

# Supplementary Information

**Jinfeng Liu, Gaetano T. Montelione and Burkhard Rost**

## **TOC for Supplementary Information:**

- Materials and Methods
- Table S1: Comparison of novel leverage at the protein level.
- Table S2: Comparison of novel leverage at the residue level.
- Figure S1: Structural coverage of UniProt database since 1985.
- Figure S2: PSI is cost efficient in obtaining novel leverage.

## Materials and Methods

### Data sets

The list of protein structures deposited by world-wide structural genomics efforts and PSI centers are obtained from TargetDB (<http://targetdb.pdb.org/>). Structures deposited before September 1, 2000 or after August 31, 2006 are excluded. ‘On hold’ structures are included in the analysis as long as the sequences are available. Sequences are taken from the SEQRES records of PDB entries. Yearly statistics are compiled according to PSI fiscal year, i.e., Y2001 means September 1, 2000 - August 31, 2001. Cost analysis of PSI centers is based on published PSI budgets of total costs (direct + indirect): \$30 million in Y2001, \$40 million in Y2002, \$53 million in Y2003, \$71 million in Y2004, and \$71 million in 2005. In addition to the statistics for all PSI centers, we also compiled the numbers for the four largest PSI centers: Joint Center for Structural Genomics (JCSG), Midwest Center for Structural Genomics (MCSG), New York Structural GenomiX Research Consortium (NYSGXRC), and Northeast Structural Genomics Consortium (NESG).

### Calculating the leverage value

We define the leverage value for a structure as the number of proteins or residues in a specified version of UniProt database that can be aligned with the query structure under certain threshold. Specifically, for each query structure  $q$ , we run PSI-BLAST against UniProt database version 7.6 using parameters “-j 3 -F T -h 5e-4 -b 3000 -v 3000”, and remove the high-scoring segment pairs (HSPs) with expect value larger than  $1e-10$  from the last iteration of the PSI-BLAST output. For each subject protein  $s$  in the alignment with one or more significant HSPs, the leverage of  $q$  with regard to  $s$  at the residue level,  $Lev_{res}(q,s)$ , is obtained by counting the number of residues in  $s$  that are covered by all significant HSPs (i.e., the union of the HSPs). The total leverage of  $q$  at the residue level,  $Lev_{res}(q)$ , is the sum of  $Lev_{res}(q,s)$  over all possible subject proteins  $s$ . Similarly, to calculate the leverage of a group of structures  $Q$ ,  $Lev_{res}(Q)$ , we first get  $Lev_{res}(Q,s)$  by taking the union of all HSPs covering  $s$  (this time from many alignments instead of just one), and then sum over all possible subject proteins  $s$ .

To calculate the leverage value of query  $q$  at the protein level, we simply count the number of subject proteins that have at least 50 residues (or 50% of the entire protein) covered by the significant HSPs.

$$\text{Lev}_{\text{prot}}(q, s) = \begin{cases} 1 & \text{if } \text{Lev}_{\text{res}}(q, s) \geq 50 \\ 0 & \text{if } \text{Lev}_{\text{res}}(q, s) < 50 \end{cases} \quad (\text{Eqn. 1})$$

$$\text{Lev}_{\text{prot}}(q) = \sum_s \text{Lev}_{\text{prot}}(q, s) \quad (\text{Eqn. 2})$$

$$\text{Lev}_{\text{prot}}(Q, s) = \begin{cases} 1 & \text{if } \text{Lev}_{\text{res}}(Q, s) \geq 50 \\ 0 & \text{if } \text{Lev}_{\text{res}}(Q, s) < 50 \end{cases} \quad (\text{Eqn. 3})$$

$$\text{Lev}_{\text{prot}}(Q) = \sum_s \text{Lev}_{\text{prot}}(Q, s) \quad (\text{Eqn. 4})$$

### Calculating the novel leverage values

The novel leverage value for a structure is defined as the number of proteins or residues in UniProt that can be aligned with the query structure, but can not be aligned with any structures deposited in PDB before the deposition date of the query structure. Operationally, we obtain the novel leverage of  $q$  with regard to  $s$  at the residue level,  $\text{NovLev}_{\text{res}}(q, s)$ , by taking the union of all significant HSPs covering  $s$  in the alignment with  $q$ , and then subtracting those residues that are in the HSPs in the alignments with all previously determined structures  $P$ . For a protein to be counted towards novel leverage, it must have less than 50 residues and less than 50% of the entire protein covered by  $P$  and more than 50 residues (or 50% of the entire protein) covered by  $q$ . The rest of procedure is similar to obtaining the total leverage values as described in the preceding paragraph.

