Protein identification with a nanopore and a binary alphabet

G. Sampath

Abstract. Protein sequences are recoded with a binary alphabet obtained by dividing the 20 amino acids into two subsets based on volume. A protein is identified from subsequences by database search. Computations on the *Helicobacter pylori* proteome show that over 93% of binary subsequences of length 20 are correct at a confidence level exceeding 90%. Over 98% of the proteins can be identified, most have multiple identifiers so the false detection rate is low. Binary sequences of unbroken protein molecules can be obtained with a nanopore from current blockade levels proportional to residue volume; only two levels, rather than 20, need be measured to determine a residue's subset. This procedure can be translated into practice with a sub-nanopore that can measure residual volumes with ~0.07 nm³ resolution as shown in a recent publication. The high detector bandwidth required by the high speed of a translocating molecule can be reduced more than tenfold with an averaging technique, the resulting decrease in the identification rate is only 10%. Averaging also mitigates the homopolymer problem due to identical successive blockade levels. The proposed method is a proteolysis-free single-molecule method that can identify arbitrary proteins in a proteome rather than specific ones.

1. Introduction
Sequencing/identification of peptides/proteins is currently based on Edman degradation, gel electrophoresis, or mass spectrometry [1-3]; it is usually done in the bulk and often followed by database search [3]. In comparison with DNA sequencing, for which several NGS (next generation sequencing) technologies are currently available [4], protein sequencing is more difficult, as it has to discriminate among 20 amino acids, as opposed to four bases with DNA. It also has to work with the available sample, as there is no amplification technique for proteins comparable to PCR for DNA [2].

Compared to protein sequencing, protein identification is simpler because it only requires finding a partial sequence and then searching through a protein sequence database to find the protein that is uniquely identified by it. Recently single-molecule methods based on proteolysis and optical or other labeling of selected residues have been proposed. In one of them a protein is cleaved into peptide fragments, pinned to a substrate, and selectively labeled [5]. The labeled residues are detected optically and a partial sequence obtained.

In contrast with these methods nanopores provide a single-molecule electrical alternative that does not require proteolysis, analyte immobilization, or labeling of any kind [6]. While their use in sequencing DNA is becoming commonplace [7], nanopore-based protein sequencing may seem like a distant prospect. Recent reviews of nanopore-based protein studies are available [8-11]. Examples of work in the area include experimental studies involving recognition tunneling [12] and use of a sub-nanometer diameter nanopore [13]. A major limiting factor has been the pore current resolution needed to discriminate among 20 different amino acids; at present this appears to be out of reach. There has been some success in nanopore-based studies of other aspects of proteins such as folding/conformation [14,15] and recognition of specific proteins or their variants [16-19].

1.1 The present work
Here it is shown by computation that proteins in a proteome can be uniquely identified from subsequences of primary sequences by using a binary code derived from a division of the amino acids into two sets based on published volume data [20]. Comprehensive computations on a sample proteome (*Helicobacter pylori*) show that the codes of subsequences 20 residues long are correct at a better than 90% confidence level, and that over 98% of the proteins in the proteome can be identified by searching for the subsequences in the binary-coded proteome database. With a nanopore this means that a pore current resolution that can discriminate between the two subsets mentioned above is sufficient. (Incidentally a binary alphabet has been used recently in DNA sequencing with a solid-state nanopore. In this approach a nucleotide is encoded with a predesigned oligonucleotide and two fluorescent label types to represent the four bases [21].)

The scheme described here can be translated into practice with existing technology (nanopore, electronic, database); an easy-to-use hand-held device similar to the MinION in genome sequencing [22] can be designed and implemented. Unlike most nanopore-based protein identification methods, the one presented here can identify arbitrary proteins in a proteome rather than specific individual targets [16-19]. As such it can also be extended to the analysis of protein mixtures (see Section 3.4).

