentration independent signature for each protein. 163

Multidimensional signatures of a set of ten proteins 164

We analysed a set of ten different proteins (Figure 1a) and used the platform described 165
above to obtain multidimensional signatures for each of them. In particular, we performed 166
n = 4 repeats on all the ten proteins using a different microfluidic device for each 167
experiment. We noted that the measured R_h values of all the proteins, which varied by 168
three orders of magnitude in molecular weight, were consistent with the values reported 169
in the literature (Supplementary Table S1). 170

In order to eliminate concentration dependence, the obtained signals in the tryptophan 171
and tyrosine imaging channels were normalised by the signal in the lysine filter. This 172
reduced the data structure to a three-dimensional signature but ensured that the obtained 173
values were independent of the concentration of the protein that was used for analysis. 174
Moreover, the measured intensities were corrected for fluctuations in the laser power 175
by also measuring the fluorescence intensities of calibration solutions in a neighbouring 176
channel, involving L-tryptophan and 4-methylumbelliferone molecules for the 280 nm 177
and the 365 nm LED, respectively (Materials and Methods). 178

The characteristic spaces that each of the analysed ten proteins occupied in a three- 179
dimensional plot are shown in Figure 3d with the 1D projections shown in Figures 3a-c 180
and the underlying data summarised in Supplementary Table S1. In particular, the 181

three-dimensional visualised ellipsoids (Figure 3d) were defined by the centres being the ₁₈₂
average of the four measurement points and their radii corresponding to the standard ₁₈₃
deviation of the four measurements. We noted that the ten analysed proteins varied in ₁₈₄
their physiochemical signatures with Figure 3d illustrating that it is likely that across a ₁₈₅
three-dimensional landscape each of the protein acquires a different signature. ₁₈₆

Protein classification ₁₈₇

In order to evaluate whether our demonstrated platform is capable of distinguishing ₁₈₈
proteins reliably and uniquely based on their signatures, we developed two models to ₁₈₉
perform sample classification. ₁₉₀

First, using the full data set of 10 classes of proteins with 4 experimental repeats ₁₉₁
for each class, leave-one-out cross-validation was used to assess the likelihood that a ₁₉₂
particular sample is classified as the correct protein. In particular, multivariate Gaussian ₁₉₃
distributions were fitted to each of the ten protein classes with the means computed ₁₉₄
from the four repeats within each class, or from the three remaining repeats for the ₁₉₅
class from which the validation sample was removed. The covariance matrices were ₁₉₆
computed by combining the group variance (using either four or three repeats similarly ₁₉₇
to the means) with the global variance involving the full dataset of 39 data points ₁₉₈
excluding the validation sample. A weighting factor of 0.9 was used for the group ₁₉₉
variance and a weighting factor of 0.1 for the global variance to avoid singular covariance ₂₀₀
matrices and ensure computational stability while simultaneously taking advantage ₂₀₁
of the extra information about the system as the variances in the same dimension ₂₀₂
between the different classes are likely to be similar. Finally, the likelihood of each of the ₂₀₃
validation samples belonging to each of the protein classes was calculated by estimating ₂₀₄
the probability density function of the individual multivariate Gaussians at that point. ₂₀₅

For each protein class, the likelihood was averaged across the four experimental ₂₀₆
repeats and the resulting values were normalised to one. Figure 3e shows a heatmap of ₂₀₇
the calculated likelihoods for assigning proteins into available classes with the actual ₂₀₈
protein being measured on the vertical axis and the protein it is likely to be identified ₂₀₉
as on the horizontal axis. We observed that, individually, 33 out of 40 samples were ₂₁₀
classified correctly. Moreover, it can be seen that on average proteins are likely to be ₂₁₁
assigned to the correct class with high confidence. ₂₁₂

