Protein structures, folds and fold spaces

2010 J. Phys.: Condens. Matter 22 033103

(http://iopscience.iop.org/0953-8984/22/3/033103)

IOP Publishing Journal of Physics: Condensed Matter

J. Phys.: Condens. Matter **22** (2010) 033103 (8pp) doi:10.1088/0953-8984/22/3/033103

TOPICAL REVIEW

Protein structures, folds and fold spaces

Michael I Sadowski and William R Taylor

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

Received 17 September 2009
Published 21 December 2009
Online at stacks.iop.org/JPhysCM/22/033103

Abstract

There has been considerable progress towards the goal of understanding the space of possible tertiary structures adopted by proteins. Despite a greatly increased rate of structure determination and a deliberate strategy of sequencing proteins expected to be very different from those already known, it is now rare to see a genuinely new fold, leading to the conclusion that we have seen the majority of natural structural types. The increase in knowledge has also led to a critical examination of traditional fold-based classifications and their meaning for evolution and protein structures. We review these issues and discuss possible solutions.

Contents

1. Introduction

Proteins are the crucial link between the processes of information and replication that take place on a genetic level and the infrastructure of living things. The spontaneous formation of specific three-dimensional structures by amino acid sequences is the property which underpins the versatility of function which has allowed proteins to become essential components of all processes of living organisms. Although it is not necessary for proteins to adopt stable structures to be functional (see [1] for a review on disordered proteins) it seems that the majority do so. It is important to understand why this is and what the options are.

The importance of tertiary structure is related to the need for specificity and efficiency in functions such as catalysis and the propensity of proteins to aggregate, which if left unchecked has serious pathological consequences [2, 3]. Since proteins present a large number of potentially reactive groups the main issue when using them for cellular processes is to prevent undesirable side-reactions with small molecules and undesirable interactions with other proteins and macromolecules whilst maintaining access to functional parts of the protein. Structure is also required to ensure that the functional residues of the protein are precisely positioned, leading to efficient function.

Evolved proteins solve these problems by adopting specific chain configurations in their native conditions, their tertiary structures. These structures are intimately related to the molecular function of each protein, yet in general there is comparatively little functional signal in protein structure [4] except what can be explained as evolutionary conservation. Once a particular structure–function pairing is found it appears to be highly conserved, with new functions being adopted by mechanisms which may require duplication, functional redundancy and possibly multifunctionality [5]. Since structural stability, as a requirement for functional efficiency, is the most significant constraint on protein sequence evolution [6], even very subtle sequence conservation is strongly predictive of structural similarity [7, 8]. Understanding protein structure is therefore essential for understanding the mechanism behind the functions of particular proteins as well as how they can evolve, which makes protein structures of central importance to much of molecular biology.

In recognition of this there has been a great deal of effort towards the goal of understanding protein structure, how it underpins protein function and how it directs the evolutionary process on a sequence level. This would provide substantial benefits for almost every area of biology, much as the application of crystallography to biological structures

has provided substantial insights into both biochemistry and genetics [9]. However, although there has been substantial progress and a great increase in knowledge there remain many questions and difficulties.

In this review we will assess the current state of knowledge of protein structure with regard to the question of classifying structural types and enumerating the structures available to proteins and those which have been chosen for use by living organisms. We start by summarizing the basics of protein structure before discussing comparison and classification of structures and some of the conceptual and practical issues arising. We will then consider the question of the abstract space of possible protein structures ('fold space') and what can be said about it. The final sections then discuss how this space looks to the evolutionary process by considering how sequences, structures and functions interrelate.

2. An overview of protein structure

The compact, relatively complex three-dimensional structures (tertiary structures) formed by proteins are characterized by the formation of extensive networks of hydrogen bonds between peptide groups, a densely packed hydrophobic core and a hydrophilic surface. The protein chain is a linear polymer of 2 amino acids formed by condensation of the amino group of one monomer with the succeeding group's carboxyl moiety to form a peptide unit. The peptide bond is partially aromatic, which constrains the torsional angle across the bond (known as the omega torsion angle) to two values: 0 and pi radians. The succeeding two torsions, the phi and psi angles, rotate freely but experience a number of steric constraints which limit their joint conformation. Using a simple hard-sphere potential function the density of occupation of phi/psi pairs by peptide units can be predicted with substantial accuracy; such density plots (Ramachandran plots) [10] are amongst the most important standard tools for determining protein model quality [11–13].

The resonance of the carbonyl oxygen and amide nitrogen makes them excellent hydrogen bond donors and acceptors. Consequently it is preferable to expose these groups to a polar environment such as the aqueous solvent. If the chain is to become compact (and hence bury substantial numbers of these groups) it is necessary for them to find hydrogen-bonding partners from within the chain. This leads to the formation of repeating structures known as secondary structures, which account for a large proportion of the structure of most proteins [14, 15].

The nature of the two structures depends on whether these bonds are made locally (to amino acids nearby in the sequence) or non-locally. The main locally bonded structure is the alpha helix, a minority of local structures are turns in the chain. Non-local bonds result in extended beta structures which in turn generate hairpins and sheets, the details of which are a consequence of the steric constraints described above and the precise geometric requirements for hydrogen bonding.

