ARTICLES
PUBLISHED ONLINE: 6 APRIL 2014 | DOI: 10.1038/NNANO.2014.54
nature nanotechnology

# Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling

Yanan Zhao[1,2†], Brian Ashcroft[2†], Peiming Zhang[2], Hao Liu[2,3], Suman Sen[2,3], Weisi Song[2,3], JongOne Im[1,2], Brett Gyarfas[2], Saikat Manna[2,3], Sovan Biswas[2,3], Chad Borges[2,3] and Stuart Lindsay[1,2,3*]

**The human proteome has millions of protein variants due to alternative RNA splicing and post-translational modifications, and variants that are related to diseases are frequently present in minute concentrations. For DNA and RNA, low concentrations can be amplified using the polymerase chain reaction, but there is no such reaction for proteins. Therefore, the development of single-molecule protein sequencing is a critical step in the search for protein biomarkers. Here, we show that single amino acids can be identified by trapping the molecules between two electrodes that are coated with a layer of recognition molecules, then measuring the electron tunnelling current across the junction. A given molecule can bind in more than one way in the junction, and we therefore use a machine-learning algorithm to distinguish between the sets of electronic 'fingerprints' associated with each binding motif. With this recognition tunnelling technique, we are able to identify D and L enantiomers, a methylated amino acid, isobaric isomers and short peptides. The results suggest that direct electronic sequencing of single proteins could be possible by sequentially measuring the products of processive exopeptidase digestion, or by using a molecular motor to pull proteins through a tunnel junction integrated with a nanopore.**

The proteome is probably a much better molecular indicator of the current health status of humans than the genome[1], but proteomic data are harder to acquire[2]. Protein sequences deduced from cDNA lack information about alternative splicing and post-translational modifications. Low concentrations of DNA and RNA are readily amplified by the polymerase chain reaction (PCR), but there is no similar technique available for proteins. Thus, there may be many rare protein variants yet to be discovered at concentrations that are well below the detection limits of current techniques[3]. In view of this, a single-molecule technique for protein sequencing is critical for the identification of biomarkers and to enable the real-time diagnostic possibilities that follow. We are currently developing recognition tunnelling (RT) as an electronic single-molecule sequencing method for DNA. Here, we show that the method also works to identify individual amino acids and peptides, and so may open the way to single-molecule protein sequencing.

## Recognition tunnelling

In recognition tunnelling (Fig. 1a), two metal electrodes, separated by a gap of ~2 nm, are covered with a layer of recognition molecules that are strongly bonded to the electrodes. The recognition molecules form weaker, non-covalent contacts with target analyte molecules. Single-molecule signals dominate when a sharp electrode is used because the signal from the shortest path is by far the largest. When a small bias (<1 V) is applied across the electrode gap, molecules captured by these non-covalent contacts produce a stochastic train of current spikes (pA–nA) at kilohertz rates[4–7], the trapped molecule remaining bound for about a second (as determined by dynamic force spectroscopy measurements[4]). Such weakly bonded complexes can remain intact for long times because of their confinement[8]. Thermal vibrations of the molecule generate current spikes (Fig. 1c,d), the distribution of which is characteristic of the bonding in the tunnel junction (Fig. 1e,f). Much as in classical spectroscopy, the temporal, spectral and amplitude information contained within a signal train ('signal features' in machine-learning terminology) can be used as an electronic 'fingerprint' with which to identify each molecule that enters the RT junction. The electronic 'fingerprints' are decoded[9] with high accuracy by a machine-learning algorithm (the 'support vector machine', SVM[10]).

## Tunnelling measurements

We used a scanning tunnelling microscope (STM), operated in buffered aqueous solution, to create a tunnel gap set to a reproducible distance by controlling the gap conductance, collecting useful signals up to 25 kHz in frequency. Palladium probes, partially insulated with polyethylene, and Pd substrates[11], functionalized with the recognition molecule, 4(5)-(2-mercaptoethyl)-1*H*-imidazole-2-carboxamide (ICA), were used as electrodes[12]. We found that a tunnel current of 4 pA at a bias of 0.5 V produced RT signals from all but two of the 20 naturally occurring amino acids (Supplementary Figs 2 and 3), while phosphate-buffer controls were almost free of signals (Supplementary Fig. 2a). This result is surprising, because ICA molecules were designed to interact with DNA bases[13]. Nonetheless, electrospray ionization mass spectrometry (ESIMS)[14,15] clearly shows the presence of 2:1 adducts of ICA molecules with all of the seven amino acids analysed in the present article (Fig. 1b, Supplementary Figs 11, 12, Supplementary Tables 6, 7).

## Identifying amino acids

We demonstrate the power of RT with three applications: distinguishing a modified amino acid, sarcosine (or *N*-methylglycine (mGly), a potential cancer marker[16]), from glycine (Gly), two enantiomers (L- and D-asparagine, Asn) and two isobaric amino acids

[1]Department of Physics, Arizona State University, PO Box 871504 Tempe, Arizona 85287, USA, [2]Biodesign Institute, Arizona State University, PO Box 875001, Tempe, Arizona 85287, USA, [3]Department of Chemistry and Biochemistry, Arizona State University, PO Box 871604, Tempe, Arizona 85287, USA, †These authors contributed equally to this work. *e-mail: stuart.lindsay@asu.edu



© 2014 Macmillan Publishers Limited. All rights reserved.

