Volumetric analysis of unmodified and modified amino acids and its application to the detection of post-translational modifications (PTMs) with a nanopore

G. Sampath
(Unaffiliated)
Email: sampath_2068@yahoo.com

Abstract. Recently a method was proposed for unambiguous identification of an amino acid (AA) in the bulk or at the single molecule (SM) level with a single binary measurement based on the superspecificity property of transfer RNAs (tRNAs) (doi: 10.36227/techrxiv.19318145.v3). The present report looks at detection of post-translational modifications (PTMs) in single AAs after their identification through their cognate tRNAs. The volumes of AAs and their PTMs are computed from crystallographic data for all 20 proteinogenic AAs and three types of PTMs (methylation, acetylation, and phosphorylation). The ratio of the spatial volume of a modified AA to that of an unmodified one is then used to assign a PTM to the former. An experimental procedure to obtain these ratios for different PTMs with a nanopore of matching diameter is proposed. AA is freed following its identification via its cognate tRNA as in the procedure mentioned above, separated by a nanofilter, and translocated through a nanopore under electrophoresis. The ratio of the blockade level due to a modified AA to that due to the unmodified AA is compared with the theoretical ratio and a PTM assigned. Because the AA has been unambiguously determined by its cognate tRNA PTM assignment is horizontal across the set of PTMs, thus the other 19 AA types and their PTMs are not involved in PTM assignment. Computational results are presented for 49 PTMs covering all 20 AAs and the three types mentioned above. Unlike most methods based on mass spectrometry this is a *de novo* method that does not require prior knowledge about the parent protein, nor does it use any sequence information from a proteome database.

1. Introduction

An understanding of protein structure, function, and interaction is central to biological research, medical diagnostics, and drug design. Biological processes at the cellular level are to a large extent determined by the 3-dimensional structure of a protein, which in turn is determined by its amino acid sequence. Protein sequencing methods that can reliably and efficiently obtain accurate sequences of known and unknown proteins have for a long time been pursued by analytical scientists [1]. These methods are based on a wide range of techniques from across analytical chemistry, physics, and, more recently, nanotechnology. The field is currently led by mass spectrometry (MS) [2], with gel/capillary electrophoresis [3] and Edman degradation followed by spectroscopy [4] playing a secondary role. MS is by and large a bulk method based on samples in the femto- to atto-liter range, which is a limitation if analyte availability is limited, as would be the case in single cell studies, where only a small number of molecules of some cellular proteins may be available. This has led to a search in recent years for methods that can sequence proteins at the single molecule (SM) level [5]. In this search, nanopores have been emerging as a potentially viable alternative, with much effort going into developing methods for the study of proteins in unfolded and folded form down to the level of a single residue [6].

While the protein's primary sequence plays a major role in the study of protein structure and the ability to artificially modify it in beneficial ways, such as in drug design, modifications of proteins that occur naturally in the cell before, during, and after translation from RNA to protein have become a major focus in recent years. These modifications occur at the residue level and are collectively referred to as post-translational modifications (PTMs) [7]. Most of PTM detection is currently based on MS, which can accurately separate ions produced from a protein or peptide and compute mass differences down to a single atom. In this it is aided considerably by database-based methods that allow assignment of PTMs to residues based on their recorded presence in peptide sequences stored in the database.

In the next section PTM detection with MS and other methods is summarized. Following this a volumetric analysis of AAs and their PTMs is introduced and is used as the basis for a proposed procedure for PTM assignment that uses the blockade current caused by AAs and their PTMs as they translocate through a nanopore. This is a SM approach that has the potential ability to assign PTMs to modified AAs in very small samples containing only a few molecules, a capability that appears to be outside the scope of MS at present. The approach to PTM detection and identification given here uses as its starting point a recently proposed method of AA identification [8] based on the 'superspecificity' [9] property of transfer RNAs (tRNAs), whereby a tRNA can only be charged with its cognate AA and not with any other. PTM identification follows AA identification, so that it is only necessary to distinguish among PTMs of the already identified AA, the other 19 AA types are not involved.

2. PTMs of amino acids and their detection and identification

PTMs can occur prior to translation of mRNA to protein, during translation, or after the protein has been synthesized in the cell. These modifications are referred to as pre-, co-, and post-translational, however it is common practice to refer to all of them as PTMs. PTMs are by and large rare occurrences in that only a few copies of a protein may undergo modifications. This requires methods of purification and enrichment to increase the sample size for analysis. When a protein itself occurs infrequently in the cell's proteome, this rarity is further compounded. This means that separation has to occur at the single cell level and hence detection and identification techniques have to be at the SM level.

A wide range of PTMs has been observed in the proteome of an organism [10]; the Uniprot knowledge base [11] contains an annotated list. The most common ones are methylation, acetylation, phosphorylation, and glycosylation. Table 1 shows a selected set of PTMs of these four types. A variety of methods have been developed for their detection. At the bulk level PTM-specific antibodies are available. By far the most widely used method is MS, which can be used for detection of almost all known PTMs and is also capable of detecting unknown ones. This versatility and power of MS derives from its ability to distinguish among molecules differing in mass by small amounts, down to the level of single atoms. However, MS is largely a bulk method and requires sample sizes with 1000s to tens of 1000s of a protein molecule to work with.

Table 1. A selection of PTMs of four types: phosphorylation, methylation, acetylation, and glycosylation.