The tertiary structure of a protein can be described by the packing of these secondary structure elements together to form a compact structure [16–18]. All naturally occurring proteins of above a certain size form a well-packed hydrophobic core [19], which suggests that in the majority of evolved proteins it is the hydrophobic effect that is likely to be the dominant driver for folding—the reduction in chain entropy experienced by the protein being counteracted by the gain in solvent entropy achieved by burying nonpolar surface area [20].

The refolding experiments of Anfinsen in the 1960s established that the protein sequence is sufficient to specify the structure of the protein [21]. This finding motivates both the study of protein structure prediction and protein folding—the former since the protein's sequence must be sufficient to predict its structure and the latter since proteins must be able to find this structure quickly, which requires a non-random process [22]. Both problems have proven extremely difficult and remain open, but fortunately for structural biologists nature has provided a shortcut: as noted above the conservation of structure means that statistically significant sequence similarity implies structural similarity. On this basis it might be possible to approximately cover all of protein structure space by sampling from each unique structural group. Whether this is possible and how easy it is depends on several features of the space, which we will discuss in section 3.

3. Protein structure classification

The principal aim in classifying protein structures is to provide a high-level view in which the dynamic details of the structure and conformational changes undergone during functional motions of the proteins are omitted. Not only does this serve as a useful aid to human understanding but it also provides a means to assess structural coverage and address questions about the nature of the interplay between physics and evolutionary contingency in protein structure. The category produced by the classification is usually known as the 'fold' of the protein, the overall goal being to determine which natural groups exist in order to provide some structure to analyses of the progress of our knowledge of protein structure as well as evolutionary questions.

Although the notion of a 'fold' as a large-scale description of structure is a useful one it has proven surprisingly difficult to pin down the concept, a fact which is immediately obvious from the existence of two authoritative and widely-used classifications of protein structure: CATH and SCOP [23, 24]. The method in both cases is very similar: sequences are grouped into homologous superfamilies, superfamilies are structurally clustered into fold groups and these are also grouped into larger-scale classes defined on the basis of their content of secondary structure elements. There are some minor differences in structure: SCOP defines more classes (proteins containing alpha and beta structure are divided into two groups), CATH inserts an extra 'architecture' level above the fold level (which it refers to as 'topology'), examples of architectures being three- or four-layer sandwich structures, up–down helical bundles and beta–alpha barrels (a few examples are shown in figure 1). Overall, however, there are more similarities than differences [25–27].

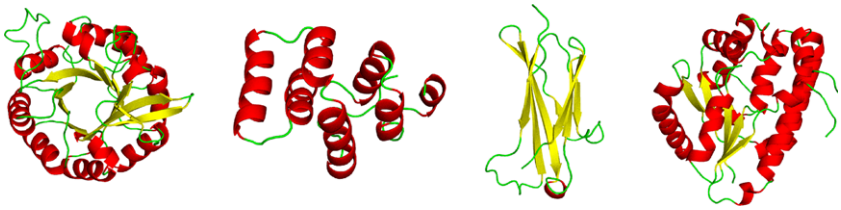

Figure 1. Examples of protein structural diversity. Four distinct protein structural types are shown. From left to right: HISF from *T. maritima*, an alpha/beta barrel; a DEATH domain from *Drosophila*, an all-alpha orthogonal bundle; an immunoglobulin from mouse, representing a two-layer all-beta sandwich domain; MOBB from *E. coli*, a three-layer alpha/beta/alpha sandwich with a Rossmann-like topology. Figure created with Pymol [28].

(This figure is in colour only in the electronic version)

The most significant contrasts between the two relate to a more subtle difference in emphasis: CATH emphasizes automatic classification on structural grounds while SCOP emphasizes manual classification and preservation of evolutionary relationships and functional similarities. This is important as it effects the way that proteins are split up into their constituent domains: SCOP prefers to maintain frequently occurring units even if they can be seen to be further decomposable into domains (e.g. the periplasmic-binding-like proteins, SCOP c.94). In contrast CATH may split functional units, leading to a set of similar structures that are not necessarily discrete functional units [25].

As a consequence of attempting to make both classifications reflect evolutionary similarity and structural similarity there may be substantial variation within particular levels [29, 30], with some families being more diverse than certain folds. This is a consequence both of the means of generating the groups (discussed below) and the underlying natural variation [31, 32].

The basis of both major classifications of structure is the comparison of two structures. To derive a classification by these means therefore requires a way to generate a score by comparison of structures and some way of scaling the score to make it objectively meaningful. This is not always an entirely straightforward procedure. We discuss this in section 4.

4. Protein structure comparison

Comparison of structures has been extensively studied, with a large number of possibilities having been explored from simple geometric methods [33–37], faster graph-based methods [38–40], machine-learning approaches [41] and topologically motivated methods [42, 43]. The core idea of most of these methods is simply to identify a set of equivalent positions in the two structures and then find the optimal rotation and translation to minimize some function of their geometric similarity, usually the RMS deviation between the two sets of points. We briefly consider this; more issues are discussed in a recent review [44].

