Some Model Notation Suggestions

Cameron A. Schmidt

25 November 2024

1. Definition of the Discrete Vector Space

The following suggestions may improve the precision of the mathematical model in the paper. Feel free to use these, or ignore as you see fit.

A protein containing R cysteine residues can occupy one of two general states in which the cysteine is either oxidized (i.e., dissulfide, sulfenide, sulfinide, sulfonide, etc) or reduced (thiol or thiolate). The state space describing the number of available cysteine redox proteoforms for a single protein ($\mathcal{I}_{\text{protein}}$) can be described:

$$\mathcal{I}_{\text{protein}} = \{\mathbf{i} \in \{0,1\}^R\} \tag{1}$$

where:

- R: Number of cysteine residues in the protein's primary structure.
- $\{0,1\}^R$: Binary vector space representing the redox states of cysteine residues (0 = reduced, 1 = oxidized).

The state space can be expanded to consider the size of the proteome, for which every genetically encoded protein has a specific number of cysteine residues R. The proteome can thus be grouped into k classes based on the respective integer value of R for each genetically encoded protein. The state space for an entire proteome, taking only one copy of every protein to account, can be described:

$$\mathcal{I}_{\text{proteome}} = \bigcup_{k=1}^{K} \bigcup_{j=1}^{J_k} \{\mathbf{i}_{k,j} \in \{0,1\}^{R_{k,j}}\} \tag{2}$$

where:

- K: Total number of possible cysteine residue classes.
- J_k: Total number of genetically encoded proteins j in residue class k.
- $\mathbf{i}_{k,j}$: State vector for one molecule of genetically encoded protein j in class k.
- $R_{k,j}$: Number of cysteine residues in protein j within class k.

Taking into account the expression levels of each protein in each class k, we have:

$$\mathcal{I}_{\text{expressed}} = \bigcup_{k=1}^{K} \bigcup_{j=1}^{J_k} \bigcup_{m=1}^{M_{k,j}} \{\mathbf{i}_{k,j,m} \in \{0,1\}^{R_{k,j}}\} \tag{3}$$

- K: Total number of possible cysteine residue classes k.
- J_k: Number of genetically encoded proteins j within class k.
- $M_{k,j}$: Total number of expressed proteins m for each genetically encoded protein j within class k.
- $\mathbf{i}_{k,j,m}$: State vector for molecule m of genetically encoded protein j in class k.
- $R_{k,j}$: Number of cysteine residues in m molecules of genetically encoded protein j within class k.

2. Size of the Discrete Vector Space

If the oxidation state of each cysteine residue in the expressed proteome is independent and identically distributed (i.i.d), and the number of oxidation states $n = 2$ then the total size of the vector space can be obtained as follows:

$$\mathcal{I}_{\text{Total}} = 2^{\sum_{k=1}^{K} \sum_{j=1}^{J_k} M_{k,j} R_{k,j}} \tag{4}$$

Where the exponent represents the total number of residues in the expressed proteome by abundance. Notably, $M_{k,j}$ constrains the upper limit on biologically accessible proteoform states, and the number of theoretically available states is much larger than those empirically observed in biological systems. The biologically accessible vector space can be denoted:

$$\mathcal{I}_{\text{Accessible}} \leq 2^{\sum_{k=1}^{K} \sum_{j=1}^{J_k} M_{k,j} R_{k,j}} \tag{5}$$

The biologically accessible state space is smaller than the theoretical space for many reasons including, but not limited to:

1. The number of expressed molecules of each protein.
2. Bias in the distribution of residue classes in the genetically encoded proteome. For example, proteins with large numbers of cysteines are rare, while proteins with only a few cysteines are comparatively abundant.
3. The probabilities of cysteine oxidation may or may not be i.i.d. and stong conditioning of the redox state of each cysteine due to common environmental conditions likely predominate.

3. The Distribution of Oxidation States

To provide a flexible model for the distribution of redox states, a binomial distribution may be considered as a reasonable starting point due to the assumption of a binary redox state vector. For a single protein, we have:

$$P(\eta_{ox}|R,p) = \binom{R}{\eta_{ox}} p^{\eta_{ox}}(1-p)^{\eta_{ox}} \tag{6}$$

where

- η_{ox}=the number of oxidized cysteines in the protein
- R=the number of cysteine residues in the protein
- p=the probability of observing a cysteine in an oxidized state

For the probability of oxidation, we expect that:

- $p \to 0$ under reducing conditions
- $p \to 1$ under oxidizing conditions

The mean of the binomial distribution is given by Rp, and the variance is $Rp(1-p)$. In the scenario where $p = 0.5$, the expected state for any given protein is that roughly half of its oxidizeable cysteines will be in an oxidized state. Interestingly, proteins containing more than one cysteine residue would exhibit $R/2$ symmetry, indicating that there are > 1 distinct proteoforms that give the same net oxidation state despite having different cysteines oxidized.

Finally, the distribution over the entire expressed proteome can be described:

$$P_{total}(\eta_{ox}) = \sum_{k=1}^{K}\sum_{j=1}^{J_k} M_{k,j} P(\eta_{ox}|R_{k,j}, p_{k,j}) \tag{7}$$

where

- $p_{k,j}$ is the probability of observing cysteine in an oxidized state in the m^{th} molecule of the j^{th} expressed protein of the k^{th} residue class.

For large R, the binomial distribution takes on a Gaussian like shape, with mean $\mu = Rp$. This property of the model may be useful for general considerations of redox equilibrium states. However, this assumption may not hold in subcellular compartments with redox environment conditions that skew the value of p. Ultimately, this model of the distribution of cysteine redox states, though highly simplified, may inform the interpretation of empirical measurements, or serve as a basis for simulation of proteoform distributions under specific biological conditions.