AA	PTM type			
	Phosphorylation	**Methylation**	**Acetylation**	**Glycosylation**
G		N,N,N-trimethylglycine N,N-dimethylglycine N-methylglycine	N-acetylglycine	N-D-glucuronoyl glycine
A		N,N,N-trimethylalanine N,N-dimethylalanine N-methylalanine	N-acetylalanine	
S	Phosphoserine	N,N,N-trimethylserine N,N-dimethylserine N-methylserine	N-acetylserine O-acetylserine	O-(2-cholinephosphoryl)serine O-linked: (DADDGlc) serine O-linked (GATDGlc) serine O-linked (Fuc) serine O-linked (FucNAc) serine
C	Phosphocysteine		N-acetylcysteine	
D			N-acetylaspartate	
T	Phosphothreonine	O-methylthreonine	N-acetylthreonine O-acetylthreonine	O-linked (Fuc) threonine
N		N4,N4-dimethylasparagine N4-methylasparagine		N-linked (GalNAc) asparagine N-linked (Hex) asparagine N-linked (HexNAc) asparagine
P		N,N-dimethylproline N-methylproline	N-acetylproline	
V			N-acetylvaline	N-linked (Glc) (glycation) valine
E			N-acetylglutamate	
Q		N5-methylglutamine		
H	Phosphohistidine			
M		N,N,N-trimethylmethionine N-methylmethionine	N-acetylmethionine	
I		N-methylisoleucine	N-acetylisoleucine	
L		N,N-dimethylleucine N-methylleucine		
K		N6-methylated lysine N6,N6-dimethyllysine	N6-acetyllysine	N-linked (Glc) (glycation) lysine N-linked (Lac) (glycation) lysine
R	Phosphoarginine	N2,N2-dimethylarginine N5-methylarginine	N2-acetylarginine	N-linked (Glc) arginine N-linked (GlcNAc) arginine N-linked (Hex) arginine N-linked (HexNAc) arginine
F		N-methylphenylalanine		
Y	Phosphotyrosine	N-methyltyrosine	N-acetyltyrosine	
W				N-linked (Man) tryptophan N-linked (Hex) tryptophan

For a quick introduction to PTMs see [12]. For a discussion of PTMs in a clinical setting see [13]. The use of tandem MS in PTM detection has been widely reviewed; see, for example, [14,15]; purification and enrichment

strategies are reviewed in [16]; database methods are covered in [17]; a single-cell method is introduced in [18]; the connection between PTMs and diseases is discussed in [19], a PTM-disease model is presented with a database containing 749 proteins, 23 PTM types, and 275 types of diseases; in [20] over 150 PTMs are obtained across the proteome of the human heart by using MS measurements and PTM identification algorithms.
In studying PTMs it is useful to know their frequencies, as this can be used to determine the level of purification and/or enrichment required. The most commonly occurring PTM is phosphorylation, with close to 60% of PTMs being phosphorylated. Other PTMs with a significant presence include acetylation, glycosylation, and methylation. The present study is limited to phosphorylation, acetylation, and methylation.
As noted earlier, PTM detection with MS is mostly based on using ionic mass as the discriminant. In the next section the possibility of using volume as the discriminant is considered.

3. Volumetric analysis of free AAs and AAs with PTMs: a computational study
Biomolecules are often modeled as a collection of spheres, one per atom, with radii equal to their van der Waals radii. The interatomic bonds between two neighboring atoms in these molecules are usually shorter than the van der Waals radii of the two atoms. As a result geometric modeling of biomolecules involves intersecting spheres. If biomolecules in a solution are considered the radius of a sphere is extended by the radius of the solvent molecule (usually water). The volume of a collection of intersecting spheres has been studied widely with methods from computational geometry [21]. Mathematical analysis of intersecting objects is often based on Voronoi diagrams and/ or Delaunay triangulation methods. For example, in [22] an exact $O(n^2)$ algorithm is given for the calculation of the union of intersecting spheres.
A quick and easy alternative approach to computational geometry is Monte Carlo simulation. This is used in the present study to compute the volumes of AAs and PTMs.

3.1 Volume determination with Monte Carlo simulation
An irregular object in 3-dimensional space like an AA, peptide, or protein can be viewed as a set of points (corresponding to the atoms in the protein) in a 3-dimensional rectangular enclosing box whose extent is given by the maximum span of the set of points. Consider an AA given by a set of K triples of the form $\mathbf{x} = (x, y, z)$:

$$P_{AA} = \{ \mathbf{x}_1, ..., \mathbf{x}_K \} \quad (1)$$

The extent of the box can be found from the spans in the x, y, and z directions as given by $[x_{min}, x_{max}]$, $[y_{min}, y_{max}]$, and $[z_{min}, z_{max}]$:

$$x_{min} = \min (x_1, ..., x_K,), y_{min} = \min (y_1, ..., y_K,), z_{min} = \min (z_1, ..., z_K) \quad (2a)$$
$$x_{max} = \max (x_1, ..., x_K), y_{max} = \max (y_1, ..., y_K,), z_{max} = \max (z_1, ..., z_K) \quad (2b)$$

The spans of the box are:

$$x_{span} = x_{max} - x_{min}, y_{span} = y_{max} - y_{min}, z_{span} = z_{max} - z_{min}, \quad (3)$$

