PROTEINS: Structure, Function, and Bioinformatics 55:646–655 (2004)

# On the Functional Significance of Electron Density Protein Structure Alignments

**Pere Constans**
*Department of Chemistry, Rice University, Houston, Texas*

***ABSTRACT*** **Electron density protein alignments are analyzed in terms of their underlying similarity measure, the density overlap. These alignments are conceptually unrelated to biochemical structural elements and, therefore, are appropriate in structure-only similarity studies. The analysis is focused on the low sequence similarity subset of protein domains. A remarkable association is found between simple, density overlap measures and the expert designed Structural Classification of Proteins (SCOP) for which functional and evolutive analogies prevail. The association found validates the functional significance of electron density alignments. Proteins 2004;55:646–655.**


**Key words: protein alignment; molecular similarity; electron density superposition; twilight zone**

## INTRODUCTION

Similarity is an especially relevant concept in protein science. Protein classification, structure, and function prediction, or homology identification, involve similarity measurement and analysis.[1–7] Similarity concretizes upon comparing and discriminating among sets of *relevant structure elements*. Albeit measurable, similarity is elusive, as is the determination of particular structure elements. Furthermore, the intricacies of the comparison and discrimination prevent single and unique measurements of similarity.[8–11]

Appropriate choices are dictated by their significance on a particular setting. Approaches relying on stripped codifications of structural characteristics[12–15] are adequate, if not imperative, to retrieve information from large databases. Once the codes or descriptors are obtained, the similarity calculation requires only a few algebraic operations. Reliable large-scale classifications were reported in this way.[16,17] On the other hand, the more demanding methods that measure similarity directly, as the proximity among equivalent elements of aligned structures, permit a detailed analysis on preselected sets. Detecting common patterns, shape chirality, or the conservation of unusual features is readily feasible through visual inspection of the resulting superpositions. These similarity alignments provide an insight into relationships between amino acid sequence and structure. Such relations are thus relevant in protein evolution and homology detection, as well as in the understanding of the physical interactions that govern the protein folding.

The structural alignment inherits the difficulties intrinsic to any similarity definition, in addition to its own computational and conceptual ones. A paradigmatic example is the root-mean-square deviation (RMSD) alignment. Similarity is derived from the RMSD between equivalent $C^{\alpha}$ coordinates. The alignment is produced by minimizing the RMSD with respect to the mutual positioning and orientation of the two structures through a simple and elegant algorithm.[18–19] It inherits from its underlying similarity definition the difficulties in establishing *equivalent* $C^{\alpha}$ correspondences, which, in principle, convert protein superposition into a combinatorial problem. The conceptual difficulties arise in connection with the intent of the alignment. If a structural alignment is intended to assist in elucidating the sequence alignment, one might question whether or not the selected $C^{\alpha}$ equivalences introduce a bias in the result. From a more general perspective the question may be: How significant is an alignment whose underlying similarity measure scarcely correlates with an expected protein classification?

In a theory on Quantum Molecular Similarity (QMS), the relevant structure elements are fuzzy electron probability density distributions.[20,21] In fact, the electron density is nowadays regarded as the fundamental observable of the molecular universe.[22,23] The measure of similarity is derived from the extent of attainable overlap between two compared molecular electron densities. Complete, global search strategies for maximizing the overlap between two density functions are known.[24] This leads to systematic and theoretically justified QMS alignments. Moreover, fitted electron densities[25] and promolecular models[26,27] significantly reduced the QMS computational cost. Recently, improved algorithms extended these established similarity techniques to macromolecules.[28] Present research on linear scaling ab initio electron density computations[29,30] may indeed provide in the future a more precise insight into macromolecular representation and similarity.

The QMS alignments, though still computationally intensive, are conceptually simple and precise. Their underlying similarity measure is the overlap between two molecular electron densities. Different from heuristic techniques, QMS introduces a systematic and encompassing approach

*Correspondence to: Pere Constans, Department of Chemistry, Rice University, Houston, TX. E-mail: constans@ruf.rice.edu

Received 16 July 2003; Accepted 18 November 2003

Published online 5 March 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20059

toward purely structural similarity measures and alignments. Therefore, they clearly posses a potential value in structure and function studies. This work is intended to assess the significance of QMS alignments or, in other words, to analyze the extent to which QMS measures are related to the biological function of proteins. Apart from its theoretical interest, such assessment introduces a practical and novel approach to protein comparison. This approach is suited for comparisons and structural alignments of proteins within the twilight zone in sequence similarity, thus being a complementary and alternative option in the situations where usual procedures present the most difficulties and discrepancies.

This article is organized as follows: The Theory and Methods section briefly describes QMS and its extension to macromolecules. Next, the Results and Discussion section presents and analyzes the data, a QMS matrix among protein domains with less than a 40% sequence identity. This focuses the analysis on a subset of potential applicability and keeps calculations at a reasonable cost. These data are contrasted with the Structural Classification of Proteins (SCOP),[31] an exceptional and expertly designed classification with respect to functional and evolutive criteria. The assessment of the QMS functional significance is performed in a deterministic and nonparametric fashion, without any kind of calibration or tuning parameters. Two procedures are considered: the one described by Levitt and Gerstein,[32] and the measuring of association for the cross-QMS single-linkage versus SCOP classification. Technical details of both procedures are included as Appendices A and B at the end of the article.

## THEORY AND METHODS

### Quantum Molecular Similarity Measures

The measure of similarity between two molecular structures $i$ and $j$ is defined as the maximum of their similarity function. The similarity function, in turn, is the projection or overlap of the two electron density functions $\rho_i(\mathbf{r})$ and $\rho_j(\mathbf{r})$, being

$$z_{ij}(\Omega) = \int \rho_i(\mathbf{r}; \mathbf{R}_i)\rho_j[\mathbf{r}; \mathbf{R}_j(\Omega)]d\mathbf{r}. \quad (1)$$

Molecular densities $\rho_i(\mathbf{r})$ and $\rho_j(\mathbf{r})$ implicitly depend on their sets of atomic coordinates $\mathbf{R}_i$ and $\mathbf{R}_j$, related to a given coordinate system. The function $z_{ij}(\Omega)$ is maximized with respect to the set of mutual displacements and orienting angles $\Omega$. Once normalized, it gives the index of molecular similarity

$$C_{ij} = z_{ij}(z_{ii}z_{jj})^{-1/2}. \quad (2)$$

The Carbó index $C_{ij}$, as defined by Carbó et al. early in the 1980s, takes a value of 1 for identical structures and tends to zero as the difference between the two structures increments.[33]

