PROTEINS: Structure, Function, and Bioinformatics Suppl 7:201–213 (2005)

# The Prediction of Protein Function at CASP6

**Simonetta Soro**[1] **and Anna Tramontano**[1,2*]
[1]*Department of Biochemical Sciences, University of Rome, "La Sapienza," Rome, Italy*
[2]*Istituto Pasteur Fondazione Cenci Bolognetti, University of Rome, "La Sapienza," Rome, Italy*

*ABSTRACT* **In the CASP6 experiment, the new "Function Prediction" category was tentatively introduced. Predictors were asked to provide functional information on the CASP targets, many of which were of unknown function. This article describes the setup of the experiment and its results, highlighting what was learned from it, and suggesting modifications to its format for the next rounds. The obvious limitation of such an experiment is that the results cannot be assessed in the standard CASP fashion, as all targets remain of unknown function. Furthermore, we had to face the expected difficulties due to the novelty of the experiment and to the problems connected with function definition. Nevertheless, and even with a limited number of participating groups, we believe that the results of the experiment can be useful both for its future and for experimentalists working on the functional assignment of the CASP6 targets. We found that, in a few cases, a consensus functional prediction could be derived for targets of unknown function. However, our analysis suggests that a general description of the method used should be made available together with the predictions so that a higher reliability can be assigned to cases where completely independent methods give the same or similar predictions. Proteins 2005;Suppl 7:201–213.** © **2005 Wiley-Liss, Inc.**

**Key words: CASP6; function prediction; binding site; GO categories**

## INTRODUCTION

One of the most important tasks of protein bioinformatics is function prediction, as demonstrated by the flourishing of computational and experimental methods developed to this aim.[1,2]

The structure of a protein is one of the ingredients of several, but not all, methods for functional assignment, and the large efforts for improving methods for the prediction of the structure of a protein, witnessed by the growing interest in the CASP experiment, find their justification also in the possibility of exploiting the information that they provide in the quest for the function of newly discovered proteins.

Many hurdles are along this path. On one side, the existence of paralogous relationships implies that a common evolutionary origin does not guarantee common function.[3,4] On the other, the discovery of moonlight proteins,[5,6] able to perform different functions in different conditions or environments, has made the complexity of the problem even more apparent. The task is a very challenging one, but also an extremely important aspect of the life sciences and therefore a field of outstanding interest.

In response to this increasing need, the CASP community set up a new challenge for itself, by adding a function prediction category. The question that the experiment wanted to answer is whether, and in which cases, computational methods are able to provide useful information about the molecular or biological function of an unknown protein.

It should be mentioned up front that this category is intrinsically different from other CASP categories for the very simple reason that, at the end of the experiment, the function of the target proteins is likely to be still unknown, and therefore there is no clear way to evaluate the performances of the different methods. The aim of the experiment is, rather, to provide potentially useful information to experimentalists working on the target proteins. It is hoped that the availability of several predictions on the same target, obtained with different independent methods, can be used by experimentalists as a tool to rank their list of experiments for functional assignment.

This first round of function prediction has been promoted, and should therefore be considered, as an exploratory experiment, aimed at learning what we need to do to face the problem. From this point of view, the initiative was successful because it provided useful suggestions for the setup of the next rounds of the experiment.

We selected to use as targets for this experiment the same proteins used in the structure prediction category. This is not the only possible choice. One could have selected a set of different targets for which functional information is likely to be available in the near future. Our choice was not only due to the objective difficulty of assessing when functional information might appear for any protein, but also to the expectation that the experiment could provide us with the opportunity to analyze to which extent the ability to predict a protein structure correlates with the ability of proposing a function for it.

---

Grant sponsor: Istituto Pasteur, Fondazione Cenci Bolognetti; Grant sponsor: the Bio-Sapiens Network of Excellence (funded by the European Commission FP6 Programme); Grant number: LHSG-CT-203-503265.

*Correspondence to: Anna Tramontano, Department of Biochemical Science, University of Rome, "La Sapienza," P.le Aldo Moro, 5, 00185 Rome, Italy. E-mail: anna.tramontano@uniroma1.it

Received 22 April 2005; Accepted 20 June 2005

Published online 26 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20738

The article was originally published online as an accepted preprint. The "Published Online" date corresponds ot the preprint version.

**TABLE I. Targets for Molecular Function Prediction in CASP6**

| | | | Molecular Function | | Biological process | | Cellular component | |
|---|---|---|---|---|---|---|---|---|
| Target | Description | CAT[a] | Source[b] | Annotation | Source[b] | Annotation | Source[b] | Annotation |
| T0196 | Hypothetical protein, *P. furiosus* | CM | EG; COG | Translation elongation factor EF-alpha (GTPase) Translation, ribosomal structure and biogenesis | TrE | protein biosynthesis | | |
| T0197 | Hypothetical protein, *P. furiosus* | FR/H | TrE | adenylate cyclase activity | TrE | cAMP biosynthesis | | |
| T0198 | Phosphate transport system regulator PhoU, putative, *T. maritima* | FR/A | | | COG | Inorganic ion transport and metabolism | SP | Cytoplasmic |
| T0199 | Heat shock operon repressor HrcA, *T. maritima* | D1:CM D2:FR/H D3:FR/A | SP | Negative regulator of class I heat shock genes (grpE-dnaK-dnaJ and groELS operons) | IP | regulation of transcription, DNA-dependent | | |
| T0200 | Conserved hypothetical protein, D. radiodurans | CM | TrE | Hydrolase | | | | |
| T0201 | Hypothetical protein, *T. maritima* | NF | | | | | | |
| T0202 | Hypothetical protein, *A. fulgidus* | D1: FR/H D2: NF | SP | Catalyzes the phosphorylation of NAD to NADP. | IP | metabolism | SP | Cytoplasmic |
| T0203 | At5g08170, *Arabidopsis* | FR/H | TrE | agmatine deiminase | TrE | polyamine biosynthesis | | |
| T0204 | At5g18200, *Arabidopsis* | CM | TrE | UDP-glucose-hexose-1-phosphate uridylyltransferase; UTP-hexose-1-phosphate uridylyltransferase | TrE | galactose metabolism; | | |
| T0205 | At2g34160, *Arabidopsis* | CM | IP | DNA/RNA-binding protein Alba | | | | |
| T0206 | BclA, *B. anthracis* | FR/H | | | TrE | phosphate transport | TrE | Cytoplasmic |
| T0208 | EFR41, *E. faecalis* | CM | SP | Mannonate dehydratase | IP | glucuronate catabolism | | |
| T0209 | IR47, *H. influenzae* | D1: FR/A D2: NF | | | | | | |
| T0211 | HR1958, Human | CM | | | | | | |
| T0212 | SOR45, *S. oncidensis* | FR/A | | | | | | |
| T0214 | Hypothetical protein, *P. furiosus* | FR/H | Pfam | ProFAR isomerase associated | | | | |
| T0215 | Hypothetical membrane protein, *T. acidophilum* | FR/A | COG | Spermidine synthase | COG; SP | Amino acid transport and metabolism; Spermidine biosynthesis | | |
| T0216 | TM0727, *T. maritima* | NF | IP | Peptidase U62, modulator of DNA gyrase | | | | |
| T0222 | Spore coat polysaccharide biosynthesis protein spsE, *B. subtilis* | D1:CM D2:FR/H | IP | ice binding | SP | carbohydrate biosynthesis; homoiothermy; response to freezing | | |
| T0223 | Putative Nitroreductase, *T. maritima* | D1:CM D2:FR/H | COG | Nitroreductase | COG | Energy production and conversion | | |
| T0224 | TM0979, *T. maritima* | FR/H | COG | involved in oxidation of intracellular sulfur | COG | Inorganic ion transport and metabolism | | |
| T0226 | TTHB84orF1199200, *T. thermophilus* | CM | IP | Beta tubulin, structural molecule activity | IP | microtubule-based movement | IP | microtubule |
| T0227 | TTHB84orF1350600, *T. thermophilus* | FR/H | | | | | | |
| T0228 | Nicotinate phosphoribosyltransferase, *S. cerevisiae* | FR/H | SP | nicotinate phosphoribosyltransferase | SP | chromatin silencing | SP | Cytoplasmic and nuclear |
| T0229 | TM0919, *T. maritima* | CM | COG | redox protein, regulator of disulfide bond formation | TrE | response to stress | | |
| T0230 | TM0487, *T. maritima* | FR/A | COG | metal-sulfur cluster biosynthetic enzyme | | | | |

