bioRxiv preprint doi: https://doi.org/10.1101/119313; this version posted April 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Protein identification with a binary code, a nanopore, and no proteolysis

G. Sampath

Abstract. If protein sequences are recoded with a binary alphabet derived from a division of the 20 amino acids into two subsets, a protein can be identified from its subsequences by searching through a recoded sequence database. A binary-coded primary sequence can be obtained for an unbroken protein molecule from current blockades in a nanopore. Only two (instead of 20) blockade levels need to be recognized to identify a residue's subset; a hard or soft detector can do this with two current thresholds. Computations were done on the complete proteome of *Helicobacter pylori* (http://www.uniprot.org; database id UP000000210, 1553 sequences) using a binary alphabet based on published data for residue volumes in the range ~0.06 nm^3 to ~0.225 nm^3. Assuming normally distributed volumes, more than 93% of binary subsequences of length 20 from the primary sequences of *H. pylori* are correct with a confidence level of 90-95%; they can uniquely identify over 98% of the proteins. Recently published work shows that a 0.7 nm diameter nanopore can measure residue volume with a resolution of ~0.07 nm^3; this makes the procedure described here both feasible and practical. This is a non-destructive single-molecule method without the vagaries of proteolysis.

1. Overview

It is shown that proteins in a proteome can be uniquely identified from subsequences of primary sequences based on a binary code. The code is derived from a division of the standard set of 20 amino acids into two subsets based on their volumes [1]. Using a sample proteome (*Helicobacter pylori*) calculations show that the codes of subsequences 20 residues long are correct at a 90-95% confidence level, and that over 98% of the proteins in the proteome can be identified by exhaustively searching for the recoded subsequences in the recoded proteome database. The scheme described here can be translated into practice with a nanopore.

The proposed method is first analyzed computationally before looking at implementation issues. The computational study is in three steps: 1) Divide the set of amino acids into two ordered subsets S_1 and S_2; 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 uniquely identify the protein to the exclusion of every other protein.

2. An amino acid partition

Table 1 shows the standard set of 20 amino acids **AA** = {G,A,S,C,D,T,N,P,V,E,Q,H,M,I,L,K,R,F,Y,W} grouped into two subsets S_1 and S_2 by volume, where

$$S_1 = \{G,A,S,C,D,T,N,P: 59.9 \leq \text{volume} \leq 123.3\} \quad (1a)$$
$$S_2 = \{V,E,Q,H,M,I,L,K,R,F,Y,W: \text{volume} \geq 138.8\} \quad (1b)$$

Table 1. The 20 amino acids in increasing order of volume. AA = Amino acid; Mean (μ) = Mean volume in nm^3; SD (σ) = Standard deviation of volume in nm^3. Shading shows division into ordered subsets S_1 and S_2; dividing line is between P and V. Data adapted from [1].

AA	Mean	SD	AA	Mean	SD	AA	Mean	SD	AA	Mean	SD
G	59.9	2.2	T	118.3	2.3	Q	145.1	5.1	K	172.7	5.9
A	87.8	2.3	N	120.1	4.1	H	156.3	6.1	R	188.2	9.6
S	91.7	1.8	P	123.3	1.8	M	165.2	1.8	F	189.7	7.4
C	105.4	5	V	138.8	3.6	I	166.1	3.4	Y	191.2	8
D	115.4	2.2	E	140.9	5.3	L	168	4.3	W	227.9	3.8

The dividing line is chosen between P and V so that the two sets have roughly the same size and the difference between the volumes of P and V is relatively large. (There are higher volume differences between Y and W and

bioRxiv preprint doi: https://doi.org/10.1101/119313; this version posted April 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

between G and A, but the resulting subset sizes are lopsided.) An amino acid X ϵ **AA** has a binary code C(X) given by

$$C(X) = i, \qquad X \in S_i;\ i = 1, 2 \qquad (2)$$

The 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 then $C(X_1)\ C(X_2)\ \ldots\ C(X_n)$.

3. Binary subsequences of primary sequences as protein identifiers

Assume that amino acid volumes are normally distributed. Figure 1 shows the distributions for the 20 amino acids based on μ and σ values in Table 1.

