Unique function words characterize genomic proteins

Andrea Scaiewicz[a,1] and Michael Levitt[a,1]

[a]Department of Structural Biology, Stanford University School of Medicine, Stanford, CA 94305

Contributed by Michael Levitt, May 21, 2018 (sent for review February 1, 2018; reviewed by Lewis Y. Geer and Christine Vogel)

Between 2009 and 2016 the number of protein sequences from known species increased 10-fold from 8 million to 85 million. About 80% of these sequences contain at least one region recognized by the conserved domain architecture retrieval tool (CDART) as a sequence motif. Motifs provide clues to biological function but CDART often matches the same region of a protein by two or more profiles. Such synonyms complicate estimates of functional complexity. We do full-linkage clustering of redundant profiles by finding maximum disjoint cliques: Each cluster is replaced by a single representative profile to give what we term a unique function word (UFW). From 2009 to 2016, the number of sequence profiles used by CDART increased by 80%; the number of UFWs increased more slowly by 30%, indicating that the number of UFWs may be saturating. The number of sequences matched by a single UFW (sequences with single domain architectures) increased as slowly as the number of different words, whereas the number of sequences matched by a combination of two or more UFWs in sequences with multiple domain architectures (MDAs) increased at the same rate as the total number of sequences. This combinatorial arrangement of a limited number of UFWs in MDAs accounts for the genomic diversity of protein sequences. Although eukaryotes and prokaryotes use very similar sets of "words" or UFWs (57% shared), the "sentences" (MDAs) are different (1.3% shared).

protein universe | genomic sequences | functional profiles | domain architecture | shared function

The size of protein sequence space as measured by the number of combinations of the 20 natural amino acids is essentially unlimited. Not every combination of amino acids can form a protein with a unique and stable fold and only a tiny part of full sequence space is occupied by protein sequences found in nature (1). Nevertheless, the number of protein sequences in the protein universe is still considerable (2, 3). Thanks to high-throughput sequencing (4), genomic sequences have been accumulating rapidly, leaving researchers struggling to comprehend the complexity of the almost 100 million nonredundant (NR) protein sequences known. Protein sequences, the building blocks of life, have been studied intensively. For some proteins, we know their structure, biological function, specificity, and reaction kinetics, but most of the vast protein sequence universe remains unexplored. Understanding how these proteins function and interact is indispensable for deciphering the language of life.

Given that functional and structural information on the protein sequence universe is so sparse, we need computational procedures to provide functional and/or structural information for unknown sequences. A powerful method for inferring function of a protein sequence is by relating it to similar proteins thereby allowing information on unknown family members to be deduced from known members. While such sequence clustering seems straightforward, it is complicated by the huge number of sequences and the difficulty of establishing a best way to match sequences. Consider the string of amino acids in a protein as analogous to the string of letters in an English sentence. For example, the two sentences "DOG IS BAD" and "BAD IS DOG" have very similar meanings and contain the same three words, but the sequences of letters themselves are different and cannot be made to match in more than 4 of 10 positions. This analogy suggests that the use of "words" is a better way to cluster protein sequences (5).

Words can be taken as the names of sequence profiles (6), which are derived from multiple sequence alignments to encapsulate information on amino acids occurring at each position along the sequence. Sequence profiles are often associated with a particular biological function giving them a meaning in analogy to words. The most common models for sequence profiles are hidden Markov models (HMMs) (7) and position-specific scoring matrices (PSSMs) (8). HMM-based profiles are used in PFAM (9) and SMART (10); PSSM-based profiles are used in CD (11) and PROSITE (12).

Sequence profiles from the different sources are used in two main databases: InterPro (13) and conserved domain architecture retrieval tool or CDART (14). Here we focus on the CDART resource (14) at the National Center for Biotechnology Information (NCBI), a PSSM-based database that includes profiles from seven different sources and matches sequences using RPS-BLAST (15), a variant of the widely used PSI-BLAST algorithm (16). Preliminary analysis of InterPro showed it less suitable than CDART for our purposes. In CDART, all profiles are matched by RPS-BLAST to all sequences using a consistent threshold, whereas in InterPro different thresholds as reported by each data source are used to match profiles to sequences. This lack of a consistent threshold makes it impossible to cluster profiles and so deduce a minimum set of unique function words as we do here. In both cases, the region of protein sequence matched by a profile is considered to be a domain.

In this study we have two classes of objects: sequences and profiles. The sequences and profiles are those matched every month or so to one another by CDART. This gives rise to domains, which are a length of a particular sequence matched to one or more profiles. The domains are named after the profile they match and this gives structural and functional information about the sequence. The patterns of domains along the polypeptide chain is called "domain architecture" and sequences can be characterized as having no domain architecture (dark matter), single (SDA), or multiple (MDA) domain architecture. The profiles are linked by strongly overlapping on the same region of sequence enabling them to be clustered by

Significance

The vast, mostly unknown protein universe can be explored by analyzing protein sequences as a string of domains. A broader coverage can be achieved when these domains, the essential blocks in protein evolution, are detected using sequence profiles. Using clustering to collapse redundant profiles into unique function words (UFWs), we find that over the years 2009–2016, the number of UFWs saturates while the number of sequences matched by a combination of two or more UFWs grows exponentially.

Author contributions: A.S. and M.L. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

Reviewers: L.Y.G., NIH; and C.V., New York University.

The authors declare no conflict of interest.

[1] To whom correspondence may be addressed. Email: michael.levitt@stanford.edu or scaiewicz@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1801182115/-/DCSupplemental.

Published online June 12, 2018.

clique-based full-linkage clustering. Each cluster can be represented by the name of just one profile in it, hence our unique functional words (UFWs).

