Proteins of the same fold and unrelated sequences have similar amino acid composition

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · JULY 2006

Impact Factor: 2.63 · DOI: 10.1002/prot.20964 · Source: PubMed

PROTEINS: Structure, Function, and Bioinformatics 64:275–279 (2006)

Proteins of the Same Fold and Unrelated Sequences Have Similar Amino Acid Composition

Yanay Ofran[1,2]* and Hanah Margalit[3]
[1]*CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York*
[2]*Columbia University Center for Computational Biology and Bioinformatics (C2B2), New York, New York*
[3]*Department of Molecular Genetics and Biotechnology, Faculty of Medicine, The Hebrew University of Jerusalem, Israel*

ABSTRACT **It is well established that there is a relationship between the amino acid composition of a protein and its structural class (i.e., α, β, α + β, or α/β). Several studies have even shown the power of amino acid composition in predicting the secondary structure class of a protein. Herein, we show that significant similarity in amino acid composition exists not only between proteins of the same class, but even between proteins of the same fold. To test conjectural explanations for this phenomenon, we analyzed a set of structurally similar proteins that are dissimilar in sequence. Based on this analysis, we suggest that specific residues that are involved in intramolecular interactions may account for this surprising relationship between composition and structure. Proteins 2006;64:275–279.**

INTRODUCTION

Since the early 1970s, it has been known that proteins of the same secondary structure class (all-α, all-β, α + β, or α/β) have similar amino acid composition, even if they are dissimilar in sequence.[1] There are several possible explanations for this phenomenon: One explanation relies on the fact that distinct types of secondary structures tend to have typical amino acid compositions.[2–4] Thus, proteins that are similar in their secondary structure content are likely to have similar amino acid compositions.[3,5–7] Alternatively, it is possible that the similarity in composition is merely a reflection of an evolutionary relatedness. A third explanation focuses on the residue–residue interactions that underlie the structure of a protein. Bahar et al.[8] suggested that the distributions of noncovalent contacts significantly differ between the different structural classes. Similarly, Dubchak et al.[9] hypothesized that the ability to predict folding types from composition implies that tertiary interactions, which require specific intramolecular contacts, are reflected in specific patterns of amino acid composition.

To further explore the relationship between amino acid composition and protein structure, we analyzed proteins that are dissimilar in sequence but share the same fold. We found that the relationship between amino acid composition and the protein structure is not confined to the structural class. Rather, proteins of the same structural fold also show significant similarity in their amino acid compositions even without a detectable sequence similarity.

When we clustered proteins using the amino acid composition of the α-helices, β-strands, or loop regions alone, we obtained even more significant discrimination between proteins of different folds. These results suggest that the common amino acid composition of structurally similar proteins could not be accounted for by the similarity of their secondary structures. Also, the dissimilarity in the sequences of the tested proteins rules out simple evolutionary relatedness. Thus, these results may support the conjecture that amino acid composition reflects specific intrachain interactions that underlie protein folding.

RESULTS

The Data Set

Based on the DALI database,[10] we built a data set of 210 groups of proteins (497 proteins all together), so that in every group the proteins share a similar structure but do not show any discernible sequence similarity [see Methods and Supplementary Material (http://www.interscience.wiley.com/jpages/0887-3585/suppmat)].

Clustering by Amino Acid Composition Versus Random Grouping

We clustered the proteins according to their amino acid composition and checked whether proteins tended to be clustered with their structural look-alikes (Fig. 1, see Methods for details). We found that in most cases proteins of the same fold were clustered into the same cluster based on their amino acid composition alone, with surprising accuracy (Fig. 2). For 90 of the 210 DALI-derived structural groups, 60% or more of the proteins in a group were classified correctly with their structural homologs, based on their amino acid composition only. Figure 2(a) demonstrates the statistical significance of this result, as evaluated by a Kolmogorov-Smirnov test that compared the actual clustering results with those generated by a random grouping ($P < 0.001$).

