Table S2. Assessment of the quality of the clusters of twin CX_9C proteins identified in this study.

The quality of the clusters of twin CX_9C proteins that were built by the procedure described in the Methods were assessed by two types of validation techniques: (i) Cophenetic Correlation (CC), and (ii) Silhouette Width (SW), as described in [1]. Both of these techniques are defined as internal quality measures, in that they estimate how well the information contained in the data (or the "natural" cluster structure of the data) is preserved in a given partitioning.

Cophenetic correlation

In CC, a coefficient is calculated that measures the correlation between two matrices, one (called the cophenetic matrix) representing the partitioning of the elements into clusters, and the other (called the distance matrix) representing the dissimilarity between each pair of elements. The partitioning of the twin CX_9C proteins into clusters was represented by a cophenetic matrix C such that $C(i, j) = 0$ if the two elements (i.e., the two proteins) i and j belong to the same cluster, and $C(i, j) = 1$ otherwise. The dissimilarity between each pair of twin CX_9C proteins was represented by a pseudo-distance matrix D such that $D(i, j) = 0$ if there is a BLAST match with an E-value $< 10^{-3}$ between the two elements (i.e., the two proteins) i and j, and $D(i, j) = 1$ otherwise. The correlation coefficient between the two matrices was calculated using Hubert's modified normalized statistic as in [2], and was found to be **0.84**. As this index ranges from -1 to 1, the value of 0.84 indicates a good correlation.

Silhouette Width

In SW, an index (called the Silhouette value) is calculated for each element i as $S(i) = (b_i - a_i) / \max(b_i, a_i)$, where a_i represents the average dissimilarity between i and all the other elements in the same cluster, and b_i represents the average dissimilarity between i and all the elements in the

closest other cluster, which is defined as the one yielding the minimal b_i[3]. It thus reflects the confidence in the cluster assignment of *i*. The average Silhouette value over all elements measures the global quality of the partitioning. Silhouette values were calculated for twin CX_9C proteins using the pseudo-distance matrix *D* defined above to represent their dissimilarity, and the average Silhouette value was found to be **0.73**. Again, as this index also ranges from -1 to 1, the value of 0.73 indicates a good clustering quality.

References

1 Handl, J.; Knowles, J.; Kell, D. B. *Bioinformatics.* 2005, **21**, 3201-3212.

2 Pal, N. R.; Biswas, J. *Pattern Recogn.* 1997, **30**, 847-857.

3 Rousseeuw, P. J. *J.Comput.Appl.Math.* 1987, **20**, 53-65.