SUPPORTING INFORMATION

Estimating the distribution of protein post-translational modification states by mass-spectrometry

Philip D. Compton[1], Neil L. Kelleher[1,2,†], Jeremy Gunawardena[3,†]

[1] Department of Chemistry, Northwestern University, Evanston, IL, USA
[2] Department of Molecular Biosciences, Northwestern University, Evanston, IL, USA
[3] Department of Systems Biology, Harvard Medical School, Boston MA, USA

[†]Corresponding authors
n-kelleher@northwestern.edu; jeremy@hms.harvard.edu, +1 (617) 432 4839

EXAMPLES AND PROOFS OF MATHEMATICAL RESULTS

Contents

Background on sets S-2
Examples of cleavage and fragment matrices S-2
Proof of Theorem 1 . . . . . S-3
Supporting references S-5

Background on sets

A *set* is any collection of previously defined mathematical entities. The entities we work with here are all numbers, $1, \cdots, n$, representing sites on a protein and our sets are all finite collections. A set can be defined by enclosing its *elements* in curly brackets, as in $S = \{1, \cdots, n\}$. Alternatively, it can be defined by specifying when an element occurs in the set, as in $S = \{i \,|\, 1 \leq i \leq n\}$, where the symbol "|" should be read as "such that". The elements of a set may be listed within the curly brackets in any order and with repetitions, so that $\{1, 2, 3\} = \{3, 2, 1, 3, 1, 2, 2\}$. This convention makes it easy to define operations on sets, as below, but it is often convenient to use *standard notation* for sets of numbers, in which the elements are listed in order without repetition.

If an entity i is (or is not) an element of a set S, this is denoted $i \in S$ ($i \notin S$, respectively). The set with no elements, called the *empty set*, is denoted $\emptyset$. The empty set plays the same role for sets as 0 does for numbers. The *size* of a set is the number of distinct elements in it and is denoted $\#U$ for a set U. Hence, $\#\emptyset = 0$ and $\#\{3, 2, 1, 3, 1, 2, 2\} = 3$.

A set U is a *subset* of a set V, denoted $U \subseteq V$, if every element in U is also in V, so that $i \in U$ implies that $i \in V$. (This allows for the possibility that U is the same as V; if that is not true, it can be denoted $U \subset V$.) The subsets of a set V can be collected together in a new set, sometimes called the *power set* of V, which is denoted $\mathcal{P}(V)$. If a set has size n, it has 2^n subsets, or, in other words, $\#\mathcal{P}(V) = 2^{\#V}$. The assertions $U \subseteq V$ and $U \in \mathcal{P}(V)$ are equivalent.

If U and V are sets then their *intersection*, denoted $U \cap V$, consists of those elements which are common to both U and V: $U \cap V = \{i \,|\, i \in U \text{ and } i \in V\}$. Their *union*, denoted $U \cup V$, consists of those elements which are in either U or V or both: $U \cap V = \{i \,|\, i \in U \text{ or } i \in V\}$. The *complement* of V in U, denoted $U \setminus V$, consists of those elements which are in U but are not in V: $U \setminus V = \{i \,|\, i \in U \text{ and } i \notin V\}$. If $U = \{2, 4, 6\}$ and $V = \{3, 4, 5\}$ then $U \cap V = \{4\}$, $U \cup V = \{2, 3, 4, 5, 6\}$, $U \setminus V = \{2, 6\}$ and $V \setminus U = \{3, 5\}$.

Examples of cleavage and fragment matrices

To exhibit the matrices corresponding to cleavage functions, an ordering must be chosen on the unit vectors. Recall that these correspond to subsets. We order subsets in blocks of increasing size and, within each such block, we order the subsets in increasing lexicographic order of sites: if $\{i_1, i_2, i_3\}$ and $\{j_1, j_2, j_3\}$ are two subsets of size 3 in standard form, then we treat $i_1 i_2 i_3$ and $j_1 j_2 j_3$ as "words" and order them as in a dictionary. Hence, $\{1, 4, 8\} < \{1, 5, 7\} < \{2, 3, 4\}$.

