Optical Single-Molecule Protein Sequencing: Feasibility Study

MAS.862 Final Project

Andrew Payne

Abstract

Single-molecule protein sequencing remains an open problem and an intense area of study. We review measurement modalities by which protein sequencing could be implemented and propose an approach to single molecule protein sequencing using light microscopic readout of synthetic fluorophore labels conjugated to amino acid side chains ("barcodes"). We describe how this could be used in shotgun contexts via conventional microscopy or in *in situ* contexts via super resolution microscopy. An idealized code space of possible two-color barcodes is characterized for the human proteome.

Contents

1 Introduction and state of the art

The central dogma of molecular biology, despite its many simplifications, has guided research for more than fifty years [1]. A substantial technological toolbox has been developed to measure the sequence information encoded by theory's biopolymers at all three levels: DNA, RNA, and protein. Moreover, impressive progress has been made in recent years utilizing optical microscopy and the polymerase chain reaction for measurement of amplified colonies of individual nucleic acids. This so-called second generation of sequencing technologies has extend DNA sequencing into the single-molecule realm, permitting high throughput whole-genome measurements of individual cells and tissues, promising concomitant scientific and biomedical payoff[2]. In contrast, the measurement of individual protein molecules has lagged behind, and investigators must rely on ensemble measurements of protein sequence information from many cells, masking cell-to-cell variations, or else measure only the most abundant proteins in single cell measurements, masking the impact of low-copy number proteins[3].

Modern mass spectrometry (MS) equipment enables routine proteome quantification through direct measurement of expressed proteins[4]. This approach exhibits attomole detection sensitivities for whole proteins and subattomole sensitivities after fractionation and stochastic sampling, implying a detection dynamic range spanning four orders of magnitude[5]. However, expression levels for a typical mammalian proteome span seven orders of magnitude, and low (1-1000) copy-number proteins making up approximately 10% of expressed protein species tend to remain undetected by this method[6]. These proteins are important despite their low frequency: at least one in four display a genetic interaction in double knockout experiments[7]. mRNA transcriptomic analysis correlates well with expressed protein levels at high copy-numbers, but is not as useful a proxy in the low copy number / single-molecule regime where stochastic effects dominate[8]. A viable single-molecule protein sequencing technology is therefore an attractive research direction.

2 Review of Measurement Modalities

We will introduce a scheme in which there exist **reporters** of amino acid identities, **labeling methods** which deliver reporters to proteins, and **sample manipulation** and **instrumentation** which read out reporters. A single-molecule protein sequencing technology will leverage one or more of these ideas to discern a protein's sequence.

2.1 Reporters

This section enumerates the most likely reporters of amino acid identities, although it does not claim to be complete.

2.1.1 Endogenous properties

The simplest possible reporter of an amino acid residue is the acid itself. Amino acid side chains have various properties which permit them to perform the myriad functions required by biological systems, and these properties can be measured. These properties include but are not limited to mass, rigidity, positive or negative charge, acidity or basicity, polarity, hydrophobicity, and endogenous fluorescence. These endogenous properties can, for example, provide contrast in a nanopore electrical measurement, provide a mass/charge fingerprint in a mass spectrometer, or provide an excitation signal in a fluorescence measurement. However, these signals are generally quite weak, and constraints on instrument sensitivity typically (although not always) means that they are only useful in bulk measurements.

2.1.2 Exogenous labels

Here, a label is defined as any agent that modifies the protein of interest in a residue- or sequence-specific manner in order to simplify downstream detection or analysis. The label could, for example, be a synthetic fluorophore to enhance brightness in a optical micrograph, a heavy metal to enhance contrast in an electron micrograph, a conductive or resistive element to enhance contrast in an electrical measurement, and so on. It is not practical to enumerate the particulars of the many possible labels in this report; rather, we will consider a label any agent that improves the sensitivity of the measurement modalities reviewed in Section 2.4.

2.2 Labeling Methods

2.2.1 Chemical reporters on side chains

Here, chemical diversity in amino acid side chains is exploited in order to covalently bond, in a residue-specific manner, a reporter that is more amenable to downstream analysis than the properties of endogenous amino acid. An ideal system of chemical label reporters would involve one reporter per amino acid; however, many amino acids are non-reactive, and so only the reactive amino acids should be considered as targets for chemical conjugation. Moreover, many (most) chemical reactions on amino acid side chains are not orthogonal: one reaction may label multiple species. It is therefore a challenge in chemical biology to find **orthogonal, residue specific** reactions on amino acid side chains.

The best characterized chemical label conjugation reactions are[9]:

1. The NHS-ester-activated amide formation reaction, which specifically targets the free primary amine side chain on lysine residues in mildly basic conditions.

2. The maleimide-activated thioether formation reaction, which specifically targets the free sulfhydryl side chain on cysteine residues in mildly acidic conditions.

3. The carbodiimide-activated amide formation reaction, which specifically targets the free carboxylic acid side chain on glutamic acid residues in mildly acidic conditions.

These three reactions can be carried out sequentially and will feature prominently in subsequent discussion. However, other candidate reactions for chemical label conjugation, including reactions on reactive residues such as tyrosine, arginine, histidine, and aspartic acid, have received a great deal of attention and can also be considered should additional diversity prove a desirable objective[10].

2.2.2 Enzymatic reporters

Here, enzymatic specificity is exploited to covalently bond, in a residue-specific manner, a reporter group. This occurs frequently in biological systems in the form of post-translational modifications, where a vast range of chemical groups are appended to amino acid side chains; phosphorylation, acylation, and glycosylation are among the most frequent examples[11]. However, the use of post-translational modifications to report amino acid sequence is not commonplace, and most efforts to develop methods for single-molecule detection of post-translational modifications have focused on measurement of endogenous modifications rather than using them as a component in a protein sequencer, such as in the case of a nanopore-based detector for phosphorylation[12].

