Emergence of *de novo* proteins from 'dark genomic matter' by 'grow slow and moult'

Erich Bornberg-Bauer*[1], Jonathan Schmitz* and Magdalena Heberlein*

*Institute for Evolution and Biodiversity, University of Muenster, Huefferstrasse 1, D48149 Muenster, Germany

Abstract

Proteins are the workhorses of the cell and, over billions of years, they have evolved an amazing plethora of extremely diverse and versatile structures with equally diverse functions. Evolutionary emergence of new proteins and transitions between existing ones are believed to be rare or even impossible. However, recent advances in comparative genomics have repeatedly called some 10 %–30 % of all genes without any detectable similarity to existing proteins. Even after careful scrutiny, some of those orphan genes contain protein coding reading frames with detectable transcription and translation. Thus some proteins seem to have emerged from previously non-coding 'dark genomic matter'. These '*de novo*' proteins tend to be disordered, fast evolving, weakly expressed but also rapidly assuming novel and physiologically important functions. Here we review mechanisms by which '*de novo*' proteins might be created, under which circumstances they may become fixed and why they are elusive. We propose a 'grow slow and moult' model in which first a reading frame is extended, coding for an initially disordered and non-globular appendage which, over time, becomes more structured and may also become associated with other proteins.

Introduction

Proteins show an amazing plethora of structures and functions and can be seen as the ubiquitous and essential toolbox that shapes cells and, ultimately, all forms of life as we know it today. Therefore, understanding protein evolution sheds light on basic principles of how all forms of life evolve. However, it is quite unclear how proteins have arisen in first place and if and by which mechanisms they descended from a possibly limited set of presumably much smaller proteins [1–3]. This lack of a comprehensive theory of protein evolution is even more intriguing because comparative analyses of extant protein inventories have suggested that the formation of new proteins, i.e. proteins with fundamentally new structures, is very unlikely [4]. The structural relationships between proteins can be imagined as islands of stable folds which are separated by large distances, corresponding to an ocean of unstable and non-functional structural intermediates which thus represent a non-permissive barrier for evolution [5]. Much of this analogy has been derived from limited knowledge comprising a relatively small set of stably folding and mostly globular proteins from a small set of organisms which crystallize well.

Over the last decade, these views have been challenged by the provision of huge amounts of data which were obtained with novel computational and experimental techniques, including massively parallel sequencing providing thousands of new genomes [6]. Furthermore, experimental techniques have improved too and it has become possible to analyse proteomes with unprecedented accuracy [7] and activities and structural properties, even *in vivo* [8].

Finally, a couple of studies have shown that, at least under somewhat specific conditions and with some prior knowledge, stable protein structures can be converted into each other, sometimes along a continuous path of viable mutations and with intermediate sequences with equal probability for two distinct structures [9–15]. Advances in computational protein folding have enabled the artificial design of a new fold which was experimentally confirmed and associated a predicted function [1]. In spite of all these successes, all evidences for structural transitions remain rather incidental and do not reveal a common pattern. Ergo, the relevance of such transitions for novel, previously non-identified structures, during evolution, must still be considered to be very low.

An important process for creating novel proteins is by rearrangement of domains rather than the accumulation of small changes (such as amino-acid substitutions and indels). Domains are functional, structural and evolutionary units of proteins. Proteins quite often change their linear order of domains (domain arrangements), leaving domains and their principal function intact. Such rearrangements induce subtle changes and an escape from evolutionary frozen links, thus facilitating novel functions [16–19]. Strictly speaking, such a rearranged protein is not new because domains often remain similar in function and structure and coding fragments are reused. Accordingly, such rearranged proteins can be well captured by data bases and rearrangements can be described by efficient algorithms [20–22].

Recent findings from comparative genomics suggest, however, that protein coding genes can also be created *de*