In 2009 (17) we used CDART profiles to describe the organization of the protein universe in terms of the order of the domains in N- to C-terminal direction along the amino acid chain, a pattern that is termed "the domain architecture of a protein sequence." Our study analyzed sequences in the protein universe in terms of families that have SDAs or MDAs. It found that MDAs grew very fast with number of added sequences while SDAs seemed to have saturated. In addition, there were known structures for a quarter of the single domain families and another quarter of sequences were dark matter (no profile matches). Inspired by our work, Chub et al. (18) confirmed a decline in the rate of discovery of new protein families. Another study (19) explored the success of structural genomics efforts in providing structural coverage of the protein universe and concluded that over the last 10 y, structural coverage at a residue level increased from 30% to 40%, with half of novel structures coming from the Structural Genomics Initiative.

Here we study how CDART has changed over 8 y since 2009. With our improved clique-based clustering, all members of a cluster are similar to one another and each cluster is represented by one unique function word, avoiding confusion caused by synonyms. The number of UFWs increases more slowly than the number of profiles. As almost all UFWs occur alone in at least one sequence, the number of SDAs is close to the number of UFWs. The trends predicted in 2009 remain valid: number of different MDAs containing two or more UFWs increases with time exponentially fast, whereas the number of SDAs containing one UFW increases very slowly. We further find that eukaryotes and prokaryotes use the same "words" (the SDAs) but combine them into different "sentences" (the MDAs).

Results

Finding UFWs. We use profiles as an indicator of function, assigning to a sequence matching a profile the properties of the particular profile. We use the matches between profiles and protein sequences provided by the CDART database. Because this database includes profiles from seven different sources, a particular region of a protein sequence can be matched by several profiles. Such redundancy hinders our attempt to comprehend the complexity of all of the proteins in the known protein universe. The matches in the CDART database offer a way to eliminate the redundancy: if two different profiles match the same region of sequence, then these profiles overlap and are synonymous names for the same function.

Fig. 1 illustrates how we eliminate redundant profiles by clustering. We judge whether two profiles overlap using the lengths of the profiles, and the length of the match (only a part of the profile may match the sequence). CDART provides two parameters to assess the strength of the match, the E value (*Eval*) and the bit score (*Bscore*). We tested values of both *Eval* and *Bscore* combined with values for PDL_x, the maximum percent difference in profile lengths, and $Frac_x$, the minimum fraction of profile length matched. We use $Eval_x = 0.001$, $PDL_x = 10\%$, and $Frac_x = 0.9$, but other values gave similar results.

When two profiles satisfy these conditions, they overlap and are considered to be linked, forming a network with possible cliques. In graph theory, a clique is a subset of vertices such that all vertices are connected to each other. Clique finding is a hard-to-solve nondeterministic polynomial time (NP)-complete problem (21). We clustered linked profiles by full-linkage clustering using a powerful and efficient algorithm to find disjoint maximum cliques of linked profiles (20).

In 2009 there were 27,038 profiles in CDART and 795 did not match any NR sequence; in 2016 there were 48,932 profiles in CDART and 426 had no match. Thus, the number of used profiles increased 85%, from 26,243 to 48,506. Increases in the number of UFWs were smaller going from 17,072 to 24,212, an increase of 42% (*SI Appendix*, Table S2). For our most recent CDART dataset (2016), clustering reduces the number of different profiles from 48,506 to 24,212, a reduction of 50%. Similar reductions are seen for all eight CDART releases analyzed here (*SI Appendix*, Fig. S1 and Table S2).

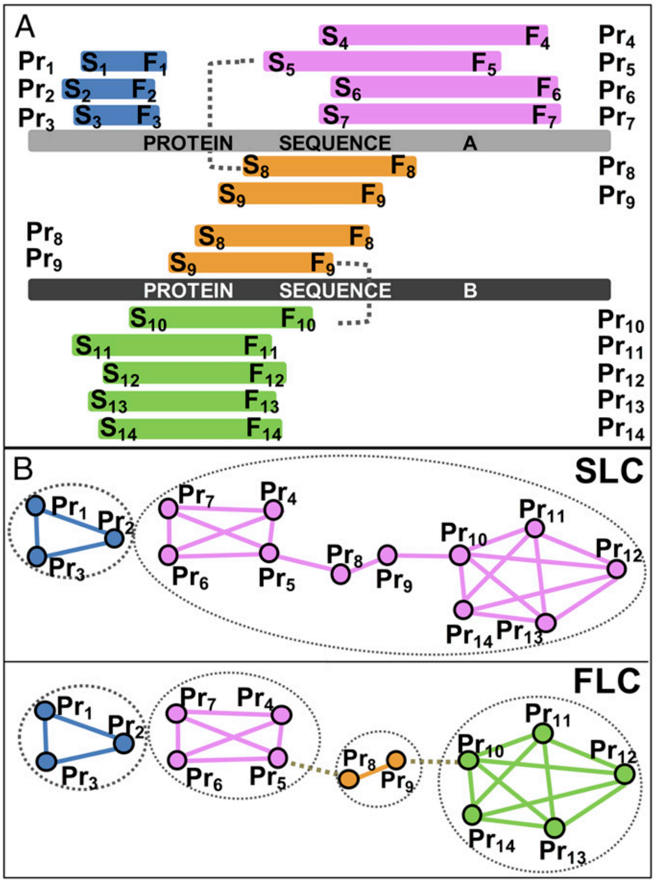
A
S_4 F_4 Pr_4
Pr_1 S_1 F_1 S_5 F_5 Pr_5
Pr_2 S_2 F_2 S_6 F_6 Pr_6
Pr_3 S_3 F_3 S_7 F_7 Pr_7
PROTEIN SEQUENCE A
S_8 F_8 Pr_8
S_9 F_9 Pr_9
Pr_8 S_8 F_8
Pr_9 S_9 F_9
PROTEIN SEQUENCE B
S_{10} F_{10} Pr_{10}
S_{11} F_{11} Pr_{11}
S_{12} F_{12} Pr_{12}
S_{13} F_{13} Pr_{13}
S_{14} F_{14} Pr_{14}
B
SLC
Pr_1 Pr_2 Pr_3 Pr_7 Pr_4 Pr_6 Pr_5 Pr_8 Pr_9 Pr_{10} Pr_{11} Pr_{12} Pr_{14} Pr_{13}
FLC
Pr_1 Pr_2 Pr_3 Pr_7 Pr_4 Pr_6 Pr_5 Pr_8 Pr_9 Pr_{10} Pr_{11} Pr_{12} Pr_{14} Pr_{13}

