BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 21 2010, pages 2664–2671
doi:10.1093/bioinformatics/btq527

Genome analysis Advance Access publication September 15, 2010

Sequencing delivers diminishing returns for homology detection: implications for mapping the protein universe

Daniel Chubb*, Benjamin R. Jefferys, Michael J. E. Sternberg and Lawrence A. Kelley

Department of Life Science, Imperial College London, London, UK

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Databases of sequenced genomes are widely used to characterize the structure, function and evolutionary relationships of proteins. The ability to discern such relationships is widely expected to grow as sequencing projects provide novel information, bridging gaps in our map of the protein universe.

Results: We have plotted our progress in protein sequencing over the last two decades and found that the rate of novel sequence discovery is in a sustained period of decline. Consequently, PSI-BLAST, the most widely used method to detect remote evolutionary relationships, which relies upon the accumulation of novel sequence data, is now showing a plateau in performance. We interpret this trend as signalling our approach to a representative map of the protein universe and discuss its implications.

Contact: daniel.chubb01@imperial.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 10, 2010; revised on August 31, 2010; accepted on September 10, 2010

1 INTRODUCTION

Only a tiny fraction of the vast space of all possible protein sequences is populated by proteins present in existing organisms. Knowledge of these populated islands can be considered as a map of the protein universe. As sequencing projects continue to provide new data, the resolution of the map increases, permitting insights into protein function and evolution. The map currently covers 1214 published genomes (Liolios *et al.*, 2008). In comparison, there are over 1.4 million known organisms (Leipe, 1996) while estimates of the total number of species on earth vary between 4 and 100 million (Crandall and Buhay, 2004). Despite the exponential growth in sequencing over the past two decades it would appear that a comprehensive map of the protein universe will not be achievable for many years. However, for many purposes the map need only contain representative sequences. A substantial proportion of the current map has been shown to be composed of very similar homologous sequences (Li and Godzik, 2006; Li *et al.*, 2002a; Park *et al.*, 2000; Suzek *et al.*, 2007) whose diversity can be captured by far fewer representative sequences. Acquiring a representative map of the protein universe will be possible in a far shorter time than achieving comprehensive coverage. Here, we investigate our progress towards such a representative map.

*To whom correspondence should be addressed.

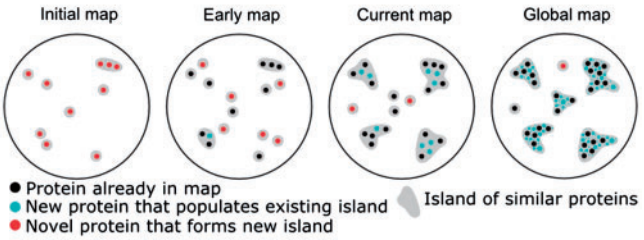

Fig. 1. Cartoon illustration of the progressive population of the protein sequence map. Early stages are characterized by the creation of new sequence islands. In contrast, later stages are characterized by the population of existing islands leading to an asymptotic approach to the complete protein map.

For the purpose of this work, we consider the protein universe to be populated by protein sequences clustered by sequence identity into islands of varying degrees of density. As more genomes are sequenced, the map of the protein universe can become more detailed either through the creation of new islands or the progressive population of existing islands. Early stages of sequencing are expected to be dominated by the discovery of novel sequence islands (Fig. 1). As a complete representative map is approached an increasing proportion of newly determined sequences are expected to fall within existing islands in sequence space and for the creation of new islands to become progressively less common. Often, the evolutionary relationship between proteins is distant: in the map representation this would be seen as points appearing far apart within an island or even in a separate island if the map has insufficient resolution. To detect these remote relationships between proteins, sophisticated homology detection tools are required. As novel sequences accumulate in the database, bridging gaps in the protein map, these tools are expected to be capable of detecting more of these remote relationships (Lobley *et al.*, 2009; Sandhya *et al.*, 2003).

Kunin *et al.* observed the growth in the number of islands while incrementally adding 83 genomes consisting of 311 256 protein sequences (Kunin *et al.*, 2003). Marsden *et al.* later used a larger dataset of 633 546 sequences from 203 genomes (Marsden *et al.*, 2006). Both investigations showed a linear increase in the number of islands with the addition of each genome. Since then there has been a nearly 6-fold increase in the number of genomes sequenced. More recent analysis (Levitt, 2009) using sequence profile matches to historical sequence databases, demonstrated that single domain sequences are now growing slowly and appear to be saturating in the sequence database. However, novelty in the form of the

re-arrangements of multi-domain architectures was shown to be growing linearly with added sequences.

In the current study, we recreated the sequence databases of the past two decades and through sequence clustering, estimated how the rate of novel sequence discovery has changed over time. Our approach considers different multi-domain architectures to lie in different islands and shows a continued increase in the number of novel sequence islands over time in agreement with Levitt (2009). However, by comparing growth in novel islands to the overall growth in sequencing we find that the *rate* of discovery is in a sustained period of decline, and predict that at least 80% of new protein sequences will fall within existing protein islands by approximately 2017.