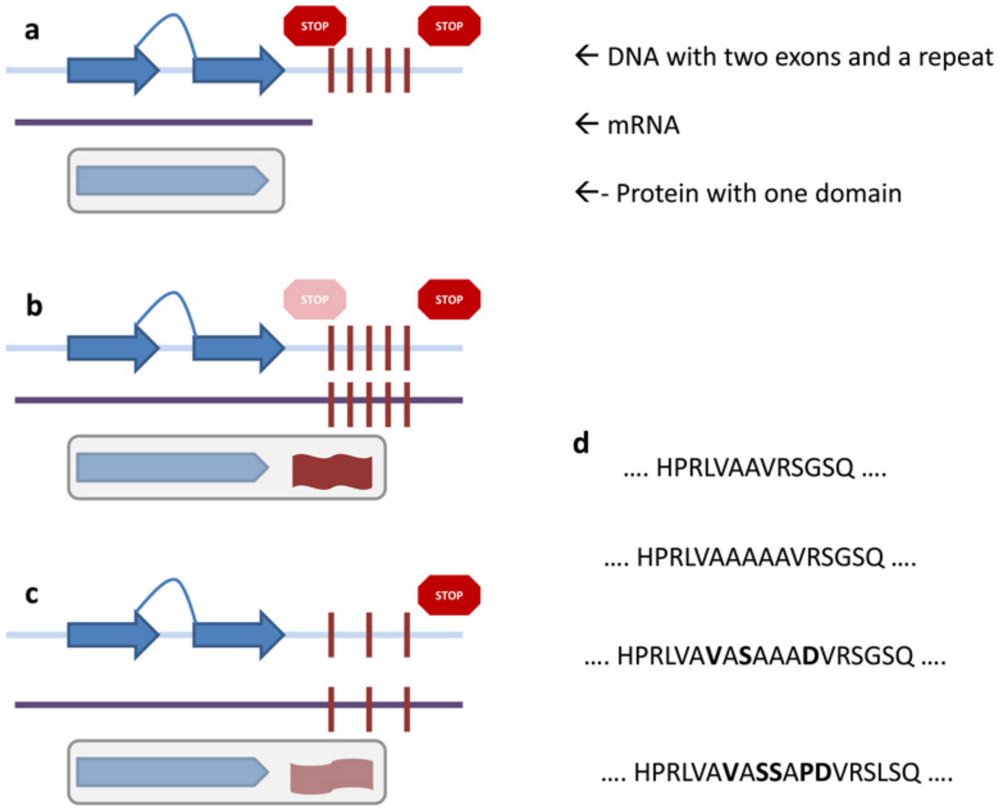
Key words: domain rearrangements, orphan genes, protein disorder, protein evolution.

Abbreviations: HCA, hydrophobic cluster analysis; HMM, Hidden Markov model; MDA, multiple domain arrangements; SDA, single domain arrangements.

[1] **To whom correspondence should be addressed (email ebb@wwu.de).**

Biochem. Soc. Trans. (2015) 43, 867–873; doi:10.1042/BST20150089

Figure 1 | Slow grow and moult model proposing a mechanism of *de novo* protein emergence


a
STOP
STOP
← DNA with two exons and a repeat
← mRNA
←- Protein with one domain
b
STOP
STOP
c
STOP
d
.... HPRLVAAVRSGSQ ....
.... HPRLVAAAAAVRSGSQ ....
.... HPRLVAVASAAADVRSGSQ ....
.... HPRLVAVASSAPDVRSLSQ ....

novo, i.e. from previously non-coding sequences on the genome [30]. Approximately 10 %–30 % of genes lack any detectable similarity to genes from previously, sometimes even very closely related, genomes. Many of these genes likely originated *de novo*. How such *de novo* proteins could emerge remains, to date, unclear and many analyses so far are contradictory or difficult to explain from a genetic and a biophysical point of view. De novo *genes* seem to emerge frequently but, the younger they are, the more likely they are lost again [23,24]. *De novo* genes are short, with fewer introns (if any) than established genes and fast evolving [25–30]. Expression of non-protein coding genes is more frequent than previously thought [31] and may lead to stable transcripts. However, the final protein products may be expressed and become functionally relevant only under very special conditions [24,27,28,32], which may be difficult to anticipate and test experimentally.

Very little is known about the properties of the proteins encoded by *de novo* genes and so far all knowledge comes from computational analyses only. From a biophysical perspective, the emergence of novel functional proteins from random DNA stretches (or randomly chosen stretches) is difficult to explain [33]. Rational protein design is still next to impossible and experimental approaches do not yield *in vivo* proteins with desired structures and *in vivo* functions. Exceptions are extremely scarce: random sequences, composed from simplified alphabets (QLR) can fold [34] and sometimes show rather low and generic functionality [35] but little else is known about designing functional proteins from scratch.

So, how can current knowledge on protein biophysics be reconciled with observations from comparative genomics? One explanation could be that *de novo* reading frames initially code for RNA genes and only later attain significant expression levels. Accordingly, selection would initially act on RNA, which is generally less specific but more versatile and only later shift to the protein. This way the requirement for the protein to immediately fold and function would be relaxed. Indeed, it has been observed that novel and functional RNA genes without further translation exist [36], that lncRNAs with significant ribosome binding UTRs are widespread and occasionally become translated [37] and that even siRNAs become occasionally translated into peptides [38]. However, since most *de novo* genes assume an ORF first ('proto-genes') [27,31,39], and become transcribed only later on, evolutionary RNA intermediates may only be marginally relevant for the creation of novel proteins.