(*Continued*)

**TABLE I. (Continued)**

| | | | Molecular Function | | Biological process | | Cellular component | |
|---|---|---|---|---|---|---|---|---|
| Target | Description | CAT[a] | Source[b] | Annotation | Source[b] | Annotation | Source[b] | Annotation |
| T0231 | Glia maturation factor gamma, Mouse | CM | SP | actin-binding proteins ADF family. GMF subfamily | EG | protein amino acid phosphorylation | IP; EG | intracellular |
| T0232 | Atu5508, *A. tumefaciens* | CM | IP | Glutathione S-transferase, | | | | |
| T0233 | Anthranilate phosphoribosyltransferase 2, Nostoc sp. pcc 7121 | CM | SP | anthranilate phosphoribosyltransferase | SP | Amino-acid biosynthesis; L-tryptophan | | |
| T0234 | Alr5027, Nostoc sp. pcc 7121 | CM | IP; Pfam | Pyridoxamine 5′-phosphate oxidase-related | Pfam | pyridoxine biosynthesis | | |
| T0235 | Ubiquitin carboxyl-terminal hydrolase 6, *S. cerevisiae* | D1: CM<br>D2: FR/A | SP | ubiquitin-specific protease | SP | protein deubiquitination | SP | proteasome regulatory particle (sensu Eukaryota) |
| T0237 | Apical merozoite antigen 1, *P. vivax* | FR/H | PM | factor H/FHL-1 binding | COG | Signal transduction | | |
| T0238 | Predicted coding region BBA68, *B. burgdorferi* | NF | | | | | | |
| T0239 | Hypothetical protein PAE0736, *P. aerophilum,* str. IM3 | FR/A | | | | | | |
| T0240 | Tonb, *E. coli* | CM/NF | SP | Periplasmic protein TonB | IP | iron ion transport; protein transport | SP | Periplasmic, Anchored to the cytoplasmic membrane |
| T0241 | Alpha-acetolactate decarboxylase, *B. brevis* | D1:NF | SP | Alpha-acetolactate decarboxylase | SP | Acetoin biosynthesis | PM | exoenzyme |
| T0242 | B116, *Sulfolobus* turreted icosahedral virus | NF | | | | | | |
| T0243 | F93, *Sulfolobus* turreted icosahedral virus | FR/H | | | | | | |
| T0244 | Galu, *E. coli* | CM | SP | UTP-glucose-1-phosphate uridylyltransferase | IP; EG | carbohydrate catabolism; response to dessication | EG | capsule (sensu Bacteria) |
| T0246 | 3-isopropylmalate dehydrogenase, *T. maritima* | CM | SP | 3-isopropylmalate dehydrogenase | SP | Amino-acid biosynthesis; L-leucine | SP | Cytoplasmic |
| T0247 | Aminomethyltransferase, *E. coli* | CM | SP | aminomethyltransferase | IP | glycine catabolism; proteolysis and peptidolysis | IP | Cytoplasmic |
| T0248 | Hypothetical protein, *M. pneumoniae* M130 | D1: FR/A<br>D2: NF<br>D3: FR/A | | | | | | |
| T0249 | Conserved hypothetical protein, *C. tepidum* TLS | FR/H | | | | | | |
| T0251 | Conserved hypothetical protein, *S. aureus* subsp. aureus N316 | FR/H | | | | | | |
| T0262 | Hypothetical protein, *A. pernix* | D1: FR/A<br>D2: FR/H | SP | RNA:NAD 2′-phosphotransferase | IP | tRNA splicing | | |
| T0263 | Hypothetical protein, *E. coli* | D1:FR/H | | | | | | |
| T0264 | Probable diphtine synthase APE0931, *A. pernix* | D1: CM | SP | diphtine synthase | SP | Diphthamide biosynthesis | | |
| T0265 | Hypothetical transcriptional regulator, *S. tokodaii* | CM | IP | transcription factor | IP | regulation of transcription, DNA-dependent | IP | intracellular |
| T0266 | Hypothetical protein APE2540, *A. pernix* | CM | | | | | | |
| T0267 | Acetyltransferase, *T. thermophilus* | CM | TrE | N-acetyltransferase | | | | |
| T0268 | mraW protein, *T. thermophilus* | CM | SP | S-adenosyl-methyltransferase | | | | |
| T0269 | Thioredoxin peroxidase, *A. pernix* | CM | SP; COG | thioredoxin peroxidase, Peroxiredoxin | COG | Cellular processes and signaling | | |

(*Continued*)

**TABLE I. (Continued)**

| | | | Molecular Function | | Biological process | | Cellular component | |
|---|---|---|---|---|---|---|---|---|
| Target | Description | CAT[a] | Source[b] | Annotation | Source[b] | Annotation | Source[b] | Annotation |
| T0271 | Hypothetical conserved protein, *T. thermophilus* | CM | | | | | | |
| T0272 | Hypothetical protein, *T. thermophilus* | FR/A | | | | | | |
| T0273 | Hypothetical cytosolic protein, *T. thermophilus* | NF | | | | | | |
| T0274 | Probable nitrilotriacetate monooxygenase component B, *T. thermophilus* | CM | TrE | nitrilotriacetate monooxygenase | TrE | electron transport | | |
| T0275 | Hypothetical conserved protein, *T. thermophilus* | CM | | | TrE | response to stress | | |
| T0276 | Conserved hypothetical protein, *T. thermophilus* | CM | IP | 5-formyltetra-hydrofolate cycloligase | TrE | metabolism | | |
| T0277 | Probable nucleotidyltransferase, *T. thermophilus* | CM | TrE | nucleotidyltransferase | | | | |
| T0279 | Uroporphyrinogen-III synthase, *T. thermophilus* | CM | TrE | lyase activity (Uroporphyrinogen-III synthase) | TrE; IP | heme biosynthesis; porphyrin biosynthesis | | |
| T0280 | Putative phosphoribosyl transferase, *T. thermophilus* | D1: CM D2:FR/A | TrE | phosphoribosyl transferase | TrE | nucleoside metabolism | | |
| T0281 | Hypothetical protein, *T. thermophilus* | D1: FR/A | COG | periplasmic or secreted lipoprotein | | | EG | periplasmic or secreted lipoprotein |
| T0282 | Formiminoglutamase, *Vibrio* cholerae O1 biovar cltor str. | CM | SP | formiminoglutamase | SP; IP | Histidine degradation; arginine catabolism | | |

