Single molecule protein sequencing based on the superspecificity of tRNA synthetases

G. Sampath
P. B. 7849, J. P. Nagar P. O., Bengaluru-560078, India

Single molecule *de novo* protein sequencing based on the 'superspecificity' of amino-acyl tRNA synthetases (aaRS) is proposed. An unfolded protein molecule is threaded through a nanopore in an electrolytic cell (e-cell) to expose the terminal residue in the e-cell's *trans* chamber. After the residue is cleaved with an exopeptidase, a set of tRNAs, their aaRSs, and ATP are added to *trans*. An aaRS charges a cognate tRNA molecule with the residue. The charged tRNA (along with the other reactants) is transferred to an extended e-cell with N ($20 \leq N \leq 61$) pores in N individual *cis* chambers and a single *trans* chamber. Each pore holds an RNA molecule ending in a unique codon that is exposed in *trans*. The charged tRNA's anticodon base-pairs with the terminal codon of an RNA. If tRNAs and residues are fluorescently tagged with two different colors, the residue can be identified from the observed position of the resulting color pair. As charging is 'superspecific' identification is unambiguous. The protein molecule in the first e-cell is advanced by one residue and the process repeated. In this approach there is no need for precise pore current or optical intensity measurements. Potential implementation issues are discussed. Other possibilities, including one in which the terminal residue is cleaved after charging, are also examined.

Keywords: protein sequencing; single molecule; nanopore; tRNA; amino acyl tRNA synthetase; codon; optical tag

1. Introduction

Biopolymer sequencing occupies a central place in the study of biological organisms and is usually focused on research or the diagnostic and therapeutic value of the information contained in a polymer (DNA, RNA, protein) sequence. DNA sequencing is now a mature area [1] and is accomplished by a number of technologies, starting with Sanger sequencing and proceeding to next generation techniques that make use of highly automated processes [2]. On the other hand protein sequencing is not as advanced, largely because there are 20 amino acids to work with, compared to four bases with DNA. It is currently done with one of the following methods: Edman degradation (ED), gel electrophoresis (GE), and mass spectrometry (MS) [3-5]. The current state of protein sequencing research is reviewed in [6].

More recently nanopores have been used to sequence DNA [7]. Briefly, an electrolytic cell [8] or e-cell consists of two chambers, named *cis* and *trans*, which contain an electrolyte, typically NaCl or KCl, and are bridged by a thin membrane containing a nanopore. An electrical potential V applied between *cis* and *trans* ionizes the electrolyte and leads to current flow through the pore. With KCl as the electrolyte, K^+ ions flow toward the cathode and Cl^- toward the anode. If a charged analyte molecule (DNA, RNA, or protein) is introduced into *cis*, it translocates through the pore in an appropriate direction. Both DNA and RNA carry a uniform negative charge along their backbone, so they flow from *cis* to *trans* in Figure 1 below. In contrast, seven of the standard 20 amino acids, namely D, E, K, R, H, C, and Y, carry a negative or positive charge whose value depends on the pH of the electrolyte [9]. This means that the net charge carried by a protein depends on the sequence of residues and also determines the direction of flow of the protein. Crucially, translocation of DNA, RNA, or protein through the pore causes a reduction in the pore current that varies with the volume of the monomer passing through the pore. This monomer-specific level of current blockade serves as the basis for current approaches to nanopore-based sequencing. Such methods are beginning to show promise and are set to compete with the more established methods mentioned above [10]. In contrast, nanopore-based protein sequencing is still in its infancy; a variety of methods, some in theory [11-14], others in practice [15-18], are known. Recently it has been reported that 13 of the 20 standard amino acids (obtained by finely grinding a protein) riding on a carrier molecule translocating through an aerolysin nanopore have long enough dwell times in the pore to allow discrimination among them based on the volume that each of them excludes in the pore [19]. This work is reminiscent of earlier theoretical work in which residues in a peptide are sequentially cleaved by an exopeptidase on the output side of the first of two pores in series (tandem pore) and then identified by the volume they exclude when they translocate through the second pore [11].

1.1 *The present work*

This report proposes a protein sequencing method in which the sequence of amino acids is obtained indirectly from a sequence of codons that code for the amino acids. An e-cell with a nanopore is used with an electrical potential (or other method, see Section 4) to hold the unfolded linear protein sequence in place while the terminal residue is exposed in *trans*. If the terminal residue is cleaved with an exopeptidase, and a set of tRNAs, their amino-acyl tRNA synthetases (aaRSs), and ATP are added to *trans*, a matching tRNA gets 'charged' with the residue by the corresponding aaRS. This charged tRNA (or aminoacyl tRNA) is transferred to a second modified e-cell with multiple nanopores each of which holds an RNA molecule ending in a codon coding for an amino acid. The anticodon in the tRNA binds with one of these RNA molecules by base-pairing with the terminal codon of the latter. With the tRNAs and residues fluorescently tagged [20] with two different colors, the occurrence of tRNA charging in the first e-cell and the identity of the residue in the second can both be detected optically. As the charging of a tRNA by a related aaRS is highly specific (the commonly used term is 'superspecific') [21-22], residue identification is unambiguous. The protein molecule in the first e-cell is advanced by one residue through the pore, and the process repeated until the protein has been sequenced. Several variations of this general scheme are also considered. In one of them a tRNA is charged with the exposed residue before it is cleaved; there are two versions, one with optical tagging and the other without.

This approach mimics the way mRNA is translated to protein *in vivo* but without involving the complexities present in the latter. E-cells with nanopores provide a controlled environment in which enzymes and processes similar to those that occur during translation can be used. Fluorescent tagging of residues and tRNAs enables the detection of salient events to drive the process in a controlled manner. Importantly precise measurements of optical intensities or nanopore currents are not necessary. Thus in the former case detection of fluorescent spots in the image is sufficient, in the latter case only the occurrence of a current blockade needs to be detected. Because no *a priori* information about the protein is used the proposed method is capable of *de novo* sequencing.