Several fast methods have been developed for finding the optimal superposition given a set of equivalences [45–50]; alignment methods differ in their choice of which to use but apart from the overall speed and possible errors arising from numerical instabilities this makes no difference to the result. The difficulty comes from the need to doubly optimize: the method must find both the optimal superposition for the equivalences and the optimal set of equivalences.

If no constraints are set on the number of equivalences it can trivially be made equal to zero by comparing only two residues. It is therefore necessary to find the relationship between RMSD and the size of the subset chosen for significant similarities. The obvious approach to this question is to model the background RMSD distribution with the size of the equivalent set as a parameter. Maclachlan [51] derived a model for the statistics of rigid-body superposition of random chains of a particular length and found that the expected RMSD grows proportionally to the square root of the chain length. Subsequently there have been many improvements made to this by sampling from real structures (e.g. [52]) or generating random-walk 'decoy' models [53].

The key issue with applying the results above to protein classification is that they derive statistical models from superpositions of contiguous chains. Although this is very useful in assessing the success of predictions of structure, where the set of equivalences is not optimized, if arbitrary gaps are permitted it is possible for an algorithm to converge on a discontiguous solution of high significance. This can lead to finding unexpected similarities between structures with unrelated topologies [54, 55], however the significance of aligning a beta strand to an alpha helix, for example, is difficult to assess. It is clear that such similarities are a consequence of the compactness of structures and the ubiquity of highly similar local structures resulting from constraints on hydrogen bonding, leading to a strong likelihood of structural convergence.

The consequences of these sorts of inconsistencies for protein classification have been analysed in several recent papers [56–58]. Particularly interesting is Pascual-Garcia *et al*'s use of transitivity violations to identify the level at which structural clustering is justified [58]. On the basis of RMSD comparisons using MAMMOTH and a number of clustering methods they argue for a disjunction between a regime in which comparisons form meaningful clusters and another which is better viewed as a continuum. This implies that if RMSD-based methods as they exist currently are used to classify there will always be a region beyond which no clustering is sensible.

Although classification and comparison of protein structures are fraught with complications and difficulties as

defined above both have been enormous sources of useful knowledge, ideas and methods for other purposes within bioinformatics—CATH and SCOP are routinely used in benchmarking novel methods. However they are also used as a basis for determining how much of fold space we have seen and what relationships exist between them. These are the subjects of the next two sections.

5. Protein fold space

When mapping protein fold space it is natural to ask how big this space is and how connected up the regions of the space are from the point of view of the evolutionary process. Several attempts to estimate the number of protein folds been made since Chothia [59] argued that there were about 1000 families of proteins on the basis of the mean number of secondary structural elements and their possible combinations. This is an interesting question since it provides an estimate of what level of structural coverage is likely to be achievable, the goal of structural genomics programmes [60].

Orengo *et al* [61] presented an early discussion of the overall statistics of protein folds and noted that nine groups dominated the set of structures then known, which they referred to as superfolds. These included alpha/beta barrels, doubly wound alpha/beta sandwich structures, Greek key double sheet sandwich proteins and up/down helical bundles. Although the databases at that point were comparatively small it could be shown that this bias was a natural effect and not sampling bias.

Their methods form the basis of other published estimates. Since each family of relatives will tend to conserve structure the number of distinct structures is at most the number of families. Therefore it is only necessary to determine the number of families, the distribution of numbers of families per fold and the sampling distribution, from which it is possible to calculate the number of folds with zero observed families.

Several estimates have been generated on this basis ranging from about 400 [62] to more than 10 000 [63]. Govindarajan *et al* [64] tested several distributional models with synthetic data and rejected the uniform and Gaussian distributions assumed by other authors in favour of a stretched exponential. On this basis they argued that there were some 4000 folds with about 2200 sufficiently likely to be found in nature. The estimate of 2000 is apparently still reasonably current [65].

Attempting to determine the completeness of our map of fold space in this way requires that our definition of a fold is at least a good approximation to a natural discrete grouping of structures. However some authors have suggested that such a discrete representation of fold space is not possible since the underlying space is continuous [66, 67] and that this underlies the problems of classification. This is to some degree underpinned by the finding that the distribution of fragments in protein structures is also highly skewed towards a small set [68] and that disparate folds can be linked in this way, a feature of proteins which is also used in structure prediction [69, 70]. Since we find that proteins are made up of a set of recurring fragments on a level larger than an individual element of secondary structure.

Such similarities are extremely important observations, however they remain similarities between substructures and the observations of significant similarity between globally different structures if either short contiguous fragments [71] or arbitrary selections of residues [54] are considered is not relevant to the comparison of global structures. Once larger-scale similarities on a topological level are considered the majority of protein fold space looks considerably more empty [72], suggesting that nature may not have exhausted the space of possible structures. We can then ask whether there are restrictions on the global structures that can be reached and whether they result from a small set of available fragments or strong biases in the use of such fragments.