Fig. 1. Profile merging procedure. CDART profiles can be redundant in that different profiles may match the same region of a protein sequence. Such overlapping profiles need to be removed to avoid confusion and give a unique set of function words. (*A*) Nine profiles (Pr_1–Pr_3 in blue, Pr_4–Pr_7 in pink, and Pr_8–Pr_9 in orange) are matched to the protein sequence A (light gray). Seven profiles (Pr_8–Pr_9 in orange and also matching sequence A, Pr_{10}–Pr_{14} in green) are matched to the protein sequence B (dark gray). S_i and F_i denote the start and final residue number of the profile matched to the sequence. Two profiles are considered linked if: (*i*) both profiles are matched to a particular sequence with an E value better than $Eval_x$; (*ii*) the percent difference in profile lengths is less than *PDLx*; and (*iii*) the profile overlap ratio ($O_{i,j}$) exceeds $Frac_x$, where $O_{i,j} = Match\_Length_{i,j}/max\{Length_i, Length_j\}$, $Match\_Length_{i,j}$ the length of sequence matched by both Pr_i and Pr_j is ($F_i - S_j + 1$) and the longest possible match is $max\{Length_i, Length_j\}$, with $Length_i = F_i - S_i + 1$. (*B*) Linked profile pairs can be clustered in many ways. In single linkage clustering (SLC, *Top*), each cluster member is connected to other members by at least one link. In full linkage clustering (FLC, *Bottom*), each cluster is a clique with every cluster member directly linked to every other member; we use Roded Sharan's (20) method to find the disjoint maximum cliques for FLC. Profiles belonging to the same cluster are consequently represented by one unique function word, which is selected for name consistency for different years of analysis.

Combinatorial Growth of MDAs. The predicted massive combinatorial growth of MDAs and very slow growth of SDAs (17) is confirmed after 7 y and with 11 times more nonredundant sequence data. Fig. 2 shows that the number of MDAs is growing rapidly with time (or added sequence data) while the number of SDAs is almost constant and seems to have saturated. Between 2009 and 2016, the number of deposited sequences increased 11-fold (from

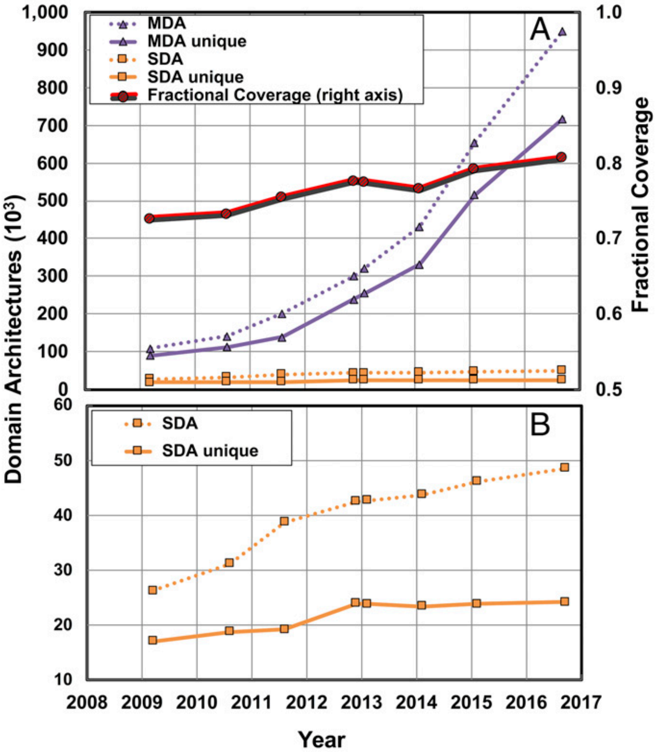

Fig. 2. Growth of SDAs and MDAs is very different. (*A*) As the number of sequences in the NR database grows, the number, *n*, of different multiple domain architecture (MDA) families (purple triangles) found by CDART increases exponentially with time in years. Specifically, $n = 108{,}000e^{0.3038(y-2009)}$ with $R^2 = 0.994$ for all profiles or $n = 77{,}000e^{0.2995(y-2009)}$ with $R^2 = 0.984$ for unique profiles). By contrast, the number of single domain architecture (SDA) families (orange squares) increases slowly and saturates with increasing time; *B* shows the same SDA plots on an expanded *y* scale. In *A*, fractional sequence coverage, the number of sequences in a SDA or MDA family divided by the total number of NR sequences, is shown on the right axis (red and black lines). It is high (over 70%) and increases with time as profiles are added to CDART. In 2009, 72.6% (5,947,106 of 7,877,467) of the sequences contained a domain recognized by a known sequence profile. By 2016, this percentage had increased to 80.1% (68,715,466 of 85,180,481), which is equivalent to a 28% drop in the amount of dark matter (unmatched sequences) from 27.5% to 19.9%. In all cases, dotted lines use all the original CDART profiles and solid lines use unique CDART sequence profiles.