The decline in the rate of novel sequence acquisition is expected to have an impact upon remote homology detection. To investigate this we employed PSI-BLAST (Altschul *et al.*, 1997), an iterative technique that uses the information in a sequence database to build statistical models, called profiles, of the mutational propensities of each position in a protein sequence of interest. In the first iteration, close homologues are gathered using the standard BLAST algorithm. The alignment of these homologues to a sequence of interest provides information on the amino acid substitutions observed at each position. This information permits the generation of a profile that can be used to search in the next iteration. This process can be continued, repeatedly refining the profile, until no further homologues are detected. This procedure is highly successful and detects more than twice as many homologous proteins with high confidence compared with BLAST, and has been the standard benchmark of remote homology detection against which new techniques are judged for the last decade. Its power stems from information in the sequence database in the form of sequence variation of homologues.

By recreating the sequence databases of the past it is possible to plot the performance of PSI-BLAST in remote homology recognition over time against a fixed test set of known homologues based on structure (Fig. 2). We show that the aforementioned decline in the rate of novel sequence discovery is reflected in a plateau in our ability to map remote homology using PSI-BLAST (Fig. 3). We suggest that this indicates an approach to a complete representative protein sequence map.

2 METHODS

Described here are the methods used to investigate the change in sequence novelty over the past two decades and its effect on homology detection. Sections 2.1 and 2.2 describe the recreation of past UniProt databases and the acquisition of metagenomic sequences that were used as the sources of sequence data. Section 2.3 describes the sequence-based clustering technique used to partition the data into sequence islands which were used to assess the change in novelty and to predict the future growth in islands. The remaining sections describe how the effect of increased sequence data on homology detection was investigated. Section 2.4 describes the design of the structure-based homology test set. The method used to create PSI-BLAST profiles for all sequences in the test set using the information in each sequence database is described in Section 2.5 and the use of these profiles to detect remote homologues within the test set is described in Section 2.6. In Section 2.7, we describe a variety of analyses used to assess the source of the observed trend. In order to investigate the effect that the order of sequencing had on the result, the PSI-BLAST analysis was repeated with random orders of sequence discovery. In addition, due to the possibility of a small number of superfamilies dominating the results an analysis was performed on a per-superfamily basis. Finally, the information content of the sequence profiles created each year was analyzed in terms of the number of sequences entering the profile and their diversity. Sub-section 2.8 describes the method for an analogous investigation to that described in Sections 2.5 and 2.6 using HHsearch, an HMM-HMM based alignment tool, in the place of PSI-BLAST. Finally, Section 2.9 describes a method of investigating the effect of sequence redundancy on the PSI-BLAST results.

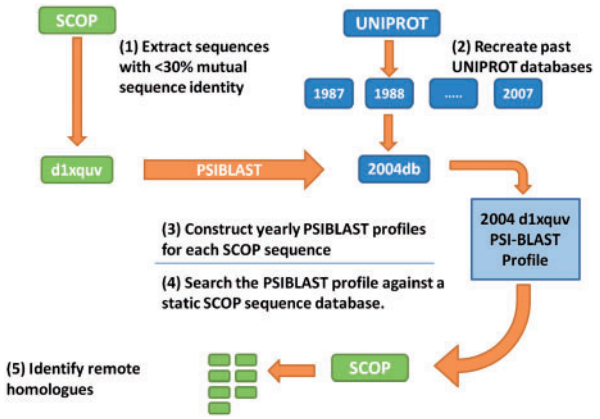

Fig. 2. Flowchart illustrating the benchmarking of PSI-BLAST performance in recognizing remote homologues relationships using profiles derived from sequence databases from 1987 to 2007.

2.1 Recreation of past databases

The UniProt (Wu *et al.*, 2006) databases from 1987 to 2007 were recreated using the deposition date field within the January 2008 UniProt_trembl.dat file and the UniProt_trembl.fasta sequence files (available from the UniProt FTP site). Any sequence found to be deposited in UniProt by December 31st in any given year was included within that year's database.

2.2 Metagenomic dataset

Metagenomic sequences were downloaded from the UniProt Metagenomic and Environmental Sequences database (UniMES) (The UniProt Consortium, 2009). UniMES contains data from the Global Ocean Sampling (GOS) expedition (Yooseph *et al.*, 2007). The downloaded file is non-redundant to a 100% threshold (i.e. contains no identical sequences) and contains approximately 6 million predicted sequences. A database was created which contained the full 2007 sequence data plus this metagenomic dataset. A 50% non-redundant version of this database was also created by clustering with CD-HIT (see below).

2.3 Sequence clustering using CD-HIT

CD-HIT (Li and Godzik, 2006) is a programme that clusters sequence databases according to a sequence identity threshold using a short word filtering heuristic. Representative sequences are selected from each cluster and are used to form a new sequence database. CD-HIT is a standard tool for creating representative databases and has been used by UniProt to create their UniRef (Suzek *et al.*, 2007) reduced redundancy databases.

CD-HIT uses greedy incremental clustering. First, a sequence database is sorted according to the sequence length and the longest sequence is chosen as the representative of the first cluster. Every other remaining sequence is then compared with the cluster representative and added to the cluster if the similarity is above a certain threshold (50% identity in this study). The next

longest remaining sequence is then selected as a representative of a new cluster and the process continues until all sequences are assigned a cluster.