[a]CM: Comparative Modeling; FR/A: fold recognition/analogous; FR/H: fold recognition/homologous; NF: new fold. When the target protein is formed by more than one domain, and two or more domains belong to different prediction categories, they are listed separately.
[b]SP = SwissProt, IP = InterPro, TrE = TrEMBL, EG = Entrez Gene, PM = PubMed.

## RESULTS

### Setup of the Experiment

The first step in the experiment was for the assessors to collect the available information about the target proteins to set the "background" of known facts. Predictors were expected to either provide more detailed information for targets for which some functional information was known, or to provide any type of information about targets of completely unknown function.

The results of this analysis was posted on the Internet at the address http://cassandra.bio.uniroma1.it/Casp6.

We had to face the nontrivial problem of defining function in a general sense, and one of the aims of the experiment was indeed to verify whether this was possible in a sensible way.

We allowed the predictors to provide, for each target, the following information:

1. GO category Molecular Function;[7]
2. GO category Biological process;[7]
3. GO category Cellular components;[7]
4. Binding;
5. Binding site;
6. Residue role;
7. Posttranslational modifications.

Predictors could add a free text comment at the end of their prediction file.

Items 1 to 3 could contain the boundary of the regions of the protein for which the prediction was being submitted. The "Binding" keyword had to be followed by a numerical code to identify whether the protein was predicted to bind DNA, RNA, another protein, or a small molecule. In the latter two cases the name of the molecule could also be added. The "Binding site" line could contain residue numbers or ranges of residue numbers. The "Residue role" could be submitted in free text. Finally, posttranslational (PT) modification types could be submitted using a predefined numerical code (http://predictioncenter.org).

### Function Prediction in Numbers

Twenty-six groups participated in the experiment, and each group could submit up to five (ranked) predictions for each target. This added up to a total of 1235 predictions.

Two hundred forty-five of them referred to targets that were, for different reasons, not assessed in the structure prediction categories in CASP6 and were also left out from this category. The remaining 990 predictions, 867 of which were designated as first models, were analyzed in the function prediction category. Most of the submissions included a Molecular Function (831), a Biological Process (657), and/or a Cellular Component (591) prediction. Fewer of the submissions predicted a binding site (337) or a residue role (199) or a PT modification (114).

## GO Molecular Function, Process, and Component Predictions

For 21 out of the 63 targets, no molecular function annotation was present in any of the publicly available databases. If we assume that the set of CASP targets is a representative sample of the proteins whose structure is solved experimentally, this implies that function is unknown for about one third of newly determined protein structures.

The remaining targets (42) were annotated at different levels of detail in publicly available databases; about 40% had a detailed annotation in SwissProt,[8,9] the remaining 60% had some annotation in TrEMBL,[10] InterPro,[11] ENTREZ,[12] Entrez Gene,[13] or Pfam,[14] or belong to a COG.[15–17]

The distribution of these targets in terms of their evolutionarily relationship with known proteins is shown in Table I. The percentage of targets for which no information is available is only 14% for domains of Comparative Model (CM) targets, but raises to more than 50% for Fold Recognition (FR) and New Fold (NF) targets.

As far as the *biological process* is concerned, the number of cases for which no functional annotation could be found in a database was 27 and only nine of the remaining (25%) were annotated in SwissProt (Table I). This highlights the fact that, although the experimental or predicted structure of a protein or its evolutionary relationships can be used in some cases for detecting its molecular function, the task of attributing a cellular role to the protein is much harder. Also in this case, the information was available more often for comparative modeling targets (76% of the times) and less for the other categories (53% for FR and 30% for NF).

In the following, we will concentrate on targets of completely unknown function. Data on the other targets are available on the CASP Web site.

We first excluded all predictions reporting nonexisting or incorrect GO numbers. These were about 10% of the predictions, and it is unclear to us whether this is due to trivial errors in computer programs, or to an intrinsic difficulty in retrieving the data. In several other cases, the prediction was "UNKNOWN," which of course, we did not consider a prediction, albeit correct!

For each target, we manually inspected the predictions to verify whether different submitted GO numbers had a common parent. The problem we encountered with this strategy is the relatively limited depth of the GO graph, so that very often the annotation of the common node between two predictions would turn out to be not very informative (e.g., "binding" or "catalysis"). In some cases, though, we could detect a consensus among the predictions and, for at least five targets, some conclusions could be derived. Of course, numbers are small here, but this is a proof of principle that, provided the number of predictions is sufficiently high, some interesting hypotheses can be derived from comparing different predictions. Here we are assuming that different predictions are based on different principles, although we cannot be certain that this is the case. As we will discuss later, because of this, we believe that it is essential that some information about the technique used is provided to the assessors in this category to provide the experimentalists with more robust conclusions.

The *cellular component* is very rarely annotated in databases (only in 12 cases, 6 of which in SwissProt) as shown in Table I. Because of this low coverage, we could not derive any conclusion from these predictions.

The results of the analysis of the GO number prediction for targets for which no annotation was found in any database is reported in Table II. As we mentioned, the correct answer is not available; however, the table might be of use to experimentalists, especially in the five cases (Table II) where multiple independent groups suggest the same functional hypothesis, and we hope it will be a useful guide for designing experiments.

As an example, no molecular function annotation is present in any sequence database for target T0226, the Hypothetical protein TTC0981 from *Thermus thermophilus*. We collected 13 predictions for this target; two were invalid (incorrect GO numbers or GO number corresponding to UNKNOWN function), two were rather general (PROTEIN BINDING and SUGAR BINDING). The other predictions were "glucose-6-phosphate isomerase," "glutamine-fructose-6-phosphate transaminase," "inorganic anion exchanger," "tRNA (adenine-N1-)-methyltransferase," and "ketoreductase." Therefore, the consensus seems to be that this protein is an enzyme (six out of seven predictions), possibly with a transferase activity, (the common GO parent for three of the six predictions).

## Binding Site, Residue Role, and Posttranslational Modifications

The submitted predictions for the location of the binding site and residue role in target proteins of unknown function were analyzed to verify whether the subsequently determined protein structure could support the suggested hypothesis, and this is clearly more informative and interesting for targets belonging to the New Fold category.