7.8 million to 85.3 million), the number of MDAs increased 8-fold (from 88,905 to 717,727), but the number of unique SDAs increased by only 1.42-fold (from 17,072 to 24,212). While the number of MDAs grows exponentially with time, the number of SDAs remains almost unchanged (Fig. 2). Interestingly, among the 24,212 unique function words, 20,241 (83.6%) are shared between both types of families, whereas 3,625 (14.9%) are seen only in SDAs and very few UFWs (341 or 1.4%) are only seen in MDAs. The corresponding values for all CDART profiles are 41,269 (80.4%), 6,496 (16.3%), and 794 (1.4%) with a total of 48,506 profiles.

Comparing Kingdoms of Life. We analyzed the growth of SDAs and MDAs in the three main kingdoms of life (Fig. 3). In both prokaryotes and eukaryotes, the number of MDAs is much higher than the number of SDAs and these MDAs are growing fast, while SDAs are saturating. Eukaryote sequences are 2.7 times more likely to give rise to a new MDA than are prokaryote sequences. This is offset by the fact that prokaryote sequences are being determined at a much faster rate, especially since 2014 when prokaryote sequences are growing 2.5 times faster than eukaryote sequences. Overall, this means that the numbers of added MDAs in eukaryotes and prokaryotes are similar (8% higher in eukaryotes as 2.7/2.5 is 1.08).

Both prokaryotes and eukaryotes use about three-quarters of the UFWs: 74% (17,915 of 24,212) in prokaryotes and 68% (16,558 of 24,212) in eukaryotes. Less than half of the UFWs occur in both prokaryotes and eukaryotes (48%, 11,674 of 24,212). By contrast, eukaryotes use more of the MDAs than do prokaryotes: 61% (436,273 of 717,727) of the MDAs are seen in eukaryotes, only 40% (285,302 of 717,727) are seen in prokaryotes. Just 2% of MDAs (13,290 of 717,727) are shared between prokaryotes and eukaryotes, which supports the finding that domain combinations give rise to new functions (22–25). Another reason for more MDAs in eukaryotes could be to ensure that functional domains are coexpressed in different cell types of these multicellular organisms.

Unlike prokaryotes and eukaryotes, viruses do not have more MDAs than SDAs (Fig. 3*C*). There are 3,223 SDAs and 3,380 MDAs with UFWs and 4,445 and 3,739, respectively, with all profiles. Both SDAs and MDAs are growing at a comparable rate. Many viruses integrate their sequences into the genome of their host for replication (26) so it is not surprising viruses would not need many MDA architectures.

Dark Matter Is Shrinking. We define dark matter as those sequences that do not match any profile in CDART. In our previous work (17) we suggested that one of the reasons for the existence of dark matter sequences could be that labor-intensive discovery of new sequence profiles lagged behind the increase in the number of sequences and many sequence profiles remained to be discovered in the dark matter. Our present analysis of seven additional CDART releases reveals that with the discovery of new profiles, more of the older sequence is being annotated. This is confirmed by the drop of the overall fraction of dark matter (number of dark matter sequences/number of NR sequences) from 27.5% in 2009 to 19.9% in 2016 (Fig. 2*A*). While the number of sequences increased 11-fold from 7.8 million to 85.2 million in these 7 y, a significant decrease in dark matter was caused by the increase of UFWs from 17,072 to 24,212 (*SI Appendix*, Fig. S8). The fall in dark matter from 27.5% to 19.9% of all sequences means that over 6 million sequences are no longer classified as dark matter (7.6% of 85.2 million). Our previous analysis found that the dark matter contained equal numbers of prokaryote and eukaryote sequences (951,101 and 927,211, respectively), but there were more eukaryote residues. In 2016, we find there are more prokaryote than eukaryote dark matter sequences: of 16.9 million dark matter sequences, 11.4 million are prokaryotes and 4.7 millon are eukaryotes. This is in accord with the number of prokaryote and eukaryote sequences in the entire NR database (63.5 million prokaryotes and 20.4 million eukaryotes). Despite having more sequences—and more dark matter sequences—fewer of the prokaryote sequences do not match a recognized profile (17.2%) than for the eukaryote sequences (20.4%). The majority of the virus sequences (93%) match a known sequence profile; their dark matter fraction is just 7%. Two recent studies of half a million Swiss-Prot sequences report similar results, with slightly higher fractions of eukaryotic sequences in the dark proteome (27, 28).

Structural Coverage Is Increasing. In 2009 we made a wild extrapolation concerning possible scenarios for the structural coverage of sequence families in 2050. Today with the benefit of many more data we can confirm some of those predictions. The NR sequence database has continued growing exponentially, doubling its size every 25 mo. Back in 2009, we presented two scenarios for the structural coverage: (*i*) Assuming continuous investment but no improvement in either experimental or computational methods, we predicted that the coverage by 2050 would reach 70%. This seems to be on the right track (*SI Appendix*, Fig. S4, cyan line) with 26% of the families containing at least one known structure

