Protein fingerprinting with a binary alphabet and a nanopore

G. Sampath

Abstract. Proteins can be partitioned into eight mutually exclusive sets of peptides and recoded with a binary alphabet obtained by dividing the 20 amino acids into two ordered sets based on volume. By searching for these binary-coded peptides in a protein sequence database, their container proteins can be identified. Over 89.7% of 20207 curated proteins in the human proteome (http://www.uniprot.org; database id UP000005640, *H. sapiens*) can be so identified. This procedure can be translated into practice. Thus standard chemical procedures can be used for partitioning and a nanopore can be used to obtain binary coded sequences for partitioned peptides. In the latter case, recently published work has shown that a sub-nanometer-diameter pore can measure residue volume with a resolution of ~0.07 nm^3. This can be used to distinguish between the two sets of residues defined above; a detector with two thresholds outputs a binary sequence for a partitioned peptide from the nanopore current signal. Using normal distributions of amino acid volume data from the literature, routine computations show that ~98% of the protein-identifying peptides in the curated human proteome have binary codes that are correct with a confidence level exceeding 85%. Similar results are presented for the proteomes of baker's yeast (*S. cerevisiae*), the pathogen *E. coli*, and the gut bacterium *H. pylori*.

1. Introduction

Protein fingerprinting is the process of identifying a protein from a subsequence by searching for the latter in a protein sequence database [1]. In the present report it is shown that in principle close to 90% of all proteins in the human proteome (*H. sapiens*) can be identified by partitioning proteins into eight mutually exclusive sets of peptides, recoding them with a binary alphabet based on amino acid volume, and searching for the recoded peptide sequences in the database. The procedure proposed here can be realized in practice with available chemical procedures for partitioning and a nanopore [2,3] to obtain the binary sequence of a partitioned peptide. Statistical calculations show that the confidence level of the binary sequences of ~98% of all protein-identifying peptides in the partition exceeds 85%. Similar results are given for three other proteomes (*S. cerevisiae*, *E. coli*, *H. pylori*).

2. A peptide partition

Consider peptide sequences of the form $X_1Z^*X_2$, where $X_1 \in \{\lambda,K\}$, $X_2 \in \{\lambda,D,E,R\}$, Z is one of the remaining 16 residue types, $Z^* \equiv 0$ or more occurrences of Z, and λ is the empty string. This leads to the peptide sequence partition **P** = {KZ^*D, KZ^*E, KZ^*R, KZ^*, Z^*D, Z^*E, Z^*R, Z^*}. Such a partition can be obtained in practice by using sequential proteolysis and electrophoretic separation, see Section 5 below. Additional information is available in [4], which also includes other practically realizable partitions.

3. Amino acid partitions

The standard 20 amino acids may be ordered on a physical or chemical property such as volume, mass, charge, diffusion constant, or mobility. They may then be divided into 2 or more ordered subsets at fixed points along the ordering. For example, Table 1 shows them ordered on volume and grouped into two subsets. The dividing line is chosen between P and V so that the two subsets have nearly the same size and the difference between the volumes of P and V is substantial. (There are higher volume differences between Y and W and between G and A, but the resulting subset sizes are lopsided.) The amino acids can now be coded with a binary code: S_1 = {G,A,S,C,D,T,N,P: $59.9 \leq$ volume ≤ 123.3} → 1, S_2 = {V,Q,E,H,I,L,M,K,R,F,Y,W: volume ≥ 138.8} → 2.

Table 1. Table of amino acids sorted by volume

AA = Amino acid; Volumes in 10^{-3} nm^3. Background shading shows grouping of amino acids into 2 subsets. Data from [5].

