THE IMPACT OF STRUCTURAL GENOMICS:

Expectations and Outcomes

John-Marc Chandonia and Steven E. Brenner

STRUCTURAL GENOMICS

- Similar to structural biology in the methods used
  - X-ray crystallography, NMR
- Attempts to determine the structure of all the proteins of a given organism
- Focuses on High Throughput Screening (HTS)
  - Robotics, data processing software, sensitive detectors
- Economy of scale

PROTEIN STRUCTURE INITIATIVE (PSI)

- Focuses on decreasing the cost and time associated with 3-D protein structure determination using structural genomics
- 10 year, 764 million budget
- Two phases
  - Phase 1 (2000-2005)
    - To develop methods that streamline the determination of protein structures
  - Phase 2 (2005-2010)
    - To use methods developed in Phase 1 to determine a large number of protein structures, and to continue to streamline the processes in structural genomics

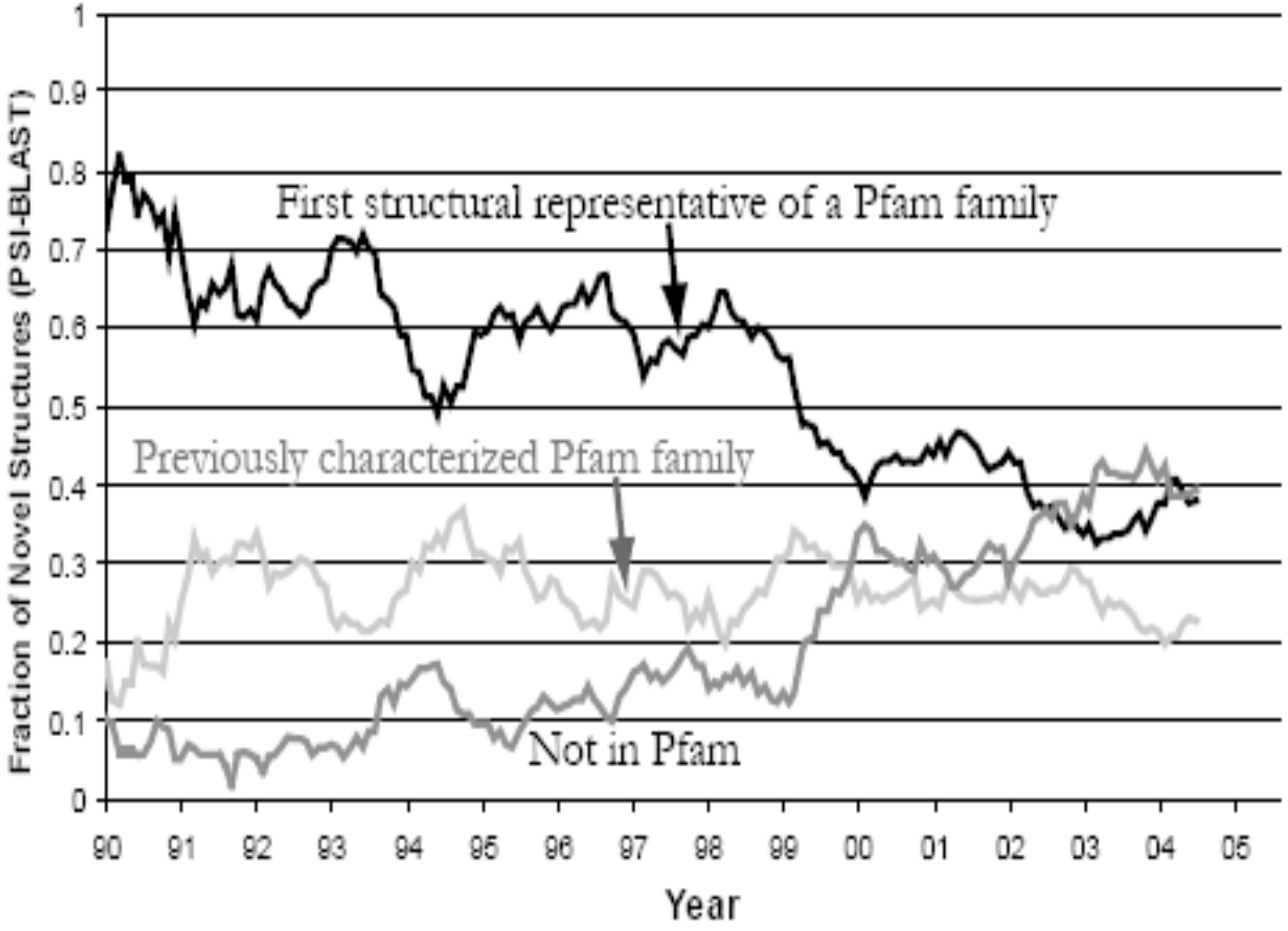
GOALS

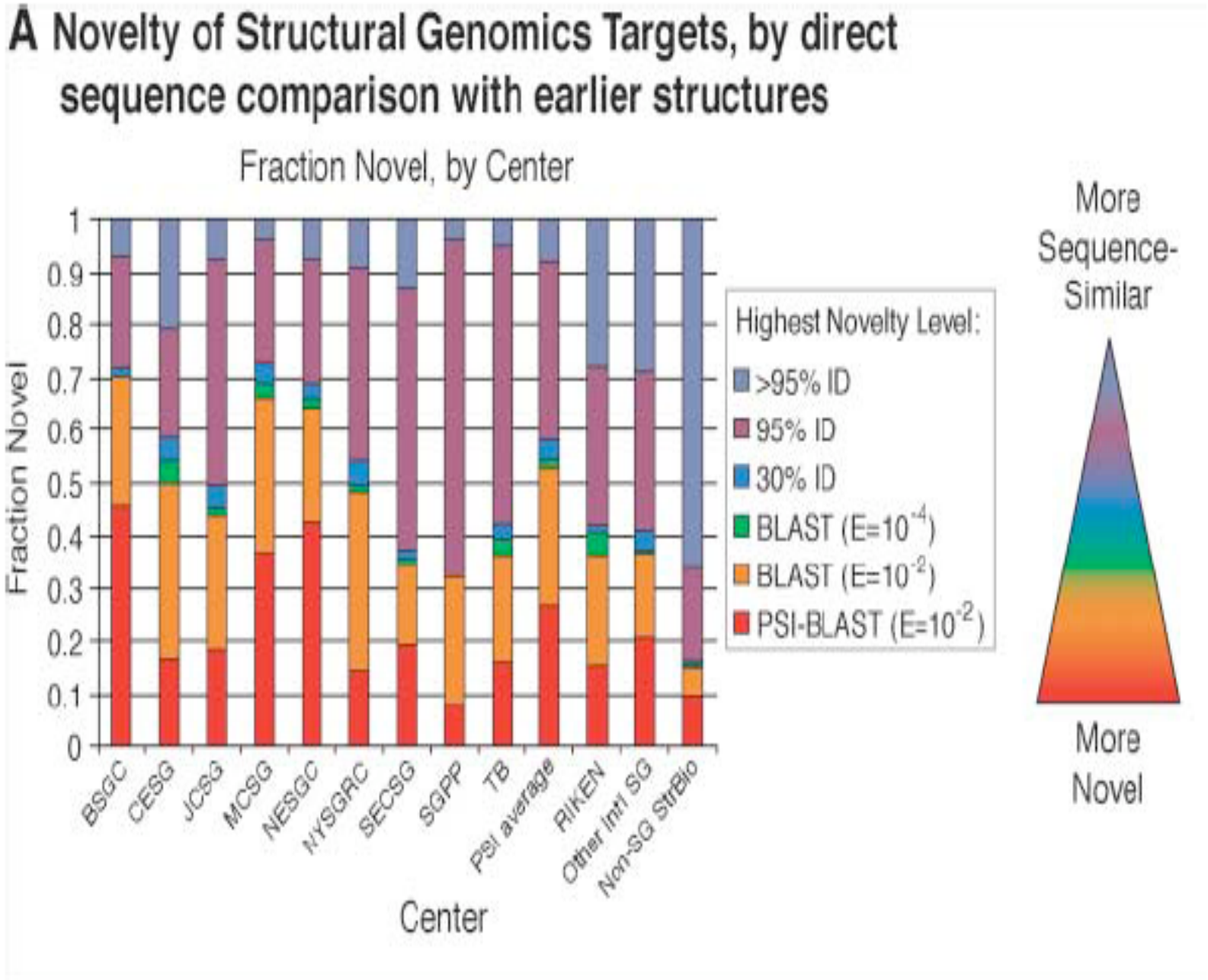
- Measures of success
  - Biological importance and difficulty
  - Novel structures
    - First protein in a family can have its structure and function evaluated, then its properties can be translated to:
      - Create comparative models
      - Find new evolutionary relationships between proteins

FINDING NEW STRUCTURES BY SEQUENCE COMPARISON

- BLAST and PSI-BLAST were used to determine sequence similarity, to help remove bias introduced by Pfam
  - Pfam does not include many species specific proteins
  - The number of novel structures has decreased over the last 15 years
    - 20% in 1990, 10% in 2005
  - According to the analysis, SG structures accounted for 44% of the total number of new structures in 2004

