# Protein function space: viewing the limits or limited by our view?

*Jeroen Raes, Eoghan Donal Harrington, Amoolya Hardev Singh, and Peer Bork*

Published in final edited form as:
Current Opinion in Structural Biology. 2007 Jun ; 17(3): 362-369 | doi: 10.1016/j.sbi.2007.05.010
Elsevier (The Netherlands) ►

Final Draft

# Protein function space: viewing the limits or limited by our view?

*Jeroen Raes*[1], *Eoghan Donal Harrington*[1], *Amoolya Hardev Singh*[1], *and Peer Bork*[1]

[1] European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany

**ABSTRACT |** Given that the number of protein functions on earth is finite, the rapid expansion of biological knowledge and the concomitant exponential increase in the number of protein sequences should, at some point, enable the estimation of the limits of protein function space. The functional coverage of protein sequences can be investigated using computational methods, especially given the massive amount of data being generated by large-scale environmental sequencing (metagenomics). In completely sequenced genomes, the fraction of proteins to which at least some functional features can be assigned has recently risen to as much as ~85%. Although this fraction is more uncertain in metagenomics surveys, because of environmental complexities and differences in analysis protocols, our global knowledge of protein functions still appears to be considerable. However, when we consider protein families, continued sequencing seems to yield an ever-increasing number of novel families. Until we reconcile these two views, the limits of protein space will remain obscured.

## Introduction

Biological function is an abstract term that can be applied to different scales — from biomolecules via cells to species and large ecosystems [1]. Although biologists are comfortable talking about the concept of function, we often struggle when it comes to quantifying it. This is partly due to our limited understanding of the underlying biological processes, which prevents us from creating a semantic framework to describe our findings. In addition, a full description of biological function has to take into account both its temporal and its spatial aspects; this has been historically hampered by the lack of adequate data [1]. Today, the most important agents of biological function — proteins — are being catalogued on a variety of scales, from macromolecular complexes at the subcellular level (e.g. [2]) to complete organisms (e.g. [3,4]) and, more recently, even entire ecosystems [5]. Therefore, we can begin to investigate how complete our understanding of protein-based biological function is.

It is generally accepted that the molecular functions or cellular roles of proteins can be described as 'known' if they can be assigned some putative function based on established classification schemes [6,7]. However, the actual fraction of possible assignments has been controversial since the first large genomic sequences became available. Early examples of such discrepancies include the analyses of the first chromosome, yeast chromosome 3 ([8] versus [9]), and the first completely sequenced bacterial genome, *Haemophilus influenzae* ([10] versus [11]). The fraction of possible assignments depends significantly on the operational definition of function, on the sources of information used to infer function, on the methods of annotation or inference used, and on the choice of parameters. Indeed, differences in the choice of parameters alone are likely to have caused the 30% difference in domain-level functional assignments for the human genome provided by the two sequencing consortia ([12] versus [13]) in 2001.

Despite these caveats, computational analysis of sequence data is probably still the most effective way to explore the dimensions of protein function space, since higher-order function is much less understood and quantifiable. As a result of various improvements in the speed, scalability and price of sequencing technology [14,15], the volume of sequence data has increased exponentially in the past 10–20 years [16] and will continue to do so. Although a variety of other large-scale data augment our knowledge of protein function [3,4,17,18••,19-22], their coverage remains considerably lower. For example, only 25–31% of human gene products are covered by determined or predicted protein structures [23], and it is estimated that only 10% of the human interactome has been documented (compared to 50% in yeast) [24•]. The vast amount of available sequence information is complemented by the increased sensitivity of function prediction methods. Beyond classical homology-based methods [16,25,26], reliable functional information can be assigned using gene context-based approaches [27-29] and by integrating data from various sources [30-37,38•]. Even though these computational function prediction techniques currently do not provide spatial and temporal aspects (although proof-of-principle papers have been published [39-43]; see also [44], and Beltrao *et al.*, and Devos and Russell in this issue for reviews) and are purely descriptive, they do provide a first level of functional understanding. Therefore, we define a protein annotated this way as functionally characterizable.

Given this operational definition of function, here we will try to assess the completeness of protein function space using computational analysis of sequence data, scaling up from the model organism *Escherichia coli* via full genome sequences and complete environmental data sets.

## Function prediction in the model organism *E. coli*: do we know all of it already?

In 2003, the fraction of proteins in the *E. coli* proteome with functional assignments had steadily increased (using both homology- and context-based approaches) to about 70% [45]. The same analysis repeated today reveals that more than 80% of *E. coli* proteins have orthologs with known functions (e.g. functionally classified clusters of orthologous groups [COGs] [46] or functionally annotated in the KEGG [47] or Uniref [48] databases; Fig.1). Moreover, when gene neighborhood, the most powerful gene context method for prokaryotes [45], is also taken into account, functions can be reliably predicted for as many as 85% of *E. coli* proteins (using a stringent confidence score of 0.7 in the STRING framework for predicting interactions [38•]; Fig.1). Predictions that integrate all of the above with knowledge from literature mining, large-scale interaction data and curated interaction databases [38•] increase the fraction of

functionally characterizable proteins to 96%, suggesting that we currently seem to have at least some clue to the functions of almost all *E. coli* proteins. Although the annotations might contain some erroneous predictions, it seems that there is not much more to be discovered at this low-level resolution of functional description (as is currently standard in genome annotation).