2. Protein identification with a nanopore
An electrolytic cell has a membrane dividing two chambers (*cis* and *trans*) containing an electrolyte. A potential difference applied across the membrane leads to an ionic current through the pore from *cis* to *trans*. An analyte molecule (such as a polymer like DNA or protein) introduced into *cis* translocates through the pore to *trans* and
causes a current blockade. A monomer may be identified from the blockade level, which may be specific to one or more contiguous monomers based on some physical or chemical property such as volume, charge, diffusion constant, etc. With proteins no enzymatic digests are required; the analyte is a single denatured unfolded protein molecule and a monomer is a residue.

Normally a blockade current resolution able to discriminate among the standard 20 residue types would be required. Such resolution is unattainable in practice, especially with noise present. This is mitigated somewhat if a four-way division of the 20 amino acids based on volume is used [13].

The method presented here makes the resolution problem manageable by reducing the measured number of blockade current signal levels from 20 to two. Subsequences in the resulting binary sequence that can uniquely identify the protein in a binary-coded proteome database are then found. As noted earlier and demonstrated below, computational analysis of a sample proteome (H. pylori) using amino acid volume data [20] indicates that almost all of the proteins therein can be identified with a confidence level exceeding 90%.

The proposed method is analyzed computationally before considering implementation issues.

3. Computational analysis and results
There are three steps: 1) Divide the set of amino acids into two ordered subsets S₁ and S₂; 2) Recode the primary sequences in a proteome with a binary code based on this two-way partition; 3) For every protein in the proteome find one or more subsequences in its binary-coded primary sequence that identify the protein uniquely.

Table 1. Amino acids sorted by volume. AA = amino acid. Mean and S.D. (Standard Deviation) in 10⁻³ nm³. Data from [20].

<table>
<thead>
<tr>
<th>AA</th>
<th>Mean</th>
<th>S.D.</th>
<th>AA</th>
<th>Mean</th>
<th>S.D.</th>
<th>AA</th>
<th>Mean</th>
<th>S.D.</th>
<th>AA</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>59.9</td>
<td>2.2</td>
<td>T</td>
<td>118.3</td>
<td>2.3</td>
<td>Q</td>
<td>145.1</td>
<td>5.1</td>
<td>K</td>
<td>172.7</td>
<td>5.9</td>
</tr>
<tr>
<td>A</td>
<td>87.8</td>
<td>2.3</td>
<td>N</td>
<td>120.1</td>
<td>4.1</td>
<td>H</td>
<td>156.3</td>
<td>6.1</td>
<td>R</td>
<td>188.2</td>
<td>9.6</td>
</tr>
<tr>
<td>S</td>
<td>91.7</td>
<td>1.8</td>
<td>P</td>
<td>123.3</td>
<td>1.8</td>
<td>M</td>
<td>165.2</td>
<td>1.8</td>
<td>F</td>
<td>189.7</td>
<td>7.4</td>
</tr>
<tr>
<td>C</td>
<td>105.4</td>
<td>5.5</td>
<td>V</td>
<td>138.8</td>
<td>3.6</td>
<td>I</td>
<td>166.1</td>
<td>3.4</td>
<td>Y</td>
<td>191.2</td>
<td>8.0</td>
</tr>
<tr>
<td>D</td>
<td>115.4</td>
<td>2.2</td>
<td>E</td>
<td>140.9</td>
<td>5.3</td>
<td>L</td>
<td>168</td>
<td>4.3</td>
<td>W</td>
<td>227.9</td>
<td>3.8</td>
</tr>
</tbody>
</table>

Table 1 shows the standard 20 amino acids grouped by volume into two subsets and shaded by group. The dividing line is between P and V so that the subsets have roughly similar sizes (8 and 12) and the difference between the volumes of P and V is maximal. The amino acids are coded as follows: S₁ = {G,A,S,C,D,T,N,P: 59.9 ≤ volume ≤ 123.3} → 1, S₂ = {V,Q,E,H,M,L,K,R,F,Y,W: volume ≥ 138.8} → 2.

It was recently shown experimentally that a sub-nanometer-diameter nanopore can measure residue volume with a resolution of 0.07 nm³ [23]. Such a level of resolution is sufficient to determine that a residue in a translocating protein belongs to S₁ or to S₂. In what follows, residue volume is used as a proxy for the blockade current.