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat

*Correspondence to: Yanay Ofran, CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032. E-mail: yanay.ofran@columbia.edu

Received 26 August 2005; Revised 12 December 2005; Accepted 5 January 2006

Published online 24 March 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20964

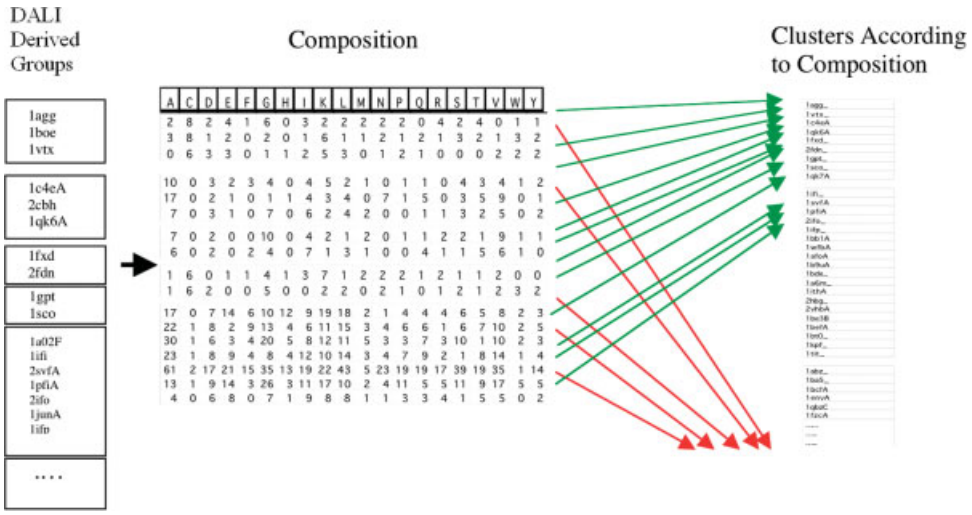

Fig. 1. Schematic description of the clustering procedure. On the **left side** are the DALI-derived groups. In every group, all the proteins (which are denoted by their PDB code) are of the same fold family, but there is no pair of proteins with detectable sequence similarity. The proteins were represented as vectors of 20 numbers representing their amino acid composition (percentage of each amino acid in the protein sequence). The proteins were clustered based on the similarities between their vectors of amino acid composition (on the **right side**). Later, we checked what percentage of every structural group was clustered together. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Clustering According to the Amino Acid Composition of Specific Secondary Structures

These results suggest that there is a strong relationship between amino acid composition and protein fold. However, this relationship may be a reflection of a trivial fact. It is widely accepted that each type of secondary structure has a typical amino acid composition. Proteins with similar folds often have similar secondary structure content and hence are likely to have similar amino acid compositions. To rule out this trivial explanation, we repeated the analysis using the amino acid composition of each secondary structure type separately. When we clustered the proteins according to the amino acid composition of the α-helices of each protein [Fig. 2(b)], we still got highly significant correspondence between composition and fold ($P < 0.001$, Kolmogorov-Smirnov). The results for β-strands and coil regions alone were also significant [Fig. 2(c,d)].

DISCUSSION

Amino acid composition of proteins has been shown to be related to some of their structural and functional characteristics such as subcellular localization[11] and stability.[12] More than three decades ago, it was shown that amino acid composition could be used to predict secondary structure content of proteins.[1] Several groups have since revisited this idea, demonstrating how amino acid composition could be used to predict structural class.[3,9,13–15] The methods used in these studies varied from multiple regression[1] through neural networks[9] to support vector machines.[13] The level of success reported varied from 60 to 100%. Eisenhaber et al.[3] analyzed what they defined to be a paradox—whereas prediction of structural class approaches 100% success, secondary structure prediction itself fails to surmount the accuracy rate of 80%. They suggested that success level for the former was overstated. Others, however, have suggested that this discrepancy can be accounted for by other explanations.

The three main explanations for the relationship between amino acid composition and structural class are: (1) evolutionary relatedness of proteins in the same class, (2) similarity in secondary structure content, and (3) intramolecular contacts that require very specific residues. Bahar et al.[8] suggested that the distributions of nonboding contacts significantly differ between folding classes. Similarly, Dubchak et al.[9] claimed that the ability to predict folding type from composition implies that tertiary interactions requiring specific intramolecular contacts are reflected by specific patterns of amino acid composition. Building on this explanation, we hypothesized that if intramolecular interactions are reflected in amino acid composition, then helices from structurally similar proteins are expected to be more similar in their composition than helices from proteins with dissimilar structures, because they are involved in similar tertiary interactions. The same should be true for strands and coils. We show that this is indeed the case. The clustering results, as shown in Figure 2, demonstrated that proteins of similar structure have similar amino acid compositions. The grouping of the sequences into random "clusters" provided the expected levels of success where there is no information in the amino acid composition that is relevant to the structure. As is shown in Figure 2(b), the effect was maintained when we used the composition of α-helices alone. Similar results were obtained for other secondary structures [Fig. 2(c,d)]. Thus, it is possible to suggest that the amino acid composition of helices, strands, or coils in distinct folds appears to contain some information regarding the exact fold of the protein.

