J Struct Funct Genomics
DOI 10.1007/s10969-007-9025-4

Structure-based inference of molecular functions of proteins of unknown function from Berkeley Structural Genomics Center

Dong Hae Shin · Jingtong Hou · John-Marc Chandonia · Debanu Das · In-Geol Choi · Rosalind Kim · Sung-Hou Kim

Received: 16 May 2007 / Accepted: 27 July 2007

Abstract Advances in sequence genomics have resulted in an accumulation of a huge number of protein sequences derived from genome sequences. However, the functions of a large portion of them cannot be inferred based on the current methods of sequence homology detection to proteins of known functions. Three-dimensional structure can have an important impact in providing inference of molecular function (physical and chemical function) of a protein of unknown function. Structural genomics centers worldwide have been determining many 3-D structures of the proteins of unknown functions, and possible molecular functions of them have been inferred based on their structures. Combined with bioinformatics and enzymatic assay tools, the successful acceleration of the process of protein structure determination through high throughput pipelines enables the rapid functional annotation of a large fraction of hypothetical proteins. We present a brief summary of the process we used at the Berkeley Structural Genomics Center to infer molecular functions of proteins of unknown function.

Keywords Structural genomics · Molecular function · X-ray crystallography · Berkeley Structural Genomics Center

Present Address:
D. H. Shin
College of Pharmacy, Ewha Womans University, Seoul 120-750, Korea

J. Hou · J.-M. Chandonia · D. Das · R. Kim · S.-H. Kim
Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

J. Hou · R. Kim · S.-H. Kim (✉)
Department of Chemistry, University of California, Berkeley, CA 94720, USA
e-mail: SHKim@cchem.berkeley.edu

J.-M. Chandonia · I.-G. Choi
Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA

Present Address:
I.-G. Choi
Division of Life Sciences and Biotechnology, College of Life Sciences, Korea University, Seoul 136-713, Korea

Introduction

At present there are about 510 complete genome sequences available in the Genomes Online Database, and the database is expanding rapidly [1]. Among all the predicted protein genes, about half have no inferable function. Since the 3-D structure of proteins is more tightly coupled to its molecular function than sequence, structural genomics approach turns out to be one of the most efficient ways to infer molecular function of the increasing number of hypothetical proteins derived from sequence genomics [2].

As a part of the Protein Structure Initiative (PSI; www.nigms.nih.gov/funding/psi.html) the BSGC has focused on obtaining the 3-D structural information of the proteins of two minimal organisms, closely related pathogens *Mycoplasma genitalium* and *M. pneumoniae* (http://www.strgen.org), which have fewer than 500 and 700 genes, respectively. The requisite to achieve this goal involved obtaining 3-D structural information for nearly all proteins, a large portion of which are hypothetical proteins, the proteins with no sequence homologies to those of known function. Now, we have 3-D structural information for near complete structural complement of *M. genitalium*. Thus, we now have a structural genomic view of protein fold usage among these and other minimal microbes [3].

Metrics and impact of BSGC structures

At the beginning of PSI initiative, about 30% of *M. genitalium* "soluble" proteins had no 3-D structural fold information. By the time of the completion of the PSI-I, about 94% of the "soluble" proteins have 3-D structural fold information, thus, achieving the mission of BSGC of obtaining a near complete structural complement of a minimal organism. Several metrics were learned from the exercise:

(1) About 1/2 of proteins that had no sequence similarity to the proteins in PDB turned out to have "new folds" and ~1/2 turned out to be "remote homologues" in which homology could only be identified through structural similarity to a known fold.
(2) About 2/3 of the 3-D structures of "hypothetical" proteins inferred testable molecular (biochemical or biophysical) functions, and some of which have since been confirmed experimentally.
(3) The overall success rate of "single-path" (low-hanging fruit) approach for clone-to-structure was <5%, and for purified protein-to-structure was ~9%.
(4) The overall success rate of "multi-path" (single-path plus "salvage path") approach for clone-to-structure was >16%, and for purified protein-to-structure was ~27%.