## Function prediction in completely sequenced genomes: whereas the vast majority of proteins can be characterized, novel unknown families continue to be discovered

As *E. coli* is one of the best-studied model organisms, a more unbiased estimate of our functional knowledge can be obtained by performing similar predictions on all completely sequenced genomes to date (we used 373 genomes as included in the STRING 7 release; Fig.1). As expected, the fraction of functional predictions averaged for all genomes is lower than for *E. coli*. However, 73% of the gene products in the average genome are functionally characterizable by homology alone and integrating other prediction methods increases this fraction to an average of 85% (Fig.1). Archaebacteria are on the low end of this scale as they are generally less characterized, but some smaller bacteria outperform even *E. coli* (Fig.1). Regardless of outliers in both directions, we roughly know at least some aspect of the function of an overwhelming majority of proteins in sequenced genomes. This view is seemingly in sharp contrast to the rate of discovery of novel families and associated protein folds that accompanies newly sequenced complete genomes. For example, Marsden *et al.* [49•] recently repeated a protein family cluster analysis originally performed by Kunin *et al.* [50] and showed that the number of novel gene families keeps growing linearly over time. Despite the large fraction of gene products in each newly sequenced genome for which functions can be assigned, the number of small novel families appears endless, suggesting that we are nowhere near the limits of protein function space.

## Towards all proteins on earth: using environmental sequence data

It could be argued that the current set of completely sequenced genomes is still biased — in eukaryotes, towards (usually) fast-evolving model organisms and, in prokaryotes, towards medically relevant strains (often with small genomes). Furthermore, prokaryotic species usually need to be grown in culture before sequencing, which is only possible for 1% of all species [51,52]. This has led to the current situation, in which fully sequenced microbial genomes represent only a miniscule fraction of all extant microbial species. There has been recent hope, however, that such biases might disappear in the near future as a result of direct large-scale environmental sequencing. The massive random shotgun sequencing of entire ecosystems, pioneered by Venter *et al.* [53] and Tyson *et al.* [54], has provided the first large-scale glimpse of the protein space of uncultured organisms and entire communities of species. Since then, several additional large-scale environmental sequencing (metagenomics) studies have been published [55-59,60••,61,62,63•]. These studies not only reveal astonishing insights into species diversity in different environments, but also are impressive just for their sheer magnitude (Fig.2) and the associated potential functional information. For instance, when the Sargasso Sea data set was released in 2004, it almost doubled the size of public protein databases [64]. Indeed, the published studies in three years of metagenomics sequencing to date have yielded a total of about four times the number of open reading frames (ORFs) produced by all genome sequencing projects (including the human genome) in their twelve years of existence (Fig.2). Moreover, given that there are currently at least 50 metagenomics projects with increasing sequence output in the pipeline [65•] and that sequencing costs continue to decrease rapidly, the associated data accumulation still appears to be in an early phase of steep exponential growth.

## Estimating novelty in environments: how to compare apples and oranges?

To estimate the impact of metagenomics sequencing on our views of protein function space, we will first try to quantify the amount of associated novelty. In absolute terms, this number is overwhelming. Based on the numbers extracted from the original reports (defining novelty as an unassignable function; Table 1), these projects have yielded almost a million 'novel' proteins so far. In relative terms, however, the fraction of reported novelty varies greatly among samples, ranging from 50% in the metagenome of communities living on deep-sea whale carcasses ('whale falls') to 75% in the mouse gut (note that most of the latter is probably due to a high fraction of data from pyrosequencing, resulting in numerous short reads of ~75 bp, compared to ~700–800 bp reads coming from classical Sanger sequencing [61]). In general, short unannotated shotgun data negatively impact homology-based function prediction as they might decrease the significance of pairwise similarities because of added noise [66]. This resembles the situation in the 1990s, when growing databases and the accompanying lower sensitivity of homology-based function prediction methods were compensated by better methods and more experimentally characterized proteins [25].

The degree of functional annotation and its variation among samples can be attributed to several factors, such as distinct phylogenetic [67] and functional [60••] complexity between environments, as well as different sampling protocols. Nevertheless, differences in annotation processes alone make direct comparisons between studies difficult. For example, the eight large-scale metagenomics studies published since 2004 (Table 1) use four assembly methods, six distinct methods of ORF calling, two distinct methods of function prediction, and almost twenty different sequence and function databases (Table 1), not to mention a plethora of different cutoffs and parameters (data not shown).

When a simple but uniform measure based on gene family clustering is used to roughly estimate the novelty yield per sample (i.e. the number of previously unseen gene families per ORF sequenced; Fig.3), clear differences can still be observed. Although these could be linked to technical issues, such as sequencing depth (leading to more assembly and larger contigs, which, in turn, improve ORF calling), they could also be due to biological factors, such as ecological complexity (e.g.

most novelty is in functionally complex habitats, such as soil [68]) or average genome size per sample (larger genomes harbor a smaller relative fraction of universal and housekeeping genes, and thus a greater chance of novelty [69,70]; (Fig.3). Indeed, there seems to be a weakly significant positive correlation between the effective genome size (EGS) and the potential for novelty (Spearmans rho = 1, p = 0.08; Fig.3). These results suggest the possibility of maximizing the amount of novelty yielded by selecting specific environments for sequencing.