2. Designing a protein sequencing method based on the superspecificity of tRNAs

The protein sequencing procedure that is the subject of this communication is motivated by the superspecificity property of aaRSs in a biological cell: a tRNA always gets charged by a related aaRS with the amino acid associated with it; error rates are on the order of 1 in 10000 [22].

Consider how translation of mRNA into protein occurs *in vivo*, that is, in a biological cell. This is a complex process that involves a host of enzymes and other factors interacting with codon-specific tRNAs to facilitate the synthesis of a peptide sequence. The following is a simplified description, see [23] for a detailed description of the dynamics of translation at the molecular structure level.

There are three major steps in translation, each of which results in an amino acid being added to a growing peptide:

T1: The enzyme aaRS activates ATP and a cognate amino acid to form an amino acid AMP complex, charges the cognate tRNA with the amino acid at the latter's carboxyl end and releases the AMP, then releases the charged or aminoacyl tRNA. (This step is independent of translation: aminoacyl tRNAs are normally available in a cell for every one of the standard 20 amino acids for translation when the cell needs them.)

T2: An aminoacyl tRNA from Step T1 with anticodon matching the next codon along the mRNA binds with the latter by Chargaff base-pair matching.

T3: The aminoacyl tRNA from the residue previously attached to the growing peptide is cleaved and the incoming residue in an aminoacyl tRNA is attached to the peptide. This last step is part of a reaction that involves the enzyme peptidyl transferase and several other factors.

Steps T2 and T3 take place in the ribosome, which consists of two subunits (named 40S and 60S in eukaryotes, and jointly named 80S), which together provide a platform on which mRNA and tRNAs are brought together to enable protein synthesis. The 30S subunit is first bound to the mRNA. There are three tRNA binding sites in 80S: the A site holds an incoming aminoacyl-tRNA, the P site holds the tRNA attached to the last residue in the growing peptide (this tRNA is known as peptidyl-tRNA), and the E site provides an exit for tRNAs that have completed amino acid delivery. After an amino acid is added to the peptide, 80S moves along the mRNA to the next codon. Translation is error-free because tRNA charging is superspecific and base pairing of an anticodon with a matching codon is unambiguous. If any errors occur the enzymes involved uncharge the incorrect tRNA [21].

A protein sequencing procedure can be designed by borrowing and adapting elements of the above translation process. Consider the following three generic steps in protein sequencing, which are repeated for each successive residue in the protein:

G1: Hold protein molecule in fixed location where terminal residue can be exposed and cleaved.
G2: Identify residue.
G3: Advance protein molecule to expose next residue.

The three steps can be realized with the following three components H1, H2, and H3.

H1: An e-cell with a nanopore provides a platform for G1. The protein molecule is threaded through the pore and held in place by a voltage across the *cis* and *trans* chambers (and/or some other means, such as optical or magnetic tweezers, see Section 4) such that only the first residue is exposed in *trans*. An exopeptidase bound on the *trans* side to the membrane containing the pore (or, alternatively, free exopeptidase in solution) cleaves the exposed residue so it is free to bind to its cognate tRNA.

H2: A cognate tRNA for the cleaved residue can be obtained by adding to *trans* a full set of tRNAs and the corresponding aaRSs, as well as ATP. This will result in one of the tRNAs being charged with the cleaved residue leading to an aminoacyl tRNA..
H3: The aminoacyl tRNA can be identified by base-pairing its anticodon with a codon in an RNA molecule. This requires a platform that holds a set of (at least) 20 RNA molecules, one (at least) for each of the 20 standard amino acids, ending in a unique triplet codon and held in a fixed location where the charged tRNA can base-pair with a matching codon. A setup similar to the one in H1 above can be used for this. It takes the form of an extended e-cell that has one *trans* chamber and 20 (or more) individual *cis* chambers and nanopores, one (or more) for a standard amino acid type, through which an RNA molecule is threaded 5' end first to expose a matching terminal codon in the common *trans* chamber. Identification can then be based on fluorescent tagging with two distinct colors, one for the tRNAs and the other for amino acids. In this case, the charged tRNA can be identified based on the two colors (which will be in close proximity) and their position in the extended e-cell. If the pores in the extended e-cell are sufficiently separated in space there is no possibility of a misread.

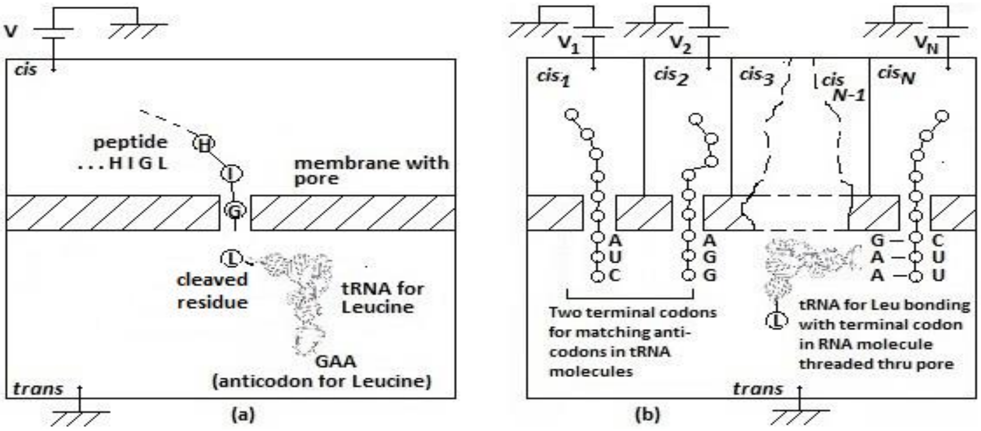

Fig 1. Electrolytic cells for sequencing a protein molecule (schematic, not to scale)
(a) Standard e-cell with *cis* and *trans* chambers. Voltage V provides electrophoretic force for analyte to translocate through pore such that first residue in protein is exposed and cleaved (residue is shown free) in *trans* by exopeptidase (not shown). Catalyzed by residue's cognate amino acyl tRNA synthetase and ATP, tRNA is charged at 3' end (terminal triplet CCA covalently binds with carboxyl terminal of residue). Occurrence of charging is detected optically with tags attached to residue and tRNA. All reactants in *trans* are then transferred to *trans* chamber of e-cell on the right.
(b) E-cell has N ($20 \leq N \leq 61$) individual *cis* chambers each with pore through common membrane. *trans* chamber is common to all of them. Voltages V_1, ..., V_N are individually applied to N *cis*-*trans* pairs, independently control translocation of i-th RNA molecule through i-th pore. Each RNA molecule ends in unique codon, one (or more, because of wobble) of which matches with anticodon in tRNA, tRNA in *trans* binds with some RNA by base-pairing. If residue and tRNAs are fluorescently tagged with two distinct colors C_1 and C_2, residue can be identified from position of color pair C_1-C_2.