BIOPHYSICS AND COMPUTATIONAL BIOLOGY

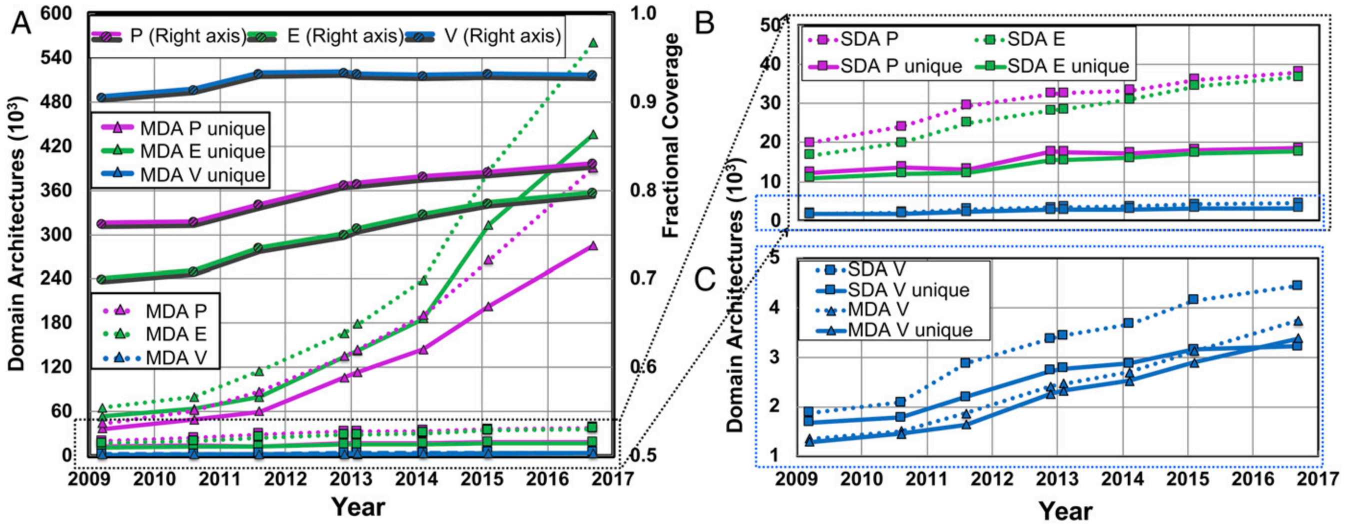

Fig. 3. Comparing growth of families in three kingdoms of life. The growth of SDA (squares) and MDA (triangles) families in prokaryotes (pink) and eukaryotes (green) shows the same trends seen for all species combined (Fig. 2). (*A*) MDAs are growing exponentially with time, whereas SDAs are saturating. In 2017, most of the MDAs are from eukaryotes (436,273 or 60.5% compared to 285,302 or 39.5% for prokaryotes). This is due to the faster growth of MDAs in eukaryotes during the last 2 y. The fractional coverage (defined as 1 − number dark matter sequences/number of sequences) is slightly higher in prokaryotes (82.8%) than in eukaryotes (79.5%). The majority of the virus sequences (92.9% or 1,041,468 of 1,126,166) have at least one match to a known profile so that the fractional coverage is very high at 93%, leaving just 7% dark matter. (*B*) The data for SDAs with an expanded *y* scale. About the same number of SDAs (18,514 or 76.4%) are seen in prokaryotes and in eukaryotes (17,287 or 71.4%). The corresponding values for all sequence profiles (dashed lines) are 560,577 or 59% and 390,640 or 41% for MDAs and 38,054 or 78.5% and 36,791 or 75.8% for SDAs. Saturation of SDAs in the three kingdoms is more evident when using unique profiles (solid lines) than original CDART profiles (dashed lines). Viruses (blue) have more SDA than MDAs, but this is difficult to see as both types of families occur much less often than in the other kingdoms of life. They are better seen in *C* with expanded *y* axis scale. In all panels, solid filled lines were obtained using unique CDART profiles, whereas dashed lines use all CDART profiles. E, eukaryotes; P, prokaryotes; and V, viruses.

in the Protein Data Bank (PDB) by July 2016. (*ii*) We also predicted that the structural coverage for structures not solved by the Protein Structural Genomics Initiative, referred to here as non-SG structures, would fall. We were unduly pessimistic. As shown in *SI Appendix*, Fig. S3, nonstructural genomics projects (no-SG) provided 65% of the 4,414 unique SDA structures solved during the last 6 y, whereas structural genomic (SG) projects provided just 35% of the unique SDA structures (*SI Appendix*, Fig. S3, yellow line). Although recent growth in unique coverage still benefits from structural genomics programs, the fraction of non-SG structures is rising faster than the SG structures (*SI Appendix*, Fig. S3, brown line). Lack of dependence of structural coverage on structural genomics programs is welcome as the NIH has phased out the program.

Common SDAs and Unique MDAs. While most SDAs are found in both prokaryotes and eukaryotes, the situation for MDAs is very different as clearly seen in the overlapping area of the Venn diagrams in Fig. 4. In 2016, a large fraction of SDAs (57% all profiles, 50% UFWs) are shared by eukaryotes and prokaryotes, while only a small percentage of MDAs (1.3% all profiles, 2.1% UFWs) are common to both kingdoms. This is not surprising since one would expect the more complex multidomain proteins to have more sophisticated functions and hence be more specific to a certain organism. As more sequences are identified, we speculate that the overlap of SDAs and MDAs between the three kingdoms of life will maintain the trends seen in Fig. 4 and *SI Appendix*, Fig. S5. Specifically, we expect to see almost all of the SDAs shared between prokaryotes and eukaryotes. Interestingly, viruses show the same behavior as prokaryotes and eukaryotes: the majority (77%) of their SDAs overlap with SDAs in either prokaryotes, eukaryotes, or both, but much fewer (23%) of their MDAs are seen in other kingdoms.

