The matrix **M** holding the canonical residue occupancy for many structures is composed of elements m_{ij}. Each row *i* in **M** specifies a structure, and each column *j* specifies a canonical residue, where $m_{ij} = 1$ indicates the canonical residue *j* is present in structure *i* and a $m_{ij} = 0$ indicates it is not. The geometric length r_i of vector M_i is

$$r_i = \sqrt{\sum_j m_{ij}^2}$$

Since $m_{ij} \in \{0,1\}$, it is clear that

$$\forall_{ij}, m_{ij}^2 = m_{ij}$$

Thus,

$$r_i = \sqrt{\sum_j m_{ij}}$$

Each of the elements of row vector m_i is divided by r_i to produce a normalized matrix **N**, with elements n_{ij}

$$n_{ij} = \frac{m_{ij}}{r_i}$$

A square matrix **C** describing the correlation (inner product) of each row n_a of **N** with another row n_b of **N** is readily calculated as the product of **N** with its transpose $\mathbf{N}^T$. That is,

$$\mathbf{C} = \mathbf{N} \cdot \mathbf{N}^T \qquad c_{ab} = c_{ba} = n_a \cdot n_b$$

Examination of **C** reveals several interesting properties. For example, if two vectors (m_a and m_b) differ by the presence/absence of a single canonical residue *j*, that is, $m_{aj} = 0$ while $m_{bj} = 1$, and m_a and m_b share all other elements in common, the element of row c_a (other than c_{aa}) that has the greatest value is c_{ab}. That is, finding the element(s) of row c_a with the greatest value corresponds to finding structures that differ from structure *a* by a single residue (if such structures exist in **M**). When multiple elements of row c_a are "tied" with the highest value, more than one structure embodied in **M** differs from *a* by a single residue (certainly a possibility for real structures).

To be more precise, the values c_{ab} in row c_a of **C** can be calculated explicitly given the number of residues *x* in row *a*, the number of residues *y* in structure *b* (being correlated to structure *a*), and the number of residues *z* shared by structures *a* and *b*.

$$c_{ab} = \frac{z}{\sqrt{xy}}$$

This equation allows structures that have a single residue more than the target residue from those having a single residue less than the target residue to be distingushed. For example, for a target residue with 5 residues, structures with 6 residues (5 shared with the target) have a correlation value of 0.91287; structures with 4 residues (all 4 shared with the target) have a correlation value of 0.89443; structures with 3 residues (all 3 shared with the target) have a correlation value of 0.77460; structures with 7 residues (5 shared with the target) have a correlation value of 0.84515; structures with 5 residues (4 shared with the target) have a correlation value of 0.8000. Thus,

structures having one more or one less residue than the target (everything else being the same) have the two highest correlation values, and these values can be predicted to ensure that the highest values seen corresond to structures that have exactly one more or one less residues than the target. If such (n +1 and n - 1) structures do not exist, the precise relationship between the target structure and the structure with the highest correlation to the target can be immediately identified by evaluating the correlation value.

This is a very rapid way to identify sequential structures along a biosynthetic or degradation pathway. The matrices can all be pre-calculated and rapidly evaluated for a particular structure by simple row scanning of matrix **C**. For example, once **C** is calculated, one can rapidly determine, "for structure *a* having *x* residues, which structures (in the set) consisting of *y* residues share exactly *z* residues with structure *a*?"