Figure 1 shows the two types of e-cells. In Figure 1(a) the terminal residue is L (Leucine); it is held in *trans* just outside the pore and then cleaved with an exopeptidase. (Only the cleaved residue is shown.) An aaRS charges a cognate tRNA with this free residue in *trans*. The charged tRNA is then transferred to the *trans* chamber of the second e-cell where its anticodon base-pairs with the terminal codon of an RNA molecule. The aminoacyl tRNA carries the anticodon GAA; the corresponding codon CUU codes for Leucine. In Figure 1(b) the codon CUU is the terminal triplet of the N-th RNA molecule. Residues are tagged with color C_1, tRNAs with C_2 . After the color pair C_1-C_2 has been sensed and its position in *trans* (given by the number of the pore, 1 to N) noted, the voltages V_1, ..., V_N can be set to suitable values to translocate all the RNA molecules (along with the base-paired tRNAs) fully into the *trans* chamber. Following this the contents of *trans* can be flushed out, the chamber refilled with electrolyte, and the pores re-threaded with RNA molecules as before for the next cycle.

The protein sequencing procedure is given next.

3. Protein sequencing procedure

The following sequence of steps is repeated for every residue in the target protein. The operations in the two e-cells can be pipelined for speedup.

Identification procedure for a single residue

Step 1. Attach fluorescent dye of some fixed color C_1 to every residue in analyte protein molecule, color C_2 to each of N ($20 \leq N \leq 61$) tRNAs.

Step 2. Advance protein molecule through pore in first e-cell (Figure 1(a)) with voltage V (or some other means, see Section 4 below) set such that the first residue (C-terminal or N-terminal) remains stationary outside pore in *trans* chamber while rest of polymer remains in pore and *cis*.

Step 3. Cleave residue with an exopeptidase (amino peptidase for N-terminal, carboxy peptidase for C-terminal), which may be bonded to membrane on *trans* side or free in *trans* solution of first e-cell.

Step 4. Add full set of tRNA molecules to *trans* of first e-cell, corresponding aaRSs, and ATP so 3' binding site (CCA terminus) of some tRNA attaches to cleaved residue (Leucine in example above). All other tRNAs remain in *trans* solution. Charged tRNA will be recognizable from tandem color pair C_1-C_2 as it diffuses around in solution. All other tRNAs will be associated only with color C_2 and can be so distinguished.

Step 5. Prepare second e-cell (Figure 1(b)). Thread RNA molecule ending in unique triplet corresponding to codon for amino acid i through pore i. Set voltages V_i such that terminal triplet in RNA molecule is in *trans* and trailing bases are in pore i and *cis*$_i$. In each case pore current is lower than free pore current because of blockading effect of RNA molecule in pore.

Step 6. Transfer all tRNAs (including charged one) and other reactants (including aaRSs) in solution out of *trans* of first e-cell into *trans* of second e-cell. Charged tRNA's anticodon base-pairs with terminal codon of one of the RNAs, say RNA_i. C_1-C_2 will remain in close vicinity of RNA_i.

Step 7. Mark position of color pair C_1-C_2 as pore number i. Other tRNAs could have their anticodons base-paired with terminal codons of other RNAs and will therefore remain in vicinity of those RNAs. Remaining tRNAs will be moving around in *trans*. If pores in second e-cell are widely separated there is no possibility of confusing color pair C_1-C_2 in position i with color C_2 of any other tRNA base-paired with an RNA or any floating tRNA. Record i as identity of residue.

Step 8. Set all V_is to appropriate values so that all threaded RNA molecules translocate fully into *trans* carrying with them any tRNAs that may have base-paired with their anticodons. This results in end of current blockade in all N pores, signaling end of read.

Step 9. Flush out *trans* of second e-cell, refill with electrolyte.

In Step 6 because of wobble in the third base of the codon it is possible for an uncharged tRNA to base-pair with the target codon thereby blocking out the charged tRNA. Such an occurrence can be detected by the absence of a fixed C_1-C_2 pair as the charged tRNA will not be stationary. When this happens the floating contents of the *trans* chamber can be saved, the threaded RNA molecules (including the one with the intruder tRNA) removed as in Step 8, and the saved floating contents reintroduced into *trans*. The maximum number of times this may have to be done is three because the maximum number of wobble variants for any codon is three.

The above procedure may be simplified considerably and the second e-cell dispensed with. Thus Step 4 can be rewritten to consist of a loop in which each of the 20 (or more) pairs of tRNA and corresponding aaRS can be added (along with ATP) one pair per pass. Both the residue and the added tRNA will be diffusing in the solution and will be recognized by their individual color tags (C_1 and C_2 respectively). Once charging occurs the charged tRNA will now be associated with the color pair C_1-C_2; uncharged tRNAs will continue to be diffusing randomly with color C_2. This event identifies the residue and the loop can be exited. Steps 5 through 9 are replaced with a single step: Flush out *trans* of the first e-cell and refill with electrolyte. Writing the standard set of amino acids as **AA** = [G, A, S, C, D, T, N, P, V, E, Q, H, M, I, L, K, R, F, Y, W], a maximum of 20 cycles, one per amino acid, would be required to identify the residue. The residue may be identified in the first cycle (as G, Glycine) in the best case and in the 20th cycle (as W, Tryptophan) in the worst case, with an average of 10 cycles (for E, Glutamic Acid).

