17. Xu, S., Falvey, D. A. & Brandriss, M. C. Roles of URE2 and GLN3 in the proline utilization pathway in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **15,** 2321–2330 (1995).
18. Marton, M. J. *et al.* Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.* **4,** 1293–1301 (1998).
19. Shamji, A. F., Kuruvilla, F. G. & Schreiber, S. L. Partitioning the transcriptional program induced by rapamycin among the effectors of the Tor proteins. *Curr. Biol.* **10,** 1574–1581 (2000).
20. Kuruvilla, F. G., Shamji, A. F. & Schreiber, S. L. Carbon- and nitrogen-quality signaling to translation are mediated by distinct GATA-type transcription factors. *Proc. Natl Acad. Sci. USA* **98,** 7283–7288 (2001).
21. Kuruvilla, F. G., Park P. J. & Schreiber, S. L. Vector algebra in the analysis of genome-wide expression data. *Genome Biol.* **3(3),** 0011.1–0011.11 (2002).
22. Bertram, P. G. *et al.* Tripartite regulation of Gln3p by TOR, Ure2p and phosphatases. *J. Biol. Chem.* **275,** 35727–35733 (2000).
23. Causton, H. C. *et al.* Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell* **12,** 323–337 (2001).
24. Gasch, A. P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11,** 4241–4257 (2000).
25. Kornberg, H. L. & Krebs, H. A. Synthesis of cell constituents from C_2 units by a modified tricarboxylic acid cycle. *Nature* **179,** 988–991 (1957).
26. Bogonez, E., Machado, A. & Satrustegui, J. Ammonia accumulation in acetate-growing yeast. *Biochim. Biophys. Acta* **733,** 234–241 (1983).
27. Edskes, H. K., Hanover, J. A. & Wickner, R. B. Mks1p is a regulator of nitrogen catabolism upstream of Ure2p in *Saccharomyces cerevisiae*. *Genetics* **153,** 585–594 (1999).
28. Edskes, H. K. & Wickner, R. B. A protein required for prion generation: [URE3] induction requires the ras-regulated mks1 protein. *Proc. Natl Acad. Sci. USA* **97,** 6625–6629 (2000).

Supplementary Information accompanies the paper on *Nature's* website (http://www.nature.com).

Acknowledgements

We thank R. Melki for providing bacterially expressed Ure2p protein. F.G.K. was supported by the NIH Medical Scientist Training Program, A.F.S. by the Howard Hughes Medical Institute predoctoral fellowship, S.M.S. by the Roche and NSF predoctoral fellowships, and P.J.H. by the American Cancer Society. S.L.S. is an Investigator at the Howard Hughes Medical Institute. This research was funded by a grant from the NIGMS (GM-38627).

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.L.S. (e-mail: sls@slsiris.harvard.edu).

A 'periodic table' for protein structures

William R. Taylor

Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK

Current structural genomics programs aim systematically to determine the structures of all proteins coded in both human and other genomes, providing a complete picture of the number and variety of protein structures that exist. In the past, estimates have been made on the basis of the incomplete sample of structures currently known. These estimates have varied greatly (between 1,000 and 10,000; see for example refs 1 and 2), partly because of limited sample size but also owing to the difficulties of distinguishing one structure from another. This distinction is usually topological, based on the fold of the protein; however, in strict topological terms (neglecting to consider intra-chain cross-links), protein chains are open strings and hence are all identical. To avoid this trivial result, topologies are determined by considering secondary links in the form of intra-chain hydrogen bonds (secondary structure) and tertiary links formed by the packing of secondary structures. However, small additions to or loss of structure can make large changes to these perceived topologies and such subjective solutions are neither robust nor amenable to automation. Here I formalize both secondary and tertiary links to allow the rigorous and automatic definition of protein topology.