CD-HIT was run on databases for every other year from 1987 to 2007 at a threshold of 50% sequence identity. This is effectively the lower limit of sequence identity achievable by CD-HIT, due to the high level of computer resources required (100 CPU weeks for this data, see Section 2.10 for hardware specifications) and the progressive drop in the efficiency of the heuristic used as the threshold is reduced (Li *et al.*, 2002b). In addition, a combined 2007 UniProt plus UniMES (metagenomic sequence) database was created and CD-HIT was run on this database of approximately 10 million sequences at the same threshold of 50% identity. For each processed database, a new database was produced, consisting of the representative sequence from each cluster, each sharing <50% sequence identity to any other sequence. The size of each of these representative databases provided our measure for the number of islands.

The ratio of islands to the total number of sequences in each database was calculated for every other year between 1987 and 2007. A power law curve was then fitted to this data and extended until 2050 (Fig. 4).

2.4 SCOP30 remote homology test set

The Structural Classification of Proteins (SCOP) (Murzin *et al.*, 1995) version 1.73 was used as a gold standard of homology based on structure. Homologous sequences sharing no >30% global sequence similarity (SCOP30) were downloaded from ASTRAL (Brenner *et al.*, 2000; http://astral.berkeley.edu). These sequences were placed into homologous groups according to superfamily membership defined within the SCOP domain classification. A PSI-BLAST searchable binary SCOP30 sequence database was created using the formatdb utility from the National Center for Biotechnology Information. This static sequence database is used for testing the power of profiles constructed for each year (Section 2.6 below).

2.5 Construction of database specific PSI-BLAST profiles

Each sequence within the SCOP30 test set was searched against each of the recreated UniProt databases from 1987 to 2007 using four iterations of PSI-BLAST with an inclusion threshold (-h parameter) of 0.001. At the end of the fourth iteration, a checkpoint file and PSSM was output. This resulted in a sequence profile for each SCOP30 sequence (6982 sequences) from each of the 21 recreated (1987–2007) databases (146 622 profiles in total) (Fig. 2). Thus, for each year it is possible to test the performance of PSI-BLAST on a fixed homology test set (SCOP30).

To subsequently test the robustness of our result, PSI-BLAST profiles were also output from 2, 3 and 5 iterations. The results of which are shown in Supplementary Figure S1. For the system with the consistently highest homologue detection rate (four iterations), another run was made with a more stringent inclusion threshold of 10^{-6} (Supplementary Fig. S2). The same trend was observed under all these conditions.

2.6 Identification of remote homologies within the SCOP30 test set

For each year there are 6982 PSI-BLAST profiles corresponding to each of the SCOP30 sequences. Each of these profiles was searched against the same static SCOP30 database (described in Section 2.5) (Fig. 2). This was accomplished by initiating a single iteration of PSI-BLAST, restarting using the checkpoint files previously created. For example, the SCOP30 sequence d1xquv will be searched against the full SCOP30 database 21 times, each search using a profile previously recreated from each recreated sequence database. In this way, the utility of each database in providing the sequence data required for remote homology detection can be assessed. Furthermore, by using the same SCOP30 database, instead of appending the test set to the recreated databases, the size of the scanned database remained constant and therefore did not affect the *E*-value score. The number of SCOP30 sequences belonging to the same superfamily as the query below an *E*-value threshold of 0.1 was recorded (see Section 3.3 and Fig. 3c). The proportion of predictions that were false positives at this threshold varied across the years between 3.3% and 6.9% with a mean of 5.1%. To ensure that the results were not dominated by a small number of larger superfamilies, the number of homologues detected after incrementally removing the largest superfamilies from the analysis (largest 5, 10, 20 and 30 superfamilies) was recorded (Supplementary Fig. S3). In addition, the percentage of all possible homologues found was calculated for all superfamilies and the results were averaged (Supplementary Fig. S4).

2.7 Profile analysis over time

It is possible that the performance of PSI-BLAST over time is a result of the order of discovery of sequences. For example, later databases might contain certain sequences that adversely affect homology detection by inducing drift in the searching by pollution of the profile with non-homologous sequences. Alternatively, there may be years when particularly novel organisms have been sequenced which could lead to sudden peaks in performance. To test this, the datasets were recreated with randomized orders of discovery, while fixing the number of sequences for a given year. The same trend was observed (Supplementary Fig. S5). It is also possible that increases in database size cause changes in the *E*-values returned by PSI-BLAST resulting in differing number of sequences entering profiles each year. The number of sequences entering profiles for each year was calculated and is shown in Supplementary Figure S6. Finally, the diversity of the resultant profiles for each year was calculated using the NEFF measure (Casbon and Saqi, 2004; Peng and Xu, 2010; Sadreyev and Grishin, 2004). The NEFF, or effective number of sequences in each alignment, is calculated as the exponential of negative entropy averaged over all columns of the alignment, resulting in a real value ranging from 1 to 20 (i.e. the number of amino acid types in nature). The larger NEFF is, the more diverse the sequence profiles. Results are shown in Supplementary Figure S7. NEFF values for each profile were provided by HHsearch after PSI-BLAST profiles were converted into HMMS (Section 2.8).

