Science Progress (2010), 93(1), 1–6
doi: 10.3184/003685009X12605492662844

An overview of the processes shaping protein evolution

ROY D. SLEATOR

ABSTRACT

From a comparatively small number of protein structural domains a staggering array of structural variants has evolved which has, in turn, facilitated an expanse of functional derivatives. Herein I review the primary mechanisms which have contributed to the vastness of our existing, and expanding, protein repertoires.

Keywords: *evolution, protein domains, gene duplication, divergence, combination, circular permutation*

Dr Roy Sleator BSc, PhD (NUI), PGCert Bioinformatics (Manchester, UK) is a lecturer at the Department of Biological Sciences at Cork Institute of Technology and a Principal Investigator at the Alimentary Pharmabiotic Centre, University College Cork. His primary interests include the bioinformatics of protein structure, function and evolution and the rational design of improved Pharmabiotics as vaccine and drug delivery vehicles. Sleator is a pioneer of the *Patho-biotechnology* concept and is the founding Editor-in-Chief of the scientific journal *Bioengineered Bugs* http://www.landesbioscience.com/journals/biobugs/ He may be contacted at E-mail: Roy.sleator@cit.ie

Introduction

"*Progress has not followed a straight ascending line, but a spiral with rhythms of progress and retrogression, of evolution and dissolution*."

Johann Wolfgang von Goethe (1749–1832).

Data from the most recent large scale sequencing projects has facilitated detailed descriptions of the constituent protein repertoires of more than 600 distinct organisms[1]. Taking protein domains

(clusters of 50–200 conserved residues) to represent units of evolution, as well as their more usual designation as structural/functional motifs, it is possible to accurately trace the evolutionary relationships of approximately half of these proteins. The primary driving force for the creation of evolutionary diverse protein families can be ascribed to three main mechanisms: the first; gene duplication, gives rise to often closely related proteins[2]. The second; divergence, further modifies the existing paralogues, leading to even more diverse protein families[3]. While the third; gene combination, results in still further and even more dramatic changes to the resulting proteins (as dictated by evolutionary pressure and the physiological fitness requirements of the organism) [4].

Herein, I review the current knowledge on protein evolution with a specific focus on how gene duplications, sequence divergence and domain combinations have shaped protein evolution.

Duplication

Of the animal genomes sequenced to date, the proportion of matched domains which are the result of duplications is estimated at between 93 and 97%[5]. Indeed, the haemoglobins, which were the first homologous proteins to have their structure determined, are perhaps the best example of how duplication (and subsequent mutational events) has given rise to subtle structural and functional variations such as oxygen binding profiles[6]. Furthermore, in addition to the generation of whole protein homologues, partial gene duplications resulting in domain duplication and elongation are also common features of protein evolution[7]. In many cases such enlargements have resulted from the addition of subdomains, variability in loop length, and/or changes to the structural core, such as beta-sheet extensions. Examples of such protein duplication events include cutinase and bovine bile-salt activated cholesterol esterase. While cutinase is the smallest enzyme of the α/β hydrolases, with five strands in the main beta-sheet[8], bovine bile-salt activated cholesterol esterase has 11 strands, and loop structures up to 79 residues in length[9].

Divergence

There are essentially two types of protein structural divergence: changes to the proteins surface or peripheral regions (*e.g.* surface loops, surfaces helices and strands on the edges of β-sheets) and the less common but far more detrimental modifications to the proteins

interior or core[10]. Indeed, it has been demonstrated that mutations in the protein surface are four times more biologically acceptable than those in the interior[1]. In support of this is the observation that pairs of homologous proteins with identities of approximately 20% have been shown to exhibit up to 50% divergence in the peripheral regions alone[11].