The organization and classification of the bewildering variety of protein structure has been approached using clustering methods. Various computer programs have been devised to measure the three-dimensional similarity of one protein coordinate set to another[3]. From these measures, similar proteins can be grouped together, given a name, and arranged in a hierarchical clustering with others that share some partial or overall similarity. Depending on the method of comparison, or the extent of expert judgements needed, differing classifications of protein structure have emerged, ranging from one that is almost completely expert-based[4], through partially automated methods[5,6] to an almost fully automatic method[7]. The drawback of these hierarchical approaches is that, although the close relationships between similar proteins are reasonably well defined, the more tentative relationships that give the large-scale structure to the hierarchy are usually beyond the ability of the computer programs to recognize and are subject to variation when defined by experts. As a result, the classifications tend to have numerous small clusters (families of super families) all roughly grouped into just a few categories that are based on overall secondary structure content and arrangement.

An important secondary problem in the classification of proteins is how large proteins should be divided into pieces (domains) that can be classified more easily. The current approach among the more automated methods is first to divide and then to classify. However, it is clear from the expert-based approach that the initial process of classification can affect how the protein is then broken into domains and so differences in domain definition are a major source of inconsistency among the current classification systems[8].

To avoid the problems associated with a hierarchy, the method outlined here is based on a set of idealized structures that are compared with all known structures. (The programs and data described in this work can be found at: http://mathbio.nimr.mrc.ac.uk/ftp/wtaylor.) The domain definition problem is less directly solved, although as the ideal structures are all of domain size, the best match can be used to define (or bias) the definition of the domains. This approach is unusual in that it shifts the classification from a clustering problem to that of finding the best set of ideal structures that can account for as much protein structure as possible. As the ideal structures will be generated from rules applied to basic 'Forms', this can be viewed as finding a minimum basis set of generating Forms. These Forms were derived from a model in

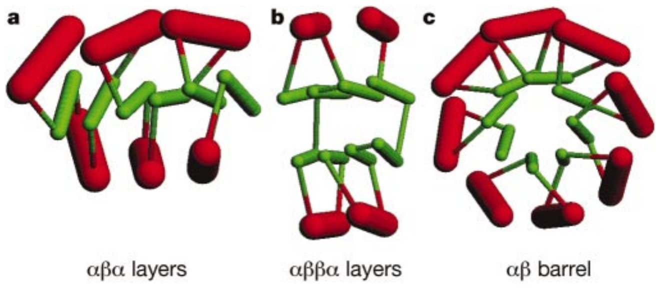

Figure 1 Stick-figure representations of the basic Forms. Each of the basic generating Forms is represented by 'stick' models in which α-helices are red and drawn thicker than the green β-strands. **a**, αβα layers. Six strands are shown, but the sheet can extend indefinitely. **b**, αββα layers. As in **a**, the sheets can be extended. (Removal of the α-layers leaves the common β-'sandwich'). **c**, Eight-fold αβ barrel. Similar barrels with 5–9 strands were constructed. (See Supplementary Information A.1 for construction details). By deleting helices and strands from these models, almost all known globular protein domains of β and βα types can be generated. Figures 1 and 4 were prepared using the program RasMol (http://www.umass.edu/microbio/rasmol).

which the hydrogen-bonded links across a β-sheet impose a layer structure onto the arrangement of secondary structures in a protein domain (Fig. 1)[9,10]. These layers can consist of either α-structure (packed α-helices) or β-structure (hydrogen-bonded β-strands). There are seldom more than four layers in any one domain and each layer tends to be exclusively composed of one of these two types of secondary structure. The spacing between the axes of packed α-helices is typically 10 Å, as is the spacing between β-sheets and between helices and sheets[11,12], whereas the spacing between the hydrogen-bonded strands in a sheet is close to 5 Å. This makes 10 Å a convenient unit with which to 'digitize' protein structure.