Suppose $n = 3$ sites, so that $S = \{1, 2, 3\}$, and there is a protease that cleaves between sites 1 and 2 only, so that the peptides correspond to the two cleavage subsets $S_1 = \{1\}$ and $S_2 = \{2, 3\}$. The cleavage matrices for $c_{\{1\}}$ and $c_{\{2,3\}}$, as described in the main text, are, respectively,

	$\emptyset$	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
$\emptyset$	1	0	1	1	0	0	1	0
$\{1\}$	0	1	0	0	1	1	0	1

	$\emptyset$	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
$\emptyset$	1	1	0	0	0	0	0	0
$\{2\}$	0	0	1	0	1	0	0	0
$\{3\}$	0	0	0	1	0	1	0	0
$\{2,3\}$	0	0	0	0	0	0	1	1

(1)

For convenience, the modforms in S and S_1 are shown in order above and to the left, respectively, in each matrix.

For an example of a fragment matrix, consider a protein with $n = 5$ sites and those modforms with $k = 2$ modifications. As described in the main text, the fragment $F = \{2, 3, 4\}$ gives rise to the fragment matrix,

	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{1,5\}$	$\{2,3\}$	$\{2,4\}$	$\{2,5\}$	$\{3,4\}$	$\{3,5\}$	$\{4,5\}$
0	0	0	0	1	0	0	0	0	0	0
1	1	1	1	0	0	0	1	0	1	1
2	0	0	0	0	1	1	0	1	0	0
3	0	0	0	0	0	0	0	0	0	0

(2)

Proof of Theorem 1

Numerical calculations originally suggested to us that the row spaces of cleavage matrices obeyed an inclusion-exclusion formula for their dimension. This was surprising, at the time, because vector space dimension does not usually follow an inclusion-exclusion formula. One of us (JG) discussed this conjecture with Professor Bernd Sturmfels (Department of Mathematics, UC Berkeley), who pointed out that it follows from a more general result of Hoşten and Sullivant—Theorem 2.6 of [1]—concerning hierarchical models in algebraic statistics [2]. This was very helpful to us at the time. To keep this paper self-contained, we present here an independent and elementary proof of Theorem 1 in the same language as used in the rest of this paper.

As above, let $S = \{1, \cdots, n\}$ be the universe of protein modification sites and recall the set notation described above, which we will use without further mention. The basic idea of the proof is to examine more closely the rows in the cleavage matrices (Eq.1). If $U \subseteq S_1$ is any peptide modform arising from the cleavage subset S_1, let ρ_{U,S_1} denote the corresponding row in the cleavage matrix of c_{S_1}. Restating the definition of the cleavage map in Eq.2 of the main text, we can write

$$\rho_{U,S_1} = \sum_{W \in \mathcal{P}(S \setminus S_1)} e_{W \cup U} \,. \tag{3}$$

If $S_1 = S$, we see that $\rho_{U,S} = e_U$. We can, in fact, generalise Eq.3 further. Suppose that $U \subseteq S_2 \subseteq S_1$. Then,

$$\rho_{U,S_2} = \sum_{W \in \mathcal{P}(S_1 \setminus S_2)} \rho_{W \cup U,S_1} \,. \tag{4}$$

This follows because the function $\mathcal{P}(S \setminus S_1) \times \mathcal{P}(S_1 \setminus S_2) \to \mathcal{P}(S \setminus S_2)$ which takes (W, W') to $W \cup W'$ is a bijection. Eq.3 follows from Eq.4 by taking $S_1 = S$.

Now suppose that $S_1, \cdots, S_N \subseteq S$ are cleavage subsets arising from any patterns of cleavage with multiple proteases. Let $X = \mathcal{P}(S_1) \cup \cdots \cup \mathcal{P}(S_N)$. Note that $\mathcal{P}(S_1 \cup S_2) \neq \mathcal{P}(S_1) \cup \mathcal{P}(S_2)$. For instance, if $S_1 = \{1, 2\}$ and $S_2 = \{2, 3\}$, then $\mathcal{P}(S_1 \cup S_2)$ contains the subset $\{1, 2, 3\}$ while $\mathcal{P}(S_1) \cup \mathcal{P}(S_2)$ does not. X is the essential object needed to understand cleavages: the number of elements in X, or $\#X$, is the quantity specified in Theorem 1.