Maximization with respect to the relative displacement and orientation of two molecules leads to a theoretically justified set of alignments. The set of maxima maps onto a corresponding set of common patterns or motifs between the compared structures. Interestingly, a systematic, global search algorithm, polynomial in time, has been established.[24] Complexity varies depending on the shape of the molecular density. The complexity is maximal and $\mathbb{O}(N_i^3N_j^3)$, with $N_i$ and $N_j$ being the number of atoms in the molecules $i$ and $j$, respectively, if $\rho_i$ and $\rho_j$ were peaked, atom-centered Dirac $\delta$ densities. On the other hand, function $z_{ij}(\Omega)$ would be constant for completely smoothed densities. The nontrivial, peak limit global search algorithm permitted devising practical and reliable searches scaling as $\mathbb{O}(N_iN_j)$, or in short $\mathbb{O}(N^2)$, appropriate for small molecules and ab initio electron densities.[24]

The $\mathbb{O}(N^2)$ scaling cost refers to similarity function evaluations. Practical evaluation of the QMS function benefits from the use of simplified density models. The Atomic Shell Approximation (ASA) densities present the general form

$$\rho_{ASA}(\mathbf{r}) = \sum_a \sum_{i\in a} n_i c_i e^{-\xi_i(\mathbf{R}_a-\mathbf{r})^2}, \quad (3)$$

with $\mathbf{R}_a$ being the atomic coordinates, $n_i$ the shell occupancies, and $c_i$ the shell normalizations.[26] Expensive 4-center overlap integrals appearing in Eq. (1) are substituted by isotropic, 2-center overlaps, thus reducing the cost from $\mathbb{O}(N^4)$ to $\mathbb{O}(N^2)$. The impact of this isotropic, pseudoatomic constraint on the resulting overlap QMS measure is reduced. The asphericity of the *atoms* upon bond formation and the interatomic charge transfer scarcely contributes to the overlap integral value. This fact permits modeling of the electron densities simply as *promolecular* ASA expansions. The accuracy of these approaches when compared to fully ab initio densities has been analyzed.[26,27]

### Macromolecular Similarity

Crystallographic macromolecular structures are fuzzy distributions around a set of preferential nuclei positions $\bar{\mathbf{R}}_n$ The usual *one-particle potential* model[34] approximates the nuclear distribution to

$$p_n(\mathbf{R}_n) = p_1(\mathbf{u}_1)p_2(\mathbf{u}_2)\cdots p_N(\mathbf{u}_N), \quad (4)$$

where, for each nucleus $a$, the displacement $\mathbf{u}_a$ is the difference $\bar{\mathbf{R}}_a - \mathbf{R}_a$. The resulting dynamic electron density $\tilde{\rho}(\mathbf{r})$ follows the convolution of the electronic and nuclear distributions,[35]

$$\tilde{\rho}(\mathbf{r}) = \rho(\mathbf{r}; \mathbf{R}_n) * p_n(\mathbf{R}_n). \quad (5)$$

Within the ASA, dynamic densities read

$$\tilde{\rho}_{ASA} = \sum_a \tilde{\rho}_a(\mathbf{r}), \quad (6)$$

with the pseudoatom contributions $\tilde{\rho}_a$ given by the convolution

$$\tilde{\rho}_a(\mathbf{r}) = \int \rho_a(\mathbf{R}_a - \mathbf{r})p_a(\bar{\mathbf{R}}_a - \mathbf{R}_a)d\mathbf{R}_a. \quad (7)$$

Upon integration, Eq. (7) is formally analogous to the previous Eq. (3) provided that nuclei move independently

as isotropic harmonic oscillators. Shell occupations remain invariant, yet the shapes of the shells appear smeared. This smearing and the collapse of the shell structure provides an effective simplification for the macromolecular electron density similarity measurement.[28] The overlap integrals are well approximated by using a minimalist, single-shell dynamic electron density model that substantially reduces the computational prefactor. The overlap QMS function becomes a double-sum of isotropic, atom–atom contributions

$$z_{ij}(\Omega) = \sum_{a\in i}\sum_{b\in j} n_a n_b \left(\frac{\mu_{ab}}{\pi}\right)^{3/2} e^{-\mu_{ab} r_{ab}^2(\Omega)}, \qquad (8)$$

with $\mu_{ab}$ being $\zeta_a\zeta_b/(\zeta_a + \zeta_b)$. Quantities $\zeta_a$ and $\zeta_b$ are the single-shell exponents that depend on the mean-square displacement of the nuclei $a$ and $b$.

The number of non-negligible integrals scales only linearly on large systems due to the exponential decay of the density overlaps. The overall cost of the macromolecular alignment is then shifted to the $\mathcal{O}(N^2)$ evaluation of the interatomic distances $r_{ab}(\Omega)$. A proposed near-neighbor list algorithm avoids unnecessary distance evaluations and permits the computation of the similarity function as a true linear scaling procedure.[28]

## RESULTS AND DISCUSSION

The assessment of the QMS scores as meaningful measures is based upon an analysis on a set of SCOP domains.[31] The selected domain structures are extracted from the Astral database,[3,36] version 1.59. Only domains within the classes α, β, α/β, and α + β with less than a 40% sequence identity are considered. Model entries and $C^\alpha$-only structures are excluded. Also excluded, are the families with less than 10 resulting entries. This gives a total of 781 domains classified into 50 different SCOP families. See Table I.

All the ATOM entries in the Protein Data Bank (PDB)[37,38] coordinate files are considered. Promolecular electron densities are constructed through atomic ASA densities, and all atomic mean displacements are taken equal to 30 Å$^2$. Similarities are evaluated using the program LSim.[39] Similarity is maximized as described.[28] The statistical analysis is performed using ad hoc software implementing the algorithms described in Appendices A and B. All calculations are performed on a PC/Linux platform equipped with two PIV/1600MHz CPUs.

### Statistical Significance

The first procedure considered for assessing the statistical significance of protein similarity scores is described in detail.[32] The $\frac{1}{2}n(n-1)$ nonredundant pair similarity values are organized according to SCOP, which is taken as the reference classification. Each pair $C_{ij}$ is labeled as *true-positive* or *true-negative,* whether structures $i$ and $j$ belong to the same or to different families, respectively, in the reference classification. The probability density functions (pdf), $f(C)$, and the cumulative distributions (cdf), $F(C)$, are then inferred for the true-positive and true-