## Unannotated ORFs: technical limits or limited knowledge?

Although an estimate of the unknown biology on earth is intellectually appealing, the sequences of novel ORFs alone tell us little of their function and role in the environment. Directed community approaches for systematic large-scale experimental protein characterization must follow, as has already been proposed and initiated for genome annotation and structural proteomics [71-73]. Without them, functional annotation of novel proteins or families by prediction alone is difficult. Indeed, given this wealth of novel proteins, detailed analysis of metagenomics sequencing has thus far yielded few discoveries of truly novel functions. Two exceptions come from studies in which metagenomics sequencing enabled the reconstruction of a complete genome. A study of an anaerobic ammonium oxidation (anammox) community dominated by *K. stuttgartiensis* proposed novel candidate genes for hydrazine metabolism and ladderane biosynthesis by combining operon analysis with the reconstruction of metabolic pathways from the genome sequence [59]. Likewise, Garcia Martin *et al.* [56], in their analysis of the reconstructed *A. phosphatis* genome from two wastewater sludge communities, discovered a novel quinol-NAD(P) reductase fusion protein that enabled a fully anaerobic Krebs cycle. In a larger-scale analysis, about 90 novel families from the Global Ocean Survey (GOS) study were attributed a Gene Ontology-based classification through neighborhood and sequence analysis [63•]. In any case, given the total number of reported novel proteins from currently available metagenomics data (see above), this discovery of 93 novel functions seems disappointingly low.

## Function prediction in environmental samples: lots of novelty, but really endless?

Despite the, in absolute terms, vast amount of novelty in complex metagenomes and the little we can currently do to characterize it, functions seem to be reliably predictable for the majority of proteins. This is possible despite the fragmentary nature of the underlying sequences and the fact that we are using information from biased genome sequences to annotate sequences from an (almost) unbiased sample of natural habitats. Our analyses of the first four samples that had been sequenced [53,54,60••] suggest that, using a combination of homology- and neighborhood-based approaches, functions can be assigned to at least two-thirds of the predicted proteins (Fig.3; ED Harrington *et al.*, unpublished).

To investigate the dependence of function prediction on sequencing depth and to assess the possibility of saturation in functional assignment (given that most widespread protein families are characterized already), we simulated a meta-environment consisting of all sequenced genomes and all four habitats (Fig.4). On the one hand, it appears that protein families containing members previously seen in genome projects are identifiable with relatively little data. On the other hand, uncharacterized 'singletons' are rarely saturating, that is, each genome or environment contains a lot of uncharacterizable ORFs (or ORF fragments). That novel families grow linearly if more ocean samples are added was also a major conclusion of a recent metagenomics ocean survey [63•], when applying a similar method to that described in [49•,50]. However, this observed 'novelty', based on the absence of homology to anything we know, could also be inflated by gene fragments that are too short to be recognized or by spurious gene predictions (although stringent criteria were applied in this study [63•]).

## Conclusions

As with completely sequenced genomes, there seem to be two possible views on functional completeness: first, that we can reliably predict functions for the majority of proteins; or second, that there is a seemingly endless repertoire of specialized families and we cannot predict whether we are approaching the limits of protein function space. Particularly in the field of metagenomics, it is still early to draw conclusions about the dimensions of protein function space on earth, even within the simplistic framework of function used here, until the different analysis pipelines are reconciled, platforms are developed and analysis methodology is improved. So far, bioinformatics research on metagenomics data is still in its infancy, trying to adapt to the overwhelming amount of data [64]. Indeed, the first web-accessible function analysis tools for the non-bioinformatics community have appeared only very recently [74,75] and usually consist of precomputed homology-based predictions. With commonly accepted frameworks and comparative analysis tools, one might be able to revisit the exciting question of where we stand regards our functional knowledge of proteins and arrive at the first realistic estimates of the protein function repertoire on earth.

## Acknowledgements

The authors would like to thank Lars Jensen and the other members of the Bork group for stimulating discussions, and apologize to all colleagues whose work could not be included because of space constraints. This work was supported by the EU 6th Framework Program (GeneFun grant contract number LSHG-CT-2004-503567). EDH is funded by the EC FP6 Marie Curie Fellowship for Early Stage Training (E-STAR) under contract number MEST-CT-2004-504640.