Databases of Sequence Profiles. To keep up with the rapid growth in the number of protein sequences, frequent updates of sequence profile databases are essential. We analyzed how the profiles and unique profile datasets have changed with time (*SI Appendix*, Tables S6 and S7). The four main sources of profiles in CDART are PFAM, PRK, CD, and COG. In 2009, most of the profiles in CDART came from PFAM (32.4%). This percentage has remained almost unchanged (33.31% in 2016). The second most common source of profiles in CDART has been PRK, and its percentage has decreased from 20.4% in 2009 to 15.8% in 2016. COG profiles represented 15.3% of CDART profiles in 2009 but dropped to 9.9% in 2016. The percentage of profiles coming from CD has increased from 12.5% in 2009 to 23.3% in 2016. We see similar trends when considering UFWs, but the contribution of PFAM is even more dominant. In 2009, 53.5% of the UFWs came from PFAM, and this percentage has remained almost unchanged at 54.1% in 2016. The changes for PRK, COG, and CD profiles are from 14.9% to 9.3%, from 18.4% to 10.1%, and from 8.4 to 10.2%, respectively.

Among all different profiles in CDART, PFAM and CD profiles seem to be updated more frequently than the others. Although the number of SMART profiles almost doubled since 2009, the SMART database seems to be updated less frequently (once a year) than the PFAM and CD databases. Note, the PSSM models we use are less well suited than HMMs to patterns containing insertions or deletions, of variable length, and with positional dependencies.

Our Clustering to Get UFWs Is Stable. In all eight independent analyses, clustering and the use of UFWs reduced the number of SDAs by approximately half but had a smaller (30%) effect on the number of MDAs probably because of the specific combinatorics (*SI Appendix*, Fig. S1). Puzzled by this difference, we simulated randomly sampled MDAs using the UFWs in our data. *SI Appendix*, Fig. S10 shows that simulated domain architecture plots are very similar to those of the real data and confirm that merging profiles into UFWs has a stronger effect on MDAs.

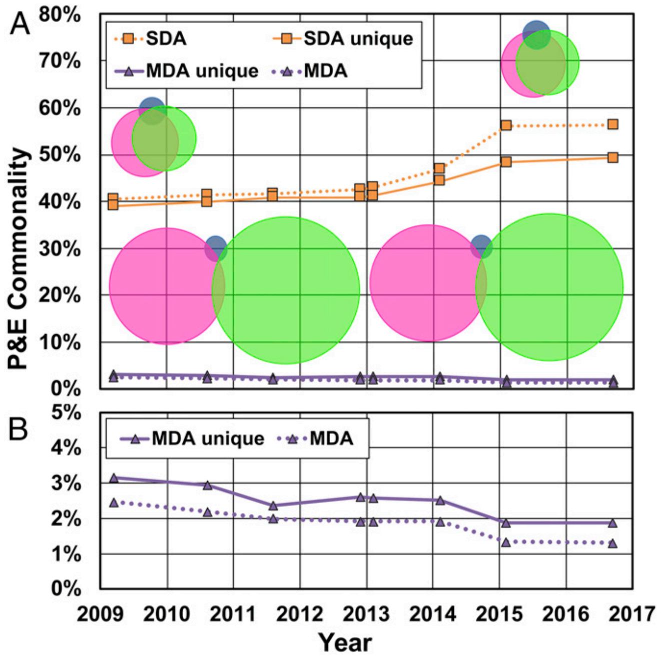

Fig. 4. Percent of SDAs used by both prokaryotes and eukaryotes increases with time, whereas percent of MDAs decreases. (*A*) The percentage of single domain architecture (SDA) and multiple domain architecture (MDA) families that are found in both eukaryotes and prokaryotes for seven different analyses. For SDAs, the percentage found in both prokaryotes and eukaryotes (P&E commonality) has increased from 40% (in 2009) to 57% (in 2016). Despite being fewer in number, SDAs in viruses behave similarly: the percentage of SDAs shared between viruses and the other two kingdoms has also increased from 56% (in 2009) to 79% (in 2016). For MDAs, the situation is very different in that there is little commonality of MDAs used in eukaryotes and prokaryotes. In 2009 only 3.1% of the MDAs were common to eukaryotes and prokaryotes and by 2016 this number has dropped to 1.3%. This suggests the commonality of MDAs is a better measure of evolutionary diversity than is the commonality of SDAs. Viruses share a much larger portion of MDAs with other kingdoms, with an increase from 18.8% in 2009 to 21.3% in 2016. Viruses also share more SDAs with other kingdoms, which is in accordance with the fact that many viruses integrate their host's genome into their genome. The Venn diagrams for 2009 (*Left*) and 2016 (*Right*) illustrate the number of SDA (*Top*) and MDA (*Bottom*) families for prokaryotes (pink), eukaryotes (green), and viruses (blue). Diagrams are scaled to keep the prokaryotes disk a constant size. The corresponding values using CDART unique profiles are 39–48% for SDAs (P&E), 54–74% (V and others), 4–2% for MDAs (P&E), and 20–24% (V and others). (*B*) MDA commonality decrease can be better appreciated (enlarged at *Bottom*).

Extrapolating the Sequence Growth. We attempted to answer the simple but important questions: What is the size of the protein universe? How many UFWs exist? How many unique sentences can be formed with these UFWs? We extrapolated the slowly growing number of SDAs as well as the exponentially growing number of sequences, all-length MDAs, and two-word MDAs over the next 50 y (Fig. 5). Assuming unchanged data trends, our data suggest that by 2066, the protein universe will contain 10^{15} nonredundant sequences. We estimate there will be between 30,000 and 50,000 SDAs, which combine to form 3×10^{12} MDAs (3×10^{11} of these composed by two SDAs).