Observations of structural continuity have been taken by some as a reason to avoid classification. In some cases interesting functional relationships [73] and insights into structural evolution [74] can be suggested regardless of global structural similarities, and representations of either fragment similarity or multiple relationships (as opposed to pairwise relationships) are substantially predictive of function [66, 75]. While this is undoubtedly true it presupposes that there is no reason to classify structure apart from functional inference, which is not the case if one is interested in structures for their own sake. Without a classification it is not possible to reason about protein structures effectively, however the somewhat vague notions of domains and folds which have been applied in the past in an attempt to simultaneously preserve evolutionary relationships and describe large-scale structural similarities have muddied the waters somewhat [32].

A significant step towards an objective fold definition was made by Taylor [76] in the definition of 'ideal forms'. These forms cover a substantial proportion of structures and permit an unambiguous definition of structures to be made. Particularly useful is the use of the forms to define a topology string as a path through a graph which represents the secondary structure elements in the protein. The similarity between two topologies can then be unambiguously defined as a distance between the two. This leads to a definition of fold space in which groups can be defined flexibly without recourse to RMSD-based measures which as we have seen can be difficult to derive a metric for. This approach allows structure comparisons to be made on a very abstract level so that a precise meaning to the difference between two domains can be given as a topological distance. This removes the need for a fold definition but allows visualization of common topologies as an aid to human understanding. Folds could then be defined as highly populated topologies which do not overlap, although the method would not provide a unique definition.

The space of protein structures is difficult to describe effectively since it requires making choices about the tradeoff between the size of a similarity and the degree of similarity. In general there is no answer to the question of whether we should consider two structures to be the same fold: if we are looking for something that is conserved then evolution is able to make large changes to global and local structures over long periods of time, and for much older superfamilies which have repeatedly

changed or modified function there will not necessarily be a universally conserved core structure. In addition there is the significant problem that some families of domains have global structures which are contained within the global structures of other family members. The similarity might be rather high and yet the two groups are best considered separately. In the limiting case we have to decide at what point two groups are considered different and at what point they are considered the same: is containing a single helix in both cases all that is required to consider them the same fold group? How much of an overlap is required? It is not possible in general to answer these questions, which shows that although the concept of a fold is useful it is nonetheless a human construct, at least unless a more precise definition can be found. Certainly trying to define it in evolutionary terms seems unlikely to work in general since evolution can change structures considerably. We consider the mechanisms of change and which changes have been observed in section 6.

6. Evolutionary trajectories through fold space

In order to fully understand protein structure we must consider the nature of the evolutionary process and its effects on how protein structures are distributed. Although it is difficult to robustly define fold categories without awkward overlaps it is undeniably true that proteins with undetectably similar sequences can adopt very similar structures and functions [61].

This leads to the question of whether the groups are likely to be divergently related. Since structures are generally very conserved the default position would be to expect that they are, but this has to be very carefully examined since deciding what level of similarity is required for an inference of homology is not straightforward.

Additionally it is necessary to ask whether it is possible for proteins to change structure substantially. This is an issue which is still receiving a great deal of attention at present. Assuming it were true, we could envisage either a discrete process of structural change with 'jumps' between folds occurring or a continuous process in which protein folds can gradually change into others. In fact both have been shown to occur, although the degree to which each has occurred in the course of evolution is not at all clear.

In the discrete case the question is essentially whether similar sequences necessarily have similar structures. Since the overwhelming experience with naturally occurring sequences is that the more similar a sequence is the more similar its structure is we should very much expect this to be the case, however these observations are biased in that they have been selected by evolution.

These questions have been inspired partly by observations that RNA 'folds' are connected extensively by neutral networks whereby simple walks of point mutations can lead from one to another [77]. Theoretical studies on cubic lattices provide the result that protein folds are densely packed in sequence space and that any two folds can be connected by a series of point mutations [78, 79]. Subsequently the existence of proteins with multiple ground-state structures has been demonstrated repeatedly [80–82] and some evidence has been found that these structures are evolutionarily relevant [83, 84]. In addition the results of a substantial number of design experiments show that proteins with entirely different secondary structure content and tertiary folds can be designed [85–87], with the most similar being 95% similar, a difference of only three amino acids [88].

It seems that such transitions do exist and are therefore available to evolution. Why, then, do structure predictions based on sequence similarities work so well? The biannual CASP protein structure prediction experiment, for example, now classifies nearly all of its targets as solvable by comparative modelling, meaning that a known structure provides a useful template for the match. In general the rule that for evolved proteins a similar sequence means a similar structure holds. If there are as many of these transitions as theory suggests then why does this work?

Assuming that we do not simply ignore the theoretical results, there are two responses to this: firstly that the transitions have occurred but that the sequences which have crossed over between folds are now in a different neutral network and so sequence information will quickly vanish as the protein diffuses along this. This is plausible since the exterior residues of the protein (most of which are likely to have changed) are under very different pressures than the interior residues; if stability is selectable then it is likely that all of the subtle patterns which allow sequence-similarity detection to function will be erased.