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

4. Peptides that uniquely identify their parent proteins in the human proteome

Consider the human proteome database UP000005640 at http://www.uniprot.org. There are 20207 curated sequences in this database. Column 2 in Table 2 gives the number of peptides in each subset of the partitioned database (as defined in Section

2). Let each protein sequence in the proteome and each peptide (subsequence) in each subset of the partition be recoded with the binary code given in Section 3. The number of proteins in the proteome that are identified by these recoded peptides in each partition subset is computed using straightforward search methods and given in Column 3. The corresponding protein identification efficiencies are given in Column 4. The total number of proteins that are uniquely identified by one or more peptides from the full partition **P** is the union of the sets of proteins identified by peptides from the individual sets in the partition (Rows 2 through 9). The size of this union and the identification efficiency are given by the entries in Row 10 of Columns 3 and 4. The corresponding data for the full alphabet and for alphabets of sizes 3 and 4 are given in Columns 5 through 10 for comparison. (The 3-character alphabet is given by the following mapping: {G,A,S: 59.9 ≤ volume ≤ 91.7} → 1, {C,D,T,N,P: 105.4 ≤ volume ≤ 123.3} → 2, {V,Q,E,H,I,L,M,K,R,F,Y,W: volume ≥ 138.8} → 3. The 4-character alphabet corresponds to: {G,A,S,C: 59.9 ≤ volume ≤ 105.4} → 1, {D,T,N,P: 115.4 ≤ volume ≤ 123.3} → 2, {V,Q,E,H,I,L,M,K}: 138.8 ≤ volume ≤ 172.7} → 3, {R,F,Y,W: volume ≥ 188.2} → 4.

Protein identification is more reliable if multiple peptides are available to identify a protein. Conversely the fewer proteins there are with a single identifying peptide the lower the false detection rate (FDR). With the binary alphabet from Section 3, the number of proteins identified by a single peptide is 2626 (13% of 20207), the number with two or more identifiers is 15507 (76.74%), and the number that cannot be identified uniquely is 2074 (10.26%).

Table 2. Protein identification efficiency with a binary alphabet for the human proteome (Uniprot database UP000005640; 20207 manually curated sequences). Data for alphabets of size 3 and 4 and the full alphabet included for comparison.

Subset of partition	Total no. of peptides in subset	Binary alphabet		Alphabet size 3		Alphabet size 4		Full alphabet	
		No. of proteins identified (a)	Identification efficiency (b)	No. of proteins identified (a)	Identification efficiency (b)	No. of proteins identified (a)	Identification efficiency (b)	No. of proteins identified (a)	Identification efficiency (b)
KZ*R	139423	5196	25.67%	8265	40.83%	9788	48.35%	14247	70.51%
KZ*D	125351	4602	22.73%	7378	36.45%	8781	43.38%	12983	64.25%
KZ*E	194024	5171	25.54%	8400	41.50%	9982	49.31%	14304	70.79%
KZ*	189713	5143	25.41%	8128	40.15%	9607	47.46%	13736	67.98%
Z*R	499784	10356	51.16%	14183	70.06%	15560	76.87%	18305	90.59%
Z*D	411189	9117	45.04%	13031	64.37%	14476	71.51%	17691	87.55%
Z*E	609872	9932	49.06%	14078	69.55%	15481	76.48%	18254	90.34%
Z*	345450	9411	46.49%	13644	67.40%	15172	74.95%	18467	91.39%
P	2514806	18133	**89.74%**	19302	**95.35%**	19532	**96.66%**	19885	**98.4%**

(b) Protein identification efficiency (%) = Total number of proteins in proteome uniquely identified by the identifying peptides in column marked (a) * 100/20207. **P** = union of all 8 subsets of the partition

Table 3 shows results obtained with the binary alphabet of Section 3 for the proteomes of three other species: baker's yeast (*Saccharomyces cerevisiae*), the common pathogen *Escherichia. coli*, and the gut bacterium *Helicobacter pylori*. (Sequence data for all three are from the Uniprot website, the respective ids are 4932, 562, and 210.)

Note that Table 2 is based on comparing a partitioned peptide sequence with protein sequences in the Uniprot database UP000005640. This database contains a curated *canonical* set of proteins as well as variant sequences for the protein from the parent gene, such as isoforms, polymorphisms, and post-translational modifications. In the Uniprot KnowledgeBase (KB) a protein sequence is *canonical* "if it satisfies one or more of the following conditions: a) It is the most prevalent sequence among all variants; b) It is the most similar to orthologous sequences found in other species; c) It provides the clearest picture among all variants in terms of the latter. If this additional information is not available, the longest sequence among the variants is taken as the canonical one." [6]. Identification consists of first comparing a partitioned peptide sequence with the set of canonical protein sequences. If there is no match or additional information is sought, comparison may be done with the variant sequences in the database.