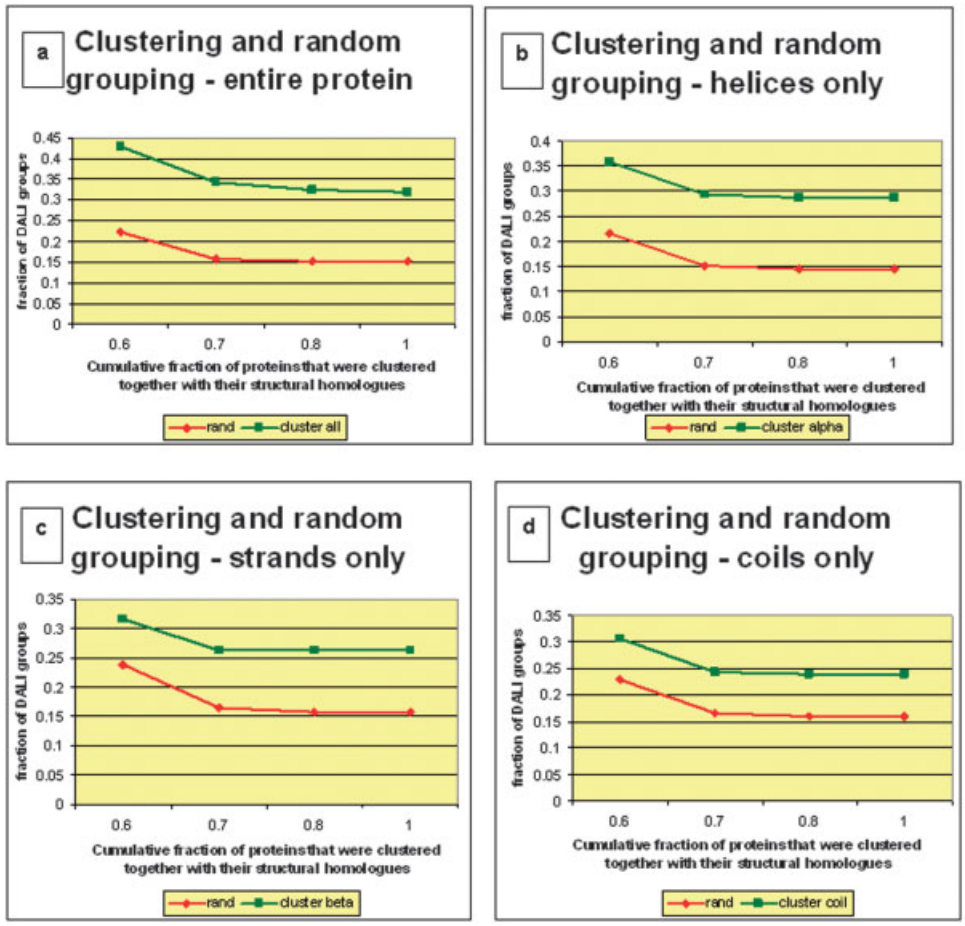

Fig. 2. Comparison between real clustering by amino acid composition and random grouping. On the *x*-axis is the fraction of proteins from a structural group that were clustered in the same cluster based on the amino acid composition. On the *y*-axis is the fraction of all structural groups in which at least this percentage of sequences were clustered together. The green line represents the results of the clustering by amino acid composition by the Bottleneck algorithm. The red line represents the average results of 100,000 random groupings (see Methods). **a:** Clustering of the amino acid composition vectors of the entire sequence of each protein. **b:** Clustering of the amino acid composition vectors of only α-helices of each protein. **c:** β-Strand. **d:** Coil regions.

The DALI-derived data set included structurally similar proteins classified to 210 different folds. We ascertained that any detectable composition similarity could not be ascribed to simple sequence similarity by excluding any proteins with similar sequences from the data set. In each DALI-derived structural group, there were no two proteins that found each other in a BLAST search. By clustering them according to their amino acid composition alone, we could check whether the composition inherits information regarding the fold. Our results, therefore, show that the similarity in amino acid composition of structurally similar proteins is not attributable to evolutionary relatedness nor is it a reflection of the similarity in secondary structure content. Thus, we refuted two of the conjectural explanations for the relationships between composition and structure.