We note that the above estimates were arrived at by assuming that the errors in the ₂₁₃
measurements in each dimension were normally distributed, so that the protein classes ₂₁₄
can be approximated by multivariate Gaussian distributions. In order to improve our ₂₁₅
analysis and derive an identification strategy that is not making an assumption about the ₂₁₆
distribution of the errors, we constructed a random forest classifier. As before, leave-one- ₂₁₇
out cross-validation was used on all 40 samples. In order to reduce variance, each random ₂₁₈
forest was trained with 1000 decision trees that were built using bootstrapping and with ₂₁₉
only 2 out of 3 variables selected at random to build each tree. The classification was ₂₂₀
performed using predictions by these ensemble models and, subsequently, predictions ₂₂₁
by all individual trees in the ensembles were collected to quantify the confidence of the ₂₂₂
ensemble model in making the predictions. For each group of four samples corresponding ₂₂₃
to the same protein class, the average number of trees in the ensemble predicting each ₂₂₄
target class were taken and normalised to sum to one for each protein. ₂₂₅

Finally, a heatmap summarising the results was constructed, similarly showing the ₂₂₆
actual protein being measured on the vertical axis and the protein it is likely to be ₂₂₇
identified as on the horizontal axis (Figure 3f). The results illustrate that the model ₂₂₈
predicts the correct class of proteins with high confidence. Moreover, on the individual ₂₂₉
level, the random forest model misclassified only 4 out of 40 samples, demonstrating ₂₃₀
a superior performance to the multivariate Gaussian model. This shows that highly ₂₃₁
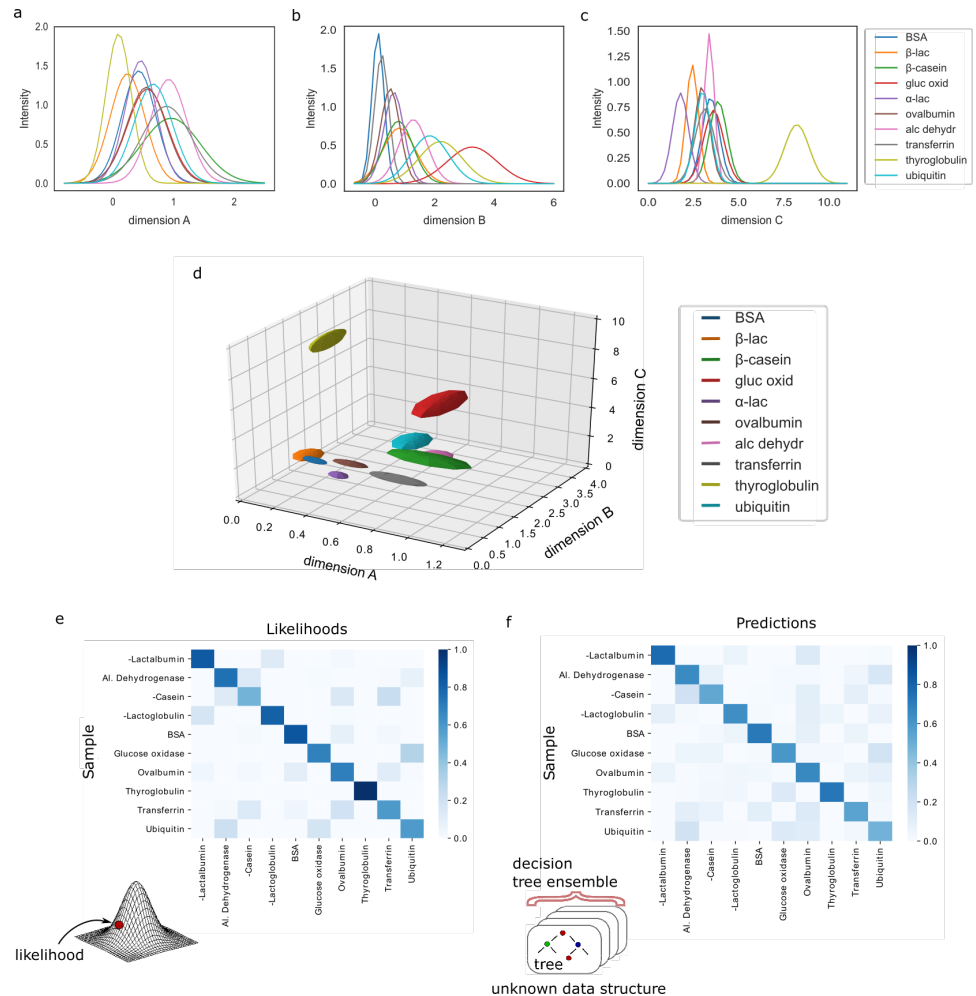
Figure 3. Protein classification from their multidimensional fingerprints. A set of ten proteins was profiled by acquiring their three-dimensional fingerprints described by **(a)** the ratio of the signals measured at the wavelengths where tyrosine and OPA fluoresce (Materials and Methods; dimension 1), **(b)** the ratio of the signals measured at the wavelengths where tryptophan and OPA fluoresce (dimension 2) and **(c)** the hydrodynamic radius, R_h (dimension 3). All these parameters are concentration independent. **(d)** Multidimensional signatures of the proteins in a 3D space. The radii of the ellipsoids correspond to one standard deviation. **(e)** The likelihoods of protein identification and misidentification in the 3D space showed in panel **(d)** assuming multivariate Gaussian model. **(f)** The confidence levels of identification process using a random forest classifier approach that assumes no underlying data distribution. The models identified correctly 82% (multivariate Gaussian) and 90% (random forest classifier) of the tested samples.