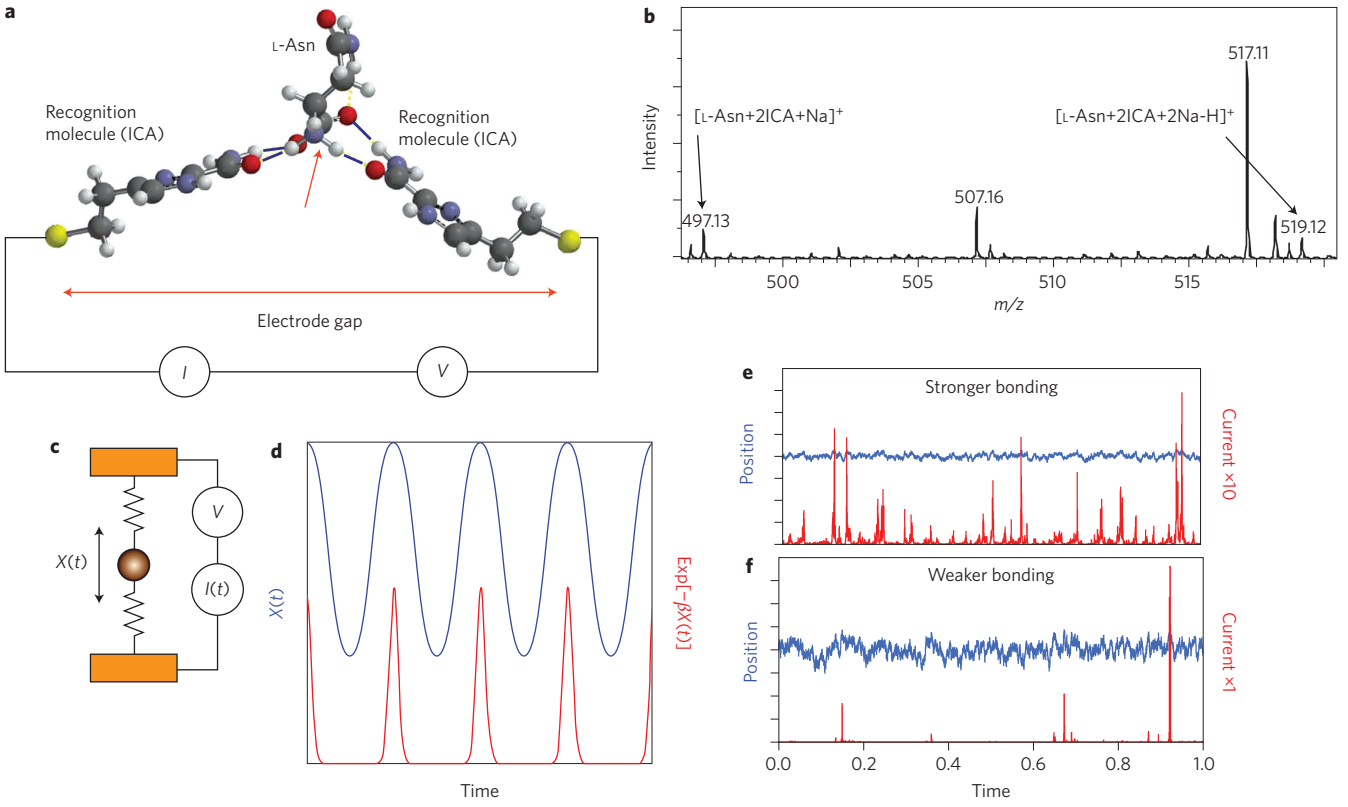


**Figure 1 | Recognition tunnelling (RT). a**, Recognition molecules (1*H*-imidazole-2-carboxamide, ICA) are strongly attached to a pair of closely spaced electrodes, displacing contamination and forming a chemically well-defined surface. An analyte (here shown as L-Asn) is captured by non-covalent interactions (blue bars show hydrogen bonds) with the recognition molecules. The bonding pattern is specific to the analyte. The red arrow shows the orientation of the molecular dipole for L-Asn. This orientation is different when D-Asn is captured (Supplementary Fig. 1). **b**, ESIMS shows that stoichiometric adducts form between reader molecules, here illustrated for 2:1 complexes of ICA and L-Asn. Data for other analytes are given in Supplementary Tables 6 and 7. **c**, Generation of RT signals. Picturing the analyte as a mass (sphere) trapped by a pair of springs that represent the non-covalent bonds, the extent of analyte motion, $X(t)$, depends on the strength of the springs. **d**, A simple sinusoidal motion of the analyte (blue trace) produces a series of sharp current spikes (red trace) because of the exponential dependence of tunnel current on position. **e**,**f**, Simulations for random thermal excitation of a strongly (**e**) and more weakly (**f**) bonded analyte, showing how the current fluctuations are much bigger when the bonding is weaker (red traces). The blue traces show the random thermal fluctuations in position of the analyte. The simulations are carried out as described by Huang and colleagues[4].

(leucine (Leu) and isoleucine (Ile)). In addition, we examined a pool of data from seven different analytes to evaluate how well any one amino acid can be identified. This is an essential first step in developing a sequence-reading technique.

In a typical RT experiment, a solution of amino acid was added into the STM liquid cell after the tunnel junction had stabilized in buffered solution for ~2 h. For each analyte, a minimum of three (usually four) separate experiments were run with freshly made probes, substrates and samples. Figure 2 presents representative signal trains. The spike shape carries significant information, and the insets show this in expanded traces. Signals occur in clusters and a computer algorithm (Fig. 2i) was developed to identify clustered data automatically. Clusters appear to correspond to single-molecule binding events for the following reasons. First, the duration of each cluster is on the order of 0.2 s (Supplementary Fig. 10b), comparable to the time for which hydrogen-bonded complexes remain bound in a nanogap[1,8,17]. Second, signals within clusters are much more strongly correlated than signals from different clusters (Supplementary Fig. 15). Finally, in signals obtained from mixed samples (Fig. 5), each cluster gave signals from only one analyte or the other.

We illustrate how signal features can distinguish pairs of analytes using data obtained from mGly and Leu (Fig. 3). Tunnel-current amplitude distributions[5,18,19] are largely overlapped (Fig. 3a). Features associated with pulse shapes (Fig. 3b,c) suffer less overlap, although the overlap still limits the accuracy of calling single-molecule events to ~70% (50% represents random calls). However, when these two signal features are used together to generate a two-dimensional map of probability densities (Fig. 3d), only a small fraction of the data are overlapped (yellow, near the origin) leading to a 95% calling accuracy if signals in the red area are assigned to mGly and in the green area to Leu. This is an illustration of Cover's theorem, which states that separability in pattern recognition increases in higher dimensions[20]. We used the SVM[9,21], a machine-learning algorithm, to discover these relationships by training on a subset of the data. The SVM assigns data between pairs of classes by partitioning the feature space into two regions. For the data shown in Fig. 3d, this partition is the curve that best separates green from red regions. In this case, SVM analysis yields an accuracy of less than 95% because a partition of the space cannot include all the 'data islands'. However, SVM analysis can be extremely accurate when a large number of features are used.

The effect of combining data from two or more parameters is even more dramatic in the case of chemically similar pairs of analytes (Fig. 4). The six feature distributions shown in Fig. 4a,b,d,e,g,h are quite overlapped for each pair of analytes (the probability of a correct identification among each pair is marked on each plot as $P$ values and values are typically only a little over 50%). Using the probability densities plotted as a function of the values of pairs of features (Fig. 4 c,f,i) increases the identification



© 2014 Macmillan Publishers Limited. All rights reserved.