For target T0201, group P0589 predicted the existence of a disulfide bonds between C16 and C62 which is topologically impossible, while group P0070 suggest that residues 15–17, 23, and 27 participate in catalysis. Interestingly, a different group, P0344, predicted a copper exporting ATPase activity for this target and the location and type of residues proposed by group P070 would be consistent with this hypothesis. The analysis of the crystal structure[18] suggests that the protein might indeed bind copper [Fig. 1(a)], supporting the prediction of group P0344. Groups P0589 and P0070 did not submit a three-dimensional model for this target, while group P0344 did, although its

**TABLE II. Molecular Function and Binding Predictions**

| Target | Group | Mod | GO | GO name | Binding | Binding molecule |
|---|---|---|---|---|---|---|
| T0198 | P0009 | 1 | 16740 | transferase activity | | |
| | P0049 | 1 | 5515 | protein binding | | |
| | P0050 | 1 | 146 | microfilament motor activity | | |
| | P0050 | 2 | 9387 | DNA topoisomerase activity | | |
| | P0050 | 3 | 4337 | geranyltransferase activity | | |
| | P0070 | 1 | 15114 | phosphate transporter activity | Small molecule | phosphate |
| | P0100 | 1 | | | Small molecule | phosphate |
| | P0261 | 1 | 156 | two-component response regulator activity | | |
| | P0589 | 1 | 3700 | transcription factor activity | Protein | probably [pstS, pstC, pstA or pstB] |
| | P0607 | 1 | 15321 | sodium-dependent phosphate transporter activity | | |
| T0201 | P0003 | 1 | 46870 | cadmium ion binding | | |
| | P0049 | 1 | 5515 | protein binding | | |
| | P0050 | 1 | 5153 | interleukin-8 receptor binding | | |
| | P0050 | 2 | 4618 | phosphoglycerate kinase activity | | |
| | P0050 | 3 | 5153 | interleukin-8 receptor binding | | |
| | P0070 | 1 | | | Small molecule | |
| | P0100 | 1 | | | Small molecule | |
| | P0261 | 1 | | | Small molecule | UDP (Uridine-5′-diphosphate) |
| | P0272 | 1 | 3735 | structural constituent of ribosome | | |
| | P0344 | 1 | 4008 | copper-exporting ATPase activity | | |
| T0206 | P0049 | 1 | 3714 | transcription corepressor activity | | |
| | P0050 | 1 | 8810 | cellulase activity | | |
| | P0050 | 2 | 4022 | alcohol dehydrogenase activity | | |
| | P0050 | 3 | 8810 | cellulase activity | | |
| | P0100 | 1 | | | Small molecule | calcium |
| T0209 | P0049 | 1 | 5515 | protein binding | | |
| | P0070 | 1 | 3700 | transcription factor activity | DNA | DNA |
| | P0100 | 1 | | | DNA | DNA |
| | P0261 | 1 | 155 | two-component sensor molecule activity | | |
| | P0631 | 1 | 18997 | electron transfer carrier | | heme |
| T0211 | P0003 | 1 | 4308 | exo-alpha-sialidase activity | Small molecule | |
| | P0009 | 1 | 5515 | protein binding | | |
| | P0049 | 1 | 4872 | receptor activity | | |
| | P0070 | 1 | 3750 | cell cycle regulator | DNA/RNA | Peptide substrate |
| | P0096 | 1 | 5534 | galactose binding | Small molecule | Galactose |
| | P0100 | 1 | | | D/R | Peptide substrate |
| | P0261 | 1 | | | Small molecule | Glycerol |
| | P0479 | 1 | 5515 | protein binding | | |
| | P0589 | 1 | 5545 | phosphatidylinositol binding | Small molecule | possibly phosphatidylinositol |
| T0212 | P0003 | 1 | 5515 | protein binding | Protein | |
| | P0049 | 1 | 5515 | protein binding | | |
| | P0070 | 1 | 46872 | metal ion binding | Small molecule | metal |
| | P0096 | 1 | 5515 | protein binding | | |
| | P0100 | 1 | | | Small molecule | metal |
| | P0261 | 1 | 179 | rRNA (adenine-N6,N6-)-dimethyltransferase activity | Protein | ATP-binding subunit |
| | P0589 | 1 | 4840 | ubiquitin conjugating enzyme activity | Protein | ubiquitin |
| | P0607 | 1 | 8471 | laccase activity | | |
| T0226 | P0003 | 1 | 5529 | sugar binding | | |
| | P0009 | 1 | 4347 | glucose-6-phosphate isomerase activity | Small molecule | |
| | P0049 | 1 | 5515 | protein binding | | |
| | P0050 | 1 | 16429 | tRNA (adenine-N1-)-methyltransferase activity | | |
| | P0050 | 2 | 5452 | inorganic anion exchanger activity | | |
| | P0050 | 3 | 16429 | tRNA (adenine-N1-)-methyltransferase activity | | |
| | P0070 | 1 | 4347 | glucose-6-phosphate isomerase activity | Small molecule | D-glucose-6-phosphate |
| | P0100 | 1 | | | Small molecule | D-glucose-6-phosphate |
| | P0237 | 1 | 5529 | sugar binding | | |
| | P0261 | 1 | 5529 | sugar binding | | |
| | P0319 | 1 | | | Small molecule | |
| | P0344 | 1 | 4360 | glutamine-fructose-6-phosphate transaminase (isomerizing) activity | | |
| | P0589 | 1 | 45703 | ketoreductase activity | Small molecule | |

(*Continued*)

