J Struct Funct Genomics (2009) 10:181–191
DOI 10.1007/s10969-008-9055-6

Structural genomics is the largest contributor of novel structural leverage

Rajesh Nair · Jinfeng Liu · Ta-Tsen Soong · Thomas B. Acton · John K. Everett · Andrei Kouranov · Andras Fiser · Adam Godzik · Lukasz Jaroszewski · Christine Orengo · Gaetano T. Montelione · Burkhard Rost

Received: 15 November 2008 / Accepted: 8 December 2008 / Published online: 5 February 2009

Abstract The Protein Structural Initiative (PSI) at the US National Institutes of Health (NIH) is funding four large-scale centers for structural genomics (SG). These centers systematically target many large families without structural coverage, as well as very large families with inadequate structural coverage. Here, we report a few simple metrics that demonstrate how successfully these efforts optimize structural coverage: while the PSI-2 (2005-now) contributed more than 8% of all structures deposited into the PDB, it contributed over 20% of all novel structures (i.e. structures for protein sequences with no structural representative in the PDB on the date of deposition). The structural coverage of the protein universe represented by today's UniProt (v12.8) has increased linearly from 1992 to 2008; structural genomics has contributed significantly to the maintenance of this growth rate. Success in increasing novel leverage (defined in Liu et al. in Nat Biotechnol 25:849–851, 2007) has resulted from systematic targeting of large families. PSI's per structure contribution to novel leverage was over 4-fold higher than that for non-PSI

R. Nair · J. Liu · T.-T. Soong · B. Rost (✉)
Department of Biochemistry and Molecular Biophysics,
Columbia University, 630 West 168th St., New York,
NY 10032, USA
e-mail: rost@rostlab.org
URL: http://www.rostlab.org/

R. Nair · J. Liu · T.-T. Soong · B. Rost
Northeast Structural Genomics Consortium (NESG) and
Columbia University Center for Computational Biology and
Bioinformatics (C2B2), Columbia University, 1130 St. Nicholas
Ave. Rm. 802, New York, NY 10032, USA

T.-T. Soong
Department of Biomedical Informatics, Columbia University,
630 West 168th St., New York, NY 10032, USA

T. B. Acton · J. K. Everett · G. T. Montelione
Center for Advanced Biotechnology, Department of Molecular
Biology and Biochemistry and Northeast Structural Genomics
Consortium (NESG), Rutgers University, 679 Hoes Lane,
Piscataway, NJ, USA

A. Kouranov
Protein Structure Initiative Knowledge Base & RCSB PDB,
Department of Chemistry and Chemical Biology, Rutgers
University, 610 Taylor Rd., Piscataway, NJ 08854-8087, USA

A. Fiser
New York SGX Research Center for Structural Genomics
(NYSGXRC), Department of Systems and Computational
Biology, Department of Biochemistry, Albert Einstein College
of Medicine, New York, NY, USA

A. Godzik · L. Jaroszewski
Joint Center for Structural Genomics (JCSG), Burnham Research
Institute, La Jolla, CA, USA

C. Orengo
Midwest Center of Structural Genomics (MCSG), Biosciences
Division, Argonne National Laboratory and Department of
Structural Biology, University College of London (UCL),
London WC1E 6BT, UK

G. T. Montelione
New York Consortium on Membrane Protein Structure
(NYCOMPS), Department of Biochemistry, Robert Wood
Johnson Medical School, University of Medicine and Dentistry
of New Jersey, Piscataway, NJ 08854, USA

B. Rost
New York Consortium on Membrane Protein Structure
(NYCOMPS), Columbia University, 1130 St. Nicholas Ave. Rm.
802, New York, NY 10032, USA

structural biology efforts during the past 8 years. If the success of the PSI continues, it may just take another ~15 years to cover most sequences in the current UniProt database.

Keywords Protein structure determination · Structural genomics · Evolution · Protein universe

Abbreviations

3D	Three-dimensional
3D structure	Here used exclusively to refer to the three-dimensional coordinates of each atom in the native conformation of a protein
HTP	High-throughput
JCSG	Joint Center for Structural Genomics
MCSG	Midwest Center for Structural Genomics
NESG	Northeast Structure Genomics Consortium
PDB	Protein Data Bank of experimentally determined 3D structures of proteins
PSI	Protein structure initiative at the NIH-NIGMS
NYSGXRC	New York Structural GenomiX Research Consortium
SG	Structural genomics
UniProt	Unification of SWISS-PROT, TrEMBL and PIR protein sequence database

Introduction

Systematic targeting of the largest families without structural coverage

The US contribution to Structural Genomics (SG), the Protein Structure Initiative (PSI), is funded by the National Institutes of Health-National Institute of General Medical Sciences (NIH-NIGMS). The second 5-year phase of the initiative, PSI-2, began in 2005. Four large-scale Structural Genomics Centers were created for high-throughput production of protein structures (JCSG, MCSG, NESG, NYSGXRC), as well as six Specialized Research Centers both charged with continuing to develop technologies needed for large-scale protein structure determination [30]. The four large-scale production centers are currently poised to generate over 3,000 entirely new experimental 3D structures of proteins for the biomedical research community in addition to the over 1,300 structures that originated from the pilot phase. At the end of the first 3 of those 5 years, the four centers had already deposited almost 2,000 new 3D structures (data from TargetDB, [9]).

Through the development and advancements of biochemical, robotic, NMR, crystallographic and computational techniques, SG centers are decreasing the cost and time required to determine a protein structure in order to advance the structural coverage of sequence space and biomedical research. The development and advancement of high-throughput protein production and protein structure determination pipelines are critical to the eventual characterization of protein structure space, expanding our understanding of molecular evolution, and to address biomedical problems such as drug discovery.

The challenges from these objectives for computational biology are mainly twofold: (1) identify targets for which each experimental structure will have a high leverage for modeling and (2) focus on those targets that will likely yield structures using current HTP methods [14, 21, 23, 37].