## Corresponding Author

Peer Bork, e-Mail: bork@embl.de

## References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest


1. P. Bork, T. Dandekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. Huynen and Y. Yuan, Predicting function: from genes to genomes and back, *J Mol Biol* 283 (1998), pp. 707–725.
2. H. Prokisch, C. Scharfe, D.G. Camp II, W. Xiao, L. David, C. Andreoli, M.E. Monroe, R.J. Moore, M.A. Gritsenko and C. Kozany *et al.*, Integrative analysis of the mitochondrial proteome in yeast, *PLoS Biol* 2 (2004), p. e160.
3. A.C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L.J. Jensen, S. Bastuck and B. Dumpelfeld *et al.*, Proteome survey reveals modularity of the yeast cell machinery, *Nature* 440 (2006), pp. 631–636.
4. N.J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta and A.P. Tikuisis *et al.*, Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*, *Nature* 440 (2006), pp. 637–643.
5. R.J. Ram, N.C. Verberkmoes, M.P. Thelen, G.W. Tyson, B.J. Baker, R.C. Blake, I.I. Shah, M. Hettich and R.L. Banfield JF, Community proteomics of a natural microbial biofilm, *Science* 308 (2005), pp. 1915–1920.
6. M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight and J.T. Eppig *et al.*, Gene ontology: tool for the unification of biology, *The Gene Ontology Consortium. Nat Genet* 25 (2000), pp. 25–29.
7. C.A. Ouzounis, R.M. Coulson, A.J. Enright, V. Kunin and J.B. Pereira-Leal, Classification schemes for protein structure and function, *Nat Rev Genet* 4 (2003), pp. 508–519.
8. S.G. Oliver, Q.J. van der Aart, M.L. Agostoni-Carbone, M. Aigle, L. Alberghina, D. Alexandraki, G. Antoine, R. Anwar, J.P. Ballesta and P. Benit *et al.*, The complete DNA sequence of yeast chromosome III, *Nature* 357 (1992), pp. 38–46.
9. P. Bork, C. Ouzounis, C. Sander, M. Scharf, R. Schneider and E. Sonnhammer, Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome III, *Protein Sci* 1 (1992), pp. 1677–1690.
10. R.D. Fleischmann, M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty and J.M. Merrick *et al.*, Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd, *Science* 269 (1995), pp. 496–512.
11. G. Casari, M.A. Andrade, P. Bork, J. Boyle, A. Daruvar, C. Ouzounis, R. Schneider, J. Tamames, A. Valencia and C. Sander, Challenging times for bioinformatics, *Nature* 376 (1995), pp. 647–648.
12. E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle and W. FitzHugh *et al.*, Initial sequencing and analysis of the human genome, *Nature* 409 (2001), pp. 860–921.
13. J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans and R.A. Holt *et al.*, The sequence of the human genome, *Science* 291 (2001), pp. 1304–1351.
14. H. Bayley, Sequencing single molecules of DNA, *Curr Opin Chem Biol* 10 (2006), pp. 628–637.
15. M.L. Metzker, Emerging technologies in DNA sequencing, *Genome Res* 15 (2005), pp. 1767–1776.
16. M. Kanehisa and P. Bork, Bioinformatics in the post-sequence era, *Nat Genet* 33 (suppl) (2003), pp. 305–310.
17. J.M. Chandonia and S.E. Brenner, The impact of structural genomics: expectations and outcomes, *Science* 311 (2006), pp. 347–351.
18. ••S.R. Collins, K.M. Miller, N.L. Maas, A. Roguev, J. Fillingham, C.S. Chu, M. Schuldiner, M. Gebbia, J. Recht and M. Shales *et al.*, Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map, *Nature* 446 (2007), pp. 806–810.
An intriguing approach to distinguish functional interactions between proteins from genetic interactions between the underlying genes
19. W.K. Huh, J.V. Falvo, L.C. Gerke, A.S. Carroll, R.W. Howson, J.S. Weissman and E.K. O'Shea, Global analysis of protein localization in budding yeast, *Nature* 425 (2003), pp. 686–691.
20. B. Neumann, M. Held, U. Liebel, H. Erfle, P. Rogers, R. Pepperkok and J. Ellenberg, High-throughput RNAi screening by time-lapse imaging of live human cells, *Nat Methods* 3 (2006), pp. 385–390.
21. J.F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G.F. Berriz, F.D. Gibbons, M. Dreze and N. Ayivi-Guedehoussou *et al.*, Towards a proteome-scale map of the human protein-protein interaction network, *Nature* 437 (2005), pp. 1173–1178.
22. U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F.H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr and S. Koeppen *et al.*, A human protein-protein interaction network: a resource for annotating the proteome, *Cell* 122 (2005), pp. 957–968.
23. L. Xie and P.E. Bourne, Functional coverage of the human genome by existing structures, structural genomics targets, and homology models, *PLoS Comput Biol* 1 (2005), p. e31.
24. G.T. Hart, A.K. Ramani and E.M. Marcotte, How complete are current yeast and human protein-interaction networks?, *Genome Biol* 7 (2006), p. 120.
An interesting study that poses a similar question to this review, but focused on the level of interactions.
25. • P. Bork and E.V. Koonin, Predicting functions from protein sequences–where are the bottlenecks?, *Nat Genet* 18 (1998), pp. 313–318.
26. E.V. Koonin, R.L. Tatusov and M.Y. Galperin, Beyond complete genomes: from sequence to structure and function, *Curr Opin Struct Biol* 8 (1998), pp. 355–363.
27. J.A. Eisen, Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis, *Genome Res* 8 (1998), pp. 163–167.
28. M. Huynen, B. Snel, W. Lathe III and P. Bork, Predicting protein function by genomic context: quantitative evaluation and qualitative inferences, *Genome Res* 10 (2000), pp. 1204–1210.
29. R.T. van der Heijden, B. Snel, V. van Noort and M.A. Huynen, Orthology prediction at scalable resolution by phylogenetic tree analysis, *BMC Bioinformatics* 8 (2007), p. 83.
30. E.M. Marcotte, M. Pellegrini, M.J. Thompson, T.O. Yeates and D. Eisenberg, A combined algorithm for genome-wide prediction of protein function, *Nature* 402 (1999), pp. 83–86.
31. M.A. Huynen, C.A. Spronk, T. Gabaldon and B. Snel, Combining data from genomes, Y2H and 3D structure indicates that BolA is a reductase interacting with a glutaredoxin, *FEBS Lett* 579 (2005), pp. 591–596.
32. W. Zhong and P.W. Sternberg, Genome-wide prediction of *C. elegans* genetic interactions, *Science* 311 (2006), pp. 1481–1484.

33. I. Friedberg, Automated protein function prediction–the genomic challenge, *Brief Bioinform* 7 (2006), pp. 225–242.

34. T. Hulsen, M.A. Huynen, J. de Vlieg and P.M. Groenen, Benchmarking ortholog identification methods using functional genomics data, *Genome Biol* 7 (2006), p. R31.

35. S. Bandyopadhyay, R. Sharan and T. Ideker, Systematic identification of functional orthologs based on protein network comparison, *Genome Res* 16 (2006), pp. 428–435.

36. N. Krishnamurthy, D.P. Brown, D. Kirshner and K. Sjolander, PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification, *Genome Biol* 7 (2006), p. R83.