The volume of the box is

$$V_{box} = x_{span} * y_{span} * z_{span} \quad (4)$$

To compute the volume of the object, a set of M random triples inside the box is generated and the number of points L that fall inside the object is determined using an inside test, given below. Random number generators usually generate an integer *rand* in the range [0,MAXINT]. A random real *rnd* in the range [0,1.0] can be obtained as

$$rnd = rand / MAXINT \quad (5)$$

A random point $\mathbf{x}_{rnd} = (x_{rnd}, y_{rnd}, z_{rnd})$ can be then be generated as

$$x_{rnd} = x_{min} + rnd * x_{span}, y_{rnd} = y_{min} + rnd * y_{span}, z_{rnd} = z_{min} + rnd * z_{span} \quad (6)$$

A random point generated is inside AA if it is within the van der Waals radius of at least one atom in the set of atoms in P_{AA}. The following test is used:

$$| \mathbf{x}_{rnd} - \mathbf{x}_i | \leq vdw_i \text{ for at least one i in } \{1, ..., K\} \quad (7)$$

where vdw_i is the van der Waals radius of atom i in P.
Let L of the M random points generated satisfy Equation 7. The volume of P is then given by

$$V_P = (L/M)\, V_{box} \quad (8)$$

If solvent molecules (water) are considered, the approximate van der Waals radius for a water molecule, 1.4 Å, is added to the van der Waals radius of an atom in the test in Equation 7 to get:

$$| \mathbf{x}_{rnd} - \mathbf{x}_i | \leq vdw_i + 1.4 \text{ for at least one i in } \{1, ..., K\} \quad (9)$$

For a list of van der Waals radii of atoms that appear in biomolecules, see [23].

3.2 Results

The coordinates of the centers of the atoms of all 20 unmodified proteinogenic AAs were downloaded from the PubChem database and the volumes computed as in Section 3.1. For each AA, 2000000 random triples were generated. The results are shown in Table 2.The PubChem id for an AA is given in columns 3 and 8.

Table 2. Volumes of unmodified amino acids from Monte Carlo simulation

AA	PubChem id	No. of atoms	Volume (Å^3)	Volume (Å^3) with solvent molecule	AA	PubChem id	No. of atoms	Volume (Å^3)	Volume (Å^3) with solvent molecule
ALA	5950	13	45.69	50.73	MET	6137	20	86.21	104.74
CYS	5862	14	49.18	54.07	ASN	6267	17	68.15	79.62
ASP	5960	16	66	77.1	PRO	145742	17	64.23	72.26
GLU	33032	19	72.44	87.77	GLN	5961	20	82.89	105.89
PHE	6140	23	96.14	110.55	ARG	6322	26	111.67	135.16
GLY	750	10	20.66	21.11	SER	5951	14	45.21	49.91
HIS	6274	20	94.33	118.85	THR	6288	17	64.78	75.32
ILE	6306	22	69.78	78.35	VAL	6287	19	66.05	75.94
LYS	5962	24	92.76	109.4	TRP	6305	27	142.47	185.31
LEU	6106	22	77.82	87.79	TYR	6057	24	101.8	118.48

The simulation was repeated with 49 PTMs with three types of modification: 17 methylated, 20 acetylated, 12 phoshporylated. Table 3 shows their volumes with and without the solvent molecule taken into account. PubChem ids are in columns 2 and 7.

Table 3. Volumes of PTMs from Monte Carlo simulation

PTM name	PubChem id	No. of atoms	Volume (Å^3)	Volume (Å^3) with solvent molecule	PTM name	PubChem id	No. of atoms	Volume (Å^3)	Volume (Å^3) with solvent molecule
N-Methyl-L-alanine	5288725	16	52.14	59.14	N-Acetyl-L-methionine	448580	25	114.92	143.6
N-acetyl-L-alanine	88064	18	69.7	83.88	N-alpha-acetyl-L-asparagine	99715	22	88.28	108.72
S-Methyl-L-cysteine	24417	17	64.9	77.19	Phospho-asparagine	54033079	22	102.12	123.74
N-Acetyl-L-cysteine	12035	19	75.79	90.58	N-acetyl-L-proline	66141	22	94.48	113.73
S-Phosphocysteine	3082729	19	99.48	118.27	Phospho-proline	91440465	22	92.62	109.56
N-methyl-D-aspartic acid	22880	19	80.4	96.68	N-acetyl-L-glutamine	182230	25	115.82	148.51
N-acetyl-L-aspartic acid	65065	21	86.16	104.08	N-phosphono-L-glutamine	126961722	25	103.93	125.42
N-Methyl-L-glutamic acid	439377	22	97.24	120.73	Methylarginine	4366	29	119.37	152.4
N-Acetyl-L-glutamic acid	70914	24	111.68	144.13	N-Acetyl-L-arginine	67427	31	142.52	192.25