Metrics of success

Several metrics of success have been developed to monitor the evolution of structural genomics during PSI [8, 18, 22]. These include (i) total numbers of PDB depositions, (ii) numbers of distinct sequences (<98% pairwise sequence identity) for which an experimental structure is determined, (iii) numbers of 'novel structures', defined as a structure for a protein having <30% sequence identity with any protein structure already in the PDB, (iv) first 3D structure from a particular domain family; (v) first 3D structure from a particular functional class of proteins, (vi) protein structures which provide a novel testable hypothesis about function, and other metrics. In the following paragraphs we outline some of these metrics relating to the value of experimental 3D structures to provide useful structural information about homologous protein sequences.

Modeling leverage of experimental structures

Homologous proteins from different organisms, defined as those that have evolved from a common recent ancestral protein, usually share similar 3D structures [10, 28, 31, 35]. Therefore, the PSI does not aim at producing structures of every protein from every organism. Instead, the PSI aims to identify structural domains in proteins, systematically organize these protein domains into sequence-structure families, and determine the 3D structure of one or a few representatives from many of these families. The ultimate goal is to attain structural coverage for every major protein domain family found in nature.

Almost 50,000 experimentally determined 3D structures have been deposited into the PDB [4]. However, this accounts for less than 1% of the ~6 million protein sequences deposited into UniProt [2]. As genome sequencing technologies advance, sequence data is being generated at an ever increasing pace, not only for complete genomes of organisms but even for entire ecologies of hundreds or thousands of microorganisms (META genomics) [12, 36, 39]. Accordingly, the rate of discovery of new protein sequences will continue to increase much faster that the rate of protein structure determination.

The fact that homologous protein domains have similar structures enables the application of homology, or comparative, modeling methods [17, 32]. Comparative modeling leverages in the information provided by each experimental structure many fold. For example, it has been proposed that experimental determination of 3D structures for one representative of the largest 1,000–2,000 protein domain families, would be sufficient to allow modeling, at some approximate level, of more than half of all the residues in all of UniProt [21, 38].

The "modeling leverage" of a particular 3D structure (modeling template) depends on several factors, including (i) the sequence similarity between the template with known experimental structure and target proteins of unknown structure, (ii) the method of comparative modeling, and (iii) the criteria by which a model is judged to be "useful". The third factor (what is good enough?) can be especially difficult to ascertain, and rather inaccurate models (e.g. just the overall fold) are sufficient for some important applications of models, while other applications may require very high accuracy models. Benchmark studies suggest that sequence similarity of >40% over >50 residues generally provide models with heavy atom root-mean-square deviations of <2.5 Å from the true experimental structure [6, 11, 16, 24–27]. However, templates that are less sequence similar to the target structure may provide even higher accuracy models, and models generated for more sequence similar templates may result in less accurate models. Leverage also must be defined with respect to what portion of the target protein can be modeled from the experimental template, leading to metrics for full protein models, protein domain models, or residue models per experimental template. Modeling leverage also needs to be defined with respect to a particular sequence database; e.g. with respect to a particular version of UniProt.

Structural coverage

The concept of modeling leverage is intimately associated with the concept of *structural coverage*; i.e. the number or percentage of a particular set of protein sequences, domains, or residues, which can be modeled from a particular set of experimental protein templates. Structural coverage of the protein universe (i.e. a particular version of UniProt), of an entire proteome of an organism (e.g. the human proteome), of an ecology of organisms (e.g. all human gut microorganisms); or of a system of co-functioning proteins (e.g. proteins associated with a particular biological process), are all key metrics in measuring the success of SG that depend on the definition of *modeling leverage*.

Novel modeling leverage and novel coverage

Related to the concept of modeling leverage is the concept of *novel modeling leverage* [22], operationally defined as the number of proteins/domains/residues that could not be modeled (based on the above specific definition of leverage) as of the date the subject experimental structure was deposited into the public PDB [22]. The novel leverage provided by a set of experimental 3D structures across a particular set of protein sequences defines the novel coverage provided by these structures. This concept of leveraging experimental structures, and particularly novel leverage, has been fundamental to the process of target selection by large-scale centers during PSI-2. In particular, the large-scale centers systematically target the largest protein domain families for which we currently have little or no structural coverage.

The need for a standard convention

The modeling leverage value of a particular experimental structure, or the coverage of a set of sequences by a set of structures, depend upon the details of thresholds defined for sequence similarity that can be expected to provide a "useful" model, as outlined above. There are also certain technical issues which may or may not be accounted for in any method of assessing *novel leverage*. Examples of such issues, not used in the current work include: (i) while a sequence may be modeled from a structure already in the PDB on the date of deposition of subject structure, the subject structure may allow *more accurate* modeling of this sequence, and (ii) one may or may not discount the novel modeling leverage of a particular structure by the modeling leverage of experimentally-determined structures subsequently deposited in the PDB. It is simply not possible to define universal thresholds or criteria of model accuracy that are appropriate for the full range of applications for which models are used. Thus, the *novel leverage* reported for the same data by different groups may vary widely. Here, we adopt as a convention the definitions and thresholds proposed by Liu et al. [22] for assessing modeling leverage, novel modeling leverage, and the corresponding metrics of novel coverage. This is a convenient measure of "modelability" that is easily reproducible with relatively modest computing resources (the analysis presented here consumed less than 2 CPU-years).

Methods

Data set

All data about the status of structural genomics targets were taken from TargetDB [9]. *Leverage*, *novel leverage*, and the corresponding metrics for *coverage* were determined by the method of Liu et al. [22]. The basic concept is the following. We begin with a fixed version of UniProt, in this case release 12.8 from Feb 2008; containing 5,678,599 protein sequences with 1,851,231,082 residues. For this version we compile the number of proteins and residues that align (PSI-BLAST E-value 10^{-10}, 3 iterations on UniProt, one on PDB with background estimates based on UniProt size; for more details see Liu et al. [22]) to any protein of experimentally determined 3D structure deposited into the PDB at a given time point $T = T_0$. Novel leverage is then everything that is not covered by this simple alignment protocol and has arisen from structures added to the PDB at $T_1 > T_0$; total leverage is computed as all structures in the PDB covered by this criteria.