A second explanation could be the revival of dormant reading frames. Genes that are pseudogenized escape purifying selection pressure and subsequently accumulate mutations such that similarity to orthologues becomes blurred. However, hardly detectable traces of hydrophobic-polar pattern might still exist and thus confer a 'head-start' to a novel protein for folding into a functional protein, if the reading frame becomes activated again. Such an idea has been proposed in the context of neofunctionalization for novel gene creation by duplication four decades earlier on [40], but no convincing example from comparative genomic studies has been provided so far for orphan gene creation.

Another possible process of *de novo* protein creation which we propose here, is a *grow slow and moult* process which starts by the extension of reading frames beyond either the N-terminus or the C-terminus of the encoded protein (Figure 1). These new stretches eventually become new domains, sometimes separating from their previously associated parent proteins and associating with other domains or proteins. This concept is supported by a couple of observations: first, proteins seem to be generally perceptive towards changes at their ends as they are likely to harbour or lose additional domains [17,41]. Second, the majority of all bona fide *de novo* domains in insects are terminal, in particular C-terminal [42]. This indicates that read through mutations, initially just phenotypic and later conserved by mutation of a stop codon, create phenotypic variations. Read through mutations frequently facilitate novel functions in proteins [43]. It is conceivable that such a process also creates novel terminal domains. Indeed, many novel domains, in particular the terminal ones [9], are more disordered than older domains. Third, the extension of reading frames may reach into highly repetitive sequences (Figures 1a and 1b), such as micro-satellites, which are enriched near genes [44,45]. Such nucleotide repeats will inevitably translate into amino acid repeats (Figures 1b and 1d), which are more likely to be disordered [46]. Over evolutionary long time scales, such disordered regions become 'tamed' (Figures 1c and 1d) and assume less regular sequence patterns. These terminal disordered regions may initially mediate some binding interface, though not necessarily a specific one and over time become more specific [47–49]). Note that intrinsic disorder in proteins is frequent and related to flexibility. Disordered regions are also often related to important physiological functions. It is worth noting that some of the unstructured regions were recently classified as regions of 'constrained disorder', which relates to a certain level of sequence conservation [50,51]. The conserved residues usually correspond to highly specific binding motifs or post-translational modification sites and can act as structural switches, e.g. when they become phosphorylated [52,53]. Finally, recent results of *de novo* domains in insects, which have been determined with a hydrophobic cluster analysis (HCA) but without prior knowledge of homology, also indicated that, the older novel domains are, the more their repertoire of hydrophobic micro-clusters resemble the repertoire of well-established globular proteins with known structures and that disorder decreases over time [42]. Such hydrophobic clusters and the decrease of disorder indicate a tighter packing and higher globularity which would entail stronger selection pressure acting on the more densely packed protein [54]. However, the opposite, i.e. younger proteins (or domains) being less disordered, has also been reported [27,55]. This may either hint at insufficient prediction methods or that, at least in some cases, disordered regions may themselves be under selection and disorder may be a derived and selected for trait.

Table 1 | List of species used in this study

All species' proteomes were downloaded from ensembl metazoa (http://metazoa.ensembl.org).

Drosophila melanogaster
Drosophila simulans
Drosophila sechellia
Drosophila erecta
Drosophila yakuba
Drosophila ananassae
Drosophila pseudoobscura
Drosophila persimilis
Drosophila willistoni
Drosophila virilis
Drosophila mojavensis
Drosophila grimshawi
Aedes aegypti
Anopheles gambiae
Culex quinquefasciatus
Apis mellifera
Nasonia vitripennis
Atta cephalotes
Solenopsis invicta
Acyrthosiphon pisum
Bombyx mori
Pediculus humanus
Tribolium castaneum
Daphnia pulex
Ixodes scapularis