37. J.D. Watson, S. Sanderson, A. Ezersky, A. Savchenko, A. Edwards, C. Orengo, A. Joachimiak, R.A. Laskowski and J.M. Thornton, Towards fully automated structure-based function prediction in structural genomics: a case study, *J Mol Biol* 367 (2007), pp. 1511–1522.

38. C. von Mering, L.J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Kruger, B. Snel and P. Bork, STRING 7–recent developments in the integration and prediction of protein interactions, *Nucleic Acids Res* 35 (2007), pp. D358–D362.
The STRING database combines an up-to-date orthologous group (OG) resource (~43 000 OGs from 373 organisms in STRING 7 versus ~10 000 OGs from 73 organisms provided by the NCBI [May 2007]) with integrated functional annotation. It is therefore well suited for metagenome annotation and the study of functional completeness.

39. M.N. Arbeitman, E.E. Furlong, F. Imam, E. Johnson, B.H. Null, B.S. Baker, M.A. Krasnow, M.P. Scott, R.W. Davis and K.P. White, Gene expression during the life cycle of *Drosophila melanogaster*, *Science* 297 (2002), pp. 2270–2275.

40. • U. de Lichtenberg, L.J. Jensen, S. Brunak and P. Bork, Dynamic complex formation during the yeast cell cycle, *Science* 307 (2005), pp. 724–727.

41. M.L. Dequeant, E. Glynn, K. Gaudenz, M. Wahl, J. Chen, A. Mushegian and O. Pourquie, A complex oscillating network of signaling genes underlies the mouse segmentation clock, *Science* 314 (2006), pp. 1595–1598.

42. L.J. Jensen, T.S. Jensen, U. de Lichtenberg, S. Brunak and P. Bork, Co-evolution of transcriptional and post-translational cell-cycle regulation, *Nature* 443 (2006), pp. 594–597.

43. S.D. Hooper, S. Boue, R. Krause, L.J. Jensen, C.E. Mason, M. Ghanim, K.P. White, E.E. Furlong and P. Bork, Identification of tightly regulated groups of genes during *Drosophila melanogaster* embryogenesis, *Mol Syst Biol* 3 (2007), p. 72.

44. P. Bork and L. Serrano, Towards cellular systems in 4D, *Cell* 121 (2005), pp. 507–509.

45. M.A. Huynen, B. Snel, C. von Mering and P. Bork, Function prediction and protein networks, *Curr Opin Cell Biol* 15 (2003), pp. 191–198.

46. R.L. Tatusov, N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumder, S.L. Mekhedov and A.N. Nikolskaya *et al.*, The COG database: an updated version includes eukaryotes, *BMC Bioinformatics* 4 (2003), p. 41.

47. M. Kanehisa, S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki and M. Hirakawa, From genomics to chemical genomics: new developments in KEGG, *Nucleic Acids Res* 34 (2006), pp. D354–D357.

48. B.E. Suzek, H. Huang, P. McGarvey, R. Mazumder and C.H. Wu, UniRef: comprehensive and non-redundant uniprot reference clusters, *Bioinformatics* (2007).

49. R.L. Marsden, D. Lee, M. Maibaum, C. Yeats and C.A. Orengo, Comprehensive genome analysis of 203 genomes provides structural genomics with new insights into protein family space, *Nucleic Acids Res* 34 (2006), pp. 1066–1080.
An update of the Kunin *et al.* [50] study, showing that protein space still grows with each added genome.

50. V. Kunin, I. Cases, A.J. Enright, V. de Lorenzo and C.A. Ouzounis, Myriads of protein families, and still counting, *Genome Biol* 4 (2003), p. 401.

51. J.T. Staley and A. Konopka, Measurement of *in situ* activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats, *Annu Rev Microbiol* 39 (1985), pp. 321–346.

52. • P. Hugenholtz, Exploring prokaryotic diversity in the genomic era, *Genome Biol* 3 (2002) REVIEWS0003.

53. J.C. Venter, K. Remington, J.F. Heidelberg, A.L. Halpern, D. Rusch, J.A. Eisen, D. Wu, I. Paulsen, K.E. Nelson and W. Nelson *et al.*, Environmental genome shotgun sequencing of the Sargasso Sea, *Science* 304 (2004), pp. 66–74.

54. G.W. Tyson, J. Chapman, P. Hugenholtz, E.E. Allen, R.J. Ram, P.M. Richardson, V.V. Solovyev, E.M. Rubin, D.S. Rokhsar and J.F. Banfield, Community structure and metabolism through reconstruction of microbial genomes from the environment, *Nature* 428 (2004), pp. 37–43.

55. E.F. DeLong, C.M. Preston, T. Mincer, V. Rich, S.J. Hallam, N.U. Frigaard, A. Martinez, M.B. Sullivan, R. Edwards and B.R. Brito *et al.*, Community genomics among stratified microbial assemblages in the ocean's interior, *Science* 311 (2006), pp. 496–503.

56. H. Garcia Martin, N. Ivanova, V. Kunin, F. Warnecke, K.W. Barry, A.C. McHardy, C. Yeates, S. He, A.A. Salamov and E. Szeto *et al.*, Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities, *Nat Biotechnol* 24 (2006), pp. 1263–1269.

57. S.R. Gill, M. Pop, R.T. Deboy, P.B. Eckburg, P.J. Turnbaugh, B.S. Samuel, J.I. Gordon, D.A. Relman, C.M. Fraser-Liggett and K.E. Nelson, Metagenomic analysis of the human distal gut microbiome, *Science* 312 (2006), pp. 1355–1359.