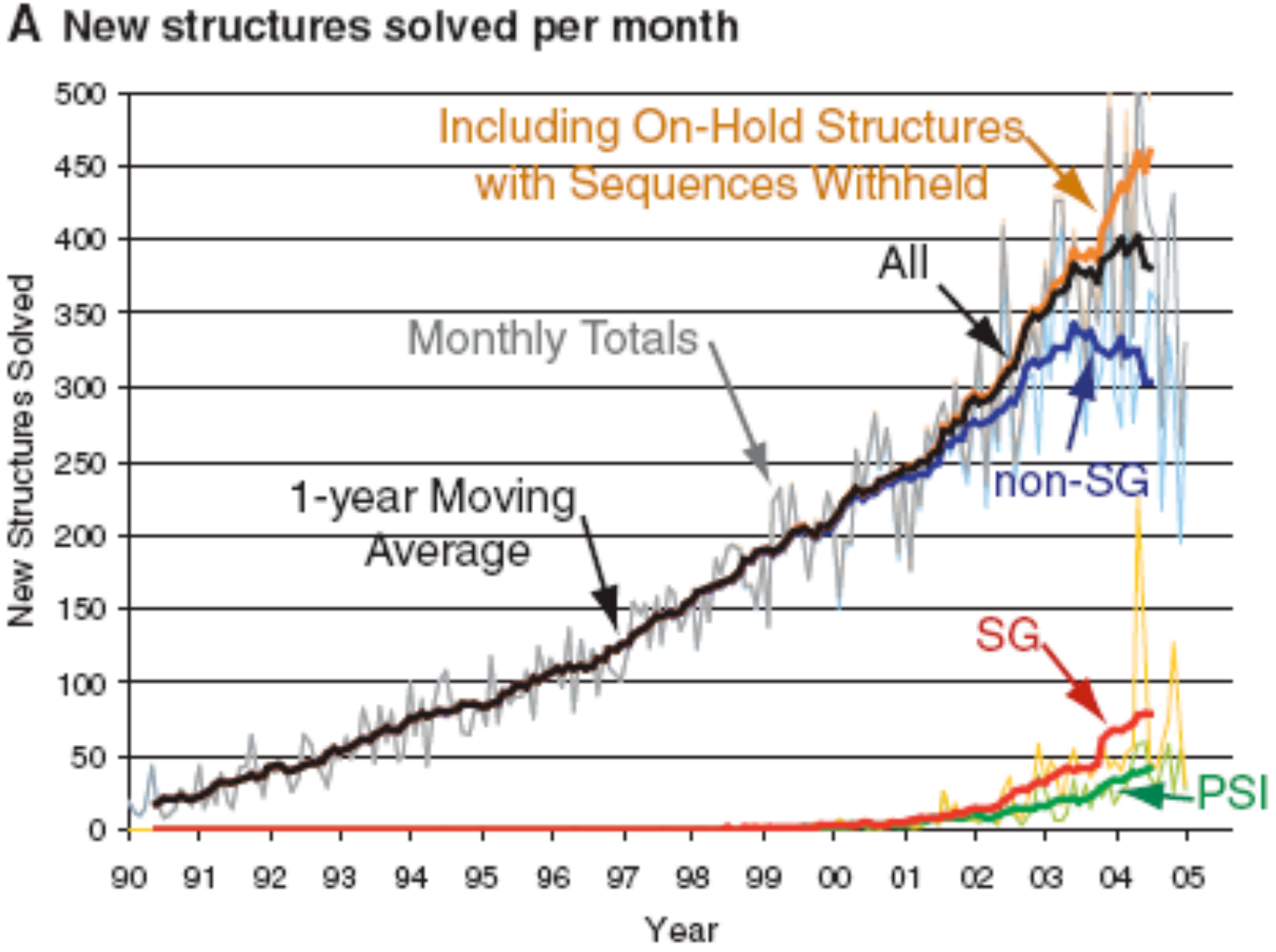
c) Overlap between PSI-BLAST and Pfam

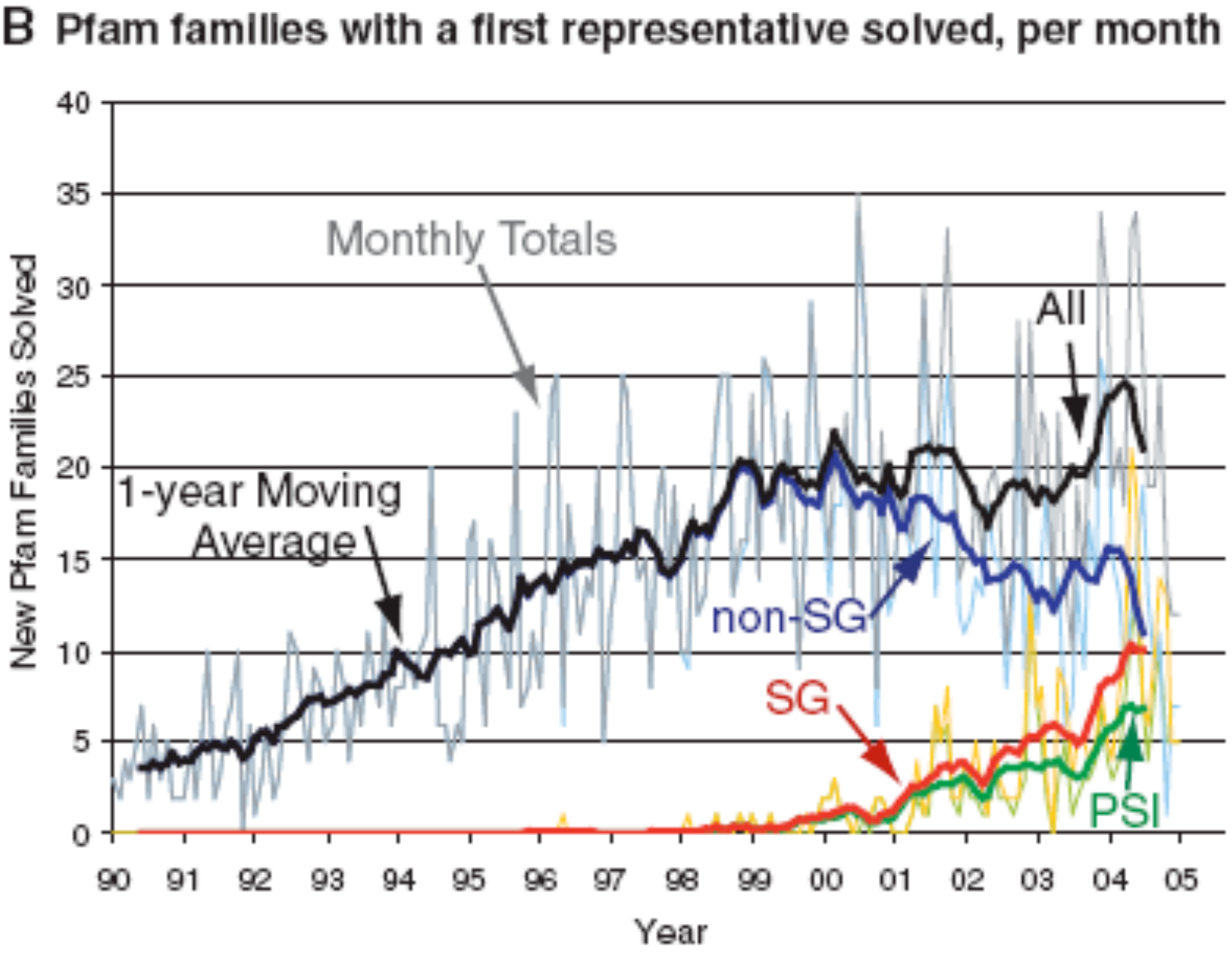



A Novelty of Structural Genomics Targets, by direct sequence comparison with earlier structures
Fraction Novel, by Center
Fraction Novel
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
BSGC
CESG
JCSG
MCSG
NESGC
NYSGRC
SECSG
SGPP
TB
PSI average
RIKEN
Other Int'l SG
Non-SG StrBio
Center
Highest Novelty Level:
>95% ID
95% ID
30% ID
BLAST (E=10-4)
BLAST (E=10-2)
PSI-BLAST (E=10-2)
More Sequence-Similar
More Novel

IMPACT OF STRUCTURAL GENOMICS ON COVERAGE OF PROTEIN FAMILIES

- Pfam families (since 2000)
  - 20.4% of structures reported by PSI represented new families
  - 5% of structures reported by non-SG labs represented new families
  - Structure determination of first structures in a Pfam family by non-SG has decreased, while SG centers have made up the difference
- SG centers now account for half of new structurally characterized families, but only make up 20% of new structures


A New structures solved per month
New Structures Solved
500
450
400
350
300
250
200
150
100
50
0
90
91
92
93
94
95
96
97
98
99
00
01
02
03
04
05
Year
Including On-Hold Structures with Sequences Withheld
All
Monthly Totals
non-SG
1-year Moving Average
SG
PSI


B Pfam families with a first representative solved, per month
New Pfam Families Solved
40
35
30
25
20
15
10
5
0
Monthly Totals
All
1-year Moving Average
non-SG
SG
PSI
90
91
92
93
94
95
96
97
98
99
00
01
02
03
04
05
Year

