Supplement
to

'*Amino acid discriminators in a nanopore and the feasibility of sequencing peptides with a tandem cell and exopeptidase*'

G. Sampath

This Supplement consists of the following sections and has some overlap with the main text:

1. Fokker-Planck model

The mathematical model for the tandem cell here is very similar to that for the tandem cell proposed for exosequencing of DNA [1]. Similar to a mononucleotide in the original tandem cell, a residue is considered to be a particle that does not interact chemically with the pore lumen or the electrolyte and moves after being cleaved by the exopeptidase through a combination of diffusion and electric drift. With most of the potential difference V_{05} dropping across the two pores (V_{05} = 0.365 V, V_{23} = 1.6 mV, V_{34} = ~0.18 V), movement of a cleaved residue through *trans*1/*cis*2, DNP, and *trans*2 is dominated by diffusion. The movement through *trans*1/*cis*2 and DNP can be studied via the trajectory of a particle whose propagator function G (x,y,z,t) is given by a linear Fokker-Planck (F-P) equation in one dimension (z) for DNP, or three (x,y,z) for *trans*1/*cis*2, containing a drift term in the z direction that arises from the voltage difference V_{05}. The drift field affects charged residues but not neutral residues. Initially each section is considered independently. The behavior at the interface between two sections is examined later.

Solution of the one-dimensional case

The F-P equation in the one-dimensional case can be solved in a straightforward way using methods from partial differentiation equations and Laplace transforms. Let μ be the mobility of the particle and D its diffusion constant. Following [1], the mean E(T) and variance $\sigma^2(T)$ of the translocation time T over a channel of length L that is reflective at the top and absorptive at the bottom with applied potential difference of V are given by

$$E(T) = (L^2/D\alpha)[1 - (1/\alpha)\,(1 - \exp(-\alpha))] \qquad (1)$$

and

$$\sigma^2(T) = (L^2/D\alpha^2)^2\,(2\alpha + 4\alpha\exp(-\alpha) - 5 + 4\exp(-\alpha) + \exp(-2\alpha)) \qquad (2)$$

where

$$\alpha = v_z L/D \qquad v_z = \mu V/L \qquad (3)$$

Here v_z is the drift velocity due to the electrophoretic force experienced by a charged particle in the z direction. For $v_z = 0$, these two statistics are

$$E_0(T) = L^2/2D; \qquad \sigma_0^2(T) = (1/6)\,(L^4/D^2) \qquad (4)$$

As discussed next, these formulas can be applied to all three relevant chambers: *trans*1/*cis*2 ($T = T_{trans1/cis2}$; $L = L_{23}$), DNP ($T = T_{DNP}$; $L = L_{34}$), and *trans*2 ($T = T_{trans2}$; $L = L_{45}$). A piecewise approach is taken, with each section considered independent of the others. The behavior at the interface between two adjoining sections is discussed below.

Translocation through DNP. A cleaved residue is treated as a particle that is released at the top of DNP at t = 0, reflected there at t > 0, and 'captured' at the bottom at t > 0. Regardless of whether a residue is charged or not the diffusion is always in the z direction because of the reflecting barrier at z = 0. With $V_{05} > 0$ α is positive for negative residues and negative for positive residues. The resulting translocation time mean for negative residues is reduced below that due to $v_z = 0$, and goes above for positive residues. In both cases the net translocation is in the positive z direction for the values of V_{05} in use. The electric field has no effect on neutral residues and their movement is entirely due to diffusion; therefore α = 0 for them. In summary all residues, charged or not, will move in the z direction and cause a current blockade in DNP; this, along with other measures (see 'Multiple discriminators in sequencing' in the main text), can be used to identify a residue. Equations 1 through 4 apply with $L = L_{34}$.

Translocation through trans*1*/cis*2*. This is modeled in three dimensions using a rectangular box-shaped region. (The tapered geometry of Figure 1 in the main text is discussed below.) A particle is released at the top center of *trans*1/*cis*2 at t = 0, reflected at the top and sides of the box at t > 0, and translocates to the bottom of the compartment where it is 'absorbed' at some t > 0. That is, the particle is considered to be detected when it reaches $z = L_{23}$ independent of x and y and to move into DNP without regressing into *trans*1/*cis*2. The propagator function G(x,y,z,t) can be written as the product of three independent propagator functions. It is shown in [1] that diffusion in the x and y directions has no effect so that the first passage time distribution in the three dimensional case reduces to that in the one-dimensional case. Thus Equations 1 through 4 apply with $L = L_{23}$. The effect of α on charged and neutral residues is the same as in DNP.

Translocation through trans*2*. This behavior can be modeled in the same way as that of a cleaved residue in *trans*1/*cis*2.

2. Volume excluded in a pore by an analyte particle (monomer)

The ratio of volume excluded (V_{excl}) by a particle treated as a cylinder of radius equal to the hydrodynamic radius R_H and height $2R_H$ in a cylindrical pore of radius r and length L to the pore volume (V_{pore}) is given by

$$V_{excl}/V_{pore} = 1 - L(A_{pore}-A_{residue})/L(A_{pore}-A_{residue}) + 2R_HA_{residue} \qquad (5)$$

where the A's are cross-section areas. Table 3 gives the ratio for each of the 20 amino acids with $L = L_{34} = 10$ nm and r = 1.5 nm. R_H values are taken from Table 1.