N-Methyl-L-phenylalanine	6951135	26	117.79	138.86	Phospho-L-arginine	92150	31	146.76	187.89
N-Acetyl-L-phenylalanine	74839	28	141.33	180.96	phosphoserine	106	19	88.01	105.07
N-Phosphono-L-phenylalanine	54382600	28	144.55	182.14	2-methyl-L-serine	7000050	17	65.53	79.06
N-methylglycine	1088	13	28.35	28.85	O-acetyl-L-serine	99478	19	77.44	88.72
N-acetylglycine	10972	15	54.79	62.07	O-methyl-L-threonine	2724875	20	75.92	91.7
N-phosphono-Glycine	11321074	15	50.05	52.76	O-Acetyl-L-Threonine	16718064	22	95.74	122.86
1-Methyl-L-histidine	92105	23	103.93	123.34	O-phospho-L-threonine	3246323	22	103.7	125.26
Acetyl histidine	75619	25	115.35	140.29	N-Methyl-L-valine	444080	22	87.04	106.93
phosphohistidine	123911	25	123.74	151.94	N-Acetyl-L-valine	66789	24	96.71	120.5
N-Methyl-L-isoleucine	5288628	25	109.02	140.2	N-Acetyl-L-tryptophan	700653	32	161.25	209.48
N-Acetyl-L-isoleucine	7036275	27	111.94	140.29	1-Methyl-L-tryptophan	676159	30	155.73	203.99
N-methyllysine	541646	27	105.46	135.02	phosphotryptophan	11472143	32	182.61	235.95
acetyllysine	6991978	29	119.89	149.6	O-Phospho-L-tyrosine	30819	29	148.57	184.24
N-Methylleucine	6951123	25	97.31	114.99	3-methyl-l-tyrosine	159657	27	118.84	140.1
N-Acetyl-L-leucine	70912	27	120.88	153.74	N-Acetyl-L-tyrosine	68310	29	146.67	187.68
N-Methyl-L-methionine	6451891	23	100.46	119.7					

The distribution by AA is shown in Table 4.

Table 4. The 49 PTMs in Table 3 distributed by AA and PTM type

	methylation		acetylation		phosphorylation	
ALA	5288725	N-Methyl-L-alanine	88064	N-acetyl-L-alanine		
CYS	24417	S-Methyl-L-cysteine	12035	N-Acetyl-L-cysteine	3082729	S-Phosphocysteine
ASP	22880	N-methyl-D-aspartic acid	65065	N-acetyl-L-aspartic acid		
GLU	439377	N-Methyl-L-glutamic acid	70914	N-Acetyl-L-glutamic acid		
PHE	6951135	N-Methyl-L-phenylalanine	74839	N-Acetyl-L-phenylalanine	54382600	N-Phosphono-L-phenylalanine
GLY	1088	N-methylglycine	10972	N-acetylglycine	11321074	N-phosphono-Glycine
HIS	92105	1-Methyl-L-histidine	75619	Acetyl histidine	123911	phosphohistidine
ILE	5288628	N-Methyl-L-isoleucine	7036275	N-Acetyl-L-isoleucine		
LYS	541646	N-methyllysine	6991978	acetyllysine		
LEU	6951123	N-Methylleucine	70912	N-Acetyl-L-leucine		
MET	6451891	N-Methyl-L-methionine	448580	N-Acetyl-L-methionine		
ASN			99715	N-alpha-acetyl-L-asparagine	54033079	Phospho-asparagine
PRO			66141	N-acetyl-l-proline	91440465	Phospho-proline
GLN			182230	N-acetyl-L-glutamine	126961722	N-phosphono-L-glutamine
ARG	4366	Methylarginine	67427	N-Acetyl-L-arginine	92150	Phospho-L-arginine
SER	7000050	2-methyl-L-serine	99478	O-acetyl-L-serine	106	phosphoserine
THR	2724875	O-methyl-L-threonine	16718064	O-Acetyl-L-Threonine	3246323	O-phospho-L-threonine
VAL	444080	N-Methyl-L-valine	66789	N-Acetyl-L-valine		
TRP	676159	1-Methyl-L-tryptophan	700653	N-Acetyl-L-tryptophan	11472143	phosphotryptophan
TYR	159657	3-methyl-l-tyrosine	68310	N-Acetyl-L-tyrosine	30819	O-Phospho-L-tyrosine

Figure 1 shows volume distribution by AA and PTM type.

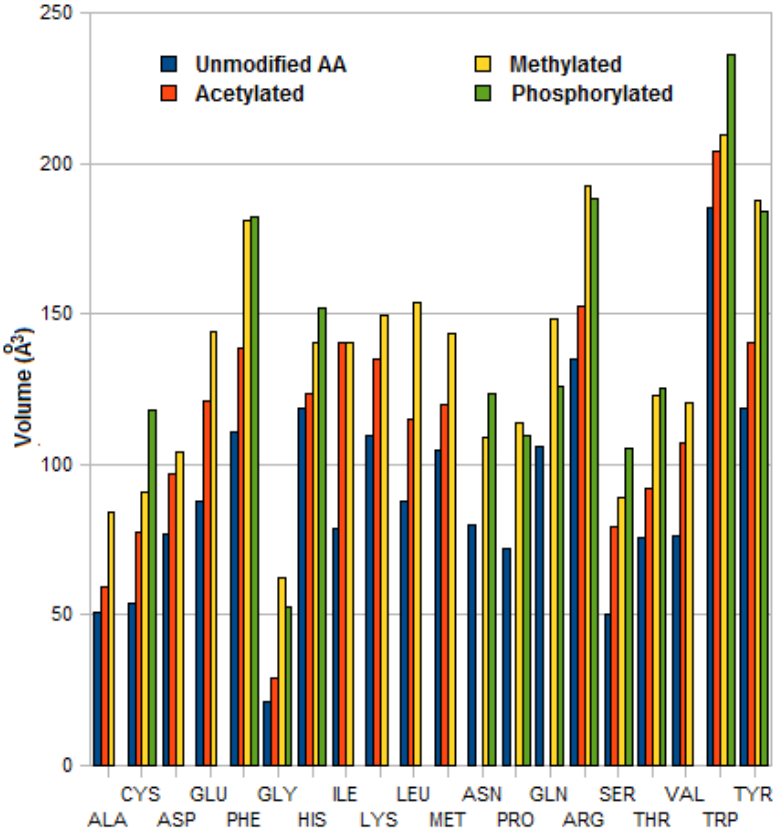

Figure 1. Distribution of volumes of unmodified AAs and their PTMs.