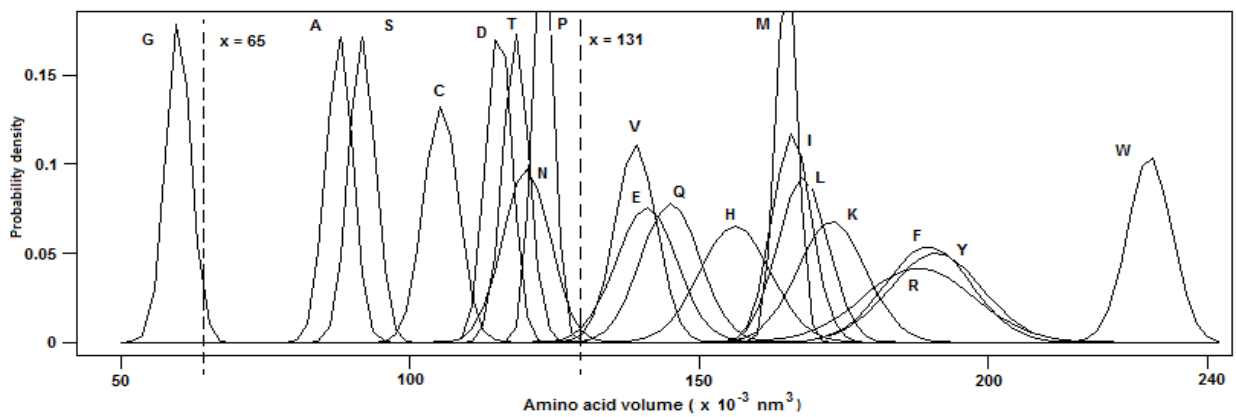

Figure 1. Normal distributions of amino acid volume with mean and standard deviation from Table 1.

Let F(x; μ, σ) be the cumulative normal distribution function with mean μ and standard deviation σ:

$$F(x;\mu,\sigma) = (1/\sqrt{2\pi}\sigma) \int_0^x . \exp(-(x-\mu)^2/2\sigma^2) \qquad (3)$$

The complement of F(x; μ, σ) is

$$G(x;\mu,\sigma) = 1 - F(x;\mu,\sigma) \qquad (4)$$

Let the mean volume and standard deviation for amino acid aa be μ_{aa} and σ_{aa} (see Table 1). If the volume of an amino acid is considered correct when it lies between two threshold levels T_1 and T_2, then $e_{aa}(T_1,T_2)$, the error for amino acid aa, is given by

$$aa \in S_1:\ e_{aa}(T_1,T_2) = F(T_1;\ \mu_{aa}, \sigma_{aa}) + G(T_2;\ \mu_{aa}, \sigma_{aa}) \qquad (5a)$$
$$aa \in S_2:\ e_{aa}(T_1,T_2) = F(T_2;\ \mu_{aa}, \sigma_{aa}) \qquad (5b)$$

Assuming for the present that successive residues in a sequence are independent the confidence level (probability that the binary code for a protein sequence **X** = $X_1\ X_2\ \ldots\ X_n$ is correct) is given by

$$c_X(T_1,T_2) = \Pi_{i=1\,..\,n}\,(1 - e_{Xi}(T_1,T_2)) \qquad (6)$$

The confidence level for a subsequence code can be computed from Equation 6 for subsequences of the primary sequences in a proteome coded as in Equation 2. The proteome of the gut bacterium *Helicobacter pylori* (Uniprot id UP000000210, 1553 sequences, http://www.uniprot.org) is used as an example. Referring to Figure 1 the widest separation near the middle (leading to roughly equal-sized subsets) occurs between the curves for P and V (this is also

bioRxiv preprint doi: https://doi.org/10.1101/119313; this version posted April 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the line of division in Table 1). This value of 0.131 nm^3 is used for T_2. A second threshold T_1 is set at 0.05 nm^3, which is below the mean for the lowest volume amino acid G. (In implementation with a nanopore, this is required, see later.) Figure 2 shows the confidence level of binary codes for all subsequences in *H. pylori* that are of length 15, 20, or 25 (data point symbol ♦) for $T_1 = 0.05$ and $T_2 = 0.131$. As expected the level drops off as subsequence length increases because of the continued multiplication in Equation 6. (This also means that the full protein sequence cannot be used as an identifier, the confidence level falls to 0 rapidly.)

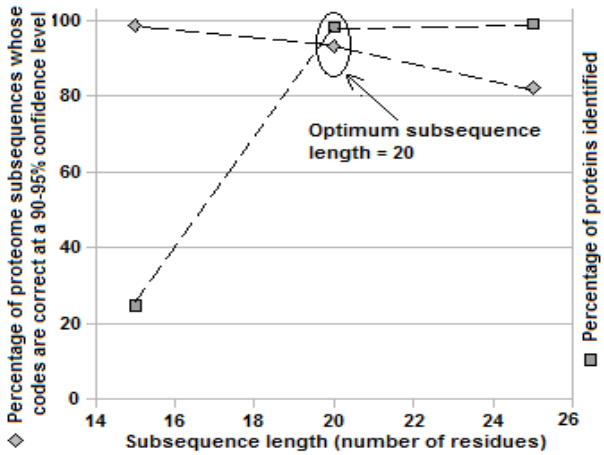

Figure 2. Percentage of protein subsequences in *H. pylori* whose volume-based binary codes have a confidence level of 90-95% vs subsequence length (♦). Percentage of proteins identified uniquely from their subsequences vs subsequence length (■). (Lines connecting data points added for visual aid.)