3. Dependence of amino acid charge on solution pH

Let the set of amino acids be **AA** = [A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V] where AA[i] is the i-th amino acid, $1 \le i \le 20$. Let the pH value of the solution (electrolyte) be p, the kA value of the i-th amino acid k[i], and the kA values of the carboxyl and amine ends of the amino acid kC and kN respectively. From [2]

$$\mathbf{k} = [\ \_,\ 12.48,\ \_,\ 3.86, 8.33,\ \_, 4.25,\ \_, 6.0,\ \_,\ \_,\ 10.53,\ \_,\ \_,\ \_,\ \_,\ \_,\ \_, 10.07, \_\ ],\ kC = 9.69,\ kA = 2.34 \qquad (6)$$

where the placeholder _ = 'not applicable'. Using the Henderson-Hasselbach method the following equation can be written for C[i], the electrical charge multiplier for amino acid i:

$$\begin{aligned} C[i] &= 10^{kC} / (10^{p} + 10^{kC}) - 10^{p} / (10^{p} + 10^{kN}); && i = 1, 3, 6, 8, 10, 11, 13, 14, 15, 16, 17, 18, 20 \\ & && \text{(A, N, Q, G, H, I, L, M, F, P, W, T, S)} \\ &= 10^{kC} / (10^{p} + 10^{kC}) - 10^{p} / (10^{p} + 10^{kN}) + 10^{k[i]} / (10^{p} + 10^{k[i]}); && i = 2, 9, 12 \text{ (R, H, K)} \\ &= 10^{kC} / (10^{p} + 10^{kC}) - 10^{p} / (10^{p} + 10^{kN}) - 10^{p} / (10^{p} + 10^{k[i]}); && i = 4, 5, 7, 19 \text{ (D, C, E, Y)} \end{aligned} \qquad (7)$$

Table 2 shows the charge multiplier for each of the 20 amino acids for pH values 7.0, 9.0, and 5.0. The charge on an amino acid is C[i]q, where $q = 1.619 \times 10^{-19}$ (coulomb) is the electron charge.

4. Statistics of translocation of amino acids

Substituting for L, μ and D, the statistics of the particle in *cis*1/*trans*2 and DNP can be calculated, where the values of D and μ for an amino acid aa with charge multiplier C_{aa} are obtained using

$$D_{aa} = k_BT_R/6\pi\eta R_{aa} \qquad \mu_{aa} = C_{aa}q/6\pi\eta R_{aa} \qquad (8)$$

Here k_B is the Boltzmann constant (1.3806×10^{-23} J/K), T_R is the room temperature (298° K), η is the solvent viscosity (0.001 Pa.s), R_{aa} the hydrodynamic radius of an amino acid (usually given in angstrom or Å), q is the electron charge (1.619×10^{-19} coulomb), and aa stands for the i-th amino acid listed in Section 3 above. D values are given in Table 1, while μ values (which depend on the charge multiplier) are given in Table 2. Table 4 shows translocation statistics for the 20 amino acids in DNP and *trans*1/*cis*2. Values for R_{aa} are taken from [3], where they are described as having been calculated from experimentally obtained values of the diffusion constants D_{aa}.

5. Conditions for entry of residues into DNP and residue occupancy of DNP

Two conditions need to be satisfied for accurate sequencing:

a) cleaved residues must enter DNP in natural order;

b) no more than one residue may occupy DNP at any time.

Both depend on the time taken by the exopeptidase to cleave the leading residue from the peptide. Since cleaving is a stochastic processs and will vary with the amino acid, let $T_{c.min\text{-}X}$ and $T_{c.max\text{-}X}$ be the minimum and maximum cleaving times for any amino acid X.

Condition (a). Let residue X_1 be cleaved at time t = 0. Its mean translocation time through *trans*1/*cis*2 is $E(T_{trans1/cis2\text{-}X1})$ and standard deviation is $\sigma_{trans1/cis2\text{-}X1}$. The next residue X_2 is cleaved no earlier than at $t = T_{c.min\text{-}X}$. Assuming 6σ support for the distribution, X_1 arrives at the entrance to DNP latest by $t = E(T_{trans1/cis2\text{-}X1}) + 3\sigma_{trans1/cis2\text{-}X1}$. The earliest that X_2 can arrive at DNP is $t = T_{c.min\text{-}X} + \max(0, E(T_{trans1/cis2\text{-}X2}) - 3\sigma_{trans1/cis2\text{-}X2})$. Therefore for X_2 to follow X_1 requires

$$E(T_{trans1/cis2\text{-}X1}) + 3\sigma_{trans1/cis2\text{-}X1} < T_{c.min\text{-}X} + \max(0, E(T_{trans1/cis2\text{-}X2}) - 3\sigma_{trans1/cis2\text{-}X2}) \quad (9)$$

Consider for example pH = 7.0 (results for the other pH values are very similar). From the data in Table 4, columns 2 and 3 (mean translocation time and standard deviation for *trans*1/*cis*2), $\max(0, E(T_{trans1/cis2\text{-}X2}) - 3\sigma_{trans1/cis2\text{-}X2}) = 0$ for any amino acid. Equation 7 reduces to

$$T_{c.min\text{-}X} > \max_X \{ E(T_{trans1/cis2\text{-}X}) + 3\sigma_{trans1/cis2\text{-}X} \} \quad (10)$$

over all X. From Table 4 (columns 2 and 3), the maximum occurs for X = K (Lys), with $E(T_{trans1/cis2\text{-}X}) = 0.86\times10^{-3}$ and $\sigma_{trans1/cis2\text{-}X} = 0.71\times10^{-3}$, leading to

$$T_{c.min} = 2.99 \text{ ms} \quad (11)$$

over all X.