5. Implementation considerations

There are two major steps in the procedure proposed above: 1) creating the partition; and 2) obtaining the binary-coded sequence of a partitioned peptide.

1) While a study of general partitions could be useful in a wider sense (and may also be of theoretical interest), the present work focuses on partitions that can be implemented in practice. Procedures based on sequential proteolysis and the use of isoelectric focusing (IEF) to separate partition subsets are discussed in [4], where a four-step process that uses four different proteases in some order for creating the 8-set partition defined in Section 2 is outlined. (Three more partitions are given in

[4].) In comparison, mass spectrometry [1,7], the most commonly used peptide/protein sequencing method, relies on the use of trypsin digests to create peptide fragments that end in K or R (or neither), effectively creating a three-set partition of the form $\{Z^*K, Z^*R, Z^*\}$. This is too coarse for the present purpose, it is also only approximate as there are many peptides in the output that violate the partition pattern [8]. (In like manner the extent to which partitioned peptides from the procedure proposed here satisfy the 8-set partition defined in Section 2 in practice can be known only when empirical data become available.)

Table 3. Protein identification efficiency with a binary alphabet for three proteomes with N sequences Yeast (UP000005640; N = 6721), *E.coli* (UP000005640, N = 4306), and *H. pylori* (UP000005640, N = 1553)

Subset of partition	Yeast *(S. cerevisiae)*			*E. coli*			*H. pylori*		
	Total no. of peptides in subset	No. of proteins identified [(a)]	Identification efficiency [(b)]	Total no. of peptides in subset	No. of proteins identified [(a)]	Identification efficiency [(b)]	Total no. of peptides in subset	No. of proteins identified [(a)]	Identification efficiency [(b)]
KZ^*R	41138	1966	29.25	14549	1214	28.19	5829	447	28.78
KZ^*D	49183	2064	30.71	15198	1294	30.05	9101	563	36.25
KZ^*E	56601	1998	29.73	16152	1278	29.68	13381	629	40.50
KZ^*	73046	2358	35.08	13851	1122	26.06	15627	777	50.03
Z^*R	93629	2937	43.70	60185	2339	54.32	11119	629	40.50
Z^*D	125007	3059	45.51	54654	2309	53.62	14361	741	47.71
Z^*E	137939	3018	44.90	61997	2370	55.04	20529	809	52.09
Z^*	114940	3600	53.56	37682	2186	50.77	22509	967	62.27
P	691483	5876	**87.43%**	274268	4052	**94.10**	112456	1451	**93.43%**

[(b)] Protein identification efficiency (%) = Total number of proteins in proteome uniquely identified by the identifying peptides in column marked [(a)] * 100/N. N = number of proteins in proteome (6721, 4306, and 1553 respectively). **P** = union of all 8 subsets of the partition

2) The most widely used method of peptide/protein sequencing, namely mass spectrometry, works in the bulk [7]. Other methods for partial sequencing of peptide fragments based on the detection of single molecules have been proposed and/or studied recently. They may be labeled (with, for example, fluorescent tags) or unlabeled (with a nanopore, for example). In the former case peptide fragments are selectively labeled with as many different dyes as the size of the reduced alphabet chosen. The labeled fragments are sequenced in one of two ways: 1) pinning the fragments to a substrate, cleaving successive residues from a fragment by Edman degradation, and using a fluorescent readout to read the cleaved residues [9]; 2) using a protein translocase to pass the fragments through a nanochannel and Förster resonance energy transfer (FRET) to detect a labeled residue as it moves past the enzyme [10]. Calculations show that as few as 2 tagged residue types may be sufficient to partially sequence a peptide; this partial sequence is then used to identify its parent protein by comparison with the set of protein sequences in a sequence database.