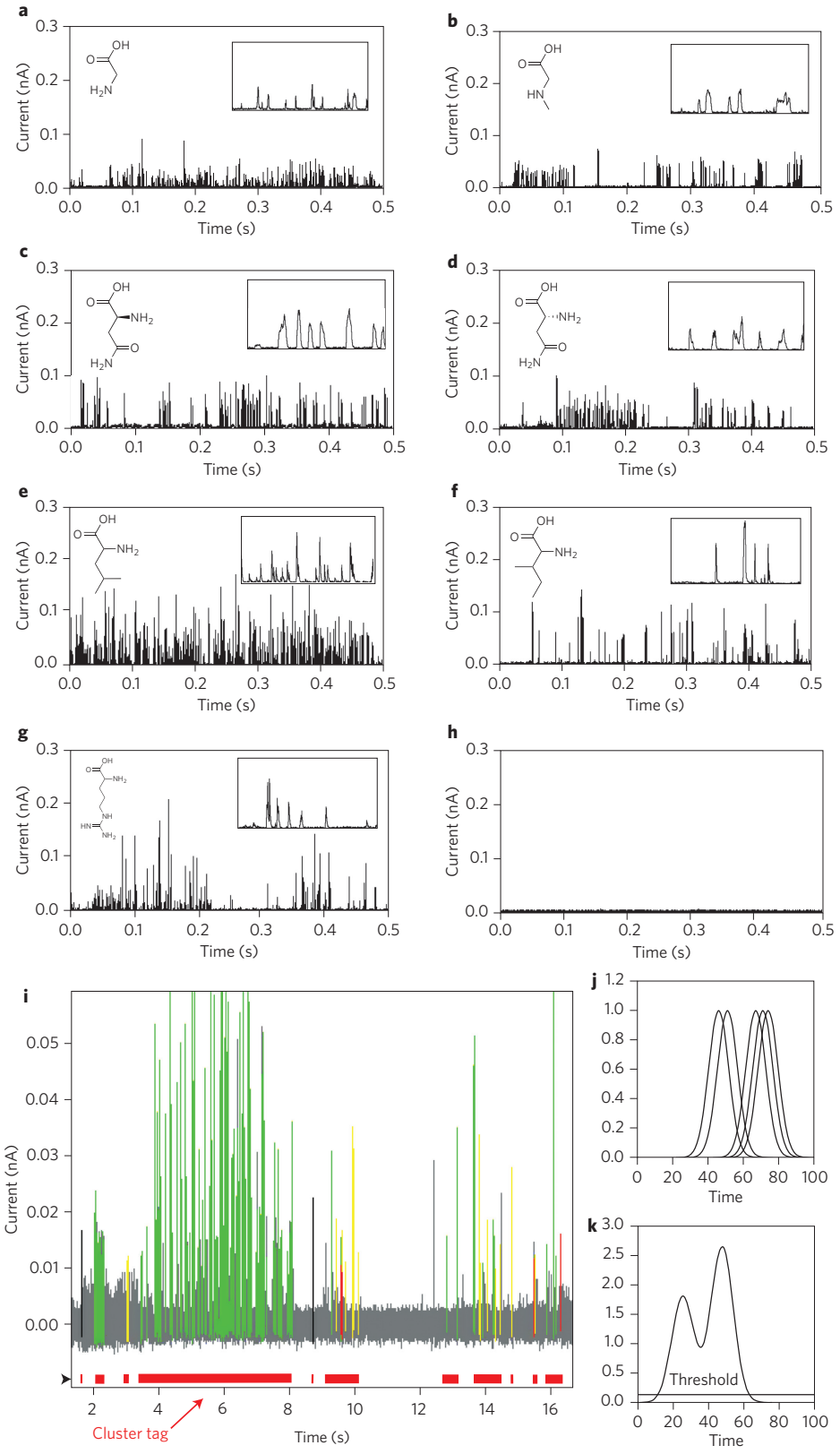


**Figure 2 | Examples of RT signals from amino acids. a**,**b**, Gly (**a**) and its *N*-methylated modification, sarcosine (mGly) (**b**). **c**,**d**, Enantiomers L-Asn (**c**) and D-Asn (**d**). **e**,**f**, Isobaric isomers Leu (**e**) and Ile (**f**). **g**, Data for the charged amino acid, Arg. **h**, Control data from buffer solution alone. Insets: Expanded traces (current scale, 150 pA; timescale, 20 ms) displaying the complex peak shapes that are important features in the analysis of these data. **i**, Signal trace for Arg, colour-coded according to the peak assignments made by a machine learning algorithm (green, correct; red, wrong call; black, 'water peak'; yellow, common to all amino acids). The red bars at the bottom mark signal clusters generated by a particular single-molecule binding event. **j**,**k**, Automatic cluster identification was carried out by placing Gaussians of unit height and full width of 4,096 data points (1 data point = 20 μs) at the location of each spike (**j**), summing them (**k**), and assigning a cluster to regions where this sum exceeds 0.05. This choice picks out obvious single-molecule events well (see Fig. 5).



© 2014 Macmillan Publishers Limited. All rights reserved.