This approach may be compared with recent methods for partial or full peptide sequencing. (With partial sequences the goal is usually identification of the peptide's parent protein.) In [13] the protein molecule is broken into short peptides and the carboxyl end of each peptide covalently attached to a glass slide. Optical tags attached to selected residue types are used for detection at the N-end, following which N-end ED is used to cleave the terminal residue. Tagging of a small number (2 or 3) of amino acid types is sufficient to identify the parent protein in a protein database. In a slightly different approach, a theoretical model is described in [24] where peptides ending in a Cysteine (Cys) residue are assumed bound to a glass slide, and the binding times of various optically tagged sensor molecules known as N-terminal amino acid binders (NAABs) to the N-terminal residue of the peptide are calculated. The results suggest that by using multiple weak binders in succession a terminal residue may be identified at a rate > 97% [24]. Effectively this could lead to the full protein sequence being known, so an

doi:10.20944/preprints201911.0318.v4

implementation of this approach would not be limited to protein identification.

In contrast, the method proposed here is aimed at protein sequencing (rather than identification). Sequencing may be done from the C-terminal or the N-terminal end of an unfolded protein molecule. A tRNA that carries the anticodon for the terminal residue attaches itself to the C-terminal end of the freed residue to charge the latter. Because charging is highly specific, multiple rounds of binding (as required in [24] to increase the probability of a correct read) and optical intensity measurements are not necessary.

4. Potential implementation issues

The proposed scheme is considerably simpler than the *in vivo* translation process. The latter is dynamically regulated by a number of processes occurring simultaneously in a restricted and crowded space. In contrast the targets in the present case, namely the cleaved residue in the first cell and the terminal codons in the N RNA molecules in the second cell, can be targeted in a more controlled manner.

In the discussion that follows several assumptions are made for the procedure in Section 3: 1) the terminal residue is exposed in *trans* while all the successor residues are in the pore and *cis*; 2) the protein is held stationary under the influence of the voltage V set to an appropriate value (and/or by some other means, discussed in Section 4); 3) the exposed residue is cleaved by a free exopeptidase in the *trans* solution; 4) an aaRS charges the residue with a cognate tRNA accurately; and 5) fluorescent tagging of amino acids or of tRNAs does not adversely affect charging of a tRNA, base-pairing of a tRNA with a terminal codon of an oligo RNA, or cleaving of the terminal residue of a peptide/protein. The rest of this section looks at potential implementation of such a scheme and problems that may arise therefrom, as well as potential solutions. There are broadly three issues to consider. 1) residue cleaving; 2) polymer position control; and 3) event detection.

4.1 Cleaving of a terminal residue

Cleaving of the exposed residue of the shrinking protein molecule is central to the proposed sequencing procedure. Proteolysis is a widely studied process and there is a vast literature on the subject; comprehensive reference works are available [25,26]. In the present context the objective is limited to cleaving the first exposed residue of a protein molecule in the pore of the first e-cell. Some applicable methods are discussed elsewhere [11] in the context of nanopore-based protein sequencing using a tandem electrolytic cell. In particular see [27] and [28] for information on cleaving at the C- and N-terminal with carboxypeptidases and aminopeptidases respectively. Recently an enzyme designed on a computer and named Edmanase that cleaves residues at the N-end of a protein has been described [29]. Naturally occurring proteins that recognize specific N-terminal amino acids have been studied for possible use in peptide sequencing, leading to variants of the adapter protein ClpS that can recognize N-terminal F (Phenylalanine) and W (Tryptophan) [30].

In the procedure in Section 3, exposure in *trans* was limited to the terminal residue. The reason for this is to prevent spurious cleaving at an internal bond. Such targeted cleaving can be done either by an exopeptidase that is covalently bonded to the membrane (or pore) on the *trans* side or free exopeptidase in the *trans* solution. However as noted below this may require precise control of the molecule by one of several ways discussed below. Such precision can be avoided if spurious cleavage does not occur, so exposure need not be limited to the terminal residue. This is discussed in the next subsection.

4.2 Polymer position control

The analyte molecule (protein in the first e-cell, RNAs in the second) translocates rapidly through a pore, which makes it difficult to measure the change in the current blockade level due to individual monomers [31]. Currently available detector bandwidths do not support such measurements. A range of methods to slow down the polymer in its passage through the nanopore have been reported, an early review can be found in [32] (but also see [33]). More recent work can be found in [34-40]. The primary aim in all of the above slowdown methods is to allow discrimination of bases in DNA and residues in proteins (and, in one case, unfolding of the protein as well [34]).

In the present case, however, there is no need to make precise blockade current measurements, it suffices to detect the occurrence of a blockade, as the goal is only to hold the protein in the pore in the first e-cell and the RNA molecules in their pores in the second e-cell such that only the terminal residue and the terminal codons are exposed, while the protein or RNA remains stationary. One possible way to achieve stationarity of the polymer is to balance the electrophoretic force due to the voltage against the motive force due to diffusion. In principle this may be easier to do with DNA and RNA, which carry a uniform negative charge along the backbone, whereas most proteins carry a weak net charge that may be positive or negative. In the latter case better control may be possible by attaching a trailing homopolymer polyZ where Z is a charged amino acid with appropriate polarity. However, as noted in [32] this is somewhat difficult because the diffusive force is spread throughout the length of the polymer both inside the pore and outside and it is hard to have stop-and-go control over the length of the polymer segment that is inside the pore. There are multiple ways to resolve this problem: a) mechanical position control, b) hydraulic position control, and c) tethering the molecule to some fixed location. They may be used singly or in combination with the electrical potential V across *cis* and *trans*.

a) *Mechanical control*
Mechanical approaches provide a viable way to hold the polymer in place with little movement. Optical or magnetic tweezers [41-43] are particularly effective. Here the electrophoretic and diffusive forces on a polymer can be exactly balanced by a mechanical force, resulting in the polymer remaining stationary, with nanometer to sub-nanometer precision. In [41,42] optical tweezers hold a double stranded DNA molecule for insertion into a nanopore with sub-nanometer positioning precision, thus allowing for base-level discrimination. Alternatively a polymer can be attached to a colloidal bead held between the poles of a magnet, leading to controlled translocation of the polymer through the pore [43]. The magnetic field counteracts the usual electric potential (and diffusion) to precisely retract the molecule after its entry into the pore. In principle, in combination with an appropriately designed feedback circuit these methods can be used to linearly stretch out an unfolded protein/peptide fully over the length of the pore so that stop-and-go control can be exercised over its translocation through the pore.