2.8 HHsearch

An analogous experiment to that conducted with PSI-BLAST (Section 2.7–2.9) was conducted using HHsearch (Söding, 2005), a highly sensitive sequence alignment method based on the pairwise comparison of profile Hidden Markov Models (HMMs). An HMM was created for each sequence in the SCOP30 test set using each UNIPROT database (1987–2007). These HMMs are analogous to the PSI-BLAST profiles constructed in Section 2.6. Two methods of HMM construction were used: (i) the default method supplied by the HHsearch package and (ii) a method that directly converts the PSI-BLAST profiles into HMMs. For the first method, the HMMs were built using the 'buildali.pl' script, downloaded as part of the HHsearch package from ftp://toolkit.lmb.uni-muenchen.de/HHsearch/. Buildali.pl constructs a multiple sequence alignment from a sequence database using a modified version of PSI-BLAST that prunes the high-scoring segment pair ends that do not achieve certain scoring criteria. This modification suppresses profile corruption coming from the ends of domains. The PSI-BLAST parameters used by the buildali.pl script in the creation of the HMMs were identical to those used in the PSI-BLAST runs (four iterations with an inclusion threshold of 0.001). For the second method, HMMs were generated by directly converting the PSI-BLAST profiles described in Section 2.6 into HMMs (using alignhits.pl and hhmake from the HHsearch package) ensuring that the same raw data was input both to PSI-BLAST and HHsearch. No secondary structure information was used in the construction of the HMMs. The HMMs generated for each SCOP30 sequence in each year were then concatenated into an HMM database. For each year, an all against all search of the HMMs was conducted and all hits within the same superfamily with a confidence (HHsearch returned probability) >95% were recorded.

The proportion of predictions that were false positives at this threshold varied between 0.2% and 4.6% over the years with a mean of 2.6%. NEFF (see Section 2.7) values were calculated for each profile during this stage. The comparison of results using both methods is shown in Supplementary Figure S8.

2.9 Homology detection using representative databases

To investigate the potential impact on remote homology detection of reducing sequence redundancy, the same PSIBLAST and HHsearch procedures described in Sections 2.7 and 2.9 were applied to the far smaller representative sequence databases formed from the island representatives that were created by CD-HIT (Supplementary Fig. S9).

2.10 Hardware

All analyses were performed on a heterogeneous computing cluster of approximately 250 AMD cores with an average of 2.2 GHz and 2 GB RAM per core. The entire analysis using the PSI-BLAST, CD-HIT and HHsearch distributions took approximately 10 CPU years with this hardware specification.

3 RESULTS

3.1 The rate of novel island formation is in decline

We define a sequence island as a set of proteins that share >50% global sequence identity to the largest sequence in the island. For each year, the database was clustered into such islands using the programme CD-HIT (Li and Godzik, 2006) (see Section 2.3). The size of these clustered databases is considerably reduced (by 60% on average) and represents an estimate of the number of sequence islands for each year. This measure of the number of sequence islands is a highly conservative upper estimate for two reasons: first, homologous protein domains may often share far lower—less than 25% (Pearson and Sierk, 2005)—sequence identity than the lowest threshold achievable by the clustering method. Secondly, the clustering technique works at the level of whole proteins and so does not take into account domain combinations. Briefly, this means that two multi-domain proteins that share one or more domains are placed in different clusters or islands if their domains are in a different order or there is at least one domain that they do not share. Thus, homologous domains will often fall into separate clusters. It has been previously shown that a large proportion of protein novelty is seen in the arrangement of domains in multi-domain proteins (Levitt, 2009). This novelty will therefore be seen in this analysis as new islands where the combination of domains is not already present in the database.

The recreated databases demonstrate exponential growth, expanding from approximately 5000 sequences in 1987 to approximately 4 million by the end of 2007 (Fig. 3a). In each database, the number of sequence islands steadily increases but is far smaller than the total number of sequences. This agrees with previous observations that sequence databases contain a substantial amount of redundancy in the form of closely related homologues with high sequence similarity (Li *et al.*, 2002a; Park *et al.*, 2000; Suzek *et al.*, 2007). While previous observations of redundant information have been focused on a single static database, the analysis presented here is broader. By comparing the size of the clustered and unclustered databases, it is possible to calculate how the rate of novel island discovery is changing over time (Fig. 3a and b). From 1987 to 1995, the ratio of the number of islands to the total number of sequences increases each year. After this point, however, the rate is in constant decline. In 1995, the number of islands was approximately half of the total number of sequences, and by 2007 this figure has fallen to just over a third.

