SUPPLEMENTARY EXPERIMENTAL PROCEDURES

Production of Full-length kp60 - Expression vector for the recombinant GST-tagged full-length kp60 of mouse was constructed by standard protocol using PCR, and ligated into *BamHI-SalI* sites of pGEX-6P3 (GE Healthcare Bioscience). Ala-substituted mutants were engineered with QuikChange site-directed mutagenesis kit (Stratagene). The fusion proteins were produced in *E. coli* JM109. Expression was induced with 0.1 mM IPTG, and LB cultures were grown overnight at 20 °C. For pull-down assays, GST-tagged proteins were bound to glutathione-Sepharose 4B (GE Healthcare Bioscience) and washed with the storage buffer (20 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1 mM $MgCl_2$, and 0.1 mM ATP) supplemented with EDTA-free protease inhibitor cocktail (Nacali tesque Inc, Kyoto, Japan) on column. GST-kp60s bound to glutathione-Sepharose were eluted with the elution buffer (50 mM Tris-HCl, 100 mM NaCl, 40 mM reduced glutathione, pH 8.0, and 5% glycerol). The eluents were further used for ATPase assays.

ATPase assays - ATPase activity was measured using an ATP regenerating system (1). The reaction mixture containing 50 mM Tris-HCl, pH 7.5, 50 mM KCl, 2 mM $MgCl_2$, 2 mM phosphoenolpyruvate, 1 mM ATP, 50 μg/ml pyruvate kinase, 50 μg/ml lactate dehydrogenase, and 0.2 mM NADH was used. The reactions were initiated by the addition of GST-kp60s (0.5 μM), and the activities were measured by monitoring the decrease of NADH absorption at 340 nm at room temperature using UV-Vis spectrophotometer, UV mini-1240 (Shimadzu, Tokyo, Japan). The data were normalized for further analysis.

Tubulin Binding Assays - 5 μg of GST-proteins bound to glutathione-Sepharose 4B (20 μl) were incubated with 10 μg of tubulin in the binding buffer (80 mM PIPES, pH 7.0, 1 mM $MgCl_2$ and 1 mM EGTA) for 30 min at 4°C. The beads were washed four times in the wash buffer (4.3 mM Na_2HPO_4, 1.47 mM KH_2PO_4, 137 mM NaCl, 2.7 mM KCl, pH 7.3, and 5% glycerol). The associated proteins were eluted in the elution buffer (50 mM Tris-HCl, 100 mM NaCl, 50 mM reduced glutathione, pH 8.0, and 5% glycerol). The eluted proteins were analyzed by SDS-PAGE and Western blotting.

Western Blotting - Proteins were resolved in SDS-PAGE and blotted onto a PVDF membrane. We detected tubulin using 1/2000 diluted anti-α-tubulin antibody (Siguma-Aldrich) followed by HRP-conjugated anti-mouse IgG secondary antibody (Promega). The proteins were visualized using an ECL-Plus kit (GE Healthcare Bioscience) and detected using LAS-1000 detector (Fuji Film, Tokyo, Japan).

Model building – A molecular model of the complex of kp60-NTD with a tubulin tetramer was constructed based on the complex between spastin-MIT and CHMP1b (PDB: 3eab). The kp60-NTD structure and the tubulin tetramer (3du7) were superimposed onto the corresponding position of spastin-MIT and the C-terminal helix of CHMP1b (174-193), respectively. The best model fully overrided on helices with binding sites was selected considering steric crash and complementary charge interactions between structures. A hexameric ring model of AAA ATPase domains of kp60 was generated by superimposing the Cα atoms of kp60 onto those of the hexameric ring structure of p97 D1 (PDB: 1s3s) using MODELLER (version 9v6) (http://salilab.org/modeller/). Finally, the complex model structure of hexameric full-length kp60 with tubulin oligomer was constructed using MOLMOL (2) by joining the components manually.