accurate identification of proteins is possible even when no assumptions are made about   232
the underlying distributions of measurement errors or data structure.   233

Protein identification   234

Having confirmed the possibility to classify an unknown protein sample by evaluating   235
which of the pre-determined multidimensional fingerprints it resembles the most (Fig-   236
ure 3), we next set out to explore if it is possible to determine the origin of each of   237
the test samples simply by performing a top-down identification process on the sample   238
without requiring prior knowledge of the fingerprints.   239

To this effect, we first derived relationships that could be used to predict the sequence   240
composition of each sample from its fingerprint. Indeed, the measured fluorescent   241
signals were constructed in a manner where they can be expected to predominantly   242
originate from the tyrosine, tryptophan and lysine residues of the proteins (Materials   243
and Methods) or, in the case of the hydrodynamic radius, be linked to its molecular   244
weight. The observed correlation between the measured fluorescent signals and the   245
amino acid content of the proteins are shown in Figures 4a-b. As before, in order to   246
eliminate any concentration dependence, ratios between the measured signals and amino   247
acid compositions were used. Figure 4c additionally outlines the relationship between the   248
measured hydrodynamic radius and the molecular weight of the proteins. We modelled   249
all the three relationships as linear regressions with zero-intercepts and estimated the   250
gradient of the line by minimising the ordinary least squares (Figure4a-c, dotted lines).   251
It is possible that when more abundant data is available, more nuanced relationships   252
between the measured signals and the sequence-specific quantities can be learned.   253

Next, the derived relationships (Figure 4a-c) were used to convert the measured   254
three-dimensional signature of our test samples into their predicted $\frac{\text{Trp}}{\text{Lys}}$ and $\frac{\text{Tyr}}{\text{Lys}}$ ratios   255
and molecular weights. To eliminate any information leakage, we re-fitted the linear   256
regression by excluding the test sample. Following this step, the z-score of the measured   257
sample being a particular protein in the database was calculated by using the estimated   258
sequence-specific properties of this sample as the mean value and the measurement noise   259
as the standard deviation. The heatmap describing the most likely origin of each of   260
the test samples is shown in Figure 4d with the data arranged such that the correct   261
sample appeared on the diagonal of the matrix. These data show that, on average, seven   262
out the ten proteins were identified correctly. On the level of individual measurements,   263
samples were identified correctly in 21 out of 40 experiments (Figure 5). These results   264
illustrate that not only can the multidimensional signatures used for classifying proteins   265
into pre-determined clusters (Figure 3e-f), it is also possible to convert the measured   266
signals into absolute sequence-specific parameters and through this process identify the   267
test samples.   268