Condition (b). Consider residue X_1 to be cleaved before X_2. Since condition (a) has to be satisfied, X_1 arrives at the entrance to DNP before X_2. Let it arrive at time t = 0. The latest it can exit DNP is at time $t = E(T_{DNP\text{-}X1}) + 3\sigma_{DNP\text{-}X1}$. The earliest that X_2 can arrive at the entrance of DNP is at $t = T_{c.min} + \max(0, E(T_{trans1/cis2\text{-}X2}) - 3\sigma_{trans1/cis2\text{-}X2}) = T_{c.min}$. Therefore for condition (b) to be satisfied

$$T_{c.min} > E(T_{DNP\text{-}X1}) + 3\sigma_{DNP\text{-}X1} \quad (12)$$

From Table 4 (columns 4 and 5), the maximum of the right hand side in Equation 10 occurs once again for X_1 = K (Lys), with $E(T_{DNP\text{-}X1}) = 14.95\times10^{-6}$ and $\sigma_{DNP\text{-}X1} = 14.89\times10^{-6}$, leading to $T_{c.min} = 5.96\times10^{-5}$ s, which is less than the value in Equation 11. Since Equation (11) has to be satisfied, it sets the minimum cleaving interval for any amino acid.

6. Sample size requirements for reliable residue identification and confidence levels for a given sample size

The two time-based discriminators discussed above are mean values. To obtain a sample mean value which approaches the population (that is, calculated) mean for amino acid X, sequencing has to be done N (= sample size) times to distinguish the sample mean of X from that for another amino acid Z. The value of N, which depends on how close the mean translocation times of two amino acids are and the desired confidence level, can be calculated using standard formulas from statistics. Thus with a population mean E and standard deviation σ, margin of error e, and confidence level α (equivalently the percentile value = 1 - α/2), the critical value $Z_{\alpha/2}$ of the normal distribution can be obtained from tables or calculated using statistical software (R was used in the present work). For example, with a confidence level of 0.95, α is 0.05, the percentile is 97.5, and the critical value is 1.96. The number of samples required for the sample mean E to approach the population mean within error e is

$$N = Z_{\alpha/2}^2 \, \sigma^2 / e^2 \quad (13)$$

For pH = 9.0, Tables 5 and 6 in the Appendix give the required sample sizes for DNP and *trans*1/*cis*2 for each amino acid X and its nearest neighbor (that is, the amino acid Z whose mean is closest to the mean of X) for three confidence levels: 90%, 80%, 70%. σ is taken from Table 4, e is set to k × min | E_X - E_Z | where Z is the amino acid in column 6 or 8 with mean E_Z nearest to the mean E_X for X, and k < 0.5. (This nearest neighbor can in most cases be identified visually in Figures 2 and 3, where the amino acids separate into ordered groups.) Figures 5 and 6 show histograms of the sample size for DNP and *trans*1/*cis*2 respectively for k = 0.4.

The value of N to use in the sequencing is the largest sample size N_{max} over all the amino acids. In determining N_{max} the discriminator to use for an amino acid is based on the smallest number of samples over all its discriminators. For example, with pH = 9.0, Asn (symbol N) has $E(T_{DNP}) = \sim 0.191\times10^{-6}$ which is 0.0038×10^{-6} from the mean time of Thr (T) and requires ~23000 samples for a confidence level of 90%. It has $E(T_{trans1/cis2}) = 0.68\times10^{-3}$ which is 0.0049×10^{3} from the mean time of Asp (D) and requires > 200000 samples. The discriminator to use for Asn is therefore $E(T_{DNP})$.

As noted in the main text, amino acid pairs whose mean times are very close to each other are the ones that effectively determine N_{max}. For DNP the problem pairs are His (H) - Trp (W), Gln (Q) - Ile (I), Met (M) - Tyr (Y), and Ala (A) – Pro (P), with N in the range 1 to 6 million; in the case of *trans*1/*cis*2 they are Glu (E) - Met (M), His (H) - Trp (W), and Gln (Q) - Ile (I), with N in the range 1 to 11 million. By excluding them from the determination of N_{max} and using error correction procedures to circumvent the resulting low confidence levels, N_{max} can be brought down significantly.

Conversely for a given maximum number of samples N_{max} one can find the confidence level for the sample mean of an amino acid X to be no farther from the population mean than e = k × min | E_X - E_Z |, where e is the distance to the nearest mean, with k < 0.5. This can be obtained from the critical value using the statistical formula

$$Z_{\alpha/2} = (e/\sigma)\sqrt{N_{max}} \qquad (14)$$

and tables (or the *pnorm* function in the software package R). For example, with pH = 9.0 and N = 10000 in DNP, consider X = A (Ala) with $\sigma = 0.13\times10^{-6}$. Its nearest mean neighbor Z = P (Pro) with distance to mean of $Z = 0.0013\times10^{-6}$. With k = 0.4 the resulting critical value $Z_{\alpha/2} = 0.40$, for which the confidence level is 0.43 (43%). Table 7 gives the confidence levels for the 20 amino acids for k = 0.4 and N = 10000 in DNP and in *trans*1/*cis*2.

7. Additional notes

This section expands on issues that were not addressed in the main text or were considered in an abbreviated form.