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Fig. 1.
Phylogenetic tree of the AAA protein superfamily. Red circle, kp60 subfamily; Blue circle, Vps4 subfamily. The tree data were calculated by ClustalX (3) and the tree was drawn with TreeView (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html).

Supplementary Fig. 2.
Ramachandran plot for the phi-psi values of the final 20 structures of kp60-NTD. This figure was produced using PROCHECK-NMR (4).

Supplementary Fig. 3.
Interactions between kp60-NTD and MTs/tubulin dimer. *A*, schematic diagram of pull-down assay using Microtubule binding protein spin down assay kit, BK029 (Cytoskeleton) to assess interactions between GST-tagged kp60-NTD and MTs/tubulin *in vitro*. GST-tagged kp60-NTD was mixed with a reaction solution after (upper left) and before (upper right) tubulin polymerization reaction, then ultracentrifuged. Tubulin was separated by molecular weight; polymerized MTs were sedimented at the bottom of tubes, non-polymerized tubulin migrated to the top of the solution (lower panel). Reaction solutions were divided into four fractions (from top to bottom) and each fraction was analyzed by SDS-PAGE. GST-tagged kp60-NTD co-sedimented with non-polymerized tubulin. *B*, the reaction solution after polymerization of only tubulin was ultracentrifuged and analyzed as a control. Lanes 1–4 correspond to fractions from top to bottom, indicated in the lower panels of *A*. *C*, pull-down assay for kp60-NTD mixed after tubulin polymerization reaction. kp60-NTD may possibly bind with a tubulin dimer rather than MTs. *D*, pull-down assays of the GST-tagged kp60-NTD mixed before tubulin polymerization reaction. SDS-PAGEs are Coomassie-stained.

Supplementary Fig. 4.
ATPase activity of full-length kp60 and interactions of kp60 with tubulin. *A*, ATPase activities of kp60s (0.5 μM) at 340 nm. Filled diamond (continuous line): wild type, filled box (dotted line): R49A, filled triangle (broken line): K67A. *B*, pull-down assays of tubulin with wild type (WT) of GST-kp60 and Ala mutants *in vitro*. Molecular size is shown in the left. Tubulin was used as the input. Only the buffer and the GST-tag mixed with tubulin as negative controls are shown in lanes 2 and 3. Recombinant proteins used for pull-down are indicated at the top of the gel. Filled and open arrowheads show tubulin and full-length kp60s, respectively. SDS-PAGE was Coomassie-stained (upper panel). Western blotting analysis of tubulin bound to full-length kp60s was visualized by ECL (lower panel).

Supplementary Fig. 5.
Comparison of structures and tubulin binding interfaces with other tubulin binding domains. Tubulin binding interfaces are indicated by black and arrows. *A*, stathmin-like domain bound to the tubulin (white) (PDB: 1sa1); *B*, EB1 CH domain (2qjz); *C*, Msps TOG2 domain (2qk2); *D*, CAP-Gly domain bound to the tubulin peptide (white) (2e4h), and *E*, tubulin-specific chaperone cofactor A (1h7c).

Supplementary Fig. 6.
Model for α–tubulin helix 12 binding with kp60-NTD. An electrostatic surface potential diagram (top), a ribbon diagram (middle), and a sequence conservation diagram (bottom) for kp60-NTD were shown. α–Tubulin helix 12 is shown as a transparent cylinder (yellow).

Supplementary Fig. 7.

Comparison between model for tubulin binding interfaces of kp60-NTD. *A*, model of kp60-NTD bound to α–tubulin at the helix 1/3 interface (see text). Ribbon diagram of the model complex between kp60-NTD and a tubulin tetramer (grey) was constructed based on the complex between spastin-MIT and CHMP1b (PDB: 3eab). α–tubulin helix 12, a putative interface to kp60-NTD, is colored yellow. *B* and *C*, side (top) and top (bottom) views of the ribbon diagram of the complex between kp60-NTD and α–tubulin using the helix 1/3 and helix 2/3 interfaces, respectively. Side chains of key residues for binding tubulin are shown (red). *D*, top view of the ribbon diagram of the complex between spastin-MIT and CHMP1b (yellow) (3eab). Side chains of the residues interacting between spastin and CHMP1b are indicated.