β-sheets normally have a twist, and the whole structure follows this twist, resulting in a staggered arrangement for the secondary structure elements in the outer layers. Sheets can also curl and incorporate a stagger between adjacent strands—combinations of these 'distortions' can result in the sheet forming a complete hydrogen-bonded cylinder (referred to as a barrel)[13]. All these parameters (twist, curl, stagger) might be varied with different numbers of layers (of different composition), but as a simple beginning, the more limited layer combinations shown in Fig. 2 were considered and the curl and stagger co-varied to give either topologically flat sheets or cylindrical sheets (barrels), with the curled sheets being represented by partial barrels. The resulting arrangement is not dissimilar to the periodic table of elements. In this loose analogy, the layers are equivalent to valence shells that become progressively filled with electrons (secondary structure elements): first, the inner β-layer (s orbitals) followed by the outer α layers (p orbitals) then repeating with a second β-layer (d orbitals). Extending the model to incorporate the permutations arising from additional α-layers would be reminiscent of the interjection of the rare-earth series into the periodic table. (See Supplementary Information A.3 for further discussion). Because there is no strict energetic difference between different protein structures, the filling of the layers with secondary structures was allowed more freedom than their electronic counterparts and, so for each of the basic Forms described above, any combination of secondary structures can be removed. For example; an eight-stranded barrel with eight helices around it can give rise to the variations shown in Fig. 3.

In order to match all ideal structures against all known protein structures, each native structure was first reduced to linear segments[14]. For each comparison, a fast bipartite graph-matching algorithm[21] was used as a pre-filter and also to 'prime' a more exhaustive double-dynamic programming comparison algorithm[15]. Each comparison began by pairing up line segments ('sticks') irrespective of their length or direction, but when a good fit was found, the lengths of the native protein 'sticks' were set to 10 Å and the connectivity of the ideal sticks were set to match the native protein. This allowed the two matching stick-structures to be directly superposed in three dimensions using a conventional comparison program and the quality of the match was taken as the root-mean-squared deviation (r.m.s.d.) between two sets of stick endpoints. From examination of the solutions, anything less than 5 Å r.m.s.d. between two structures is a good fit, and between 5 and 6 Å r.m.s.d. is acceptable. An example of two solutions in these ranges is shown in the Supplementary Information A.1.

The method was initially tested on the single domains defined to have distinct folds in the CATH database[5]. From these, the all-α

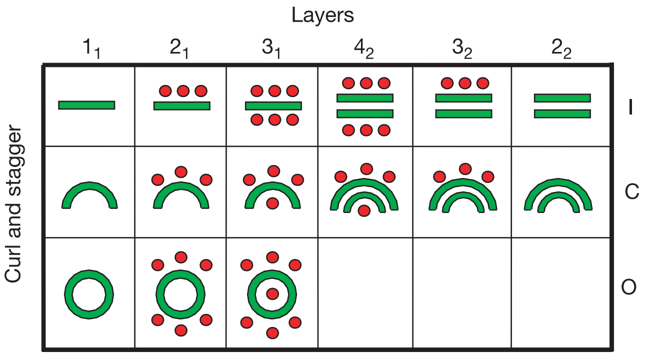

Figure 2 Simplified layer structure of proteins. Layers of secondary structure (β, green; α, red) are combined to make globular protein domains. The β-sheets are represented as bars and circles, as they would appear when viewed looking along their component strands. Each sheet has a left-handed twist between the strands (not depicted) onto which can be added curl and stagger. This allows the sheets to progressively 'deform' from a topologically flat sheet into a cylinder (or barrel). The two endpoints and one intermediate stage are represented by the rows in the figure and indexed as I (flat), C (curled) and O (barrel). For each of these, up to four layers of secondary structure are shown (α-helices are drawn as red dots viewed end-on). For simplicity, not all possible layer combinations have been represented; in particular, those with adjacent α-layers have been omitted because the boundary between these is not well defined. Most biologically important structures can be generated from three of those represented above: this set of three is referred to as the basis set. (See Fig. 1 for three-dimensional 'stick' figures of the basis set.) Using the I, C, O index plus layer number: $I3_1$ can generate $I2_1$ and $I1_1$ (by the deletion of helices); $O2_1$ can generate $O1_1$, and with the removal of strands from the barrel, also $C2_1$ and $C1_1$. Similarly, $I4_2$ can produce an $\alpha\beta\beta$ and the common $\beta\beta$ layer structures. Of those remaining, only $C3_1$ is biologically important and will be reconsidered later. All possible deletions are made for each basic structure: those for O2 are shown in Fig. 3.

