# Can protein folds be automatically and objectively defined? An analysis based on transitivity

Alberto Pascual-García, Enrique García de Bustos, David Abia, Angel Ramírez Ortiz and Ugo Bastolla*
Centro de Biología Molecular 'Severo Ochoa', (CSIC-UAM), Cantoblanco, 28049 Madrid, Spain

**Equivalence relationship; Transitivity**

In mathematical terms, a protein fold is an equivalence class of protein structures. The question that we investigate here is to which extent the fold can be defined only based on a quantitative measure of structural similarity and a threshold. Mathematically, an equivalence relationship based on a similarity measure is automatically endowed with the reflexive property (any object $a$ is similar to itself) and with the symmetric property (if $a$ is similar to $b$, then $b$ is similar to $a$). The *transitive* property is more problematic. For transitivity to hold, if $a$ is similar to $b$ and $b$ is similar to $c$, then $a$ must also be similar to $c$. This property is essential for a similarity measure to give rise to an objective equivalence relationship.

**Uniparental evolution satisfies transitivity**

There is a simple reason why one should expect that protein structural similarity fulfils the transitive property. Most genes coding for related proteins are related through gene duplication[1], and they can be represented as the leaves of a phylogenetic tree. The distance across such a tree, i.e. the time spent since the divergence of two lineages, is ultrametric[2], and therefore it is naturally endowed with the transitive property. In fact, if the lineages leading to $a$ and $b$ and those leading to $b$ and $c$ both splitted less than $t$ million years ago, the same is true also for the lineages leading to $a$ and $c$, for whatever value of the dissimilarity threshold $t$. Therefore, a phylogenetic tree naturally implies a hierarchical classification for every similarity threshold. If we can find a structural dissimilarity measure between pairs of proteins linearly correlated with their time of divergence, as it happens for suitable sequence distances, the transitivity property will approximately hold for such a distance.

**Fragment assembly violates transitivity;**

Nevertheless, single gene duplication is not the only possible mechanism for the evolution of protein domains. It is well known that complex proteins are formed from a combination of individual domains with independent evolutionary history. For this reason, the domain and not the complete protein is the basic unit for protein classification. However, there is increasing evidence that protein domains are not always fundamental units, but they may be formed by smaller fragments below the domain level[4,5], and it has been observed that many structurally unrelated proteins share common substructures[6,7,5].

If two domains $a$ and $b$ are similar because of a partial substructure $A$, while $b$ and $c$ are similar because of a different partial substructure $C$, then $a$ and $c$ are not similar and transitivity is violated. Several authors refer to this kind of situation stating that protein space is continuous, since one can connect two different structures $a$ and $c$ through two small steps passing through $b$. In this case, there is no classification simulataneously compatible with all the pairwise similarity relationships. Borrowing a term from statistical physics, we can say that the classification problem is *frustrated*[3] if transitivity is violated. We can expect that, if this situation is common for many triplets, there is an exponentially large number of substantially different classifications that are almost optimal, in the sense that they violate a small and similar number of pairwise relationships, whereas, if the transitive property approximately holds, a well-defined unique globally optimal classification given by the ultrametric tree exists, and all sub-optimal classifications are very similar to the optimal one. We expect that the transitive property depends on the threshold used for defining structural similarity: When this threshold is very large, only domains that share most of their structure are regarded as similar, and we expect that transitivity approximately holds, and the structure space is made of discrete clusters. In contrast, for less stringent thresholds, two domains may be regarded as similar due to partial substructures. We propose here a measure for assessing violations of the transitive property as a hiererchical clustering algorithm joins proteins into clusters. In this way, we aim at detecting a cross-over point beyond which the hierarchical clustering is not justified anymore. We propose to identify the clusters defined up to this point as intrinsic equivalence classes of protein domains. At smaller structural similarity, the protein structure space should be rather regarded as continuous, and the similarity relationships between clusters should be represented as a network rather than as a tree.

[*] ubastolla@cbm.uam.es

[1] S. Ohno, Evolution by gene duplication (Springer, 1970).

[2] R. Rammal, G. Toulouse and M.A. Virasoro Ultrametricity for physicists Rev. Mod. Phys. 58, 765 - 788 (1986).

[3] G. Toulouse, Theory of the frustration effect in spin glasses: I *Comm. Phys.* **2**, 115-119 (1977).

[4] Tsai, Maizel and Nussinov, Anatomy of protein structures: Visualizing how a one-dimensional protein chain folds into a three dimensional shape, PNAS 97: 12038-12043 (2000).

[5] JD Szustakowski, S Kasif, Z Weng (2005) Less is more: towards an optimal universal description of protein folds. *Bioinformatics* **S2**: ii66-ii71.

[6] Efimov AV. Structural trees for protein superfamilies. PROTEINS: Structure, Function and Genetics. 1997;28:241-60.

[7] A. Harrison, F. Pearl, R. Mott, J. Thornton, C. Orengo,Quantifying the similarity within fold space, J. Mol. Biol. 323:909-26 (2002).