1) *Behavior of a particle at the interface between two sections in the tandem cell*. Residues at the interface between *trans*1/*cis*2 and DNP experience a drift field inside both regions that depends on their net charge. Using formal probabilistic arguments [1] it can be shown that with sufficiently large V_{05} a residue with a substantial negative charge will eventually pass into DNP, such passage being aided indirectly by the reflecting boundaries in *trans*1/*cis*2. The behavior at the interface between DNP and *trans*2 is similar. The tapered geometry of *trans*1/*cis*2 shown in Figure 1 in the main text aids passage into DNP. Similarly the abrupt increase in cross-section from DNP to *trans*2 decreases the probability of a detected particle regressing into DNP from *trans*2. Residues that are substantially positive experience a negative drift field inside both regions. Because of this there is a non-zero probability that such a residue may ultimately not enter DNP and therefore may be 'lost' to diffusion in *trans*1/*cis*2. Also on entering DNP it may be trapped inside and neither regress into *trans*1/*cis*2 nor exit into *trans*2. A solution to the first problem that is based on repeating the sequencing with the voltage reversed was considered in the main text. A second possible solution is to design the pore lumen so as to prevent regression of the residue once it has entered DNP. One can also consider use of a hydraulic pressure gradient to prevent entry into DNP; however the hydrodynamic radius of an amino acid is too small for the pressure to be comparable to the electric field. (Compare this with the behavior of polyethylene glycol (PEG) in a nanopore with combined electric field and hydraulic pressure gradients [4]: 12 kDa PEG molecules with a length of 0.35 nm have a hydrodynamic radius of 3.2 nm, which is ~10 × average radius for an amino acid [3].) At the interface between *trans*1/*cis*2 and DNP residues that have very little charge are not affected by the electric field in either region. They are subject entirely to diffusion. In this case the tapered geometry of *trans*1/*cis*2 in Figure 1 is useful in promoting entry from *trans*1/*cis*2 into DNP and also reduces the probability of permanent regression into *trans*1/*cis*2 from DNP. Although a hydraulic gradient could be used to assist entry into DNP, the improvement is minimal because its effect is small for reasonable values of hydraulic pressure in the range 5-10 atm for solid-state membranes [4] (1 atm = 1.01325×10^{5} Pa.). The behavior at the interface between DNP and *trans*2 can be similarly understood, combined with the fact that the abrupt change in diameter from DNP to *trans*2 acts as a deterrent to regression from *trans*2 into DNP. (The behavior is also impacted by the pH value. Looking at Table 3, a pH value of 9.0 results in 18 of the 20 amino acids having (significant) negative charge and only two, Arg (R) and Lys (K), being positive. However, this has to be considered along with the effect that the pH value has on exopeptidase efficiency. See discussion in the main text.)

2) *Ensuring entry of the correct end of a peptide into UNP*. If the incorrect end (amino or carboxy) has entered UNP this will be known from the absence of characteristic blockades; the peptide remains intact when it enters *trans*2. It can be recycled to *cis*1 for another attempt at detection, this to be repeated until residue-driven blockades are detected. With two identical copies of the peptide, two sequencers, one with amino exopeptidase and the other with carboxy, can be used to increase the probability of successful sequencing. An alternative approach that dispenses with any dependence on the peptide's random orientation when entering DNP may be based on two tandem cells in tandem, the first with amino exopeptidase and the second with carboxy. The device would then have the structure [*cis*1, UNP with amino exopeptidase, *trans*1/*cis*2, DNP with carboxypeptidase, *trans*2/*cis*3, third (sensing) nanopore (TNP), *trans*3]. To guarantee detection in the second stage of a peptide that was not sequenced in DNP because it entered UNP C-terminal first, the unsequenced polymer has to enter DNP C-terminal first. This can be ensured if the poly-X leader (which entered UNP C-terminal first) is longer than the length of *trans*1/*cis*2 so that the trailing polymer is still inside UNP and the leader (with its free C-terminal in front) enters

DNP C-terminal first. (High enough voltages that are within the breakdown limit may ensure such entry. Up to 0.7 V can be applied across a biological nanopore of length 10 nm [1].) This ensures that the leading residue is cleaved by the carboxypeptidase attached to the downstream side of DNP. When sequencing occurs in the first stage spurious signals from cleaved residues that try to enter TNP after detection in DNP can be avoided by flushing them out after they have entered *trans*2 (thus effectively deactivating TNP). Yet another possible, and somewhat simpler, alternative (although it requires an additional step) is to attach a capping molecule (similar to a biotin-streptavidin tether [5]) to the trailer at either the C-end or N-end of the peptide to prevent that end from entering UNP.

3) *N× sequencing with one copy of the peptide*. This may be possible by recycling the cleaved residues after their detection in the tandem cell back into *cis*1 for translocation through UNP and *trans*1/*cis*2 to DNP for another round of detection. This recycling can be done N_{max} times; this assumes that the recycled residues are not affected by the exopeptidase attached to UNP. For short peptides the value of N_{max} can be set adaptively after the first few sample runs have yielded a tentative sequence. The possibility of a tandem cell with recycling capability that uses a hydraulic gradient to 'pump' detected residues back to *cis*1 is currently being examined.

4) *Sequencing a whole protein*. A folded protein could be loaded into the tandem cell and unfolded by an unfoldase enzyme [6] like ClpX before cleaving and sequencing. The unfoldase, which acts as a motor that both unfolds and translocates the protein, could be attached to the upstream side of UNP in *cis*1 so that the protein enters UNP unfolded and is then cleaved by the exopeptidase attached to the downstream side of UNP. Alternatively the unfoldase could be attached to the downstream side (similar to [6]) of a precursor nanopore in a double tandem cell with the structure [*cis*0, precursor UNP with ClpX, *trans*0/*cis*1, UNP with exopeptidase, *trans*1/*cis*2, DNP, *trans*2]. Here the unfolded protein translocates to UNP after it has passed through ClpX, following which its behavior would be similar to that in the basic tandem cell. Such a modified tandem cell may be used to sequence a whole protein.