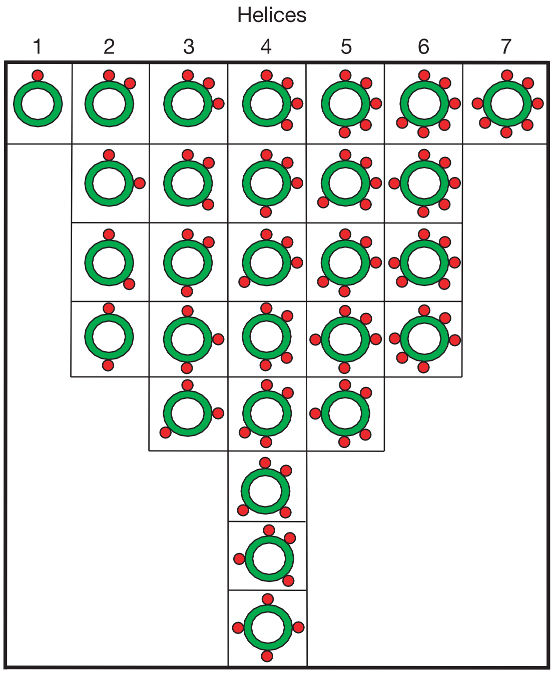

Figure 3 β-barrel helix packing variations. The basic barrel Form (O2 in Fig. 2), with eight strands and eight helices, can generate 30 packing variations through the deletion of helices (28 depicted above plus the two not shown with eight and zero helices). A similar combinatorial enumeration of variants was made for barrels of 5–9 strands and flat sheets of 3–13 strands. The intermediate curled strands (row C in Fig. 2) were made by successive deletion of all but three strands from each of the barrels (and the helices similarly permuted). The application of symmetry considerations greatly reduced the number of possible combinations, but for some larger structures it was necessary to impose a limit. Variations on the large 'flat' layers were ranked by compactness and (for alphabetic reasons) just the 26 most compact combinations were considered for matching. Variations on the larger barrels were also limited by allowing only one break in the sheet layer and similarly restricted to 26 variants on any given combination of secondary structures. In total, 12,640 variations were generated.

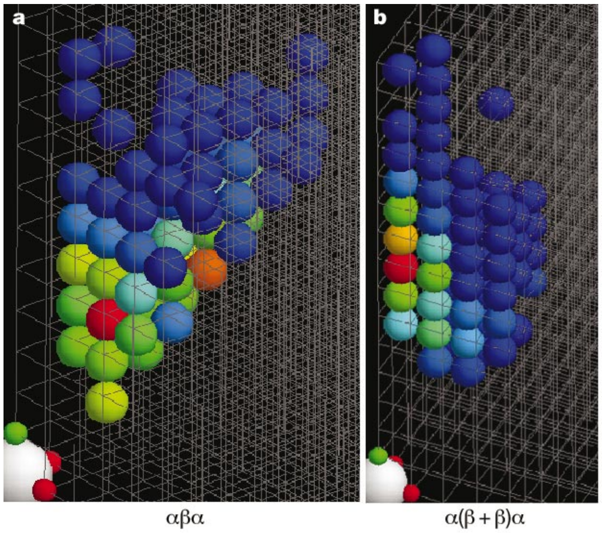

Figure 4 Structure tables for flat sheets. The number of helices and strands in each secondary structure layer of the ideal Forms is represented as a grid: **a**, αβα; and **b**, αββα. The large white sphere marks the origin of the grid with the directions of increasing β and α indicated by small green and red attached spheres. (For the αββα structures, the number of β-strands in each sheet have been combined). Grid cells contain a sphere if their corresponding ideal structure constitutes the best match to a native protein domain. This is coloured to indicate the number of proteins that match (red, most; blue, least). In **a**, the most populated cell has the Form 0-4-2 (α-β-α) with 48 members, closely followed by the 2-5-3 Form (42 members). The former included 25 different topologies while the latter includes only 18 (being dominated by the common flavodoxin fold). The most populated cell in the αββα table (83 members) has seven strands and no helices, corresponding to the 3-on-4 β-sandwich of the immunoglobulin fold. The table for the barrel structures (not shown) has a single highly populated cell for the $(\beta\alpha)_8$ barrel (8-8.8).