In addition to subtle changes resulting from missense point mutations leading to single amino acid substitutions and the resulting gradual divergence in structure and function, more radical divergence of structure, mediated by domain shuffling (recombination or permutation) has also been reported[12]. Circular permutations (CPs) in particular represent a specific form of recombination event which is characterised by the presence of the same protein sub-sequences in the same linear order but different positions of the N and C termini[13], in essence CP of a protein can be visualised as if its original termini were linked and new ones created elsewhere. First observed in plant lectins[14], a substantial number of natural examples of CP have been reported; indeed, some 120 protein clusters which appear to have segments of their sequences in different sequential order are reported in the Circular Permutation Database[15]. In addition to natural evolutionary processes, artificial CPs have been engineered in an effort to study protein folding properties as well as the design of more efficient enzymes[16]. A circularly permuted streptavidin for example has been designed to remove the flexible polypeptide loop that undergoes an open to closed conformational change when biotin is bound. The original termini have been joined by a tetrapeptide linker, and four loop residues have been removed, resulting in the creation of new N- and C-termini[17].

While domain shuffling may have dramatic effects on protein structure, protein homologues usually conserve their catalytic mechanisms *i.e.* the relative positions of their functional active sites or catalytic residues may shift but they retain their functional activity. This usually occurs when divergence induces structural changes in the catalytic region, thus necessitating a reconfiguration of the position of the catalytic residues in order to maintain function[18]. In several cases, whilst the functionally equivalent residues are located at non-homologous positions on the protein's 3D structure, the catalytic residues themselves are identical. An example of this is chloramphenicol acetyltransferase (PaXAT) and UDP-*N*-acetylglucosamine acyltransferase (LpxA) both of which contain an essential histidine residue thought to be involved in deprotonation of a hydroxyl group in their individual substrates.

However, these residues are located at different points within the protein fold; in LpxA, the histidine is located in the core of the domain[19], whereas in PaXAT, it occurs in a loop extending from the solenoid structure.

Thus, two proteins may have quite divergent structures and/or sequences while retaining similar function; such proteins are said to be functional analogs. Such analogs may also arise as a result of convergent evolution; that is they do not diverge from a common ancestor but instead arise independently and converge on the same active configuration as a result of natural selection for a particular biochemical function. L-aspartate aminotransferase and D-amino acid aminotransferase provide excellent examples of convergently evolved functional analogues. Despite having a strikingly similar arrangement of residues in their active sites, the two proteins have completely different architectures; differing in size, amino acid sequence and in the fold of the protein domains.

Conversely, certain proteins share significant sequence and/or structure similarity but differ in terms of substrate specificity or indeed catalytic function. An example of such structural analogs, which arise by means of divergent evolution from a single ancestor, include Human IL-10 (hIL-10); a cytokine that modulates diverse immune responses and the Epstein-Barr virus (EBV) IL-10 homologue (vIL-10). Although vIL-10 suppresses inflammatory responses like hIL-10, it cannot activate many other immunostimulatory functions performed by the cellular cytokine[20].

Combination

While the evolutionary impact of duplication and divergence on protein sequence, structure and function is obvious, multi-domain proteins are for the most part the result of gene combinations[21]. Such combinations can give rise to domain recruitment and enlargement and can significantly affect both protein structure-/stability and function. For example in the case of domain recruitment the addition of an accessory domain may affect protein function by modulating substrate selectivity; achieved either by the addition of a binding site, or, by playing a purely structural role, shaping the existing active site to accommodate substrates of different shapes and/or sizes[18]. For example, prokaryotic methionine aminopeptidase exists as a monomeric single-domain protein while creatinase, is a two-domain protein. The additional domain of the second subunit of creatinase caps the active site allowing the binding of the small molecule creatine[22].

Conclusion

While the genesis of protein evolution most have necessitated the synthesis of new proteins 'from scratch'[1], such an *ab initio* invention step now appears to be largely absent, replaced with the much faster process of shaping new proteins with modified functions by the processes of gene duplication, sequence divergence and domain combinations[23]. Herein we have discussed how these mechanisms have shaped protein evolution and how the retention of sequence and/or structural domains has facilitated the tracking of this evolutionary process through the millennia. With the development of metagenomics[24] and the discovery of new and previously uncharacterised microbes, and their constituent protein repertoires, it is entirely likely that additional domain families will continue to be identified and new chapters of the protein evolution story will continue to be written.