b) *Hydraulic control*
The use of hydraulic pressure to control the movement of a polymer through a pore was first studied in [44]. This approach has been investigated in detail recently [45]. With pressures in the range 0.5 to 2 atmospheric pressures (atm) applied with a plunger mounted in the *cis* or *trans* chamber to act against the electric field over the length of the pore, blockade current changes on the order of 5-10 nA have been measured with the analyte trapped for considerably long periods of time. (In comparison, in most nanopore experiments the currents measured are around 100-200 pA [8].)

c) *Chemical tethering*
Covalent bonding makes it possible to hold one end of a molecule in position. In [13] and [24] this is done with a glass slide acting as an anchor, and terminal residues identified via total internal reflection fluorescence (TIRF) microscopy. In [46] a DNA strand is bound to the vestibule of an alpha hemolysin pore, in [47] to an adapter attached at the pore entrance.

In the present case, all three approaches may be suitable. For example, similar to [13] and [24] a peptide can be tethered to a glass slide in the *cis* chamber of the first e-cell, then drawn into a nanopore with an electric field. With a sufficiently large field the peptide can be stretched fully through the pore, and the glass slide controlled with a plunger in *cis* to which hydraulic pressure may be used to insert/retract the peptide into/from the pore. This enables the controlled exposure of only the terminal residue in the *trans* chamber while the rest of the tethered molecule remains in the pore and *cis*. A similar setup can be used with optical or magnetic tweezers.

With the second e-cell the RNA molecules can be controlled with tweezers or tethered to the membrane in the respective *cis* chamber near the pore entrance on the *cis* side (see Figure 1(b)). In the latter case to ensure correct base pairing of the anticodon on a tRNA with the terminal codon of an RNA molecule in the second e-cell an RNA oligo should have k+3 bases where k * length of base ≈ length of pore such that only the terminal triplet is exposed in *trans*. Once again with a suitable $V_i > 0$ the i-th RNA molecule, which carries a uniform negative charge, is drawn into the the i-th pore and remains fully stretched inside while only the terminal triplet is exposed. However, the length of the pore places a rigid constraint on the value of k. One possible solution is to use a tethered DNA molecule with a dangling single strand 3-base extension consisting of only the triplet codon (or its equivalent, since in DNA base T replaces base U).

4.3 Event detection

Four salient events occur in sequence in the procedure in Section 3, the times of their occurrence are instrumental in driving the identification procedure. These are

a) cleaving of the exposed residue in the *trans* chamber of the first e-cell;
b) charging of a tRNA with the cleaved residue;
c) transfer of the charged tRNA to the second cell; and
d) binding of the transferred tRNA in the second e-cell to the terminal codon of some RNA molecule.

With optical tagging all four events are easy to detect. With color C_1 associated with residues in the protein molecule and color C_2 with tRNAs, event a is flagged by color C_1. This generates a moving spot in the image as the free residue diffuses around in the *trans* chamber of the first e-cell. Following this the addition of tRNAs to *trans* causes a number of moving spots of color C_2. When a cognate tRNA attaches itself to the residue, the C_1 and C_2 spots form a C_1-C_2 pair, this flags the occurrence of event b. When the transfer in c occurs, it leads to a single C_1-C_2 pair and a bunch of C_2 spots due to the uncharged tRNAs in the *trans* chamber of the second e-cell, all of them diffusing around in *trans*. When the charged tRNA binds with one of the RNAs by base-pairing of the anticodon in the former with the terminal codon of the latter the C_1-C_2 pair location becomes more or less fixed. This flags event d. (Some of the uncharged tRNAs may base-pair with the other RNAs thereby fixing some of the C_2 spots while the C_2 spots of the remaining uncharged tRNAs will continue to move around. None of this affects residue identification.)

4.4 Relaxing the constraint on residue or terminal codon exposure

The condition that only the terminal residue be exposed in the *trans* chamber of the first e-cell can be relaxed if an exopeptidase that reliably cleaves only the terminal residue is used. In this case part or all of the target peptide can be allowed into *trans* while the part of the molecule that is in the pore remains stationary. This provides more flexibility in polymer position control. In this case if a poly-X trailer of length m, where X is one of the charged residues (D, E, K, R, H, C, Y), is attached to the tail of the target peptide the peptide can be paid out into *trans* one residue at a time with one of the control methods discussed above.

Thus let the peptide sequence R_1 R_2 ... R_n going from N-terminal to C-terminal (or *vice versa*) be extended by a poly-X trailer of length m, where X is a charged residue and m*length of residue X > length of pore in the first electrolytic cell. Let X be the negatively charged residue D. If X_1 is tethered to the membrane near the pore entrance in *cis*, then with V set to a large enough positive value (see Figure 1(a)) the extended molecule D D ... D_m R_1 R_2 ... R_n is drawn into the pore and remains fully stretched through the pore. Although two or more residues from R_1 R_2 ... R_n are now exposed in *trans*, the terminal R gets cleaved and is attached to a corresponding tRNA. In the following cycles the next R is cleaved and becomes available for charging, etc. Any of the above position control methods, tweezer, hydraulic, or tethering, can be used to ensure that the analyte remains close to the pore on the *trans* side so that charging and cleaving events occur in a known small neighborhood where they can be detected fairly easily with optical tagging.

5. Discussion

1) In the procedure in Section 3 it was assumed that the terminal residue is cleaved before a tRNA is charged with it. This is to prevent any potential steric clashes with the pore or the membrane during charging. However, if it can be determined experimentally that a tRNA can be charged with the terminal residue before cleaving, some distinct advantages can accrue. First, the use of optical/magnetic tweezers or tethering to control the position of the peptide becomes simpler. Second, if the charging time is known (it can be determined experimentally) then there is no need for optical tags at all. Third, because the tRNA is bound to the exposed terminal residue so that after charging occurs all other reactants can be flushed out of the *trans* chamber of the first e-cell before cleaving of the residue. Following this the charged tRNA can be transferred to the second e-cell (along with any free peptidase used for cleaving). Since there are no other reactants, this is the only transfusion into the second e-cell. The anticodon in the charged tRNA binds to the codon of one of the RNAs by base-pairing. As in the case of charging, prior knowledge of base-pairing times can be used to determine when the next step, namely identification of the residue using optical or non-optical means, can take place.