Supplementary Fig. 8.
Proposed model for tubulin binding with full-length kp60. Model complex between tubulin origomer (grey) and hexameric full-length kp60, composed of kp60-NTD and AAA ATPase domain (violet) is shown. AAA ATPase domains form hexameric ring. Five of the six kp60-NTDs on the hexameric AAA ATPase domains were not drawn for clarity. One of the tubulin C-terminal tail is shown in yellow. The tail on the surface of MT may bind to the pore of the hexameric AAA ATPase domain of kp60.

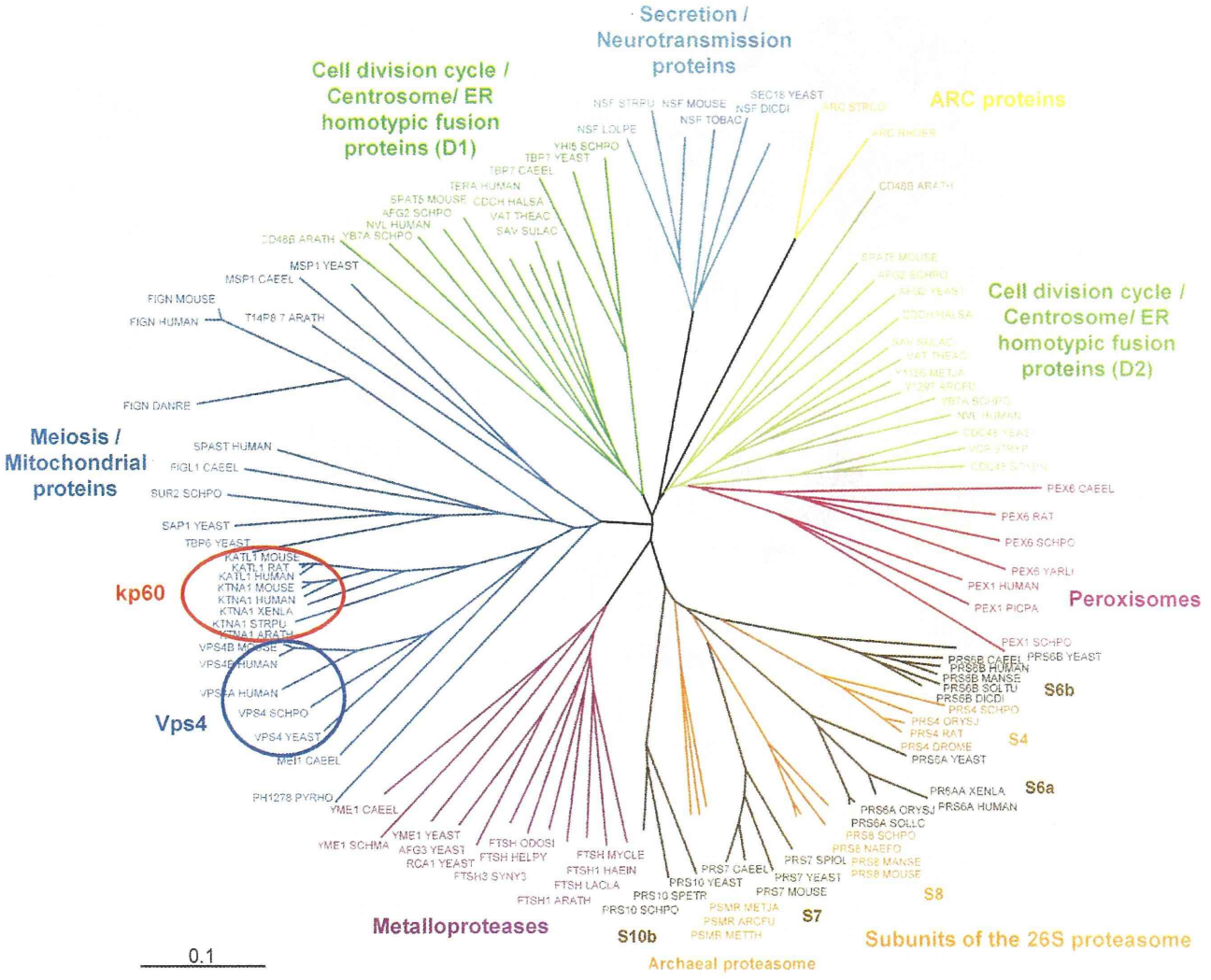
Supplementary Table 1.