**TABLE I. The Data Set Specifying the Detailed Number of Domains $n_d$ for Each of the Selected SCOP Families**

| SCOP ID | $n_d$ | Family |
|---|---|---|
| a.1.1.2 | 19 | Globins |
| a.26.1.1 | 10 | Long-chain cytokines |
| a.3.1.1 | 14 | Monodomain cytochrome c |
| a.45.1.1 | 15 | Glutathione s-transferases, C-terminal domain |
| a.74.1.1 | 10 | Cyclin |
| b.1.1.1 | 28 | V set domains |
| b.1.1.2 | 23 | C1 set domains |
| b.1.1.4 | 37 | I set domains |
| b.1.1.5 | 28 | E set domains |
| b.1.2.1 | 38 | Fibronectin type III |
| b.10.1.2 | 16 | Plant virus proteins |
| b.10.1.4 | 15 | Animal virus proteins |
| b.34.2.1 | 15 | SH3-domain |
| b.40.4.3 | 11 | Single-strand DNA-binding domain, SSB |
| b.40.4.5 | 11 | Cold shock DNA-binding domain-like |
| b.47.1.2 | 20 | Eukaryotic proteases |
| b.6.1.1 | 10 | Plastocyanin/azurin-like |
| b.6.1.3 | 17 | Multidomain cupredoxins |
| b.60.1.1 | 12 | Retinol binding protein-like |
| b.71.1.1 | 14 | α-Amylases, C-terminal beta-sheet domain |
| c.1.4.1 | 10 | FMN-linked oxidoreductases |
| c.1.8.1 | 14 | α-Amylases, N-terminal domain |
| c.1.8.3 | 16 | β-Glycanases |
| c.2.1.2 | 29 | Tyrosine-dependent oxidoreductases |
| c.2.1.3 | 12 | Glyceraldehyde-3-phosphate dehydrogenase-like, N-terminal domain |
| c.2.1.5 | 10 | Lactate & malate dehydrogenases, N-terminal domain |
| c.23.1.1 | 10 | CheY-related |
| c.26.1.1 | 11 | Class I aminoacyl-tRNA synthetases (RS), catalytic domain |
| c.3.1.5 | 23 | FAD/NAD-linked reductases, N-terminal and central domains |
| c.36.1.1 | 10 | Pyruvate oxidase and decarboxylase |
| c.37.1.13 | 22 | Extended AAA-ATPase domain |
| c.37.1.1 | 16 | Nucleotide and nucleoside kinases |
| c.37.1.8 | 16 | G proteins |
| c.47.1.5 | 12 | Glutathione s-transferases, N-terminal domain |
| c.61.1.1 | 12 | Phosphoribosyltransferases (PRTases) |
| c.67.1.3 | 10 | Cystathionine synthase-like |
| c.67.1.4 | 11 | ω-Amino acid: pyruvate aminotransferase-like |
| c.93.1.1 | 13 | L-arabinose binding protein-like |
| c.94.1.1 | 21 | Phosphate binding protein-like |
| c.95.1.1 | 11 | Thiolase-related |
| d.104.1.1 | 13 | Class II aminoacyl-tRNA synthetase (aaRS)-like, catalytic domain |
| d.131.1.2 | 10 | DNA polymerase processivity factor |
| d.144.1.1 | 14 | Serine/threonin kinases |
| d.153.1.4 | 17 | Proteasome subunits |
| d.162.1.1 | 11 | Lactate & malate dehydrogenases, C-terminal domain |
| d.169.1.1 | 15 | C-type lectin domain |
| d.185.1.1 | 11 | MPP-like |
| d.19.1.1 | 13 | MHC antigen-recognition domain |
| d.58.7.1 | 12 | Canonical RBD |
| d.93.1.1 | 13 | SH2 domain |
| Total | 781 | |

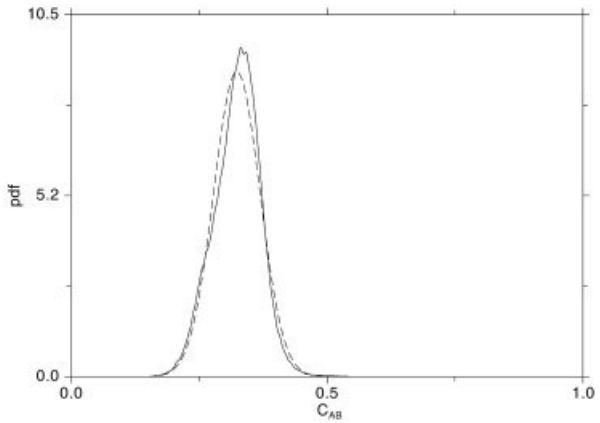


Fig. 1. True-negative probability density function (pdf) of similarity indices. The solid line depicts the nonparametric kernel estimate, and the dashed line the adjusted beta distribution.

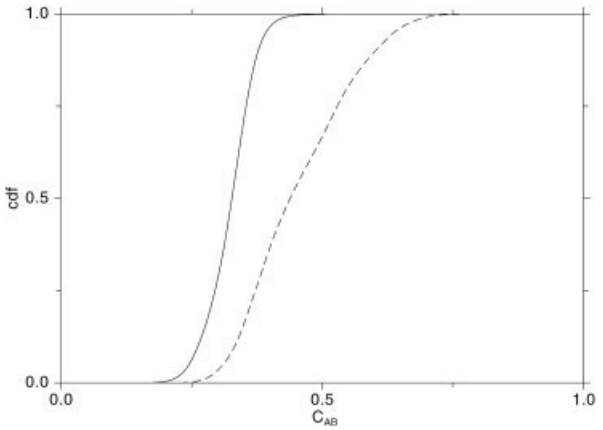


Fig. 2. True-negative, in solid line, and true-positive, in dashed line, for the cumulative distribution function (cdf) of similarity indices.

**TABLE II. Statistical Significance**

| $\mu_{tn}$ | $\sigma_{tn}$ | $P_{tn}$ | $c_k$ | $1 - F_{tp}(c_k)$ | $c_\beta$ | $1 - F_{tp}(c_\beta)$ |
|---|---|---|---|---|---|---|
| 0.324 | 0.045 | 0.95 | 0.392 | 0.67 | 0.400 | 0.63 |
| | | 0.99 | 0.426 | 0.54 | 0.433 | 0.52 |

Quantities $\mu_{tn}$ and $\sigma_{tn}$ are the true-negative mean and standard deviation, $P_{tn}$ is the probability threshold or $P$-value, $c_k$ and $c_\beta$ are the respective similarity values, as obtained using the kernel and β-estimates, and $1 - F_{tp}(c)$ is the coverage for true positives at these similarity values.

negative subsets of $C_{ij}$ values. The knowledge of the true-negative and true-positive score distributions permits defining two measures of statistical significance, the *P-value* and the *coverage*.

The $P$-value $P(C_{ij} \geq c)$ is the probability that the alignment of any two structures $i$ and $j$ from different families will give a similarity measurement $C_{ij}$ greater than or equal to $c$. The coverage indicates the extent of true-positives found at a given $P$-value and $c$. It measures, in essence, the overlap between $f_{tn}(C)$ and $f_{tp}(C)$. See Appendix A for the details on density estimation and definitions of significance.