Nevertheless optical tags are useful to precisely determine the times when charging occurs in the first e-cell and when base-pairing of the tRNA with an RNA occurs in the second e-cell. Thus if the tRNAs are optically tagged with a single color C_2 the occurrence of charging can be detected when the color C_2 of the charged tRNA spot is bound to the terminal residue and remains stationary (all other C_2 spots will be diffusing around in the e-cell). Since these other reactants are flushed out, the moving C_2 spots will disappear, leaving only the fixed C_2 of the charged tRNA. After cleaving, when the tagged tRNA is transferred to the second e-cell, the event of base-pairing of its anticodon to the codon of an RNA (RNA_i) is flagged by the C_2 spot becoming stationary in the vicinity of pore i. The spatial position of C_2 in the *trans* chamber of the second e-cell is sufficient to identify the residue.

Based on the above discussion, an alternative residue identification procedure is given next, it is to be repeated for every residue in the target protein. As with the procedure in Section 3, the two e-cells can be pipelined.

Alternative identification procedure for a single residue

Step A1. Attach fluorescent dye of some fixed color color C_2 to each of N ($20 \leq N \leq 61$) tRNAs. (This step is optional.)

Step A2. Same as Step 2 in Section 3 except that protein molecule enters pore C-terminal first.

Step A3. Same as Step 4 in Section 3. Some tRNA gets charged with exposed uncleaved residue. If optical tagging is done as in Step A1, this leads to stationary C_2 spot near exit of pore in *trans*, and moving C_2 spots due to all uncharged tRNAs.

Step A4. Flush these remaining tRNAs (and all other reactants) out of *trans* while refilling with electrolyte. If optical tagging is done only C_2 spot due to charged tRNA remains in *trans* of first e-cell.

Step A5. Same as Step 5 in Section 3.

Step A6. Cleave residue with a carboxyl peptidase. If optical tagging is used in Step A1, C2 spot due to charged tRNA will start diffusing around in *trans*.

Step A7. Transfer charged tRNA from *trans* of e-cell 1 to *trans* of e-cell 2.
If optical tagging is used in Step A1, C_2 spot due to charged tRNA will start diffusing around in *trans* of e-cell 2. When anticodon in tRNA base-pairs with terminal codon of RNA_i C_2 will remain stationary near RNA_i; residue is identified by pore number i; go to Step A9.

Step A8. (Optical tagging not used) Set voltages V_i in second e-cell to values such that all the RNAs retract into their *cis* chambers, except the one with whose terminal codon charged tRNA has base-paired. In latter case current through pore is still at blockade level, in all other cases current blockade ceases. Number of blockaded pore identifies residue.

Step A9. Flush out *trans* chamber of second e-cell and refill with electrolyte. Rethread RNA molecules through all N pores as before with terminal codon exposed. (Alternatively with optical tagging base-paired tRNAs can be detached from RNAs with appropriate enzyme before flushing, in which case rethreading is not needed. Such detachment would be required if tethering or tweezers are used for position control.)

The procedure can be further simplified similar to the simplification discussed in Section 3 following the residue identification procedure in that section. Add tRNA-aaRS pairs individually in sequence until charging occurs. As before a maximum of 20 cycles, one per amino acid, and an average of 10 cycles would be required (refer to the amino acid array **AA** given earlier). With optical tagging only the tRNA needs to be tagged, and there is no need for the second e-cell. The residue (with the attached tRNA) is cleaved, followed by flushing of *trans* and refilling with electrolyte. Without optical tags, the contents of *trans* in the first e-cell are transferred to *trans* in the second e-cell and Steps A8 and A9 are executed.

If tRNA charging before cleaving is problematic then the possibility of modifying aaRSs to enable such charging can be considered. Such modification may be designed similar to [48], wherein lists of the primary sequences of redesigned aaRSs for use as NAABs in protein sequencing are given.

2) Although the sequencing procedure described above may broadly apply to full proteins, it may be difficult in practice because of the tendency of proteins to fold outside the nanopore. This can be avoided with a shotgun sequencing approach similar to that used for DNA sequencing. Thus the protein can be broken into fragments that are short enough that the likelihood of their forming folds is minimal. These peptides can be sequenced and the resulting sequences assembled with an assembly algorithm such as one based on Eulerian paths [49]. Unlike with DNA such assembly is much easier as the probability of long homopolymers occurring in a protein sequence is very low.

3) Most of the voltage drop in an electrolytic cell is across the pore [8]. The drop in the *cis* and *trans* chambers is only ~2% so the motive force in these chambers is mostly diffusion. Because of this an analyte molecule may have to be drawn into the pore by some agency such as dielectrophoretic trapping [50]. With proteins a possible solution is to use a sodium dodecyl sulphate (SDS) sheath to give the protein a uniform negative charge that gets stripped out as the protein is drawn into the pore [16,51].

4) In the case when charging is done before cleaving, the protein molecule must enter the pore in the first e-cell carboxyl end first as the tRNA attaches to this end. To ensure entry at the correct end the amino end can be capped with a small molecule of a size large enough that it cannot enter the pore. This also requires a polyZ trailer of length b, where Z is a selected amino acid and b*(length of Z) > length of pore. Similarly an RNA molecule in the extended e-cell has to enter a pore at the 3' end. Once again capping can be used for this. An alternative approach in both cases would be to reverse the roles of *cis* and *trans*, which can effectively be achieved by reversing the voltage polarity. Entry into a pore at the wrong end can be detected optically or when the desired event (charging of tRNA, binding of tRNA to codon) does not occur within some expected time.