58. S.J. Hallam, N. Putnam, C.M. Preston, J.C. Detter, D. Rokhsar, P.M. Richardson and E.F. DeLong, Reverse methanogenesis: testing the hypothesis with environmental genomics, *Science* 305 (2004), pp. 1457–1462.

59. M. Strous, E. Pelletier, S. Mangenot, T. Rattei, A. Lehner, M.W. Taylor, M. Horn, H. Daims, D. Bartol-Mavel and P. Wincker *et al.*, Deciphering the evolution and metabolism of an anammox bacterium from a community genome, *Nature* 440 (2006), pp. 790–794.

60. S.G. Tringe, C. von Mering, A. Kobayashi, A.A. Salamov, K. Chen, H.W. Chang, M. Podar, J.M. Short, E.J. Mathur and J.C. Detter *et al.*, Comparative metagenomics of microbial communities, *Science* 308 (2005), pp. 554–557.
The first large-scale paper comparing the protein content of different environments.

61. P.J. Turnbaugh, R.E. Ley, M.A. Mahowald, V. Magrini, E.R. Mardis and J.I. Gordon, An obesity-associated gut microbiome with increased capacity for energy harvest, *Nature* 444 (2006), pp. 1027–1031.

62. D.B. Rusch, A.L. Halpern, G. Sutton, K.B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J.A. Eisen, J.M. Hoffman and K. Remington *et al.*, The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific, *PLoS Biol* 5 (2007), p. e77.

63. S. Yooseph, G. Sutton, D.B. Rusch, A.L. Halpern, S.J. Williamson, K. Remington, J.A. Eisen, K.B. Heidelberg, G. Manning and W. Li *et al.*, The Sorcerer II Global Ocean Sampling Expedition: expanding the universe of protein families, *PLoS Biol* 5 (2007), p. e16.

This study surprisingly shows a still linear increase in the number of protein families, even after adding six million proteins to the known combined proteome.

64. ••M.Y. Galperin, Metagenomics: from acid mine to shining sea, *Environ Microbiol* 6 (2004), pp. 543–545.

65. K. Liolios, N. Tavernarakis, P. Hugenholtz and N.C. Kyrpides, The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide, *Nucleic Acids Res* 34 (2006), pp. D332–D334.
A valuable resource that keeps a finger on the pulse of sequencing projects.

66. M.L. Tress, D. Cozzetto, A. Tramontano and A. Valencia, An analysis of the Sargasso Sea resource and the consequences for database composition, *BMC Bioinformatics* 7 (2006), p. 213.

67. C. von Mering, P. Hugenholtz, J. Raes, S.G. Tringe, T. Doerks, L.J. Jensen, N. Ward and P. Bork, Quantitative phylogenetic assessment of microbial communities in diverse environments, *Science* 315 (2007), pp. 1126–1130.

68. • R. Daniel, The metagenomics of soil, *Nat Rev Microbiol* 3 (2005), pp. 470–478.

69. J. Raes, J.O. Korbel, M.J. Lercher, C. von Mering and P. Bork, Prediction of effective genome size in metagenomic samples, *Genome Biol* 8 (2007), p. R10.

70. E. van Nimwegen, Scaling laws in the functional content of genomes, *Trends Genet* 19 (2003), pp. 479–484.

71. • P.D. Karp, Call for an enzyme genomics initiative, *Genome Biol* 5 (2004), p. 401.

72. R.J. Roberts, Identifying protein function–a call for community action, *PLoS Biol* 2 (2004), p. E42.

73. A.F. Yakunin, A.A. Yee, A. Savchenko, A.M. Edwards and C.H. Arrowsmith, Structural proteomics: a tool for genome annotation, *Curr Opin Chem Biol* 8 (2004), pp. 42–48.

74. V.M. Markowitz, N. Ivanova, K. Palaniappan, E. Szeto, F. Korzeniewski, A. Lykidis, I. Anderson, K. Mavromatis, V. Kunin and H. Garcia Martin *et al.*, An experimental metagenome data management and analysis system, *Bioinformatics* 22 (2006), pp. e359–e367.

75. R. Seshadri, S.A. Kravitz, L. Smarr, P. Gilna and M. Frazier, CAMERA: a community resource for metagenomics, *PLoS Biol* 5 (2007), p. e75.

76. S. Van Dongen, Graph Clustering by Flow Simulation, University of Utrecht (2000).

77. F.E. Angly, B. Felts, M. Breitbart, P. Salamon, R.A. Edwards, C. Carlson, A.M. Chan, M. Haynes, S. Kelley and H. Liu *et al.*, The marine viromes of four oceanic regions, *PLoS Biol* 4 (2007), p. e368.