Figure 1 shows that there is sufficient variability in the volumes of many of the AAs and their PTMs to allow the use of analyte volume as a discriminator in the detection of PTMs. In the next section a method for PTM detection following AA identification based on superspecificity is described. This reduces PTM assignment to selecting among the PTMs of a given AA. The other 19 AA types and their PTMs are not involved, thus reducing the effort required in PTM assignment.

4. Amino acid identification with tRNA and PTM assignment with a nanopore

As noted earlier, protein/peptide sequencing is commonly based on one of the following three methods: 1) Edman degradation (ED); 2) capillary or gel electrophoresis; and 3) mass spectrometry (MS). Most protein sequencing is currently based on MS.

AA identification with MS, whether of the unmodified or of the modified form, is based on breaking a peptide obtained by proteolysis (bottom-up MS) or a whole intact protein (top-down MS) into ion fragments. The extreme precision that comes with MS is possible because all (or most) possible fragments are obtained from the ionization step, which gives the ability to calculate minute differences between the masses of fragments down to the level of a single atom. However this precision comes at a price, namely the required solution sample size, which is in the femto- to atto-liter range and corresponds to thousands to millions of copies. The advantage of MS more or less disappears if only a few molecules are available, which is often the case with single cells. In studying the contributions of different proteins in a cell to and modeling of the cell's processes there is wide variation in the number of molecules that may be available to the analyst. Often only a few copies of some of the cell's proteins may be available. MS is not equipped to handle the full dynamic range of variation in an organism's proteome, although efforts are ongoing to change this situation.

Sparsely occurring proteins need to be handled by procedures that can examine them at the single molecule (SM)

level. Several new approaches have been developed; see [1,5].
One emerging SM approach is based on nanopores [6]. An electrolytic cell (e-cell) has a nanopore in a membrane that divides two chambers filled with an electrolyte and known as *cis* and *trans*. With an electrical potential applied across the membrane the electrolyte is ionized, leading to an ionic current which can be measured with a detector. When an analyte is inserted into *cis* it translocates into *trans* by a combination of electrophoresis and diffusion, sometimes aided by electro-osmotic forces due to electrical charges on the wall of the pore. The intrusion of the analyte results in a decrease or blockade in the normal ionic current. The size of this blockade is roughly proportional to the volume of the analyte. Often it is possible to identify the analyte from the level of the blockade. If the analyte is restricted to one of a small set of molecules it would be possible to distinguish among the members of the set based on the blockade level. For example, if the set consists of the 20 proteinogenic AAs, it may be possible to identify an AA from the measured blockade level [24], and this may be used to sequence a peptide [25].
The field of nanopore-based protein sequencing is reviewed in [26]. Biological pores are reviewed in [27], solid-state in [28]. Simultaneously nanopore-based identification of PTMs has been studied. In [29] an AA is modified artificially with a label to identify its presence in a peptide. Labeled and unlabeled AAs are distinguished based on mass, geometry, charge, and hydrophobicity. In [30] the site of a phosphorylated residue in a protein is detected with a nanopore. In [31] differences in blockade level and dwell time are used to identify a PTM in a protein without using labels. In [32] molecular dynamics simulations are used to show that acetylated and methylated residues in an intact peptide from histone protein H4 can be detected based on differences in the electrical resistance measured during passage of the peptide through the pore.

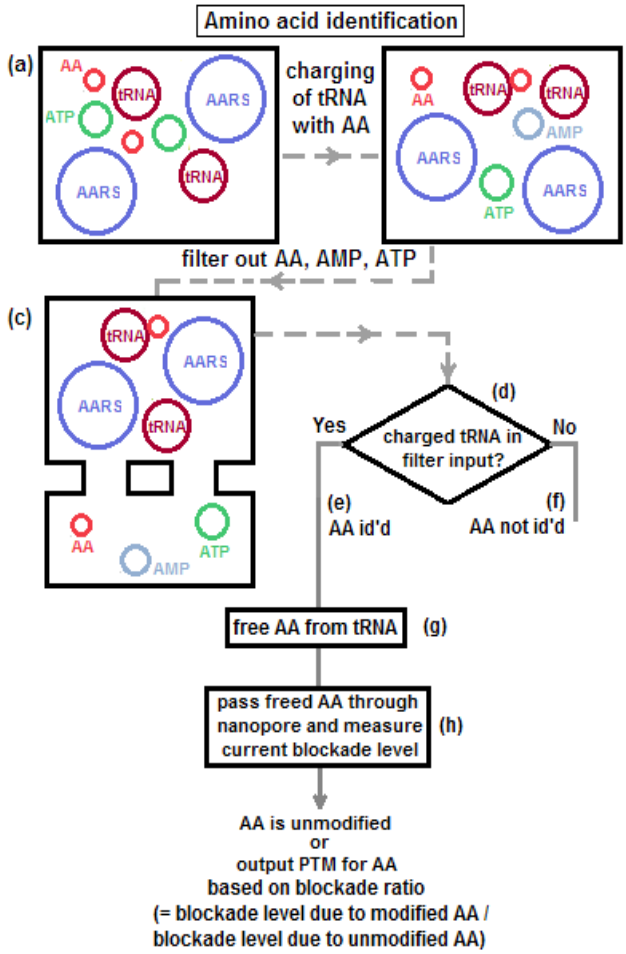

Figure 2. Schematic showing AA identification with tRNA followed by PTM identification with nanopore