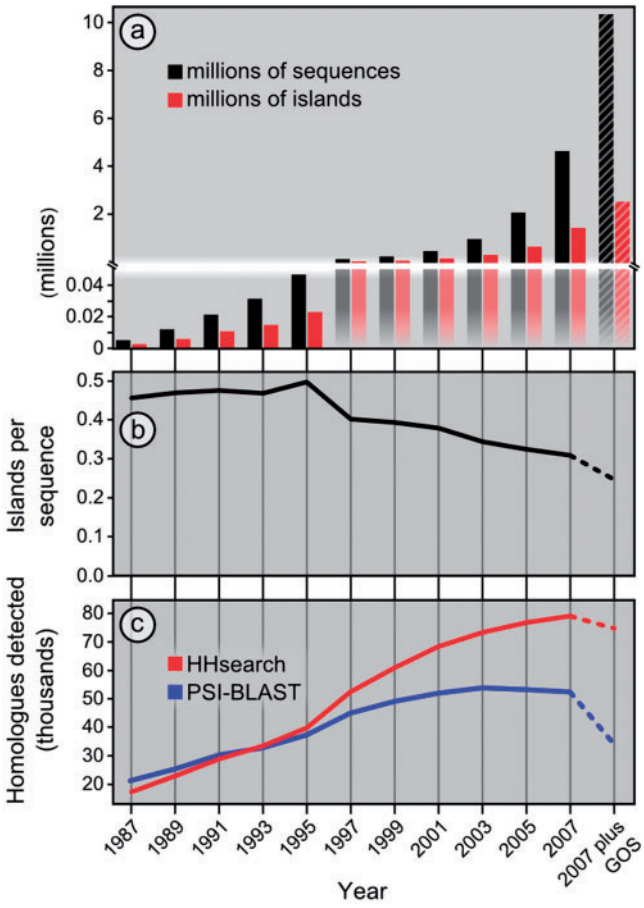

Fig. 3. Plots (**a**—**c**) show three different views of the change in sequence space over the last two decades. All three plots use the same horizontal axis. 2007 plus GOS indicates the combination of the 2007 sequence database with the Global Ocean Survey metagenomics data. (a) The protein sequence database (black bars) has grown exponentially over the past two decades. A much smaller increase is seen in the number of sequence islands (grey bars) at a level of 50% sequence identity. This is particularly the case with the metagenomic data (striped) which appears to have high redundancy. Note that the vertical axis uses a different scale for the top and bottom halves, in order to show growth both in the early sequence databases and the more recent ones. (b) The black line indicates the ratio of the number of islands to the total number of sequences, representing the rate of novel island discovery. Until 1995 this rate was steady or growing. Since that time this rate has been falling. There is a sharp drop on the addition of metagenomic data. (c) This plot shows the change in the ability of PSI-BLAST (black line) and HHsearch (grey line) to detect homology using profiles built from the databases of each year. It is clear that the more computationally intense HHsearch detects more homologues in the majority of cases. Although there is a general improvement in both methods over time, this improvement slows and in the case of PSI-BLAST, it plateaus and even begins to decline. The addition of metagenomic data adversely affects both methods.

An initial explanation for this trend is bias in sequencing projects towards highly similar organisms. To investigate this effect, the analysis was extended using metagenomic data derived

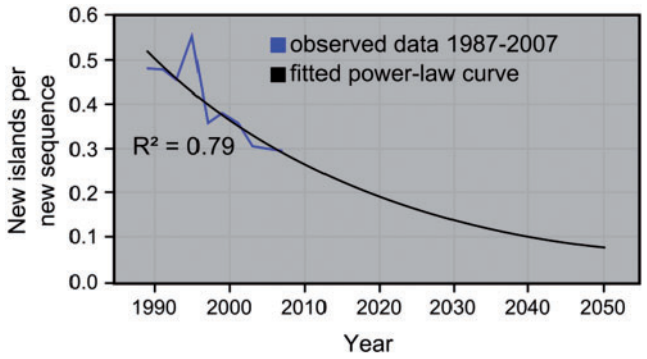

Fig. 4. A prediction of future island growth is made by fitting a power law curve to the number of new islands per new sequence in each year from 1987 to 2007. The GOS metagenomic data is not included in our projection. We predict that by approximately 2017, 80% of new sequences will fit within an existing island.

from environmental sequencing. Metagenomic projects sample a substantial diversity of extant sequences across habitats. In this work, the Global Ocean Survey (Yooseph *et al.*, 2007) data were merged with the 2007 UniProt sequence database (Sections 2.2 and 2.3) and the novel sequence island contribution was calculated (Fig. 3a and b, rightmost sample). An even steeper decline in the rate of novel island discovery than seen to date is observed. However, the implications of this result are ambiguous. The Global Ocean Survey has sampled environments across the world, albeit only within the oceans. In addition, there are likely to be biases in the locations sampled and the ability to sequence those samples. Finally, the metagenomic dataset is likely to contain a substantial number of artefactual sequences (Li *et al.*, 2008), which will artificially increase the total number of sequences and islands leading to an overestimate of the rate of novel island discovery. It is thus premature to suggest that further metagenomic data could not reverse the trend observed in this work.

3.2 Prediction of future island growth

These results show an increasing proportion of newly determined sequences falling within existing islands, which may indicate an approach to the representative map of the protein universe. If this trend continues, by approximately 2017 at least 80% of new sequences will fall within an existing island (Fig. 4) that is, have a sequence identity >50% with sequences already present in the database. This does not imply that the remaining 20% are entirely novel. This is a conservative estimate because the analysis does not cluster together sequences with <50% sequence identity that may be homologous, or those that are simple rearrangements of the same set of domains.