The nonredundant $\frac{1}{2}(781 \times 780)$ pairs of similarities $C_{ij}$ are accordingly divided into true negatives and true positives. The true-negative density $f_{tn}(C)$ is plotted in Figure 1. The pdf is calculated by a nonparametric kernel estimate and, in addition, adjusted to a beta distribution, as described in the Appendix A. Table II reports the sample driven means and standard deviation used to infer the pdf. Table II also reports the similarity $c$ for the typical $P$-values of 95% and 99%, given by the nonparametric and beta pdf's, and the respective values of the coverage. A 5% of true-negative alignments will have a similarity $C_{ij} > 0.40$, and a 1% a similarity $C_{ij} > 0.43$. The coverage at these $P$-values (i.e., the portion of true-positive alignments with a similarity greater than 0.40 and 0.43), is, respectively, above 60% and 50%. The plots of the nonparametric cumulative distributions $f_{tn}(C)$ and $f_{tp}(C)$ are presented in Figure 2.

The quality of these results is in accordance with other structure-only analysis of SCOP classification,[32,40,41] though an exact comparison would be infeasible due to the very nature of the procedural variations. It is worth noting, however, that this study does not rely on tuning parameters, data calibration, or domain-size preclassifications.

## Clustering

An additional insight into the assessment of QMS significance is gained by examining the proximity relationships among the set of structures. A simple and intuitive clustering is the single linkage procedure.[42] Intuitively, a single-linkage cluster is the subset or portion of elements that can be visited through steps shorter than a threshold $c$. Single linkage is deterministic, hierarchical, and conservative in the sense that it might not detect all clusters but will detect the modes separated by a sufficiently deep valley in similarity space.[43]

This analysis of proximities is summarized in Table III, while the procedural details are described in the Appendix B. The single-linkage clustering is performed for all the relevant similarity values $c$ within 0 and 1. This gives a hierarchical classification or dendrogram. Interestingly, dendrograms impose a partial ordering in the set of objects that is independent of any particular $c$. The number of consecutive occurrences of 10 or more equally SCOP-labeled structures (i.e., runs $R_{\geq 10}$) is reported. Proposed in Appendix B, the number of runs $R_{\geq 10}$ is an approximate measure of association for the hierarchy derived from QMS and the SCOP classification. An additional measure of association for cross-classification, the Matthews coefficient, $Q_M$, is also reported. The Matthews coefficient depends on a particular clustering or similarity threshold $c$. Table III reports similarity $c$ that maximizes $Q_M$, as well as the number of single linkage clusters, $n_{sl}$, and the percentage of singletons or "outliers" at this particular $c$ value.

To facilitate the analysis, the clustering is performed on 9 different combinations of the SCOP classes α, β, α/β, and α + β. The cluster structure is stable with respect to this

**TABLE III. Classification**

| Sets | $n_d$ | $n_{sf}$ | $R_{\geq 10}$ | $c$ | $n_{sl}$ | Outliers % | $Q_M$ |
|---|---|---|---|---|---|---|---|
| A | 68 | 5 | 5 | 0.419 | 5 | 0.0 | 1.000 |
| B | 295 | 15 | 9 | 0.515 | 32 | 23.7 | 0.596 |
| C | 289 | 20 | 13 | 0.436 | 30 | 12.5 | 0.763 |
| D | 129 | 10 | 9 | 0.426 | 13 | 1.6 | 0.924 |
| ABC | 652 | 40 | 27 | 0.515 | 85 | 27.6 | 0.581 |
| ABD | 492 | 30 | 22 | 0.515 | 53 | 22.0 | 0.650 |
| ACD | 486 | 35 | 26 | 0.442 | 49 | 9.7 | 0.807 |
| BCD | 713 | 45 | 30 | 0.515 | 92 | 26.1 | 0.597 |
| ABCD | 781 | 50 | 35 | 0.515 | 99 | 25.9 | 0.608 |

A, B, C, and D in Sets stand for the classes α, β, α/β and α + β, respectively. The quantity $n_d$ is the number of domains at each combination of classes, $n_{sf}$ is the number of SCOP families, $R_{\geq 10}$ is the counted number of runs greater than or equal to 10. The similarity value $c$ is the one that maximizes the Matthews coefficient $Q_M$, while $n_{sl}$ is the corresponding number of single linkage clusters, and the percent outliers refers to the singletons.

**TABLE IV. Excerpt From the Group B Contingency Table, Class β, at a Similarity 0.442**

| | $c_1$ | $c_2$ | $c_3$ | Total | Singletons |
|---|---|---|---|---|---|
| b.1.1.1 | 27 | | | 27 | 1 |
| b.1.1.2 | 22 | | | 22 | 1 |
| b.1.1.4 | 37 | | | 37 | |
| b.1.1.5 | 18 | | | 22 | 6 |
| b.1.2.1 | 38 | | | 38 | |
| b.10.1.2 | | | | 9 | 7 |
| b.10.1.4 | | | | 10 | 5 |
| b.34.2.1 | 15 | | | 15 | |
| b.40.4.3 | | | | 7 | 4 |
| b.40.4.5 | 5 | | | 7 | 4 |
| b.47.1.2 | | 20 | | 20 | |
| b.6.1.1 | 9 | | | 9 | 1 |
| b.6.1.3 | 15 | | | 17 | |
| b.60.1.1 | | | 12 | 12 | |
| b.71.1.1 | 14 | | | 14 | |
| Total | 200 | 20 | 12 | 266 | 29 |

**TABLE V. Excerpt From the Group B Contingency Table, Class β, at a Similarity 0.515**

| | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | Total | Singletons |
|---|---|---|---|---|---|---|---|
| b.1.1.1 | 27 | | | | | 27 | 1 |
| b.1.1.2 | | | 22 | | | 22 | 1 |
| b.1.1.4 | 31 | | | | | 33 | 4 |
| b.1.1.5 | | | | | | 12 | 16 |
| b.1.2.1 | 1 | 32 | | | | 33 | 5 |
| b.10.1.2 | | | | | | 7 | 9 |
| b.10.1.4 | | | | | | 10 | 5 |
| b.34.2.1 | | | | | 12 | 12 | 3 |
| b.40.4.3 | | | | | | 2 | 9 |
| b.40.4.5 | | | | | | 6 | 5 |
| b.47.1.2 | | | | 20 | | 20 | |
| b.6.1.1 | | | | | | 8 | 2 |
| b.6.1.3 | | | | | | 14 | 3 |
| b.60.1.1 | | | | | | 9 | 3 |
| b.71.1.1 | | | | | | 10 | 4 |
| Total | 59 | 32 | 22 | 20 | 12 | 225 | 70 |

addition and subtraction of classes, as it emerges from the number of runs. Also, the class β, hereafter subset B, reflects a greater fractioning with 60% of recovered runs, while the whole set recovers 70%. This different behavior manifests as well in the Matthews coefficient. The maximum correlation in subset B is 0.60, at a similarity value $c = 0.515$, contrasting with a 0.81 correlation at $c = 0.442$ for the α, α/β, and α + β combination, hereafter subset ACD. Whenever a combination contains B, the similarity increases and the correlation found reduces. The following discussion treats these two groups separately.