2.2.3 Affinity reagents

Here, non-covalent binding interactions are exploited in order to attach a reporter to a sequence of one or more amino acids. This is traditionally embodied in the form of an antibody[13], but aptamers[14], nanobodies[15], and other emerging classes of affinity reagent[16] can also be considered; moreover, single-molecule counting methods are also under development[17]. However, specificity and reproducibility are often major difficulties with this type of measurement, and the typical size of an epitope - many sequential residues - naively requires one reporter per type of protein [18, 13]. One promising method to circumvent this requirement is to use a affinity reagent specific to the N-terminus of a particular amino acid, and combine multiple N-terminal measurements with sample degradation[19]. However, research in this area is still in its infancy.

2.3 Sample Manipulation

In addition to reporter-based labeling of proteins to enable measurements (“additive methods”), subtractive methods of sequential or targeted protein degradation can unintuitively impart additional information. Here, the investigator is able to take measurements before and after degradation, and by knowing something about how the degradation occurs (e.g. which amino acids could have fragmented), can infer information that may not be encoded by the reporters themselves.

2.3.1 Chemical Degradation

The best-characterized method of chemical protein degradation is the so-called Edman degradation, in which an isothiocyanate is selectively reacted with the N-terminus of the target protein to form a reactive phenylthiocarbamyl (PTC) intermediate. In the presence of anhydrous acid the amide bond closest to the intermediate is cleaved, releasing the intermediate and revealing the N-terminus of the next amino acid in the sequence. Historically, the reaction was carried out en-masse at the e.g. picomole scale and the cleaved PTC derivatives were collected and analyzed in bulk [20]. This yields the sequences of proteins in bulk, but is unsuitable for the kind of single molecule measurements we are interested in here. One approach that *is* consistent with our requirements, proposed in [21], involves taking a single molecule measurement (e.g. via TIRF microscopy, see section 2.4.1) of a sequence of length N (i.e. integrating information from all reporters), followed by Edman degradation and a second single molecule measurement of a sequence of length N-1 and inferring the nature of the cleaved reporter via ratiometric comparision to the original measurement. The effect of cyanogen bromide on proteins may also be of interest, being a means of selective cleavage at methionine residues[22].

2.3.2 Enzymatic Degradation

Many enzymes - such as proteases or aminopeptidases - have evolved to cleave proteins in exquisitely selective fashion, and the additional information imparted via fragmentation can again be harnessed to ease protein sequence analysis. An exhaustive inventory falls outside the scope of this report. However, we note that trypsin, which cleaves c-terminal to lysine and arginine residues [23], and GluC, which cleaves c-terminal to glutamic acid residues [24], are both well characterized tools. Aminopeptidases may also be of interest of a chemical Edman degradation cannot be carried out [25].

2.4 Instrumentation

The preceding sections enumerating the possible additive and subtractive transformations that can occur on amino acid polymers are the set-up for the main event: the sensitive single-molecule measurement. There is tremendous diversity in single-molecule methods, and several promising candidates will be described here.

2.4.1 Light microscopy

If the reporter can be excited in order to emit photons (i.e. it fluoresces), it can be detected with a light microscope if it is sufficiently bright. The commonplace epifluorescence microscope, however, excites the whole sample volume, and the resulting background often makes single-molecule measurements impossible. We thus turn to two specialized types of light microscopes: the confocal microscope and the TIRF microscope. In confocal microscopy, contrast is enhanced by blocking out of focus light with a pinhole [26], and this is widely used for single-molecule protein measurements[27], although sequencing remains aspirational. In total internal reflection (TIRF) microscopy, the excitation light is fully reflected at glass-sample interface to produce an exponentially decaying evanescent wave that only excites fluorophores within approximately 100 nm of the surface [28]. This is also effective at enhancing contrast, again yielding single-molecule protein measurements as well as a proposal for single-molecule sequencing[21].

2.4.2 Electrical (nanopore)

Here, a molecular machine is used to unfold and ratchet a protein through a pore; current across the pore is continuously measured and current modulations due to amino acid reporters can yield the acid's identity. This is one of the more experimentally substantiated approaches to protein sequencing, with a recent demonstration of protein translocation and epitope determination[29]. It should be noted that the reporter need not only modulate electrical measurements; a nanopore can be simultaneously monitored optically to detect e.g. cooperative interactions between fluorescent reporter dyes on the protein and on the nanopore [30].

2.4.3 Electron microscopy

Electron microscopes can measure electron density, and sensitive electron microscopes can discern differences in electron density between amino acids; these differences can also be accentuated by stains or other reporters. Impressive progress has been made recently determining 3D protein structure using cryo-electron microscopy, however, due to sample processing requirements it has yet to be proposed as serious or scalable method of protein sequencing [31].

2.4.4 Mass spectrometry

Mass spectrometry is the workhorse of current bulk protein sequencing efforts, and efforts are being made to extend its capabilities into the single molecule regime. Here, either a whole or fragmented protein is ionized and introduced into a mass analyzer (e.g. quadrupole, time-of-flight). The resulting mass/charge ratio can fingerprint the fragment and bioinformatics approaches can verify its sequence [32]. Typically femtomole quantities of the protein or fragment is needed to generate a detectable signal; however, a recent landmark study permitted a single-molecule mass spectrum measurement of antibodies by tracking vibrational modes of a nanoelectromechanical resonator[33].

2.4.5 Force spectroscopy

Force-induced unfolding of proteins by optical tweezers or atomic force microscopy are well characterized single molecule methods which can generate force spectra which readily identify particular folding states and domains [34]. However, extending the analysis to reveal sequence information is challenging, and the highly technical nature of these experiments will make widespread adoption difficult.

2.4.6 NMR spectroscopy

Nuclear magnetic resonance spectroscopy can identify chemical information in molecules based on absorption and emission of RF pulses by nuclei in a magnetic field. This has been an important tool for structural studies of e.g. protein conformations, as well as a routine analytic tool in chemical synthesis, medicine, etc [35]. However, extending NMR into the single molecule regime is daunting, although efforts involving nitrogen vacancy centers in nanodiamonds shows promise [36]. That said, there is no clear path from the state of the art to a practical means of extracting protein sequence information.

2.5 Scope of the project

While the review above illustrates the design space for single molecule protein measurements to be large indeed, we will limit the scope of this report to involve single molecule measurements of **(1) exogenous labels** (Section 2.1.2) in the form of **(2) fluorescent reporters conjugated to amino acid side chains** (Section 2.2.1) that are detected by **(3) light microscopy** (Section 2.4.1). The sequence and identities of the reporters will be read out one label at at time from the N-terminus using chemical degradation (Section 2.3.1) in the fashion described in [21].