4.1 AA identification with a tRNA
The present work takes a somewhat different approach, it proposes using a nanopore to identify PTMs of AAs that

have already been identified. In [8] a procedure is described for the unambiguous identification of free AAs cleaved from a peptide. It is based on the superspecificity property of tRNAs: a tRNA can get charged only with its cognate AA and not with any other AA. By confining each of 20 copies of an AA with a different tRNA, only one of them gets bound with its cognate tRNA. the charging event can be detected optically with TIRF (total internal fluorescence spectroscopy) if the AA is tagged with a fluorescent dye, or by translocating it through a nanopore after the bound AA is freed from the cognate tRNA. The detection is of a binary nature, so precise measurements are not needed. The upper part of Figure 1 shows how AA is identified from its cognate tRNA, the lower part shows the steps in PTM identification that follow AA identification. Figure 2 shows the detection of AA bound to its cognate tRNA. There are two ways to detect the AA. Figure 2a shows detection of a fluorescent tag attached to the AA, Figure 2b shows detection with a nanopore after the bound AA has been deacylated and is free to translocate through a nanopore and cause the current blockade that is used to detect it. Notice that detection of the presence of AA is sufficient, identification has already implicitly occurred when tRNA gets charged with AA. (This occurs only in the identification unit with the cognate copy of AA. In the other 19 units there is no output.) For the details see [8], where a series version that can work with a single copy of AA is also outlined.

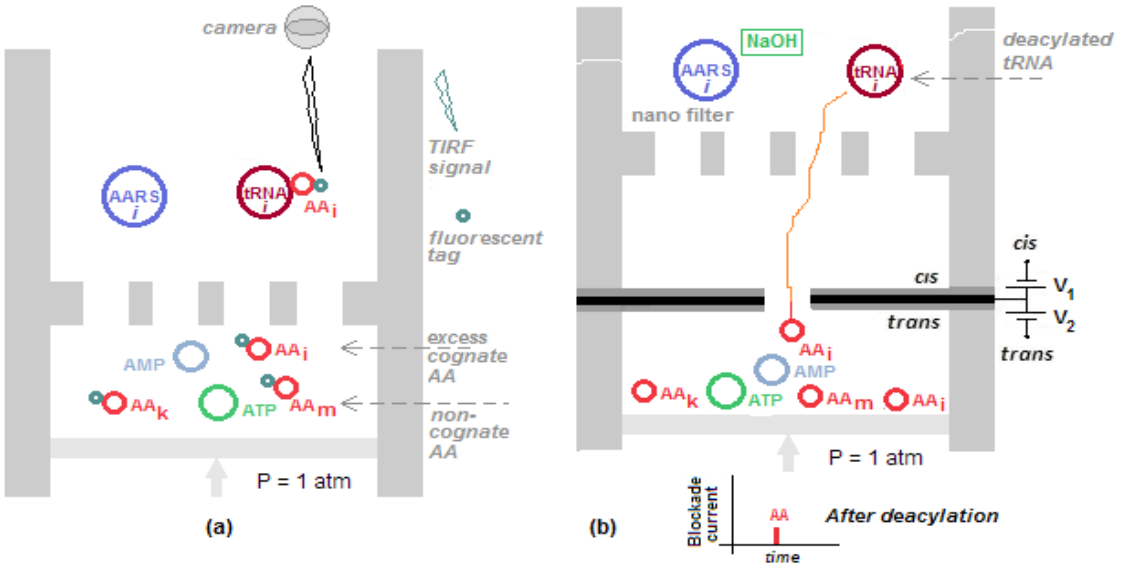

Figure 3. Detection of AA bound to cognate tRNA. (a) Optical detection of tag attached to AA with TIRF; (b) Electrical detection of AA after AA is freed from tRNA and passed through nanopore where occurrence of current blockade signals its presence.

4.2 PTM identification with a nanopore

Following identification the released AA is translocated through a nanopore in an e-cell and the blockade level measured. PTM assignment can be done horizontally across the set of PTMs for an AA without involving the other AAs. The blockade level due to a PTM is compared with the reference blockade level for the corresponding unmodified AA. The ratio of the two levels is used to assign a PTM to the AA if it has been modified. Table 5 shows the volume ratios used for such assignment.

Table 5. Volume ratios of PTMs with respect to volume of unmodified AA

AA	Unmodified AA volume ratio	Volume ratio for PTM type		
		Methylated	Acetylated	Phosphorylated
ALA	1	1.17	1.65	
CYS	1	1.43	1.68	2.18
ASP	1	1.25	1.35	
GLU	1	1.38	1.64	
PHE	1	1.26	1.64	1.65
GLY	1	1.37	2.94	2.5
HIS	1	1.04	1.18	1.28
ILE	1	1.79	1.79	
LYS	1	1.23	1.37	
LEU	1	1.31	1.75	

MET	1	1.14	1.37	
ASN	1		1.37	1.55
PRO	1		1.57	1.51
GLN	1		1.4	1.18
ARG	1	1.13	1.42	1.39
SER	1	1.58	1.78	2.1
THR	1	1.22	1.63	1.65
VAL	1	1.41	1.59	
TRP	1	1.1	1.13	1.27
TYR	1	1.18	1.58	1.55

4.3 Nanopore diameters required for detecting PTMs

The minimum nanopore diameter required for an AA or PTM to translocate through can be approximated by the length of the middle of the three axes of the minimum covering ellipsoid for the set of points corresponding to the atom centers for the analyte. This ellipsoid was calculated with the R program function *EllipsoidHull*. The results are shown in Table 6 and Figure 4.