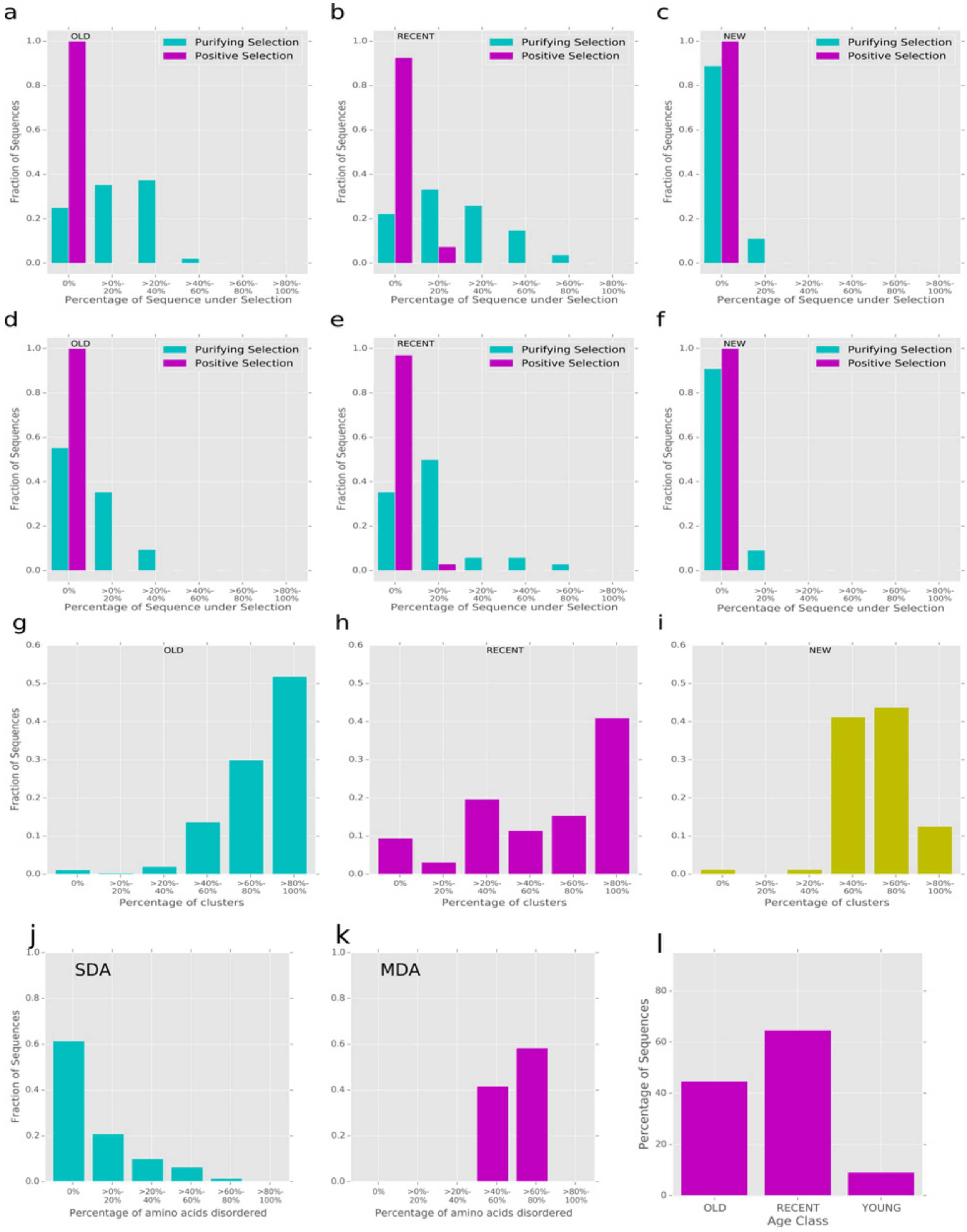
Taken together, recent results support that some domains may have emerged via grow slow and moult and call for further integrated analyses in which the biophysical parameters are concordantly interpreted with signals of adaptation and genetic mechanisms underlying the creation of novel genes. Such a brief complementary study is presented here, on a dataset of 29 novel domains which have arisen in insects and are recorded in domain databases.

Materials and methods

Arthropod proteomes with annotation in OrthoDB were downloaded from Ensemble metazoa. See Table 1 for a list of the species used in this study. Proteins containing the new domains found in [17] where found using Pfamscan [59]. In those proteins, disorder was computed using IUPRED [60]. Hydrophobic clusters were determined using Seq-HCA [61]. These hydrophobic clusters were compared with a

Figure 2 | **Analysis of novel, terminal domains**

(**a-c**) Analysis of residues under selection in protein sequences, different plots shown for the age classes. (**d-f)** Analysis of residues under selection in domain sequences, different plots shown for the age classes. (**g-i)** Analysis of hydrophobic clusters found in novel domain sequences. For each age group, the number of clusters also found in globular proteins is shown. (**j** and **k**): Protein disorder in domains found in different types of domain arrangements. For SDA and MDA each, the percentage of disordered sequence is shown. (**l**) Domains with at least one residue under selection, by age class.



dataset containing clusters known to occur in globular protein sequences [62], as described in [42]. OrthoDB version 7 [63] orthologous clusters containing proteins with new domains were extracted. Those proteins containing new domains where aligned using TranslatorX and Muscle [64,65]. SLR [66] was used to search for patterns of selection. For this purpose, a phylogenetic tree based on [67] was used as well.