3 Review of Light Microscopy Modalities

A light-microscopy-based protein sequencing technology could be used either **ex situ** or **in situ**, depending on the instrumentation. (We neglect the *in vivo* case, since any widespread modification of protein side chains would inevitably introduced unacceptable toxicity). In the ex situ case, there is a great deal of control over the placement and density of individual molecules, permitting standard diffraction limited optics to be used. If the protein is to be sequenced in situ, the crowded environment of cells or tissue necessitate the use of **super-resolution** methods to resolve the individual molecules. An attractive optical protein sequencing technology would ideally be compatible with one or more of these methods.

3.1 Diffraction limited optics

A collection of densely spaced fluorophores, such as what one may find for a sample of labelled proteins fixed to a surface (or, in general, a biological specimen of interest), can be well-modeled in one dimension by a diffraction grating with a spacing d equal to the fluorophore-to-fluorophore distance. This arrangement will produce a series of diffraction orders (Young's fringes) which, after refraction by a lens, will produce an image that is their diffraction pattern (Figure 1). The microscope aperture must therefore be wide enough to admit at least the zeroth and first orders of the grating (or else periodicity will be absent and the existence of multiple point sources cannot be inferred). This corresponds to a requirement of (from the diagram) $sin(\theta_1) = \lambda/d < sin(a)$ or, for a given aperture, a minimum grating resolution of

$$d_{min} = \frac{\lambda}{sin(a)} \tag{1}$$

This is the Rayleigh limit. In practice this can be improved by a factor of two (since only half the first order fringe need be collected to construct the diffraction patter), and that gives the well-known Abbe diffraction limit of:

$$d_{min} = \frac{\lambda}{2\,n\,sin(a)} = \frac{\lambda}{2NA} \tag{2}$$

where $NA = n * sin(a)$ is the numerical aperture of the lens. **This is a fundamental physical limit on the resolution that can be achieved by conventional microscopy,** and microscopies that circumvent this limit are referred to as "super-resolution" methods. For a standard 1.41 NA lens and a 600 nm emitter, a resolution of approximately 200 nm can be achieved[37].

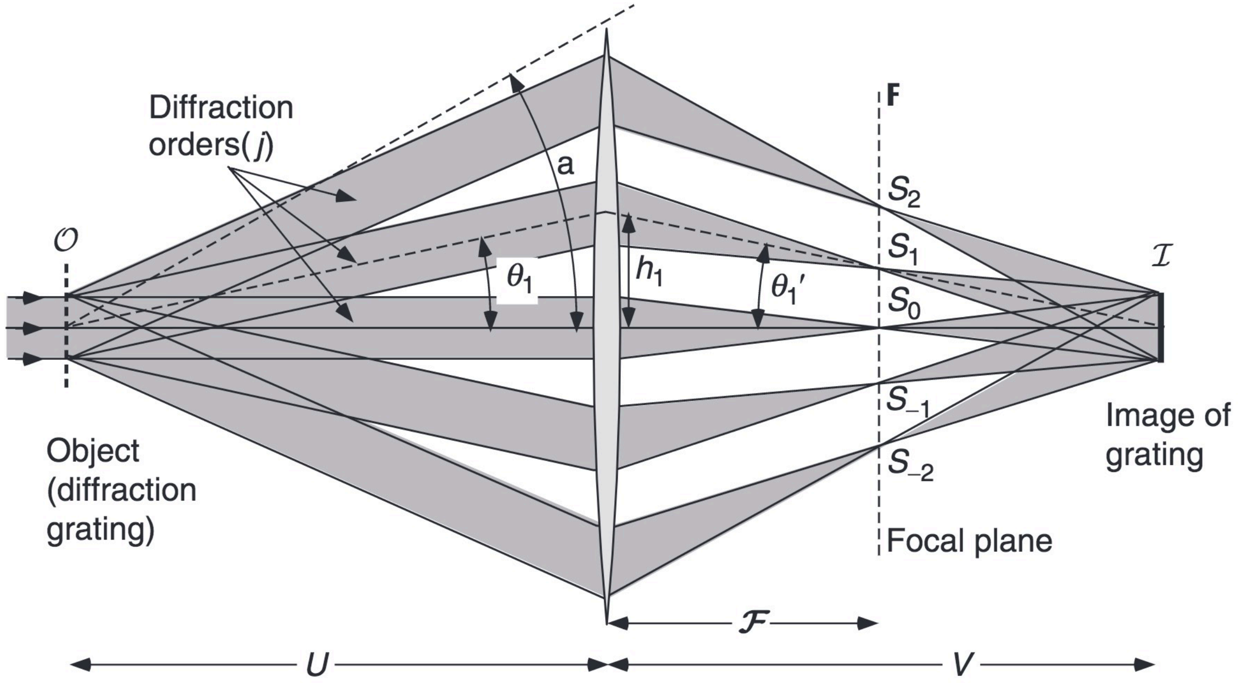

Figure 1: Interference pattern of a diffraction grating. Figure reproduced from [37].

3.2 Stochastic photoswitching approaches to super-resolution

This class of technique includes but is not limited to stochastic optical reconstruction microscopy (STORM)[38], points accumulation by nanoscale topography (PAINT)[39], and photoactivated localization microscopy (PALM)[40]. Here, a collection of fluorophores in a diffraction limited area stochastically switch between bright and dark states such that only a single fluorophore is typically emitting at a given point in time. The single-fluorophore diffraction patterns (point spread functions) have a well defined mean position which can be fit by e,g. a gaussian, and by repeatedly sampling single point spread functions and approximating their means, the emitter's location can be localized at an arbitrary position (Figure 2). In practice, since fluorophores photobleach after repeated measurements, a resolution of 20-30 nm laterally can be achieved.