3.3 Improvements in remote homology detection are slowing or declining

This trend has important implications for the role of sequence-based homology detection in characterizing the structure, function and evolutionary relationships of a protein. Increases in the power of methods such as PSI-BLAST to detect remote homology currently rely on the steady discovery of novel sequence information. If, as indicated above, this rate of novel sequence discovery is declining, then one may expect this to be reflected in a slowing in improvements in the detection of remote homology. To investigate this, the performance of PSI-BLAST on the sequence databases of the past 20 years was determined.

We asked whether there has been an improvement over time in the performance of PSI-BLAST in identifying remote homologues in a set of 6982 proteins from the Structural Classification of Proteins (SCOP) (Murzin *et al.*, 1995) sharing less than 30% sequence identity (SCOP30). SCOP is a database of protein structural domains curated by experts that provides an authoritative classification of remote homologues on the basis of sequence and known structure, exploiting the fact that structure is more conserved than sequence in evolution. PSI-BLAST profiles were created for each of these sequences by scanning against each database from 1987 to 2007. Each sequence profile was then used to search the same static SCOP30 sequence database and the number of detected homologues was recorded (see Section 2.5–2.6, Fig. 2). In contrast to the CD-HIT clustering method, PSI-BLAST is able to routinely identify relationships with <25% sequence identity and as it operates on local alignments, and is thus capable of considering independent domains.

As expected, in the early phases of database growth a steady rise in remote homology detection is observed (Fig. 3c). However, this improvement does not directly scale with the massive increase in available sequences particularly evident in the last decade. Moreover, homology detection plateaus in 2004 and subsequently shows a slight decline. To determine the cause of this slowing and decline, a variety of analyses were undertaken.

3.4 Potential causes of the observed decline

To control for the possibility that the order of deposition of the proteins may affect homology detection, this order was randomized while holding the database size fixed for each year. This demonstrated the same trend (Supplementary Fig. S5). As sequence databases increase in size, so do the *E*-values for a given alignment with the same score. It is therefore possible that more recent, larger databases are performing more poorly as fewer hits are included in alignments, resulting in less information entering the profiles. To investigate this, the number of sequences entering the profile of each sequence for each year was calculated. Supplementary Figure S6 shows that the number of sequences being included in profiles is growing exponentially, mirroring the growth in sequence data.

It is also possible that, although the number of sequences entering the profiles is increasing, the diversity of sequences in the profile is decreasing, thus reducing the power of the profile to detect remote homologues. Using the NEFF measure of sequence diversity described in Section 2.7, we observe a continued increase in the effective number of sequences throughout the 20 years of database growth, indicating that sequence diversity is continuing to increase in the profiles (Supplementary Fig. S7). However, in comparison with the absolute number of sequences entering the profile, the diversity is slowing rapidly.

It is possible that the trend observed in this analysis may be caused by a small number of protein superfamilies that contain many members. Performing the same analysis while removing the largest 5, 10, 20 and 30 superfamilies, continues to show the same trend of a slowing in improvements of homology detection. However, it is important to note that the *decline* seen in PSI-BLAST performance is no longer apparent (Supplementary Fig. S3).

If a lack of sequence diversity is responsible for the observed trend, it may be expected that novel sources of sequence data from metagenomics would alleviate the problem. The same method was applied to the combined 2007 UniProt and UniMES metagenomic databases (Fig. 3c, rightmost sample) and substantially fewer homologues were detected from SCOP30. As discussed earlier, this is likely to be due to the large number of hypothetical sequences and sequence fragments within metagenomic datasets, and potential biases in sampling in sequencing, which have previously been shown to adversely affect the quality of PSI-BLAST profiles (Tress *et al.*, 2006).

Many of the most successful current approaches for remote homology detection are based on matching HMMs and such methods achieve substantially superior performance to PSI-BLAST. The same historical analysis was performed with a typical example of such a programme, HHsearch (Söding, 2005). While not plateauing, performance improvements over time are slowing, and this slowing is particularly evident when compared with the scale of sequence discovery (Fig. 3c). Almost identical results are seen for both the HMM construction methods described in Section 2.8 (Supplementary Fig. S8).

The combination of exponential database growth with a decline in the rate of novel sequence acquisition implies a concomitant increase in redundancy in the database. Past investigations have implicated sequence redundancy and sub-optimal sequence weighting as having a negative effect on homology detection using PSI-BLAST (Li *et al.*, 2002a; Park *et al.*, 2000). This affect is particularly evident in large diverse protein superfamilies where highly redundant families can trap sequence profiles, stopping them from detecting more remote superfamily relationships (Li *et al.*, 2002a). Thus, the trend reported here might be the result of algorithmic problems of PSI-BLAST and to a lesser extent HHsearch, in handling large amounts of redundancy. The entire analysis was repeated using databases from each year that have had redundancy reduced by CD-HIT to $<50\%$ sequence identity. These databases are typically 60% of the size of the full database from which they were derived. Using these databases, we observe a slight improvement in recognition from 1997 to 2007 and the decline in PSI-BLAST performance is no longer apparent (Supplementary Fig. S9). However, the same trend of a plateau in performance is observed. This indicates that although high levels of sequence redundancy are a factor in the performance of these techniques, it is not the primary cause of the overall observed trend. This is compatible with the results obtained when the largest superfamilies were removed from the analysis. These superfamilies are likely to contain the most amount of redundancy and removing them eliminates the reduction in performance while maintaining the plateau. Due to the previously discussed limitations of the clustering method, it was not possible to determine whether using representative database with mutual sequence identity $<50\%$ would improve homology detection further. However, in a previous investigation into representative databases and homology detection, databases with $<50\%$ sequence redundancy provided worse performance (Park *et al.*, 2000).