The overall impact of BSGC structures to the functional inference is summarized below:

- 66 BSGC structures belong to 51 protein sequence families.
- There are 13,171 total protein sequences in these 51 families with an average of ~260 sequences/family.
- Of these, molecular functions of 12,618 (96%) protein sequences can be inferred based on their 3-D structures.
- 3-D structures did not provide possible functions for 553 (4%) sequences.

Since a large portion of the 3-D structures we have determined are for hypothetical proteins (proteins with no sequence homologies to those with known functions) of the organisms, we present a few examples each for five categories of inferring functions from 3-D structures, where 3-D structural information provided functional inferences, and some of which were experimentally verified. Overall scheme of the functional inference process is shown in Fig. 1.

"Remote homologue" proteins

The majority of the structures belong to this category, where structure-based inference of a molecular function is immediately possible. Inferring molecular function of an uncharacterized protein is straightforward when the new structure resembles one or more protein structures whose functions are known. The new structure is a "remote homologue," a structural homologue of a characterized protein structure despite the remoteness of its sequence similarity. BSGC examples of this category are MJ0882 from *Methanococcus jannaschii* (GI number 1499712) [4], MJ0936 from *M. jannaschii* (GI 1499771) [5], MG027 (GI 3844637) from *M. genitalium* [6], SP_1288 (GI 15675166) from *Streptococcus pyogenes* [7], MPN555 from *M. pneumoniae* (GI 1673958) [8], ScpB in *Chlorobium tepidum* (GI 21646405) [9], AQ_1354 (GI 2983779) from *Aquifex aeolicu* [10], PhoU proteins (GI 4982311) from *Thermotoga maritima* [11], YodA protein from *E. coli* (GI 16129919), TA1145 from *Themoplasma acidophilum* (GI 16082162) [12], R1281 from *Deinococcus radiodurans* (GI 6459028), and TM0651 (GI 4981173) from *T. maritima* [13]. All are the homologues of *M. pneumoniae* and *M. genitalium* proteins. A few examples with different validation processes are described below.

The crystal structure of MJ0882 from *M. jannaschii* (GI number 1499712) revealed that it has the same protein fold as many methyltransferases such as catechol O-methyltransferase from *Rattus norvegicus* [4]. The methyltransferase activity of MJ0882 inferred from the structure was subsequently confirmed by biochemical experiments.

MJ0936 from *M. jannaschii* (GI 1499771) is a hypothetical protein of unknown function with over 50 sequence homologues found in many bacteria and archaea. Since its crystal structure revealed structural homology to nucleases, phosphatases, and nucleotidases, a series of biochemical screens for a catalytic activity was performed and a novel phosphodiesterase activity was detected with an absolute requirement for divalent metal ions, Ni^{2+} and Mn^{2+} [5].

The crystal structure of a hypothetical protein TA1145 from *T. acidophilum* (GI 16082162) revealed that the functional domain has a type II phosphoribosyltransferase fold first found in quinolinic acid phosphoribosyltransferase [12]. Based on the structural information, the complex structures with phosphoribosylpyrophosphate or nicotinate mononucleotide were solved to deduce molecular function of TA1145. The complex structures clearly suggested that TA1145 is nicotinate phosphoribosyltransferase [12].

DR1281 from *D. radiodurans* (GI 6459028; Fig. 2) is a protein of unknown function with over 170 homologues. The crystal structure shows that DR1281 has two domains, a small α domain, and a putative catalytic domain with a phosphohydrolase fold formed by a four-layered structure of two β-sheets flanked by five α-helices on both sides (PDB ID: 1T70). A panel of general enzymatic assays of