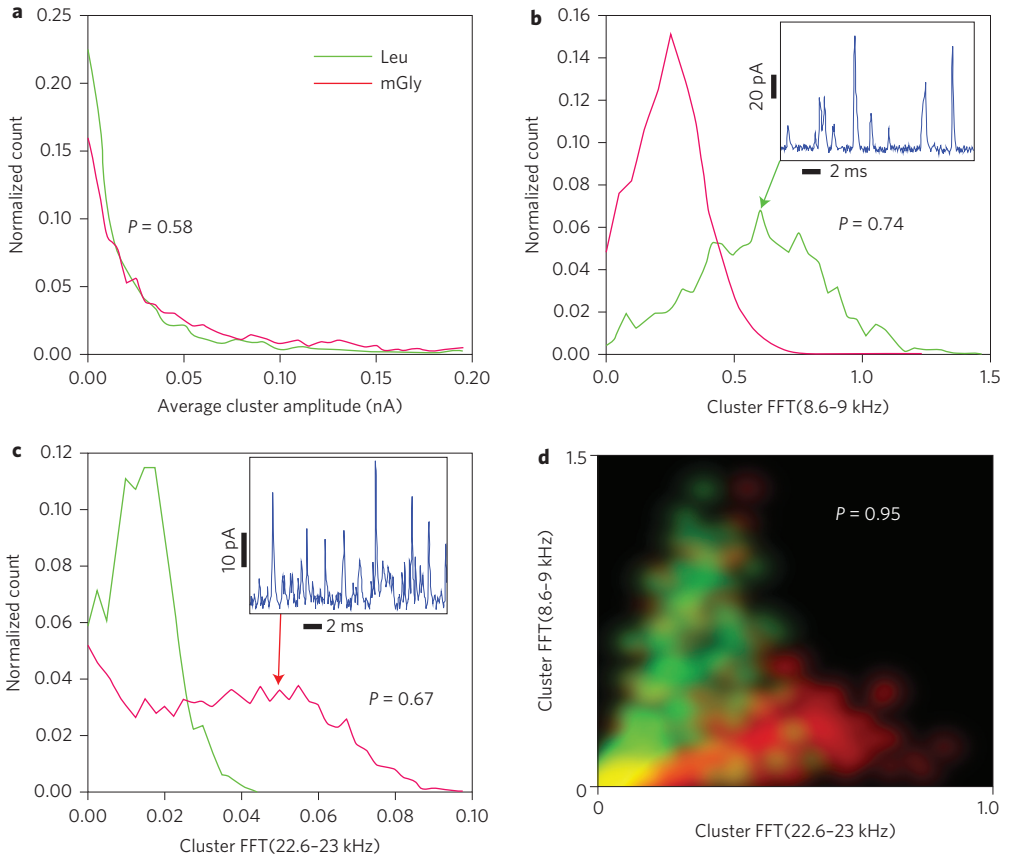


**Figure 3 | Signal features identify analytes. a**, Peak amplitudes are exponentially distributed so provide little discrimination. Assigning the larger spikes to mGly (red curve) yields an accuracy ($P = 0.58$) only slightly better than random (0.5). **b**,**c**, Particular Fourier components (Supplementary Table 1) of the clusters show more separation, producing 74% (**b**) and 67% (**c**) accuracies if called solely on the more probable value of the feature. The way in which these Fourier components reflect peak shapes in a cluster is illustrated by the signal traces inset in **b** and **c**, each trace having the feature value indicated. The high amplitude of high-frequency components of the mGly signals (inset in **c**) is evident in the sharper spikes. Accuracy improves when multiple features are used together. **d**, Two-dimensional plot of probability density as a function of the two FFT feature values. The colour scale shows mGly data points as red and Leu points as green. Calling all the spikes with pairs of feature values that fall in the green regions as Leu and all the spikes with pairs of features that fall in the red regions as mGly produces a correct call 95% of the time. Only the yellow regions yield ambiguous calls.

accuracy to 80% or greater in all three cases. (The separation of stereoisomers is presumably a consequence of the local-chiral adsorption geometry on surfaces[22].)

SVM analysis with a large set of signal features is carried out as follows. Each signal spike, represented by $N$ feature values, is plotted in $N$-dimensional space. A subset of known data is used to find the support vectors (of $N-1$ dimension) that best partition the known data and thus train the SVM. Data from subsequent analyses are then identified according to on which side of the partition they reside. Thus far, we have described the SVM as a binary classifier, separating data into one of two classes, but multiclass SVMs are readily constructed. In the version of the SVM used here, training for multiple analytes works by determining the support vector set that best separates signals from each analyte from a pool of signal feature values taken from the remainder of the analytes. This process was repeated seven times to cover each of the analytes studied, generating seven different support vectors. Once trained, signals from an unknown sample are fed to all seven SVMs sequentially, and a confidence level returned for assignment to each amino acid (as opposed to the remainder). Each signal spike is assigned to the amino acid corresponding to the highest confidence level.

## Reproducibility of the SVM analysis

The key questions are 'How reproducible are the tunnelling data?' and 'How transferrable is the SVM training?' To address these questions, we have analysed multiple sets of data for each analyte, selecting signal features and settings for the SVM parameters that give robust results across multiple data sets. Each spike and the cluster that contains it were characterized by values of 161 signal features (Supplementary Table 1). This large number of features includes parameters that describe amplitudes and amplitude fluctuations in both individual spikes and clusters as well spike and cluster shapes as described by Fourier and cepstrum[23] components. (Fourier and cepstrum components were corrected for the frequency response of the instrument; see Methods and Supplementary Fig. 8.) A total of 30,000 data spikes (corresponding to about 3,000 clusters) were collected for each of the seven analytes. A correlation analysis (Supplementary Fig. 4) was used to identify groups of signal features that are linearly dependent, with each group represented by one of the strongly correlated signal features. This reduced the total feature set by 40 (Supplementary Table 2) to 121. A second correlation analysis identified features that vary from experimental run to run on the same analyte, and those that do not vary from one analyte to another. Fifteen such unreliable features were found (Supplementary Table 3), and removing them reduces the sensitivity to experimental artefacts and reduces the feature set to 106. Noise spikes (1–15% of the total data, varying from run to run) were eliminated by training the SVM to find signals common to all seven analytes. This last stage of noise filtering was adjusted by varying the soft margin (broadening of the partition



© 2014 Macmillan Publishers Limited. All rights reserved.