Oligonucleotides used as primers for Ala substitution.

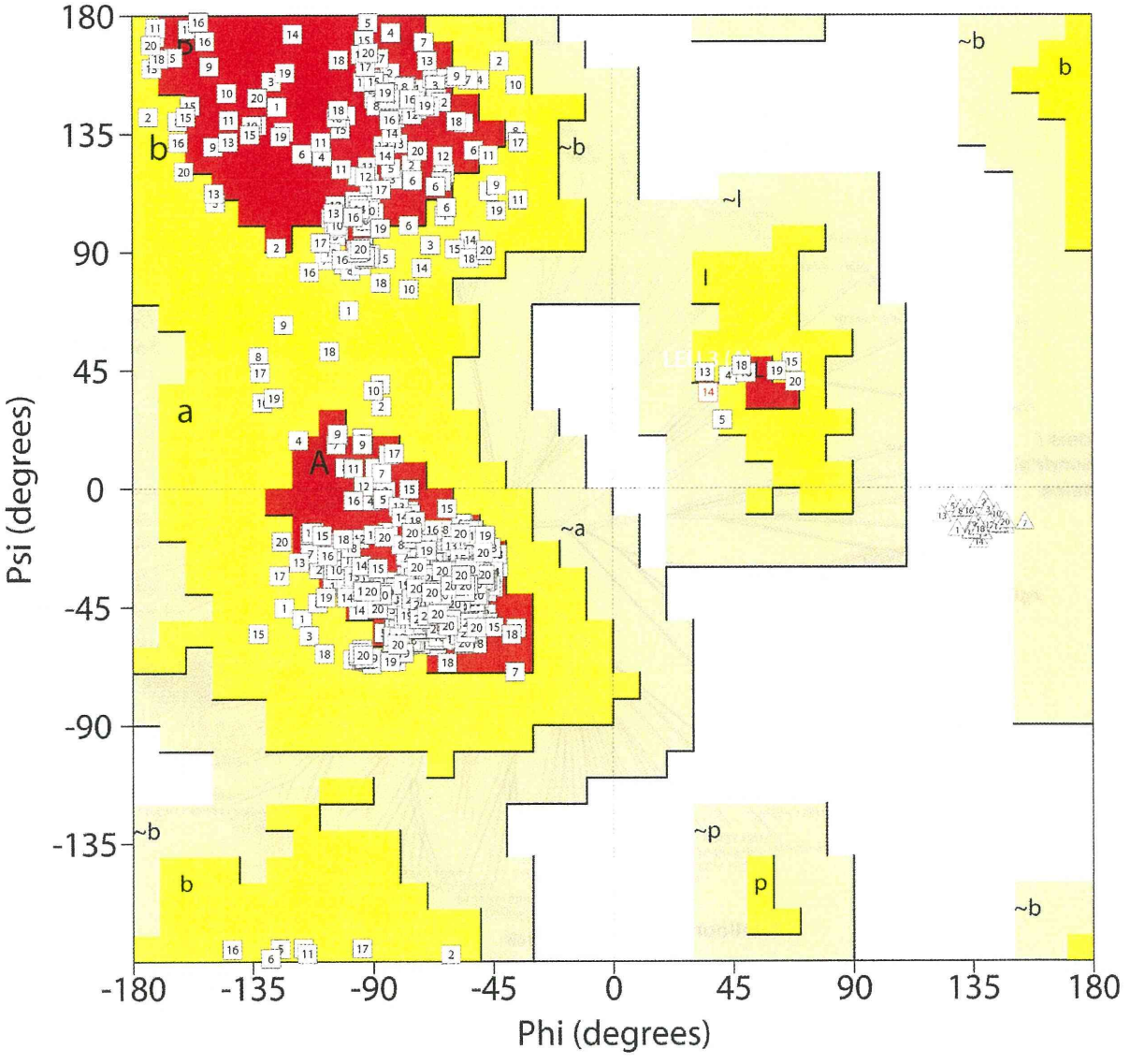
Primer	Sequence
Q35A_F	CAGGGAGTTCTTGAC**GCC**ATGAACAAGTACCTGTACTCAGTC
Q35A_R	GACTGAGTACAGGTACTTGTTCAT**GGC**GTCAAGAACTCCCTG
N37A_F	CAGGGAGTTCTTGACCAAATG**GCC**AAGTACCTGTACTCAGTC
N37A_R	GACTGAGTACAGGTACTT**GGC**CATTTGGTCAAGAACTCCCTG
D45A_F	CTGTACTCAGTCAAA**GCC**ACACACCTCCGTCAGAAATGG
D45A_R	CCATTTCTGACGGAGGTGTGT**GGC**TTTGACTGAGTACAG
R49A_F	GTCAAAGATACACACCTC**GCC**CAGAAATGGCAACAG
R49A_R	CTGTTGCCATTTCTG**GGC**GAGGTGTGTATCTTTGAC
Q53A_F	CTCCGTCAGAAATGG**GCC**CAGGTTTGGCAGGAAATAAATGTG