**TABLE II. (Continued)**

| Target | Group | Mod | GO | GO name | Binding | Binding molecule |
|---|---|---|---|---|---|---|
| Cons | | | | Transferase activity/Isomerase activity | | |
| T0227 | P0003 | 1 | 3677 | DNA binding | | |
| | P0003 | 2 | 156 | two-component response regulator activity | | |
| | P0049 | 1 | 5215 | transporter activity | | |
| | P0050 | 1 | 16429 | tRNA (adenine-N1-)-methyltransferase activity | | |
| | P0050 | 2 | 5452 | inorganic anion exchanger activity | | |
| | P0050 | 3 | 16429 | tRNA (adenine-N1-)-methyltransferase activity | | |
| | P0070 | 1 | 3677 | DNA binding | DNA/RNA | |
| | P0261 | 1 | | | Small molecule | AMP/ATP/APC |
| | P0319 | 1 | | | Small molecule | |
| | P0344 | 1 | 3948 | N4-(beta-N-acetylglucosaminyl)-L-asparaginase activity | | |
| T0237 | P0049 | 1 | 5515 | protein binding | | |
| | P0050 | 1 | 42054 | histone methyltransferase activity | | |
| | P0050 | 2 | 46974 | histone lysine N-methyltransferase activity (H3-K9 specific) | | |
| | P0050 | 3 | 16428 | tRNA (cytosine-5-)-methyltransferase activity | | |
| | P0100 | 1 | | | Protein | erythrocytes membrane proteins |
| | P0261 | 1 | 3793 | defense/immunity protein activity | Small molecule | PLP (PYRIDOXAL-5′-PHOSPHATE) |
| | P0589 | 1 | 3793 | defense/immunity protein activity | | |
| T0238 | P0009 | 1 | | | Protein | |
| | P0049 | 1 | 4672 | protein kinase activity | | |
| | P0050 | 1 | 217 | DNA secondary structure binding | | |
| | P0050 | 2 | 9387 | DNA toposiomerase activity | | |
| | P0050 | 3 | 217 | DNA secondary structure binding | | |
| | P0261 | 1 | 5524 | ATP binding | Protein | Human complement regulators ‘rFHL-1’ and factor-H′ |
| | P0272 | 1 | 5201 | extracellular matrix structural constituent | | |
| | P0344 | 1 | 4519 | endonuclease activity | | |
| | P0589 | 1 | 42277 | peptide binding | DNA | maybe |
| T0239 | P0003 | 1 | 42123 | Glucanosyltransferase activity | | |
| | P0049 | 1 | 5198 | structural molecule activity | | |
| | P0050 | 1 | 4909 | interleukin-1, Type 1, activating receptor activity | | |
| | P0050 | 2 | 4751 | ribose-5-phosphate isomerase activity | | |
| | P0050 | 3 | 4909 | interleukin-1, Type 1, activating receptor activity | | |
| | P0344 | 1 | 4175 | endopeptidase activity | | |
| T0242 | P0049 | 1 | 5515 | protein binding | | |
| | P0050 | 1 | 19948 | SUMO activating enzyme activity | | |
| | P0050 | 2 | 9387 | DNA toposiomerase activity | | |
| | P0050 | 3 | 9041 | uridylate kinase activity | | |
| | P0237 | 1 | 166 | nucleotide binding | | |
| | P0261 | 1 | | | Small molecule | MAGNESIUM ION/MANGANESE (II) ION |
| | P0589 | 1 | | | Protein | |
| T0243 | P0049 | 1 | 4871 | signal transducer activity | | |
| | P0050 | 1 | 30911 | TPR domain binding | | |
| | P0050 | 2 | 3980 | UDP-glucose:glycoprotein glucosyltransferase activity | | |
| | P0050 | 3 | 4581 | dolichyl-phosphate beta-glucosyltransferase activity | | |
| | P0070 | 1 | 3700 | transcription factor activity | DNA | DNA binding |
| | P0100 | 1 | | | DNA | DNA binding |
| | P0237 | 1 | 3677 | DNA binding | | |
| | P0261 | 1 | | | Small molecule | FUSICOCCIN |
| | P0344 | 1 | 3677 | DNA binding | | |
| | P0589 | 1 | 46789 | host cell surface receptor binding | Protein | |
| | P0607 | 1 | 4087 | carbamoyl-phosphate synthase (ammonia) activity | | |
| Cons | | | | DNA binding/Transferase activity | | |
| T0248 | P0003 | 1 | 8658 | penicillin binding | | |
| | P0049 | 1 | 5515 | protein binding | | |
| | P0050 | 1 | 4364 | glutathione transferase activity | | |
| | P0050 | 2 | 49 | tRNA binding | | |

(*Continued*)

**TABLE II. (Continued)**

| Target | Group | Mod | GO | GO name | Binding | Binding molecule |
|---|---|---|---|---|---|---|
| | P0050 | 3 | 4231 | insulysin activity | | |
| | P0070 | 1 | | | DNA/RNA | |
| | P0100 | 1 | | | DNA/RNA | |
| | P0237 | 1 | 5319 | lipid transporter activity | | |
| | P0344 | 1 | 5478 | intracellular transporter activity | | |
| | P0589 | 1 | 15616 | DNA translocase activity | DNA | |
| T0249 | P0003 | 1 | 3677 | DNA binding | | |
| | P0049 | 1 | 5515 | protein binding | | |
| | P0050 | 1 | 48365 | Rac GTPase binding | | |
| | P0050 | 2 | 30159 | receptor signaling complex scaffold activity | | |
| | P0050 | 3 | 42768 | ecdysteroid 2-hydroxylase activity | | |
| | P0070 | 1 | 3700 | transcription factor activity | | DNA |
| | P0096 | 1 | 30528 | transcription regulator activity | | |
| | P0100 | 1 | 3700 | transcription factor activity | | DNA |
| | P0237 | 1 | 3677 | DNA binding | | DNA |
| | P0261 | 1 | 8289 | lipid binding | | |
| | P0344 | 1 | 3677 | DNA binding | | |
| | P0479 | 1 | 5515 | protein binding | | |
| | P0589 | 1 | 3700 | transcription factor activity | DNA | |
| Cons | | | | DNA binding/Transcription factor activity | | |
| T0251 | P0049 | 1 | 5200 | structural constituent of cytoskeleton | | |
| | P0050 | 1 | 19948 | SUMO activating enzyme activity | | |
| | P0050 | 2 | 3980 | UDP-glucose:glycoprotein glucosyltransferase activity | | |
| | P0050 | 3 | 19948 | SUMO activating enzyme activity | | |
| | P0070 | 1 | 8794 | arsenate reductase (glutaredoxin) activity | Small molecule | arsenate |
| | P0100 | 1 | | | Small molecule | arsenate |
| | P0237 | 1 | 5489 | electron transporter activity | | |
| | P0319 | 1 | | | Protein | |
| | P0344 | 1 | 5489 | electron transporter activity | | |
| T0263 | P0049 | 1 | 3754 | chaperone activity | | |
| | P0050 | 1 | 5098 | Ran GTPase activator activity | | |
| | P0050 | 2 | 5098 | Ran GTPase activator activity | | |
| | P0050 | 3 | 3969 | RNA editase activity | | |
| | P0070 | 1 | 4497 | monooxygenase activity | Small molecule | small ligand |
| | P0100 | 1 | | | Small molecule | small ligand |
| | P0237 | 1 | 16491 | oxidoreductase activity | | |
| | P0261 | 1 | | | Small molecule | MTX (METHODTREXATE) |
| | P0319 | 1 | | | Small molecule | |
| | P0344 | 1 | 3676 | nucleic acid binding | | |
| | P0589 | 1 | 19845 | exotoxin activity | | |
| Cons | | | | Binding/Oxidoreductase activity/Enzyme regulator activity | | |
| T0266 | P0003 | 1 | 3723 | RNA binding | | |
| | P0049 | 1 | 3754 | chaperone activity | | |
| | P0050 | 1 | 4587 | ornithine-oxo-acid transaminase activity | | |
| | P0050 | 2 | 4477 | methenyltetrahydrofolate cyclohydrolase activity | | |
| | P0050 | 3 | 4047 | aminomethyltransferase activity | | |
| | P0070 | 1 | | | DNA/RNA | tRNA |
| | P0096 | 1 | 4827 | proline-tRNA ligase activity | | |
| | P0100 | 1 | | | DNA/RNA | tRNA |
| | P0237 | 1 | 166 | nucleotide binding | | |
| | P0261 | 1 | 5215 | transporter activity | | |
| | P0589 | 1 | | | DNA | |
| | P0726 | 1 | 4812 | tRNA ligase activity | | |
| Cons | | | | Binding/Transport activity/Transferase activity | | |
| T0271 | P0003 | 1 | 3677 | DNA binding | | |
| | P0049 | 1 | 5524 | ATP binding | | |
| | P0050 | 1 | 4587 | ornithine-oxo-acid transaminase activity | | |
| | P0050 | 2 | 8750 | NAD(P)+ transhydrogenase (AB-specific) activity | | |
| | P0050 | 3 | 4587 | ornithine-oxo-acid transaminase activity | | |
| | P0100 | 1 | | | Protein | protein partner |
| | P0261 | 1 | 4601 | peroxidase activity | Small Molecule | RP5 (RIBOSE-5-PHOSPHATE, PYRANOSE FORM) |