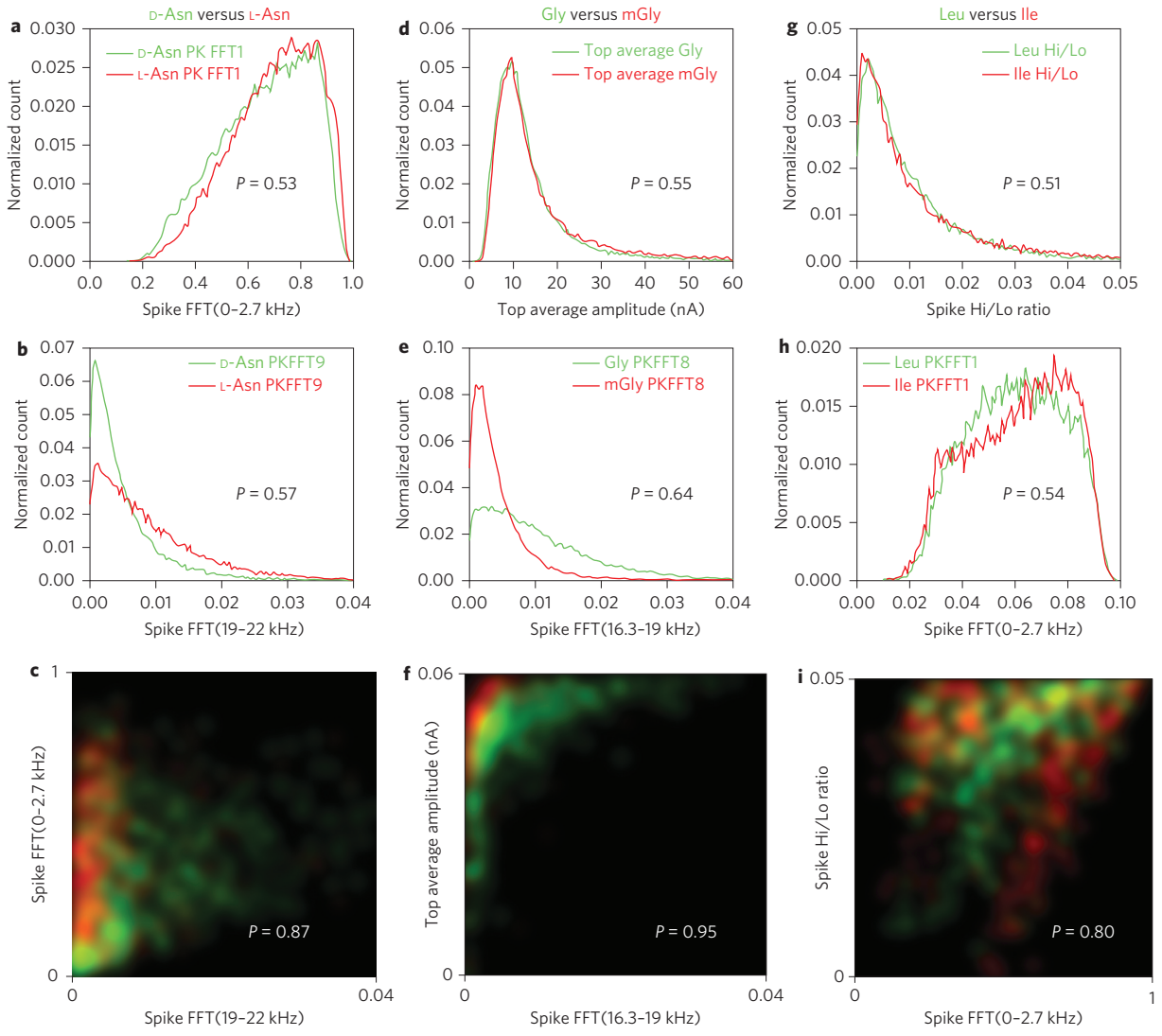


**Figure 4 | Closely related pairs of analytes can be significantly separated (>80%) using just two signal features together.** All data are for pure solutions of one analyte. **a–i**, Chiral enantiomers D-Asn and L-Asn (**a–c**), Gly and mGly (**d–f**) and the isobaric isomers Leu and Ile (**g–i**) are quite well separated in two-dimensional probability density maps (**c**,**f**,**i**), even when the distributions of any one signal feature are almost completely overlapped in one dimension (**a**,**b**,**d**,**e**,**g**,**h**; see Methods and Supplementary Table 1 for a description of these features). The two-dimensional maps plot probability densities for the analyte pairs (colour coded as listed at the top) as a function of both features, which, by themselves, produce separations only a little above random (0.51 to 0.64). Probabilities of making a correct call based on the probability densities are marked on **c**, **f** and **i**, and calculated as described in the caption for Fig. 3.

boundaries) of the SVM parameters. Increasing the soft margin improved accuracy at the cost of rejecting more signals (Supplementary Fig. 5). The SVM was then trained on a small subset (~10%) of the data and then tested on the remainder. This process was repeated using randomly chosen training data to ensure that fluctuations in the outcomes were small. Finally, the analysis was repeated with smaller numbers of signal features to see how the final accuracy depended on the number of features used. Table 1 shows how a single signal spike can be assigned to any one of the seven analytes with 95% accuracy, compared to 14% probability of a correct random call. Spikes within a cluster are highly correlated (Supplementary Fig. 15), so sampling multiple peaks within a cluster cannot be used to improve accuracy. However, spikes from different clusters constitute independent reads and can be used to improve accuracy when reads are known to come from molecules of the same type as in the output from a chromatography column (or some other sequential separation). Cluster correlations were removed by randomizing the order of the spikes, and analytes assigned based on a majority vote of successive spikes (Table 1) to yield accuracies that approach 100%.

Accuracy is reduced in the more challenging (and more realistic) case where the SVM is trained on data from one run and tested on data from other runs, because data filtering is not as stringent. In this case, accuracies of 90% or greater could be attained for pooled data from all seven analytes, based on a single read. Once again, accuracies increased rapidly when repeated reads were made of spikes from different clusters.

## Analysing mixtures of analytes

Thus far, we have confined our analysis to pure samples of one type of analyte, distinguishing one pure sample from six other pure samples. Can the SVM, trained on a pure analyte, recognize it in a mixture? To address this question, we made mixtures of L- and D-Asn with stoichiometric ratios of 1:1, 2:1 and 3:1, repeating measurements at each concentration, twice. We then used the support vectors developed for pure L- and D-Asn to assign the spikes in the signal train obtained from mixtures. Figure 5a shows



© 2014 Macmillan Publishers Limited. All rights reserved.

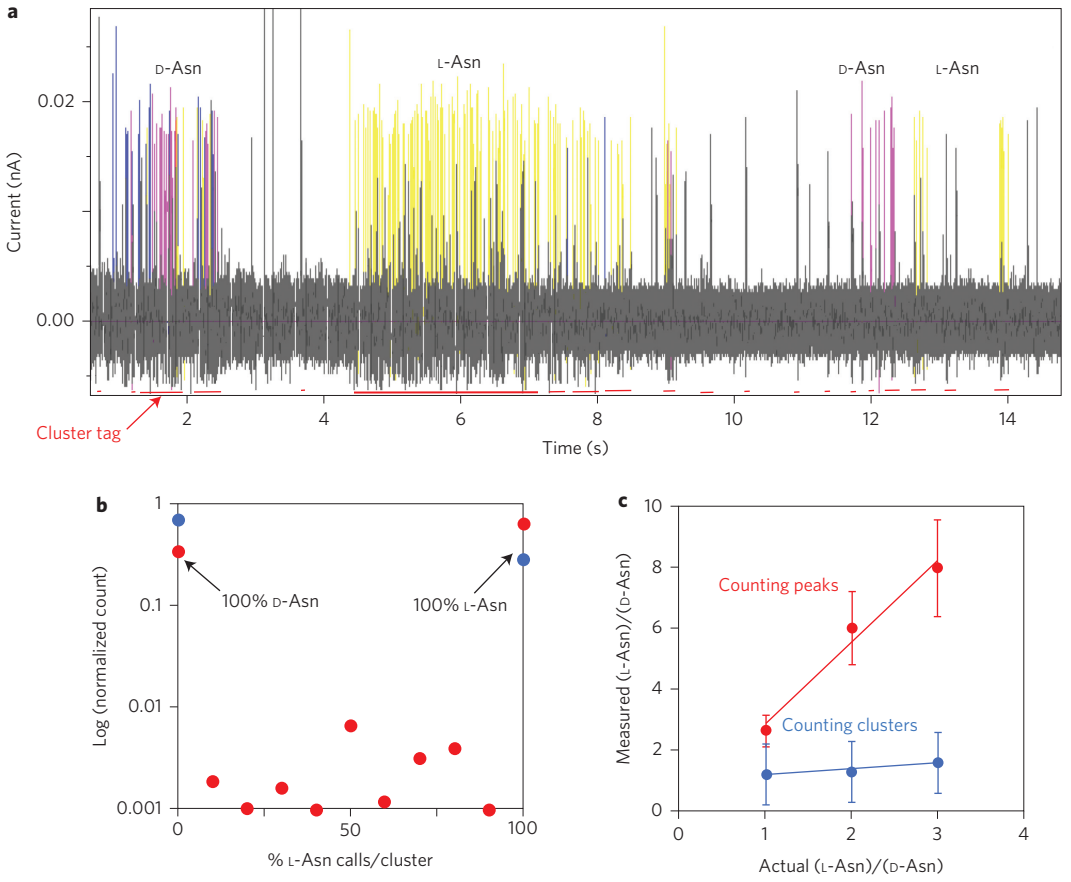


**Figure 5 | A mixture produces alternating cluster signals as different molecules diffuse into and out of the gap. a**, Signal trace obtained with a 1:1 mixture of L- and D-asparagine. The SVM assignments are coded purple (D-Asn) and yellow (L-Asn) (black spikes are unassigned). **b**, Each cluster (red tags in **a**) contains only one type of signal, as shown statistically. The red points are for 556 raw data clusters and the blue points are for 400 clusters that remain after filtering for common signals. After filtering (blue points), no mixed clusters survive, with all of the clusters being 100% L- or D-Asn signals. **c**, Quantification of the L/D ratio using SVM trained on pure samples. The measured ratio increases with actual ratio in the samples, but the calibration depends on whether the number of signal spikes (red) or clusters (blue) is used, probably reflecting differential binding. Error bars are from repeated runs and repeated samplings.