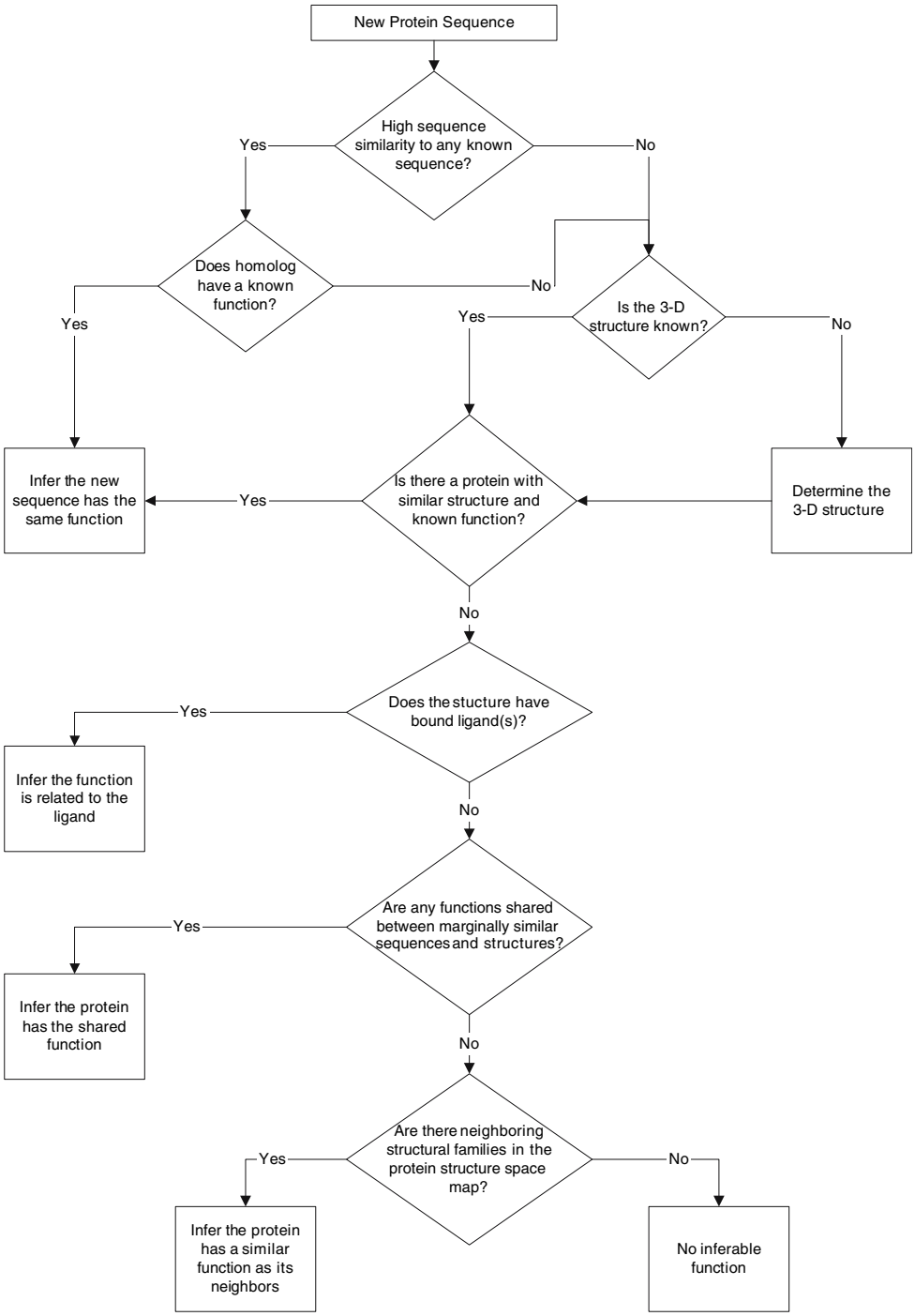

Fig. 1 Decision tree used by BSGC for inferring molecular function

DR1281 revealed a metal-dependent catalytic activity toward model substrates for phosphatases and phosphodiesterases. Subsequent secondary enzymatic screens with natural substrates demonstrated a significant activity toward 2′, 3′ -cAMP. Thus, a combination of structural and enzymatic studies have identified the biochemical function of DR1281 as a novel phosphatase/phosphodiesterase (unpublished results).

