Getting past appearances: the many-fold consequences of remote homology

Olivier Lichtarge

In the absence of other biological information, the detection of remote homology is a prerequisite step toward understanding the function of a new protein. A novel method based on structure comparison improves our ability to do this automatically and systematically.

The benefit of large-scale, high-throughput approaches to address biological questions lies in the wealth of data they provide. Sequence and structure databases are growing exponentially, as are those for gene expression and protein–protein interactions. But the price we pay for this bounty is an abundance of sequences and structures without known functions, and patterns of expression and interactions with unclear physiological relevance. To recover the biological context required to interpret these data, the essential first step is nearly always to find a related protein for which biological information is available. Hence genome sequence annotation, protein structure prediction, and the elucidation of protein–protein interactions all stand to gain from improved methods to recognize remote homology — that is, common evolutionary origin. For these reasons, and given the accelerating pace at which new structures are being solved, the innovative, structure-based approach to remote homology detection proposed by Dietmann and Holm[1] on pages 953 of this issue of *Nature Structural Biology* is especially timely.

Protein homology is most often established by telltale amino acid identities between sequences. However, chance variations in the sequence invariably take place, even at important functional residues, such that over time nonconservative substitutions chisel away at the original sequence — so much so that eventually the sequence identity between divergent proteins is no longer significantly greater than that among random proteins. This threshold, near 20% sequence identity, defines the 'twilight zone', at which point the structure becomes essential for identifying homology. Even as protein sequences evolve beyond recognition, the core of their structure remains relatively spatially invariant[2]; therefore related proteins most likely have similar folds. The converse statement, that unrelated proteins have unrelated folds, is unfortunately not true — there are many fewer ways to pack α-helices and β-sheets tightly in three dimensions than there are proteins, so unrelated proteins often have duplicate folds[3]. How, then, is one to establish whether a common fold indicates divergence or chance? One approach is through detailed examination of each structure.

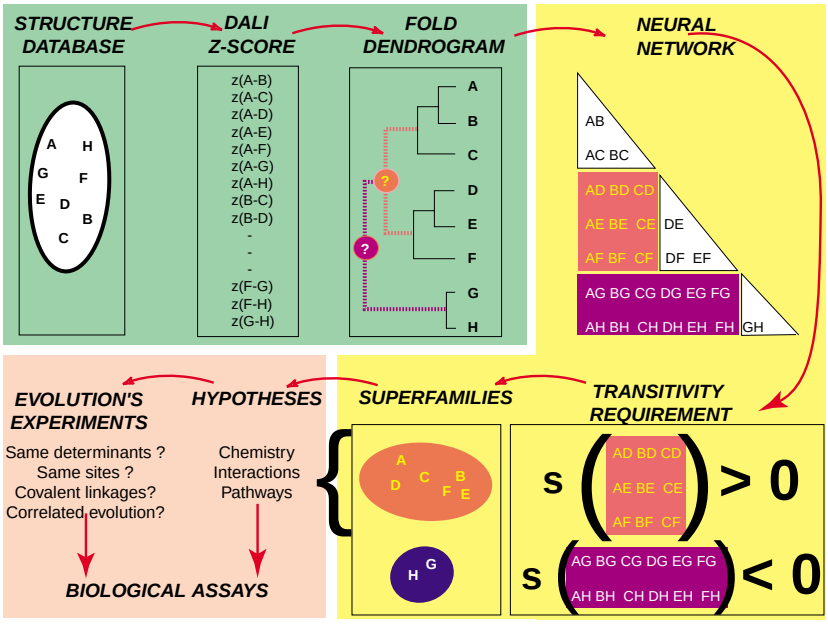

Fig. 1 From structure to function. This pathway illustrates a possible way to link structures produced by the Structural Genomics Initiative to biological functions. The green section shows how, until now, it was possible to sort various structures in a database (designated A to H) by computing a DALI Z-score[14], which is how much the similarity of two structures deviates from the expected mean similarity of unrelated proteins. This information could then be summarized in a fold dendrogram. Dietman and Holm[1] have now added two new steps, shown in the yellow panel. First, a neural network estimates the likelihood of homology between all pairs of proteins in the fold dendrogram. Second, the authors introduce a 'transitivity' requirement, function s, such that two branches of homologous structures may be joined into a single one only if all possible structure pairings between the two branches give rise to a positive value s. As shown here, s({A, B, C},{D, E, F}) is positive so these structures may all be joined into a single superfamily. On the other hand, the negative value of s when branch {G, H} is also considered suggests these two structures should be in their own superfamily. The implications of such an automated and systematic classification scheme is shown in the pink panel. If proteins A to F are remote homologs, a reasonable hypothesis is that they share some of their functional characteristics. This can be tested by direct experiment, or by analyzing the results of evolution's own experiments — for example, by determining which proteins in a homologous family share common functional sites[23–28,30], or by examining genomes for evidence that two distinct domains are on occasion covalently linked or that their evolution is correlated[20,21], suggesting a possible protein–protein interaction.