Results and discussion

We used those domains from [56] that only occurred alone or at the end of domain arrangements. These domains are grouped into three age bins: old domains emerged between the root of the insect tree (~430 Ma) and the common ancestor of diptera (~225 Ma), recent domains which arose thereafter until the emergence of Drosophilidae (~40 Ma) and new domains are even younger. We first calculated signals of selection, specifically the degree of purifying and positive selection by comparing the d_N/d_S values by aligning the underlying coding DNA sequences (see 'Materials and methods' for details). In this process, we aligned the whole proteins and counted positions under selection for whole proteins, as well as domain sequences. Although there are no significant signs of positive selection, we find purifying selection acting more often on older domains and, even more so, in the overall proteins which harbour these new domains (Figures 2a–2e). The reason for not finding positive selection acting on the new domains could be that they are already relatively established and their dataset is rather small (four domains with on average 12 copies). This result indicates that the proteins are under some selection pressure, presumably to maintain their structure against the backdrop of a newly recruited and evolving domain which may have a destabilizing effect on the overall protein.

Second, we analysed the emergence of hydrophobic clusters (Figures 2g–2i). Although the domains studied here are all recorded as Hidden Markov models (HMMs) profiles in a database (Pfam) and therefore evolutionarily well-established, they still show the same signal of evolutionary dynamics as an earlier study on newly detected domains [42]. Old domains have a much higher percentage of such micro-clusters which are also found in well-established globular proteins and, therefore, seem to have assumed a more compact and possibly globular structure.

Finally, we used a disorder prediction on the domains as has been performed on a slightly different dataset in [55]. Here, we find that the terminal domains tend to be more disordered, if they appear in combination with other domains [multiple domain arrangements (MDA)], rather than alone in single domain arrangements (SDA, Figures 2j and 2k). This result indicates that, in SDAs, disorder is selected against, possibly to maintain the overall globularity of the protein. In MDAs, with other, probably globular domains, this selective pressure could be weaker. In combination with the study, this result indicates that novel domains tend to become more compact and globular and that their hosting proteins are under pressure to maintain their function.

Conclusions and outlook

We have proposed here a solution to the notorious problem of *de novo* protein emergence which materialized with the recent and rapid growth of genomic data and puts at odds genetic observations and biophysical reasoning. The 'grow slow and moult' model will describe just one among several mechanisms of how previously non-coding genetic material can be turned into functional proteins as nature almost always has a multitude of (mutually non-exclusive) solutions to any given challenge. However, the model may well describe a dominant process because it allows for accumulation of non-deleterious mutations which are more likely to occur in a non-structurally constrained region than, e.g. in an α-helix. Accordingly, fully structured and functional proteins (or domains, in first place) do not need to be built from scratch. This is further supported by the dominant means of protein evolution which involves domain rearrangements. These in turn result from three major genetic processes; gene duplication, gene fusion and terminal domain loss by insertion of stop codons or loss of start codons [17].

Future research will need to broaden the type of analyses presented here to include more genomes and, in particular, more genomes from recently split species and diverging populations [57]. Also, the reconstruction of ancestral sequences [58] will help to capture the evolutionary trajectories along lineages in more detail and to understand how selection has acted on shaping the properties, such as disorder and hydrophobic clusters, of extant proteins. Finally, these reconstructed proteins and many of the extant proteins can be synthesized and subjected to scrutiny using, e.g. NMR or CD.

Upcoming insights will undoubtedly help obtain a better understanding how nature has shaped today's protein repertoire and possibly help to develop new strategies for protein design which exploit this knowledge about protein evolution.

Author contribution

Magdalena Heberlein provided data from a preliminary study and computational support. Jonathan Schmitz performed all computations. Erich Bornberg-Bauer conceived the study and wrote the paper. All authors interpreted data and edited the paper.

Acknowledgements

We thank Tristan Bitard-Feildel and Steffen Klasberg for computational support.