Using routine search methods subsequences of given lengths from every protein in the recoded proteome may be exhaustively compared with subsequences in every other protein to determine if they uniquely identify their container proteins. To reduce computation time search subsequences from a protein sequence of length S are spaced $\Delta = 5$ residues apart instead of at every position along the primary sequence. Thus subsequences of length L starting at positions 0, 5, 10, ..., S-4 are used. The percentages of proteins so identified are given for L = 15, 20, and 25 in Figure 2 (data point symbol ■). As expected, with larger L more proteins are identified. The number of proteins identified goes from ~25% with L = 15 to ~98% with L = 20 and ~99% with L = 25. The increase from L = 20 to L = 25 is less than 1%; thus subsequences longer than 20 have diminishing returns. (Using $\Delta = 5$ is justified because the number of identifiers is close to 100%; the gains from reducing Δ are minimal going from L = 20 to L = 25. With L = 15, setting Δ to 1 leads to the number of identified proteins going from 135 (8.69%) to 384 (24.66%), which is still a very low identification rate.)

From Figure 2 it can be concluded that L = 20 is a near-optimal length for identifier subsequences as it simultaneously optimizes the number of proteins identified and the confidence level of their subsequence codes.

The above computational procedure can be translated into practice by using a nanopore to measure residue volume.

4. Using a nanopore to identify proteins from binary subsequences

Nanopore sequencing of proteins [2] uses an electrolytic cell in which two chambers *cis* and *trans* containing an electrolyte like KCl or NaCl are separated by a membrane with a nano-sized pore. A potential difference across the membrane results in an ionic current through the pore. When a protein molecule is introduced into *cis* it translocates through the pore and causes a blockade of the ionic current. By measuring the level of the blockade, a residue can be identified. With proteins no enzymatic digests are required; the analyte is a denatured (unfolded) form of unbroken protein. Normally this would require a blockade current resolution that can discriminate among 20 residue types (the standard set of amino acids is assumed). Such resolution is virtually unattainable in practice, especially with noise present.

The method proposed here directly addresses this problem and resolves it by reducing protein identification to measuring two (rather than 20) levels in the blockade current signal generated by an intact translocating protein.

bioRxiv preprint doi: https://doi.org/10.1101/119313; this version posted April 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Following this subsequences in the resulting binary sequence are found such that they can uniquely identify the protein in a binary-coded proteome database. As the blockade level is directly related to the volume of one or more residues translocating through the pore, residue volume is used here as a proxy for the blockade current.

It was recently shown experimentally that a sub-nanometer-diameter (0.7 nm) nanopore can measure residue volume with a resolution of 0.07 nm^3 [3]. The threshold values T_1 and T_2 chosen earlier can be understood in light of this. If T_2 is considered in isolation, residues with mean volume > T_2 will be identified as belonging to the residue subset S_2, and those < T_1 as belonging to S_1. However blockades due to residues in the lower volume group have to be detected/differentiated from the quiescent ionic current (the current when there is no protein in the pore either fully or partially). This is why a second and lower threshold T_2 corresponding to a volume of ~0.05 nm^3 was set in the computational analysis in Section 3.

5. Discussion

Some potential implementation-related issues are now considered.

1) The method described here works on single unbroken protein molecules without any proteolysis. It is thus free from the vagaries of the latter. There is no degradation of the sample, which can be reused.

2) In practice, matching a measured subsequence with a subsequence in the (recoded) proteome will require a forward match as well as a reverse match because the protein may enter the pore C- terminal or N-terminal first.

3) Equation 6 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; see [4] for example.

4) Almost every protein in the H. *pylori* proteome has a large number of identifying subsequences so this reduces the false detection rate (FDR) considerably.

5) Depending on the primary sequence a protein may carry only a weak charge, if at all, so that entry into and translocation through the pore may be a problem. One solution [5] is to attach a strongly charged carrier molecule like DNA (which has a uniform charge along its backbone) to the protein molecule.

6) Charged residues on the pore wall tend to interfere with the passage of an analyte, especially 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.) One possible solution is to neutralize the wall charge in some way. With DNA as the analyte a lipid coat has been shown to have this effect [6].