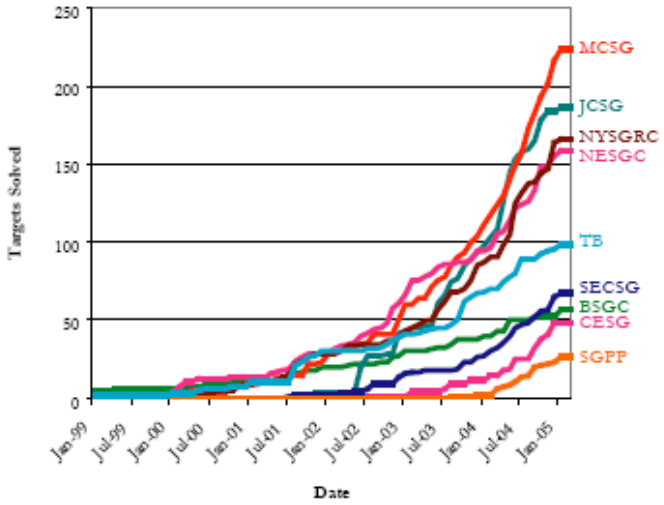
NOVEL STRUCTURES SOLVED

Group or SG center	Targets and nonidentical chains	New Pfam families (total family size)	Novel structures (30% ID)	New SCOP folds	New SCOP fold or superfamily
SG centers					
Berkeley Structural Genomics Center (BSGC)	57 (57 chains)	22 (5757)	41	4	6
Center for Eukaryotic Structural Genomics (CESG)	48 (48 chains)	7 (387)	28	0	0
Joint Center for Structural Genomics (JCSG)	186 (187 chains)	32 (4875)	92	3	4
Midwest Center for Structural Genomics (MCSG)	224 (229 chains)	55 (5512)	163	18	25
Northeast Structural Genomics Consortium (NESGC)	159 (159 chains)	52 (4811)	108	15	26
New York Structural Genomics Research Consortium (NYSGRC)	166 (171 chains)	27 (3982)	90	6	9
Southeast Collaboratory for Structural Genomics (SECSG)	67 (67 chains)	6 (1079)	25	0	1
Structural Genomics of Pathogenic Protozoa Consortium (SGPP)	26 (26 chains)	1 (19)	8	2	2
TB Structural Genomics Consortium (TB)	99 (99 chains)	9 (3938)	42	0	1
PSI centers (total of 9 centers above)	1032 (1043 chains)	211 (30,360)	597	48	74
Japanese center (RIKEN)	686 (718 chains)	50 (6860)	289	10	20
Other international SG (total, excluding all centers above)	169 (183 chains)	33 (5877)	69	6	9
Non-SG groups (since 2000)					
Non-SG structural biology (total)	17,096 (23,747 chains)	928 (249,171)	2,521	269	478
Steitz group	46 (559 chains)	23 (4190)	31	7	12
Huber group	185 (273 chains)	8 (679)	38	5	10
Iwata group	14 (54 chains)	14 (7960)	20	2	3

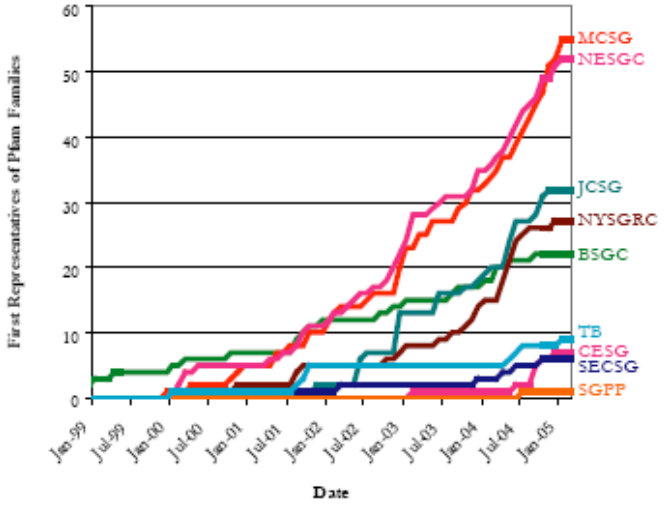
IMPACT OF STRUCTURAL GENOMICS ON THE STRUCTURAL CLASSIFICATION OF PROTEINS

- SCOP classifications
  - Family
    - Clear common evolutionary origin, one family member can be used to construct comparative models
  - Superfamily
    - Groups of families that have similar structure or functions that imply a common evolutionary origin
  - Fold
    - Superfamilies that share similar secondary structures, but have little evidence of common evolutionary origin