Funding statement

EBB's ORCID is 0000-0002-1826-3576, ResID is A-1563-2013. This work was supported by DFG [grant numbers BO2544/2-2 and HFSP RGP006/2013 (to E.B.B.)].

References

1 Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T.B., Montelione, G.T. and Baker, D. (2012) Principles for designing ideal protein structures. Nature **491**, 222–227 CrossRef PubMed

2 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 CrossRef PubMed

3 Dokholyan, N.V., Shakhnovich, B. and Shakhnovich, E.I. (2002) Expanding protein universe and its origin from the biological big bang. Proc. Natl. Acad. Sci. U.S.A. **99**, 14132–14136. PMID: CrossRef

4 Chothia, C. (1992) Proteins. one thousand families for the molecular biologist. Nature **357**, 543–544

5 Kolodny, R., Pereyaslavets, L., Samson, A.O. and Levitt, M. (2013) On the universe of protein folds. Annu. Rev. Biophys. **42**, 559–582 CrossRef PubMed

6 Yandell, M. and Ence, D. (2012) A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. **13**, 329–342 CrossRef PubMed

7 Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H. et al. (2014) Mass-spectrometry-based draft of the human proteome. Nature **509**, 582–587 CrossRef PubMed

8 Hamatsu, J., O'Donovan, D., Tanaka, T., Shirai, T., Hourai, Y., Mikawa, T., Ikeya, T., Mishima, M., Boucher, W., Smith, B.O. et al. (2013) High-resolution heteronuclear multidimensional NMR of proteins in living insect cells using a baculovirus protein expression system. J. Am. Chem. Soc. **135**, 1688–1691 CrossRef PubMed

9 Bryan, P.N. and Orban, J. (2010) Proteins that switch folds. Curr. Opin. Struct. Biol. **20**, 482–488 CrossRef PubMed

10 Dalal, S., Balasubramanian, S. and Regan, L. (1997) Protein alchemy: Changing β-sheet into α-helix. Nat. Struct. Mol. Biol. **4**, 548–552 CrossRef

11 Gambin, Y., Schug, A., Lemke, E.A., Lavinder, J.J., Ferreon, A.C.M., Magliery, T.J., Onuchic, J.N. and Deniz, A.A. (2009) Direct single-molecule observation of a protein living in two opposed native structures. Proc. Natl. Acad. Sci. U.S.A. **106**, 10153–10158 CrossRef PubMed

12 Farías-Rico, J.A., Schmidt, S. and Höcker, B. (2014) Evolutionary relationship of two ancient protein superfolds. Nat. Chem. Biol. **10**, 710–715 CrossRef PubMed

13 Alexander, P.A., He, Y., Chen, Y., Orban, J. and Bryan, P.N. (2009) A minimal sequence code for switching protein structure and function. Proc. Natl. Acad. Sci. U.S.A. **106**, 21149–21154 CrossRef PubMed

14 Sikosek, T., Bornberg-Bauer, E. and Chan, H.S. (2012) Evolutionary dynamics on protein bi-stability landscapes can potentially resolve adaptive conflicts. PLoS Comput. Biol. **8**, e1002659 CrossRef PubMed

15 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. U.S.A. **105**, 5057–5062 CrossRef PubMed

16 Forslund, K. and Sonnhammer, E. L.L. (2008) Predicting protein function from domain content. Bioinformatics **24**, 1681–1687 CrossRef PubMed

17 Moore, A.D., Björklund, Å.K., Ekman, D., Bornberg-Bauer, E. and Elofsson, A. (2008) Arrangements in the modular evolution of proteins. Trend. Biochem. Sci. **33**, 444–451 CrossRef PubMed

18 Bornberg-Bauer, E., Huylmans, A.-K. and Sikosek, T. (2010) How do new proteins arise? Curr. Opin. Struct. Biol. **20**, 390–396 CrossRef PubMed

19 Yu, Y. and Lutz, S. (2011) Circular permutation: a different way to engineer enzyme structure and function. Trends Biotechnol **29**, 18–25 CrossRef PubMed

20 Weiner, J., Thomas, G. and Bornberg-Bauer, E. (2005) Rapid motif-based prediction of circular permutations in multi-domain proteins. Bioinformatics **21**, 932–937 CrossRef PubMed

21 Terrapon, N., Weiner, J., Grath, S., Moore, A.D. and Bornberg-Bauer, E. (2014) Rapid similarity search of proteins using alignments of domain arrangements. Bioinformatics **30**, 274–281 CrossRef PubMed