From Table 1, a blockade current threshold T₂ corresponding to a volume of ~0.13 nm³ can distinguish residues in S₁ from those in S₂. To distinguish blockades due to residues in S₁ from the open pore current a second threshold T₁ corresponding to a volume of ~0.05 nm³ is set. Thus if the measured volume of a residue is between T₁ and T₂ the output is '1'; if > T₂ the output is '2'.

Errors in the resulting binary codes for subsequences of different lengths can be estimated for different volume thresholds. Assuming residue volumes to be normally distributed with mean μ and standard deviation σ, the errors can be computed with the normal (Gaussian) error function. Fig. 1 shows normal distributions of the 20 amino acids based on data from Table 1.

Let the mean volume and standard deviation for amino acid aa be μₐa and σₐa (Table 1). Let the error in reading the volume of amino acid aa be εₐa(T₁,T₂). If F(x; μ, σ) is the cumulative normal distribution function with mean μ and standard deviation σ, the errors for the 20 amino acids are given by

\[ εₐa(T₁,T₂) = F(T₂; μₐa, σₐa) + 1 - F(T₁; μₐa, σₐa) \ ]
\[ εₐa(T₁,T₂) = F(T₂; μₐa, σₐa) \ ] (1a)
\[ εₐa(T₁,T₂) = F(T₁; μₐa, σₐa) \ ] (1b)
Assuming that blockades due to successive residues are independent the probability that the measured binary code for a protein sequence $X = X_1 X_2 \ldots X_n$, where $X_i$ is one of the 20 amino acids, is correct (that is, its confidence level) is given by

$$c_x(T_1, T_2) = \Pi_{i=1}^n (1 - e^{X_i(T_1, T_2)})$$  \hspace{1cm} (2)

The proteome of the gut bacterium *H. pylori* (Uniprot id UP000000210, 1553 sequences, www.uniprot.org) is used as an example. Fig. 2 (symbol ♦) shows the percentage of binary-coded subsequences with confidence level $> 90\%$ versus subsequence length.

Subsequences from every protein in a proteome are exhaustively compared with every other protein to determine if they uniquely identify their container proteins. To reduce computation time candidate subsequences used are spaced $\Delta = 5$ residues apart. The percentages of proteins identified are given for subsequence length $L = 15, 20$, and $25$ in Fig. 2 (symbol ■). The number of proteins identified goes from $\sim 25\%$ with $L = 15$ to $\sim 98\%$ with $L = 20$ and $\sim 99\%$ with $L = 25$. $L > 20$ yields diminishing returns; gains from reducing $\Delta$ are minimal. As seen in Fig. 2, $L = 20$ is a near-optimal length for subsequences.

A complete list of protein identifying subsequences for all the proteins of *H. pylori* is given in ESM (File 1).

### 3.1 Reduced false detection rate

A significant majority of proteins in the *H. pylori* proteome have a large number of identifying subsequences; this reduces the false detection rate (FDR) considerably. Fig. 3 shows the distribution of the number of proteins vs the
number of subsequences of length 20 that identify them. Of the 1553 proteins in the proteome 1501 proteins have more than one identifier, 23 have one, and 29 have none. Twenty proteins have more than 100 identifiers.

Fig. 3. Number of proteins vs number of protein identifying subsequences of length 20 in a protein in H. pylori (1553 sequences)

3.2 Reducing detector bandwidth

Normally an analyte like a protein molecule translocates through the pore rapidly (diffusion is the major cause, although the electrophoretic force due to the applied potential and other factors also play a role [24]). This makes it difficult for a detector with insufficient bandwidth to detect changes in the blockade current level. (The limit at present is $\sim$1 MHz [25].) One way to reduce the required bandwidth is to compute an average over the raw pore current with hardware or software. This is essentially a smoothing technique that also leads to a decrease in the number of proteins that can be identified. It is shown next that with this approach the required bandwidth can be reduced by a factor of 10 or more with only a 10% decrease in the number of proteins identified.