If $U \in X$ is some peptide modform arising from the cleavages, let $\phi(U) \in \{1, \cdots, N\}$ be the index for an arbitrary choice of cleavage subset which gives rise to this modform, so that $U \in \mathcal{P}(S_{\phi(U)})$. The function $\phi : X \to \{1, \cdots, N\}$ is otherwise arbitrary. The row vectors $\rho_{U,S_{\phi(U)}}$ for $U \in X$ are then in one-to-one correspondence with the elements of X.

As for the quantity we want to determine, the number of linearly independent equations, this corresponds to the dimension of the subspace

$$\text{Row}(c_{S_1}) + \cdots + \text{Row}(c_{S_N}) \,, \tag{5}$$

where $\text{Row}(c_{S_i})$ is the subspace of $\mathbb{R}^{\mathcal{P}(S)}$ generated by the rows of the cleavage matrix of c_{S_i}. Theorem 1 can now be reformulated more precisely as follows.

Restatement of Theorem 1. For any choice of function $\phi : X \to \{1, \cdots, N\}$, the vectors $\rho_{U,S_{\phi(U)}}$ for $U \in X$ form a basis for the subspace in Eq.5, whose dimension is therefore $\#X$.

Proof: Given the vector $\rho_{U,S_i} \in \mathbb{R}^{\mathcal{P}(S)}$, recall that $\rho_{U,S_i}(V)$ denotes the component of the vector at the element $V \in \mathcal{P}(S)$.

It will be helpful to introduce the function $\theta : X \to \mathcal{P}(\{1, \cdots, N\})$ to pick out those cleavage subsets which contain a given peptide modform U, $\theta(U) = \{1 \leq j \leq N \,|\, U \in \mathcal{P}(S_j)\}$, so that $\phi(U) \in \theta(U)$, and the function $\delta : X \to \{1, \cdots, N\}$, which gives the number of those subsets, $\delta(U) = \#\theta(U)$, so that, for instance, $\theta(\emptyset) = \{1, \cdots, N\}$ and $\delta(\emptyset) = N$.

If $\rho_{U,S_i}(V) = 1$, then $V \cap S_i = U$, so that $U \subseteq V$. Hence, $\theta(U) \supseteq \theta(V)$. Conversely, if $\theta(V) \setminus \theta(U) \neq \emptyset$, then $\rho_{U,S_i}(V) = 0$. This provides the clue needed to show that the vectors $\rho_{U,S_{\phi(U)}}$ for $U \in X$ are linearly independent.

Consider any ordering of the subsets $U \in X$ which is non-decreasing in $\delta(U)$. In other words, if $\delta(U) < \delta(V)$, then U comes before V in the ordering, while if $\delta(U) = \delta(V)$, U and V may be in either order with respect to each

other. It is clear that such orderings exist and we can choose any one of them. Given such an ordering, consider the $\#X \times \#X$ submatrix of the cleavage matrix corresponding to the entries $\rho_{U,S_{\phi(U)}}(V)$ for $U, V \in X$. If $\delta(V) > \delta(U)$, then $\rho_{U,S_{\phi(U)}}(V) = 0$ because, if not, so that $\rho_{U,S_{\phi(U)}}(V) = 1$, then, by the clue above, $\theta(U) \supseteq \theta(V)$ and so $\delta(V) \leq \delta(U)$. Hence, the submatrix is block lower-triangular, where the blocks are demarcated by the distinct values of δ. Furthermore, if U and V are in the same block, so that $\delta(U) = \delta(V)$, then distinct subsets of $\{1, \cdots, N\}$ of the same size cannot be mutually contained in each other, so the only way in which $\theta(U) \supseteq \theta(V)$ is if $U = V$. Hence, if $\delta(U) = \delta(V)$, then $\rho_{U,S_{\phi(U)}}(V) = 1$ if, and only if, $U = V$, so that the diagonal blocks are the identity. It follows that the submatrix is lower-triangular with 1's on the main diagonal, so that the rows defined by $\rho_{U,S_{\phi(U)}}$ are necessarily linearly independent.