References

1. Chothia, C. and Gough, J. (2009) Genomic and structural aspects of protein evolution. *Biochem. J.*, **419**, 15–28.
2. Brenner, S.E., Hubbard, T., Murzin, A. and Chothia, C. (1995) Gene duplications in H. influenzae. *Nature*, **378**, 140.
3. Teichmann, S.A., Park, J. and Chothia, C. (1998) Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. *Proc. Natl. Acad. Sci. USA*, **95**, 14658–14663.
4. Volff, J.N. and Brosius, J. (2007) Modern genomes with retro-look: retrotransposed elements, retroposition and the origin of new genes. *Genome Dyn.*, **3**, 175–190.
5. Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C. and Gough, J. (2009) SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucl. Acids Res.*, **37**, D380–386.
6. Blanchetot, A., Wilson, V., Wood, D. and Jeffreys, A.J. (1983) The seal myoglobin gene: an unusually long globin gene. *Nature*, **301**, 732–734.
7. Moore, A.D., Bjorklund, A.K., Ekman, D., Bornberg-Bauer, E. and Elofsson, A. (2008) Arrangements in the modular evolution of proteins. *Trends Biochem. Sci.*, **33**, 444–451.
8. Longhi, S., Czjzek, M., Lamzin, V., Nicolas, A. and Cambillau, C. (1997) Atomic resolution (1.0 A) crystal structure of Fusarium solani cutinase: stereochemical analysis. *J. Mol. Biol.*, **268**, 779–799.
9. Chen, J.C., Miercke, L.J., Krucinski, J., Starr, J.R., Saenz, G., Wang, X., Spilburg, C.A., Lange, L.G., Ellsworth, J.L. and Stroud, R.M. (1998) Structure of bovine pancreatic cholesterol esterase at 1.6 A: novel structural features involved in lipase activation. *Biochemistry*, **37**, 5107–5117.
10. Gerstein, M., Sonnhammer, E.L. and Chothia, C. (1994) Volume changes in protein evolution. *J. Mol. Biol.*, **236**, 1067–1078.

11. Chothia, C. and Lesk, A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.
12. Kawashima, T., Kawashima, S., Tanaka, C., Murai, M., Yoneda, M., Putnum, N.H., Rokhsar, D.S., Kanehisa, M., Satoh, N. and Wada, H. (2009) Domain shuffling and the evolution of vertebrates. *Genome Res*. (in press).
13. Vogel, C. and Morea, V. (2006) Duplication, divergence and formation of novel protein topologies. *Bioessays*, **28**, 973–978.
14. Lindqvist, Y. and Schneider, G. (1997) Circular permutations of natural protein sequences: structural evidence. *Curr. Opin. Struct. Biol*., **7**, 422–427.
15. Lo, W.C., Lee, C.C., Lee, C.Y. and Lyu, P.C. (2009) CPDB: a database of circular permutation in proteins. *Nucl. Acids Res*., **37**, D328–332.
16. Heinemann, U., Ay, J., Gaiser, O., Muller, J.J. and Ponnuswamy, M.N. (1996) Enzymology and folding of natural and engineered bacterial beta-glucanases studied by X-ray crystallography. *Biol. Chem*., **377**, 447–454.
17. Chu, V., Freitag, S., Le Trong, I., Stenkamp, R.E. and Stayton, P.S. (1998) Thermodynamic and structural consequences of flexible loop deletion by circular permutation in the streptavidin-biotin system. *Protein Sci*., **7**, 848–859.
18. Todd, A.E., Orengo, C.A. and Thornton, J.M. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol*., **307**, 1113–1143.
19. Wyckoff, T.J. and Raetz, C.R. (1999) The active site of Escherichia coli UDP-N-acetylglucosamine acyltransferase. Chemical modification and site-directed mutagenesis. *J. Biol. Chem*., **274**, 27047–27055.
20. Yoon, S.I., Jones, B.C., Logsdon, N.J. and Walter, M.R. (2005) Same structure, different function crystal structure of the Epstein-Barr virus IL-10 bound to the soluble IL-10R1 chain. *Structure*, **13**, 551–564.
22. Apic, G., Gough, J. and Teichmann, S.A. (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol*., **310**, 311–325.
23. Chothia, C., Gough, J., Vogel, C. and Teichmann, S.A. (2003) Evolution of the protein repertoire. *Science*, **300**, 1701–1703.
24. Sleator, R.D., Shortall, C. and Hill, C. (2008) Metagenomics. *Lett. Appl. Microbiol*., **47**, 361–366.