Q53A_R	CACATTTATTTCCTGCCAAACCTG**GGC**CCATTTCTGACGGAG
V55A_F	CTCCGTCAGAAATGGCAACAG**GCC**TGGCAGGAAATAAATGTG
V55A_R	CACATTTATTTCCTGCCA**GGC**CTGTTGCCATTTCTGACGGAG
E58A_F	CAGAAATGGCAACAGGTTTGGCAG**GCC**ATAAATGTGGAAGCTAAG
E58A_R	CTTAGCTTCCACATTTAT**GGC**CTGCCAAACCTGTTGCCATTTCTG
K64A_F	GTTTGGCAGGAAATAAATGTGGAAGCT**GCC**CAAGTTAAGGATATCATG
K64A_R	CATGATATCCTTAACTTG**GGC**AGCTTCCACATTTATTTCCTGCCAAAC
K67A_F	GTGGAAGCTAAGCAAGTT**GCC**GATATCATGAAAACATAATAGAGC
K67A_R	GCTCTATTATGTTTTCATGATATC**GGC**AACTTGCTTAGCTTCCAC
D68A_F	GTGGAAGCTAAGCAAGTTAAG**GCC**ATCATGAAAACATAATAGAGC
D68A_R	GCTCTATTATGTTTTCATGAT**GGC**CTTAACTTGCTTAGCTTCCAC