4 DISCUSSION

Our map of the protein universe is reliant on the sequences available from various sequencing projects. As novel data is acquired, the resolution of this map is expected to increase and eventually become a global map fully representing the space of all existing sequences. Obtaining a representative map will be achievable long before we have sequenced all of Earth's biodiversity because of the similarities inherent in proteins due to sharing a common evolutionary origin. These similarities mean that proteins will exist in islands, the distribution, shape and density of which reflect their relationship to each other. When a point is reached where any new sequence can be placed in a pre-existing island we will have obtained the representative map. In this study, we investigated the progress towards such a map over the last 20 years of sequence data acquisition.

By comparing the number of islands to the absolute number of added sequences over time we show that the rate of novel island discovery is in a sustained period of decline. An ever-increasing proportion of newly sequenced genomes are highly similar to proteins already in the database. Thus, the general structure of the sequence map is changing more slowly as it is dominated by those protein families common to a wide range of species. This is not to say that novel sequences are no longer being found. It is clear from Figure 3a that the number of islands continues to increase to the present day. However, the effect of these sequences on the global picture of sequence space is diminishing.

Methods for remote homology detection such as PSI-BLAST are able to routinely identify relationships with sequence identity $<25\%$. Being able to detect these remote relationships is vital to our ability to map the protein universe and predict the function and structure of proteins. The PSI-BLAST paper (Altschul *et al.*, 1997) has over 30 000 citations in the literature making it the most highly cited of the past decade. This reflects the crucial role played by remote homology detection for the accurate inference of the relationships between protein sequence, structure, function and evolution. It has been assumed (Lobley *et al.*, 2009; Sandhya *et al.*, 2003) that the continued growth of the sequence database will bring with it a steady improvement in our ability to detect remote homology with methods such as PSI-BLAST. Here, we have shown that the slowing in the gain of novel sequence data is associated with a plateau in our ability to detect remote homologues with PSI-BLAST.

The recent *decline* in PSI-BLAST performance (since 2004) is of particular importance. Here, we eliminated this decline by either the removal of the most populated superfamilies from the analysis or by a global reduction in sequence redundancy in the databases. Thus, the decline appears to be related to errors regarding the handling of redundant data that have been previously identified (Li *et al.*, 2002a; Park *et al.*, 2000). Given the expectation of continued exponential growth in the sequence database, problems caused by redundancy are expected to become more acute. In addition, this growth is substantially increasing the computational burden of sequence analysis faster than progress in computational power. The removal of sequence redundancy can alleviate both of these issues. However, this approach does not alter the overall trend of substantial slowing or a plateau in performance over recent years.

The utility of representative databases for remote homology detection will apply outside of the test set used here. For example, in the CASP8 assessment the Phyre structure prediction server (Kelley and Sternberg, 2009) used sequence profiles built using a 50% non-redundant sequence database similar to that described here and was placed in the top four servers (Cozzetto *et al.*, 2009).

One of the sources of information on which powerful methods of remote homology detection rely is *bridging* sequences that connect distant-related protein families. If, as it appears, we are approaching a representative picture of the global sequence map, then the rate of discovery of such bridging sequences will progressively decline while the vast majority of sequences will fall into pre-existing clusters. There are obvious parallels between these bridging sequences and the 'missing links' in paleontology. As with the missing links, many of these bridging sequences will be transitional forms, or present in extinct lineages that will never be sequenced. Although sequence space is continuous, its population by evolution is not.

That our current sampling of sequence space may already be approaching a representative map may not be wholly surprising. In terms of protein 3D structure, such a map appears to be near completion. Fold space, the space of distinct 3D protein topologies, appears to be populated by a relatively small number of protein folds, variously estimated at between 1000 and 10 000 (Wolf *et al.*, 2000). Although there is some debate regarding the discreteness or continuity of fold space, it is clear that the majority of protein structures fall within a limited range of fold islands (Sadreyev *et al.*, 2009). With the aid of structural genomics initiatives, the number of experimentally determined protein structures has been growing rapidly while the rate of novel fold discovery is slowing considerably (Levitt, 2009). This indicates our view of fold space is changing ever more slowly and that we are approaching a full representation of the protein structural repertoire (Zhang *et al.*, 2006).