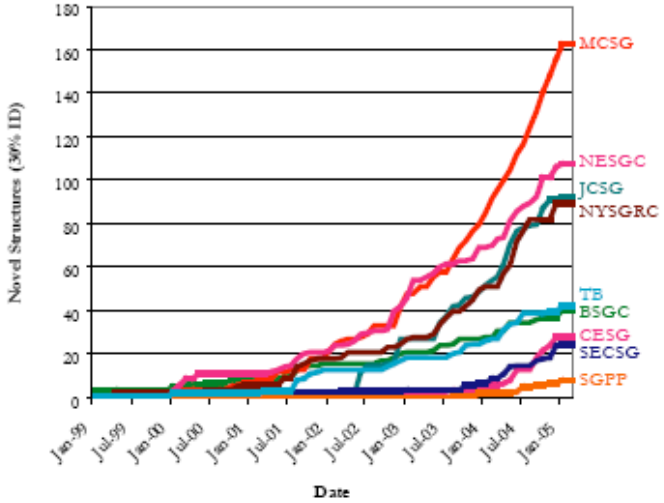
a) Number of Targets Solved at PSI Pilot Centers



b) Number of First Representatives of Pfam families



c) Number of Novel Structures (30% ID)



d) Number of New SCOP Folds or Superfamilies

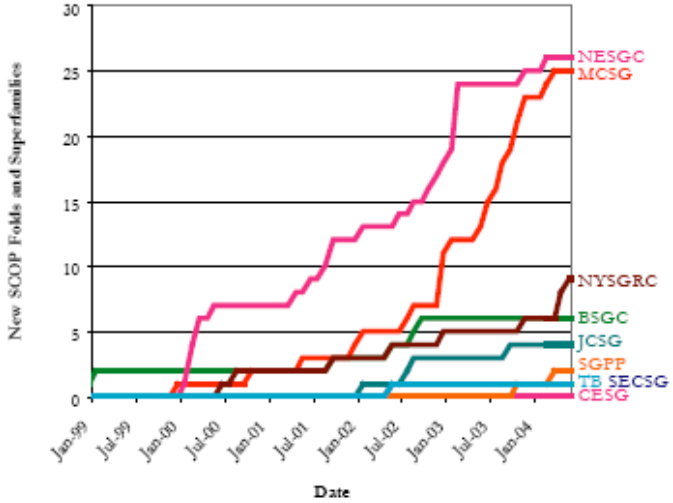
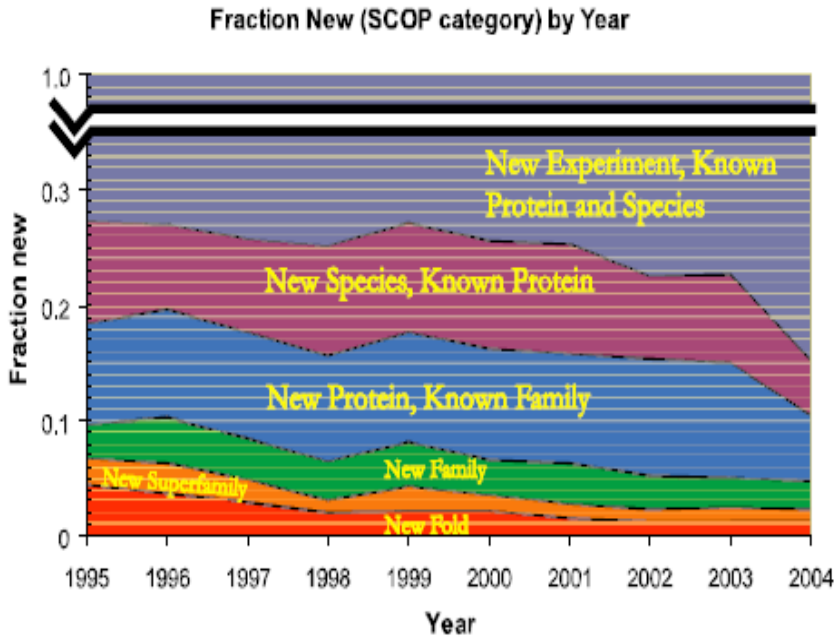
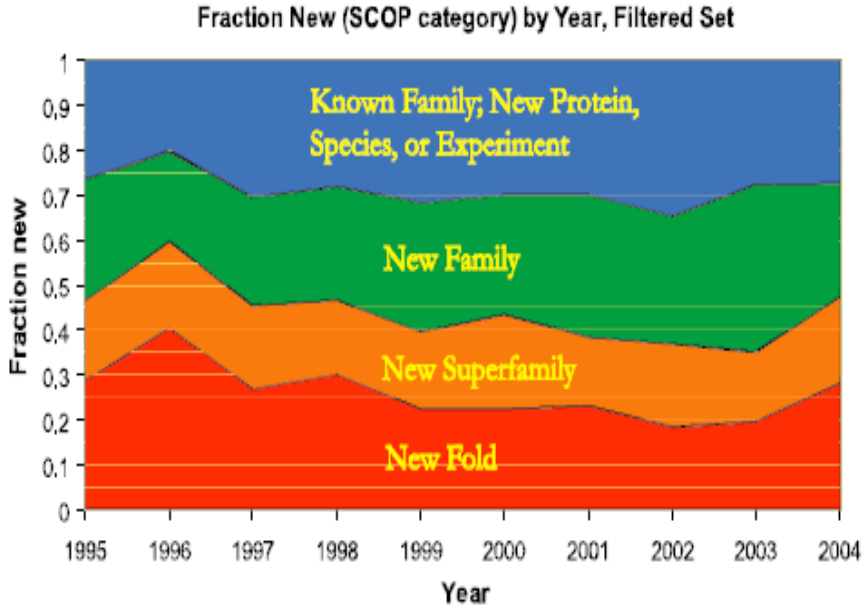