Structure-based classification

This approach is the premise behind the Structural Classification Of Proteins (SCOP)[4], a database curated by a few experts who sort proteins into a four-level hierarchy. The first level, Class, describes the type and distribution of secondary structural elements in each protein. The second level, Fold, specifies the connectivity and spatial relationship between these secondary structural elements. The third level, Superfamily, groups together proteins with sufficient functional or structural similarity to infer a common origin. The fourth level, Family, encompasses proteins whose sequence identity leaves

no doubt as to their homology and functional similarity.

Importantly, the structural and functional clues that help define SCOP superfamilies are subtle. For example, even though the sequence identity between ribosome anti-association factor IF6[5] and L-arginine:glycine amidinotransferase[6] is insignificant (8%), both proteins have a similar fold of five ββαβ repeats arranged in approximate five-fold symmetry[7]. These proteins were initially considered to belong to distinct superfamilies because they could at best be superposed over only half of one protein, due to insertions, and even then the root mean squared (r.m.s.) deviation remained relatively poor over the backbone atoms (3.5 Å) due to deviations from five-fold symmetry in both proteins[5]. The superposition of individual repeats, however, shows that the two helices and three strands pair up well, as do the three loops between them. Thus, a common pattern does emerge, supporting a joint classification into one SCOP superfamily, but only after both proteins are broken into smaller structural units and compared piecewise[7,8]. Such careful analysis depends on visual reasoning and establishes SCOP as the gold standard of remote homologies. Yet how long can a manual approach keep pace with high throughput structure determination?

To address this issue, two other well-known structural classifications resort to automation. One is CATH (each letter stands for one of its four hierarchical levels: Class, Architecture, Topology, and Homology)[9], in which algorithms define the boundaries of protein domains[10] and the program SSAP assigns them to homologous superfamilies[11] based on the similarities in the three-dimensional environment of residues. Other automated procedures are being added to speed up the intake of new structures[12]. Another popular classification scheme is FSSP[13], Families of Structurally Similar Proteins. FSSP is fully automated and unlike SCOP and CATH, makes no attempt to define any *a priori* hierarchy. Rather, it uses the program Dali[14] to measure structural similarity between proteins in terms of a Z-score (Fig. 1, green panel). The Z-score for a given pair of structures is the number of standard deviations by which their similarity exceeds the mean of the distribution of an all-against-all comparison of similarity among a selected set of structurally unrelated proteins. Z-scores can be used to construct a 'tree' of protein structures, with similar structures (having high Z-scores amongst themselves) clustered together in the terminal branches of the tree. High Z-scores increase the likelihood that two proteins had a common evolutionary origin; when sequence identity is greater than 25% and the Z-score is above 6, FSSP structure pairs match SCOP or CATH superfamilies nearly perfectly. However, the agreement falls to nearly half when identity is below 20%, and much less if the Z-score is below 6 (ref. 15). Thus, a fully automated classification of structures into divergent groups has remained elusive, until now.

Automated identification of homology

In their paper, Dietmann and Holm[1] introduce novel and objective measures to partition the tree of structures into superfamilies (Fig. 1, yellow panel). This is accomplished by way of two innovations. First, the authors build a neural network to estimate whether pairs of proteins in the tree are homologous. The network's input describes attributes of each protein's sequence, structure and even function when available, and the output is a score from 0 (no homology) to 1 (perfect homology). Second, they develop an objective criterion to decide when two branches of homologous sequences ought to be joined into a single superfamily at their common node. This is done only if the neural network indicates that most members of one branch are significantly homologous to members of the second branch. The rationale is that homology is transitive: two proteins that each shares ancestry with a third protein must be related. Although simple, this constraint proves powerful enough to improve the reliability of predictions compared to either the Dali Z-score or the neural network taken alone. The authors can retropredict with near perfect reliability almost three quarters of SCOP superfamilies, the visual inspection-based gold standard.

How is this useful for the functional annotation of novel structures? This is an important question since the function of most structures to come out of the Structural Genomics Initiative will be unknown[16]. Dietmann and Holm test this in two ways. First, they remove the functional information from the neural network input and observe that this degrades the prediction of SCOP superfamilies only slightly (6%). Second, they apply their method to 15 Structural Genomics targets whose structures have been determined recently. Four are entirely new folds and three others are known folds but with little evidence of an evolutionary link to known superfamilies. The other eight fall within well-characterized superfamilies, and thereby can be associated with plausible biochemical functions. For example, MTH152, a protein that was solved as part of a structural proteomics effort in *M. thermoautotrophicum* has now been associated with the activity of a ferric reductase (1I0R). This link remains to be tested experimentally.