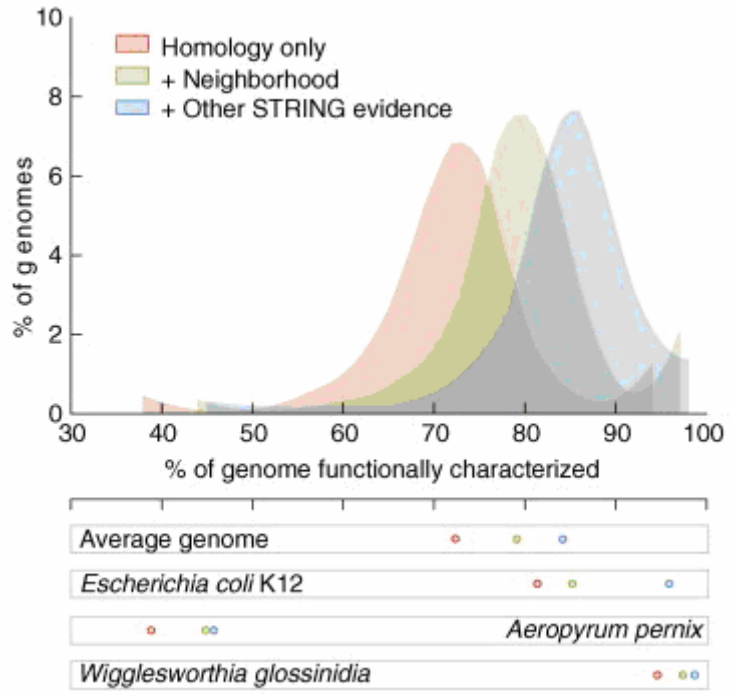


**Figure 1.** Assessment of novelty in fully sequenced genomes by computational methods. Our knowledge of function space is unevenly spread across the tree of life. The 338 prokaryotic genomes in the STRING database (version 7) were classified according to the proportion of proteins for which some inference of function is possible using three different criteria. Using simple homology, we considered functional inference possible for a protein if it can be mapped to a KEGG pathway, a characterized COG or a UniRef90 cluster. We then added neighborhood evidence with a score greater than 0.7 from the STRING database to infer function for those proteins in the same neighborhood as those characterized by homology. Similarly, we added all combined evidence from STRING to infer function for the remaining proteins.

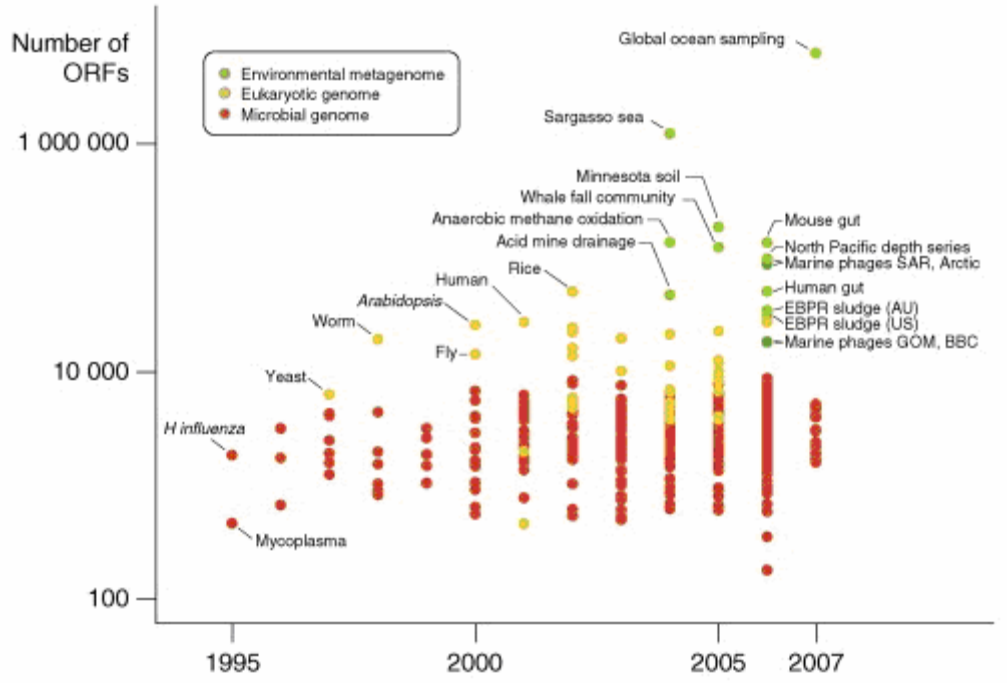


**Figure 2.** Number of ORFs generated by genome sequencing projects (red: bacteria, orange: eukaryotic) and metagenomics projects (light green: microbial, dark green: viral). Data were taken from the GOLD database [65*].

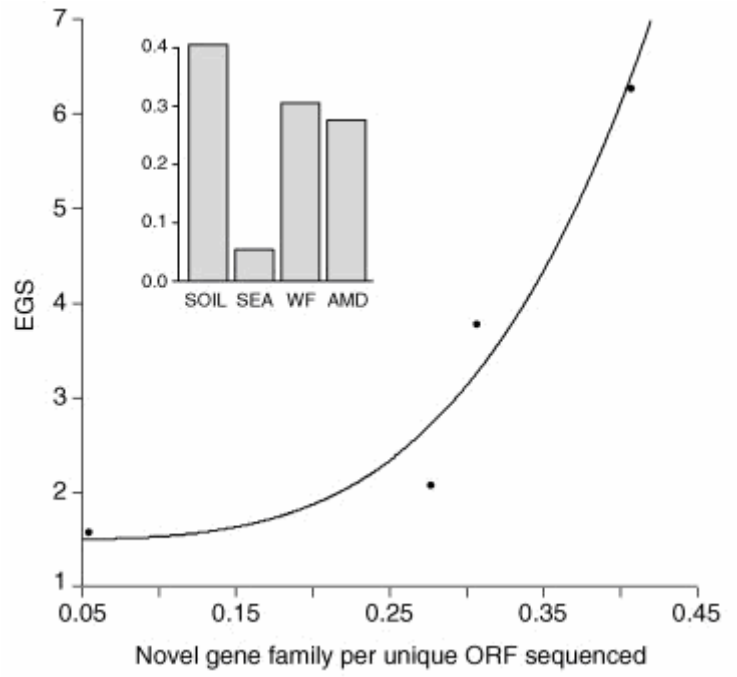


**Figure 3.** Analysis of novelty in metagenomics samples based on family clustering of the predicted proteins of four samples (Sargasso Sea [SEA], whale-fall [WF], acid mine drainage [AMD], Minnesota soil [SOIL]) and the predicted proteins of complete genomes in the STRING database (version 6.3; clustering performed with MCL [76] at I = 1.1, BLAST bit score > 60). The insert shows the number of previously unseen gene families (i.e. those without a member within the STRING database) per unique ORF sequenced — an estimate of the number of novel proteins discovered when compared to known genomes. The main figure shows the correlation between this measure and the EGS of each sample. This relationship fits a power law ($y = 1.51 + 128.70x^{3.63}$, $R^2 = 0.85$).