(*Continued*)

**TABLE II. (Continued)**

| Target | Group | Mod | GO | GO name | Binding | Binding molecule |
|---|---|---|---|---|---|---|
| | P0283 | 1 | 3676 | nucleic acid binding | | |
| | P0344 | 1 | 8781 | N-acylneuraminate cytidylyltransferase activity | | |
| | P0589 | 1 | 15619 | thiamin pyrophosphate-transporting ATPase activity | | |
| T0272 | P0003 | 1 | 3677 | DNA binding | | |
| | P0049 | 1 | 5524 | ATP binding | | |
| | P0050 | 1 | 16429 | tRNA (adenine-N1-)-methyltransferase activity | | |
| | P0050 | 2 | 30156 | benzodiazepine receptor binding | | |
| | P0050 | 3 | 16429 | tRNA (adenine-N1-)-methyltransferase activity | | |
| | P0070 | 1 | 3678 | DNA helicase activity | DNA/RNA | DNA |
| | P0100 | 1 | | | DNA/RNA | DNA |
| | P0283 | 1 | 5137 | | | |
| | P0344 | 1 | 3779 | actin binding | | |
| | P0589 | 1 | 4034 | aminomethyltransferase activity | Small molecule | |
| T0273 | P0049 | 1 | 5515 | protein binding | | |
| | P0050 | 1 | 30156 | benzodiazepine receptor binding | | |
| | P0050 | 2 | 30156 | benzodiazepine receptor binding | | |
| | P0050 | 3 | 17050 | D-erythro-sphingosine kinase activity | | |
| | P0100 | 1 | | | DNA/RNA | DNA |
| | P0237 | 1 | 3677 | DNA binding | | |
| | P0261 | 1 | 15036 | disulfide oxidoreductase activity | | |
| | P0283 | 1 | 3746 | translation elongation factor activity | | |
| | P0344 | 1 | 3723 | RNA binding | | |
| | P0589 | 1 | 4175 | endopeptidase activity | Small molecule | |
| | P0726 | 1 | 4812 | tRNA ligase activity | | |
| T0275 | P0003 | 1 | 5524 | ATP binding | | |
| | P0049 | 1 | 5198 | structural molecule activity | | |
| | P0050 | 1 | 50333 | thiamin-triphosphatase activity | | |
| | P0050 | 2 | 50333 | thiamin-triphosphatase activity | | |
| | P0050 | 3 | 16429 | tRNA (adenine-N1-)-methyltransferase activity | | |
| | P0070 | 1 | | | Small molecule | unknown ligands |
| | P0100 | 1 | | | Small molecule | unknown ligands |
| | P0237 | 1 | 5524 | ATP binding | | |
| | P0261 | 1 | 5524 | ATP binding | Small molecule | ATP |
| | P0272 | 1 | | | Small molecule | |
| | P0319 | 1 | | | Small molecule | |
| | P0344 | 1 | 5524 | ATP binding | | |
| | P0589 | 1 | 16491 | oxidoreductase activity | Small molecule | |
| | P0726 | 1 | 155 | two-component sensor molecule activity | | |

When more predictions share a common GO parent, the latter is reported in the line labeled "Cons."

model only had a GDT-TS (the distance based measure used in CASP to evaluate the overall quality of a three-dimensional model[19]) of about 36. Interestingly, in this latter model the orientation of the copper binding Cys23 and His16 side chains is not compatible with the function proposed by the same group. It is possible that their prediction could have been improved had they taken the function prediction into account.

For target T0209, group P0070 suggests a DNA binding role for a long list of residues, some of which (residues 193–197 and 75–85) are missing from the X-ray structure. The prediction of group P0631 also includes two residues that are not visible in the X-ray structure[20] (K70 and K156), plus residues K17 and K19, which are indeed exposed to solvent. These residues could be compatible with the nucleic acid binding function of the protein, suggested by group P070 [Fig. 1(b)].

Target T0216 is part of a COG including Zn-dependent proteases. Group P0070 suggests that residues 271, 282, 287, 288, and 289 are involved in catalysis, but they are buried in the protein structure. (Group P0100 also submitted three-dimensional models for this target. The region 282–289 is missing in all of them, while residue 271 is present, and predicted to be exposed, in some of the models.) Group P0580 predicted a functional role for His 363 and some of its neighbors. The three-dimensional structure of this target[21] shows that it is a homodimer. His 363 faces the equivalent residue from the other chain and could indeed be involved in metal binding [Fig. 1(c)].

Target T0238 is the BbCRASP-1 from *Borrelia burgdorferii* and is not annotated in any of the sequence databases. However, an article by Kraiczy et al.[22] found by searching PubMed with the protein name, states that this class of