From homology to function

The biological implications of such evolutionary linkages in terms of protein chemistry must be interpreted with care, however. Database studies show that the biochemical activities of homologues with greater than 40% sequence identity are nearly identical, and that broad functional classes are often preserved down to 25% identity[17]. Nevertheless, different members of a single superfamily can support entirely different biochemical activities. For example, 25% of enzyme superfamilies contain proteins that carry out unrelated functions, possibly through entirely different chemistries[18] at different functional surfaces[19].

Similarly, the possible interactions hinted at by a protein's history must be evaluated critically. Clearly, knowledge of a protein domain's superfamily may suggest possible protein–protein interactions. For example, the fact that two domains are part of a single polypeptide chain in one genome has predictive value for their possible interaction in genomes where they are not covalently linked[20,21]. Moreover, such putative coupling hypotheses can be fairly specific, since, overall, 91% of domain superfamilies are linked in the same polypeptide chain to no more than two other superfamilies. Yet, some domains are known to recombine promiscuously with different partners in different genomes[22], and the number of plausible couplings for any one domain is bound to increase as more genomes are sequenced.

Thus, determining to which superfamily a novel protein (the query) belongs is only the first step in a series of hypotheses that ultimately aim to focus experiments on its most likely functions (Fig. 1, pink panel). When the query belongs to a functionally characterized family, a number of possible functions and binding interactions may be suggested, each of which will each require testing. But as an increasing number of structures are solved and superfamilies grow larger, one must ask whether enough sensitive biochemical assays will be designed for every reasonable hypothesis in every single superfamily. This large experimental burden could be reduced by further computational pre-filtering. One

approach is to determine the location of the active sites in the various members of a superfamily[23–28] and then check whether they correspond to the same location and preserve the character and geometry of the key residues associated with one of these functions. This strategy is able to distinguish between DNA binding domains of nuclear hormone receptors that homodimerize head-to-head over palindromic response elements from those that dimerize head-to-tail over repeated response elements[29]. A recent study suggests the viability of this strategy on a large scale[30].

Until these additional computational and experimental studies can be performed on a genomic scale, the danger exists that functional hypotheses driven by the recognition of remote homology will be taken at face value. If so, erroneous functional annotations are bound to occur and, in turn, propagate throughout databases as they become the sources for further false hypotheses[31]. This is one reason why manual oversight of protein homology classification will remain essential, not only as a gold standard but also to minimize automated error propagation. At the same time, the automation of this process is critical to cope with a massive inflow of data and to shed light on the biases and inconsistencies that humans will inevitably introduce. Dietman and Holm[1] make an important step in this direction.

Olivier Lichtarge is in the Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA. email: lichtarge@bcm.tmc.edu