5) *Transverse recognition tunneling and the tandem cell*. If an amino acid can be uniquely identified by a transverse recognition tunneling (RT) current [7], a cascade of 21 nanopores could be used to fully sequence a peptide. In such a tandem cascade the first tandem stage is used to cleave residues in the peptide, followed by 20 pores each one of which is designed to recognize a unique amino acid. Such a system can sequence a peptide without having to depend on ionic current blockades and the extreme measurement precision required to distinguish among their closely spaced values in the presence of noise. Alternatively a single DNP with 20 recognizers and 20 pairs of transverse electrodes may also be possible. In either case the length of the pore is no longer a crucial issue as it is in most nanopore sequencing approaches. Correlations among the 20 transverse current records can be used not only to improve residue calling accuracy but also to extract other kinds of peptide-related information. The order of the recognizers may also be optimized to maximize discrimination among the residues.

6) *Recovering the original peptide?* The tandem cell approach to peptide sequencing as described above is a destructive process as the peptide is broken down into its constituent amino acids. Unlike exonuclease-based DNA sequencing, where re-sequencing of the original strand from the cleaved bases can be done using the individual cleaved nucleotides and a template with an enzyme motor attached to a nanopore [8], there is no simple way to re-synthesize the peptide that can be integrated with the tandem cell.

7) *Other two-pore systems in peptide analysis*. There appears to be one other reported instance in the literature of a system with twin nanopores for protein analysis. In [9] two nanopores in series are used to measure mobility and particle sizes to identify specific proteins. The nanopores are comparatively larger, with cross-section dimensions that are several 10's of nm. The system is structurally and procedurally different from the tandem cell described here. (For two-pore systems used in DNA sequencing see Supplement to [1].)

References

[1] G. Sampath, "A tandem cell for nanopore-based DNA sequencing with exonuclease," *RSC Adv.*, 2015, **5**, 167-171.

[2] D. L. Nelson and M. M. Cox, *Lehninger's Principles of Biochemistry*, 4th Edition, W. H. Freeman and Company, New York, 2005.

[3] M. W. Germann, T. Turner, and S. A. Allison, "Translational diffusion constants of the amino acids: measurement by NMR and their use in modeling the transport of peptides," *J. Phys. Chem. A*, 2007, **111**, 1452-1455.

[4] B. Lu, D. P. Hoogerheide, Q. Zhao, H. Zhang, Z Tang, D. Yu , and J. A. Golovchenko, "Pressure-controlled motion of single polymers through solid-state nanopores," *Nano Lett*., 2013, **13**, 3048–3052.

[5] L. Movileanu, S. Howorka, O. Braha, and H. Bayley, "Detecting protein analytes that modulate transmembrane movement of a polymer chain within a single protein pore,"*Nature Biotech*., 2000, **18**, 1091-1095.

[6] J. Nivala, D. B. Marks and M. Akeson, "Unfoldase-mediated protein translocation through an a-hemolysin nanopore," *Nature Biotech*., 2013, **31**, 247–250.

[7] Y. Zhao, B. Ashcroft, P. Zhang, H. Liu, S. Sen, W. Song, J. Im, B. Gyarfas, S. Manna, S. Biswas, C. Borges, and S. Lindsay, "Single-molecule spectroscopy of amino acids and peptides by recognition tunneling," *Nature Nanotech*., 2014, 9, 466–473.

[8] K. R. Lieberman, G. M. Cherf, M. J. Doody, F. Olasagasti, Y. Kolodji, and M. Akeson, "Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase," *J. Am. Chem. Soc*., 2010, **132**, 17961–17972.

[9] Z. D. Harms, D. G. Haywood, A. R. Kneller, L. Selzer, A. Zlotnick, and S. C. Jacobson, "Single-particle electrophoresis in nanochannels," *Anal. Chem*., 2014, Article ASAP. (DOI: 10.1021/ac503527d)

Next page: Appendix

Appendix

Table 1 - Hydrodynamic radii and diffusion constants

AA	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
R_{aa} [(a)]	2.66	3.60	2.98	3.02	2.86	3.23	3.14	2.32	3.49	3.24	3.39	3.69	3.08	3.35	2.68	2.76	3.04	3.50	3.57	3.32
D_{aa} [(b)]	8.21	6.06	7.32	7.23	7.63	6.76	6.95	9.41	6.25	6.74	6.44	5.91	7.09	6.52	8.14	7.91	7.18	6.24	6.11	6.57

AA= Amino acid   [(a)] Values (10^{-10} m) from [3]   [(b)] Values (10^{-8} m/Vs) computed from Equation 8

Table 2 - Charge multiplier and mobility for three values of pH

pH	AA	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
7.0	Multiplier	-0.0020	0.9980	-0.0020	-1.0013	-0.0467	-0.0020	-1.0002	-0.0020	0.0889	-0.0020	-0.0020	0.9977	-0.0020	-0.0020	-0.0020	-0.0020	-0.0020	-0.0020	-0.0029	-0.0020
	Mobility	-0.0064	2.3560	-0.0057	-2.8177	-0.1388	-0.0053	-2.7072	-0.0074	0.2165	-0.0053	-0.0051	2.2978	-0.0056	-0.0051	-0.0064	-0.0062	-0.0056	-0.0049	-0.0068	-0.0052
9.0	Multiplier	-0.1696	0.8301	-0.1696	-1.1695	-0.9934	-0.1696	-1.1695	-0.1696	-0.1686	-0.1696	-0.1696	0.8018	-0.1696	-0.1696	-0.1696	-0.1696	-0.1696	-0.1696	-0.2480	-0.1696
	Mobility	-0.5417	1.9597	-0.4835	-3.2912	-2.9520	-0.4461	-3.1654	-0.6211	-0.4105	-0.4447	-0.4251	1.8466	-0.4678	-0.4301	-0.5377	-0.5221	-0.4740	-0.4117	-0.5904	-0.4340
5.0	Multiplier	0.0022	1.0022	0.0022	-0.9303	0.0017	0.0022	-0.8469	0.0022	0.9113	0.0022	0.0022	1.0022	0.0022	0.0022	0.0022	0.0022	0.0022	0.0022	0.0022	0.0022
	Mobility	0.0069	2.3658	0.0062	-2.6179	0.0050	0.0057	-2.2921	0.0079	2.2190	0.0057	0.0054	2.3081	0.0060	0.0055	0.0069	0.0067	0.0060	0.0053	0.0051	0.0055