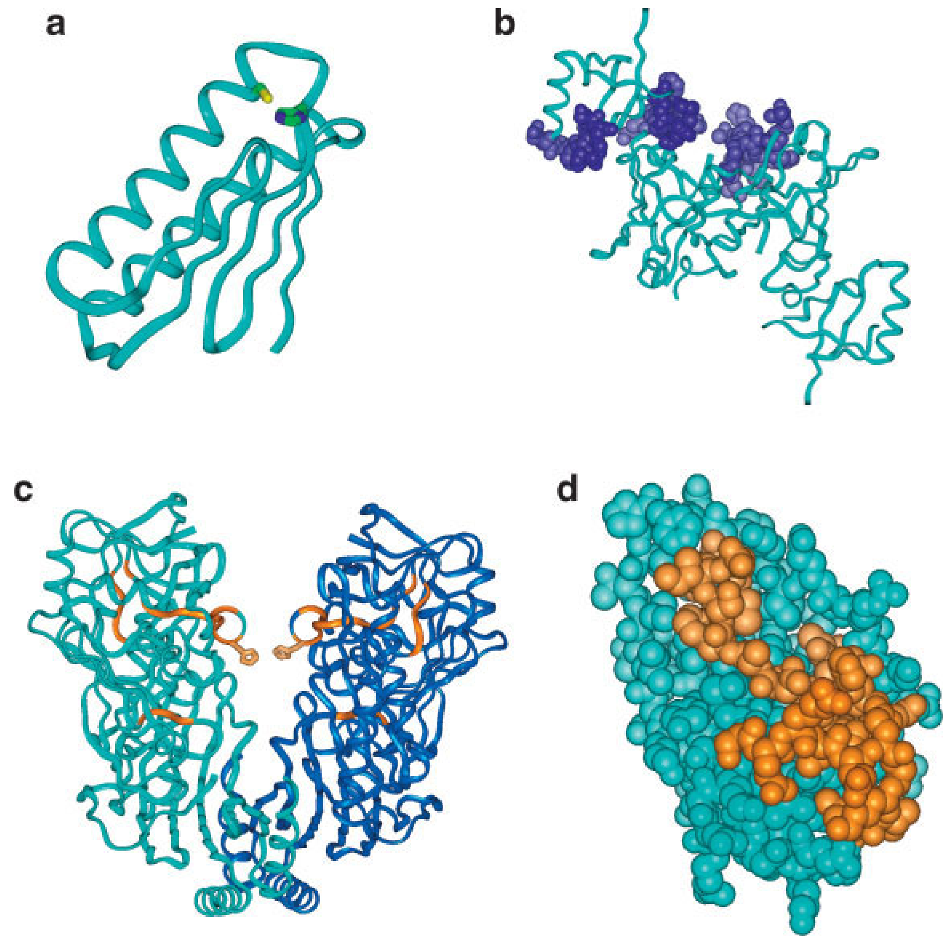


Fig. 1. (**a**) Ribbon representation of the structure of target T0201 (pdb code 1S12), H16, and C23, predicted to bind metal by group P0344, are highlighted. (**b**) Ribbon representation of the structure of target T0209 (pdb code 1XQB). The binding site residues predicted by group P0070 are shown in CPK representation. (**c**) Ribbon representation of the dimer of target T0216 (pdb code 1VI4). Residues (226–229, 329–331, 355, 357–363, and 365–367) predicted by group P0580 are in orange, H363 in sticks; (**d**) CPK representation of the structure ot target T0242 (pdb code 2BLK). Binding site residues predicted by group P0589 are shown in orange.

proteins binds components of the complement regulatory system, factor H, and/or factor H-like protein 1. Only group P0261 submitted the prediction "Human complement regulators 'rFHL-1' and factor-H." Other predictions for the same protein did not take into account these experimental data.

We could find no supporting structural evidence for the binding site predictions submitted for target T0239. On the other hand, there are three disulfide bridges in this protein that have been correctly predicted by group 100.

For target T0241, group P0100 predicted that H174, H176, H187, and N190 are part of this protein binding site, and group P0176 predicted that H174, H176, and H187 are possible active site residues. Other groups (P0237 and P0589) also predicted a binding function for residues H176 and H187. These latter two residues, together with H174, do indeed bind metal (Zn), as revealed by the experimental structure.

The binding site predictions for target T0242 include the suggestion, put forward by group P0261, that the binding site is formed by two pairs of residues (I24, D25 and H51, D52); however, they lie on opposite faces of the protein. Similar is the case for the prediction of group P0580. On the contrary, the regions suggested to be involved in binding by group P0589 (residues 8–15, 47–55, 87–96) do cluster on the surface of the protein[23] [Fig.1(d)]. Also, this protein contains a disulfide bridge that has not been predicted by any of the participating groups.

For target T0248, groups P0580, P0070, and P0100 predicted, as binding site, residues belonging to three different domains. These residues are located on the surface surrounding a negatively charged patch or in a cleft present on one side of the protein. Groups P0070, P0100, and P0237 also predicted a functional role for K210. This residue is among the conserved residues, and is located at the end of a alpha-helix, on the surface surrounding the cleft present on one site of the protein.[24] Group P0050 (model 3) predicted "Insulysin activity" (i.e., insulin protease activity) as Molecular Function for this target. In the article describing the T0248 crystal structure,[24] the authors performed an array of functional assays for the protein and could not detect any protease activity.

**TABLE III. Residue PT Modification Predictions for the Targets for Which No Functional Information Is Available**

| Target | Group | PT modification[a] | Solvent accessibility | | |
|---|---|---|---|---|---|
| | | | Exposed | Buried/partially buried | Undetermined |
| T0198 | P0261 | NG | N38, N124 | | |
| T0198 | P0479 | Ph | Y153, S159, Y176 | S30 | |
| T0201 | P0479 | Ph | Y51 | | |
| T0209 | P0479 | Ph | T94, E178 | S21 | T74, I233 |
| T0211 | P0479 | Ph | S79 | | |
| T0211 | P0589 | Ph | T19, T71, T78, S121 | | |
| T0212 | P0479 | Ph | S22, Y28 | | |
| T0212 | P0589 | Ph | S7, S21 | T87 | S2 |
| T0214 | P0009 | Ph | Y27, S59, S78 | | |
| T0214 | P0479 | Ph | S59, S78 | | |
| T0214 | P0589 | Ph | K19, S59, S78 | | |
| T0226 | P0589 | Ph | T86, T122, S130, T159, S236 | S197 | |
| T0227 | P0479 | Ph | | | V60 |
| T0227 | P0589 | Ph | R29 | | V60 |
| T0237 | P0272 | SS[b] | S52 | | P205 |
| T0237 | P0479 | SS[b] | Q402, I390, K392 | N346 | N407, Y409 |
| T0237 | P0589 | Ph | L156, N254, N358, P387, I439 | S136, P248 | Y6, K307 |
| T0239 | P0589 | Ph | R19(e), T46(e/b) | | |
| T0239 | P0100 | SS[b] | C22, C24, C34, C54, C80, C83 | | |
| T0242 | P0589 | Ph | S87, S97, S111 | | |
| T0248 | P0589 | Ph | S21, S65, T150, K158, T191, T281 | | |
| T0249 | P0589 | Ph | S18, T37, T55, T58, T72, R90, S95, T110 | | S164 |
| T0251 | P0589 | Ph | T18, S38, T42, T54, R63 | T75 | |
| T0263 | P0261 | Ph | S44, T70 | T42 | |
| T0263 | P0479 | Ph | S93 | | |
| T0263 | P0589 | Ph | S44, T70 | T42 | |
| T0266 | P0589 | Ph | T15, T27, S38, S70, S117, S135 | | |
| T0271 | P0589 | Ph | S51, S59 | T26 | |
| T0272 | P0589 | Ph | S70, T118, T152 | T63, T191 | T110 |
| T0273 | P0261 | Ph | S12, R17, T41, S45, S99 | | |
| T0273 | P0589 | Ph | S12, R17, T41, S45, S99 | | |
| T0275 | P0589 | Ph | | S122 | |

Solvent accessibility of the residue side chains has been assessed by visual inspection.
[a]SS = disulfide bond; NG = cotranslational N glycosylation in ER and Golgi apparatus; Ph = phosphorylation.
[b]These disulfide bridges are correctly predicted.