3.3 Deterministic photoswitching approaches

This class of technique involves photoswitching of fluorophores in a selective rather than stochastic manner and includes techniques such as stimulated emission depletion (STED) microscopy and ground state depletion (GSD) microscopy. Here, the emitter is simultaneously excited by a standard laser and de-excited by a donut-shaped beam. Co-localization of the two beams guarantees that only a fluorophore inside the well-defined inner region of the donut will emit; if a signal is detected,

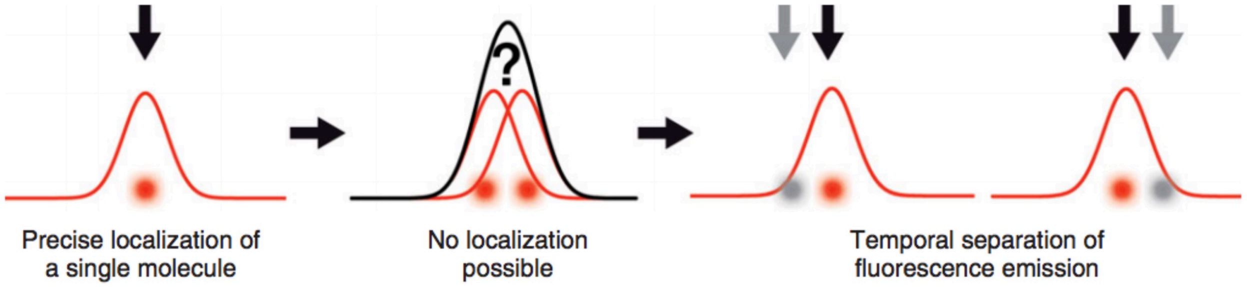
?
Precise localization of a single molecule
No localization possible
Temporal separation of fluorescence emission

Figure 2: Sub-diffraction-limited localization via stochastic photoswitching. Figure reproduced from [41].

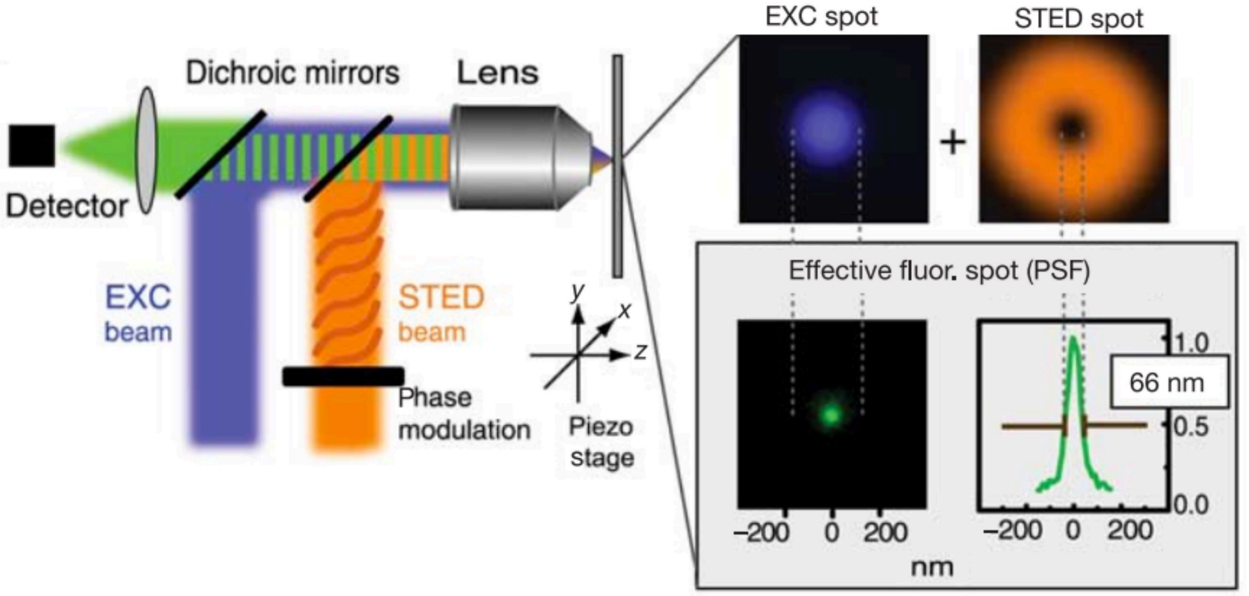

Figure 3: Figure reproduced from [43].

the fluorophore is within the donut hole. By scanning the two beams across the sample a super-resolution image can be reconstructed. In this approach, lateral resolutions on the order of 30 nm can be achieved[42].

3.4 Structured illumination approaches

This class of technique involves overlaying line patterns (structure) on the sample. The emitted light will thus be the product of the sample emitter distribution and the structured light, and this product will contain moire fringes which contain additional high-frequency information (Figure 4). The information can be processed to yield additional information about the sample, in practice extending the microscope resolution by about a factor of two (approx. 150 nm lateral resolution)[44].

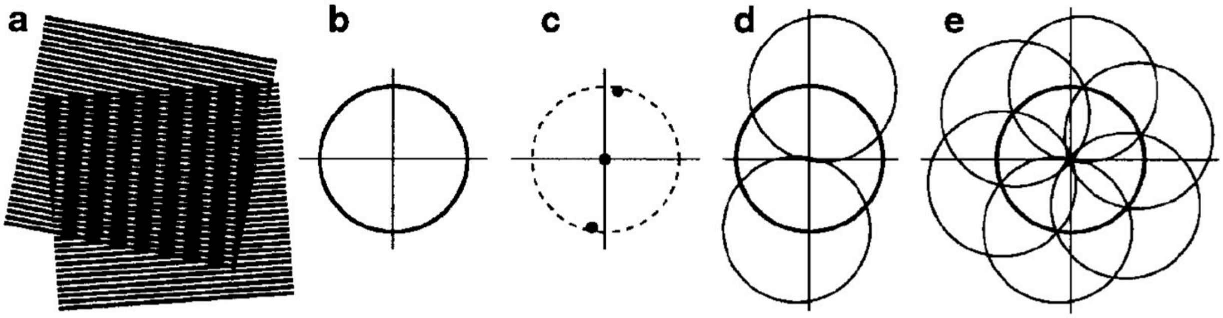

Figure 4: Additional information contained in Moire fringes. Figure reproduced from [44].