A hypothetical protein TM0651 (GI 4981173) from *T. maritima* is a member of the haloacid dehalogenase (HAD) superfamily with unknown function. The crystal structure indicate the presence of the characteristic hydrolase fold of the HAD family and a new tertiary fold having many aromatic residues in the interface of the two domains [13]. Thus, the crystal structure immediately suggests a phosphatase function of TM0651 with a

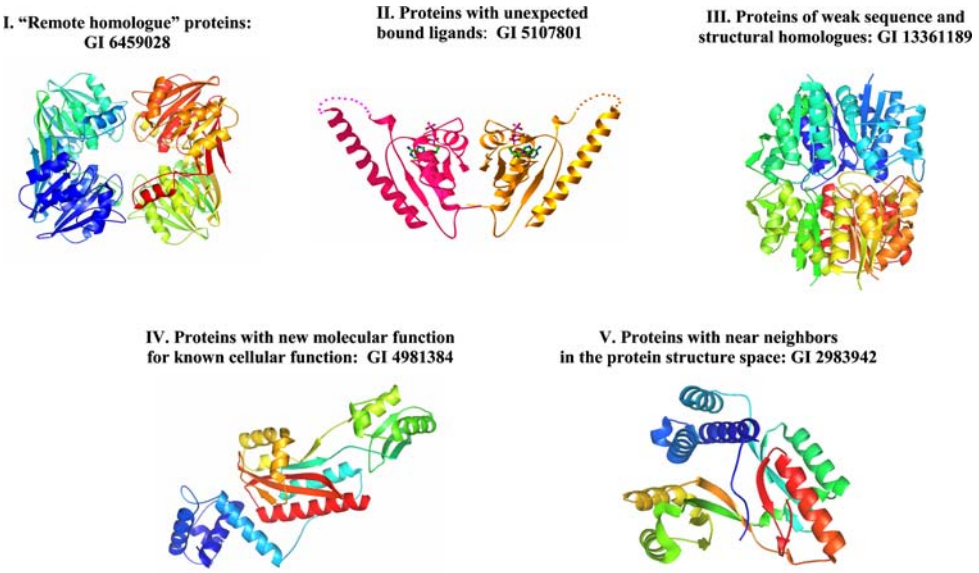

Fig. 2 One sample BSGC structure for each category of structure-based functional inference

carbohydrate molecule as a substrate. The molecular function was proved later by biochemical assays with a TM0651 homologue, YbiV from *E. coli* which hydrolyzes a phosphate from various sugar-like substrates [14].

Proteins with unexpected bound ligands

This is the second most frequent category, where the presence of an unexpected bound ligand the 3-D structure leads the direct inference of the molecular function of a hypothetical protein. BSGC structures belonging to this category are MJ0577 (GI 5107801, Fig. 2) from *M. jannaschii* [15], TM841 (GI 4982034) from *T. maritima* [16], TM1717 from *T. maritima* (GI 4982294) [17], AF2373 (GI 2650718) from *Archaeoglobus fulgidus* [18], all of which are the homologues of *M. pneumoniae* and *M. genitalium* proteins. Two examples are described below.

A hypothetical protein TM841 (GI 4982034) from *T. maritima* was found to belong to a large protein family, DegV in the Pfam database [19] or COG1307 [20] of unknown function. Though its crystal structure does not show a clear resemblance to any know protein structures, the electron density maps revealed a clear density for a bound fatty acid molecule (one palmitate) in a pocket [16]. Thus, the structure indicates that TM841 has the molecular function of fatty acid binding and may play a role in the cellular functions of fatty acid transport or metabolism.

The crystal structure of a hypothetical protein AF2373 (GI 2650718) from *A. fulgidus* with 148 family members revealed a bound NADP near the GGDG motif and a Gly-rich motif [18]. Consequently, ATP, NAD and NADP bound structures were solved to find out the molecular function of AF2373. The complex structures suggested that AF2373 may be a NAD kinase and a possible phosphate transfer mechanism was also proposed based on structures. Subsequent biochemical assays showed that AF2373 had an ATP-NAD kinase activity [18].

Proteins of weak sequence and structural homologues

In the third category, no strong functional inferences can be obtained from either the sequence or the structure of a protein. However, some coincidence can be found among the list of molecular functions of weak sequence homologues and weak structural homologues, providing a clue for the molecular function of a protein. Two examples are described below.