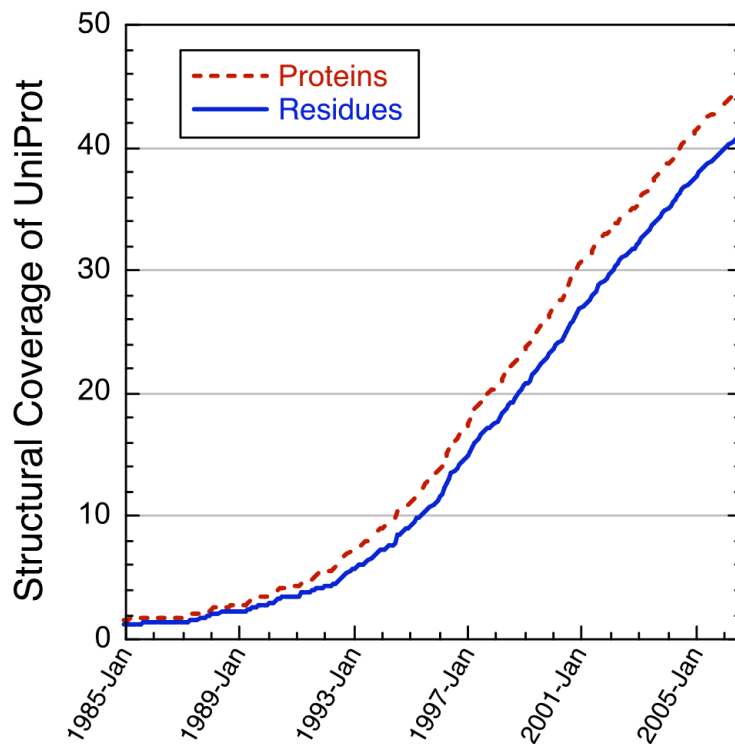
$$\text{NovLev}_{\text{prot}}(q, s) = \begin{cases} 1 & \text{if } \text{Lev}_{\text{res}}(q, s) \geq 50 \text{ and } \text{Lev}_{\text{res}}(P, s) < 50 \\ 0 & \text{otherwise} \end{cases} \quad (\text{Eqn. 5})$$

**Table S1: Comparison of novel leverage at the protein level**

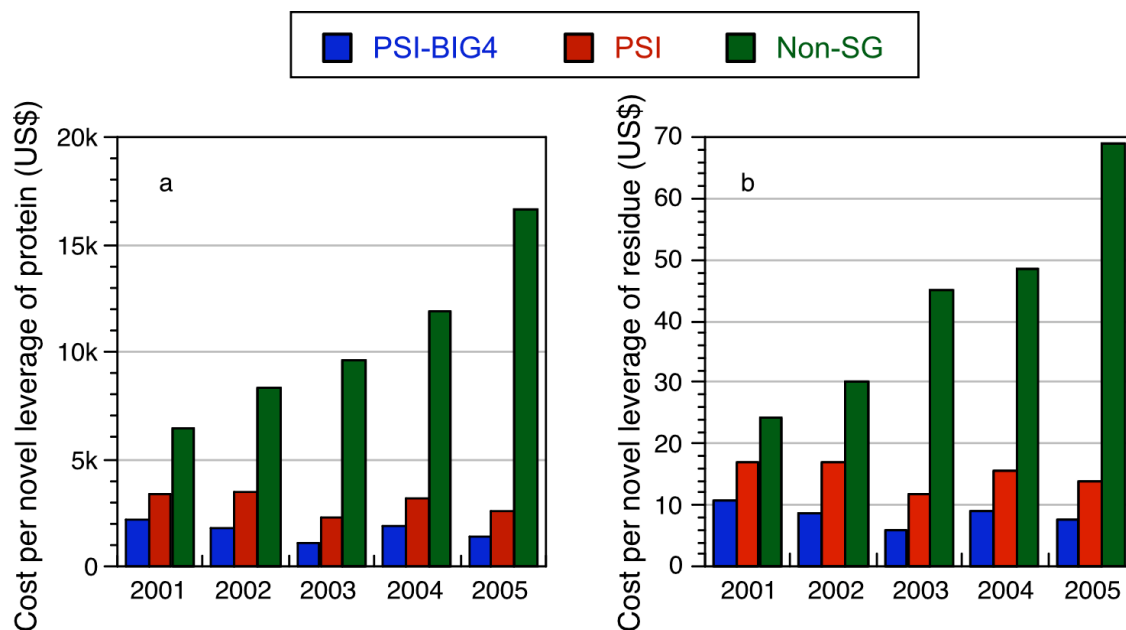
	<b>PSI-BIG4</b>	<b>PSI</b>	<b>SG</b>	<b>Non-SG</b>	<b>PDB</b>
All	95,188	113,262	161,947	460,390	600,519
Eukaryotic	12,501	16,674	34,833	153,877	182,852
Prokaryotic	81,907	95,778	126,070	270,832	381,218
Human	746	943	3,282	11,557	14,209

**Table S2: Comparison of novel leverage at the residue level**

	<b>PSI-BIG4</b>	<b>PSI</b>	<b>SG</b>	<b>Non-SG</b>	<b>PDB</b>
All	19,042,587	22,394,741	31,401,140	122,677,081	153,143,111
Eukaryotic	2,602,848	3,336,699	5,927,079	42,166,320	47,932,240
Prokaryotic	16,319,753	18,927,547	25,314,771	72,810,121	97,376,060
Human	153,857	189,353	489,183	2,993,960	3,471,716



**Fig. S1: Structural coverage of UniProt database (release 7.6) since 1985.** Structural coverage is defined as the percentage of proteins and residues in UniProt that can be aligned to PDB structures by PSI-BLAST (Supplementary methods) at the expect value threshold of  $1e-10$ .



**Fig. S2: PSI is cost efficient in obtaining novel leverage.** We evaluated the cost of structure determination in the context of obtaining novel leverage. Lacking good estimates for the amount spent on structural biology worldwide, we based our comparison on the often quoted assumption that the average total cost (including overhead costs) of solving a protein structure by traditional means is about \$250,000<sup>1,2</sup>. It should be noted that structural genomics (SG) and traditional structural biology (non-SG) have different focus and the cost for non-SG may include cost of functional characterization of the proteins. **PSI-BIG4**: four largest centers of PSI (JCSG, MCSG, NESG, and NYSGXRC). The cost per novel leverage of **(a)** protein and **(b)** residue for non-SG structures has been increasing constantly; in contrast, it has been decreasing for PSI structures.

1. Chandonia, J.M. & Brenner, S.E. *Science* **311**, 347-351 (2006).
2. Service, R. *Science* **307**, 1554-1558 (2005).