SUPPLEMENTARY REFERENCES

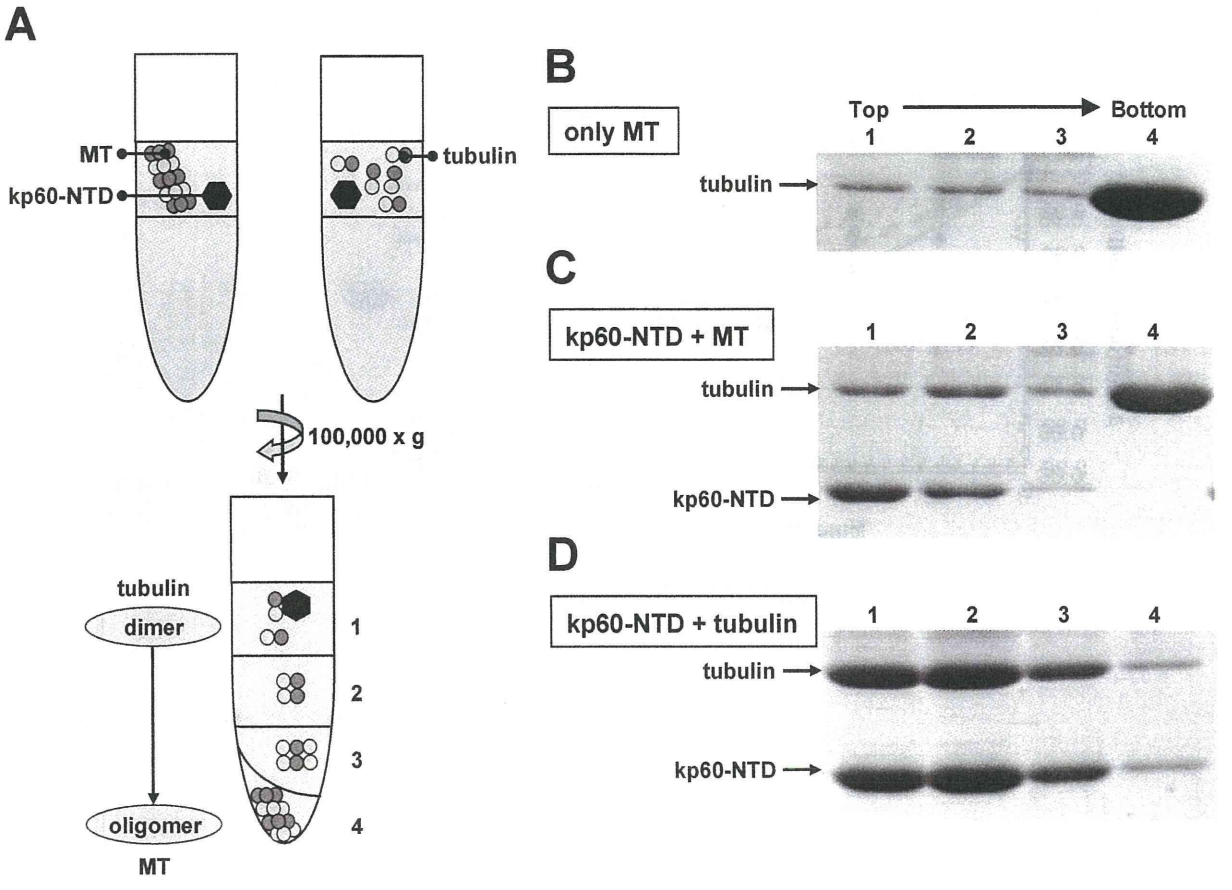
1. Hackney, D. D., and Jiang, W. (2001) *Methods Mol Biol* **164**, 65-71
2. Koradi, R., Billeter, M., and Wuthrich, K. (1996) *J Mol Graph* **14**, 29-32
3. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997) *Nucleic Acids Res* **25**, 4876-4882
4. Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996) *J Biomol NMR* **8**, 477-486