### *Subset B*

Within the single-linkage hierarchy, the families with SCOP codes b.47.1.2 and b.60.1.1 first separate from the bulk at a relatively high similarity, 0.417. Later, at the similarity range 0.452–0.515, the families b.34.2.1, b.71.1.1, and b.1.2.1 separate. They form spherical clusters close in similarity space. The two immunoglobulin families, b.1.1.1 and b.1.1.4, however, appear connected up to a similarity threshold of 0.595. Classical multidimensional scaling[44] and stochastic proximity embedding projections[45] show alike two enlarged clusters with some of their domains spread across, like a loosely woven bridge between the two.

Excerpts from the QMS versus SCOP contingency are shown in Tables IV and V. Only the clusters containing 10 or more domains are reported. Table IV considers the QSM clustering at similarity 0.442, and Table V at similarity 0.515. There is a noticeable agglomeration at similarity 0.442. At similarity 0.515, all SCOP families except b.1.1.1 and b.1.1.4 are differentiated. This connectedness shifts the maximum Matthews coefficient to high similarity values. Thus, singletons and split families are numerous and the correlation degrades.

The families b.6.1.3, b.60.1.1, and b.71.1.1, accounted for by the $R_{\geq 10}$ test, separate at a greater than 0.442 similarity, but are broken at 0.515 similarity threshold. The remaining families appears fragmented and spread over the similarity space. This includes b.1.1.5 and the viral b.10.1.2 and b.10.1.4 families. The latter two are composed of domains of significantly varying shapes and sizes that will not bunch up under global measures of similarity.

### *Subset ACD*

Clusters within this group are neat and clear. Their identification with SCOP families is relevant. The $R_{\geq 10}$ test counts 26 of the 35 runs, approximately 70% of the SCOP clusters. The maximal correlation is equal to 0.807 at a similarity value 0.442, according to the Matthews coefficient.

The contingency table for the QMS clustering at a 0.442 similarity is shown in Table VI. The most noteworthy differences from SCOP classification are the two dehydrogenase families, c.2.1.3 and c.2.1.5, that appear clustered together. Also, the two domains d1dpga1 and d1qkia1 from the c.2.1.3 family are included in the c.2.1.5 family. Both families belong to the NAD(P)-binding Rossmann-fold domain superfamily.

**TABLE VI. Excerpt From the Contingency Table for the ACD Combination at a Similarity 0.442**

| | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_6$ | $c_7$ | $c_8$ | $c_9$ | $c_{10}$ | $c_{11}$ | $c_{12}$ | $c_{13}$ | $c_{14}$ | $c_{15}$ | $c_{16}$ | $c_{17}$ | $c_{18}$ | $c_{19}$ | $c_{20}$ | $c_{21}$ | $c_{22}$ | Total | Unclassified |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| a.1.1.2 | | | 19 | | | | | | | | | | | | | | | | | | | | 19 | |
| a.26.1.1 | | | | | | | | | | | | | | | | | | | | | | | 9 | 1 |
| a.3.1.1 | | | | | | | | | | 13 | | | | | | | | | | | | | 13 | 1 |
| a.45.1.1 | | | | | | | | 15 | | | | | | | | | | | | | | | 15 | |
| a.74.1.1 | | | | | | | | | | | | | | | | | | | | | | 10 | 10 | |
| c.1.4.1 | | | | | | | | | | | | | | | | | | | | | | | 9 | 1 |
| c.1.8.1 | | | | | | | | | | | | | | | | 12 | | | | | | | 12 | 2 |
| c.1.8.3 | | | | | | | 15 | | | | | | | | | | | | | | | | 15 | 1 |
| c.2.1.2 | | 22 | | | | | | | | | | | | | | | | | | | | | 29 | |
| c.2.1.3 | | | | | | | | | | | | | 2 | | | | | | | | | | 9 | 3 |
| c.2.1.5 | | | | | | | | | | | | | 10 | | | | | | | | | | 10 | |
| c.23.1.1 | | | | | | | | | | | | | | | | | | | | 10 | | | 10 | |
| c.26.1.1 | | | | | | | | | | | | | | | | | | | | | | | 9 | 2 |
| c.3.1.5 | 23 | | | | | | | | | | | | | | | | | | | | | | 23 | |
| c.36.1.1 | | | | | | | | | | | | | | | | | | | 10 | | | | 10 | |
| c.37.1.13 | | | | | | | | | | | | | | | | | | | | | | | 12 | 10 |
| c.37.1.1 | | | | | | | | | | | | 13 | | | | | | | | | | | 13 | 3 |
| c.37.1.8 | | | | | | | | | | | | | | | | | | | | | | | 10 | 6 |
| c.47.1.5 | | | | | | | | | | | | | | | 12 | | | | | | | | 12 | |
| c.61.1.1 | | | | | | | | | | | | | | | | | | | | | | | 9 | 3 |
| c.67.1.3 | | | | 8 | | | | | | | | | | | | | | | | | | | 10 | |
| c.67.1.4 | | | | 9 | | | | | | | | | | | | | | | | | | | 10 | 1 |
| c.93.1.1 | | | | | | | | | | | | | | | | | | | | | | | 11 | 2 |
| c.94.1.1 | | | | | | | | | | | | | | 12 | | | | | | | | | 16 | 5 |
| c.95.1.1 | | | | | | | | | | | | | | | | | | | | | | | 11 | |
| d.104.1.1 | | | | | | | | | | | | | | | | | | | | | | | 10 | 3 |
| d.131.1.2 | | | | | | | | | | | | | | | | | | | | | | | 9 | 1 |
| d.144.1.1 | | | | | | | | | 14 | | | | | | | | | | | | | | 14 | |
| d.153.1.4 | | | | | 17 | | | | | | | | | | | | | | | | | | 17 | |
| d.162.1.1 | | | | | | | | | | | | | | | | | 11 | | | | | | 11 | |
| d.169.1.1 | | | | | | 15 | | | | | | | | | | | | | | | | | 15 | |
| d.185.1.1 | | | | | | | | | | | | | | | | | | 11 | | | | | 11 | |
| d.19.1.1 | | | | | | | | | | | | | | | | | | | | | | | 13 | |
| d.58.7.1 | | | | | | | | | | | | | | | | | | | | | 10 | | 10 | 2 |
| d.93.1.1 | | | | | | | | | | | 13 | | | | | | | | | | | | 13 | |
| Total | 23 | 22 | 19 | 17 | 17 | 15 | 15 | 15 | 14 | 13 | 13 | 13 | 12 | 12 | 12 | 12 | 11 | 11 | 10 | 10 | 10 | 10 | 439 | 47 |