7) A persistent problem in nanopore-based sequencing is homopolymer recognition, which in the present case means successive residues from the same subset generating 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 such values may be difficult to identify, although correlations in the measured signal can often provide useful information. Thus in [7] the blockade current was found to correlate well with four contiguous residues. On the other hand the severity of the problem can be reduced by using a single atom layer of graphene [8] or molybdenum sulphide (MoS_2) [9] for the membrane, or a biological pore with a narrow constriction as in MspA [10] or in CsgG [11], or an adapter such as β-cyclodextrin in αHL [12]. In this case roughly one residue will be resident in the pore or its constriction (or in the adapter) any time during translocation. Software based on hidden Markov models [4] or the Viterbi algorithm [13] can also be used to computationally separate successive residues with near identical blockade levels.

8) The high speed with which an analyte translocates through the pore makes it difficult for a detector with insufficient bandwidth to detect changes in the blockade current level [14]. One possible solution to this is the use of a room-temperature ionic liquid (RTIL) for the electrolyte. RTIL is a high viscosity liquid that can slow down an analyte by a factor of about 200 [9]. Another solution is to use an opposing hydraulic pressure field [15].

9) The method described in this report is aimed at identification of any protein belonging to an arbitrarily large set such as a proteome, rather than particular ones [16-18].

10) The approach described here targets known proteins. Thus *de novo* identification is outside its scope, at least in its present form.

References

[1] S. J. Perkins, "Protein volumes and hydration effects", *Eur. J. Biochem.*, 1986, **157**, 169-180.

[2] S. Acharya, S. Edwards, and J. Schmidt, "Nanopore protein detection and analysis", *Lab on a Chip*, 2015. DOI: 10.1039/c5lc90076j.

bioRxiv preprint doi: https://doi.org/10.1101/119313; this version posted April 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[3] 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 Nanotech*., 2016, **11**, 968-976. doi:10.1038/nnano.2016.120.
[4] J. Schreiber and K. Karplus, "Analysis of nanopore data using hidden Markov models", *Bioinformatics*, 2015, **31**, 1897–1903.
[5] N. A. W. Bell and U. F. Keyser, "Specific protein detection using designed DNA carriers and nanopores", *J. Am. Chem. Soc*., 2015. DOI: 10.1021/ja512521w.
[6] A. Sischka, L. Galla, A. J. Meyer, A. Spiering, S. Knust, M. Mayer, A. R. Hall, A. Beyer, P. Reimann, A. Gölzhäuser, and D. Anselmetti, "Controlled translocation of DNA through nanopores in carbon nano-, silicon-nitride- and lipid-coated membranes", *Analyst*, 2015. DOI: 10.1039/c4an02319f.
[7] M. Kolmogorov, E. Kennedy, Z. Dong, G. Timp, and P. Pevzner, "Single-molecule protein identification by sub-nanopore sensors", 2016, *arXiv.org*:1604.02270v1 [q-bio.QM].
[8] M. Drndic, "Sequencing with graphene pores", *Nature Nanotech*., 2014, **9**, 743.
[9] J. Feng, K. Liu, R. D. Bulushev, S. Khlybov, D. Dumcenco, A. Kis, and A. Radenovic. "Identification of single nucleotides in MoS2 nanopores", *Nature Nanotech*., 2015, doi: 10.1038/nnano.2015.219.
[10] T. Z. Butler, M Pavlenok, I. M. Derrington, M Niederweis, and J. H. Gundlach, "Single-molecule DNA detection with an engineered MspA protein nanopore," *PNAS*, 2008, **105**, 20647–20652.
[11] P. Goyal, P. V. Krasteva, N. VanGerven, F. Gubellini, I. Van Den Broeck, A. Troupiotis-Tsailaki, et al., "Structural and mechanistic insights into the bacterial amyloid secretion channel CsgG", *Nature*, 2014, **516**, 250-253.
[12] 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*., 2009, **4**, 265-270.
[13] W. Timp, J. Comer, and A. Aksimentiev, "DNA base-calling from a nanopore using a Viterbi algorithm", *Biophys. J*., 2012, **102**, L37-39.
[14] S. Carson, and M. Wanunu, "Challenges in DNA motion control and sequence readout using nanopore devices", *Nanotech*., 2015, **26**, 074004.
[15] 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.
[16] R. Wei, V. Gatterdam, R. Wieneke, R. Tampe, and U. Rant, "Stochastic sensing of proteins with receptor-modified solid-state nanopores". *Nature Nanotech*. 2012, **7**, 257–263.
[17] C. B. Rosen, D. Rodriguez-Larrea, H. Bayley, "Single-molecule site-specific detection of protein phosphorylation with a nanopore", *Nature Biotechnol*. 2014, **32**, 179–181.
[18] J. Nivala, L. Mulroney, G. Li, J. Schreiber, and M. Akeson, "Discrimination among protein variants using an unfoldase-coupled nanopore", *ACS Nano*, 2014, 8, 12365–12375.