Let the detector time resolution be $\tau$, the corresponding detector bandwidth $B = 1/2\tau$. With this up to $L$ blockade pulses can be identified in a pore current signal interval of width $2L\tau$. With a bandwidth of $B/L$ (or equivalently a time resolution of $L\tau$), a pore current pulse of width $L\tau$ can be detected and an average over this interval computed. The result is a sequence of average signal values from which in principle the binary-coded primary sequence can be extracted. Alternatively the average over an interval of width $L\tau$ can be approximated by the number of 2's (or 1's) in that interval. Let $N_2(i)$ be the number of 2's in the interval $[(2i\tau, (2i+1)\tau], i = 0,1,2,...$. The sequence of (continuous) average values over alternating intervals of width $L\tau$ can now be approximated by the sequence of integers $\{N_2(i); i = 0,1,2,...\}$ corresponding to the number of 2's in $[0, L\tau], [2L\tau, 3L\tau], [4L\tau, 5L\tau], etc$. Since $0 \leq N_2(i) \leq L$, this is a sequence of numbers in base $L+1$. Subsequences of length $K$ from the sequence, that is, $N_2(j) N_2(j+1) ... N_2(j+(K-1)$, $j \geq 0$, can now be used as identifiers if they identify the parent protein uniquely. ESM File 2 has two examples showing how this works.

A second sequence $\{N_2'(i); i = 0,1,2,...\}$ can be defined with sample intervals $[L\tau, 2L\tau], [3L\tau, 4L\tau], [5L\tau, 6L\tau], ...$ (For example, in protein 0 above this gives _4_3_1_5_3_5, where _ stands for the sample interval $[2k, (2k+1)L\tau], k \geq 0$). By counting 1's instead of 2's two more sequences $\{N_1(i); i = 0,1,2,...\}$ and $\{N_1'(i); i = 0,1,2,...\}$ can be defined. From these four sequences subsequences of length $K$ can be examined to determine if they are identifiers. The total number of identified proteins is then the cardinality of the union of the four sets of proteins identified by subsequences of the four sequences $\{N_2(i); i = 0,1,2,...\}, \{N_2'(i); i = 0,1,2,...\}, \{N_1(i); i = 0,1,2,...\}, and \{N_1'(i); i = 0,1,2,...\}$. More generally sample intervals over which averaging is done can start anywhere along the primary sequence.

Table 2 gives the results for different values of $L$ and $K$; it shows a trade off between the bandwidth and the number of proteins identified. With $L = 5$ and $K = 8$ the bandwidth is reduced by a factor of 10 while the number of proteins identified in $H. pylori$ falls from 1524 (98.13%) to 1372 (88.35%). (This process can be repeated with 1's instead of 2's, but the increase is marginal; with $L = 5$ and $K = 8$ the number goes up from 1372 to 1375.) ESM File 3 contains a complete list of protein identifiers for $H. pylori$ based on averaged subsequences for the case $L = 5$ and $K = 8$.

For other hardware-based approaches to bandwidth reduction see Item 6 in Section 4.
Table 2. Bandwidth reduction with averaging. Average over alternating windows of width \( L \) (= length of subsequence) is given by number of 2's in subsequence binary code. Resulting sequence of averages is an \((L+1)\)-ary sequence; an id is an \((L+1)\)-ary subsequence thereof of length \( K \). Data for \( H. pylori \) (1553 sequences). Id’d = Number of proteins identified in proteome.

<table>
<thead>
<tr>
<th>L</th>
<th>5</th>
<th>5</th>
<th>5</th>
<th>6</th>
<th>6</th>
<th>6</th>
<th>6</th>
<th>8</th>
<th>8</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>6</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Id’d</td>
<td>849</td>
<td>1372</td>
<td>1346</td>
<td>1054</td>
<td>1337</td>
<td>1271</td>
<td>1185</td>
<td>1197</td>
<td>1227</td>
<td>1118</td>
</tr>
</tbody>
</table>

3.3 The homopolymer problem
The homopolymer problem refers here to the difficulty in identifying successive residues from the same subset as they generate the same (binary) output value. With a thick (8-10 nm) biological or synthetic pore, multiple (typically 4 to 8) residues are resident in the pore at any time during translocation so that the boundary between two successive values may be hard to detect, although correlations in the measured signal can often provide useful information. Thus in [13] the blockade current was found to correlate well with four contiguous residues.