Fig. S4: Time Course of Results for PSI Centers

a) Novelty of non-SG StrBio PDB entries in SCOP



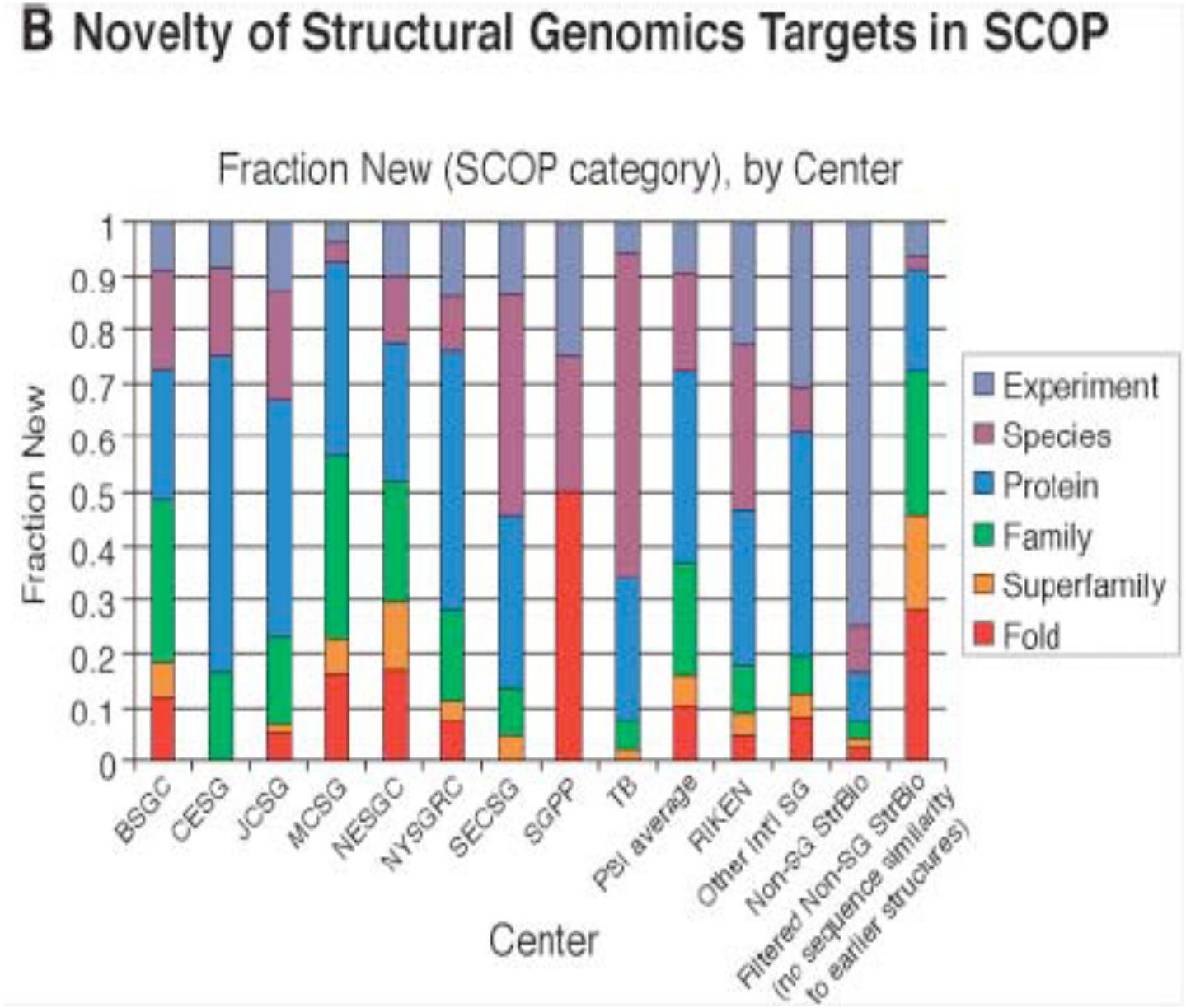
b) Novelty of non-SG StrBio PDB entries without Sequence Similarity to Previously Solved Structures