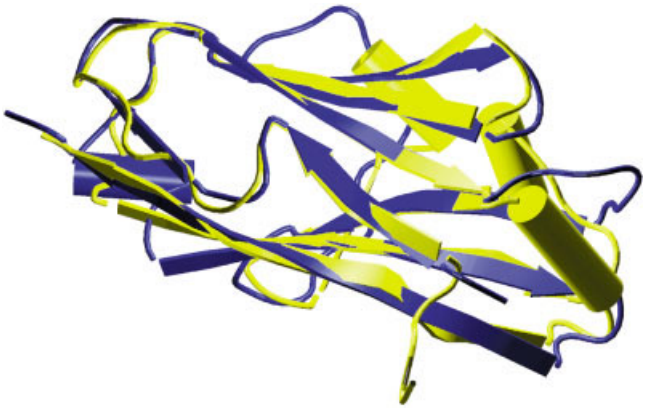

Fig. 3. Alignment of the fibronectin type III domain d1i1ra1, in yellow, with the immunoglobulin domain d1epfa1, in blue. Plotted using VMD[57] and Raster3D.[58]

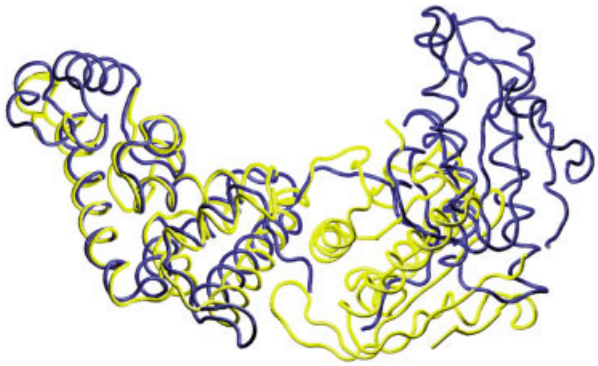

Fig. 4. Alignment of the domains d1uaaa2, in blue, and d1pjr_2 in yellow. Local similarity and functional analogies prevailed in classifying these two extended AAA-ATPase domains. Plotted using VMD[57] and Raster3D.[58]

### Alignments

I provide here a direct understanding of the intricacies of these classifications by presenting two peculiar structure alignments.

Immunoglobulin families appear already unconnected at a similarity value 0.515, except for the I and V set domains that are still linked. Also linked to them is the Fibronectin type III domain d1i1ra1. Among the Fibronectin type III domains, d1i1ra1 is closer to d1fna_, with a similarity of 0.456. The closest to d1i1ra1, however, are the immunoglobulin I set domains d1cs6a3, d1g1ca_, and d1epfa1, with similarities 0.519, 0.529 and 0.553, respectively. Although d1i1ra1 and d1epfa1 belong to two different SCOP superfamilies, their structural resemblance is extraordinary. Their electron density alignment is plotted in Figure 3.

It is easily perceived from Table VI that the family c.37.1.13, extended AAA-ATPase domains, has 10 singletons and a considerable fragmentation at a similarity of 0.442. The SCOP criteria rely preferentially on functionality, while structural similarity is kept to only local patterns of the protein domain. This tendency to emphasize localized structural features has also been noted by Hou et al.[46] The alignment of the two extended AAA-ATPase domains, d1uaaa2 and d1pjr_2, is displayed in Figure 4. The two structures are almost identical at one end of the chain, yet the other end shows more of a chiral image of the two domains compared. Their global electron density measure of similarity is 0.367 only.

## CONCLUSIONS

The maximization of the overlap between simplified electron density models of proteins is a reliable approach to structural alignments. An encompassing and systematically inclusive calculation toward a complete and global maximization of the density overlap or similarity function exists. This permits detecting global and local common patterns in a systemic procedure. Such structural similarity analysis is conceptually unrelated to preconceived biochemical structure elements, which thus facilitates a clear distinction between structure-only common patterns and common biochemical motifs. Nonetheless, a remarkable correlation between the simple overlap measure and the knowledge-based SCOP criteria is found. Such a degree of correlation emphasizes the functional significance and usefulness of electron density alignments. It sketches a surprising perspective of the present theories in the understanding of molecules.

## ACKNOWLEDGMENT

I am grateful to S. Vega for her careful reading of the manuscript.

## REFERENCES

1. Koehl P. Protein structure similarities. Curr Opin Struct Biol 2001;11:348–353.
2. Lo Conte L, Ailey B, Hubbard TJP, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res 2000;28:257–259.
3. Brenner SE, Koehl P, Levitt R. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 2000; 28:254–256.
4. Shindyalov IN, Bourne PE. A database and tools for 3-D protein structure comparison and alignment using the combinatorial extension (CE) algorithm. Nucleic Acids Res 2001;29:228–229.
5. Bray JE, Todd AE, Pearl FMG, Thornton JM, Orengo CA. The CATH dictionary of homologous superfamilies (DHS): a consensus approach for identifying distant structural homologues. Protein Eng 2000;13:153–165.
6. Dietmann S, Holm L. Identification of homology in protein structure classification. Nat Struct Biol 2001;8:953–957.
7. Irving JA, Whisstock JC, Lesk AM. Protein structural alignments and functional genomics. Proteins 2001;42:378–382.
8. Feng ZK, Sippl MJ. Optimum superimposition of protein structures: ambiguities and implications. Fold Des 1996;1:123–132.
9. Godzik A. The structural alignment between two proteins: Is there a unique answer? Protein Sci 1996;5:1325–1338.
10. Jaroszewski L, Rychlewski L, Godzik A. Improving the quality of twilight-zone alignments. Protein Sci 2000;9:1487–1496.
11. Yang AS, Honig B. An integrated approach to the analysis and modeling of protein sequences and structures: II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. J Mol Biol 2000;301:679–689.
12. Randic M, Krilov G. Characterization of 3-D sequences of proteins. Chem Phys Lett 1997;272:115–119.
13. Randic M, Krilov G. On a characterization of the folding of proteins. Int J Quantum Chem 1999;75:1017–1026.

