Appendix S1. Finding the singular value decomposition through eigendecomposition.

The columns of U and V and the values of Σ can be found by forming the matrix product FF^T and F^TF and then finding the eigenvectors and eigenvalues of these two products. Depending on the order of multiplication of F with its transpose, two different matrices are obtained.

$$C = FF^T \tag{S1A}$$

$$D = F^TF \tag{S1B}$$

We refer to C as a "sequence identity matrix" and to D as a "residue pair count matrix" for reasons that will become clear below. C and D are symmetric with dimensions m by m and 20ℓ by 20ℓ respectively.

Each element in the C matrix gives the number of identities in a pair of aligned sequences. The diagonal elements, which result from the dot product of each MSA sequence vector $\vec{s}_i$ with itself, are equal to the length of sequence i (ℓ minus the number of gaps). The off-diagonal elements, which result from the dot product of pairs of sequences, are equal to the number of identical residues between sequences i and j.

Each element of the D matrix gives the number of sequences that contain a pair of residues at two positions. Each element d_{ii} on the main diagonal (resulting from the dot product of each residue vector $\vec{r}_i$ with itself) is a count of how many times a residue of a given type occurs at a given position. There are two types of off-diagonal elements in D. The off-diagonal elements that are contained within 20 by 20 block matrices running along the main diagonal of D all have values of zero because they count the number of impossible sequences where two different residue types (e.g., Ala and Leu) occur at the same position. Outside of these blocks, elements d_{ij} correspond to the number of sequences that have a residue of one type at position i, and a residue of another type at position j. If these two residue types occur frequently in positions i and j in the same sequences, d_{ij} will be a large number.

The singular value decomposition of the F matrix is closely related to the eigendecomposition of the C and D matrices (which are the matrices on which principal component analysis is performed[10]). Each of these symmetric matrices can be written in a diagonalized form:

$$C = X \wedge_C X^{-1} \tag{S2A}$$

$$D = Y \wedge_D Y^{-1} \tag{S2B}$$

[10] Though typically PCA is performed on a data matrix in which the columns of matrix F are shifted to have a mean of zero. When this "centering" is applied, the first component of the PCA (and SVD) no longer reflects the conservation among sequences.

X and Y are square matrices of dimensions m by m and 20ℓ by 20ℓ, respectively. Their columns ($\vec{x}^{(i)}$ and $\vec{y}^{(i)}$) are the eigenvectors of C and D, which we will refer to as sequence and residue eigenvectors, respectively. $\wedge_C$ and $\wedge_D$ are diagonal matrices with the same dimensions as C and D, respectively, with eigenvalues λ_i on their diagonals. Equations S2A and S2B are rearranged versions of the eigenvalue equations

$$C\vec{x}^{(i)} = \lambda_i\vec{x}^{(i)} \quad \text{(S3A)}$$

$$D\vec{y}^{(i)} = \lambda_i\vec{y}^{(i)} \quad \text{(S3B)}$$

which state that for each eigenvector/eigenvalue pair, multiplying the sequence and residue eigenvectors $\vec{x}^{(i)}$ and $\vec{y}^{(i)}$ by matrices C and D does not change the direction of $\vec{x}^{(i)}$ and $\vec{y}^{(i)}$, but simply scales them by their eigenvalues[11]. As in SVD, the eigenvectors $\vec{x}^{(i)}$ and $\vec{y}^{(i)}$ are normalized to have unit length, and the eigenvalues are adjusted accordingly.

Because the C and D matrices are symmetric (i.e., they are equal to their transposes), their eigenvectors are orthogonal. Thus, X and Y are orthogonal matrices. Since the inverse of an orthogonal matrix is equal to its transpose (e.g., $X^T = X^{-1}$), Equations S2A and S2B can be written as

$$C = X\wedge_C X^T \quad \text{(S4A)}$$

$$D = Y\wedge_D Y^T \quad \text{(S4B)}$$

To relate the eigendecomposition of C and D and the SVD of F, we can substitute the singular value decomposition of F (Equation 1) into Equations S1A and S1B. For example,

$$C = FF^T \quad \text{(S5)}$$

[11] Note that Equations S3A and S3B seem like more than just simple rearrangements of the Equations S2A and S2B because the former are matrix equations and the latter are vector equations, Equations S3A and S3B can be obtained by right-multiplying Equations S2A and S2B by their eigenvalue matrices and evaluating a column at a time.