Finally, for target T0273, correctly predicted as an "endonuclease fold" by group P0100,[25] group P0070 suggests a DNA binding role, as does group P0100. However, the former proposes a binding site mainly formed by buried, hydrophobic residues that do not cluster in the protein structure. Group P0237 (who also proposed a DNA binding function for this protein) suggests that the binding site is located in the region 78–86. The sequence of this region is SILSPDAAF, but only residues 78–80 are fully exposed on the surface; therefore, this prediction is likely to be incorrect.

In the PT modification category 101 out of 114 predictions concerned phosphorylation sites. Table III shows a summary of the predictions for targets without any prior functional information. Among these predictions, 80% of the residues indicated as phosphorylated are Ser or Thr. Most of them (76%) are exposed in the crystal structure.

## DISCUSSION

We believe that the function prediction test performed in CASP6, and described here, confirms the importance and feasibility of a function category prediction in CASP. As expected, this first trial round highlighted a number of relevant issues that will be taken in due account in the next rounds of the experiment and that might be useful for other similar initiatives.

We hope that the information gathered during this experiment, reported here and available in the CASP Web site, will reveal to be useful to experimentalists for designing experiments on the targets of unknown function. On our side, we will continue to follow the literature and the database annotations on the CASP6 target proteins to verify whether new information appears that could allow us to evaluate the prediction results in a more traditional way.

In general, there seemed to be not many cases where groups predicting the three-dimensional structure of a protein used this information for also predicting its function and vice versa. This can be due to the limited time available for the experiment, or to the novelty of this aspect of the experiment. We hope that this will change in future experiments, as, in the few cases where structure and function were predicted by the same groups, and when the structure was reasonably well predicted, function predictions seem to be more likely to be correct and therefore useful to the biological community.

At the end of the CASP meeting, there was a discussion among predictors to decide, in light of the results reported here, whether the experiment had to be continued and, if so, in which form. The overall consensus was that the experiment should be repeated, with some adjustments to take into account the outcome of the first round. We report here the results of that discussion, to stimulate suggestions and ideas from a more extended audience.

Molecular function prediction, it was felt, had to be limited to enzymes, or proteins predicted to be enzymes, and collected in terms of EC numbers rather than GO annotations. This would simplify the unambiguous comparison among different predictions making it easier to derive a consensus.

At the same time, the community agreed that, to be useful to the biological community, at least a general description of the method used should be made available to the assessor, to give her or him the possibility of assigning the appropriate weight, in terms of putative reliability, to cases where completely independent methods give the same or similar predictions.

The assessor mentioned, and maintain here, that a rigorous adhesion to a predefined format is essential to allow a proper compilation of the results and to attempt to derive information that could be of use to the experimental community.

We are extremely grateful to the community for participating in this experiment, and for doing so in a very cooperative and understanding way. Almost the same number of groups took part in this experiment as in the structure prediction category of CASP1 and, although we did meet some format problem, this was really marginal compared to the difficulties met by the structure prediction assessors in CASP1! We hope that the same impressive progression of participating group that has been observed for the structure prediction categories will occur in this area. This would be of paramount importance, we think, for speeding up the function discovery process and for assessing whether the parallel use of different function prediction methods, be them based on sequence, structure, or both, is likely to substantially increase the number of proteins for which functional predictions can be made computationally.

## ACKNOWLEDGMENTS

S.S. is the recipient of a "FIRB Bioinformatica per la Genomica e la Proteomica" fellowship. The authors are grateful to Krzystof Fidelis, Alfonso Valencia, Roland Dunbrack, B.K. Lee, Claudia Bertonati, Sharon Goldsmith-Fishman, and Burkhard Rost for useful suggestions, and to Domenico Cozzetto for help in processing the data.

## REFERENCES

1. Jones S, Thornton JM. Searching for functional sites in protein structures. Curr Opin Chem Biol 2004;8:3–7.
2. Gabaldon T, Huynen MA. Prediction of protein function and pathways in the genome era. Cell Mol Life Sci 2004;61:930–944.
3. Devos D, Valencia A. Practical limits of function prediction. Proteins 2000;41:98–107.
4. Rost B. Enzyme function less conserved than anticipated. J Mol Biol 2002;318:595–608.
5. Jeffery CJ. Moonlighting proteins: old proteins learning new tricks. Trends Genet 2003;19:415–417.
6. Jeffery CJ. Multifunctional proteins: examples of gene sharing. Ann Med 2003;35:28–35.
7. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R, Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 2004;32:D258–D261.
8. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 2003;31:3784–3788.
9. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. The Universal Protein Resource (UniProt). Nucleic Acids Res 2005;33:D154–D159.
10. Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31: 365–370.
11. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJ, Silventoinen V, Studholme DJ, Vaughan R, Wu CH. InterPro, progress and status in 2005. Nucleic Acids Res 2005;33:D201–D205.
12. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JU, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2005;33:D39–D45.
13. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 2005;33:D54–D58.
14. Bateman A, Coin L, Durbin RD, Finn R, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C, Eddy SR. The Pfam Protein Families Database. Nucleic Acids Res 2004;32:D138–D141.
15. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 2001;29:22–28.
16. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin

JJ, Natale DA. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 2003;4:41–54.

17. Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin AV, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 2004;5:R7–R34.
18. Shin DH, Lou Y, Jancarik J, Yokota H, Kim R, Kim S-H. Crystal structure of Tm1457. Forthcoming.
19. Zemla A, Venclovas E, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins 1999;S3:22–29.
20. Benach J, Lee I, Forouhar F, Kuzin AP, Keller JP, Itkin A, Xiao R, Acton T, Montelione GT, Hunt JF. X-ray structure of Yaeb from *Haemophilus influenzae*. Northeast Structural Genomics Research Consortium (Nesgc) Target Ir47. Forthcoming.
21. Joint Center for Structural Genomics: crystal structure of Pmba-related protein (Tm0727) from *Thermotoga maritima* at 1.95 A resolution. Forthcoming.
22. Kraiczy P, Hartmann K, Hellwage J, Skerka C, Kirschfink M, Brade V, Zipfel PF, Wallich R, Stevenson B. Immunological characterization of the complement regulator factor H-binding CRASP and Erp proteins of *Borrelia burgdorferi*. Int J Med Microbiol 2004;37:152–157.
23. Larson E, Reiter D, Young M, Lawrence CM. The structure of B116 from sulfolobus turreted icosahedral virus, a ubiquitous crenarchaeal viral protein. Forthcoming.
24. Das D, Oganesyan N, Yokota H, Pufan R, Kim R, Kim SH. Crystal structure of the conserved hypothetical protein Mpn330 (Gi: 1674200) from *Mycoplasma pneumoniae*. Proteins 2004;58:504–508.
25. Vincent JJ, Tai C-H, Sathyanarayana BK, Lee BK. Assessment of CASP6 predictions for new and nearly new fold targets. Proteins 2005;Suppl 7:67–83.