14. Mezey PG. Quantum chemistry of macromolecular shape. Int Rev Phys Chem 1997;16:361–388.
15. Mezey PG, Fukui K, Arimoto S, Taylor K. Polyhedral shapes of functional group distributions in biomolecules and related similarity measures. Int J Quantum Chem 1998;66:99–105.
16. Carugo O, Pongor S. Protein fold similarity estimated by a probabilistic approach based on C-alpha-C-alpha distance comparison. J Mol Biol 2002;315:887–898.
17. Rogen P, Fain B. Automatic classification of protein structure by using Gauss integrals. Proc Natl Acad Sci USA 2003;100:119–124.
18. McLachlan AD. A mathematical procedure for superimposing atomic coordinates of proteins. Acta Crystallogr A 1972;28:656–657.
19. Kabsch W. Solution for best rotation to relate 2 sets of vectors. Acta Crystallogr A 1976;32:922–923.
20. Carbó R, editor. Molecular similarity and reactivity: from quantum chemical to phenomenological approaches. Amsterdam: Kluwer Academic; 1995.
21. Carbó-Dorca R, Mezey PG, editors. Advances in molecular similarity. Vol. 1. Greenwich, CT: JAI Press; 1996.
22. Kohn W, Becke AD, Parr RG. Density functional theory of electronic structure. J Phys Chem 1996;100:12974–12980.
23. Kohn W. Nobel lecture: Electronic structure of matter-wave functions and density functionals. Rev Mod Phys 1999;71:1253–1266.
24. Constans P, Amat L, Carbó-Dorca R. Toward a global maximization of the molecular similarity function: superposition of two molecules. J Comput Chem 1997;18:826–846.
25. Mestres J, Solà M, Duran M, Carbó R. On the calculation of ab-initio quantum molecular similarities for large systems—fitting the electron-density. J Comput Chem 1994;15:1113–1120.
26. Constans P, Carbó R. Atomic shell approximation—electron-density fitting algorithm restricting coefficients to positive values. J Chem Inf Comput Sci 1995;35:1046–1053.
27. Constans P, Amat L, Fradera X, Carbó R. Quantum molecular similarity measures and the atomic shell approximation. Vol. 1. Greenwich, CT: JAI Press; 1996. p 187.
28. Constans P. Linear scaling approaches to quantum macromolecular similarity: evaluating the similarity function. J Comput Chem 2002;23:1305–1313.
29. Ayala PY, Scuseria GE. Linear scaling second-order Moller–Plesset theory in the atomic orbital basis for large molecular systems. J Chem Phys 1999;110:3660–3671.
30. Exner TE, Mezey PG. Ab initio quality properties for macromolecules using the ADMA approach. J Comput Chem 2003;24:1980–1986.
31. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP—a Structural Classification of Proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
32. Levitt M, Gerstein M. A unified statistical framework for sequence comparison and structure comparison. Proc Natl Acad Sci USA 1998;95:5913–5920.
33. Carbó R, Leyda L, Arnau M. How similar is a molecule to another?: an electron-density measure of similarity between two molecular structures. Int J Quantum Chem 1980;17:1185–1189.
34. Coppens P. X-ray charge densities and chemical bonding. New York: Oxford University Press; 1997.
35. Coulson CA, Thomas MW. The effect of molecular vibrations on apparent bond lengths. Acta Crystallogr B 1971;27:1354–1359.
36. Chandonia JM, Walker NS, Conte LL, Koehl P, Levitt M, Brenner SE. ASTRAL compendium enhancements. Nucleic Acids Res 2002;30:260–263.
37. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. Protein Data Bank—computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
38. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
39. Constans P. LSim: a quantum macromolecular similarity program, v 1.1. 2001–2003.
40. Yang AS, Honig B. An integrated approach to the analysis and modeling of protein sequences and structures: I. Protein structural alignment and a quantitative measure for protein structural distance. J Mol Biol 2000;301:665–678.
41. Ortiz AR, Strauss CEM, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 2002;11:2606–2621.
42. Sneath PHA. The application of computers to taxonomy. J Gen Microbiol 1957;17:201–226.
43. Hartigan JA. Consistency of single linkage for high-density clusters. J Am Stat Assoc 1981;76:388–394.
44. Mardia KV, Kent JT, Bibby JM. Multivariate analysis. London: Academic Press; 1979.
45. Agrafiotis DK, Xu HF. A self-organizing principle for learning nonlinear manifolds. Proc Natl Acad Sci USA 2002;99:15869–15872.
46. Hou JT, Sims GE, Zhang C, Kim SH. A global representation of the protein fold space. Proc Natl Acad Sci USA 2003;100:2386–2390.
47. Simonoff JS. Smoothing methods in statistics. New York: Springer-Verlag; 1996.
48. Harris JW, Stocker H. Handbook of mathematics and computational science. New York: Springer-Verlag; 1998.
49. Gower JC, Ross GJS. Minimum spanning trees and single linkage cluster analysis. Appl Stat-J Roy Stat Soc 1969;18:54–64.
50. Sibson R. Order invariant methods for data analysis. J Roy Stat Soc B 1972;34:311–349.
51. Hartigan JA. Clustering algorithms [Wiley series in probability and mathematical statistics]. New York: Wiley; 1975.
52. Gordon AD. A review of hierarchical classification. J Roy Stat Soc A 1987;150:119–137.
53. Morgan BJT, Ray APG. Non-uniqueness and inversions in cluster analysis. Appl Stat-J Roy Stat Soc 1995;44:117–134.
54. Murtagh F. Counting dendrograms—a survey. Discrete Appl Math 1984;7:191–199.
55. Knuth DE. The art of computer programming. seminumerical algorithms. Vol 2, 3rd edition. Reading, MA: Addison-Wesley; 1998.
56. Matthews BW. Comparison of predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–451.
57. Humphrey W, Dalke A, Schulten K. VMD—visual molecular dynamics. J Mol Graph 1996;14:33–38.
58. Merritt EA, Bacon DJ. Raster3D: Photorealistic molecular graphics. Methods Enzymol 1997;277:505–524.

## APPENDIX A: STATISTICAL SIGNIFICANCE

### Probability Density Function Estimates

The distribution of true-negative similarity index values is inferred from both a nonparametric kernel estimate and as an adjusted beta distribution.

#### *Nonparametric density estimation*

Nonparametric or model free density estimates present the general form

$$\hat{f}_K(C) = \frac{1}{nh} \sum_{i<j}^{N} K\left(\frac{C - C_{ij}}{h}\right). \quad \text{(A1)}$$

The kernel $K$ is taken here as the Gaussian distribution $(2\pi)^{-1/2} e^{-(C-C_{ij})^2/2}$. The bandwidth or smoothing parameter $h$ is simply derived from data as

$$\hat{h} = 1.059 \sqrt{\frac{\hat{\sigma}}{n^{1/5}}}, \quad \text{(A2)}$$

where $\hat{\sigma}$ is the usual standard deviation estimate, and $n$ is the number of nonredundant $C_{ij}$ measurements.[47]

***Beta distribution***

The beta distribution is used to model distributions for random variables whose values are bounded, being

$$\hat{f}_\beta(C) = C^{\alpha-1}\frac{(1-C)^{\beta-1}}{B(\alpha,\beta)} \quad \text{(A3)}$$

for $0 \le C \le 1$, and zero otherwise.[48] $B(\alpha,\beta)$ is the beta function. The mean of this distribution is

$$\mu = \frac{\alpha}{\alpha+\beta}, \quad \text{(A4)}$$

and the variance is

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \quad \text{(A5)}$$

Estimates $\hat{\mu}$ and $\hat{\sigma}$ permit, after solving Eqs. (A4) and (A5), derivation of the $\hat{\alpha}$ and $\hat{\beta}$ parameters of the beta distribution.