Table 6. Lengths of middle axes of minimum volume covering ellipsoids of AAs and PTMs

AA	Middle axis of covering ellipsoid – unmodified AA (Å)	Middle axis of covering ellipsoid of PTM (Å)		
		Methylated	Acetylated	Phosphorylated
ALA	2.62	5.8	5.04	
CYS	4.36	5.96	5.16	5.88
ASP	5.28	4.48	7.84	
GLU	5.5	5.12	10.74	
PHE	8.12	5.72	9.68	8.7
GLY	1.86	1.74	3.58	2.18
HIS	5.1	4.64	8.82	4.84
ILE	7.24	7.08	7.96	
LYS	4.5	5.72	6.34	
LEU	6.56	6.74	8.88	
MET	5.48	5.98	9.66	
ASN	4.6		10.72	5.78
PRO	3.5		6.86	7.88
GLN	4.98		9.18	5.72
ARG	6.28	6.8	10.54	9.94
SER	4.56	3.3	4.56	5.12
THR	5.56	6.84	6.62	6.56
VAL	5.4	6.64	8.24	
TRP	9.42	10.92	11.66	9.14
TYR	7.56	7.96	9.68	7.58

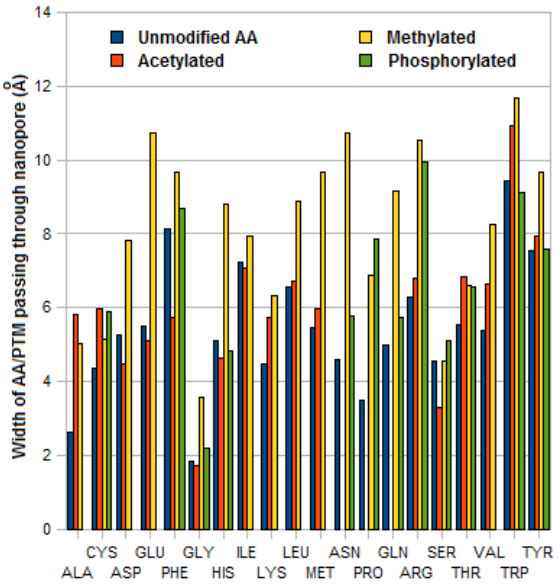

Figure 4. Distribution of AA/PTM widths

In principle the diameters in Table 6 can be used to separate PTMs and detect them with a series of nanopores in the pore with diameter closest to the analyte diameter. This provides an additional discriminator. Currently the smallest diameter solid-state pore is about 0.5 to 1 nm [25,33]. Considering the diameters in Table 6, at present such a separation technique can be used with a few of the PTMs in the table. Other possible discriminants that could be used include the dwell tine of an analyte inside the pore, approximately equal to the blockade pulse width. This aspect is being studied, the results of simulation will be made available later.

6. Discussion

The work reported here is an attempt to reduce the complexity of PTM detection by separating the process of identifying an amino acid in a peptide/protein from that of detecting a PTM associated with the amino acid and assigning the PTM type to it. This is possible because of the silo-like nature of the AA identification step (see discussion in [8]). Essentially it allows optimization of the characteristics of any device that comes after identification of AA has occurred through charging of a cognate tRNA. Thus if AA identification is one dimension and PTM identification another, then the latter is a horizontal process because only those PTMs that are associated with the identified AA are involved, the other 19 AA types and their PTMs do not play a significant role if any. If PTM identification is enhanced by database search methods then search procedures would be significantly less complex with the PTM detection/identification approach presented here.

Supplementary Information. Data files for 20 AAs and 49 PTMs in SDF format, slightly modified with added annotation on the first line.

References

[1] 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.

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

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

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

[5] J. A. Alfaro, P. Bohländer, M. Dai, M. Filius, C. J. Howard, X. F. van Kooten, S. Ohayon, A. Pomorski, S. Schmid, A. Aksimentiev, E. V. Anslyn, G. Bedran, C. Cao, M. Chinappi, E. Coyaud, C. Dekker, G. Dittmar, N. Drachman, R. Eelkema, D. Goodlett, S. Hentz, U. Kalathiya, N. L. Kelleher, R. T. Kelly, Z. Kelman, S. H. Kim, B. Kuster, D. Rodriguez-Larrea, S. Lindsay, G. Maglia, E. M. Marcotte, J. P. Marino, C. Masselon, M. Mayer, P. Samaras, K. Sarthak, L. Sepiashvili, D. Stein, M. Wanunu, M. Wilhelm, P. Yin, A. Meller, and C. Joo. "The emerging landscape of single-molecule protein sequencing technologies", Nature Methods 18, 604–617, 2021. doi: 10.1038/s41592-021-01143-1

[6] L. Reynaud, A. Bouchet-Spinelli, C. Raillon, and A. Buhot, "Sensing with nanopores and aptamers: a way forward", Sensors 20, 4495, 2020.