CONCLUSIONS

Amino acid composition is a very noisy measure. On the one hand, it might not be sensitive enough to detect short local similarities between sequences; on the other hand, very different sequences, with very different structures, may exhibit similar amino acid compositions. Yet, as we showed here, there is a significant similarity in amino acid composition between proteins of the same structure, even when they exhibit no detectable sequence similarity. At the conceptual level, these results suggest that despite the ostensible dissimilarity in their sequences, proteins of the same fold share some elements of their sequence that lead them to the same structures. In other words, these results support the view that even seemingly different sequences may encode similar "folding instructions." At the practical level, our results suggest that when training machine-learning tools to identify structure, one should use even very different sequences that share the same structure. A prediction tool that will incorporate the amino acid composition measure might achieve a higher success in predicting the structure from the sequence.

METHODS

Choosing Database

The DALI database[16] is based on exhaustive all-versus-all structural alignments of proteins in the Protein Data Bank (PDB) database. The classification and assignment are automatically maintained and continuously updated using the DALI program.[17] Each DALI entry has a single structural representative, against which all structurally similar proteins are aligned in decreasing order of structural similarity. DALI structural groupings are not mutually exclusive, i.e., a particular protein can appear in more than one group. This redundancy, namely, the existence of individual structures that have a low root-mean-square deviation (RMSD) to more than one representative, creates a problem in defining a structural fold. We tried to minimize the effect of this problem by eliminating duplicates (see next section).

Data

From each fold family of DALI we selected all proteins that fulfill the following criteria:

1. A significant structural alignment above a certain threshold ($z > 4.5$), which is higher than the cutoff of DALI itself and ascertains higher structural similarity.
2. Sequence identity is lower than 25%.
3. The structural alignment spans >75% of the longer sequence.
4. The RMSD between the two structures is <2.5 Å.

A protein that appeared in more than one fold was maintained only in one family.

To ensure that there is no sequence similarity between the proteins in each structural group, we ran BLAST[18] against PDB for each protein to ensure that it did not detect any of its structural homologs in standard sequence comparison ($E < 100$).

At the end of the process, we had 497 proteins grouped into 210 structural groups. In each group, there were no two proteins with sequence similarity that could be detected by BLAST.

Clustering Algorithm

Each protein was represented as a vector of 20 numbers, representing the composition of the 20 amino acids. Specifically, every number $P(a|p)$ in each of the vectors is the probability of seeing the amino acid a in the protein p. This representation of the composition ascertains that the clustering results will not be biased by the similarity in length between proteins in the same DALI group. Next, the vectors were clustered based on their amino acid composition, using an implementation of the agglomerative information bottleneck algorithm.[19] This is a bottom-up method that attempts to maximize the mutual information between the protein clusters and the amino acid composition. The algorithm starts with a trivial partition in which each element (in our case, the amino acid composition of a single protein) forms a separate cluster. Then, at each step, two clusters are merged into a new cluster. The merged clusters are chosen so that the loss of mutual information between the amino acid composition and the protein clusters will be minimal. This process is repeated until reaching a predefined number of clusters (10 clusters). At the heart of the algorithm lies the Jensen-Shannon measure of divergence,[20] which, as Slonim and Tishby show, can be used to assess the loss of mutual information. The Jensen-Shannon measure of divergence[20] is given by the equation:

$$JS(p_1,p_2) = H(\pi_1 p_1 + \pi_2 p_1) - \pi_1 H(p_1) - \pi_2 H(p_2)$$
$$\text{with: } \pi_1 \text{ and } \pi_2 \geq 0, \text{ and } \pi_1 + \pi_2 = 1$$
$$\text{and } H(p) = -\sum_i p(x_i)\log_2(x_i) \quad (1)$$

where π_1 and π_2 are the weights of the two probability distributions p_1 and p_2, respectively, and H(x) is the Shannon entropy.[21]

Using the bottleneck algorithm, we clustered the composition vectors. The final clustering may have grouped together proteins from different structural groups (Fig. 1). However, if amino acid composition is relevant to the fold, we expect that proteins from the same DALI group will be clustered together more than expected at random.