Conclusions   269

We developed a strategy for obtaining multidimensional physicochemical signatures of   270
individual proteins on a single microfluidic chip and showed that this strategy can be   271
used for protein classification as well as top-down identification. Specifically, we achieved   272
this objective by designing a chip on which the hydrodynamic radius of a protein sample   273
could be obtained simultaneously with signals describing its tryptophan, tyrosine and   274
lysine content. We showed that this approach generated unique fingerprints for all the   275
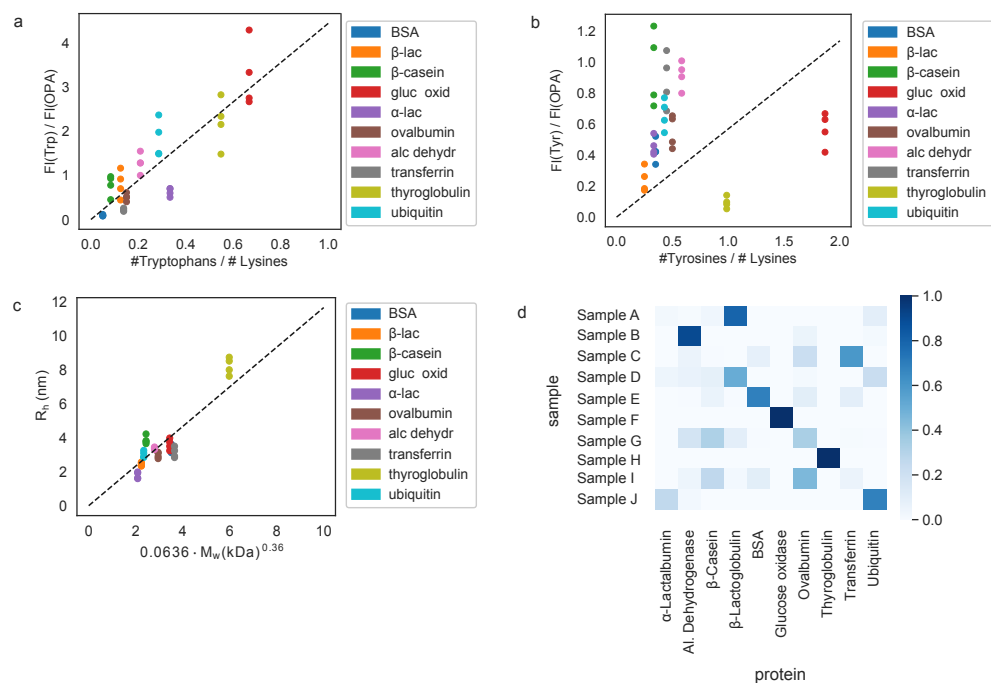ten proteins in our test set and, moreover, that the signatures can be used to identify   276