domains were excluded, as were integral membrane proteins. Although ideal structures can be described for these classes[16,17], a comprehensive nomenclature has not been devised. Some small domains that do not have packed secondary structure were also excluded, leaving a sample of 418 domains. Of these domains, only 34 had no acceptable fit to an instance of the ideal Forms, and most of these were small, with fragmentary beta structure. Of the remaining 384 domains 85% of the matches accounted for over 50% of the structure with more than 70% of the structure being matched in over half of the domains. Partially matched domains comprised large β-propellor structures and viral coat proteins that have long 'unstructured' loops. The remainder was typically composed of α-helices which were either packed in a distinct subdomain or suggested an additional α-layer. Rather than adapt the current model to account for these elements, it was clear that it would first be necessary to reassess the definition of domains. This was done using an automatic (Ising-like) method[18] in which it was possible to bias the matched portion of a protein to remain distinct from the rest. Any remaining material was then presented again for matching to the ideal structures and the process repeated. The full chains from a non-redundant set of 2,230 protein structures were then processed by this algorithm, resulting in an improved coverage by the ideal Forms which, on average, now accounted for 80% of each protein, with 70% of the structure being matched in 75% of the domains (previously just over 50%).

Each Form can be indexed by the number of secondary structures in each layer (see Supplementary Information A.2 for details) and from the index of its matching Form, each domain can be allocated a grid reference and plotted in space. Any step in the planar-β grids represents the addition or removal of a secondary structure, and one of the dimensions in the barrel grid represents the opening of the barrel (for a fixed number of strands). Using this representation, the full range of known protein structure can be visualized (Fig. 4).

