

Protein structure space is much more than the sum of its folds

To the Editor:

A recent editorial in *Nature Structural & Molecular Biology* echoes questions being raised in the structural biology community and elsewhere as to whether the expense of The Protein Structure Initiative (PSI) is justified, given the current budget crunch at the US National Institutes of Health. Although I believe that you have accurately represented both sides of the argument as they pertain to methods development in experimental structure determination, I feel that your portrayal of some of the underlying conceptual and computational issues is not consistent with recent developments in structural bioinformatics. Specifically, the goal of creating a library of all protein folds is highlighted in a number of places in the editorial, yet this goal has largely been abandoned by the PSI, in part owing to the realization that it is essentially impossible to classify or enumerate folds. In fact, there is no clear quantitative measure available as to how a fold should be defined.

The problem goes beyond issues of definition. There are many examples of significant geometric relationships between proteins that at first glance appear to have quite different structures and that have been classified as having different folds. Geometric alignment programs

often detect large overlapping structural regions in proteins that have been classified as belonging to different folds. Moreover, the existence of such regions can reveal important functional relationships. Increasingly, it appears that fold space should be regarded as continuous, which of course obviates any attempt at classification. Indeed, changes underway in both the CATH and SCOP databases reflect the difficulties associated with a hierarchical classification based on geometric similarity.

This alternative perspective on protein-structure space has the potential to significantly modify prevailing views of the evolution of protein structure and function. On a practical level, it will lead to new tools for the detection of previously unknown relationships among proteins. This set of goals has emerged, for the most part, from researchers in the structural genomics community who are motivated by the need to extract maximal information from each structure that is solved. In parallel, the use of homology models — and hence structural information — in all areas of biology is growing, a process that is being accelerated by the close coupling of modeling with target selection and analysis in structural genomics initiatives. As pointed out in your editorial, much remains to be done in improving model accuracy, but

current models are often accurate enough to reveal novel biological insights — for example, when they are applied to the identification of common features, or specificity-determining features, among members of a protein family.

Structural genomics initiatives are leading to innovative science that could have a major impact both on the elucidation of general principles of protein structure and function and on the use of structure in a variety of biological applications. Alternative uses of the funds invested in these initiatives, such as the determination of increasingly more complex biological structures, are of course extremely valuable, and I do not wish here to make a case favoring one goal over others. However, I hope that recognition that structural genomics is accomplishing much more than counting folds may facilitate a more balanced discussion of the underlying issues.

Barry Honig

*Department of Biochemistry and Molecular Biophysics, Center for Computational Biology and Bioinformatics, and Howard Hughes Medical Institute, Columbia University, New York, New York 10032, USA.
e-mail: bh6@columbia.edu*

COMPETING INTERESTS STATEMENT

The author declares no competing financial interests.