***Cumulative distribution functions***

The probability that a particular similarity measurement $C_{ij}$ will be greater than or equal to $c$ is given by the cumulative distribution function $F(c)$,

$$P_{tn}(C_{ij} \ge c) = 1 - F_{tn}(c) = 1 - \int_0^c f_{tn}(C)dC. \quad \text{(A6)}$$

Given a $P_{tn}$ value for the true-negative distribution $F_{tn}(c)$, similarity thresholds $c_k$ or $c_\beta$ are numerically obtained from the integral in Eq. (A6), using the estimates in Eqs. (A1) or (A3), respectively. The coverage at a given $P_{tn}$ value and its corresponding similarity $c_p$ is the probability

$$P_{tp}(C_{ij} \ge c_p) = 1 - F_{tp}(c_p) = \int_{c_p}^1 f_{tp}(C)dC. \quad \text{(A7)}$$

## APPENDIX B: CLUSTERING

### Single Linkage Clustering

Let $Q$ be a set of $n$ objects and $C$ an $n$ by $n$ matrix with

$$C_{ij} = C_{ji}, \quad \text{(B1)}$$

$$0 \le C_{ij} \le 1, \quad \text{(B2)}$$

and

$$C_{ii} = 1, \quad \text{(B3)}$$

establishing the proximity or similarity relationships among the objects. Consider an object $i \in Q$ and a neighborhood radius $c$ such that $0 \le c \le 1$. Then, the partition $P^c_\mu$ of $Q$ given by

$$P^c_\mu = \{j|c_{ij} \ge c;\ \forall j\} \cup \{k|c_{jk} \wedge c_{ij} \ge c;\ \forall j,k\} \cup \cdots \quad \text{(B4)}$$

is the subset that contains object $i$ and its relatives by a single linkage. Alternative formulations and statistical properties of the single linkage clustering are found elsewhere.[42,43,49–53]

Partitions drawn by definition Eq. (B4) are disjoint, that is,

$$P^c_\mu \cap P^c_\nu = \varnothing \ \forall \mu \ne \nu, \quad \text{(B5)}$$

and consequently,

$$Q = \bigcup_\mu P^c_\mu. \quad \text{(B6)}$$

There may exist a proximity radius $c'$, such that $c' > c$, for which $P^c_\mu$ is partitioned into disjoint subsets $P^{c'}_{\mu'}$, that is,

$$P^c_\mu = \bigcup_{\mu'} P^{c'}_{\mu'}, \quad \text{(B7)}$$

thus producing a hierarchy of partitions. Single linkage hierarchies are unique, independent of agglomerative or divisive procedures, and independent of random starting seeds.

### Randomness and Dendrogram Ordering

Let $P^c_1 \cdots P^c_m$ be all the $m$ partitions of $Q$ at a proximity value $c$, and denote by $n_1 \ldots n_m$ their respective cardinalities. Let's label all $n_\mu$ objects in partition $P^c_\mu$ as being of class $C^c_\mu$. Then, from the hierarchy definition, Eqs. (B6) and (B7), there are exactly $m$ runs in the dendrogram series of objects (i.e., occurrences of contiguous objects equally labeled). The lengths of the runs will be the cardinalities $n_\mu$. Moreover, the occurrence of runs in the series is an order invariant from all equivalent dendrograms.†

Considered an outer or reference classification $S$ for the objects in $Q$, with $m'$ classes having $n'_1 \ldots n'_{m'}$ elements, then, the counting of runs on the dendrogram series checks for the clustering of data and the extent of association between classifications $S$ and $C$. For instance, if the two classifications $S$ and $C$ were unrelated, the probability of $q$ contiguous, equally $S$-labeled objects would be, approximately and for equally populated classes, $(1/m')^q$. In case of having 50 different $S$ labels, as in this study, and if the ordering were random, the chances of seeing one run of length equal to 10 would be as low as $10^{-17}$. On the other hand, if classification $S$ were equivalent to one of the $C^c$, and each partition had 10 or more elements, the number of runs seen greater than or equal to 10 would be exactly 50.

There are several known tests for checking randomness of series, yet very few for checking the clustering. The proposed test, though only approximately invariant in general, provides qualitative evidence of data clustering and association of an a priori classification $S$ with a hierarchical one, $C$, while considering the whole range of proximity values $c$. See Murtaugh[54] for dendrogram counting and Knuth[55] for the serial and run tests as applied to randomness checking.

†The set $Q$ can be sorted in $n!$ different ways. This amount reduces to $m!n_1!\ldots n_m!$ If preserving classification $C^c$. An additional ordering is induced on the objects in $Q$ by the dendogram; there are fewer $2^{l-1}$ equivalent dendograms for a given hierarchy of $l$ leaves or nodes.

**Measure of Association for Cross Classification**

Hierarchical clustering produces a countable series of relation matrices $C^c$ defined as

$$C^c_{ij} = \begin{cases} 1, & \text{if } i \wedge j \in P^c_\mu \\ 0, & \text{if } i \in P^c_\mu \wedge j \in P^c_\nu. \end{cases} \quad \text{(B8)}$$

To analyze some relevant classifications with respect to an outer classification having a relation matrix $S$, the association of each $C^c$ with $S$ is evaluated by generalizing the Matthews correlation coefficient.[56] Correlation is expressed in terms of the usual terminology true-positive (tp), true-negative (tn), false-positive (fp), and false-negative (fn) cases. The number of cases $n_{tp}$, $n_{tn}$, $n_{fp}$, and $n_{fn}$ are defined as the occurrences of $C^c_{ij} = Sij = 1$, $C^c_{ij} = S_{ij} = 0$, $C^c_{ij} = 1$, while $S_{ij} = 0$, and $C^c_{ij} = 0$, while $S_{ij} = 0$, for all $i < j$, respectively. The correlation between the two classification matrices $C^c$ and $S$ is

$$Q_M = \frac{n_{tp}n_{tn} - n_{fp}n_{fn}}{\sqrt{(n_{tp} + n_{fn})(n_{tp} + n_{fp})(n_{tn} + n_{fp})(n_{tn} + n_{fn})}}, \quad \text{(B9)}$$

where superscript $c$ is omitted for notation clarity. The values taken by the Matthews coefficient are within the interval $-1 \leq Q_M < 1$. The coefficient is 0 for completely uncorrelated classifications and close to 1 or $-1$ for correlated or anticorrelated cases, respectively.