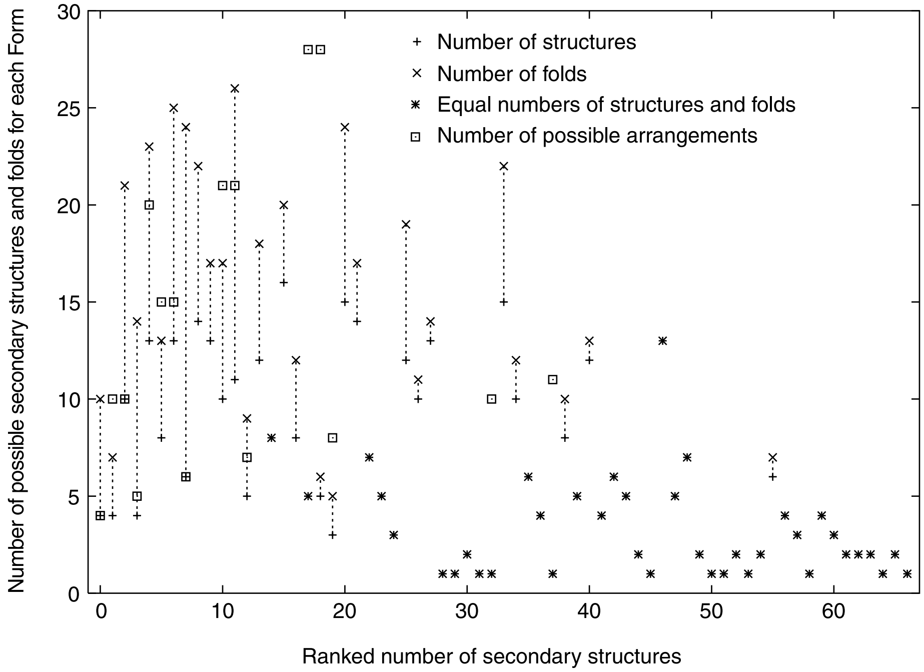

Figure 5 Fold numbers for ranked Forms. The three-layer Forms (Fig. 4a) were ranked by the number of secondary structures they contain. Those of equal size were ordered by the number of β, then α elements. Against this is plotted the number of different secondary structure arrangements seen for each Form (upright crosses) and the number of different folds (tilted crosses). (See Supplementary Information A.2 for details). The number of observed arrangements and the number of folds are linked by a dashed line and when they are equal, appear as an asterisk. For reference, the number of possible arrangements is plotted as a box but this only appears for some of the smaller Forms as, for most, it is a very large number.

This classification of structure has not explicitly considered topology but has concentrated on the prerequisite of defining the secondary and tertiary links. From this basis, the previously difficult issue of topology becomes almost trivial, because two proteins matching the same ideal Form will either have the same connection of their secondary structures or will not. (See Supplementary Information A.2 for nomenclature details). Topology descriptions were written for each protein in each grid-cell and sorted for uniqueness, thus giving the number of folds in each cell. This value was then plotted along with the number of unique secondary structure strings for the three-layer Forms (Fig. 5). This simple analysis shows that for the smallest Forms, all linear arrangements of secondary structure have been observed and there are typically twice as many folds as secondary structure arrangements. With over ten secondary structures (around 35 in the ranked Forms in Fig. 5), this balance changes, and almost every secondary structure arrangement corresponds to a unique fold. This suggests that with these larger βα proteins, sequence-structure matching (threading) methods[19] need only concentrate on the correct prediction of secondary structure to find an unknown fold and not on their three-dimensional interaction.

If a structure matches, say, a 2-5-3 (α-β-α) Form then it will also match any substructure at least as well (2-5-2, 1-5-3, 2-4-2 . . .) and two proteins which have a different topology when matched against the 2-5-3 Form might have the same core 2-5-2 topology. The question of whether two proteins have the same fold now becomes relative and should be posed thus: what is the largest ideal Form under which two proteins have the same fold? Although the largest Form with common topology will be of greatest interest, it is also informative to view all the subsidiary matches in the format of the grids used in Fig. 4. Such a representation allows not only pairs, but also whole groups of proteins to be analysed and it is simple to determine both visually and automatically if they share a common core. Each step through the grid from one protein to another (via subsidiary matches) represents a deletion or addition of a secondary structure element. This allows a path to be defined between any two structures in the same grid (no matter how dissimilar) and a minimum path length to be computed.

Besides the all-α class of protein, the only structures not 'captured' by the set of three basic Forms used above were β proteins that contain internal repetition. These included not only the series of propellor proteins but also those with a triangular arrangement of structure, including β-trefoils, β-prisms and β-helices[20]. Clearly triangles do not map well onto layers (or cylinders) and the only solution for these may be to generate a 'triangular' version of Fig. 2, or else to treat them as exceptions because of their detectable internal sequence repeats. For the moment, the latter route will be followed but a series of propellor Forms will be added to the basis set, along with the neglected $C3_1$ Form from Fig. 2 (which includes the β-grasp structures).

Thus, by breaking down structures into their basic Forms I have opened up a new approach to the classification and analysis of protein structure that is both flexible and automatic. Unlike clustering methods, it does not require the comparison of all structures to each other and can therefore be incrementally updated as new structures are determined. The major result from this approach is that it allows the difficult problem of protein topology to be rigorously addressed. The fold of any individual protein can be specified as that of the largest matching Form, and for pairs or groups of proteins, the fold is that of the largest common Form. This implies that the topology of a protein can only be defined under the specification of a given ideal Form which, in turn, means that proteins do not have a unique intrinsic topology or a unique position in any classification that is based on topology. □

Received 12 October 2001; accepted 28 January 2002.