Another possible response is that they are very rare events since to observe a transition the protein must become fixed in the population, which is very unlikely unless it has some sort of function. Although it seems that it is not difficult for proteins to develop new functions (functional convergence is apparently quite common), particularly given that they may all be multifunctional, we would require that the protein found a new structure and function (or maintained its old function) and that this new function was sufficiently useful to become fixed and finally that the protein was not outcompeted by another with the same function.

It is not possible to test these ideas until a model which suggests where to look for these transitions is well established. It is also necessary to point out that the results above only tell us what could happen; real proteins experience a number of constraints from genome structure, mutational mechanisms, function and need for interaction which may well close off large areas of the network.

Perhaps less radical is the concept of continuous change. In order to change without large discontinuities it is only necessary to add and remove regions of structure until the goal is reached. Obviously since evolution has no goal it is better to describe this as a stochastic process in which certain insertions and deletions are more likely to have acceptable consequences for the structure. This is somewhat easier to imagine than the mechanisms described above since it does not require a large-scale change to have occurred at any point, and it is relatively easy to conceive of a structure adding a few bits while maintaining the old, finding a use for the new bits and then losing the old as a redundant copy is created.

That such a process could lead to a change of fold seems intuitively fairly obvious, however it is only quite recently that a full set of examples was brought together and reviewed [89]. Common examples include circular permutation, in which the sequence is rotated through the structure by a process of duplication and deletion, strand invasion and withdrawal and changes of elements from one type to another. Many of the larger topological changes can be rationalized as resulting from processes of gene duplication followed by deletion. Naturally these deletions do not have to be complete, and this may lead to insertion of elements. Alternatively the deletions may occur asymmetrically and lead to another fold.

Other mechanisms accounting for such changes might include exonization of intronic regions, asymmetric recombination and small-scale versions of the 'discrete' mechanism described above (since a local change need not necessarily have any effect on the function of the protein) in which local regions mutate from one type into another. This is easy to spot in many cases, particularly where deletions have made a helix into a strand-like region. A few examples which might indicate this sort of overlap have now been published [90, 91].

Another intriguing possibility is that the set of modern folds has arisen from several origins [92]. In this model the earliest proteins were oligomers of small fragments. These fragments, which might form individual strands or helices or larger regions of supersecondary structures, eventually joined in a variety of combinations to form the ancestors of the modern folds. The observation that fragment-based methods are effective in structure prediction and can be used to join quite different fold groups [68] might be seen to suggest this, but at present it is not clear whether there is a physical reason for this or whether it is simply convergence. Naturally the models might all have been used in the evolutionary process, although one may be the dominant force for change. In this case it remains only to develop the models sufficiently to determine the path which has been used by each family.

7. Conclusion

There remains considerable debate over the nature of the space of possible folds, in particular the question of continuity [66, 67, 75, 93–95]. In general it seems very common for a strong global similarity to convergently evolve and that a classification which is both structurally accurate and representative of evolutionary relationships may not be possible. Instead it may be preferable to derive a purely structural classification and use this as a standpoint from which to examine structural change in search of a more natural description of protein fold space. The ideal forms for protein structures [76] introduced in 2002 offer solutions to the problem of defining domains and folds objectively, and in combination with more traditional comparison methods should be useful in advancing our understanding.

Within the limitations of current methods we can summarize the state of knowledge as follows: protein structures exhibit considerable similarity on the level of substructures, which in turn may lead to certain structures being considerably more common through convergent evolution. Additionally the number of possible sequences stably adopting a given structure can vary considerably between structures, which further increases the bias towards particular large-scale structures since this increases the chance of finding a useful function and being evolutionarily selected. Consequently certain protein folds are substantially more common than others. It is not clear how many folds there are in total, nor how many there are which have been used by nature, although a number on the order of 2000 seems likely from previous estimates. From the experience of the CASP competition as well as three recent analyses it seems likely that we are reaching the limit of the most widely-used folds in nature, although it is not certain whether this is the entire range available to polypeptides, a question that the field of protein design seems ideally placed to answer. The challenge now is to find explanations for how proteins can move in this space and how we can use this to improve our understanding of protein structure and function, ultimately leading to improved tools for prediction and design.

Acknowledgments

The work was supported by the Medical Research Council (UK).

References

[1] Dunker A K, Oldfield C J, Meng J, Romero P, Yang J Y, Chen J W, Vacic V, Obradovic Z and Uversky V N 2008 The unfoldomics decade: an update on intrinsically disordered proteins *BMC Genomics* **9** S1

[2] Luheshi L M and Dobson C M 2009 Bridging the gap: from protein misfolding to protein misfolding diseases *FEBS Lett.* **583** 2581–6

[3] Sideras K and Gertz M A 2009 Amyloidosis *Adv. Clin. Chem.* **47** 1–44

[4] Sadowski M I and Jones D T 2009 The sequence-structure relationship and function prediction *Curr. Opin. Struct. Biol.* **19** 357–62

[5] James L J and Tawfik D S 2001 Catalytic and binding poly-reactivities shared by two unrelated proteins: the potential role of promiscuity in enzyme evolution *Protein Sci.* **10** 2600–7