1. Dietman, S. & Holm, L. *Nature Struct. Biol.* **8**, 953–957(2001).
2. Chothia, C. & Lesk, A.M. *EMBO J.* **5**, 823–826 (1986).
3. Chothia, C. Proteins. *Nature* **357**, 543–544 (1992).
4. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. *J. Mol. Biol.* **247**, 536–540 (1995).
5. Groft, C.M., Beckmann, R., Sali, A. & Burley, S.K. *Nature Struct. Biol.* **7**, 1156–1164 (2000).
6. Humm, A., Fritsche, E., Steinbacher, S. & Huber, R. *EMBO J.* **16**, 3373–3385 (1997).
7. Paoli, M. *Nature Struct. Biol.* **8**, 744 (2001).
8. Teichmann, S.A., Murzin, A.G. & Chothia, C. *Curr. Opin. Struct. Biol.* **11**, 354–363 (2001).
9. Orengo, C.A. *et al. Structure* **5**, 1093–1108 (1997).
10. Jones, S. *et al. Protein Sci.* **7**, 233–242 (1998).
11. Taylor, W.R. & Orengo, C.A. *Protein Eng.* **2**, 505–519 (1989).
12. Pearl, F.M. *et al. Nucleic Acids Res.* **29**, 223–227 (2001).
13. Holm, L., Ouzounis, C., Sander, C., Tuparev, G. & Vriend, G. *Protein Sci.* **1**, 1691–1698 (1992).
14. Holm, L. & Sander, C. *J. Mol. Biol.* **233**, 123–138 (1993).
15. Hadley, C. & Jones, D.T. *Structure Fold. Des.* **7**, 1099–1112 (1999).
16. Montelione, G.T. & Anderson, S. *Nature Struct. Biol.* **6**, 11–122 (1999).
17. Wilson, C.A., Kreychman, J. & Gerstein, M. *J. Mol. Biol.* **297**, 233–249 (2000).
18. Todd, A.E., Orengo, C.A. & Thornton, J.M. *J. Mol. Biol.* **307**, 1113–1143 (2001).
19. Russell, R.B., Sasieni, P.D. & Sternberg, M.J. *J. Mol. Biol.* **282**, 903–918 (1998).
20. Enright, A.J., Iliopoulos, I., Kyrpides, N.C. & Ouzounis, C.A. *Nature* **402**, 86–90 (1999).
21. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. & Eisenberg, D. *Nature* **402**, 83–86 (1999).
22. Apic, G., Gough, J. & Teichmann, S.A. *J. Mol. Biol.* **310**, 311–325 (2001).
23. Sowa, M.E. *et al. Nature Struct. Biol.* **8**, 234–237 (2001).
24. Lichtarge, O., Bourne, H.R. & Cohen, F.E. *J. Mol. Biol.* **257**, 342–358 (1996).
25. Casari, G., Sander, C. & Valencia, A. *Nature Struct. Biol.* **2**, 171–178 (1995).
26. Jones, S. & Thornton, J.M. *J. Mol. Biol.* **272**, 133–143 (1997).
27. Olmea, O. & Valencia, A. *Folding Des.* **2**, S25–32 (1997).
28. Landgraf, R., Xenarios, I. & Eisenberg, D. *J. Mol. Biol.* **307**, 1487–1502 (2001).
29. Lichtarge, O., Yamamoto, K.R. & Cohen, F.E. *J. Mol. Biol.* **274**, 325–337 (1997).
30. Aloy, P., Querol, E., Aviles, F.X. & Sternberg, M.J. *J Mol Biol* **311**, 395–408 (2001).
31. Brenner, S.E. *Trends Genet.* **15**, 132–133 (1999).

A snapshot of Nature's favorite pump

Philip J. Thomas and John F. Hunt

The 4.5 Å map of the MsbA protein, a putative lipid A transporter from *Escherichia coli*, provides the first detailed structural model for the transmembrane domain and cytoplasmic 'loops' of an ABC transporter and the geometric relationship of these regions to the ATP-binding cassette motor domain. Based on this structure, specific hypotheses for the mechanics of the pump can now be formulated and tested.

Nature has evolved a variety of protein machines to convert the chemical potential energy present in ATP into the osmotic work of moving solutes from one side of the lipid bilayer to the other. ATP-binding cassette (or ABC) transporters represent the most common and ancient solution to this problem. This class of proteins forms the largest gene superfamily in many of the completely sequenced microbial genomes[1] and contains a number of members whose function or dysfunction is central to serious human pathologies such as cystic fibrosis, hypercholesterolemia, adrenoleukodystrophy, Stargardt's disease and multidrug resistance[2]. Unlike other ATP-driven pumps, such as the F-, V- and P-type ATPases, which catalyze the ATP-driven transport of cations exclusively, the ABC transporters couple ATP hydrolysis to the movement of a staggeringly diverse set of solutes, including large proteins, peptides, lipids, pigments, sugars, amino acids and anions. In addition to being fundamentally interesting, the ABC transporters play a central role in bacterial virulence and human genetic disorders, thereby motivating the practical need for a structural understanding of their functional mechanism.

As reported in a recent issue of *Science*, Chang and Roth[3] provided a major milestone in this effort by characterizing the crystal structure of a complete ABC transporter, the MsbA pump from *Escherichia coli*, at 4.5 Å resolution. MsbA is hypothesized to function in flipping lipid A — and possibly other anionic lipids — across the cytoplasmic membrane; this hypothesis is partially based on the observation that MsbA mutants cease to transport lipids between the inner and outer membranes at nonpermissive temperatures[4].

Structure of the ABC transporters

The MsbA structure provides important insight into the three-dimensional organization of all of the ABC transporters, as they share a common architecture comprising two ATP-binding cassettes (ABC's) and two transmembrane (TM) domains, each with 6–8 α-helices connected by a set of large cytoplasmic 'loops'[1–3]. The conservation of the ABC sequences is striking in that transporters as diverse as MsbA and the human cystic fibrosis transmembrane conductance regulator (CFTR) share 27% identity and 54% similarity in their cassettes. The conservation of the TM domains is weaker but still significant, as MsbA and CFTR share 18% identity and 38% similarity in the first