1. Chothia, C. One thousand families for the molecular biologist. *Nature* **357,** 543–544 (1992).
2. Orengo, C. A., Jones, D. T. & Thornton, J. Protein superfamilies and domain superfolds. *Nature* **372,** 631–634 (1994).
3. Eidhammer, I., Jonassen, I. & Taylor, W. R. Structure comparison and structure patterns. *J. Comput. Biol.* **7,** 658–716 (2000).
4. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536–540 (1995).
5. Orengo, C. A. *et al.* CATH—a hierarchic classification of protein domain structures. *Structure* **5,** 1093–1108 (1997).
6. Mizuguchi, K., Deane, C. M., Blundell, T. L. & Overington, J. P. HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* **7,** 2469–2471 (1998).
7. Holm, L. & Sander, C. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* **25,** 231–234 (1997).
8. Hadley, C. & Jones, D. T. A systematic comparison of protein structure classifications SCOP, CATH and FSSP. *Structure* **7,** 1099–1112 (1995).
9. Chothia, C. & Finkelstein, A. V. The classification and origins of protein folding patterns. *Ann. Rev. Biochem.* **59,** 1007–1039 (1990).
10. Finkelstein, A. V. & Ptitsyn, O. B. Why do globular proteins fit the limited set of folding patterns?*Prog. Biophys. Mol. Biol.* **50,** 171–190 (1987).
11. Cohen, F. E., Sternberg, M. J. E. & Taylor, W. R. Analysis and prediction of protein β-sheet structures by a combinatorial approach. *Nature* **285,** 378–382 (1980).
12. Cohen, F. E., Sternberg, M. J. E. & Taylor, W. R. Analysis and prediction of the packing of α-helices against a β-sheet in the tertiary structure of globular proteins. *J. Mol. Biol.* **156,** 821–862 (1982).
13. Murzin, A. G., Lesk, A. M. & Chothia, C. Principles determining the structure of β-sheet barrels in proteins: I, A theoretical analysis. *J. Mol. Biol.* **236,** 1396–1381 (1994).
14. Taylor, W. R. Defining linear segments in protein structure. *J. Mol. Biol.* **310,** 1135–1150 (2001).
15. Taylor, W. R. Searching for the ideal forms of proteins. *Biochem. Soc. Trans.* **28,** 264–269 (2000).
16. Murzin, A. G. & Finkelstein, A. V. General architecture of the α-helical globule. *J. Mol. Biol.* **204,** 749–769 (1988).
17. Taylor, W. R., Jones, D. T. & Green, N. M. A method for α-helical integral membrane protein fold prediction. *Protein Struct. Funct. Genet.* **18,** 281–294 (1994).
18. Taylor, W. R. Protein structure domain identification. *Protein Eng.* **12,** 203–216 (1999).
19. Jones, D. T., Taylor, W. R. & Thornton, J. M. A new approach to protein fold recognition. *Nature* **385,** 86–89 (1992).
20. Chothia, C. & Murzin, A. G. New folds for all-β proteins. *Structure* **1,** 217–222 (1993).
21. Taylor, W. R. Protein structure comparison using bipartite graph matching and its application to protein structure classification. *Mol. Cell. Proteom.* advance online publication 4 March 2002 (DOI 10.1074/mcp.T200001-MCP200).

Supplementary Information accompanies the paper on *Nature's* website (http://www.nature.com).

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to the author (e-mail: wtaylor@nimr.mrc.ac.uk).

correction

Transition of Mount Etna lavas from a mantle-plume to an island-arc magmatic source

Pierre Schiano, Roberto Clocchiatti, Luisa Ottolini & Tiziana Busà

Nature **412**, 900–904 (2001).

We acknowledge that the source of the data for major-element composition of sample 27.2 in the Supplementary Information should have been cited as the work of Kamenetsky and Clocchiatti (ref. 1). We also acknowledge that the discussion of our model for the origin of Mount Etna magmatism would have benefited from referencing the model of multi-stage melting contemporaneous with metasomatic enrichment presented in ref. 1. □

1. Kamenetsky, V. & Clocchiatti, R. Primitive magmatism of Mt. Etna: insights from mineralogy and melt inclusions. *Earth Planet. Sci. Lett.* **142,** 553–572 (1996).