**Table 1 | Accuracy with which any one of seven pure analytes is identified from the total pool of data taken from all seven pure samples using 52 signal features together.**

| Number of spikes | Arg | D-Asn | L-Asn | Gly | Ile | Leu | mGly |
|---|---|---|---|---|---|---|---|
| 1 | 95.14 | 94.99 | 96.99 | 97.24 | 96.87 | 94.36 | 96.45 |
| 3 | 98.77 | 99.62 | 99.99 | 99.62 | 99.99 | 99.55 | 99.99 |
| 5 | 99.99 | 99.99 | 99.99 | 99.99 | 99.99 | 99.99 | 99.99 |

Results in the first row are based on a single spike. The subsequent rows are based on a majority vote using three and five spikes taken from different signal clusters. These results were obtained with the noise-filter soft margin set to reject ~70% of the data spikes.

a stream of raw data that has been colour-coded according to this assignment (yellow, L-Asn; purple, D-Asn; black, common). The red bars at the bottom of the trace mark the identified clusters and it is clear that each cluster corresponds to just one analyte or the other. This is summarized statistically for 556 clusters in Fig. 5b. Essentially all of the clusters consist of all L- or all D-Asn spikes, with less than 1% in total containing more than one type of spike. When the common noise filter is applied, only pure clusters remain (blue points). This further supports the view that clusters reflect single-molecule binding events.

The measured stoichiometric ratio is quite sensitive to signal filtering and the method used to count molecules. Figure 5c shows the measured L/D ratio based on the total number of spikes of each type (red data points) and the total number of clusters of each type (blue data points). Spikes overcount the L-Asn content (the slope of the linear fit in Fig. 5a is 2.7), and clusters undercount it (slope of 0.2). In the case where spikes are counted, the excess assigned to L-Asn is a consequence of longer clusters, probably reflecting stronger binding of this analyte to the recognition molecules. In the case where clusters are counted, the undercounting of L-Asn molecules may reflect a local reduction in concentration owing to preferential binding of L-Asn on the surfaces of the electrodes. Nonetheless, the relationship between measured and actual stoichiometry is monotonic and reproducible to better than ~20%.

## RT signals from peptides

The obvious hydrogen-bonding sites for amino acids are the zwitterionic centres (Fig. 1a). In a peptide, N and C termini are more spatially separated, so it is not at all clear that amino acids, as parts of peptides, will produce RT signals. We found that 100 μM



© 2014 Macmillan Publishers Limited. All rights reserved.

solutions of the short peptides GGGG and GGLL readily produced RT signals (Supplementary Fig. 6). Interestingly, the SVM trained on pure amino acids did not recognize either of these peptides as their constituent amino acids (Supplementary Table 5a). However, each peptide produced distinctive signals, allowing one to be distinguished from the other and also from any of the amino acids (Supplementary Table 5a). Thus, the binding motifs of the amino-acid residues in a peptide are clearly different from those of the same amino acids free in solution. We also obtained signals from the trimer, GGG. An SVM analysis of all three peptides (GGG, GGGG and GGLL) together (Supplementary Table 5b) yielded >90% accuracy (with 65% of the signals rejected as common, as might be expected given the sequence homology). Thus, multiple peptides may be separated from each other, even when the difference is just one residue in four. This suggests that amino-acid variants of proteins can be detected. In addition, single-molecule sequencing of proteins may be possible, particularly if residues can be presented to a tunnel gap sequentially.

In this pilot study, high concentrations (~100 μM) were used to ensure rapid diffusion of analytes into the tunnel gap (although concentrations down to 1 μM worked). Sample concentrations can easily be reduced by micro- or nanofluidic injection of samples into the tunnel gap. Of significant interest is incorporation of tunnel junctions into nanopores, where capture assisted by electrophoresis[24] or electroosmosis[25], could reduce this concentration to the picomolar range.

## Bonding in the RT junctions

All seven analytes form stoichiometric adducts with one or two ICA molecules, as demonstrated by ESIMS (Supplementary Figs 11, 12 and Supplementary Tables 6, 7). ICA was designed to bond DNA bases, but the density functional theory calculated structures in Fig. 1a and Supplementary Fig. 1 show that amino acids can be captured by hydrogen-bonding to aminium and carboxylate groups of their zwitterionic centres. However, the observation that peptides generate RT signals that are different from those generated by amino acids (Supplementary Table 5a), suggests that other types of binding motif are possible. How many such motifs might occur? To address this question for the case of the amino acids, we used an algorithm that identifies clusters of data[26] to locate such clusters in the 24-dimensional space occupied by the most significant signal features from single spikes (Supplementary Fig. 16). Three distinct clusters were found for six of the amino acids (Supplementary Table 9), with only two for *N*-methylglycine, methylation presumably blocking a bonding site.

We examined bonding further using force spectroscopy. A dipeptide (Cys–Gly) was attached to the tip of an atomic force microscope through a polyethyleneglycol (PEG) linker via the thiol of the cysteine residue and single-molecule rupture forces were recorded (a hexane-terminated PEG was used as control). When the peptide retracted from an ICA-coated gold surface, bond-rupture events were observed. The distribution of rupture forces (Supplementary Fig. 14) is consistent with two hydrogen bonds[17].

## Conclusions

RT generates complex signals that can form the basis of a new type of molecular spectroscopy for identifying a potentially vast range of chemicals at the single-molecule level. It discriminates between members of molecular classes, such as enantiomers and isobaric isomers, analytes that present challenges to other analytical techniques. RT has the potential to demonstrate numerous significant advantages over a variety of current types of instrumentation and analytical methods that require chemical labelling or complex and expensive instrumentation, such as mass spectrometers. For example, instrumentation for single-molecule analysis and protein sequencing that integrates RT and nanopore technologies on a solid-state device platform would likely be substantially smaller, less expensive, have lower operating costs and be more robust. Individual amino acids can be identified with high accuracy. In the near term, a microreactor containing an exo-peptidase should be able to identify the terminal sequence of proteins by feeding the digest to the tunnel gap and analysing the time-dependent signal (Supplementary Fig. 7). We have also shown that peptide chains generate distinctive and reproducible signals. With RT tunnel junctions integrated into nanopores, it may very well prove possible to carry out continuous strand sequencing of proteins (a molecular motor being used to feed entire proteins into nanopores[27]). The real power of RT as a chemical spectroscopy lies in the possibility of massively parallel detection using large-scale integration of solid-state devices. Such devices are under development in our laboratory.