5) The procedure as described above assumes that post-translational modification of an amino acid in the sequence does not adversely affect the process of charging of a tRNA with a cleaved residue. The nature of such modifications, if any, will have to be determined by some other method, after the residue has been identified in the second e-cell. One possibility would be to detach the tRNA (along with the residue with which it has been charged) from the terminating codon of an RNA in the second e-cell that it has base-paired with and then subject it to analysis with a method like GE. With the latter, in principle the distance traveled by the fluorescently tagged tRNA-residue complex may be detected in the gel from the position of the tag. The distance traveled by the analyte (tRNA + amino acid + tag + any post-translational modification) would be roughly proportional to the sum of the four masses, from which the mass of the modified part of the amino acid (e.g., phophorylate, acetylate) may be estimated and the modifier identified. However this also raises the question whether current methods of GE are sensitive enough to do this with a single tRNA molecule.

6) Nanopores play a secondary role in the proposed approach. Their only purpose is to linearize and keep unfolded a protein/RNA molecule such that only the terminal residue or codon is exposed to chemical reactions, which, in the case of protein, minimizes the possibility of an exopeptidase acting like an endopeptidase and cleaving the protein internally. Thus such properties as pore length (or equivalently membrane thickness), pore diameter, electro-osmotic effects due to charged residues on the inner wall of the pore lumen, and similar properties that affect conventional nanopore-based sequencing are not important. The only exception is that the pore diameter be sufficiently small that the protein does not fold inside the pore and residues get exposed in *trans* one at a time in sequence order.

References

[1] J. M. Heather and B. Chain. “The sequence of sequencers: the history of sequencing DNA”. *Genomics* **107**, 1–8, 2016.

[2] E. Pettersson, J. Lundeberg, and A. Ahmadian. “Generations of sequencing technologies”. *Genomics* **93**, 105–11, 2009. doi:10.1016/j.ygeno.2008.10.003

[3] R. J. Simpson. *Proteins and Proteomics: A Laboratory Manual*, CSHL Press, 2008.

[4] E. de Hoffmann and V. Stroobant. *Mass Spectrometry: Principles and Applications*, 3rd edn., Wiley, 2007.

[5] T. Rabilloud and C. Lelong. “Two-dimensional gel electrophoresis in proteomics: a tutorial”. *J. Proteomics* **74**, 1829-1841, 2011.

[6] N. Callahan, J. Tullman, Z. Kelman, and J. Marino. "Strategies for development of a next-generation protein sequencing platform". *Trends Biochem. Sci*. 2019. doi:10.1016/j.tibs.2019.09.005.

[7] H. Bayley. “Nanopore sequencing: from imagination to reality”. *Clin. Chem*. **61**, 25–31, 2015.

[8] J. E. Reiner, A. Balijepalli, J. W. F. Robertson, J. Campbell, J. Suehle, and J. J. Kasianowicz. “Disease detection and management via single nanopore-based sensors”. *Chem. Rev*. **112**, 6431-6451, 2012.

[9] D. L. Nelson and M. M. Cox. *Lehninger's Principles of Biochemistry*, 4th edn., W H Freeman, 2005.

[10] D. Deamer, M. Akeson, and D. Branton. “Three decades of nanopore sequencing”. *Nature Biotechnol*. **34**, 518–524, 2016.

[11] G. Sampath. “Amino acid discrimination in a nanopore and the feasibility of sequencing peptides with a tandem cell and exopeptidase”. *RSC Adv*. **5**, 30694–30700, 2015.

[12] J. Swaminathan, A. A. Boulgakov, and E. M. Marcotte. “A theoretical justification for single molecule peptide sequencing”. *PLOS Comput. Biol*. **11**, e1004080, 2015.

[13] P. Boynton and M. Di Ventra. “Sequencing proteins with transverse ionic transport in nanochannels”. *Sci. Rep*. **6**, 25232, 2016.

[14] J. Wilson, L. Sloman, Z. He, and A. Aksimentiev. “Graphene nanopores for protein sequencing”. *Adv. Funct. Mater*. **26**, 4830–4838, 2016.

[15] 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 Nanotechnol*. **9**, 466–473, 2014.

[16] E. Kennedy, Z. Dong, C. Tennant, and G. Timp. “Reading the primary structure of a protein with 0.07 nm^3 resolution using a subnanometre-diameter pore”. *Nature Nanotechnol*. **11**, 968–976, 2016.

[17] J. Swaminathan, A. A. Boulgakov, E. T. Hernandez, A. M. Bardo, J. L. Bachman, J. Marotta, A. M. Johnson, E. V. Anslyn, and E. M. Marcotte. “Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures”. *Nature Biotechnol*. **36**, 1076-1082, 2018.

[18] L. Restrepo-Pérez, C. Joo, and C. Dekker. “Paving the way to single-molecule protein sequencing”. *Nature Nanotechnology* **13**, 786-796, 2018. https://doi.org/10.1038/s41565-018-0236-6

[19] H. Ouldali, K. Sarthak, T. Ensslen, F. Piguet, P. Manivet, J. Pelta, J. C. Behrends, A. Aksimentiev, and A. Oukhaled. "Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore". *Nature Biotech*., December 16, 2019.

[20] C. Joo, H. Balci, Y. Ishitsuka, C. Buranachai, and T. Ha. “Advances in single-molecule fluorescence methods for molecular biology”. *Annu. Rev. Biochem*. **77**, 51–76, 2008.