For example, the crystal structure of a hypothetical protein MJ0226 (GI 6980392), a homologue of a *M. pneumoniae* protein, revealed a homodimeric structure with a new fold [21]. Since there were coincident molecular function of nucleotide binding among the proteins with weak sequence and structural homologies, biochemical analysis was performed and found that MJ0226 protein is a novel nucleotide triphosphatase, not for standard nucleotides but for non-standard nucleotide triphosphates such as XTP or ITP in the presence of Mg^{2+} or Mn^{2+} ions. Combined with the observation that MJ0226 is a weak sequence homologue to yeast HAM1 protein, the molecular function of MJ0226 has been experimentally confirmed that the protein removes non-standard nucleotide triphosphates and prevents mutations by protecting DNA from incorporation of modified bases such as Xanthine and Inosine (unpublished results).

YchN from *E. coli* (GI 13361189; Fig. 2), another homologue of a *M. pneumoniae* protein, belongs to the

Cluster of Orthologous Group COG1553 [20]. The crystal structure indicated that this protein has a new fold with no obvious similarity to those of known protein structures. The protein quaternary structure consists of a dimer of trimers with six putative active sites being positioned along the equatorial surface of a hexamer. In the putative active site, two characteristic cysteines found in YchN members have been positioned similar to those of oxidoreductases. Recently, the sulfurtransferase activity of the cysteine residue has been reported in the case of TusBCD [22] which has the same fold and the quaternary structure found in the *E. coli* YchN.

Proteins with new molecular function for known cellular function

In the fourth category, the 3-D structure of a protein of known cellular function can provide a clue to the molecular function of the protein. A cellular function is the result of the combination of many molecular functions, and the 3-D structure of a protein known to participate in the cellular function can identify a particular molecular function among many the protein is responsible. BSGC structures of this category are MJ0285 from *M. jannaschii* (GI 5822407) [23], MPN625 (GI 1673883) from *M. pneumoniae* [24], a *M. jannaschii* protein (GI 15669898) [25], and HrcA from *T. maritima* (GI 4981384; Fig. 2) [26]. Two examples are described below.

MJ0285 from *M. jannaschii* (GI 5822407) is annotated as having the cellular function of a small heat shock protein (sHSP) usually induced under cellular stress. This sHSP has been known to protect other proteins from thermal denaturation. The crystal structure of the protein revealed a homomeric complex of 24 subunits having an overall structure of a multi-windowed hollow sphere with an external diameter of $\sim$120 Å and an internal diameter of $\sim$65 Å with six square windows of $\sim$17 Å across and eight triangular windows of $\sim$30 Å across. This immediately suggested that the partially denatured proteins may bind to the inside or the "widows" of the sphere for renaturation. A series of biochemical experiments such as protease digestion, antibody binding, and electron microscopy were performed using purified single-chain monellin as a substrate. The results strongly implicate that the partially denatured cellular proteins under stress are bound on the outer surface of the sphere, thus preventing them from forming aggregation and resultant inactivation.

A DNA sequence specificity subunit from *M. jannaschii* (GI 15669898) is one of components of type I restriction-modification enzymes. Though it was annotated as a part of type I restriction-modification system, its molecular role was not known. The crystal structure of a specificity subunit revealed that two highly conserved regions in the middle and at the C-terminus form a coiled–coil of long anti-parallel α-helices [25]. The coiled–coil structure of conserved regions acts as a molecular ruler for the separation between two recognized DNA sequences. Furthermore, the relative orientation of the two DNA binding clefts suggests kinking of bound dsDNA and exposing of target adenines from the recognized DNA sequences. Therefore, the crystal structure clearly helps to understand its molecular role in type I restriction-modification enzyme complex.

Proteins with near neighbors in the protein structure space

There are many protein structures whose molecular functions cannot be inferred by any of the methods described above so far. However, as more annotated sequence information and structures become available, their molecular function may become predictable following the processes described for the categories I–IV. For the proteins outside of the categories, one can still attempt to find a list of potential molecular functions to be experimentally tested by identifying the protein structure families that map close to the target protein structure in the protein structure space, "the protein structure universe map" [27]. These close neighbors represent the protein structure families whose structural similarity is not strong enough to be detected by the method such as Dali [28], yet are closer than other structural families. When there is one or more coincidence in molecular functions represented by several members of the close neighbor structural families, those functions are considered as good candidate to experimentally test to find the correct molecular function of the target protein. Examples of inferences of catalytic functions by this approach, only one of which (PDB ID: 1TM9) are experimentally tested (unpublished results), are listed in Table 1.