3.5 Physical Expansion

In contrast to the techniques presented so far, one approach to super resolution involves engineering the sample rather than the microscope. Here, the sample is embedded in a swellable hydrogel polymer matrix. After fluorophore transfer to the matrix, the sample is homogenized and the hydrogel is swollen via dialysis activation of the polyelectrolyte effect (Figure 5. Samples swollen by a factor of 4.5 in linear dimension have been reported, resulting in an effective resolution of 300 nm / 4.5 ≈ 65 nm[45].

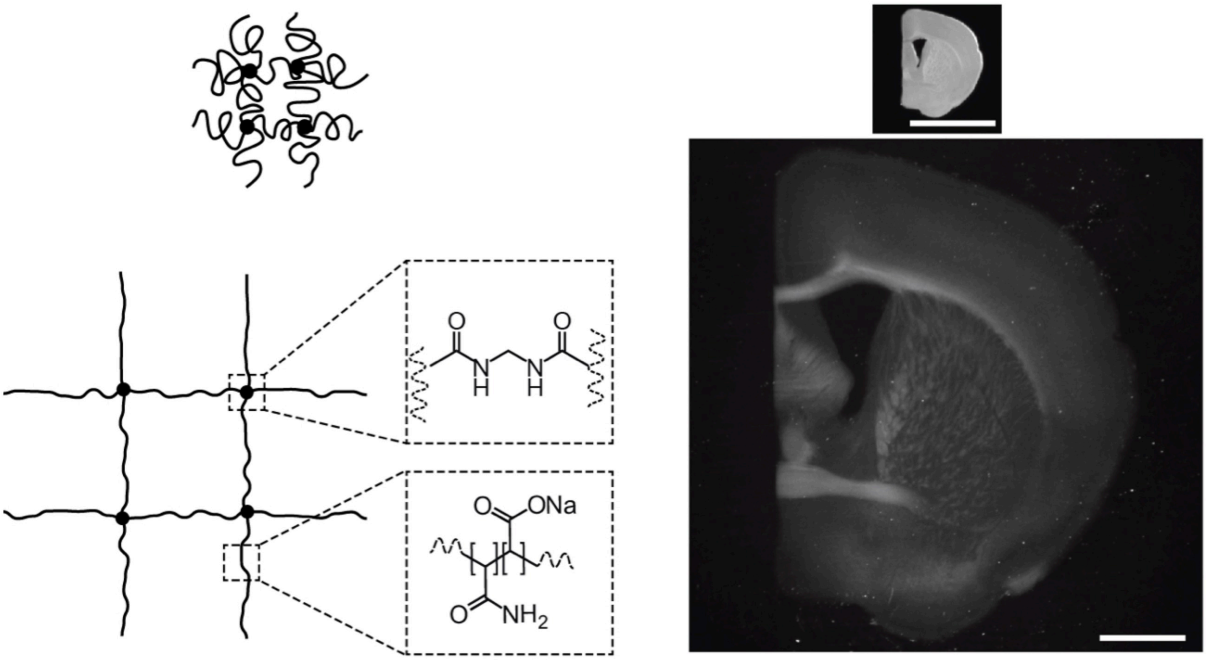

Figure 5: Swellable polymers in Expansion Microscopy. Figure adapted from [45].

4 Analysis

As we have stated in Section 2.5, we seek to make single molecule measurements of **(1) exogenous labels** in the form of **(2) fluorescent reporters conjugated to amino acid side chains** that are detected by **(3) light microscopy.** We will refer to a particular sequence of fluorescent reporters as an **m-color barcode of length n**, where we use color interchangeably with a particular labelled amino acid and we count labels from the N-terminus. For example, if we label the amino acids E and Y in the sequence MYTARGETPRQTEIN, we would get "YEE," a 2-color barcode of length 3.

We will now explore the "code space" for this arrangement. Specifically, we would like to ask, given a particular instantiation of (1), (2), and (3), which proteins can be uniquely identified from a given pool. We will assume our pool is the **UniProtKB/Swiss-Prot canonical human proteome** [46], which is a hand-annotated set of all proteins known to be expressed by the canonical human genome (i.e. one representative protein for each gene). We will neglect protein isoforms and post-translational modifications in our analysis, but this additional diversity should be considered in a future work. We thus have a pool of approximately 20,000 protein sequences to analyze.

4.1 Information in the human proteome

If we are to construct a coding scheme for the human proteome, which we will call X our first question should be: what is the entropy $H(X)$ of the human proteome? We can easily find an upper bound for this value if we assume that we have no a priori knowledge about the likelihood of observing a particular sequence x. Then, we will maximize the entropy function

$$H(X) = -\sum_{x \in X} p(x)\, log_2\, p(x) \tag{3}$$

There are N =20139 proteins in the 03-15-16 release of the UniProtKB/Swiss-Prot human proteome, so $p(x) = 1/N$ and we find that

$$H(X) = 14.29 \approx 15 \text{ bits/protein} \tag{4}$$

So, we will need at least 15 bits of information in our fluorescent barcodes in order to uniquely identify a protein; in fact, this is just a restatement that $2^{14} < N < 2^{15}$. We reiterate that this is the worst case estimate when $H(X)$ is maximized; a less naive but more strenuous approach can take into account protein frequency (structural proteins will, for example occur more often than functional ones). This could be realized on a whole-organism level by shotgun methods or in a more nuanced manner by in-situ methods.

4.2 Information in 2-color barcodes

Given an $m = 2$ color barcoding scheme, how much information can we expect to extract from individual proteins, and will it exceed the average of 15 bits required for unique identification? We will in this section examine several figures of merit ultimately summarized in Table 1.