[21] O. O. Favorova. “Superspecificity of aminoacyl-tRNA-synthases”. *Mol. Biol*. (Mosk). **18**, 205-226, 1984.
[22] D. Goodsell. “Aminoacyl-tRNA Synthetases”. *PDB Molecule of the Month*, April 2001. doi:10.2210/rcsb_pdb/mom_2001_4
[23] T. A. Steitz. “A structural understanding of the dynamic ribosome machine”. *Nature Reviews* **9**, 242-253, 2008.
[24] S. Rodriques, A. Marblestone, and E. Boyden. “A theoretical analysis of single molecule protein sequencing via weak binding spectra”. *PloS One* **14**, e0212868, 2019. doi: 10.1371/journal.pone.0212868
[25] A. J. Barrett, N. D. Rawlings, and J. F. Woessner. (eds.) *Handbook of Proteolytic Enzymes*, Academic Press, 1998.
[26] D. L. Crimmins, S. M. Mische, and N. D. Denslow. “Chemical cleavage of proteins on membranes”. *Curr. Protoc. Protein Sci*., 2001. doi: 10.1002/0471140864.ps1105 s19
[27] K. Breddam and M. Ottesen, “Determination of c-terminal sequences by digestion with serine carboxypeptidases: the influence of enzyme specificity”. *Carlsberg Res. Commun*. **52**, 55-63, 1987.
[28] A. Taylor, “Aminopeptidases: structure and function”. *FASEB J*. **7**, 290-298, 1993.
[29] B. Borgo and J. J. Havranek. "Computer-aided design of a catalyst for Edman degradation utilizing substrate-assisted catalysis". *Protein Sci*. **24**, 571–579, 2015.
[30] J. Tullman, N. Callahan, B. Ellington, Z. Kelman, and J. P. Marino. "Engineering ClpS for selective and enhanced N-terminal amino acid binding". *Appl. Microbiol. and Biotechnol*. **103**, 2621–2633, 2019.
[31] S. Carson and M. Wanunu. “Challenges in DNA motion control and sequence readout using nanopore devices”. *Nanotechnol*. **26**, 074004, 2015.
[32] U. F. Keyser. "Controlling molecular transport through nanopores". *J. R. Soc. Interface* **8**, 1369–1378, 2011. doi:10.1098/rsif.2011.0222
[33] C. Plesa, S. W. Kowalczyk, R. Zinsmeester, A. Y. Grosberg, Y. Rabin, and C. Dekker. "Fast translocation of proteins through solid state nanopores". *Nano Lett*. **13**, 658-663, 2013. dx.doi.org/10.1021/nl3042678
[34] J. Nivala, D. B. Marks, and M. Akeson. “Unfoldase-mediated protein translocation through an alpha-hemolysin nanopore”. *Nat Biotechnol*. **31**, 247–250, 2013.
[35] L. Mereuta, M. Roy, A. Asandei, J. K. Lee, Y. Park, I. Andricioaei, and T. Luchian. "Slowing down single-molecule trafficking through a protein nanopore reveals intermediates for peptide translocation". *Sci. Rep*. **4**, 3885, 2014. doi: 10.1038/srep03885
[36] S. W. Kowalczyk, D. B. Wells, A. Aksimentiev, and C. Dekker. "Slowing down DNA Translocation through a Nanopore in Lithium Chloride". *Nano Lett*. **12**, 1038-1044,2012. dx.doi.org/10.1021/nl204273h
[37] Z. Tang, Z. Liang, B. Lu, J. Li, R. Hu, Q. Zhao, D. Yu. "Gel mesh as “brake” to slow down DNA translocation through solid-state nanopores". *Nanoscale* **7**, 13207–13214, 2015.
[38] F. Cecconi, M. A. Shahzad, U. M. B. Marconi, and A. Vulpiani. "Frequency-control of protein translocation across an oscillating nanopore". *Phys.Chem.Chem.Phys*. **19**, 11260, 2017. DOI:10.1039/c6cp08156h
[39] A. Asandei, I. S. Dragomir, G. Di Muccio, M. Chinappi, Y. Park, and T. Luchian. "Single-molecule dynamics and discrimination between hydrophilic and hydrophobic amino acids in peptides, through controllable, stepwise translocation across nanopores". *Polymers* **10**, 885, 2018. doi:10.3390/polym10080885
[40] X. Liu, Y.Zhang, R. Nagel, W. Reisner, and W. B. Dunbar. "Controlling DNA tug-of-war in a dual nanopore device". *arXiv*:1811.11105v1 [physics.bio-ph], 27 Nov 2018.
[41] U. F. Keyser, J. van der Does J, C. Dekker, and N. H. Dekker. “Optical tweezers for force measurements on DNA in nanopores”. *Rev. Sci. Instrum*. **77**, 105105, 2006. doi:10.1063/1.2358705.
[42] E. H. Trepagnier, A. Radenovic, D. Sivak, P. Geissler, and J. Liphardt. "Controlling DNA capture and propagation through artificial nanopores". *Nano Lett*. **7**, 2824-2830, 2007.
[43] H. Peng and X. S. Ling. "Reverse DNA translocation through a solid-state nanopore by magnetic tweezers". *Nanotech*. **20**, 185101, 2009. doi:10.1088/0957-4484/20/18/185101
[44] 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*. **13**, 3048–3052, 2013.
[45] H. Zhang, Q. Chen, Y. Wu, Y. Wang, X. Bei, and L. Xiao. "The temporal resolution and single-molecule manipulation of a solid-state nanopore by pressure and voltage". *Nanotech*. **29**, 495501, 2018. doi:/10.1088/1361-6528/aae190
[46] S. Howorka and H. Bayley. “Probing distance and electrical potential within a protein pore with tethered DNA”. *Biophys. J*. **83**, 3202–3210, 2002.
[47] J. Clarke, H. C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley. “Continuous base identification for single-molecule nanopore DNA sequencing”. *Nature Nanotech*. **4**, 265–70, 2009.
[48] J. J. Havranek and B. Borgo. "Molecules and methods for iterative polypeptide analysis and processing", 2013 U. S. Patent US20140273004A1. Patent Assignee: Washington University in St Louis.
[49] P. A. Pevzner, H. Tang, and M. S. Waterman. “An Eulerian path approach to DNA fragment assembly,” *PNAS* **98**, 9748–9753, 2001.
[50] K. J. Freedman, L. M. Otto, A. P. Ivanov, A. Barik, S.-H. Oh, and J. B. Edel. “Nanopore sensing at ultra-low concentrations using single-molecule dielectrophoretic trapping”. *Nature Commun*. **7**, 10217, 2016. doi: 10.1038/ncomms10217
[51] L. Restrepo-Pérez, S. John, A. Aksimentiev, C. Joo, and C. Dekker. "SDS-assisted protein transport through solid-state nanopores". *Nanoscale* **9**, 11685–11693, 2017.

Email: sampath_2068@yahoo.com