Dark proteins important for cellular function

Andrea Schafferhans [1,2,*], Seán I. O'Donoghue [3,4,5], Michael Heinzinger [1], & Burkhard Rost [1,6]

1 TUM (Technical University of Munich) Department of Informatics, Bioinformatics & Computational Biology - i12, Boltzmannstr. 3, 85748 Garching/Munich, Germany

2 Department of Bioengineering Sciences, University of Applied Sciences, Freising, Germany

3 CSIRO Data61, Sydney, Australia

4 Division of Genomics & Epigenetics, Garvan Institute of Medical Research, Sydney, Australia

5 School of Biotechnology & Biomolecular Sciences, University of New South Wales (UNSW), Sydney, NSW, Australia

6 Institute for Advanced Study (TUM-IAS), Lichtenbergstr. 2a, 85748 Garching/Munich & TUM School of Life Sciences Weihenstephan (WZW), Alte Akademie 8, Freising, Germany

* Corresponding author: schafferhans@rostlab.org, http://www.rostlab.org/
Tel: +49-289-17-811 (email rost: assistant@rostlab.org)

Abstract

Despite substantial and successful projects for structural genomics, many proteins remain for which neither experimental structures nor homology-based models are known for any part of the amino acid sequence. These have been called *dark proteins*, in contrast to non-dark proteins, in which at least part of the sequence has a known or inferred structure. We hypothesized that non-dark proteins may be more abundantly expressed than dark proteins which are known to have much fewer sequence relatives. Surprisingly, we observed the opposite: human dark and non-dark proteins had quite similar levels of expression, in terms of both mRNA and protein abundance. Such high levels of expression strongly indicate that dark proteins – as a group - are important for cellular function. This is remarkable, given how carefully structural biologists have focused on proteins crucial for function, and highlights the important challenge posed by dark proteins in future research.

Key words: Dark proteome, Dark proteins, Protein expression, Intrinsically disordered proteins, Transmembrane proteins.

Abbreviations used: 3D, three-dimensional; **3D structure**, three-dimensional coordinates of protein structure;

Received: 06 19, 2018; Revised: 09 14, 2018; Accepted: 10 12, 2018

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pmic.201800227.

Statement of significance:

This study takes a detailed look at the expression of so-called dark proteins, i.e. those for which neither experimental structures nor homology-based models are known for any part of the amino acid sequence. The dark proteins in human are compared against non-dark proteins, i.e., all other proteins, in which at least part of the sequence has a known or inferred structure. At the beginning of the study, we originally hypothesized that non-dark proteins may be more abundantly expressed than dark proteins which are known to have much fewer sequence relatives. Surprisingly, we observed the opposite: human dark and non-dark proteins had quite similar overall levels of expression, in terms of both mRNA and protein abundance. This strongly indicates that dark proteins – as a group – are important for cellular function. This is remarkable, given how carefully structural biologists have focused on proteins crucial for function, and highlights the important challenge posed by dark proteins in future research.

Introduction

Genome sequencing is transforming biomedical research, but it is only one of many steps needed to understand biological systems. Subsequent steps include the determination of three-dimensional (3D) structures and functions for all proteins, and the study of protein interactions. One long-term goal of Structural Genomics (SG) has been to make high-resolution 3D atomic-level structures easily obtainable for most proteins from their corresponding DNA sequences. In particular, the Protein Structure Initiative (PSI) from the National Institutes of Health (NIH) in the USA has expanded the impact of the Human Genome Project [1] by large-scale structure determination [2]. From the start, the project included computational biology to optimize target selection. One particular optimization criterion was to experimentally determine high-resolution structures for those protein families for which structures were neither known experimentally nor could be modeled through comparative modeling [3]. An additional objective was the prioritization of those families according to the highest leverage of each experimental nugget [4-6]. Put simply: determine structures for the largest families for which no structural knowledge is available. Despite substantial funding and impressive successes in advancing the streamlining of determining high-resolution structures [7], most families that were experimentally pursued did not yield high-resolution structures [6].

'Dark' proteins have been defined as proteins for which no 3D structural information is available today, neither from experiments nor from models [8]. Many proteins have dark regions that are long enough to independently fold as domains for which no structure is available. The dark proteome of structural biology is the union of all dark proteins and all proteins with dark regions. It has been shown [8] that what makes the dark proteome dark is largely not explained by regions notoriously difficult for structural biology, such as transmembrane and coiled-coil regions, regions of low complexity, and disordered regions such as found in intrinsically disordered proteins (IDPs). Instead,

the fractions of those regions are similar for proteins with and without 3D annotations in SWISS-PROT [9].