4.2.1 Average information from amino acid frequencies and protein lengths

One way we can approximate this is from examination of the proteome's length distribution and amino acid frequencies. The length distribution is shown in Figure 6. The canonical human proteome is composed of approximately 10^7 amino acids, and their frequency distribution is also

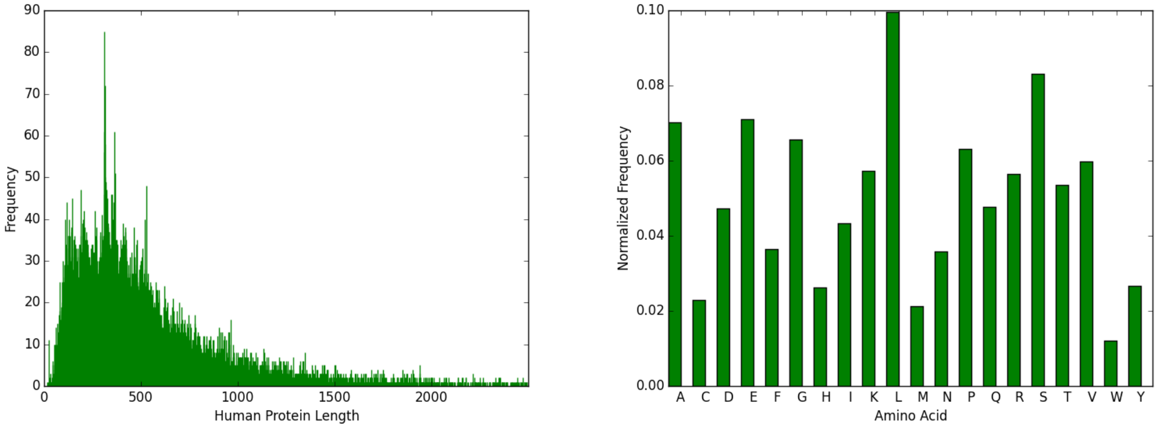

Figure 6: **Left:** Distribution of protein lengths (number of amino acids) in the human proteome. **Right**: occurrence frequency of amino acids.

characterized in Figure 6. The first distribution yields an average protein length of $\langle l \rangle$ =561 amino acid residues. Combining this information with the information in the second distribution we can make rough predictions regarding the expected information in a two-bit barcode. Taking the two most frequently occurring amino acids, L and S, occurring with frequencies f_L and f_S, the average number that occur per protein will be $\langle LS \rangle = (f_L + f_S)\langle l \rangle$ or 102 L and S residues. Similarly, for the least frequent residues, M and W, we have $\langle MW \rangle$= 18. Therefore, to a first approximation, we have between 18 and 102 bits to form our barcode with, and this is sufficient, **on average**, to meet the minimum barcoding requirement of 15 bits/protein regardless of the two colors chosen. However, it is evident from Figure 6 that there is a long tail of protein lengths which suggests that reliance on average information is of dubious value.

4.2.2 Accessible 2-color barcodes

We can construct all 2-color barcodes for the human proteome for several representative colors (without yet making any statements about uniqueness of barcodes) and examine their length distributions. If we call N_l the number of proteins of length l, and we call N_l^m the number of uniquely possible m-color barcodes of length l, it will be impossible to uniquely identify all proteins when the condition

$$\frac{N_l}{N_l^m} > 1 \tag{5}$$

This condition is indeed met for small l and is plotted alongside their distributions (Figures 7, 8, 9). Here we find that virtually all LS barcodes are of sufficient length to uniquely identify their proteins, whereas 89.9% and 59.4% of CK and MW barcodes are of sufficient length to uniquely identify their generating proteins, respectively, to uniquely identify their generating proteins.

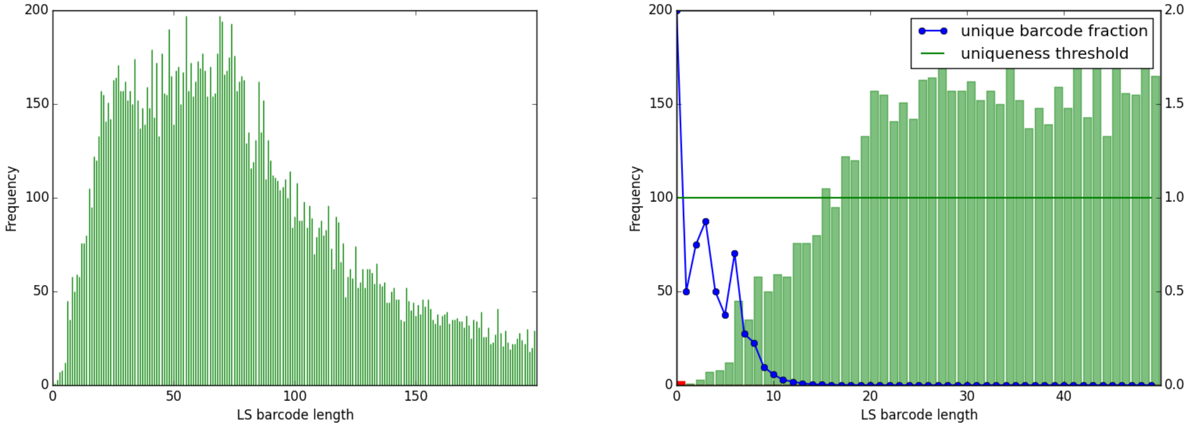

Figure 7: **Left:** LS barcode length distribution**. Right:** $\frac{N_l}{N_l^m}$ and unity threshold plotted against that distribution.. All proteins can in principle be uniquely barcoded.

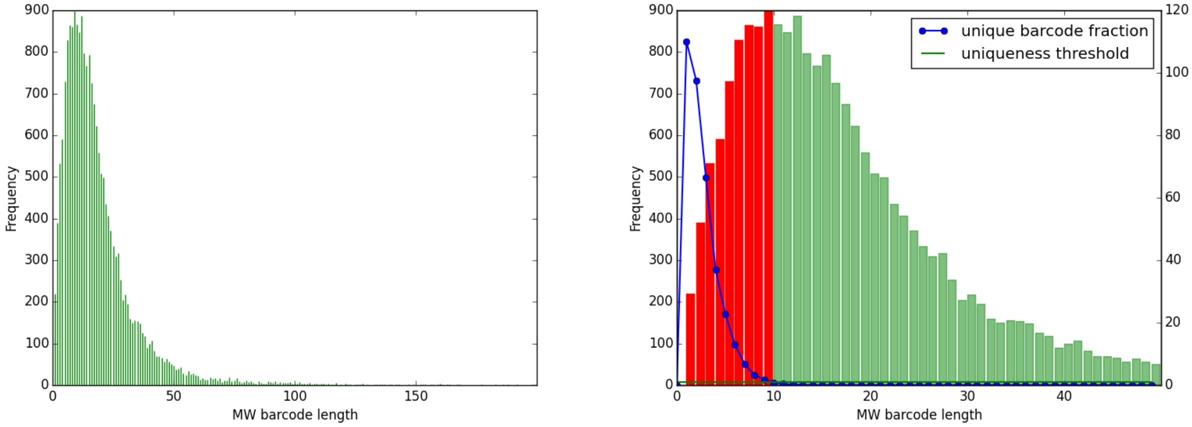

Figure 8: **Left:** MW barcode length distribution**. Right:** $\frac{N_l}{N_l^m}$ and unity threshold plotted against that distribution. 59.4% of proteins can be uniquely barcoded.