[6] Bloom J D and Glassman M J 2009 Inferring stabilizing mutations from protein phylogenies: application to influenza hemagglutinin *PLOS Comput. Biol.* **5** e1000349

[7] Chothia C and Lesk A M 1986 The relation between the divergence of sequence and structure in proteins *EMBO J.* **5** 823–6

[8] Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T and Tramontano A 2007 Critical assessment of methods of protein structure prediction—round vii *Protein Struct. Funct. Genet.* **S8** 3–9

[9] Laskowski R A and Thornton J M 2008 Understanding the molecular machinery of genetics through 3d structures *Nat. Rev. Genet.* **9** 141–51

[10] Ramachandran G N, Ramakrishnan C and Sasisekharan V 1963 Stereochemistry of polypeptide chain configurations *J. Mol. Biol.* **7** 95–9

[11] Laskowski R A, Macarthur M W, Moss D S and Thornton J M 1993 Procheck—a program to check the sterochemical quality of protein structures *J. Appl. Crystallogr.* **26** 283–91

[12] Vriend G 1990 What if—a molecular modeling and drug design program *J. Mol. Graph.* **8** 52–6

[13] Sadowski M I and Jones D T 2007 Benchmarking template selection and model quality assessment for high-resolution comparative modeling *Protein Struct. Funct. Genet.* **69** 476–85

[14] Pauling L, Corey R B and Branson H R 1951 The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain *Proc. Natl Acad. Sci. USA* **37** 205–10

[15] Pauling L and Corey R B 1951 The structure of synthetic polypeptides *Proc. Natl Acad. Sci. USA* **37** 241–50

[16] Flores T P, Moss D S and Thornton J M 1994 An algorithm for automatically generating protein topology cartoons *Protein Eng.* **7** 31–7

[17] Westhead D R, Hatton D C and Thornton J M 1998 An atlas of protein topology cartoons available on the world-wide web *Trends Biochem. Sci.* **23** 35–6

[18] Johannissen L O and Taylor W R 2004 Protein fold comparison by the alignment of topological strings *Protein Eng.* **16** 949–55

[19] Orengo C A and Thornton J M 2005 Protein families and their evolution—a structural perspective *Annu. Rev. Biochem.* **74** 867–900

[20] Banavar J R and Maritan A 2007 Physics of proteins *Annu. Rev. Biophys. Biomol. Struct.* **36** 261–80

[21] Anfinsen C B 1973 Principles that govern the folding of protein chains *Science* **181** 223–30

[22] Levinthal C 1969 *Mossbauer Spectroscopy in Biological Systems* ed P Debrumer, J C M Tsibris and E Munck (Urbana: University of Illinois Press) pp 22–4

[23] Orengo C A, Michie A D, Jones S, Jones D T, Swindells M B and Thornton J M 1997 CATH—a hierarchic classification of protein domain structures *Structure* **5** 1093–108

[24] Murzin A G, Brenner S E, Hubbard T and Chothia C 1995 SCOP: a structural classification of proteins database for the investigation of sequences and structures *J. Mol. Biol.* **247** 536–40

[25] Hadley C and Jones D T 1995 A systematic comparison of protein structure classifications SCOP CATH and FSSP *Structure* **7** 1099–112

[26] Day R, Beck D A, Armen R S and Daggett V 2003 A consensus view of fold space: combining scop, cath and the dali domain dictionary *Protein Sci.* **12** 2150–60

[27] Csaba G, Birzele F and Zimmer R 2009 Systematic comparison of scop and cath: a new gold standard for protein structure analysis *BMC Struct. Biol.* **9** 23

[28] Delano W L 2002 *The Pymol Molecular Graphics System* (Paolo Alto, CA: Delano Scientific)

[29] Harrison A, Pearl F, Mott R, Thornton J and Orengo C 2002 Quantifying the similarities within fold space *J. Mol. Biol.* **323** 909–26

[30] Reves G A, Dallman T J, Redfern O C, Akpor A and Orengo C A 2006 Structural diversity of domain superfamilies in the CATH database *J. Mol. Biol.* **360** 725–41

[31] Cuff A L, Sillitoe I, Lewis T, Redfern O C, Garratt R, Thornton J and Orengo C A 2008 The cath classification revisited—architectures reviews and new ways to characterize structural divergence in superfamilies *Nucleic Acids Res.* **37** 310D–4D

[32] Taylor W R 2007 Evolutionary transitions in protein fold space *Curr. Opin. Str. Biol.* **17** 354–61

[33] Holm L and Sander C 1997 Dali/FSSP classification of three-dimensional protein folds *Nucleic Acids Res.* **25** 231–4

[34] Taylor W R 2000 Protein structure comparison using SAP *Protein Structure Prediction* (*Methods in Molecular Biology* vol 143, ed J M Walker) ed D M Webster (Totowa, NJ: Humana Press) pp 19–32

[35] Ortiz A R, Strauss C E M and Olmea O 2002 Mammoth (matching molecular models obtained from theory): an automated method for model comparison *Protein Sci.* **11** 2606–21