- 70% of Non-SG structures in the last 10 years represent a new experiment on a protein with a previously determined structure
- The percentage of domains that represent a new family in SCOP has decreased from 9.6% in 1995 to 4.4% in 2004
  - Structural biologist tend to work on known proteins more often

RESULTS VS. EXPECATIONS

- For PSI, the number of domains that represented a new SCOP domain or superfamily was 16%
  - Higher than non-SG average of 4%, but not at the target of 40%
  - Can use BLAST or PSI-BLAST to avoid finding structures for homologs of known proteins
  - Higher sequence novelty correlates with higher rates of discovery for new folds, superfamilies and families


B Novelty of Structural Genomics Targets in SCOP
Fraction New (SCOP category), by Center
Fraction New
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
BSGC
CESG
JCSG
MCSG
NESGC
NYSGRC
SECSG
SGPP
TB
PSI average
RIKEN
Other Intl SG
Non-SG StrBio
Filtered Non-SG StrBio (no sequence similarity to earlier structures)
Center
Experiment
Species
Protein
Family
Superfamily
Fold

COST EFFECTIVENESS

- Average costs for determining a structure with <95% sequence similarity
  - Non-SG: $250,000 to $300,000
  - PSI: $211,000 (overall)
    - 2004-2005: $138,000
      - Most productive center (MCSG):$67,000
    - Adjusted for protein size and composition
      - Larger complexes are more difficult
      - 66% to 85% of non-SG

COST EFFECTIVENESS

- Costs per novel structure
  - 30% sequence ID
    - Non-SG: $532,000 to $1.9 million
    - PSI: $364,000
  - New Pfam family
    - Non-SG: $1.5 to $5.5 million
    - PSI: $1.0 million
  - SCOP superfamily or fold
    - Non-SG: $2.0 to $7.3 million
    - PSI: $2.2 million
- Since most structural biology labs focus on proteins with similar sequences to those already solved, the cost associated is much higher for new structures
  - Due to focus on function

COMPARISONS WITH LEADING STRUCUTRAL BIOLOGISTS

- Large Structural Biology labs
  - Good at solving large, challenging complexes
  - Robert Huber: proteasome, DNA primase, light harvesting complexes
  - So Iwata: photosystem II complex
  - Tom Steitz: protein-nucleic acid complexes
    - Budget
      - $1.5 vs. $5.7 million for PSI centers
      - Also comparable in cost efficiency at solving novel structures

Non-SG groups (since 2000)					
Non-SG structural biology (total)	17,096 (23,747 chains)	928 (249,171)	2,521	269	478
Steitz group	46 (559 chains)	23 (4190)	31	7	12
Huber group	185 (273 chains)	8 (679)	38	5	10
Iwata group	14 (54 chains)	14 (7960)	20	2	3

CONFOUNDING FACTORS

- Factors not taken into consideration
  - Many SG centers collaborate with structural biology groups
    - Causes some of the cost protein structure determination and materials to be shifted onto structural biologists
  - SG centers included structures in the lists solved before their funding in 2000
  - However, a lot of capital was invested into new technologies that may not have given a return yet
  - SG centers had to put money into additional costs
    - Computation, data reporting and analysis
  - Many structural biology projects benefited from prior work on the proteins they study, which is important for more complicated projects

EVALUATION OF STRUCTURE IMPACT

- Can be crudely determined by number of subsequent citations
  - 104 SG structures published between 2001 and 2002
    - 11 average, 4 median
    - The two most cited (107 and 61) describe the overall work of a SG center
  - Randomly selected 104 non-SG publications
    - 21 average, 11.5 median
    - However, novel structures were cited more often
  - This may have to do with the fact that traditional structural biologists usually biochemically characterize their protein as well

CRITICISM

- Gives little functional information of determined proteins
  - Open vs. closed Bcr-Abl tyrosine kinase
- Very pricey
  - Budget could fund 100-200 traditional structural biology labs
  - Uses public money
- Author bias
  - Chandonia and Brenner are affiliated with BSCG, a PSI lab

IS IT WORTH IT?

- Depends on different factors
  - SG tends to focus on smaller, easier to determine structures
  - Relatively little functional information may be known about a given protein
  - However, it allows for a greater rate of discovery of novel protein families
- Good for overall knowledge, not necessarily for specific problems