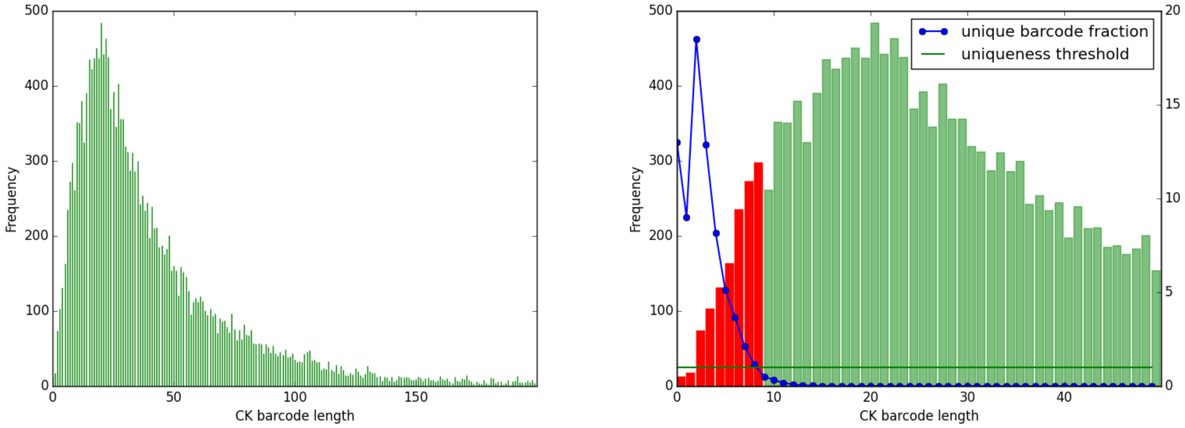

Figure 9: **Left:** CK barcode length distribution**. Right:** $\frac{N_l}{N_l^m}$ and unity threshold plotted against that distribution. 89.9#% of proteins can be uniquely barcoded.

4.3 Uniquely identifiable proteins for 2-color barcodes

Of the barcodes which can in principle uniquely identify their generating proteins, how often does this in fact occur instead of a barcode collision? The number of barcodes with a unique parent protein, N_b, is compared to the total number of proteins in the proteome, N. We define uniqueness in this idealized model as

$$U = \frac{N_b}{N} \tag{6}$$

Here, we examine all 20^2 possible pairs of colors (Figure **??**). Discarding the diagonal, we find that indeed a 'LS' fingerprint maximizes the number of uniquely identifiable human proteins (97.9%) while a 'MW' fingerprint minimizes the number of uniquely identifiable human proteins (57.9%). If we consider only fingerprints with tractable synthetic strategies 'CK' fingerprint, we find that 86.4% of human proteins are in principle uniquely identifiable. These values differ only slightly from the estimates presented in Section 4.2.2, implying that the majority of unidentifiable proteins are indeed of small$l$ and that there are few collisions once the space of possible barcodes becomes large.

4.4 Summary

The conclusions of this section are summarized in Table 1.

Barcode fraction estimate	**LS barcode**	**MW barcode**	**CK barcode**
Sec. 4.2.1: average AA frequency ($(f_L + f_S)\langle l \rangle$)	100%	100%	100%
Sec. 4.2.2: accessible fraction ($\frac{N_l}{N_l^m} > 1$)	99.9%	59.4%	89.9%
Sec. 4.3: uniquely identifiable fraction ($U = \frac{N_b}{N}$)	97.9%	57.9%	86.4%

Table 1: Summary of proteome barcode estimates for three representative pairs of colors.

5 Conclusions and Future Work

We have shown that, in principle, a large fraction of the human proteome is amenable to single-molecule optical readout using two-color barcodes. In this analysis, between 57.9% and 97.9% of the human proteome can be uniquely barcoded depending on colors chosen. If colors with tractable labeling strategies are chosen, $>$ 86.4% of the proteome is still uniquely barcodable. Future work may want to examine additional colors to determine what additional gains may be realized.

However, this analysis does not take into account the significant technical difficulties involved in practically realizing such a scheme. The inevitable insertions, deletions and transpositions which would result from imperfect labelling and degradation efficiencies or photobleaching, for example, will expand the space of observable barcodes and cause collisions. Future work may seek to, for example, simulate the space of possible barcodes in a fashion similar to [21]. We have nonetheless established an upper bound on the problem.

References

[1] Crick, F. *et al.* Central dogma of molecular biology. *Nature* **227**, 561–563 (1970).

[2] Schuster, S. C. Next-generation sequencing transforms today's biology. *Nature* **200**, 16–18 (2007).

[3] Hamdan, M. & Righetti, P. G. Modern strategies for protein quantification in proteome analysis: advantages and limitations. *Mass spectrometry reviews* **21**, 287–302 (2002).

[4] Schieltz, D. M. & Washburn, M. P. Analysis of complex protein mixtures using multidimensional protein identification technology (mudpit). *Cold Spring Harbor Protocols* **2006**, pdb–prot4555 (2006).

[5] Thakur, S. S. *et al.* Deep and highly sensitive proteome coverage by lc-ms/ms without prefractionation. *Molecular & Cellular Proteomics* **10**, M110–003699 (2011).

[6] Picotti, P., Bodenmiller, B., Mueller, L. N., Domon, B. & Aebersold, R. Full dynamic range proteome analysis of s. cerevisiae by targeted proteomics. *Cell* **138**, 795–806 (2009).

[7] Butland, G. *et al.* esga: E. coli synthetic genetic array analysis. *Nature methods* **5**, 789–795 (2008).