Figure 4. Protein identification from their sequence. The correlations between the measured signals and the sequences of the analysed proteins, specifically **(a)** the ratio of the measured tryptophan and OPA signals as a function of the tryptophan and lysine composition of the proteins, **(b)** the ratio of the measured tyrosine and OPA fluorescence signals as a function of their tyrosine and lysine composition and **(c)** the measured hydrodynamic radius, R_h, as a function of the molecular weight. In all cases, the dotted line shows the best fit linear regression function with the intercept set to 0. **(d)** The measured signals for each of the ten samples (A–J) were converted to estimates of their sequence-composition using the relationships outlined in panels **(a)**-**(c)** and the latter estimates were used to evaluate the probabilities of each of the ten samples being any one of the ten proteins in our dataset by using Gaussian mixture models. The data are shown such that the correct sample appears on the diagonal of the matrix. Individual samples were identified correctly on 21 out of 40 occasions. When averaging the results over $n = 4$ repeats, seven out of ten proteins were identified correctly.
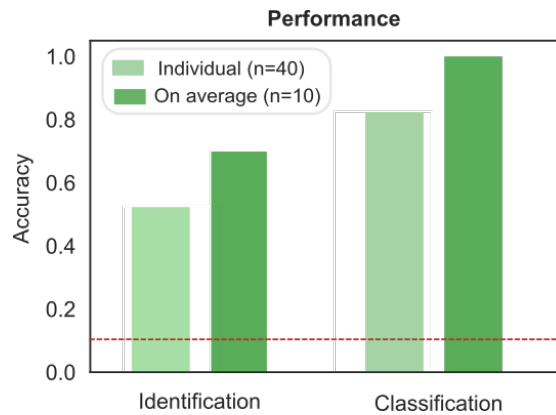
Figure 5. Comparison of the performance of the protein classification (Figure 3) and identification (Figure 4) strategies. When identifying a measured sample directly from its sequence, samples were identified correctly on 53% of the occasions or on 70% of the occasions when the results were averaged across the four repeats performed on each samples. When pre-determined fingerprints were used, proteins were classified correctly on 83% of the occasions or on 100% of the occasion when the results were averaged across the repeats. The red dotted line corresponds to the case where the classification or identification was performed by a process of random guessing.

proteins in a top-down manner. Our results suggest that an on-chip multidimensional protein characterisation strategy could serve as a powerful probe-free approach for on-chip biomarker profiling using only microlitre sized samples.

Author Contributions

Y.Z., M.A.W., K.L.S., P.C. and T.P.J.K. designed the study and conceived the experiments, P.C. built the experimental hardware, Y.Z. and M.A.W. acquired the data and processed the micrographs for which Q.A.E.P. contributed relevant software. K.L.S. and A.S.M. analysed and interpreted the data and built the machine learning models. Y.Z., K.L.S., A.S.M. and T.P.J.K. wrote the original manuscript and M.A.W., P.C., S.D., Q.A.E.P. and C.M.D. reviewed it.

Conflicts of interest

Parts of this work have been the subject of a patent application filed by Cambridge Enterprise Limited, a fully owned subsidiary of the University of Cambridge.

Acknowledgements

References

1. Alberts, B. *et al.* Molecular biology of the cell (2002).

2. Berg, J. M., Tymoczko, J. L. & Stryer, L. Biochemistry (2002).

3. Alberts, B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **92**, 291–294 (1998).

4. Collins, F. S., Green, E. D., Guttmacher, A. E. & Guyer, M. S. A vision for the future of genomics research. *Nature* **422**, 835 (2003).

5. Coscia, F. *et al.* Integrative proteomic profiling of ovarian cancer cell lines reveals precursor cell associated proteins and functional status. *Nature Communications* **7**, 12645 (2016).

6. Genovese, F. *et al.* Mass spectrometric/bioinformatic identification of a protein subset that characterizes the cellular activity of anticancer peptides. *Journal of proteome research* **13**, 5250–5261 (2014).

7. Kang, S. *et al.* In situ identification and localization of igha2 in the breast tumor microenvironment by mass spectrometry. *Journal of proteome research* **11**, 4567–4574 (2012).

8. Oláh, Z. *et al.* Proteomic analysis of cerebrospinal fluid in alzheimer's disease: wanted dead or alive. *Journal of Alzheimer's Disease* **44**, 1303–1312 (2015).

9. Puchades, M. *et al.* Proteomic studies of potential cerebrospinal fluid protein markers for alzheimer's disease. *Molecular Brain Research* **118**, 140–146 (2003).

10. Sultana, R. *et al.* Proteomics analysis of the alzheimer's disease hippocampal proteome. *Journal of Alzheimer's Disease* **11**, 153–164 (2007).

11. Savaryn, J. P., Catherman, A. D., Thomas, P. M., Abecassis, M. M. & Kelleher, N. L. The emergence of top-down proteomics in clinical research. *Genome Medicine* **5**, 53 (2013).

12. Chen, R. & Snyder, M. Promise of personalized omics to precision medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **5**, 73–82 (2013).