Novel structures

We loosely referred to an experimental structure (more precisely the structure specified by a particular PDB identifier) as a *novel structure* if at least 50 residues of this structure could be used to create novel leverage. This implies in particular, that novelty was not at all constrained by any particular definition in terms of the similarity of this new coordinate set in terms of structure to any other structure already in the PDB. When compiling per-residue estimates for novel leverage, we did not apply any such threshold, instead, any single residue that could not have been modeled before counted.

Novel leverage versus novel coverage

Leverage and coverage are related metrics that differ essentially only in the perspective they provide:

$$\begin{aligned}\text{Leverage} = {} & \text{Number of proteins/residues in database DB}\\ & \text{that can be modeled at threshold } E = E_0\\ & \text{based on a structure added at time } T = T_0.\end{aligned} \tag{1A}$$

$$\begin{aligned}\text{Coverage} = {} & \text{Percentage of proteins/residues in dataset DS}\\ & \text{that can be modeled at threshold } E = E_0\\ & \text{based on a structure added at time } T = T_0.\end{aligned} \tag{1B}$$

In the context of this work, we used the DB = UniProt 12.8 (Feb. 2008), E_0 = PSI-BLAST E-value $< 10^{-10}$. Coverage often is compiled with respect to the same database as leverage, i.e. DS = DB. In fact, this is the metric that we compiled for this work. However, we have also compiled coverage values for the set of proteins in particular organisms, e.g. focusing on the structural coverage for the human proteome [29]. In principle, *leverage* and *coverage* are symmetric: both can be compiled on the same data set, and the only essential difference is that one counts numbers, the other percentages.

Both leverage and coverage can be computed on a per-structure, on a per-residue or on per-annum base. Frequently, we also compiled those numbers as sums over all PSI structures in light of the sum over all PDB structures and/or over all PDB structures without those PSI-structures.

The measures for *leverage* and *coverage* as defined above have a severe problem: they do not distinguish at all between structures that provide new information and those that simply confirm the information we already have in the PDB. This effectively implies that the measures as defined above do not capture a scientifically relevant reality. This problem is easy to fix: all we need to do is to compile the leverage/coverage at a given time and to then define the novelty provided by new structures as the added leverage and coverage. We have introduced this simple metric as "novel leverage" and "novel coverage", and defined them by:

$$\begin{aligned}\text{Novel leverage} = {} & \text{Number of proteins/residues in}\\ & \text{database DB that could } \textit{first} \text{ be}\\ & \text{modeled at threshold } E = E_0 \text{ based}\\ & \text{on a structure added at time } T = T_0.\end{aligned} \tag{2A}$$

$$\begin{aligned}\text{Novel coverage} = {} & \text{Percentage of proteins/residues in}\\ & \text{database DB that could } \textit{first} \text{ be}\\ & \text{modeled at threshold } E = E_0 \text{ based}\\ & \text{on a structure added at time } T = T_0.\end{aligned} \tag{2B}$$

With the same choices as above: DB = UniProt 12.8, and E_0 = PSI-BLAST E-value $< 10^{-10}$. The deposition date in the PDB entry decides whether or not a structure is novel. One important and desired consequence of this definition is the following. Assume you solved a structure that has high impact in the sense that many groups use it as a basis for molecular replacement to do more accurate structures of the same or of a similar protein sequence. Then the first structure in this family of structures is recognized for the novel information it provided on the date it was deposited in the PDB. The problem that remains and that we have not addressed convincingly, yet, is how to measure the benefit of a structure that allows to build better models for proteins for which we can already build models.

As indicated by Eq. 2A, only sequences that match to the sequence of the template with the minimal threshold (E-value $< 10^{-10}$) count.

Results and discussion

Every other novel structure from the USA now from the PSI

A primary goal of the PSI has been the development of automation and robotics for large-scale protein structure determination. It took a few years to scale the pipelines up to reaching "high-throughput" levels; currently some 600–800 protein structures per year (i.e. two structures per day). Progress is evident: over 1,300 structures originated from the pilot phase PSI-1 (2000–2005), and after 3 of the 5 years of PSI-2, the large-scale centers have already deposited almost 2,000 new 3D structures. This success is also evident in the increased contribution from PSI to all experimental structures deposited into the PDB: over the course of PSI-2 (2005/07/01–2008/09/19), PSI centers have contributed almost 9% of all structures world-wide and all structural genomics (SG) centers have contributed almost 18% of all structures (data not shown). As the PSI is entirely financed by the NIGMS at the NIH in the USA, its contribution should be compared directly to structures deposited into the PDB from US-based laboratories: in the first 3.25 years of PSI-2 (labeled 2005–2009, where 2009 represents only the first quarter of Year 4), PSI-2 centers alone had contributed about 18% of all structures deposited by US structural biology groups (Fig. 1a, cumulative sum over gray bars). When comparing the annual contribution from the PSI of *novel* structures (i.e., $< 30\%$ sequence identity with any other structure in the PDB at the time of deposition) to that from all other sources (Fig. 1c), it is noticeable that the PSI fraction has continued to increase over its entire duration. Over the last several years, the US contribution to all structures increases, although the non-PSI fraction from the US shrinks. Without PSI, the US structures and even more significantly the US novel structures would be on decline: the US non-PSI reduction from 2001 to 2009 is more significant than the reduction of all other non-PSI contributors over the same time period (Fig. 1c).

Given that novel leverage is an important criterion in PSI-2 target selection, PSI-generated structures also possess more novel leverage than structures from non Structural Genomic groups (Figs. 1b, d, 3a). The concept of structural leverage has also been employed in target selection by non-US Structural Genomics efforts such as the RIKEN project. Despite competition to structurally characterize unique sequences, the PSI deposits now almost as many novel structures as all the other depositors in the US combined (Fig. 1a), and about 30% of all novel leverage contributed worldwide (Fig. 1d); the fraction of novel structures per structure solved is 2–5 times higher for PSI than for all other depositors (data not shown). In fact, since 2005, half of the novel leverage generating structures from the USA was determined by the PSI-2 centers (Fig. 1a, cumulative sum over blue bars). The contribution of PSI to the generation of novel leverage is equally impressive (Fig. 1b, discussed in more detail below). Worldwide SG contributed about 18% of the structures since 2005 and about 37% of all novel protein leverage (171/459 K, Table 1); more than three-quarters of the novel leverage since 2005 came from PSI-2 centers ($30\% = 135/459$ K, Table 1).