The averaging technique of Section 3.2 mitigates the homopolymer problem. As averaging is done over a whole interval, the boundary between two successive residues is no longer relevant.

For other solutions based on hardware or software, see Item 7 in Section 4 below.

3.4 Quantifying proteins in a mixture
The procedure described here can be used to quantify proteins in a mixture of proteins \( \{M_i; i = 1, 2, \ldots\} \), where \( M_i \) is the number of molecules of the \( i \)-th protein. Let \( M_{\text{tot}} = \sum_i M_i \). If molecules enter the pore in some random order, then after a sufficiently long run \( M_i/M_{\text{tot}} \) can be estimated as \( \tilde{M}_i/\tilde{M}_{\text{tot}} \) where \( \tilde{M}_i \) and \( \tilde{M}_{\text{tot}} \) are the measured number of molecules of protein \( i \) and the total measured molecules over the run. Quantification time can be reduced significantly by using an array of pores.

4. Discussion
Some potential implementation and other relevant issues are now considered.
1) The method described here works on single unbroken protein molecules, no proteolysis is done; it is thus free from the vagaries of the latter [3]. As there is no degradation the sample can be reused/resequenced.
2) Searching for a measured subsequence in the proteome will require both forward and reverse matches because the protein may enter the pore C- or N-terminal first.
3) Equation 2 assumes that successive residues in a protein are independent. This is not true in practice as there are inherent correlations. The latter can be extracted from the pore current signal and used in error correction, this leads to increased reliability. Software used in nanopore-based DNA sequencing routinely uses this kind of information to improve sequencing accuracy [26,27].
4) Depending on the primary sequence a protein may carry only a weak charge so that entry into and/or translocation through the pore may be a problem. One solution [16] to this is to attach a negatively charged carrier molecule like DNA to the protein molecule; another may be based on dielectrophoretic trapping [28].
5) Charged residues on the pore wall tend to interfere with the passage of an analyte when the latter has charged residues. (Seven 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.) To resolve this the wall charge can be neutralized in some way. With DNA as the analyte a lipid coat has been shown to do this [29].
6) Hardware solutions to the bandwidth reduction problem of Section 3.2 include use of: a) a room-temperature ionic liquid (RTIL), which is a high viscosity electrolyte that can slow down an analyte by a factor of \( \sim 200 \) [30]; b) an opposing hydraulic pressure field [31]; c) an enzyme (‘ unfoldase’) to unfold the protein molecule before it enters into and thus slow its passage through [32] the pore; or d) ligands attached to the protein or the pore [33].
7) The homopolymer problem of Section 3.3 has also been addressed with hardware. If a single atom layer of graphene [34] or molybdenum sulphide (MoS\(_2\)) [30] is used for the membrane, or a biological pore with a narrow constriction (MspA, CsgG) [35,36], or an adapter such as \( \beta \)-cyclodextrin in \( \alpha \)-HL [37], roughly one residue will be resident in the pore or its constriction or in the adapter during translocation. Software based on a Hidden Markov Model [26] or Viterbi algorithm [27] may also be used to computationally separate successive residues with near identical (binary-coded) blockade levels.

5. Conclusion
Unlike most recent work in protein identification, which is usually aimed at identifying specific single proteins or
their variants, the method proposed in this Letter can identify an arbitrary protein in a large set such as a proteome. The availability of sub-nanopores capable of measuring residue volumes with adequate resolution makes the proposed method both feasible and practical. Reducing the number of current blockade levels to be measured from 20 to two effectively removes a major obstacle to the use of nanopores for protein identification. All of this points to the near-term prospect of implementing an easy-to-use hand-held device that can identify and quantify proteins in mixtures while requiring minimal sample preparation without any proteolysis.

References