In contrast, no analyte immobilization or labeling of any kind is used in nanopore-based sequencing [2,3]. This is an electrical method based on the use of an electrolytic cell in which a membrane with a nano-sized pore bridges two compartments (usually referred to as *cis* and *trans*) containing an electrolyte, typically KCl. A potential difference across *cis* and *trans* causes an ionic current. When an analyte such as a single protein or peptide molecule is introduced into *cis* it causes a blockade of the quiescent ionic current as it translocates through the pore. The magnitude of the blockade due to a residue in the translocating peptide is partly determined by residue volume, so the blockade current level can in principle be used to identify the residue. This approach can lead to a sequencing device that is significantly smaller than one based on fluorescent methods and is easier to operate. However, with 20 amino acids to distinguish in the peptide the required current resolution cannot be easily realized (especially in the presence of noise). One way to deal with this seemingly insurmountable problem is to reduce the size of the alphabet, as discussed in the next section.

The next two sections show how a binary coded sequence for a partitioned peptide can be obtained with a nanopore and used for protein identification.

6. Nanopore sequencing of partitioned peptides with small alphabets

A recent report [11] has shown that by using a sub-nanometer-diameter nanopore a residue can be detected with a volume resolution of 0.07 nm^3. Assuming that the blockade due to a residue is directly related to its volume, the blockade current can be used to call residues in a peptide. However with 20 amino acids to resolve, the required current resolution is virtually unattainable, and noise makes it worse. In [12] an attempt was made to partially resolve this problem by dividing the amino acids into four subsets labeled *Minuscule*, *Small*, *Intermediate*, and *Large* (which is almost identical to the 4-character alphabet in Section 3); the labels are descriptive of the volume excluded in the pore by the corresponding residue in

the protein as it translocates through the pore. The level of the measured signal is mapped to one of the four subsets and the resulting peptide encoded with the alphabet {M, S, I, L}. Further improvements can be had by recognizing that protein sequences generally carry inherent correlations. One such is obtained by reading current blockades due to four successive residues (4-mers) rather than individual ones in sequence. While with the full alphabet this would require the ability to distinguish among 20^4 4-mers, combining this approach with the 4-character alphabet {M, S, I, L} results in only 35 unique volume combinations, a considerable drop from 160000. This approach makes it possible to read one protein out of 9999 human proteins with < 10 % error from a single 4-mer blockade [11]. While the result is significant the gain is not of much practical use.

On the other hand if the goal is protein identification rather than peptide sequencing the pessimism implied by the above result is considerably diminished if not dispelled by using the binary alphabet of Section 3. Read rates approaching practical levels can be achieved: as shown below, over 98% of protein-identifying peptides can be read correctly during measurement of the pore current with confidence levels exceeding 85%. This makes nanopore-based protein identification both feasible and practical.

In what follows, residue volume is used as a proxy for the blockade current.

7 Protein identification from binary sequences obtained with a subnanometer-diameter nanopore

As noted in Section 6, a residue volume resolution of 0.07 nm^3 is possible with a sub-nanometer-diameter nanopore. Referring to Table 1, with a blockade current threshold corresponding to a volume of ~0.13 nm^3 residues can be distinguished as belonging to one or the other of the two sets S_1 and S_2 shown shaded in the table. However blockades due to residues in the lower volume group have to be detected/differentiated from the quiescent ionic current (= current when there is no peptide present in the pore either fully or partially). For this reason a second and lower threshold needs to be set. Given a volume resolution of 0.07 nm^3, it can be set to ~0.06. (This also increases the error rate for G residues but the effect is not very significant, as seen below.)

The errors that result from using a binary code can be roughly estimated for different volume thresholds. Assuming residue volumes to be normally distributed with mean μ and standard deviation σ errors can be computed with the normal (Gaussian) error function. Normal distributions for the 20 amino acids based on the data in Table 1 (μ and σ) are shown in Figure 1. The widest separation occurs between the curves for P and V, this is also the line of division in Table 1, and it leads to a threshold value of 131 for the detector.