22 Moore, A.D., Held, A., Terrapon, N., Weiner, J. and Bornberg-Bauer, E. (2014) DoMosaics: software for domain arrangement visualization and domain-centric analysis of proteins. Bioinformatics **30**, 282–283 CrossRef PubMed

23 Palmieri, N., Kosiol, C. and Schlötterer, C. (2014) The life cycle of drosophila orphan genes. Elife **3**, e01311 CrossRef PubMed

24 Wissler, L., Gadau, J., Simola, D.F., Helmkampf, M. and Bornberg-Bauer, E. (2013) Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biol. Evol. **5**, 439–455 CrossRef PubMed

25 Domazet-Loso, T. and Tautz, D. (2003) An evolutionary analysis of orphan genes in drosophila. Genome Res **13**, 2213–2219 CrossRef PubMed

26 Tautz, D. and Domazet-Lošo, T. (2011) The evolutionary origin of orphan genes. Nat. Rev. Genet. **12**, 692–702 CrossRef PubMed

27 Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., Charloteaux, B., Hidalgo, C.A., Barbette, J., Santhanam, B. et al. (2012) Proto-genes and *de novo* gene birth.. Nature **487**, 370–374 CrossRef PubMed

28 Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. and Bosch, T.C. (2009) More than just orphans: are taxonomically-restricted genes important in evolution? Trend. Genet. **25**, 404–413 CrossRef PubMed

29 Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X. and Albà, M.M. (2009) Origin of primate orphan genes: A comparative genomics approach. Mol. Biol. Evol. **26**, 603–612 CrossRef PubMed

30 Neme, R. and Tautz, D. (2013) Phylogenetic patterns of emergence of new genes support a model of frequent *de novo* evolution.. BMC Genomics. **14**, 117 CrossRef PubMed

31 Wilson, B.A. and Masel, J. (2011) Putatively noncoding transcripts show extensive association with ribosomes. Genome Biol. Evol. **3**, 1245–1252 CrossRef PubMed

32 Neme, R. and Tautz, D. (2015) Entire genome transcription across evolutionary time exposes non-coding DNA to *de novo* gene emergence.. bioRxiv 017152

33 DePristo, M.A., Weinreich, D.M. and Hartl, D.L. (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. Nat. Rev. Genet. **6**, 678–687 CrossRef PubMed

34 Davidson, A.R., Lumb, K.J. and Sauer, R.T. (1995) Cooperatively folded proteins in random sequence libraries. Nat. Struct. Mol. Biol. **2**, 856–864 CrossRef

35 Keefe, A.D. and Szostak, J.W. (2001) Functional proteins from a random-sequence library. Nature **410**, 715–718 CrossRef PubMed

36 Heinen, T.J. A.J., Staubach, F., Häming, D. and Tautz, D. (2009) Emergence of a new gene from an intergenic region. Curr. Biol. **19**, 1527–1531 CrossRef PubMed

37 Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S. et al. (2015) The landscape of long noncoding RNAs in the human transcriptome. Nat. Genet. **47**, 199–208 CrossRef PubMed

38 Lauressergues, D., Couzigou, J.-M., Clemente, H.S., Martinez, Y., Dunand, C., Bécard, G. and Combier, J.-P. (2015) Primary transcripts of microRNAs encode regulatory peptides. Nature **520**, 90–93 CrossRef PubMed

39 Zhao, L., Saelao, P., Jones, C.D. and Begun, D.J. (2014) Origin and spread of *de novo* genes in drosophila melanogaster populations. Science **343**, 769–772 CrossRef PubMed

40 Ohno, S. (1970) Evolution by gene duplication, 1st edn., Springer-Verlag, New York CrossRef

41 Weiner, J., Beaussart, F. and Bornberg-Bauer, E. (2006) Domain deletions and substitutions in the modular protein evolution. FEBS J **273**, 2037–2047 CrossRef PubMed

42 Bitard-Feildel, T., Heberlein, M., Bornberg-Bauer, E. and Callebaut, I Detection of orphan domains in drosophila using "hydrophobic cluster analysis. Biochimie, in the press

43 Rockah-Shmuel, L., Tóth-Petróczy, Á., Sela, A., Wurtzel, O., Sorek, R. and Tawfik, D.S. (2013) Correlated occurrence and bypass of frame-shifting insertion-deletions (InDels) to give functional proteins. PLoS Genet **9**, e1003882 CrossRef PubMed