Novelty of BSGC structures

A study of the impact of structural genomics during the pilot phase of PSI found that by 2005, structural genomics centers contributed nearly half of all the novel structures (i.e., those without sequence similarity to previously solved structures) that were solved in the previous year by all structural biology groups worldwide [29]. In that study, the fraction of BSGC structures that represented the first structure solved in their protein family [30] was 39%, the highest of all 9 PSI pilot centers, and more than double the average of 19% at the other 8 pilot centers. This degree of

Table 1 Inferred enzymatic functions of BSGC structures with unknown function

GI number	PDB ID	E.C. number	Candidate functions of E.C. 1st hierarchy	Candidate functions of E.C. 2nd hierarchy
1674217	1N0E	3.1.	Hydrolase	Acting on ester bonds
		6.3.	Ligase	Forming carbon–nitrogen bonds
		6.4.	Ligase	Forming carbon–carbon bonds
		3.5	Hydrolase	Acting on carbon–nitrogen bonds, other than peptide bonds
2983942	1LFP*	2.7	Transferase	Transferring phosphorous-containing groups
		2.1	Transferase	Transferring one-carbon groups
		1.1	Oxidoreductase	Acting on the CH–OH group of donors
		2.8	Transferase	Transferring sulfur-containing groups
4982022	1S12	3.4	Hydrolase	Acting on peptide bonds (peptide hydrolases).
		2.7.	Transferase	Transferring phosphorous-containing groups
		5.4	Isomerase	Intramolecular transferases (mutases)
		3.5	Hydrolase	Acting on carbon-nitrogen bonds, other than peptide bonds
3844938	1TM9	3.1	Hydrolase	Acting on ester bonds
		5.3	Isomerase	Intramolecular oxidoreductases

Four BSGC structures indexed by their GI (GenInfo Identifier) numbers and PDB (Protein Data Bank) IDs are listed in the leftmost two columns. For each BSGC structure, 2–4 enzymatic functions are inferred from the protein structure space (27), as denoted by their E.C. (Enzyme Commission) numbers

novelty was partly the result of our target de-selection process [31], in which work on a target was usually stopped if the structure of a similar protein was solved elsewhere.

The first structure of a protein family is particularly important, not only because it may reveal a previously unknown molecular function or evolutionary relationship, but also because it allows the fold of other proteins in the family to be inferred, and detailed comparative models to be constructed for the most similar proteins in the family [32]. An efficient strategy for expanding structural coverage of the universe of protein sequences is to choose targets from amongst the largest families with unknown structure [33, 34], and large families are therefore a focus in the production phase of PSI, which commenced in October 2005 [35]. Interestingly, we found that in the pilot phase of PSI, the families that were first structurally characterized by the BSGC averaged twice the size of the families characterized by the other 8 pilot centers, containing an average of 262 members, vs. 130 members for the other centers [29]. This is presumably a result of focusing on a minimal organism, as a large fraction of *M. genitalium*'s genes are thought to be essential for life, and therefore nearly ubiquitous across a wide range of species [36].

Summary

In summary, structural genomics can provide inference for molecular functions of a large number of proteins, which were functionally uncharacterized based on sequence homology methods, as well as hypothetical proteins. In addition, 3-D structures also can validate or suggest alternative molecular functions inferred by sequence homology methods.

Acknowledgements We thank all the component members of BSGC for their efforts towards accomplishing the BSGC objectives. We gratefully acknowledge the supports of the NIH grant GM62412 for most of the structures cited in this article, NIH (R01-GM073109) and the U.S. Department of Energy under contract DE-AC02-05CH11231.