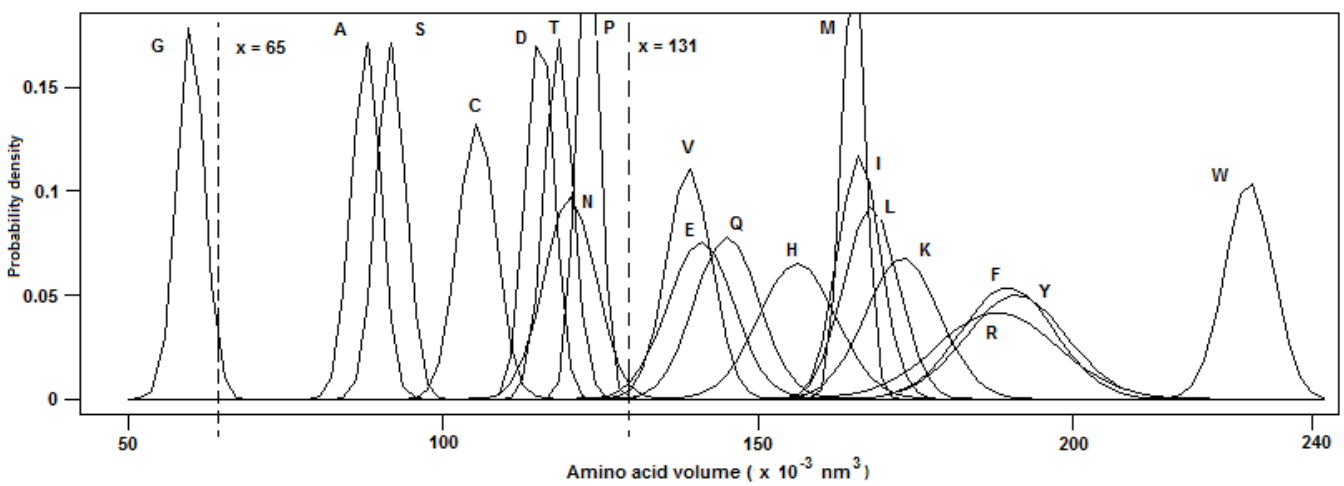

Figure 1. Normal distributions of amino acid volume with mean and standard deviation from Table 1.

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

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

and

$$G(x; \mu, \sigma) = 1 - F(x; \mu, \sigma) \quad (2)$$

Let thresholds be set at T_1 and T_2. Let the mean volume and standard deviation for amino acid aa be μ_{aa} and σ_{aa} (see Table 1), and the error in reading the volume of amino acid aa $e_{aa}(T_1,T_2)$. The errors for the 20 amino acids can be written as

$$aa = G: \quad e_G(T_1,T_2) = G(T_1;\ \mu_G, \sigma_G) \tag{3a}$$
$$aa \in S_1 - G: \quad e_{aa}(T_1,T_2) = F(T_1;\ \mu_{aa}, \sigma_{aa}) + G(T_2;\ \mu_{aa}, \sigma_{aa}) \tag{3b}$$
$$aa \in S_2: \quad e_{aa}(T_1,T_2) = G(T_1;\ \mu_{aa}, \sigma_{aa}) \tag{3c}$$

The probability that the measured binary code for a peptide $\mathbf{X} = X_1 X_2 \ldots X_n$ where X_i is one of the 20 amino acids is correct is given by

$$c_{\mathbf{X}}(T_1,T_2) = \Pi_{i=1..n}(1 - e_{Xi}(T_1,T_2)) \tag{4}$$

Figure 2 shows $c_{\mathbf{X}}(T_1,T_2)$ for $T_1 = 65$ and $T_2 = 130$ for four proteomes: human, yeast, *E. coli*, and *H. pylori*. The results indicate that close to 80% of all binary sequence codes for protein-identifying partition peptides are correct with a confidence level exceeding 90%. The percentage rises to 98% for confidence levels exceeding 85%.

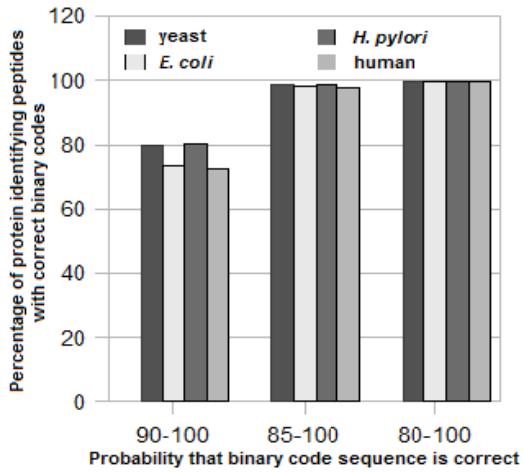

Figure 2. Fraction of protein identifying peptides obtained from a nanopore with two thresholds vs confidence level of binary code obtained