In this work, we focused on characterizing how entirely 'dark' proteins differ from all other 'non-dark' proteins, i.e., proteins containing at least some regions that are not 'dark'. It has been shown that dark proteins can be important for function [8]. However, dark proteins also tend to have many fewer known sequence relatives than non-dark proteins [8]. This implies that dark proteins are likely to have fewer paralogs than non-dark proteins. The reach of homology-based or comparative modeling increases with family size and diversity, i.e. the larger the family, the more likely it is that a structurally related protein in the PDB can be identified. In addition, the PDB is heavily over-represented in large families [4]. Together, both of these trends partly contribute to the greatly reduced family size of dark proteins compared with that of non-dark proteins. Unfortunately, we do not have the resources to gauge how big the difference would have to be to rule out that these two biases contribute significantly to the observation. Hence we can base our assumption only on today's observation from the comparatively unbiased SWISS-PROT: dark protein families were much smaller than non-dark families (median 27 for dark vs 397 for non-dark).

Assume we compared expression levels between two sets of proteins randomly drawn in the following way: set H contains 1,000 proteins with many paralogs, and set L contains 1,000 proteins with few paralogs. We expect expression levels to be higher for H than for L, simply because more proteins in H are bi-functional in the sense that they maintain the original function from which they diverged through duplication [10]. Therefore, our hypothesis was that dark proteins are expressed at much lower levels (referred to as "less expressed", for simplicity) than non-dark proteins. Given that some dark proteins have important functions, we also expected some outliers from this general trend. In this work we compared data on mRNA and protein expression levels; we found that dark and non-dark proteins have very similar expression levels, hence we could clearly reject our hypothesis.

Methods

Dark human proteome. We identified all human proteins for which 3D structures were neither available from experiments nor from models as described previously [Dataset S1 from 8]. The total set comprised 20,209 human proteins of which 4,403 (20%) were considered as dark proteins. The corresponding number would be much higher if partly dark proteins where also included. In this work, we focused on studying the properties of fully dark proteins versus all other proteins (i.e., versus all proteins containing non-dark regions).

Expression and mass-spectrometry data. We used mRNA expression datasets all obtained from public resources. All expression values for the E-MTAB-2919 experiment [11] were downloaded from *ArrayExpress* [12]. The experiments come from different laboratories and combine multiple assays to build a joint dataset on expression of human genes in 53 different tissues. 18,371 genes were mapped from ENSEMBL to UniProt identifiers. The GTEx project provides expression values for

18,256 of the 20,203 human proteins (82%) studied in this paper. We found expression data for 3,571 dark (20% of 18,256) and 14,685 non-dark proteins (80% of 18,256).

The mass-spectrometry data used were obtained from ProteomicsDB [13]. ProteomicsDB contains protein expression measurements from 16,857 liquid chromatography tandem-mass-spectrometry (LC- MS/MS) experiments involving human tissues, cell lines and body fluids. Normalized intensity values for each protein were retrieved from ProteomicsDB using the *proteinexpression* API. Empty values were counted as 0. We found non-zero mass-spectrometry data for 3,012 dark and 13,044 non-dark proteins. For each protein, we counted the number of tissues in which the respective protein was found to be expressed at any detectable level.

Normalization and analysis of data: Proteins might be undetected in either mRNA or proteomics experiments due to many different reasons, some of which are unrelated to the actual expression properties of the protein. Therefore, we ignored proteins without any experimentally detected levels for all subsequent analyses. For the mRNA data we used the Expression Atlas resource (https://www.ebi.ac.uk/gxa/home) with the recommended cutoff of 0.5 for the normalized expression level (also known as the 'FPKM' score) to filter for significant expression. Since the ratio of proteins with non-zero expression values differed between dark and non-dark proteins, the differences in coverage between both sets had to be used to normalize the raw data. For mass spectrometry data, we used the normalized intensity values that measure the relative abundance of peptides belonging to each protein in a specific sample on a logarithmic scale, compared to all peptides in the same sample. The raw values for the GTEx data differed by several orders of magnitude. Therefore, these data were analyzed and plotted on a logarithmic scale (note the proteomics data was already provided on such a scale). For calculating the overlap between distributions we used the 'overlapping' package in R [14]. For assessing the significance of differences between distributions, we used a one-sided Kolmogorov–Smirnov test [15].

Results

Detectable expression for most dark proteins. Significant mRNA expression in at least one tissue (*ArrayExpress*, Methods) was observed for 14,158 non-dark proteins (90%) and for 3,455 dark proteins (78%). Thus, while the fraction of non-dark proteins with significant expression was higher than that for the dark proteins, the majority of dark proteins were expressed in at least one tissue.