Structural coverage of sequence space continues to increase

We froze a version of the entire sequence space known in Feb. 2008 (UniProt 12.8) and then estimated to what extent the structures deposited into the PDB at a given time point could have been used to structurally cover this sequence universe. We compiled two separate values, one estimating the per-protein coverage that considers a new arrival to cover a new protein when at least 50 consecutive residues were aligned above the threshold (E-value $< 10^{-10}$, Methods; orange in Figs. 1b, 2a), the second a per-residue coverage, simply an estimate of which fraction of all residues can be structurally covered (purple in Figs. 1b, 2a). While the former per-protein measure is intuitive, it requires the definition of an ad hoc threshold (50 residues). This has to be done because most structures in the PDB contain single domains while over 75% of proteins in nature appear to contain multiple domains [19–21, 34]. If one domain of a protein can be modeled then this constitutes an important advance and ought to be considered.

The PSI contribution to the coverage added by US structures is now exceeding the 50% mark, i.e. PSI-2 contributes more novel leverage, and hence more coverage than all other US efforts (Fig. 1b). With this increase, the US contribution to the novel leverage worldwide continues to increase (Fig. 1b). Interestingly, the contribution of non-PSI SG, which peaked in 2004–2007, has contributed relatively little to the worldwide annual novel leverage, while novel leverage contributions from non-US, non-SG groups has been relatively constant at $\sim 40\%$ annually.

Overall, the structural coverage of UniProt 12.8 increased slowly, up until about 1992 (Fig. 2a). After that, structural coverage increased at roughly a constant annual rate. The growth slowed down slightly toward the onset of structural genomics, because, despite the continued annual increase in the number of structures determined, it is getting

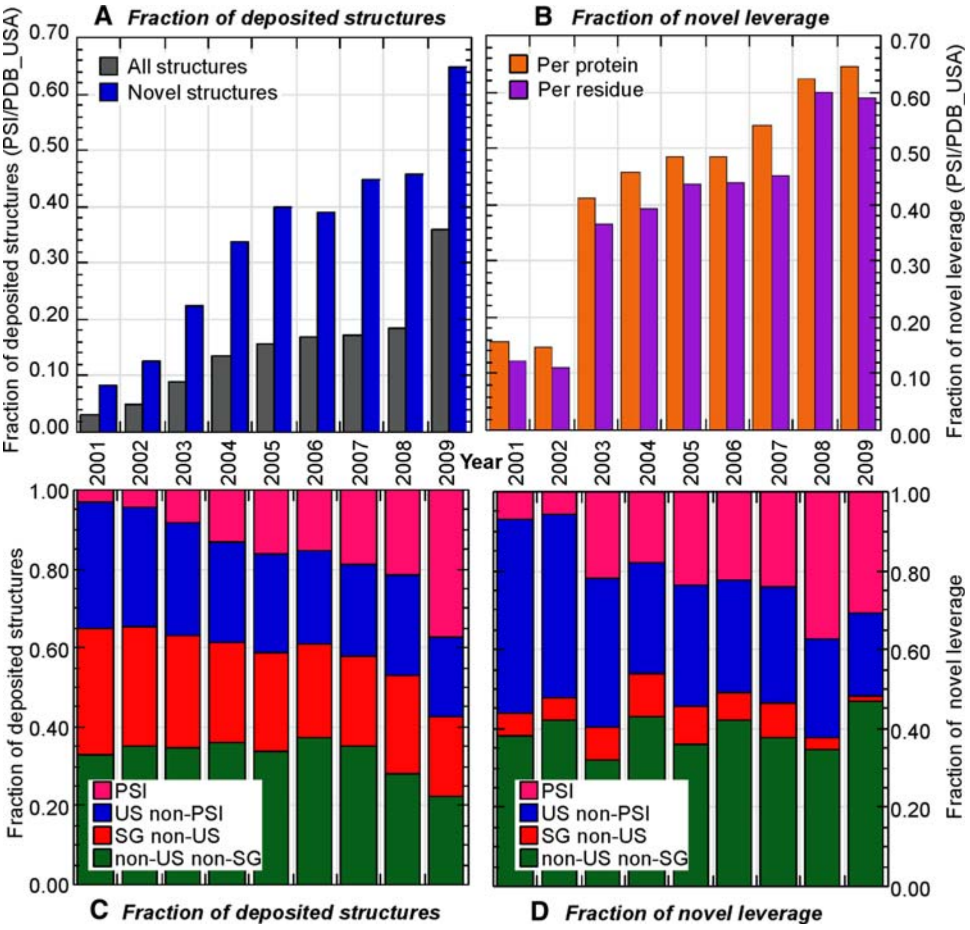

Fig. 1 PSI annual throughput as percentage of the worldwide PDB and the US-PDB. **a** Annual statistics for the fraction of structures determined by the PSI (Protein Structure Initiative at NIH's NIGMS) distinguishing between the contribution to all structures deposited in a given year (*gray bars*), and the contribution toward novel structures (*blue bars*). In this context, we considered any structure that yielded novel leverage for at least 50 consecutive residues as a "novel structure". The PSI contribution to novel leverage is 2–3 times higher than its contribution to all structures. 100% marks all structures determined by US-laboratories. **b** While panel (**a**) shows the fractions of structures, panel (**b**) shows the fraction of novel leverage added in each year (i.e. PSI novel leverage/US-PDB novel leverage), in terms of per-protein (*orange*) and per-residue (*purple*) values. Panels (**c**) and (**d**) distinguish between the contribution from the PSI, from the US without PSI, from structural genomics (SG) without the PSI and from all other depositors. In particular, we distinguish the contribution to all structures (**c**) and that to all novel leverage (**d**). Note that in all figures the years refer to PSI grant years, e.g. 2001 refers to the period of July 2000–June 2001. The last entry (labeled 2009) marks an incomplete year from July 2008–September 2008 corresponding to the first quarter of year 4 of PSI-2

increasingly difficult to succeed for proteins that have so far eluded structure determination. Novel leverage becomes an increasingly evasive objective. The advent of structural genomics countered this development and returned the growth in structural coverage to almost constant annual rates. During the course of PSI, the overall structural coverage for UniProt 12.8 has approximately doubled (Table 1) from ~22 to ~45% per-protein coverage (Fig. 2a).