13. Petricoin, E., Wulfkuhle, J., Espina, V. & Liotta, L. A. Clinical proteomics: revolutionizing disease detection and patient tailoring therapy. *Journal of proteome research* **3**, 209–217 (2004).

14. Zhu, H. *et al.* Global analysis of protein activities using proteome chips. *Science* **293**, 2101–2105 (2001).

15. Liotta, L. A., Kohn, E. C. & Petricoin, E. F. Clinical proteomics: personalized molecular medicine. *JAMA: The Journal of the American Medical Association* **286**, 2211–2214 (2001).

16. Zhu, H. *et al.* Analysis of yeast protein kinases using protein chips. *Nature Genetics* **26**, 283 (2000).

17. MacBeath, G. Protein microarrays and proteomics. *Nature Genetics* **32**, 526 (2002).

18. LaBaer, J. & Ramachandran, N. Protein microarrays as tools for functional proteomics. *Current Opinion in Chemical Biology* **9**, 14–19 (2005).

19. Marshall, A. G., Hendrickson, C. L. & Jackson, G. S. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrometry Reviews* **17**, 1–35 (1998).

20. Valaskovic, G. A., Kelleher, N. L. & McLafferty, F. W. Attomole protein characterization by capillary electrophoresis-mass spectrometry. *Science* **273**, 1199–1202 (1996).

21. Mann, M., Hendrickson, R. C. & Pandey, A. Analysis of proteins and proteomes by mass spectrometry. *Annual Review of Biochemistry* **70**, 437–473 (2001).

22. Frank, A. M., Savitski, M. M., Nielsen, M. L., Zubarev, R. A. & Pevzner, P. A. De novo peptide sequencing and identification with precision mass spectrometry. *Journal of Proteome Research* **6**, 114–123 (2007).

23. Vyatkina, K. *et al.* De novo sequencing of peptides from top-down tandem mass spectra. *Journal of Proteome Research* **14**, 4450–4462 (2015).

24. Reid, G. E. & McLuckey, S. A. Top down protein characterization via tandem mass spectrometry. *Journal of Mass Spectrometry* **37**, 663–675 (2002).

25. Resing, K. A. & Ahn, N. G. Proteomics strategies for protein identification. *FEBS letters* **579**, 885–889 (2005).

26. Swaminathan, J. *et al.* Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. *Nature biotechnology* **36**, 1076 (2018).

27. Collins, B. C. & Aebersold, R. Proteomics goes parallel. *Nature biotechnology* **36**, 1051 (2018).

28. Saar, K. L. *et al.* On-chip label-free protein analysis with downstream electrodes for direct removal of electrolysis products. *Lab on a Chip* **18**, 162–170 (2018).

29. Challa, P. K. *et al.* Real-time intrinsic fluorescence visualization and sizing of proteins and protein complexes in microfluidic devices. *Analytical Chemistry* **90**, 3849–3855 (2018).

30. Duffy, D. C., McDonald, J. C., Schueller, O. J. & Whitesides, G. M. Rapid prototyping of microfluidic systems in poly (dimethylsiloxane). *Analytical Chemistry* **70**, 4974–4984 (1998).

31. Arosio, P. *et al.* Microfluidic diffusion analysis of the sizes and interactions of proteins under native solution conditions. *ACS Nano* **10**, 333–341 (2015).

32. Saar, K. L. *et al.* Rapid two-dimensional characterisation of proteins in solution. *Microsystems & Nanoengineering* **5**, 33 (2019).

33. Yates, E. V. *et al.* Latent analysis of unmodified biomolecules and their complexes in solution with attomole detection sensitivity. *Nature Chemistry* **7**, 802 (2015).

34. Zhang, Y. *et al.* A microfluidic strategy for the detection of membrane protein interactions. *Lab on a Chip* **20**, 3230–3238 (2020).

35. Zhang, Y. *et al.* On-chip measurements of protein unfolding from direct observations of micron-scale diffusion. *Chemical Science* **9**, 3503–3507 (2018).