The proteomics data showed a similar trend: Out of the 20,203 human proteins queried in the *proteomicsDB* (4,403 dark, 15,806 non-dark), expression values were available for 3,012 of the dark (68%) and 13,044 of the non-dark (83%) proteins. While the fraction of dark proteins for which no expression had been measured in the proteomics data was higher than for the non-dark proteins, proteomics found the majority of dark proteins to be expressed in at least one tissue. Thus, both experimental methods, mRNA expression and proteomics, confirmed that most dark proteins are expressed, suggesting most dark proteins are important for cellular function.

Dark and non-dark proteins similar in mRNA expression. As expected, given the relatively large sample sizes, the distributions of mRNA expression for dark and non-dark proteins (Fig. 1A) had statistically significant differences (signed Kolmogorov-Smirnov test, *p*-value of 3×10^{-5}). Notable differences were seen in the tails of the distribution at high expression levels, where non-dark proteins dominate. Overall, however, the two distributions are very similar, with a 79% overlap (indicated in gray), and with only a slight difference in median mRNA expression levels (average normalized value of 3.4 versus 3.8 for dark and non-dark proteins, respectively). This strong similarity in overall mRNA expression contradicted the hypothesis that dark proteins are expressed substantially less than the non-dark proteins. When considering the level of mRNA expression as a sign of the functional importance, we picture dark proteins to be similar in importance to non-dark proteins.

>>> Fig. 1 <<<

Dark similar to non-dark in protein intensity. A similar overall pattern was seen when considering protein expression levels (Fig. 1B): the distributions for dark and non-dark proteins were significantly different (signed Kolmogorov-Smirnov test, *p*-value below 2.2×10^{-16}), with notable differences at high expression levels (dominated by non-dark proteins) and at very low levels (dominated by dark proteins). Although disorder, composition-bias, and membrane proteins by no means explain darkness, the set of all dark proteins contains slightly more of those types than the set of non-dark proteins [8]. Due to their biophysical features some of those proteins might simply escape experimental detection. If so: the substantial difference between dark and non-dark for the lowest intensity level (essentially 'not detected'), might be explained more by experimental challenges than by differences in the actual levels in the cell. However, as with mRNA, the distributions for protein expression were fairly similar overall, with 85% overlap (shown in gray), and with fairly similar median values (average normalized expression of 0.38 versus 0.68 for dark and non-dark proteins, respectively). These data also refuted our original hypothesis: dark proteins have remained structurally uncharacterized although they are as important for function as non-dark proteins that have been explicitly selected due to their high importance (since experimental structure determination is so expensive, structural biologists have done an outstandingly successful job in prioritizing proteins of extraordinary relevance).

Dark similar to non-dark in ubiquity of tissue expression. Next, we considered the number of different tissues in which each protein was expressed (Fig. 1C). Again, we saw a similar pattern; the distributions for dark and non-dark proteins were significantly different (signed Kolmogorov-Smirnov test, *p*-value below 2.2×10^{-16}), with notable differences at the tails, but with overall similarity. The distributions had a 85% overlap (gray regions), a fairly similar median number of tissues with measurable expression (6 versus 11 tissues for dark and non-dark proteins, respectively).

Dark proteins strongly expressed in testis. Our initial hypothesis was refuted most dramatically for testis, were dark proteins had greater overall mRNA expression than non-dark proteins (Fig. 1D). Here, once again, the distributions for dark and non-dark proteins were significantly different (signed

Kolmogorov-Smirnov test, *p*-value below 2.2 x 10^{-16}), with notable differences at the tails: for testis, high expression levels were dominated by dark proteins, and low expression levels by non-dark proteins. However, as before, the overall features of the distributions were quite similar, with 85% overlap (gray), and a similar median values (average normalized mRNA expression of 6 versus 4 for dark and non-dark proteins, respectively).

Conclusion: dark proteins constitute an important challenge. We clearly refuted our hypothesis: dark proteins have overall similar distribution of expression as non-dark proteins. Thus, the many proteins that have proven impenetrable to structure biology despite immense recent efforts constitute an important challenge to be addressed in the future. Simply put: there are important unchartered territories out there.

Acknowledgements

Thanks primarily to Tim Karl, but also to Guy Yachdav and Laszlo Kajan for invaluable help with hardware and software; to Inga Weise for support with many other aspects of this work. This work was supported by a grant from the Alexander von Humboldt foundation through the German Ministry for Research and Education (BMBF: Bundesministerium fuer Bildung und Forschung). Last, not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases.

References

[1] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A.

Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, J. Szustakowki, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, Nature 2001, 409, 860.