If we reset the coverage clock to zero at the beginning of PSI, and compute the gain over the structural coverage in a given year (Fig. 2b), we note that between 2000 and 2008 the per-protein structural coverage of UniProt 12.8 increased by about 26 percentage points (Fig. 2b: sum over all contributions; Table 1: 1,485/5,679 K) corresponding, by 2008, to an overall per-protein coverage around 45%. Some 22% of the increase in per-residue and per-protein structural coverage provided by all structures deposited worldwide came essentially from four PSI large-scale centers, compared with ~34 and ~40% increase in the structural coverage of UniProt 12.8 by all non-PSI US and all non-SG, non-US groups, respectively, in the same time frame (Fig. 2b). Note that the precise values here depend crucially on the parameters chosen. Our restriction to E-values $\leq 10^{-10}$ implies that the inferred structural models are of relatively high reliability and cover most of the aligned regions [16, 25]; higher leverage and coverage can be achieved at the expense of accuracy [3, 17, 27, 33].

Table 1 Structural coverage of UniProt

	Number of proteins (in Kilo)	Number of residues (in Mega)	PSI novel leverage proteins (Kilo)	SG novel leverage proteins (Kilo)	PSI novel leverage residues (Mega)	SG novel leverage residues (Mega)
UniProt 12.8	5,679	1,851	–	–	–	–
3D coverage 2000	1,389	403	–	–	–	–
New 3D 2000–2008	1,485	349	319	457	64	90
New 3D 2005–2008	459	104	135	171	27	33

All values are compiled with respect to UniProt version 12.8; *3D coverage 2000*, marks the structural coverage compiled as specified in Methods (10-3 PSI-BLAST) that could have been achieved on UniProt 12.8 with the structures in the PDB by June 30, 2000; *New 3D 2000–2008*, marks the addition of structural coverage over the course of the PSI (from July 1, 2000 to September 16, 2008); *New 3D 2005–2008*, marks the addition of structural coverage over the course of PSI-2 (from July 1, 2005 to September 16, 2008)

PSI per-protein gain in novel leverage is 3–4 fold higher than PDB without PSI

The success of PSI-2 in increasing novel leverage in a competitive environment is being demonstrated most clearly when we compile the annual increase in novel structural coverage *per deposited structure*. The per-structure leverage of PSI has consistently been 5–8 times higher than the corresponding number for non-SG structures (Fig. 3a). At the same time, the novel leverage of a deposited experimental structure has decreased significantly over the past 8 years, both for the PSI, SG non-PSI, and non-SG structures: it is getting increasingly difficult to achieve the high novel leverage values that PSI structures obtained in the earlier years of the program.

Although novel leverage per structure has been dropping, the total number of novel structures solved by PSI groups has increased in each year of the program. Has this sufficed to counterbalance the increase in the difficulty of the task? One answer is provided by Fig. 2b: while the rate of coverage for non-SG and non-US non-SG groups are plateauing, the PSI curve has continued to remain almost

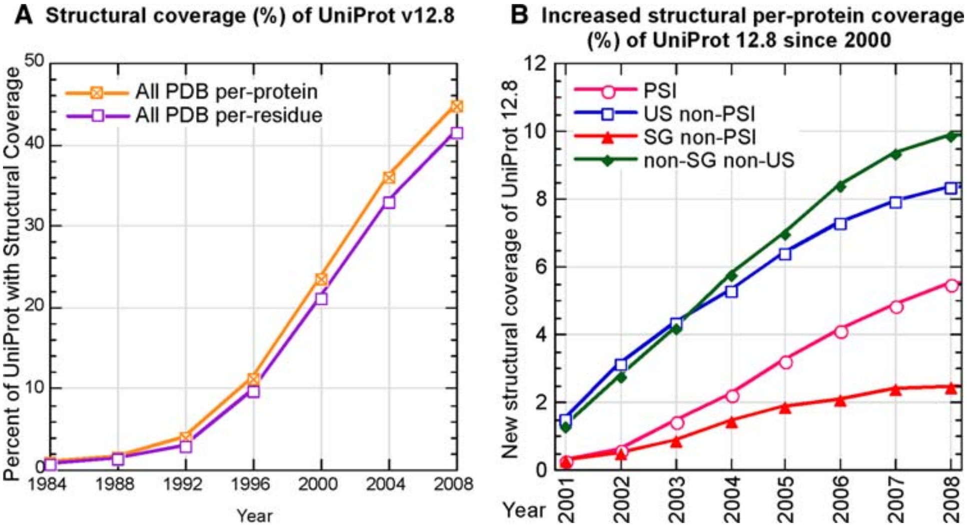

Fig. 2 Increase of structural coverage of UniProt. Plotted are the percentage of proteins (*orange* with *crossed squares*) and residues (*purple* with *open squares*) in the entire UniProt database (release 12.8 Feb. 2008) that potentially be modeled using one of the structures in the PDB as a template, where "modelability" is based on PSI-BLAST alignments (E-value < 10^{-10}) between the sequence of the target and the sequence of the template of known structure. Panel (**a**) shows the percentage of UniProt with structural coverage, per year, while panel (**b**) on the right (coloring as in Fig. 1) zooms in to showing the gain in coverage with respect to the onset of PSI (July 2000). Note that the absolute values of coverage depend crucially on the values chosen for what is considered to be an acceptable model. Our choices of E-values < 10^{-10} provide relatively conservative estimates for high-accuracy models