[36] Zhang Y and Skolnick J 2005 TM-align: a protein structure alignment algorithm based on the TM-score *Nucleic Acids Res.* **33** 2302–9

[37] Pandit S B and Skolnick J 2008 Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score *BMC Bioinformatics* **9** 531

[38] Mitchell E M, Artymiuk P J, Rice D W and Willett P 1989 Use of techniques derived from graph theory to compare secondary structure motifs in proteins *J. Mol. Biol.* **212** 151–66

[39] Harrison A, Pearl F, Sillitoe I, Slidel T, Mott R, Thornton J and Orengo C 2003 Recognizing the fold of a protein structure *Bioinformatics* **19** 1748–59

[40] Krissinel E and Henrick K 2004 Secondary-structure matching (ssm), a new tool for fast protein structure alignment in three dimensions *Acta Crystallogr.* D **60** 2256–68

[41] Redfern O C, Harrison A, Dallman T, Pearl F M G and Orengo C A 2007 Cathedral: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures *PLOS Comput. Biol.* **3** 2333–47

[42] Bostick D L, Shen M and Vaisman I I 2004 A simple topological representation of protein structure: implications for new, fast, and robust structural classification *Protein Struct. Funct. Genet.* **56** 487–501

[43] Rogen P and Fain B 2003 Automatic classification of protein structure by using gauss integrals *Proc. Natl Acad. Sci. USA* **100** 119–24

[44] Hasegawa H and Holm L 2009 Advances and pitfalls of protein structural alignment *Curr. Opin. Struct. Biol.* **19** 341–8

[45] McLachlan A D 1972 A mathematical procedure for superimposing atomic coordinates of proteins *Acta Crystallogr.* A **28** 656–7

[46] Kabsch W 1976 A solution for the best rotation to relate two sets of vectors *Acta Crystallogr.* A **32** 922–3

[47] Diamond R 1976 On the comparison of conformations using linear and quadratic transformations *Acta Crystallogr.* A **32** 1–10

[48] Kabsch W 1978 A discussion of the solution for the best rotatation to relate two sets of vectors *Acta Crystallogr.* A **34** 827–8

[49] McLachlan A D 1982 Rapid comparison of protein structures *Acta Crystallogr.* A **38** 871–3

[50] Theobald D L 2005 Rapid calculation of rmsds using a quaternion-based characteristic polynomial *Acta Crystallogr.* A **61** 478–80

[51] McLachlan A D 1984 How alike are the shapes of two random chains? *Biopolymers* **23** 1325–31

[52] Betancourt M R and Skolnick J 2001 Universal similarity measure for comparing protein structures *Biopolymers* **59** 305–9

[53] Taylor W R 2006 Transcription and translation in an RNA world *Phil. Trans. R. Soc.* B **361** 1751–60

[54] Kihara D and Skolnick J 2003 The PDB is a covering set of small protein structures *J. Mol. Biol.* **334** 793–802

[55] Zhang Y, Hubner I A, Arakaki A K, Shakhnovich E and Skolnick J 2006 On the origin and highly likely completeness of single-domain protein structures *Proc. Natl Acad. Sci. USA* **103** 2605–10

[56] Sam V, Tai C H, Garnier J, Gibrat J F, Lee B and Munson P J 2006 ROC and confusion analysis of structure comparison methods identify the main causes of divergence from manual protein classification *BMC Bioinformatics* **7** 206

[57] Sam V, Tai C H, Garnier J, Gibrat J F, Lee B and Munson P J 2008 Towards an automatic classification of protein structural domains based on structural similarity *BMC Bioinformatics* **9** 74

[58] Pascual-Garcia A, Abia D, Ortiz A R and Bastolla U 2009 Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures *PLOS Comput. Biol.* **5** e1000331

[59] Chothia C 1992 Proteins—1000 families for the molecular biologist *Nature* **357** 543–4

[60] Chandonia J M and Brenner S E 2006 The impact of structural genomics: expectations and outcomes *Science* **311** 347–51

[61] Orengo C A, Jones D T and Thornton J M 1994 Protein superfaimiles and domain superfolds *Nature* **372** 631–4

[62] Wang Z X 1996 How many fold types of protein are there in nature? *Protein Struct. Funct. Genet.* **26** 186–91

[63] Coulson A F W and Moult J 2001 A unifold, mesofold and superfold model of protein fold use *Protein Struct. Funct. Genet.* **46** 61–71

[64] Govindarajan S, Recabarren R and Goldstein R A 1999 Estimating the total number of protein folds *Protein Struct. Funct. Genet.* **35** 408–14

[65] Levitt M 2006 Growth of novel protein structural data *Proc. Natl Acad. Sci. USA* **104** 3183–8

[66] Kolodny R, Petery D and Honig B 2006 Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction *Curr. Opin. Struct. Biol.* **16** 393–8

[67] Skolnick J, Arakaki A K, Lee S Y and Brylinski M 2009 The continuity of protein structure space is an intrinsic property of proteins *Proc. Natl Acad. Sci. USA* **106** 15690–5

[68] Friedberg I and Godzik A 2005 Fragnostic: walking through protein structure space *Nucleic Acids Res.* **33** W249–51