To show that $\rho_{U,S_{\phi(U)}}$ also span the subspace in Eq.5, it is sufficient to show that any other vector ρ_{U,S_j} with $U \in \mathcal{P}(S_j)$ is linearly dependent on the vectors $\rho_{U,S_{\phi(U)}}$. We show this by a "backwards" induction. Suppose, to the contrary, that ρ_{U,S_j} is a vector that is linearly independent of $\rho_{U,S_{\phi(U)}}$ and that this vector has been chosen so that $\delta(U)$ is minimal. This means that if ρ_{W,S_q} is any vector with $W \in \mathcal{P}(S_q)$ for which $\delta(W) < \delta(U)$, then ρ_{W,S_q} is linearly dependent on the vectors $\rho_{U,S_{\phi(U)}}$. Let $i = \phi(U)$. If $i = j$, then ρ_{U,S_j} is one of the vectors $\rho_{U,S_{\phi(U)}}$, which is a contradiction. Hence $i \neq j$. It follows from Eq.4 applied to $S_i \cap S_j \subseteq S_i$, and noting that $\mathcal{P}(S_i \setminus (S_i \cap S_j)) = \mathcal{P}(S_i \setminus S_j)$, that

$$\begin{aligned} \rho_{U,S_i \cap S_j} &= \textstyle\sum_{Q \in \mathcal{P}(S_i \setminus S_j)} \rho_{Q \cup U, S_i} \\ &= \rho_{U,S_i} + \left(\textstyle\sum_{\emptyset \neq Q \in \mathcal{P}(S_i \setminus S_j)} \rho_{Q \cup U, S_i} \right) . \end{aligned} \tag{6}$$

Choose any $\emptyset \neq Q \in \mathcal{P}(S_i \setminus S_j)$ occurring in the term in brackets on the right of Eq.6. Note that $Q \notin \mathcal{P}(S_j)$. Since $U \subseteq Q \cup U$, $\theta(Q \cup U) \subseteq \theta(U)$. However, $Q \cup U \notin \mathcal{P}(S_j)$, for otherwise $Q \in \mathcal{P}(S_j)$, and, since $U \in \mathcal{P}(S_j)$, it must be that $\theta(Q \cup U) \neq \theta(U)$. It follows that $\delta(Q \cup U) < \delta(U)$. Hence, by the minimality hypothesis, all the terms within the brackets on the right of Eq.6 are linearly dependent on $\rho_{U,S_{\phi(U)}}$. We can now apply Eq.4 to $S_i \cap S_j \subseteq S_j$, to get,

$$\rho_{U,S_i \cap S_j} = \rho_{U,S_j} + \left(\sum_{\emptyset \neq Q' \in \mathcal{P}(S_j \setminus S_i)} \rho_{Q' \cup U, S_j} \right) , \tag{7}$$

in which, by the same argument, the terms within brackets are linearly dependent on $\rho_{U,S_{\phi(U)}}$. Since the left-hand sides of Eqs.6 and 7 are the same, we can cancel the common term $\rho_{U,S_i \cap S_j}$, to get, recalling that $i = \phi(U)$,

$$\begin{aligned} \rho_{U,S_j} &= \rho_{U,S_{\phi(U)}} + \left(\textstyle\sum_{\emptyset \neq Q \in \mathcal{P}(S_i \setminus S_j)} \rho_{Q \cup U, S_i} \right) \\ &\quad - \left(\textstyle\sum_{\emptyset \neq Q' \in \mathcal{P}(S_j \setminus S_i)} \rho_{Q' \cup U, S_j} \right) . \end{aligned}$$

Here, all the terms on the right-hand side are linearly dependent on the vectors $\rho_{U,S_{\phi(U)}}$ for $U \in X$. Hence, so is ρ_{U,S_j}, contrary to assumption. Therefore, no such linearly independent vector can exist and the vectors $\rho_{U,S_{\phi(U)}}$ for $U \in X$ span the subspace in Eq.5. It follows that $\dim(\text{Row}(c_{S_1}) + \cdots + \text{Row}(c_{S_N})) = \#X$, as required. This completes the proof.

■

We note that the number of elements in the set X is given by the usual inclusion-exclusion formula for sets. Consequently, by Theorem 1, inclusion-exclusion also holds for the dimension of sums of row spaces in Eq.5, as originally conjectured.

Supporting references

[1] S. Hoşten and S. Sullivant. Gröbner bases and polyedral geometry of reducible and cyclic models. *J. Comb. Theory A*, 100:277–301, 2002.

[2] L. Pachter and B. Sturmfels, editors. *Algebraic Statistics for Computational Biology*. Cambridge University Press, Cambridge, UK, 2005.