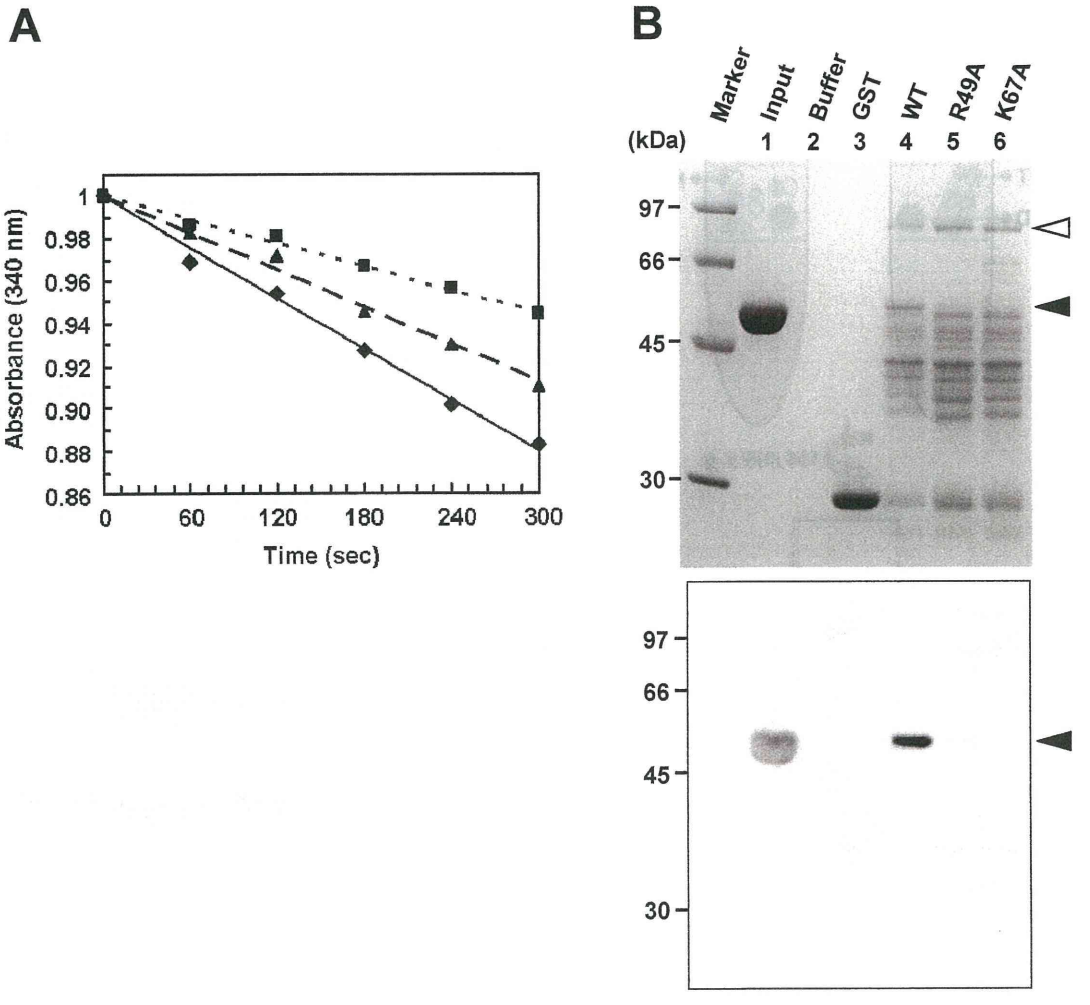
Supplementary Figure 1



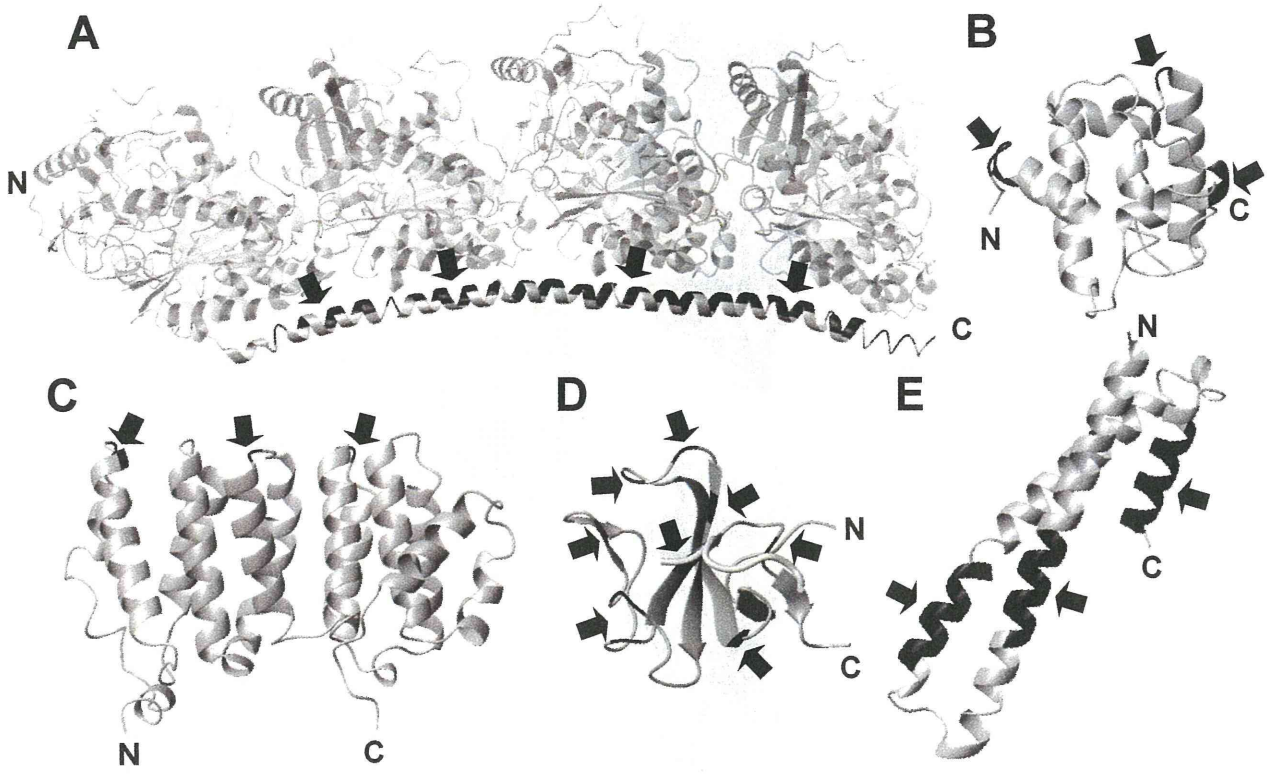