$$= U\Sigma V^T (U\Sigma V^T)^T$$

$$= U\Sigma V^T (U\Sigma V^T)^T$$

Because $V^T = V^{-1}$ (since V is orthogonal),

$$C = U\Sigma V^{-1}\Sigma^T \mathrm{U}^T \quad \text{(S6)}$$

$$= U\Sigma\Sigma^T \mathrm{U}^T$$

Comparing Equations 8 and 3A gives

$$U\Sigma\Sigma^T \mathrm{U}^T = X \wedge_C X^{-1} \quad \text{(S7)}$$

An analogous relationship involving V and Y can be obtained from Equations 1 and S1B using a derivation analogous to Equations S5-S7:

$$V\Sigma^T\Sigma \mathrm{V}^T = Y \wedge_D Y^{-1} \quad \text{(S8)}$$

Equations S7 and S8 imply that

$$U = X \quad \text{(S9A)}$$

$$V = Y \quad \text{(S9B)}$$

that is, the SVD matrices U and V of F are equal to the eigenvector matrices X and Y of FF^T and F^TF. Additionally, Equations S7 and S8 imply that

$$\Sigma\Sigma^T = \wedge_C \quad \text{(S10A)}$$

$$\Sigma^T\Sigma = \wedge_D \quad \text{(S10B)}$$

Multiplying out the products on the left-hand sides of Equations S10A and S10B gives the result that $\sigma_i^2 = \lambda_i$. This relationship requires the same ordering of eigenvectors and values as described above for SVD (i.e., decreasing values of λ_i along the diagonal of $\wedge$). Equations S10A and S10B require that the eigenvalues associated with the sequence identity matrix C are the same as those associated with residue pair count matrix D, even though $\wedge_C$ and $\wedge_D$ differ in dimension. The connection between eigendecomposition and SVD is summarized in Figure S2.

In addition to providing a means to obtain a singular value decomposition of F, the eigenvalue approach forms the basis of PCA.

Appendix S2. Visualizing MSA sequences in SVD space.

The coordinates of sequences in SVD space can be obtained by separating the U and V matrices in the SVD Equation 1. Right-multiplying Equation 1 by the V matrix gives[12]

$$FV = U\Sigma V^T V = U\Sigma V^{-1} = U\Sigma \quad \text{(S11)}$$

The left side of Equation S11 is a matrix in which each element is the dot product of a sequence vector $\vec{s}^{(i)}$ from F and a residue eigenvector $\vec{v}^{(k)}$ (Figure S3). The right side of Equation S11 is a matrix in which the columns are the sequence eigenvectors weighted by their corresponding singular values (e.g., the k^{th} column is $\sigma_k \vec{u}^{(k)}$). The equality between the FV and $U\Sigma$ product matrices requires that each of their corresponding elements (for example, the red and blue boxed elements in Figure S3) be equal, giving a set of m equations of the form of Equation 2 in the main text:

$$\vec{s}_i \cdot \vec{v}^{(k)} = \sigma_k u_i^{(k)} \quad \text{(S12)}$$

Appendix S3. Scripts for SVD analysis of multiple sequence alignments.

To run SVD on a family of sequences, an MSA must be prepared. Although the guidelines for MSA preparation provided above have performed well in revealing functionally relevant clustering, they may not be optimal for all protein families. It may be wise to try a range of identity filters, length cut-offs, and gap filters. Once an MSA is prepared, SVD can be performed using a collection of Python scripts that we have made available in GitHub

[12] Again, we use the equality $V^T = V^{-1}$ which results from the orthogonality of V.

(https://github.com/barricklab-at-jhu/SVD-of-MSAs). To help streamline the process, we have combined various calculations into a single Jupyter Notebook. The major steps in the process are show in in Figure S11, along with the various output files.