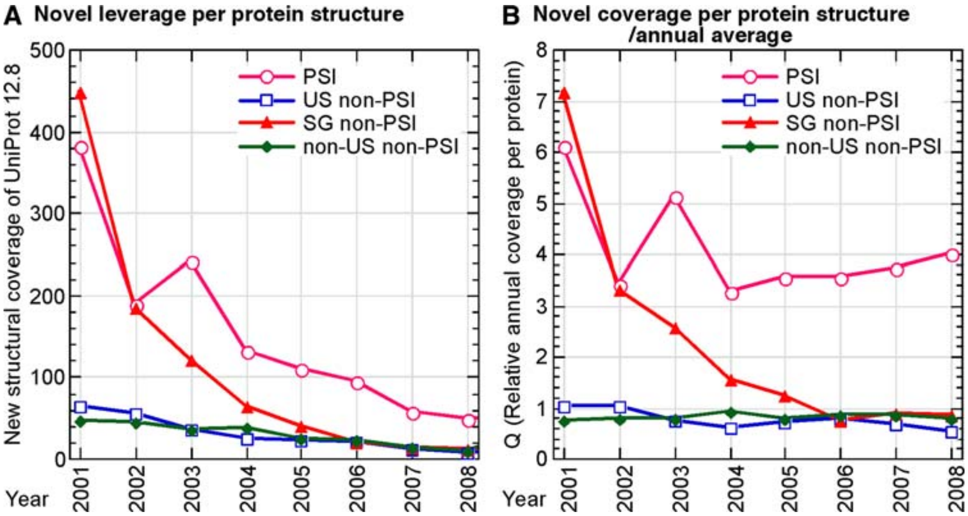

Fig. 3 Per-structure estimates of novel leverage. The left panel (**a**) demonstrates how the non-cumulative (annual) novel leverage for UniProt 12.8 per deposited structure decreases over time because the task of generating high novel leverage becomes increasingly difficult. The right panel (**b**) reports the relative annual coverage per deposited structure (Q, Eq. 3). Values Q below 1 mark contributions below the average over the entire PDB in the year. While the relative values given in Fig. 1 vary little with the particular threshold for what is considered to be a "useful model", the absolute values given in Fig. 2 and Fig. 3 depend crucially on the values chosen for what is considered to be an acceptable model. Coloring as in Fig. 1: *pink* with *open circle*: PSI alone; *blue* with *open squares*: structures from US labs excluding structures claimed by PSI; *red* with *filled triangle*: SG structures from non-PSI efforts; *green* with *filled diamonds*: structures from outside the US not claimed by any SG consortium

linear. Another answer is provided by the contribution to the coverage per deposited structure with respect to the average annual contribution by the following ratio of fractions:

$$Q = \frac{\left(\frac{\text{novel coverage generated by effort } X \text{ in year } Y}{\text{number of protein structures deposited by } X \text{ in } Y}\right)}{\left(\frac{\text{novel leverage generated by PDB in } Y}{\text{number of proteins deposited in } Y}\right)} \qquad (3)$$

shown annually for various efforts in Fig. 3b. Values below 1 imply that effort X contributed less to the coverage per structure than the average over the entire PDB. PSI has consistently contributed over 3 times more than average (Fig. 3b). The contribution of non-PSI SG has also been very high, but in recent years non-PSI SG has reduced to levels just above 1.

PSI has by now targeted and worked on most of the largest 16,787 sequence-structure families with prokaryotic representatives. PSI-2 continues to pick the largest remaining families, however, those become smaller. The novel leverage of all non-PSI structures in the PDB is also decreasing. This is partly due to the same reason: the largest families are either structurally covered or continue to evade structure determination. Furthermore, as already discussed, the generation of novel leverage becomes increasingly challenging.

Does this imply that attempts at experimentally determining structures for new sequence-structure families will be doomed? Despite efforts in optimizing novel leverage and providing structures for as yet uncharacterized domain sequence families, structural genomics has not discovered many truly novel structures [1, 5, 7, 8, 13, 15, 18]. Indeed, the discovery of previously unobserved protein structure space (new geometries and principles not seen before) is becoming increasingly difficult [22]. This implies that (i) we now know most protein structure geometries or folds and (ii) on average, staying within the vicinity of known structures is more likely to result in a successful structure determination. By design, PSI-2 has been attempting and succeeding in targeting proteins which are not similar to proteins with known structures, i.e. to increase the odds of discovering new territories through their development of high-through pipelines and technologies. To rephrase this in a common analogy: by focusing on protein domain families with no structural representatives PSI-2 has systematically targeted and succeeded in reaching "higher-hanging fruits".

Many other criteria for success

Structural genomics, by design, is a hypothesis-generating instead of a hypothesis-driven endeavor. It shares this aspect with many new high-throughput genomics projects in the evolving molecular biology discipline although—unlike other genomics projects—structural genomics continues to generate very high-resolution, detailed molecular data. The success of the PSI is reflected by many aspects which range from increasing the speed of structure determination and deposition (both dramatically increased during this decade), through high literature impact and extreme reduction in the number of papers per structure to the push of automation and robotics which increases the diverse biophysical measurements readily available to researchers in related fields with different expertise.

Objective criteria that allow the monitoring of the degree to which scientific endeavors deliver what they promised are naturally becoming integral parts of a landscape in which the funding for science shrinks, while the challenges for the scientist arguably increase, and in which an increasing fraction of all science is funded by temporal grants. Here, we have demonstrated that PSI-2 has been extremely successful by the aims it posed at the start: it contributed substantially toward the increase of novel leverage to the extent that a future without PSI will clearly imply a considerable lengthening of the time needed to cover today's protein universe.

Given that the PSI was successful in meeting the milestones that the PSI commission posed, the aim now is to finish with a wider perspective that considers the optimization of structural coverage as a means and not as an end. One aspect of structural genomics is the adventure of mapping unknown spaces. We seek connections to create maps. These objectives require the coincidence of a wealth of sequences and structures in spaces that have hardly been experimentally covered (i.e. families of unknown function) but appear to be extremely important, as demonstrated by the annotations for the universal family of EVE/PUA/PUA-like proteins enriched by structural genomics [5]. All these connections contribute to the understanding of protein evolution. The PSI has covered an immense fraction of the prokaryotic sequence-space in terms of generating protocols, reagents, and experimental data. This wealth is available today through the PSI Materials Repository (http://www.hip.harvard.edu/gateway/) and through the PSI Knowledge Base (http://kb.psi-structuralgenomics.org/). A relatively small fraction of the target families have so far yielded experimental structures, but this "small" fraction now contributes over one-third of the novel leverage worldwide, providing structural templates for over 300,000 new reliable protein structure models.