Charge multiplier values computed from Equation 7, charge on amino acid AA = multiplier $\times 1.619 \times 10^{-19}$ coulomb; mobility (10^{-10} m^2/s) from Equation 8

Table 3 - Volume exclusion ratio

AA	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
VER	0.0863	0.2196	0.1223	0.1274	0.1078	0.1568	0.1437	0.0568	0.1993	0.1583	0.1821	0.2371	0.1354	0.1756	0.0883	0.0966	0.1300	0.2011	0.2139	0.1707

VER = Volume exclusion ratio, computed from Equation 5; R_{aa} from Table 1; L = L_{34} = 10 nm, r = 1.5 nm

Table 4 - Statistics of translocation times in DNP and *trans*1/*cis*2 for three values of pH

	pH = 7.0				pH = 9.0				pH = 5.0			
Amino acid	**Mean trans-location time in *trans*1/*cis*2 [(a)] (10^{-3} s)**	**Std deviation of translocation time in *trans*1/*cis*2 [(h)] (10^{-3} s)**	**Mean trans-location time in DNP [(c)] (10^{-6} s)**	**Std deviation of trans-location time in DNP [(d)] (10^{-6} s)**	**Mean trans-location time in *trans*1/*cis*2 [(a)] (10^{-3} s)**	**Std deviation of trans-location time in *trans*1/*cis*2 [(h)] (10^{-3} s)**	**Mean trans-location time in DNP [(c)] (10^{-6} s)**	**Std deviation of trans-location time in DNP [(d)] (10^{-6} s)**	**Mean trans-location time in *trans*1/*cis*2 [(a)] (10^{-3} s)**	**Std deviation of trans-location time in *trans*1/*cis*2 [(h)] (10^{-3} s)**	**Mean trans-location time in DNP [(c)] (10^{-6} s)**	**Std deviation of trans-location time in DNP [(d)] (10^{-6} s)**
Ala A	0.609373	0.497551	0.242606	0.197900	0.607233	0.495455	0.170212	0.127737	0.609373	0.497551	0.244986	0.200232
Arg R	0.842077	0.690397	14.602984	14.548098	0.839119	0.687496	6.422961	6.364767	0.842151	0.690470	14.914867	14.860067
Asn N	0.682681	0.557407	0.271791	0.221707	0.680284	0.555058	0.190689	0.143104	0.682681	0.557407	0.274457	0.224320
Asp D	0.677680	0.551018	0.067638	0.033834	0.675343	0.548730	0.059285	0.027816	0.678670	0.551987	0.071886	0.037017
Cys C	0.654556	0.534344	0.235677	0.188191	0.641881	0.521927	0.064478	0.032355	0.655191	0.534961	0.263117	0.215004
Gln Q	0.739953	0.604169	0.294592	0.240307	0.737355	0.601623	0.206686	0.155110	0.739953	0.604169	0.297482	0.243139
Glu E	0.704623	0.572927	0.070387	0.035224	0.702178	0.570534	0.061641	0.028922	0.706850	0.575107	0.080662	0.043045
Gly G	0.531484	0.433954	0.211596	0.172604	0.529617	0.432126	0.148456	0.111410	0.531484	0.433954	0.213672	0.174638
His H	0.800994	0.654251	0.398006	0.338117	0.796725	0.650068	0.223751	0.168003	0.814864	0.667847	9.198090	9.143185
Ile I	0.742244	0.606040	0.295504	0.241051	0.739638	0.603486	0.207326	0.155590	0.742244	0.606040	0.298403	0.243891
Leu L	0.776607	0.634097	0.309185	0.252211	0.773880	0.631425	0.216925	0.162793	0.776607	0.634097	0.312218	0.255183
Lys K	0.863124	0.707652	14.946021	14.889757	0.859587	0.704183	5.763911	5.703731	0.863205	0.707731	15.287508	15.231338
Met M	0.705590	0.576112	0.280912	0.229147	0.703112	0.573684	0.197088	0.147906	0.705590	0.576112	0.283667	0.231847
Phe F	0.767444	0.626615	0.305537	0.249235	0.764749	0.623975	0.214365	0.160872	0.767444	0.626615	0.308534	0.252172
Pro P	0.613955	0.501292	0.244430	0.199388	0.611799	0.499180	0.171492	0.128698	0.613955	0.501292	0.246828	0.201737
Ser S	0.632282	0.516256	0.251726	0.205340	0.630062	0.514081	0.176611	0.132540	0.632282	0.516256	0.254196	0.207759
Thr T	0.696427	0.568630	0.277263	0.226171	0.693981	0.566234	0.194528	0.145986	0.696427	0.568630	0.279983	0.228836
Trp W	0.801807	0.654673	0.319218	0.260395	0.798991	0.651914	0.223964	0.168076	0.801807	0.654673	0.322349	0.263463
Tyr Y	0.817843	0.667766	0.324958	0.264971	0.813647	0.663656	0.197938	0.142535	0.817843	0.667766	0.328790	0.268726
Val V	0.760571	0.621004	0.302801	0.247003	0.757900	0.618387	0.212445	0.159432	0.760571	0.621004	0.305771	0.249913