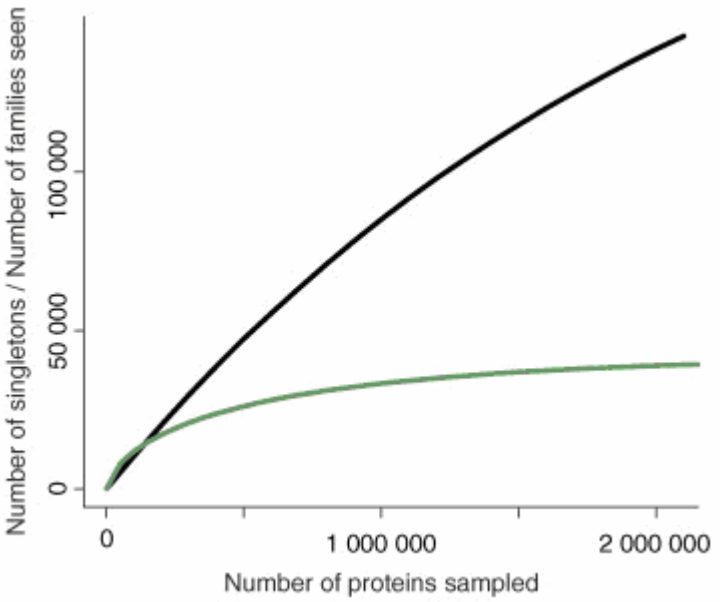


**Figure 4.** Analysis of gene family and singleton enrichment when sampling random proteins (with replacement) from four metagenomics data sets and the proteomes of fully sequenced genomes. Green, previously seen gene families (see Fig.3 for details); black, singletons (uncharacterizable ORFs).

**Table 1. Methods used in selected large-scale metagenomics studies.**

| Publication year | Environment (location) | Number of OFRs (Mpb) | Number of novel ORFs (% of metagenome) | ORF-calling procedure (database searched) | Functional annotation procedure (database searched) | References |
|---|---|---|---|---|---|---|
| 2004 | Acid mine (California) | 46 862 (76) | 34 301 (73.2%) | FGENESB pipeline (nr) | blastp (COG, nr) | [54] |
| 2004 | Surface sea water (Sargasso Sea, samples 1–4) | 1 001 987 (779) | 649 608 (64.8%) | Evidence based, using translation start and stop sites (bacterial portion of nraa) | blast (TIGR Role Category) | [53] |
| 2005 | Deep-sea whale-fall (Pacific, Antarctic) | 122 147 (75) | 63 021 (51.6%) | FGENESB pipeline (nraa) | blastp (extCOG version 6, KEGG) | [60••] |
| 2005 | Farm soil (Minnesota) | 183 536 (100) | 114 301 (62.3%) | FGENESB pipeline (nraa) | blastp (extCOG version 6, KEGG) | [60••] |
| 2006 | Subtropical ocean gyre (North Pacific) | NA[a] (64) | – | ORF calling not performed | blastx, blastp, blastn (KEGG, SEED, COG, Sargasso data set) | [55] |
| 2003 | Four oceanic viral metagenomes | NA[a] (181) | – | ORF calling not performed | blastx, tblastx, tblastn (SEED, Environments) | [77] |
| 2006 | Human gut | 50 164 (78) | 34 504 (68.8%) | Evidence based, using translation start and stop sites (AllGroup.niaa, in-house non-redundant protein repository[b]) | blast (COG, KEGG, STRING) | [57] |
| 2006 | Wastewater sludge (US, Australia) | 65 328 (176) | 47 032 (72.0%) | FGENESB pipeline (nraa) | IMG/M pipeline (KEGG) | [56] |
| 2006 | Mouse gut | 134 189 (160) | 100 599 (75.0%) | Glimmer software tool version 3.01 (InterPro) | blastx (nr, extCOG version 6.3, KEGG) | [61] |
| 2007 | Global Ocean Sampling (Northwest Atlantic through Eastern Tropical Pacific, including Sargasso Sea) | 6 123 395 (6250) | 95 455 (~1.5%)[c] | Using translation start and stop sites, dense subgraph clustering, filtering of shadow ORFs and non-coding sequences | blastp (nr, SWISS-PROT, PDF, PIR, PRF, TIGR Gene Indices, Ensembl, psi-blast of profile HMMs (TIGRFAM) | [62,63] |

[a] ORF calling not performed.
[b] Sourced from GenBank, Uniprot, Protein Research Foundation, Protein Data Bank and Omnium.
[c] S Yooseph, personal communication. The values for the GOS survey should be considered as lower bounds and are difficult to compare with other samples, as only a subfraction of the data was considered (only non-redundant proteins belonging to GOS-only gene families >20 members ['type-II clusters']).