Two primary factors have governed our progress to date in remote homology detection and the insights it generates into the relationships between protein sequence, structure, function and evolution: novel algorithm development and the growth in available sequence information. The evidence presented here suggests that the rate of accumulation of sequences that are sufficiently novel to enable detection of new remotely homologous relationships is diminishing. If transient bridging sequences are no longer present in extant species, it may now be necessary to focus attention on attempts to generate these missing links artificially. Some groups have created extra diversity in the sequence databases by creating artificial sequences using structural and functional constraints (Pei *et al.*, 2003) or using phylogenies as a guide to re-create ancestral sequences (Cai *et al.*, 2004). When these sequences are added to sequence profiles, an improvement is seen in remote homology detection.

It is important to recognize that the findings reported here could be modified by the sequencing of radically different organisms to those already analyzed. Unarguably, sequencing projects demonstrate some degree of bias in their choice of organism to sequence (Kyrpides, 2009) and this may equally apply to metagenomics projects. Although the number of unique protein domain sequences is vast, it is nonetheless finite. It is inevitable that at some point the discovery of truly novel sequences will become an extremely rare event and eventually all but cease. The analysis presented here suggests that sequencing is already beginning to approach a representative map of the protein universe.

ACKNOWLEDGEMENTS

Thanks to Drs Mark Wass and John Pinney for their advice on the manuscript.

Funding: Biotechnology and Biological Sciences Research Council (BBS/S/E/2006/13187 to D.C., BBG0039121 to B.R.J., BB/G022569/1 to L.A.K.).

Conflict of Interest: M.J.E.S. is a founder director of Equinox Pharma Ltd, holds shares in the company, and has obtained remuneration from the company. Equinox Pharma Ltd manufactures and markets bioinformatics software.

REFERENCES

Altschul,S. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Brenner,S.E. *et al.* (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.*, **28**, 254–256.

Cai,W. *et al.* (2004) Reconstruction of ancestral protein sequences and its applications. *BMC Evol. Biol.*, **4**, 33.

Casbon,J. and Saqi,M. (2004) Analysis of superfamily specific profile-profile recognition accuracy. *BMC Bioinformatics*, **5**, 200.

Cozzetto,D. *et al.* (2009) Evaluation of template-based models in CASP8 with standard measures. *Proteins Struct. Funct. Bioinformatics*, **77**, 18–28.

Crandall,K.A. and Buhay,J.E. (2004) EVOLUTION: genomic databases and the tree of life. *Science*, **306**, 1144–1145.

Kelley,L.A. and Sternberg,M.J.E. (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.*, **4**, 363–371.

Kunin,V. *et al.* (2003) Myriads of protein families, and still counting. *Genome Biol.*, **4**, 401.

Kyrpides,N.C. (2009) Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. *Nat. Biotechnol.*, **27**, 627–632.

Leipe,D.D. (1996) Biodiversity, genomes, and DNA sequence databases. *Curr. Opin. Genet. Dev.*, **6**, 686–691.

Levitt,M. (2009) Nature of the protein universe. *Proc. Natl Acad. Sci. USA*, **106**, 11079–11084.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Li,W. *et al.* (2002a) Sequence clustering strategies improve remote homology recognitions while reducing search times. *Protein Eng.*, **15**, 643–649.

Li,W. *et al.* (2002b) Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*, **18**, 77–82.

Li,W. *et al.* (2008) Probing metagenomics by rapid cluster analysis of very large datasets. *PLoS ONE*, **3**, e3375.

Liolios,K. *et al.* (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, **36**, D475–D479.

Lobley,A. *et al.* (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics*, **25**, 1761–1767.

Marsden,R.L. *et al.* (2006) Comprehensive genome analysis of 203 genomes provides structural genomics with new insights into protein family space. *Nucleic Acids Res.*, **34**, 1066–1080.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Park,J. *et al.* (2000) RSDB: representative protein sequence databases have high information content. *Bioinformatics*, **16**, 458–464.

Pearson,W.R. and Sierk,M.L. (2005) The limits of protein sequence comparison? *Curr. Opin. Struct. Biol.*, **15**, 254–260.

Pei,J. *et al.* (2003) Using protein design for homology detection and active site searches. *Proc. Natl Acad. Sci. USA*, **100**, 11361–11366.

Peng,J. and Xu,J. (2010) Low-homology protein threading. *Bioinformatics*, **26**, i294–i300.

Sadreyev,R.I. and Grishin,N.V. (2004) Quality of alignment comparison by COMPASS improves with inclusion of diverse confident homologs. *Bioinformatics*, **20**, 818–828.

Sadreyev,R.I. *et al.* (2009) Discrete-continuous duality of protein structure space. *Curr. Opin. Struct. Biol.*, **19**, 321–328.

Sandhya,S. *et al.* (2003) Effective detection of remote homologues by searching in sequence dataset of a protein domain fold. *FEBS Lett.*, **552**, 225–230.

Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.

Suzek,B.E. *et al.* (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**, 1282–1288.

The UniProt Consortium (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.

Tress,M.L. *et al.* (2006) An analysis of the Sargasso Sea resource and the consequences for database composition. *BMC Bioinformatics*, **7**, 213.

Wolf,Y.I. *et al.* (2000) Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.*, **299**, 897–905.

Wu,C.H. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.

Yooseph,S. *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.*, **5**, e16.

Zhang,Y. *et al.* (2006) On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl Acad. Sci. USA*, **103**, 2605–2610.