(a), (b), (c), (d)   Values computed from Equations 1-4

Table 5 - Sample sizes for three confidence levels for DNP, pH = 9.0

DNP		Confidence level = 0.9		Confidence level = 0.8		Confidence level = 0.7	
Amino Acid	**Nearest Amino Acid [a]**	**Difference in means (10^{-6} s)**	**Sample size [b]**	**Difference in means (10^{-6} s)**	**Sample size [b]**	**Difference in means (10^{-6} s)**	**Sample size [b]**
A	P	0.001280	168458	0.001280	102261	0.001280	66883
R	K	0.659050	1577	0.659050	957	0.659050	626
N	T	0.003839	23491	0.003839	14260	0.003839	9327
D	E	0.002356	2356	0.002356	1430	0.002356	935
C	E	0.002837	2199	0.002837	1335	0.002837	873
Q	I	0.000640	993560	0.000640	603131	0.000640	394477
E	D	0.002356	2547	0.002356	1546	0.002356	1011
G	A	0.021756	443	0.021756	269	0.021756	176
H	W	0.000212	10571896	0.000212	6417570	0.000212	4197403
I	Q	0.000640	999721	0.000640	606871	0.000640	396923
L	F	0.002560	68401	0.002560	41522	0.002560	27157
K	R	0.659050	1266	0.659050	768	0.659050	502
M	Y	0.000850	512330	0.000850	311005	0.000850	203412
F	V	0.001920	118750	0.001920	72086	0.001920	47148
P	A	0.001280	171001	0.001280	103804	0.001280	67893
S	P	0.005119	11335	0.005119	6880	0.005119	4500
T	M	0.002560	55006	0.002560	33391	0.002560	21839
W	H	0.000212	10580977	0.000212	6423082	0.000212	4201008
Y	M	0.000850	475793	0.000850	288825	0.000850	188906
V	F	0.001920	116633	0.001920	70801	0.001920	46307

[a] Amino acid with closest mean translocation time [b] Computed from Equation 13

Table 6 - Sample sizes for three confidence levels for *trans*1/*cis*2, pH = 9.0

*trans*1/*cis*2		Confidence level = 0.9		Confidence level = 0.8		Confidence level = 0.7	
Amino Acid	**Nearest Amino Acid [a]**	**Difference in means (10^{-3} s)**	**Sample size [b]**	**Difference in means (10^{-3} s)**	**Sample size [b]**	**Difference in means (10^{-3} s)**	**Sample size [b]**
A	P	0.004566	199129	0.004566	120879	0.004566	79061
R	K	0.020468	19078	0.020468	11581	0.020468	7574
N	D	0.004941	213398	0.004941	129541	0.004941	84726
D	N	0.004941	208560	0.004941	126604	0.004941	82805
C	S	0.011819	32975	0.011819	20017	0.011819	13092
Q	I	0.002283	1174455	0.002283	712941	0.002283	466298
E	M	0.000934	6305602	0.000934	3827756	0.000934	2503538
G	A	0.077616	524	0.077616	318	0.077616	208
H	W	0.002266	1391248	0.002266	844544	0.002266	552372
I	Q	0.002283	1181738	0.002283	717363	0.002283	469190
L	F	0.009131	80855	0.009131	49082	0.009131	32102
K	R	0.020468	20015	0.020468	12150	0.020468	7946
M	E	0.000934	6375422	0.000934	3870140	0.000934	2531259
F	V	0.006848	140371	0.006848	85211	0.006848	55732
P	A	0.004566	202134	0.004566	122703	0.004566	80254
S	C	0.011819	31991	0.011819	19420	0.011819	12701
T	E	0.008197	80688	0.008197	31991	0.008197	32036
W	H	0.002266	1399162	0.002266	849348	0.002266	555515
Y	W	0.014656	34671	0.014656	21047	0.014656	13765
V	F	0.006848	137868	0.006848	83691	0.006848	54738

[a] Amino acid with closest mean translocation time [b] Computed from Equation 13

Table 7 - Confidence levels for DNP and *trans*1/*cis*2 for a sample size of 10000, pH = 9.0

Amino Acid	DNP				*trans*1/*cis*2			
	Nearest Amino Acid [a]	Difference in means (10^{-6} s)	$Z_{a/2}$ [b]	Confidence level	Nearest Amino Acid [a]	Difference in means (10^{-3} s)	$Z_{a/2}$ [b]	Confidence level
A	P	0.001280	0.400757	0.429120	P	0.004566	0.368604	0.397832
R	K	0.659050	4.141863	1.000000	K	0.020468	1.190856	0.907842
N	T	0.003839	1.073168	0.870907	D	0.004941	0.356067	0.385426
D	E	0.002356	3.388176	0.999998	N	0.004941	0.360173	0.389501
C	E	0.002837	3.506890	0.999999	S	0.011819	0.905797	0.799803
Q	I	0.000640	0.165018	0.184526	I	0.002283	0.151778	0.169957
E	D	0.002356	3.258653	0.999996	M	0.000934	0.065503	0.073807
G	A	0.021756	7.811308	1.000000	A	0.077616	7.184600	1.000000
H	W	0.000212	0.050588	0.057034	W	0.002266	0.139452	0.156341
I	Q	0.000640	0.164508	0.183966	Q	0.002283	0.151310	0.169441
L	F	0.002560	0.628917	0.626224	F	0.009131	0.578458	0.586679
K	R	0.659050	4.621886	1.000000	R	0.020468	1.162636	0.899868
M	Y	0.000850	0.229801	0.254810	E	0.000934	0.065144	0.073403
F	V	0.001920	0.477320	0.500345	V	0.006848	0.439024	0.465317
P	A	0.001280	0.397766	0.426242	A	0.004566	0.365853	0.395119
S	P	0.005119	1.544947	0.971103	C	0.011819	0.919621	0.806584
T	M	0.002560	0.701325	0.678716	E	0.008197	0.579056	0.587162
W	H	0.000212	0.050567	0.057010	H	0.002266	0.139057	0.155904
Y	M	0.000850	0.238461	0.264060	W	0.014656	0.883364	0.788432
V	F	0.001920	0.481633	0.504212	F	0.006848	0.442991	0.469002