[69] Bonneau R, Tsai J, Ruczinski I, Chivian D, Strauss C E M and Baker D 2001 Rosetta in CASP4: progress in *ab initio* protein structure prediction *Proteins* **S5** 119–26

[70] Jones D T 2001 Predicting novel protein folds by using FRAGFOLD *Proteins* **S5** 127–32

[71] Friedberg I and Godzik A 2005 Connecting the protein structure universe by using sparse recurring fragments *Structure* **13** 1213–24

[72] Taylor W R, Chelliah V, Hollup S M, Macdonald J T and Jonassen I 2009 Probing the dark matter of protein fold space *Structure* **17** 1244–52

[73] Shindyalov I N and Bourne P E 2001 An alternative view of protein fold space *Proteins: Struct. Funct. Bioinf.* **38** 247–60

[74] Hou J, Sims G E, Zhang C and Kim S-H 2003 A global representation of protein fold space *Proc. Natl Acad. Sci. USA* **100** 2386–90

[75] Petrey D and Honig B 2009 Is protein classification necessary? Toward alternative approaches to function annotation *Curr. Opin. Struct. Biol.* **19** 363–8

[76] Taylor W R 2002 A periodic table for protein structure *Nature* **416** 657–60

[77] Schultes E A and Bartel D P 2000 One sequence, two ribozymes: implications for the emergence of new ribozyme folds *Science* **289** 448–52

[78] Bornberg-Bauer E 1997 How are model protein structures distributed in sequence space? *Biophys. J.* **73** 2393–403

[79] Babajide A, Farber R, Hofacker I L, Inman J, Lapedes A S and Stadler P F 2001 Exploring protein sequence space using knowledge-based potentials *J. Theor. Biol.* **212** 35–46

[80] Cordes M H J, Burton R E, Walsh N P, McKnight C J and Sauer R T 2000 An evolutionary bridge to a new protein fold *Nat. Struct. Biol.* **7** 1129–32

[81] Meier S, Jensen P R, David C N, Chapman J, Holstein T W, Grzesiek S and Ozbeck S 2007 Continuous molecular evolution of protein-domain structures by single amino acid changes *Curr. Biol.* **17** 173–8

[82] Tuinstra R L, Peterson F C, Kutlesa S, Elgin E S, Kron M A and Volkman B F 2008 Interconversion between two unrelated protein folds in the lymphotactin native state *Proc. Natl Acad. Sci. USA* **105** 5057–62

[83] Roessler C G, Hall B M, Anderson W J, Ingram W M, Roberts S A, Montfort W R and Cordes M H J 2008 Transitive homology-guided structural studies lead to discovery of Cro proteins with 40% sequence identity but different folds *Proc. Natl Acad. Sci. USA* **105** 2343–8

[84] Van Dorn L O, Newlove T, Chang S M, Ingram W M and Cordes M H J 2006 Relationship between sequence determinants of stability for two natural homologous proteins with different folds *Biochemistry* **35** 10542–53

[85] Jones D T, Moody C M, Uppenbrink J, Viles J H, Doyle P M, Pearl L H, Sadler P J and Thornton J M 1996 Towards meeting the paracelsus challenge: the design, synthesis, and characterization of paracelsin-43, an α-helical protein with over 50% sequence identity to an all-β protein *Protein Struct. Funct. Genet.* **24** 502–13

[86] Dalal S, Balasubramanian S and Regan L 1997 Protein alchemy: changing β-sheet into α-helix *Nat. Struct. Biol.* **4** 548–52

[87] Alexander P A, He Y, Chen Y, Orban J and Bryan P N 2007 The design and characterization of two proteins with 88percent sequence identity but different structure and function *Proc. Natl Acad. Sci. USA* **104** 11963–8

[88] He Y, Chen Y H, Alexander P, Bryan P N and Orban J 2008 Nmr structures of two designed proteins with high sequence identity but different fold and function *Proc. Natl Acad. Sci. USA* **105** 14412–7

[89] Grishin N V 1985 Fold change in evolution of protein structures *J. Struct. Biol.* **134** 167–85

[90] Krishna S S and Grishin N V 2005 Structural drift: a possible path to protein fold change *Bioinformatics* **21** 1308–10

[91] Cheng H and Grishin N V 2005 DOM-fold: a structure with crossing loops found in DmpA ornithine acetyltransferase, and molybdenum cofactor-binding protein domain *Protein Sci.* **14** 1902–10

[92] Lupas A N, Ponting C P and Russell R B 2001 On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion or relics of an ancient peptide world? *J. Struct. Biol.* **134** 191–203

[93] Sadreyev R I, Kim B-H and Grishin N V 2009 Discrete-continuous duality of protein structure space *Curr. Opin. Struct. Biol.* **19** 321–8

[94] Sippl M J 2009 Fold space unlimited *Curr. Opin. Struct. Biol.* **19** 310–1

[95] Valas R E, Yang S and Bourne P E 2009 Nothing about protein structure classification makes sense except in the light of evolution *Curr. Opin. Struct. Biol.* **19** 329–34