Another long-term impact is the contribution toward making structure become an integral part of molecular biology and toward converting structure determination from an amazing art mastered by few into a pipeline accessible to many. Clearly the cost reduction, the development of sophisticated semi-automated high throughput pipelines contributed immensely to making this happen. Without structural genomics, today's level of automation would not have been reached at all. The development of cheaper sequencing techniques was certainly no goal of the human sequencing project. But those techniques have been changing biology immensely over the last decade.

16–20 years to go to complete coverage of sequence universe?

How much more is left to do? The following rationale provides an over simplified answer. Firstly, we have estimated that at least 20% of all residues in proteomes are not viable targets for structural genomics because they encode complex integral membrane proteins, long continuous coiled-coils regions, long regions that are natively unstructured, and leftovers from partial models (e.g. model A covers domain D1 from residues 6–55, model B covers domain D2 from residues 61–100 in a protein of 100 residues; this leaves 10 residues 1–5 and 56–60 as non-viable targets) [21]. Most of these 20% of the residues are in short regions not assigned to a particular domain and are probably some sort of domain linkers and embellishments. Put differently, 80% per-residue coverage implies "completion". Secondly, today's coverage is about 40%, i.e. 40% (80–40) remains to be done. Thirdly, extrapolating from Fig. 2a, we might estimate the average annual per-residue growth in coverage of UniProt 12.8 to be about 2.5%. Assuming this rate to hold for the future, we would estimate 16 years (40/2.5 = 16) to structurally cover whatever remains of the UniProt 12.8 sequence database. While sequence space continues to grow, much of this new growth maps to domain families covered by this 80% of current proteins sequence universe.

Clearly, the assumption of identical growth is overly optimistic: the rate has been kept at a linear growth only due to the focused effort of structural genomics. Given that PSI-2 has already cloned almost all the largest viable families, it is clear that the future leverage will be lower. Moreover, as new genomes are sequenced, only a fraction of these sequences map to known protein domain families, and the uncovered protein universe continues to grow.

Furthermore, it might be argued that the 40% of the residues that remain to be structurally explored will constitute proteins that are much more challenging for structure determination than those in the 40% of the residues that are covered today. If so, structural genomics methods might fail to capture those residues in these much more challenging classes of proteins, and our assumption of a constant growth rate might be inappropriate. True, this might be so, and we have no scientific argument to dispel this concern. However, we can move back into the past and pretend to estimate for what was then the future: e.g. if we had taken the growth rate from 1994 to 2000 to estimate the coverage of 2008, we would have been completely right (Fig. 2a).

Where from here?

We have established structural genomics as an extremely efficient way to discover new areas in the protein universe that will undoubtedly continue to invoke testable hypothesis for years to come. Will the trend continue? Can we extrapolate from today's data, or will we need something completely different to efficiently cover what remains? Clearly, we have to improve structural determination for sequence-structure families from eukaryotes. Today, it requires some 5–10-fold more resources to determine the structure for an average eukaryotic protein than for an average prokaryotic protein. A considerable fraction of the untouched sequence space falls into sequence-structure families that exclusively represent eukaryotes. Clearly, targeting this important domain becomes an important objective. Another fact of PSI-2 was that structure determination has so far succeeded for less than 30% of all families targeted. Developing techniques that allow a substantial increase in this yield appears to be another important goal.

The final question seems to be hovering around the issue of how much will the part of the universe without structural coverage differ from the part we cover today? Clearly, we need to find ways to make structural genomics work for types of proteins for which it has so far had only limited success, including membrane proteins, eukaryotic proteins, and secreted proteins. Are there any new *structural principles* out there that remain to be discovered and that totally elude today's techniques for structure determination? Biology is so full of innovation and surprise that the answer will clearly be in the affirmative. To which extent this will be the case remains utter speculation. However, we have strong evidence that a considerable part of what is left falls into the category of proteins that are unusually flexible, or intrinsically unstructured and that possibly do not adopt regular structures without a binding partner. Do we therefore have to step up in terms of complexity and attack the problem of a structural genomics for complexes? Clearly, this will be one of the important challenges for both the short-term and long-term of structural genomics.

Acknowledgments Thanks to Marco Punta (Columbia) for critical comments on the manuscript and work; to Paul Glick and Guy Yachdav (Columbia) for computer assistance, and to all those who contribute to making structural genomics become an exciting scientific breakthrough. This work was supported primarily by the grant U54-GM074958-01 to the Northeast Structural Genomics consortium (NESG). PSI would not be possible without the dedication and continued support of grant agencies and in particular of those administrating grants. We specially like to acknowledge many at NIGMS, in particular, of Jeremy Berg, John Norvell, Charles Edmonton, and Jerry Li. Last, not least, thanks to Amos Bairoch (SIB, Geneva), Rolf Apweiler (EBI, Hinxton), Phil Bourne (San Diego Univ.), John Westbrook (Rutgers), Helen Berman (Rutgers), and their crews for maintaining excellent databases and to all experimentalists who enabled this analysis by making their data publicly available.