8 Discussion

One of the more persistent problems with nanopore-based sequencing approaches is homopolymer recognition, which arises from successive residues belonging to 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. The severity of the problem can be reduced by using a single atom layer of graphene [13] or molybdenum sulphide (MoS_2) [14] for the membrane, or a biological pore with a narrow constriction (MspA [15], CsgG [16]) or an adapter (β-cyclodextrin in αHL [17]). In this case roughly one residue will be resident in the pore (or its constriction) during translocation. Another possibility is to use an electrolytic cell with a tandem pair of nanopores and an exopeptidase [18]. Here the enzyme, which is situated below the first pore, cleaves successive residues from a peptide emerging from the first pore; the cleaved residues diffuse-drift through the second pore and cause distinct current blockades, one per residue; only one residue will be resident in the second pore at any time during translocation. Multiple discriminators, including the blockade level and the time of translocation through the second pore, are available in this approach. Software based on hidden Markov models [19] or the Viterbi algorithm [20] can also be used to computationally separate successive residues with identical discriminator values. Incidentally, the work described in [14] also includes a workable solution to another problem in nanopore sequencing, namely the high speed with which a peptide translocates through the pore, which makes it difficult for a detector with insufficient bandwidth to detect changes in the blockade current level.

References

[1] M.A. Baldwin, "Protein identification by mass spectrometry", *Mol. & Cellular Proteomics*, 2004, **3**, 1-9.

[2] W. Timp, A. M. Nice, E. M. Nelson, V. Kurz, K. McKelvey, and G. Timp, "Think small: nanopores for sensing and synthesis', *IEEE Access*, 2014, **2**, 1396-1408.

[3] A. Oukhaled, L. Bacri, M. Pastoriza-Gallego, J.-M., Betton, and J. Pelta. "Sensing proteins through nanopores: fundamental to applications." *ACS Chem. Biol.*, 2012, **7**, 1935-1949.

[4] G. Sampath, "Peptide sequence partitions and protein identification: a computational analysis", *biorxiv.org*, 10.1101/069526, 15 August 2016.

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

[6] http://www.uniprot.org/.
[7] H. Steen and M. Mann. "The ABC'S (and XYZ's) of peptide sequencing." Nature Reviews, 2004, **5**, 699-711.
[8] http://www.srmatlas.org/human/.
[9] J. Swaminathan, A. A. Boulgakov, and E. M. Marcotte, "A theoretical justification for single molecule peptide sequencing", *PLoS Comput. Biol*., 2015, **11**, e1004080.
[10] Y. Yao, M. Docter, J. van Ginkel, D de Ridder, and C. Joo, "Single-molecule protein sequencing through fingerprinting: computational assessment", *Phys. Biol*., 2015, **12**, 055003.
[11] 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.
[12] M. Kolmogorov, E. Kennedy, Z. Dong, G. Timp, and P. Pevzner. "Single-molecule protein identification by sub-nanopore sensors", *arXiv*, 1604.02270v1 [q-bio.QM], 8 April 2016.
[13] M. Drndic, "Sequencing with graphene pores", *Nature Nanotech*., 2014, **9**, 743.
[14] J. Feng, K. Liu, R. D. Bulushev, S. Khlybov, D. Dumcenco, A. Kis, and A. Radenovic. "Identification of single nucleotides in MoS_2 nanopores", *Nature Nanotech.*, 21 September 2015, doi: 10.1038/nnano.2015.219.
[15] 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.
[16] 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.
[17] 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.
[18] G. Sampath, "Amino acid discrimination in a nanopore and the feasibility of sequencing peptides with a tandem cell and exopeptidase", *RSC Adv.*, 2015, **5**, 30694-30700.
[19] J. Schreiber and K. Karplus, "Analysis of nanopore data using hidden Markov models", *Bioinformatics*, 2015, **31**, 1897–1903.
[20] W. Timp, J. Comer, and A. Aksimentiev, "DNA base-calling from a nanopore using a Viterbi algorithm", *Biophys. J.*, 2012, **102**, L37-39.