## Methods

**Preparation of analytical solutions.** Amino acids were obtained from Sigma Aldrich (>98% purity) and dissolved in 1 mM phosphate buffer (pH 7.4), made using water from a Milli-Q system with specific resistance of ~18 MΩ cm and total organic carbon contamination below 5 ppb. Peptides were obtained from CPC Scientific and solutions prepared as for the amino acids.

**Preparation of probes and substrates.** Palladium substrates were deposited on a 750 μm silicon wafer using electron-beam evaporation of 100 nm Pd onto a 10 nm Ti adhesion layer. Probes were etched[11] from 0.25 mm Pd wire (California Fine Wires) and insulated with polyethylene to leave the metal end open with a linear dimension of a few tens of nanometres. Probes were tested to ensure that leakage current was <1 pA in the standard buffer solution at 0.5 V bias. This is because ionic leakage current cannot be simply subtracted from the signal because of its distance dependence[28], so leaky probes result in errors in the set point current. For functionalization, insulated probes and the Pd substrates were first cleaned by rinsing them with ethanol and $H_2O$, dried with nitrogen and then immersed in a solution of ICA[13] (0.5 mM) in ethanol. After ~16 h, the probe and substrate were removed, rinsed with ethanol, gently dried with nitrogen and used immediately.

**Tunnelling measurements.** We used two different PicoSPMs (Agilent Technologies) equipped with custom LabView interfaces for data acquisition. Tunnel current was sampled at 50 kHz. The −3 dB bandwidth of the current-to-voltage converter was 7 kHz, but useful signals were obtained out to the Nyquist limit of 25 kHz after correction for the instrumental response (Supplementary Fig. 8). The liquid cells were cleaned in Piranha (note: these solutions are potentially explosive and must be handled with extreme care) and rinsed with Milli-Q water and ethanol. The current set point was set to 4 pA with 0.5 V bias applied (probe positive, as this results in less leakage) and the probe approached with integral and proportional gains set to 1.0. The surface was scanned to ensure that the grain structure of the Pd was clearly visible (Supplementary Fig. 9). The microscope was left to stabilize for at least 2 h before signals were recorded, and the integral and proportional gains were then reduced to 0.1. The control (1.0 mM phosphate buffer at pH 7.4) was run before an amino-acid solution was measured. Recordings were distorted by movement of the Z transducer during runs in which a series of high-amplitude spikes were recorded, but this artefact was common to all analytes and incorporated into the training of the SVM. We used different batches of substrates and probes for each run, usually recording four runs for each analyte. We also alternated measurements between different instruments. In this way, the influence of small changes in experimental conditions could be removed from the final analysis.

**SVM analysis.** We used the kernel-mode SVM[10] available from https://github.com/vjethava/svm-theta. Each spike above 15 pA in amplitude was characterized using the features listed in Supplementary Table 2. The shape of each spike was characterized by constructing a fast Fourier transform (FFT). The resulting Fourier amplitude distribution was then downsampled using linear interpolation into nine bins of equal frequency intervals from zero to 25 kHz. FFT amplitudes (before downsampling) were averaged across three equally spaced frequency intervals (0–2.7 kHz, 8.4–11.1 kHz and 22.3–25 kHz), and these averages were used as additional features, as was the ratio of the highest to lowest FFT bins useful (peak hi/lo ratio, Fig. 4g).

Clusters contain additional information. They were identified with a Gaussian-broadening algorithm as described in Fig. 2 (ref. 9). The peaks used to locate the clusters were subject to a 15 pA threshold, but once a cluster was identified, all the data in it were used for the analysis, so amplitudes down to the baseline were included (see Fig. 3a). Distributions of cluster lengths for various analytes are shown in Supplementary Fig. 10a,b. We also developed a series of features to describe these clusters (Supplementary Table 2). These included the spike frequency



© 2014 Macmillan Publishers Limited. All rights reserved.

within a cluster, as well as the Fourier spectrum of the whole cluster (deconvolved for instrumental response by spectral division). Clusters contain many more data points than individual spikes, so the downsampling of the FFT was much finer, with a total of 61 bins used (each one corresponding to 25 kHz/61 or 410 Hz in width). The method of Noll[23] was used to calculate the cepstrum amplitudes from the Fourier transform of the power spectrum, downsampling again to 61 frequency bins.

So as not to bias the analysis towards features with bigger numerical values and ranges, we rescaled all features as follows. The distribution of each signal feature was measured for one amino acid (in this case, arginine for the amino-acid analysis and glycine for the peptide analysis). The scale factor and additive constant required to move the mean of the distribution to zero and the standard deviation to 1.0 were calculated. Feature values for all the other analytes were remapped using the same linear transformation.

Feature selection was performed in three stages. First, those features that showed too much linear correlation were removed. The normalized correlation between different pairs of features $(x, y)$ was defined in the usual way, $\sigma_{xy} = \langle (x - \bar{x})(y - \bar{y}) \rangle$ where we normalized the components using $\sigma_{xx} = 1$. All the data from the entire pool were used to generate a correlation matrix where correlations are shown by off-diagonal elements (Supplementary Fig. 4). Trial and error resulted in rejecting all feature combinations for which $\sigma_{xy} \geq 0.7$. We chose one feature from each overly correlated set to represent the set in the next stage of analysis.

Second, a comparison was performed for each feature for its variation over repeated experiments on the same analyte versus the variation between the different analytes. Histograms of all feature values (see Figs 3 and 4) were compiled for each experimental run for a given analyte. The absolute values of the differences between the normalized histograms were accumulated to give an 'in-group' fluctuation. The same procedure was carried out for all possible pairs of analytes to give an 'out-group' measure of fluctuation. Parameters were then ranked by the magnitude of the ratio of out-group to in-group fluctuation and the bottom 15 parameters dropped (Supplementary Table 3). Finally, the usefulness of the remaining features was evaluated by determining the identification accuracy obtained with a randomly selected group of them. A tree search was used to maximize the efficiency of this process. This led to the 52 features (Supplementary Table 4) used in the analysis reported in Table 1.

Full details of the SVM (written in Matlab) can be found in a download of the data analysis code available from https://svmsignalanalysis.codeplex.com/.

**ESIMS.** Solutions of amino acids in 1:1 and 2:1 molar ratios of ICA to amino acid were prepared by dissolving these chemicals into water for a final concentration of 100 μM amino acid. Samples were injected into a Bruker MicrOTOF-Q electrospray ionization quadrupole time-of-flight (ESI-Q-TOF) mass spectrometer, and tandem mass spectrometry was used to confirm the composition of mass peaks from adducts. We checked that the lack of buffer did not hinder the acquisition of RT signals (Supplementary Fig. 13). Full details of the methods and analysis are given in the Supplementary Information.