[a] Amino acid with closest mean translocation time [b] Critical value of normal distribution (Equation 14)

Table 8 - Statistics of nucleotides (for comparison)

Base	Nucleotide volume [a] V_N (10^{-30} m^3)	Hydrodynamic radius [b] R_{aa} (10^{-10} m)	Diffusion coefficient [c] D_{aa} (10^{-10} m^2/s)	Mobility [d] μ_{aa} (10^{-8} m/Vs)	Mean translocation time in DNP [e] (10^{-6} s)	Std deviation of translocation time in DNP [f] (10^{-6} s)	Mean translocation time in *trans*1/*cis*2 [g] (10^{-3} s)	Std deviation of translocation time in *trans*1/*cis*2 [h] (10^{-3} s)	Volume exclusion ratio [j]
A	349	4.878957	4.473436	1.741885	0.019994	0.019919	1.141287	0.935719	0.412390
T	339	4.808550	4.538936	1.767390	0.019705	0.019632	1.124817	0.922216	0.399273
C	324	4.700962	4.642815	1.807839	0.019264	0.019193	1.099650	0.901582	0.379756
G	359	4.948362	4.410692	1.717454	0.020278	0.020203	1.157522	0.949030	0.425593

(a) Volumes in column 2 from: M. Zwolak and M. DiVentra, "Physical approaches to DNA sequencing and detection," *Rev. Mod. Phys.*, 2008, **80**, 141-165.
(b) Calculated from ellipsoid of length 7 Å (= length of stretched mononucleotide, same for all 4 types) and circular cross-section from volume in column 2
(c), (d) Values computed from Equation 8 (e), (f), (g), (h) Values computed from Equations 1-4 (V_{23} = 1.6 mV, V_{34} = 0.18 V)
(j) Values computed from Equation 5 (DNP: L = L_{34} = 10 nm, r = 1.5 nm; *trans*1/*cis*2: L = L_{23} = 1 µm, r = 0.5 µm)

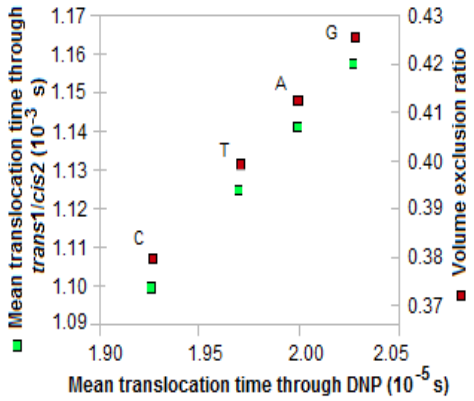
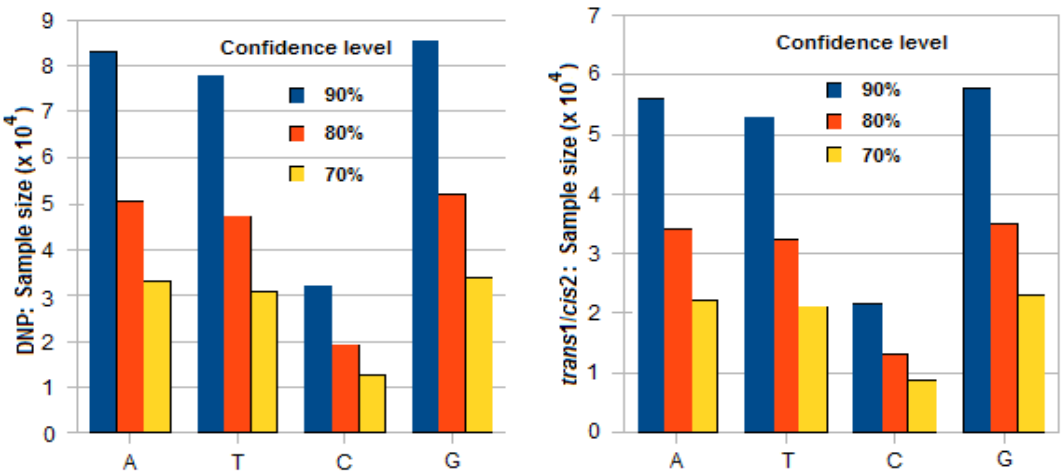


Figure 7. *Scatter chart of mean translocation time through DNP and mean translocation time through trans1/cis2 for nucleotides*



(a) DNP (b) *trans*1/*cis*2

Figure 8. *Histograms of sample sizes for three confidence levels*

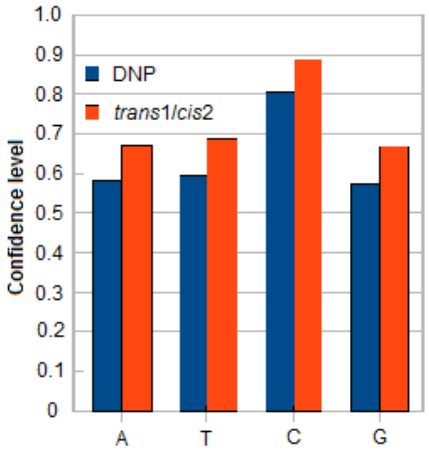


Figure 9. *Histograms of confidence levels for sample size = 10000*