References

1. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2008) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36:D419–D425. doi:10.1093/nar/gkm993
2. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M et al (2004) UniProt: the universal protein knowledgebase. Nucleic Acids Res 32:D115–D119. doi:10.1093/nar/gkh131
3. Berman HM, Burley SK, Chiu W, Sali A, Adzhubei A, Bourne PE, Bryant SH, Dunbrack RL Jr, Fidelis K, Frank J et al (2006) Outcome of a workshop on archiving structural models of biological macromolecules. Structure 14:1211–1217. doi:10.1016/j.str.2006.06.005
4. Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 35:D301–D303. doi:10.1093/nar/gkl971
5. Bertonati C, Punta M, Fischer M, Yachdav G, Forouhar F, Zhou W, Kuzin AP, Seetharaman J, Abashidze M, Ramelot TA et al (2008) Structural genomics reveals EVE as a new ASCH/PUA-related domain. Proteins. doi:10.1002/prot.22287
6. Bhattacharya A, Wunderlich Z, Monleon D, Tejero R, Montelione GT (2008) Assessing model accuracy using the homology modeling automatically software. Proteins 70:105–118. doi:10.1002/prot.21466
7. Bourne PE, Allerston CK, Krebs W, Li W, Shindyalov IN, Godzik A, Friedberg I, Liu T, Wild D, Hwang S, et al. (2004) The status of structural genomics defined through the analysis of current targets and structures. Pac Symp Biocomput 9:375–386
8. Chandonia JM, Brenner SE (2005) Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches. Proteins 58:166–179. doi:10.1002/prot.20298
9. Chen L, Oughtred R, Berman HM, Westbrook J (2004) TargetDB: a target registration database for structural genomics projects. Bioinformatics 20:2860–2862. doi:10.1093/bioinformatics/bth300
10. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5:823–826
11. Fernandez-Fuentes N, Rai BK, Madrid-Aliste CJ, Fajardo JE, Fiser A (2007) Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments. Bioinformatics 23:2558–2565. doi:10.1093/bioinformatics/btm377
12. Fraser-Liggett CM (2005) Insights on biology and evolution from microbial genome sequencing. Genome Res 15:1603–1610. doi:10.1101/gr.3724205
13. Gerstein M, Edwards A, Arrowsmith CH, Montelione GT (2003) Structural genomics: current progress. Science 299:1663. doi:10.1126/science.299.5613.1663a
14. Grant A, Lee D, Orengo C (2004) Progress towards mapping the universe of protein folds. Genome Biol 5:107. doi:10.1186/gb-2004-5-5-107
15. Harrison A, Pearl F, Sillitoe I, Slidel T, Mott R, Thornton J, Orengo C (2003) Recognizing the fold of a protein structure. Bioinformatics 19:1748–1759. doi:10.1093/bioinformatics/btg240
16. Koh IYY, Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Narayanan E, Grana O, Valencia A, Sali A, Rost B (2003) EVA: evaluation of protein structure prediction servers. Nucleic Acids Res 31:3311–3315. doi:10.1093/nar/gkg619

17. Kopp J, Schwede T (2004) The SWISS-MODEL repository of annotated three-dimensional protein structure homology models. Nucleic Acids Res 32:D230–D234. doi:10.1093/nar/gkh008
18. Levitt M (2007) Growth of novel protein structural data. Proc Natl Acad Sci USA 104:3183–3188. doi:10.1073/pnas.0611678104
19. Liu J, Rost B (2003) Domains, motifs, and clusters in the protein universe. Curr Opin Chem Biol 7:5–11. doi:10.1016/S1367-5931(02)00003-0
20. Liu J, Rost B (2004) CHOP: parsing proteins into structural domains. Nucleic Acids Res 32:W569–W571. doi:10.1093/nar/gkh481
21. Liu J, Hegyi H, Acton TB, Montelione GT, Rost B (2004) Automatic target selection for structural genomics on eukaryotes. Proteins 56:188–200. doi:10.1002/prot.20012
22. Liu J, Montelione GT, Rost B (2007) Novel leverage of structural genomics. Nat Biotechnol 25:849–851. doi:10.1038/nbt0807-849
23. Marsden RL, Orengo CA (2008) Target selection for structural genomics: an overview. Methods Mol Biol 426:3–25. doi:10.1007/978-1-60327-058-8_1
24. Marti-Renom MA, Stuart A, Fiser A, Sanchez R, Melo F, Sali A (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29:291–325. doi:10.1146/annurev.biophys.29.1.291
25. Marti-Renom MA, Madhusudhan MS, Fiser A, Rost B, Sali A (2002) Reliability of assessment of protein structure prediction methods. Structure 10:435–440. doi:10.1016/S0969-2126(02)00731-1
26. Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A (2005) Critical assessment of methods of protein structure prediction (CASP)-round 6. Proteins 61:3–7. doi:10.1002/prot.20716
27. Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A (2007) Critical assessment of methods of protein structure prediction-round VII. Proteins 69(Suppl 8):3–9. doi:10.1002/prot.21767
28. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540
29. Nair R, Fajardo E, Fiser A, Godzik A, Jaroszewski L, Marsden R, Orengo C, Rost B (2008) Progress at PSI—milestones measuring the success of structural genomics in the USA. Columbia University, New York
30. Norvell JC, Berg JM (2007) Update on the protein structure initiative. Structure 15:1519–1522. doi:10.1016/j.str.2007.11.004
31. Orengo CA, Michie AD, Jones DT, Swindells MB, Thornton JM (1997) CATH—a hierarchic classification of protein domain structures. Structure 5:1093–1108. doi:10.1016/S0969-2126(97)00260-8
32. Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis FP, Stuart AC, Mirkovic N, Rossi A, Marti-Renom MA, Fiser A et al (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 32:D217–D222. doi:10.1093/nar/gkh095
33. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D et al (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 34:D291–D295. doi:10.1093/nar/gkj059
34. Redfern OC, Harrison A, Dallman T, Pearl FM, Orengo CA (2007) CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. PLoS Comput Biol 3:e232. doi:10.1371/journal.pcbi.0030232
35. Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68. doi:10.1002/prot.340090107
36. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43. doi:10.1038/nature02340
37. Watson JD, Todd AE, Bray J, Laskowski RA, Edwards A, Joachimiak A, Orengo CA, Thornton JM (2003) Target selection and determination of function in structural genomics. IUBMB Life 55:249–255. doi:10.1080/1521654031000123385
38. Yeats C, Lees J, Reid A, Kellam P, Martin N, Liu X, Orengo C (2008) Gene3D: comprehensive structural and functional annotation of genomes. Nucleic Acids Res 36:D414–D418. doi:10.1093/nar/gkm1019
39. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W et al (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 5:e16. doi:10.1371/journal.pbio.0050016