Discussion

We characterized the protein universe—the collection of all proteins of every biological species that lives or has lived on earth—by classifying 85 million known sequences into families based on their having the same domain architectures. The two main types of domain architectures, SDAs and MDAs, are the main components of the language of life. Proteins speak using words (UFWs or SDAs) linked in an ordered way into sentences (MDAs). All 85 million NR sequences are characterized by just 24,212 UFWs and short sentences are made by combining these words. The number of UFWs in the language of proteins is saturating, and it looks like there are very few to be discovered. Structural coverage of the existing UFWs is high (27% of the UFWs have structures) and increasing rapidly (about 1% per year), thanks to equal contributions from both structural genomic projects and conventional crystallography. However, the number of sentences, which are mainly different combinations of words, is growing linearly with the number of sequences, but it too is expected to saturate. Although initial estimates of the protein universe size were much smaller (29, 30), fewer data were available.

These findings are consistent with the concept of protein evolution proceeding in large part by the creation of new proteins by new arrangements of existing protein domains to form novel multidomain proteins (25, 31). One can estimate the age of a domain family by finding the largest group of organisms within it (32). Different kingdoms of life share the same vocabulary, but use different sentences. Most words are used in both prokaryotes and eukaryotes, and this commonality increases with time, or number of sequences added. However, sentences made by combining words are much more organism specific and are rarely shared, with only 1.3% of the sentences used both in prokaryotes and eukaryotes. Sharing drops from 57% to 1.3% in going from SDAs to MDAs, which supports the finding that domain combinations give rise to new function (33–36). Remarkably, viruses share the majority (93%) of their words with either prokaryotes, eukaryotes, or both, but far fewer (7%) of their sentences are seen in other kingdoms. Viral proteins have a comparable number of single and multiple domain architectures. Many viruses integrate their genome into the host cell, which exempts them from carrying refined MDA machinery. On the other hand, for completing a successful

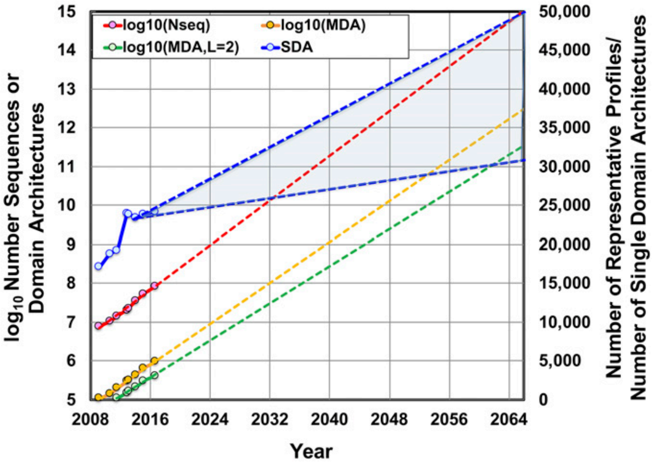

Fig. 5. Extrapolating growth over next 50 y. The number of sequences (NR), number of multiple domain architectures (MDAs), and number of two-word multiple domain architectures (MDA, L = 2) shown on the *Left* axis all grow exponentially with time as evidenced by the linear growth of their logarithms. The number of unique single domain architectures (SDAs) shown on the *Right* axis grows linearly and slowly with time. By the year 2066, we predict that there will be 10^{15} known sequences, 3×10^{12} different MDAs, 3×10^{11} different MDAs of length 2. For the number of unique SDAs, the expected number is between 30,000 and 50,000. More precisely, the linear plots of $\log_{10}$ of NR, MDA) and MDA, L = 2 against time have different slopes of 0.1442, 0.1319, and 0.1197, respectively. The fact that MDA and MDA, L = 2 grow more slowly that NR means that there is a finite but very large number of MDAs and a smaller finite number of MDAs, L = 2. In 2016, an average of 17 different MDA = 2 occur for each UFW, far fewer than the maximum possible value of 24,212 (the number of UFWs).

infection cycle, viruses must cope with the cell machinery for entry, replication, and translation while hiding from the host immune system so they still need some MDAs (37).

The dark matter of the protein universe are those sequences that do not have any match to a known sequence profile, or, following the language analogy, silent proteins. Although the absolute size of the dark matter in the protein universe has been growing with the number of deposited sequences in NR, its relative size is slowly decreasing. CDART and NR databases are well maintained and updated. Due to the admirable computational coverage of the protein sequence space, the unexplored regions of the protein sequence universe is slowly but consistently being reduced in its relative size. Although the protein universe has been increasing its size at a very fast pace, we still are orders of magnitude away from covering the entire feasible sequence space. Our insights about the protein universe are based on the sample of organisms that have been sequenced and we do not know how well this subset represents the protein universe. Still, the steady trends we have seen during a period when data increased 11-fold lend support to the views suggested here.

Conclusions

The functional language of proteins uses few words (fewer than 25,000 UFWs) and new words are added very slowly. Complexity of protein sequences comes for sentences made with these words. Different kingdoms of life use the same words but in different sentences. For UFWs, new sentences are being discovered rapidly with one per 196 added NR sequences in eukaryotes and one per 312 added NR sequences in prokaryotes. Corresponding values for all profiles are 156 and 227, respectively. Although our conclusions are based on the development of 7 y of protein sequencing data, we attempted to extrapolate these findings to millions of years of evolution across the entire protein repertoire; this could benefit from targeting sequencing efforts toward more diverse species.