[7] A.Holtz, N. Basisty, and B. Schilling, "Quantification and identification of post-translational modifications using modern proteomics approaches". In: Marcus, K., Eisenacher, M., Sitek, B. (eds) Quantitative Methods in Proteomics. Methods in Molecular Biology, vol 2228. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-1024-4_16

[8] G. Sampath, "A binary/digital approach to peptide sequencing and protein identification". TechRxiv.org preprint, August 2022. https://doi.org/10.36227/techrxiv.19318145.v3

[9] M. Ibba and D. Söll. “The renaissance of aminoacyl-tRNA synthesis”. EMBO Reports, 2, 382-387, 2001.

[10] Anonymous. "Post-translational modifications of amino acids", www.cellsignal.com/ptmscan.

[11] Anonymous. “Controlled vocabulary of post-translational modifications", ptmlist.txt at www.uniprot.org.

[12] K. W. Barber and J. Rinehart, "The ABCs of PTMs", Nat Chem Biol. 2018 February 14; 14(3): 188–192. doi:10.1038/nchembio.2572

[13] J. Hermann, L. Schurgers, and V. Jankowski, "Identification and characterization of post-translational modifications: Clinical implications", Molecular Aspects of Medicine 86, 101066, 2022. doi: 10.1016/j.mam.2022.101066

[14] A. M. N. Silva, R. Vitorino, M. Rosário, M. Domingues, C. M. Spickett, and P. Domingues, "Post-translational modifications and mass spectrometry detection", Free Radic. Biol. Med. 65, 925-941, 2013. doi: 10.1016/j.freeradbiomed.2013.08.184

[15] M. R. Larsen, M. B. Trelle, T. E. Thingholm, and O. N. Jensen, “Analysis of posttranslational modifications of proteins by tandem mass spectrometry: Mass spectrometry for proteomics analysis”, BioTechniques 40, 790-798, 2006. https://doi.org/10.2144/000112201
[16] S. Doll and A. L. Burlingame, "Mass spectrometry-based detection and assignment of protein post-translational modifications", ACS Chem. Biol. 2015, 10, 63-71. DOI: 10.1021/cb500904b
[17] S. Ramazi and J. Zahiri, "Post-translational modifications in proteins: resources, tools and prediction methods", Database, baab012, 1–20, 2021. doi:10.1093/database/baab012
[18] Y. Zhang, C. Sohn, S. Lee, H. Ahn, J. Seo, J. Cao, and L. Cai , "Detecting protein and post-translational modifications in single cells with iDentification and qUantification sEparaTion (DUET)", Communications Biology 3, 420, 2020.
[19] H. Xu, Y. Wang, S. Lin, W. Deng, D. Peng, Q. Cui, and Y. Xue, "PTMD: A database of human disease-associated post-translational modifications", Genomics Proteomics Bioinformatics 16, 244–251, 2018. doi: 10.1016/j.gpb.2018.06.004
[20] N. Bagwan, H. H. El Ali, and A. Lundby, "Proteome-wide profiling and mapping of post translational modifications in human hearts", Scientific Reports, 11, 2184, 2021. doi: 10.1038/s41598-021-81986-y
[21] M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars. Computational Geometry: Algorithms and Applications. (3rd edn.), Springer-Verlag, 2008.
[22] D. Avis, B. K. Bhattacharya, and H. Imai, "Computing the volume of the union of spheres", The Visual Computer 3, 323-328,1988.
[23] S. S. Batsanov, "Van der Waals radii of elements", Inorganic Materials 37, 871–885, 2001.
[24] 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., 38, 176–181, 2020. doi: 10.1038/s41587-019-0345-2
[25] X. Liu, Z. Dong, and G. Timp, "Calling the amino acid sequence of a protein/peptide from the nanospectrum produced by a sub-nanometer diameter pore", bioRxiv.org, preprint, October 18, 2021. doi: https://doi.org/10.1101/2021.10.17.464717
[26] 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
[27] A. Crnkovi´c, M. Srnko, G. Anderluh, "Biological nanopores: Engineering on demand". Life 11, 27, 2021. doi: 10.3390/life11010027
[28] Y. Luo, L. Wu, J. Tu, Z. Lu, “Application of solid-state nanopore in protein detection”, Int. J. Mol. Sci. 21, 2808, 2020. doi:10.3390/ijms21082808
[29] L. Restrepo-Pe´rez, G. Huang, P. R. Bohla¨nder, N. Worp, R. Eelkema, G. Maglia, C. Joo, and C. Dekker, "Resolving chemical modifications to a single amino acid within a peptide using a biological nanopore", ACS Nano 13, 13668-13676, 2019. DOI: 10.1021/acsnano.9b05156
[30] C. B. Rosen, D. Rodriguez-Larrea, and H. Bayley, "Single-molecule site-specific detection of protein phosphorylation with a nanopore", Nat. Biotechnol. 32, 179–181, 2014. doi: 10.1038/nbt.2799
[31] L. Restrepo-Pe´rez, C. H. Wong, G. Maglia, C. Dekker, and C. Joo, ""Label-free detection of post-translational modifications with a nanopore", Nano Lett. 19, 7957-7964, 2019. DOI: 10.1021/acs.nanolett.9b03134
[32] T. Ensslen, K. Sarthak, A. Aksimentiev, and J. C. Behrends, "Resolving isomeric posttranslational modifications using a nanopore", bioRxiv preprint, November 28, 2021. doi: 10.1101/2021.11.28.470241
[33] 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.