bioRxiv preprint doi: https://doi.org/10.1101/119313; this version posted March 22, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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 fingerprinting with a binary alphabet

G. Sampath

Abstract. Proteins can be identified by partitioning a proteome into eight mutually exclusive sets of peptides, recoding them with a binary alphabet obtained by dividing the 20 amino acids into two ordered sets based on amino acid volume, and searching for the recoded peptides in a protein sequence database. With this approach over 89.7% of all protein sequences in the human proteome (http://www.uniprot.org; database id UP000005640, 20207 curated sequences) can be uniquely identified. Implementation issues are briefly discussed. In particular, nanopore-based sequencing of partitioned peptides becomes less difficult as the signal processing involved is largely a matter of thresholding the current blockade signal due to a translocating peptide and generating a binary sequence from it.

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 can be identified from peptide subsequences coded with a binary protein alphabet. The proposed procedure can be implemented in practice with available chemical procedures, this is discussed toward the end of this report.

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^*}.

3. An amino acid partition based on volume

Table 1 shows the 20 standard amino acids ordered on volume. The amino acids can be grouped into two subsets by dividing the table at a suitable point. The dividing line was chosen between P and V so that the two subsets are of nearly the same size and the difference between the volumes of P and V is substantial. (There are much higher volume differences, such as between Y and W or between G and A, but the two subset sizes are lopsided.) The amino acids can now be coded with a binary code: {G,A,S,C,D,T,N,P: volume $\leq$ 123.3} $\rightarrow$ 1, {V,Q,E,H,I,L,M,K,R,F,Y,W: volume $\geq$ 138.8} $\rightarrow$ 2.

Table 1. Table of amino acids ranked by volume
AA = Amino acid; Volumes in nm^3. Background shading shows grouping of amino acids into 2 subsets. Data based on [2].

AA	G	A	S	C	D	T	N	P	V	E
Mean volume	59.9	87.8	91.7	105.4	115.4	118.3	120.1	123.3	138.8	140.9
Standard deviation	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	145.1	156.3	165.2	166.1	168	172.7	188.2	189.7	191.2	227.9
Standard deviation	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 partition for it 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 calculated using a straightforward algorithm 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 Columns 3 and 4, Row 10. 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: volume $\leq$ 91.7} $\rightarrow$ 1, {C,D,T,N,P: 105.4 $\leq$ volume $\leq$ 123.3} $\rightarrow$ 2, {V,Q,E,H,I,L,M,K,R,F,Y,W: volume $\geq$ 138.8} $\rightarrow$ 3. The 4-character alphabet corresponds to the one used in a

bioRxiv preprint doi: https://doi.org/10.1101/119313; this version posted March 22, 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.

recent report on nanopore-based sequencing of proteins [3], see the last paragraph in Section 5 below for more information on the work reported therein.)

5. Implementation notes and discussion

The most commonly used fingerprinting method is based on mass spectrometry [4]. More recently theoretical methods have been proposed based on optical or other labeling of selected residues. In these methods a protein is proteolytically cleaved into peptide fragments and selectively labeled with as many different dyes as the size of the alphabet. The labeled fragments are sequenced in one of two ways: 1) by 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 [5]; 2) using a protein translocase to pass the fragments through a nanochannel followed by Förster resonance energy transfer (FRET) to detect a labeled residue as it moves past the enzyme [6]. As few as 2 tagged residue types among the possible 20 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.

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

Subset of partition	Total no. of peptides in subset	Binary alphabet		Full alphabet		Alphabet size 3		Alphabet size 4	
		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%	14247	70.51%	8265	40.83%	10060	49.78%
KZ*D	125351	4602	22.73%	12983	64.25%	7378	36.45%	9847	44.63%
KZ*E	194024	5171	25.54%	14304	70.79%	8400	41.50%	10185	50.40%
KZ*	189713	5143	25.41%	13736	67.98%	8128	40.15%	9832	48.65%
Z*R	499784	10356	51.16%	18305	90.59%	14183	70.06%	15770	78.04%
Z*D	411189	9117	45.04%	17691	87.55%	13031	64.37%	14700	72.74%
Z*E	609872	9932	49.06%	18254	90.34%	14078	69.55%	15690	77.64%
Z*	345450	9411	46.49%	18467	91.39%	13644	67.40%	15370	76.06%
P	2514806	18133	**89.74%**	19885	**98.4%**	19302	**95.35%**	19581	**96.90%**

(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

In contrast the method proposed in the present work is designed to identify proteins without the use of analyte immobilization or labeling of any kind. Instead an electrical method based on the passage of a single protein or peptide molecule through a nanopore [7] is to be employed. As in [5] and [6] partitioned peptides are partially sequenced, but sequencing consists merely of determining which of the two ordered groups of residue types defined in Section 3 a residue in the peptide belongs to. This is done by translocating the peptide through a nanopore in a membrane and assigning one of two values to the resulting blockade current. This turns sequencing into a binary electrical current detection process and makes the procedure considerably simpler. Additionally the sequencing device is significantly smaller than one based on fluorescent methods. A downside to this approach (as in most other nanopore-based sequencing approaches) is the homopolymer problem, which arises from successive residues belonging to the same subset generating the same binary output value. With a thick (8-10 nm) pore, biological or synthetic, 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 is usuallydifficult to identify. The severity of the problem can be reduced by using a single layer of graphene [8] or molybdenum sulphide (MoS_2) [9] for the membrane. In this case only one residue will be resident in the pore at any time during translocation. Alternatively an electrolytic cell with a tandem pair of nanopores and an exopeptidase may be used [10]. Here the enzyme, which is situated below the first pore, cleaves successive residues from a peptide emerging from the first pore; these 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. Software based on hidden Markov models can also be used to computationally separate successive

bioRxiv preprint doi: https://doi.org/10.1101/119313; this version posted March 22, 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.

residues with identical blockade values [11]. (The work described in [9] also reports on 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.)

Additional information is available in [12] on potential implementation procedures, including peptide partitioning of proteins based on sequential proteolysis, the use of isoelectric focusing (IEF) to separate partition subsets, and the use of single molecule methods (including nanopores and flurorescent tagging) for sequencing of partitioned subsequences. Also see [3] for recent work on nanopore-based sequencing of proteins in which the 20 amino acids are divided into four subsets labeled *Minuscule*, *Small*, *Intermediate*, and *Large*; 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 these four subsets and the resulting peptide encoded with the 4-character alphabet {M, S, I, L}.

References

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

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

[3] 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.

[4] E. M. Marcotte, "How do shotgun proteomics algorithms identify proteins?", *Nature Biotech*., **25**, 755-757, 2007.

[5] J. Swaminathan, A. A. Boulgakov, E. M. Marcotte, "A theoretical justification for single molecule peptide sequencing", *PLoS Comput. Biol*., **11**, e1004080, 2015.

[6] Y. Yao, M. Docter, J. van Ginkel, D de Ridder, and C. Joo, "Single-molecule protein sequencing through fingerprinting: computational assessment", *Phys. Biol*., **12**, 055003, 2015.

[7] D. Wu, S. Bi, L. Zhang, and J. Yang. "Single-molecule study of proteins by biological nanopore sensors", *Sensors* **14**, 18211-18222, 2014.

[8] M. Drndic, "Sequencing with graphene pores", *Nature Nanotech*., **9**, 743, 2014.

[9] 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.

[10] 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.

[11] J. Schreiber and K. Karplus, "Analysis of nanopore data using hidden Markov models", *Bioinformatics* **31**, 1897–1903, 2015.

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