[8] Taniguchi, Y. *et al.* Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–538 (2010).

[9] Hermanson, G. T. *Bioconjugate techniques* (Academic press, 2013).

[10] Baslé, E., Joubert, N. & Pucheault, M. Protein chemical modification on endogenous amino acids. *Chemistry & biology* **17**, 213–227 (2010).

[11] Khoury, G. A., Baliban, R. C. & Floudas, C. A. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Scientific reports* **1** (2011).

[12] Rosen, C. B., Rodriguez-Larrea, D. & Bayley, H. Single-molecule site-specific detection of protein phosphorylation with a nanopore. *Nature biotechnology* **32**, 179–181 (2014).

[13] Taussig, M. J. *et al.* Proteomebinders: planning a european resource of affinity reagents for analysis of the human proteome. *Nature Methods* **4**, 13–17 (2007).

[14] Ellington, A. D. & Szostak, J. W. In vitro selection of rna molecules that bind specific ligands. *nature* **346**, 818–822 (1990).

[15] Rothbauer, U. *et al.* Targeting and tracing antigens in live cells with fluorescent nanobodies. *Nature methods* **3**, 887–889 (2006).

[16] Skerra, A. Alternative non-antibody scaffolds for molecular recognition. *Current opinion in biotechnology* **18**, 295–304 (2007).

[17] Tessler, L. A., Reifenberger, J. G. & Mitra, R. D. Protein quantification in complex mixtures by solid phase single-molecule counting. *Analytical chemistry* **81**, 7141–7148 (2009).

[18] Baker, M. Blame it on the antibodies. *Nature* **521**, 274–6 (2015).

[19] Havranek, J. J. & Borgo, B. Molecules and methods for iterative polypeptide analysis and processing (2014). US Patent App. 14/211,448.

[20] Laursen, R. A. Solid-phase edman degradation. *European Journal of Biochemistry* **20**, 89–102 (1971).

[21] Swaminathan, J., Boulgakov, A. A. & Marcotte, E. M. A theoretical justification for single molecule peptide sequencing. *PLoS computational biology* **11**, e1004080–e1004080 (2015).

[22] Gross, E. The cyanogen bromide reaction. *Methods in enzymology* **11**, 238–255 (1967).

[23] Chen, E. I., Hewel, J., Felding-Habermann, B. & Yates, J. R. Large scale protein profiling by combination of protein fractionation and multidimensional protein identification technology (mudpit). *Molecular & Cellular Proteomics* **5**, 53–56 (2006).

[24] Drapeau, G. R., Boily, Y. & Houmard, J. Purification and properties of an extracellular protease of staphylococcus aureus. *Journal of Biological Chemistry* **247**, 6720–6726 (1972).

[25] Sträter, N., Sun, L., Kantrowitz, E. & Lipscomb, W. N. A bicarbonate ion as a general base in the mechanism of peptide hydrolysis by dizinc leucine aminopeptidase. *Proceedings of the National Academy of Sciences* **96**, 11151–11155 (1999).

[26] Minsky, M. Memoir on inventing the confocal scanning microscope. *Scanning* **10**, 128–138 (1988).

[27] Schuler, B., Lipman, E. A. & Eaton, W. A. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature* **419**, 743–747 (2002).

[28] Axelrod, D., Thompson, N. L. & Burghardt, T. P. Total internal reflection fluorescent microscopy. *Journal of microscopy* **129**, 19–28 (1983).

[29] Nivala, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation through an [alpha]-hemolysin nanopore. *Nature biotechnology* **31**, 247–250 (2013).

[30] Soni, G. V. *et al.* Synchronous optical and electrical detection of biomolecules traversing through solid-state nanopores. *Review of Scientific Instruments* **81**, 014301 (2010).

[31] Dubochet, J. *et al.* Cryo-electron microscopy of vitrified specimens. *Quarterly reviews of biophysics* **21**, 129–228 (1988).

[32] Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **5**, 976–989 (1994).

[33] Naik, A., Hanay, M., Hiebert, W., Feng, X. & Roukes, M. Towards single-molecule nanomechanical mass spectrometry. *Nature nanotechnology* **4**, 445–450 (2009).

[34] Carrion-Vazquez, M. *et al.* Mechanical and chemical unfolding of a single protein: a comparison. *Proceedings of the National Academy of Sciences* **96**, 3694–3699 (1999).

[35] Dyson, H. J. & Wright, P. E. Unfolded proteins and protein folding studied by nmr. *Chemical reviews* **104**, 3607–3622 (2004).

[36] Perunicic, V. S., Hall, L. T., Simpson, D. A., Hill, C. D. & Hollenberg, L. C. Towards single-molecule nmr detection and spectroscopy using single spins in diamond. *Physical Review B* **89**, 054432 (2014).

[37] Lipson, A., Lipson, S. G. & Lipson, H. *Optical physics* (Cambridge University Press, 2010).

[38] Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm). *Nature methods* **3**, 793–796 (2006).

[39] Jungmann, R. *et al.* Single-molecule kinetics and super-resolution microscopy by fluorescence imaging of transient binding on dna origami. *Nano letters* **10**, 4756–4761 (2010).

[40] Betzig, E. *et al.* Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313**, 1642–1645 (2006).

[41] Jungmann, R., Scheible, M. & Simmel, F. C. Nanoscale imaging in dna nanotechnology. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology* **4**, 66–81 (2012).

[42] Zhong, H. Photoactivated localization microscopy (palm): an optical technique for achieving˜ 10-nm resolution. *Cold Spring Harbor Protocols* **2010**, pdb–top91 (2010).

[43] Willig, K. I., Rizzoli, S. O., Westphal, V., Jahn, R. & Hell, S. W. Sted microscopy reveals that synaptotagmin remains clustered after synaptic vesicle exocytosis. *Nature* **440**, 935–939 (2006).

[44] Gustafsson, M. G. Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy. *Journal of microscopy* **198**, 82–87 (2000).

[45] Chen, F., Tillberg, P. W. & Boyden, E. S. Expansion microscopy. *Science* **347**, 543–548 (2015).

[46] Consortium, U. *et al.* Uniprot: a hub for protein information. *Nucleic acids research* gku989 (2014).