Materials and Methods

We downloaded from ftp://ftp.ncbi.nih.gov/pub/mmdb/cdart/ the CDART data and from ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz the NR sequences. PDB entries solved by structural genomics were downloaded from targetdb.rcsb.org/target_files and taxonomy from ftp://ftp.ncbi.nih.gov/pub/taxonomy/gi_taxid_prot.dmp.gz. We did our analysis on: (*i*) February 17, 2009; (*ii*) June 16, 2010; (*iii*) June 15, 2011; (*iv*) September 24, 2012; (*v*) January 9, 2013; (*vi*) January 13, 2014; (*vii*) January 1, 2015; and (*viii*) July 3, 2016. Each analysis used a different version of NR, CDART, PDB, and taxonomy files, downloaded on that day.

ACKNOWLEDGMENTS. Computations were done on Stanford University's Bio-X2 computer cluster (National Science Foundation Award CNS-0619926). This work was supported by NIH Award R35GM122543 (to M.L., who is the Robert W. and Vivian K. Cahill Professor of Cancer Research).

1. Kolodny R, Pereyaslavets L, Samson AO, Levitt M (2013) On the universe of protein folds. *Annu Rev Biophys* 42:559–582.
2. Mora C, Tittensor DP, Adl S, Simpson AG, Worm B (2011) How many species are there on Earth and in the ocean? *PLoS Biol* 9:e1001127.
3. Godzik A (2011) Metagenomics and the protein universe. *Curr Opin Struct Biol* 21: 398–403.
4. Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* 94:441–448.
5. Scaiewicz A, Levitt M (2015) The language of the protein universe. *Curr Opin Genet Dev* 35:50–56.
6. Gribskov M, Lüthy R, Eisenberg D (1990) Profile analysis. *Methods Enzymol* 183: 146–159.
7. Eddy SR (2004) What is a hidden Markov model? *Nat Biotechnol* 22:1315–1316.
8. Gribskov M, McLachlan AD, Eisenberg D (1987) Profile analysis: Detection of distantly related proteins. *Proc Natl Acad Sci USA* 84:4355–4358.
9. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R (1998) Pfam: Multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* 26: 320–322.
10. Schultz J, Milpetz F, Bork P, Ponting CP (1998) SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc Natl Acad Sci USA* 95: 5857–5864.
11. Marchler-Bauer A, et al. (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res* 43:D222–D226.
12. Sigrist CJ, et al. (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res* 38:D161–D166.
13. Hunter S, et al. (2012) InterPro in 2011: New developments in the family and domain prediction database (vol 40, pg D306, 2011). *Nucleic Acids Res* 40:4725.
14. Geer LY, Domrachev M, Lipman DJ, Bryant SH (2002) CDART: Protein homology by domain architecture. *Genome Res* 12:1619–1623.
15. Marchler-Bauer A, et al. (2013) CDD: Conserved domains and protein three-dimensional structure. *Nucleic Acids Res* 41:D348–D352.
16. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
17. Levitt M (2009) Nature of the protein universe. *Proc Natl Acad Sci USA* 106: 11079–11084.
18. Chubb D, Jefferys BR, Sternberg MJ, Kelley LA (2010) Sequencing delivers diminishing returns for homology detection: Implications for mapping the protein universe. *Bioinformatics* 26:2664–2671.
19. Khafizov K, Madrid-Aliste C, Almo SC, Fiser A (2014) Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. *Proc Natl Acad Sci USA* 111:3733–3738.
20. Sharan R, et al. (2000) CLICK: A clustering algorithm with applications to gene expression analysis. *Proc Int Conf Intell Syst Mol Biol* 8:307–316.
21. Garey MRJ, David S (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness* (W. H. Freeman, Gordonsville, VA).
22. Zhang L, Gaut BS, Vision TJ (2001) Gene duplication and evolution. *Science* 293:1551.
23. Bornberg-Bauer E, Beaussart F, Kummerfeld SK, Teichmann SA, Weiner J, 3rd (2005) The evolution of domain arrangements in proteins and interaction networks. *Cell Mol Life Sci* 62:435–445.
24. Hurles M (2004) Gene duplication: The genomic trade in spare parts. *PLoS Biol* 2:E206.
25. Björklund AK, Ekman D, Light S, Frey-Skött J, Elofsson A (2005) Domain rearrangements in protein evolution. *J Mol Biol* 353:911–923.
26. Van Vliet K, et al. (2009) Poxvirus proteomics and virus-host protein interactions. *Microbiol Mol Biol Rev* 73:730–749.
27. Perdigão N, et al. (2015) Unexpected features of the dark proteome. *Proc Natl Acad Sci USA* 112:15898–15903.
28. Bitard-Feildel T, Callebaut I (2017) Exploring the dark foldable proteome by considering hydrophobic amino acids topology. *Sci Rep* 7:41425.
29. Chothia C (1992) Proteins. One thousand families for the molecular biologist. *Nature* 357:543–544.
30. Orengo CA, Jones DT, Thornton JM (1994) Protein superfamilies and domain superfolds. *Nature* 372:631–634.
31. Jacob F (1977) Evolution and tinkering. *Science* 196:1161–1166.
32. Weiner J, 3rd, Moore AD, Bornberg-Bauer E (2008) Just how versatile are domains? *BMC Evol Biol* 8:285.
33. Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* 307:1113–1143.
34. Koonin EV, Aravind L, Kondrashov AS (2000) The impact of comparative genomics on our understanding of evolution. *Cell* 101:573–576.
35. Vogel C, Teichmann SA, Pereira-Leal J (2005) The relationship between domain duplication and recombination. *J Mol Biol* 346:355–365.
36. Chothia C, Gough J, Vogel C, Teichmann SA (2003) Evolution of the protein repertoire. *Science* 300:1701–1703.
37. Horie M, et al. (2010) Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* 463:84–87.