[2] M. Punta, J. Love, S. Handelman, J. F. Hunt, L. Shapiro, W. A. Hendrickson, B. Rost, Journal of Structural and Functional Genomics 2009, in press; T. B. Acton, K. C. Gunsalus, R. Xiao, L. C. Ma, J. Aramini, M. C. Baran, Y. W. Chiang, T. Climent, B. Cooper, N. G. Denissova, S. M. Douglas, J. K. Everett, C. K. Ho, D. Macapagal, P. K. Rajan, R. Shastry, L. Y. Shih, G. V. Swapna, M. Wilson, M. Wu, M. Gerstein, M. Inouye, J. F. Hunt, G. T. Montelione, Methods Enzymol 2005, 394, 210; L. Slabinski, L. Jaroszewski, A. P. Rodrigues, L. Rychlewski, I. A. Wilson, S. A. Lesley, A. Godzik, Protein Sci 2007, 16, 2472.

[3] J. Liu, H. Hegyi, T. B. Acton, G. T. Montelione, B. Rost, Proteins: Structure, Function, and Bioinformatics 2004, 56, 188.

[4] J. Liu, G. T. Montelione, B. Rost, Nature Biotechnology 2007, 25, 849.

[5] U. Pieper, A. Schlessinger, E. Kloppmann, G. A. Chang, J. J. Chou, M. E. Dumont, B. G. Fox, P. Fromme, W. A. Hendrickson, M. G. Malkowski, D. C. Rees, D. L. Stokes, M. H. Stowell, M. C. Wiener, B. Rost, R. M. Stroud, R. C. Stevens, A. Sali, Nat Struct Mol Biol 2013, 20, 135.

[6] B. H. Dessailly, R. Nair, L. Jaroszewski, J. E. Fajardo, A. Kouranov, D. Lee, A. Fiser, A. Godzik, B. Rost, C. Orengo, Structure 2009, 17, 869.

[7] S. K. Burley, A. Joachimiak, G. T. Montelione, I. A. Wilson, Structure 2008, 16, 5; J. Love, F. Mancia, L. Shapiro, M. Punta, B. Rost, M. Girvin, D. N. Wang, M. Zhou, J. F. Hunt, T. Szyperski, E. Gouaux, R. MacKinnon, A. McDermott, B. Honig, M. Inouye, G. Montelione, W. A. Hendrickson, J Struct Funct Genomics 2010, 11, 191.

[8] N. Perdigao, J. Heinrich, C. Stolte, K. S. Sabir, M. J. Buckley, B. Tabor, B. Signal, B. S. Gloss, C. J. Hammang, B. Rost, A. Schafferhans, S. I. O'Donoghue, Proceedings of the National Academy of Sciences of the United States of America 2015.

[9] E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, P. Bansal, A. J. Bridge, S. Poux, L. Bougueleret, I. Xenarios, Methods Mol Biol 2016, 1374, 23.

[10] S. Mika, B. Rost, PLoS Computational Biology 2006, 2, e79; N. L. Nehrt, W. T. Clark, P. Radivojac, M. W. Hahn, PLoS Comput Biol 2011, 7, e1002073.

[11] T. G. Consortium, Nature genetics 2013, 45, 580.

[12] N. Kolesnikov, E. Hastings, M. Keays, O. Melnichuk, Y. A. Tang, E. Williams, M. Dylag, N. Kurbatova, M. Brandizi, T. Burdett, K. Megy, E. Pilicheva, G. Rustici, A. Tikhonov, H. Parkinson, R. Petryszak, U. Sarkans, A. Brazma, Nucleic Acids Research 2014, 43, D1113.

[13] M. Wilhelm, J. Schlegl, H. Hahne, A. Moghaddas Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J.-H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber, B. Kuster, Nature 2014, 509, 582.

[14] M. Pastore, 2017; R Core Team, R Foundation for Statistical Computing, 2015.

[15] N. V. Smirnov, Bull. Math. Univ. Moscou 1939, 2, 3.

Figure captions

<Figure 1>

Fig. 1: Dark and non-dark proteins equally important for function. Dark (no knowledge of 3D structure) and non-dark (3D known by experiment or modeling) human proteins were selected as described previously [8]. The areas under the curves (normalized to 1) represent the complete respective sets of proteins: black for dark, white for non-dark, gray to show overlap between the curves. The dashed vertical lines indicate the medians of the distributions (white for non-dark; black for dark). Our initial hypothesis that non-dark proteins have considerably higher expression levels than dark proteins was clearly refuted because the gray areas essentially dominate all plots. The panels differ in their x-axes: (A) protein intensity from mass-spectrometry averaged across all tissues, (B) the number of tissues with clearly detectable protein intensity, (C) mRNA expression levels averaged across all tissues, and (D) differential mRNA expression levels for testis as the tissue for which dark were most above non-dark proteins as indicated by the median for dark being higher than that for non-dark.