Statistical Significance

To check whether the clustering maintained the original DALI-derived groups (the fold families), we compared the clustering results to a random clustering of the proteins into groups. This procedure enabled us to assess the significance of the actual clustering of composition vectors compared with random clustering. This was done as follows:

1. After finishing the clustering, each of the DALI-derived groups was checked to see what percent of the group was in one cluster. Then we analyzed the distribution of the successfully clustered proteins (for how many structural groups 10% or more of the proteins were classified into the same cluster, for how many structural groups 20% or more of the proteins were classified into the same cluster, etc).
2. The composition vectors were randomly grouped into groups corresponding in size and number to the clusters produced by the bottleneck algorithm.
3. Similar to stage 1, for the random clustering we recorded the number of DALI groups of which at least 10% of the proteins ended up in the same random group. Then we recorded the number of groups of which at least 20% of the proteins were randomly grouped together, and so forth. This process was repeated 100,000 times and the average and median number of groups for each percentile were calculated.
4. At the end of the process, the accumulated distribution of the real clustering results was compared with the random one by a Kolmogorov-Smirnov test (see Fig. 2). This test provided the assessment of the significance of the clustering results.

Analysis of Secondary Structures

The assignment of secondary structures was done using DSSP.[22] Then the sequences of only α-helices were extracted. Every protein was now represented as a vector of 20 numbers, representing the amino acid counts of its α-helices. Then, these vectors were clustered using the bottleneck algorithm, as described above. The same process was repeated for β-strands and coils.

ACKNOWLEDGMENTS

The authors deeply thank Prof. Norman Grover (HUJI) for his help with the statistical analysis and his insightful comments, Noam Slonim and Prof. Naftaly Tishbi (both at HUJI) for their help with the bottleneck algorithm, and Profs. Burkhard Rost (Columbia University) and Michael Levitt (Stanford University) for helpful and insightful discussions.

REFERENCES

1. Krigbaum WR, Knutton SP. Prediction of the amount of secondary structure in a globular protein from its amino acid composition. Proc Natl Acad Sci USA 1973;70(10):2809–2813.
2. Chou PY, Fasman GD. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry 1974;13(2):211–222.
3. Eisenhaber F, Frommel C, Argos P. Prediction of secondary structural content of proteins from their amino acid composition alone. II. The paradox with secondary structural class. Proteins 1996;25(2):169–179.
4. Koehl P, Levitt M. Structure-based conformational preferences of amino acids. Proc Natl Acad Sci USA 1999;96(22):12524–12529.
5. Cai Y, Zhou G. Prediction of protein structural classes by neural network. Biochimie 2000;82(8):783–785.
6. Dubchak I, Muchnik I, Holbrook SR, Kim SH. Prediction of protein folding class using global description of amino acid sequence. Proc Natl Acad Sci USA 1995;92(19):8700–8704.
7. Lin Z, Pan XM. Accurate prediction of protein secondary structural content. J Protein Chem 2001;20(3):217–220.
8. Bahar I, Atilgan AR, Jernigan RL, Erman B. Understanding the recognition of protein structural classes by amino acid composition. Proteins 1997;29(2):172–185.
9. Dubchak I, Holbrook SR, Kim SH. Prediction of protein folding class from amino acid composition. Proteins 1993;16(1):79–91.
10. Holm L, Sander C. Mapping the protein universe. Science 1996; 273(5275):595–603.
11. Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 2001;17(8):721–728.
12. Avbelj F, Baldwin RL. Role of backbone solvation in determining thermodynamic beta propensities of the amino acids. Proc Natl Acad Sci USA 2002;99(3):1309–1313.
13. Cai YD, Liu XJ, Xu XB, Chou KC. Prediction of protein structural classes by support vector machines. Comput Chem 2002;26(3):293–296.
14. Liu W, Chou KC. Prediction of protein secondary structure content. Protein Eng 1999;12(12):1041–1050.
15. Zhang CT, Chou KC. Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition. Biophys J 1992;63(6):1523–1529.
16. Holm L, Sander C. The FSSP database of structurally aligned protein fold families. Nucleic Acids Res 1994;22(17):3600–3609.
17. Holm L, Sander C. Dali: a network tool for protein structure comparison. Trends Biochem Sci 1995;20(11):478–480.
18. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389–3402.
19. Slonim N, Tishbi N. Agglomerative information bottleneck. In: Solla S, Leen T, Muller K, editors. Advances in neural information processing. Cambridge, MA: MIT Press; 1999. p 617–623.
20. Lin J. Divergence measures based on the Shannon entropy. IEEE Trans Inform Theory 1991;37(1):145.
21. Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948;27:379–423, 623–656.
22. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577–2637.