44 Toll-Riera, M., Radó-Trilla, N., Martys, F. and Albà, M.M. (2012) Role of low-complexity sequences in the formation of novel protein coding sequences. Mol. Biol. Evol. **29**, 883–886 CrossRef PubMed

45 Radó-Trilla, N. and Albà, M. (2012) Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. BMC Evol. Biol. **12**, 155 CrossRef PubMed

46 Simon, M. and Hancock, J.M. (2009) Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. Genome Biol **10**, R59 CrossRef PubMed

47 Chouard, T. (2011) Structural biology: Breaking the protein rules. Nature **471**, 151–153 CrossRef

48 Marsh, J.A. and Teichmann, S.A. (2014) Protein flexibility facilitates quaternary structure assembly and evolution. PLoS Biol **12**, e1001870 CrossRef PubMed

49 Abrusán, G. (2013) Integration of new genes into cellular networks, and their structural maturation. Genetics **195**, 1407–1417 CrossRef PubMed

50 vander Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T. et al. (2014) Classification of intrinsically disordered regions and proteins. Chem. Rev. **114**, 6589–6631 CrossRef PubMed

51 Colak, R., Kim, T., Michaut, M., Sun, M., Irimia, M., Bellay, J., Myers, C.L., Blencowe, B.J. and Kim, P.M. (2013) Distinct types of disorder in the human proteome: functional implications for alternative splicing. PLoS Comput. Biol. **9**, e1003030 CrossRef PubMed

52 Espinoza-Fonseca, L.M., Kast, D. and Thomas, D.D. (2007) Molecular dynamics simulations reveal a disorder-to-order transition on phosphorylation of smooth muscle myosin. Biophys. J. **93**, 2083–2090 CrossRef PubMed

53 Metskas, L.A. and Rhoades, E. (2015) Folding upon phosphorylation: translational regulation by a disorder-to-order transition. Trends Biochem. Sci. **40**, 243–244 CrossRef PubMed

54 Tompa, P. (2011) Unstructural biology coming of age. Curr. Opin. Struct. Biol. **21**, 419–425 CrossRef PubMed

55 Bornberg-Bauer, E. and Albà, M.M. (2013) Dynamics and adaptive benefits of modular protein evolution. Curr. Opin. Struct. Biol. **23**, 459–466 CrossRef PubMed

56 Moore, A.D. and Bornberg-Bauer, E. (2012) The dynamics and evolutionary potential of domain loss and emergence. Mol. Biol. Evol. **29**, 787–796 CrossRef PubMed

57 Chain, F.J.J., Feulner, P.G.D., Panchal, M., Eizaguirre, C., Samonte, I.E., Kalbe, M., Lenz, T.L., Stoll, M., Bornberg-Bauer, E., Milinski, M. and Reusch, T.B.H. (2014) Extensive copy-number variation of young genes across stickleback populations. PLoS Genet **10**, e1004830 CrossRef PubMed

58 Harms, M.J. and Thornton, J.W. (2010) Analyzing protein structure and function using ancestral gene reconstruction. Curr. Opin. Struct. Biol. **20**, 360–366 CrossRef PubMed

59 Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J. et al. (2012) The pfam protein families database. Nucleic Acids Res. **40**, D290–D301 CrossRef PubMed

60 Dosztányi, Z., Csizmok, V., Tompa, P. and Simon, I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics **21**, 3433–3434 CrossRef PubMed

61 Faure, G. and Callebaut, I. (2013) Comprehensive repertoire of foldable regions within whole genomes. PLoS Comput. Biol. **9**, e1003280 CrossRef PubMed

62 Eudes, R., Tuan, K.L., Delettré, J., Mornon, J.-P. and Callebaut, I. (2007) A generalized analysis of hydrophobic and loop clusters within globular protein sequences. BMC Struct. Biol. **7**, 2 CrossRef PubMed

63 Waterhouse, R.M., Tegenfeldt, F., Li, J., Zdobnov, E.M. and Kriventseva, E.V. (2013) OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Res. **41**, D358–365 CrossRef PubMed

64 Abascal, F., Zardoya, R. and Telford, M.J. (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. **38**, W7–W13 CrossRef PubMed

65 Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**, 1792–1797 CrossRef PubMed

66 Massingham, T. and Goldman, N. (2005) Detecting amino acid sites under positive selection and purifying selection. Genetics **169**, 1753–1762 CrossRef PubMed

67 Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., Beutel, R.G., Niehuis, O. et al. (2014) Phylogenomics resolves the timing and pattern of insect evolution. Science **346**, 763–767 CrossRef PubMed

Received 27 April 2015
doi:10.1042/BST20150089