**Force spectroscopy.** A Cys–Gly dipeptide was ligated to a PEG tether ($N = 36$), which was in turn attached to a SiN AFM probe (VeecoProbes) using click chemistry[29]. Force curves (Supplementary Fig. 14) were collected in aqueous buffer over a gold-coated mica substrate covered with a monolayer of ICA. Full details of the sample preparation, data acquisition and analysis are given in the Supplementary Information.

Received 6 March 2013; accepted 18 February 2014; published online 6 April 2014

## References


1. Uhlen, M. & Ponten, F. Antibody-based proteomics for human tissue profiling. *Mol. Cell. Proteom.* **4,** 384–393 (2005).
2. National Research Council (US) Committee on Intellectual Property Rights in Genomic and Protein Research and Innovation. *Reaping the Benefits of Genomic and Proteomic Research: Intellectual Property Rights, Innovation, and Public Health* (National Academies Press, 2006).
3. Archakov, A. I., Ivanov, Y. D., Lisitsa, A. V. & Zgoda, V. G. AFM fishing nanotechnology is the way to reverse the Avogadro number in proteomics. *Proteomics* **7,** 4–9 (2007).
4. Huang, S. *et al.* Identifying single bases in a DNA oligomer with electron tunneling. *Nature Nanotech.* **5,** 868–873 (2010).
5. Chang, S. *et al.* Electronic signature of all four DNA nucleosides in a tunneling gap. *Nano Lett.* **10,** 1070–1075 (2010).
6. Huang, S. *et al.* Recognition tunneling measurement of the conductance of DNA bases embedded in self-assembled monolayers. *J. Phys. Chem. C* **114,** 20443–22044 (2010).
7. Lindsay, S. M. *et al.* Recognition tunneling. *Nanotechnology* **21,** 262001 (2010).
8. Friddle, R. W., Noy, A. & De Yoreoa, J. J. Interpreting the widespread nonlinear force spectra of intermolecular bonds. *Proc. Natl Acad. Sci. USA* **109,** 13573–13578 (2012).
9. Chang, S. *et al.* Chemical recognition and binding kinetics in a functionalized tunnel junction. *Nanotechnology* **23,** 235101 (2012).
10. Chang, C. C. & Lin, C. J. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2,** 27–52 (2011).
11. Chang, S. *et al.* Palladium electrodes for molecular tunnel junctions. *Nanotechnology* **23,** 425202 (2012).
12. tuchband, M., He, J., Huang, S. & Lindsay, S. M. Insulated gold scanning tunneling microscopy probes for recognition tunneling in an aqueous environment. *Rev. Sci. Instrum.* **83,** 015102 (2012).
13. Liang, F., Li, S., Lindsay, S. M. & Zhang, P. Synthesis, physicochemical properties, and hydrogen bonding of 4(5)-substituted-1H-imidazole-2-carboxamide, a potential universal reader for DNA sequencing by recognition tunneling. *Chem. Eur. J.* **18,** 5998–6007 (2012).
14. Daniel, J. R. M., Friess, S. D., Rajagopalan, S., Wendt, S. & Zenobi, R. Quantitative determination of noncovalent binding interactions using soft ionization mass spectrometry. *Int. J. Mass Spectrom.* **216,** 1–27 (2002).
15. Nesatyy, V. J. Mass spectrometry evaluation of the solution and gas-phase binding properties of noncovalent protein complexes. *Int. J. Mass Spectrom.* **221,** 147–161 (2002).
16. Sreekumar, A. *et al.* Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* **457,** 910–914 (2009).
17. Fuhrmann, A. *et al.* Long lifetime of hydrogen-bonded DNA basepairs by force spectroscopy. *Biophys. J.* **102,** 2381–2390 (2012).
18. tsutsui, M., Taniguchi, M., Yokota, K. & Kawai, T. Identification of single nucleotide via tunnelling current. *Nature Nanotech.* **5,** 286–290 (2010).
19. Zwolak, M. & Di Ventra, M. Electronic signature of DNA nucleotides via transverse transport. *Nano Lett.* **5,** 421–424 (2005).
20. Cover, T. M. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Electron. Comput.* **EC-14,** 326–334 (1965).
21. Jethava, V., Martinsson, A., Bhattacharyya, C. & Dubhashi, D. The Lovász $\theta$ function, SVMs and finding large dense subgraphs. *Neural Inf. Proc. Syst.* 1169–1177 (2012).
22. Kühnle, A., Linderoth, T. R., Hammer, B. & Besenbacher, F. Chiral recognition in dimerization of adsorbed cysteine observed by scanning tunnelling microscopy. *Nature* **415,** 891–893 (2002).
23. Noll, A. M. Short-time spectrum and cepstrum techniques for vocal-pitch detection. *J. Acoust. Soc. Am.* **36,** 296–302 (1964).
24. Wanunu, M., Morrison, W., Rabin, Y., Grosberg, A. Y. & Meller, A. Electrostatic focusing of unlabelled DNA into nanoscale pores using a salt gradient. *Nature Nanotech.* **5,** 160–165 (2010).
25. Keyser, U. Controlling molecular transport through nanopores. *J. R. Soc. Interface* **8,** 1369–1378 (2011).
26. Chiu, S. Fuzzy model identification based on cluster estimation. *J. Intell. Fuzzy Syst.* **2,** 267–278 (1994).
27. Nivaia, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation though an alpha-hemolysin pore. *Nature Biotechnol.* **31,** 247–250 (2013).
28. Bard, A. J., Fan, F. R. F. & Mirkin, M. V. in *Electroanalytical Chemistry* Vol. 18 (ed. Bard, A. J.) 243–371 (CRC Press, 1994).
29. Senapati, S., Manna, S., Lindsay, S. M. & Zhang, P. Application of catalyst-free click reactions in attaching affinity molecules to tips of atomic force microscopy for detection of protein biomarkers. *Langmuir* **29,** 14622–14630 (2013).


## Acknowledgements

S. Chang assisted in the original survey of amino acids. The authors thank P. Pang, P. Krstic, C. Hernandez-Suarez and W. Offenberg for useful discussions. This work was supported in part by a DNA sequencing technology grant from the NHGRI (HG 006323).

## Author contributions

Y.Z. and H.L. carried out tunnelling measurements with assistance from S.S., W.S. and J.I., B.A. wrote the SVM code and analysed data. B.G. contributed to the analysis. S.M. carried out force spectroscopy experiments. C.B. and S.B. carried out the electrospray MS., P.Z. and S.L. designed experiments and S.L. wrote the paper.

## Additional information

Supplementary information is available in the online version of the paper. Reprints and permissions information is available online at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.L.

## Competing financial interests

Y.Z., P.Z. and S.L. are named as inventors in a patent application (patent application number PCT/US13/24,130; title: Device for reading amino acid sequence inventors). S.L. is co-founder of a company that has a licence option on the technology.



© 2014 Macmillan Publishers Limited. All rights reserved.