References

1. Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC (2006) Nucleic Acids Res 34:D332–4
2. Kim SH, Shin DH, Choi IG, Schulze-Gahmen U, Chen S, Kim R (2003) J Struct Funct Genomics 4:129–135
3. Chandonia JM, Kim SH (2006) BMC Struct Biol 6:7
4. Huang L, Hung L, Odell M, Yokota H, Kim R, Kim SH (2002) J Struct Funct Genomics 2:121–127
5. Chen S, Yakunin AF, Kuznetsova E, Busso D, Pufan R, Proudfoot M, Kim R, Kim SH (2004) J Biol Chem 279:31854–31862
6. Liu J, Yokota H, Kim R, Kim SH (2004) Proteins 55:1082–1086
7. Oganesyan V, Pufan R, DeGiovanni A, Yokota H, Kim R, Kim SH (2004) Acta Crystallogr D Biol Crystallogr 60:1266–1271
8. Schulze-Gahmen U, Aono S, Chen S, Yokota H, Kim R, Kim SH (2005) Acta Crystallogr D Biol Crystallogr 61:1343–1347
9. Kim JS, Shin DH, Pufan R, Huang C, Yokota H, Kim R, Kim SH (2006) Proteins 62:322–328
10. Oganesyan V, Busso D, Brandsen J, Chen S, Jancarik J, Kim R, Kim SH (2003) Acta Crystallogr D Biol Crystallogr 59:1219–1223
11. Liu J, Lou Y, Yokota H, Adams PD, Kim R, Kim SH (2005) J Biol Chem 280:15960–15966
12. Shin DH, Oganesyan N, Jancarik J, Yokota H, Kim R, Kim SH (2005) J Biol Chem 280:18326–18335

13. Shin DH, Roberts A, Jancarik J, Yokota H, Kim R, Wemmer DE, Kim SH (2003) Protein Sci 12:1464–1472
14. Roberts A, Lee SY, McCullagh E, Silversmith RE, Wemmer DE (2005) Proteins 58:790–801
15. Zarembinski TI, Hung LW, Mueller-Dieckmann HJ, Kim KK, Yokota H, Kim R, Kim SH (1998) Proc Natl Acad Sci USA 95:15189–15193
16. Schulze-Gahmen U, Pelaschier J, Yokota H, Kim R, Kim SH (2003) Proteins 50:526–530
17. Shin DH, Lou Y, Jancarik J, Yokota H, Kim R, Kim SH (2004) Proc Natl Acad Sci USA 101:13198–13203
18. Liu J, Lou Y, Yokota H, Adams PD, Kim R, Kim SH (2005) J Mol Biol 354:289–303
19. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR (2004) Nucleic Acids Res 32:D138–141
20. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV (2001) Nucleic Acids Res 29:22–28
21. Hwang KY, Chung JH, Kim SH, Han YS, Cho Y (1999) Nat Struct Biol 6:691–696
22. Numata T, Fukai S, Ikeuchi Y, Suzuki T, Nureki O (2006) Structure 14:357–366
23. Kim KK, Kim R, Kim SH (1998) Nature 394:595–599
24. Choi IG, Shin DH, Brandsen J, Jancarik J, Busso D, Yokota H, Kim R, Kim SH (2003) J Struct Funct Genomics 4:31–34
25. Kim JS, DeGiovanni A, Jancarik J, Adams PD, Yokota H, Kim R, Kim SH (2005) Proc Natl Acad Sci USA 102:3248–3253
26. Liu J, Huang C, Shin DH, Yokota H, Jancarik J, Kim JS, Adams PD, Kim R, Kim SH (2005) J Mol Biol 350:987–996
27. Hou J, Sims GE, Zhang C, Kim SH (2003) Proc Natl Acad Sci USA 100:2386–2390
28. Holm L, Sander C (1998) Nucleic Acids Res 26:316–319
29. Chandonia JM, Brenner SE (2006) Science 311:347–51
30. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A (2006) Nucleic Acids Res 34:D247–51
31. Chandonia JM, Kim SH, Brenner SE (2005) Proteins 62:356–70
32. Baker D, Sali A (2001) Science 294:93–6
33. Chandonia JM, Brenner SE (2005) Proteins 58:166–79
34. Chandonia JM, Brenner SE (2005) Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, Shanghai, China, 751–55
35. Service R (2005) Science 307:1554–8
36. Hutchison CA, Peterson SN, Gill SR, Cline RT, White O, Fraser CM, Smith HO, Venter JC (1999) Science 286:2165–9