Supplementary Figure 2



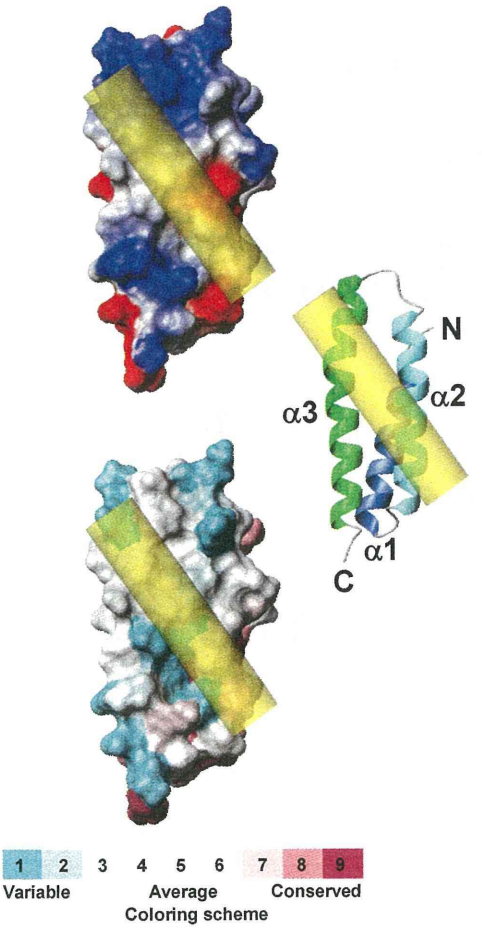
Supplementary Figure 3



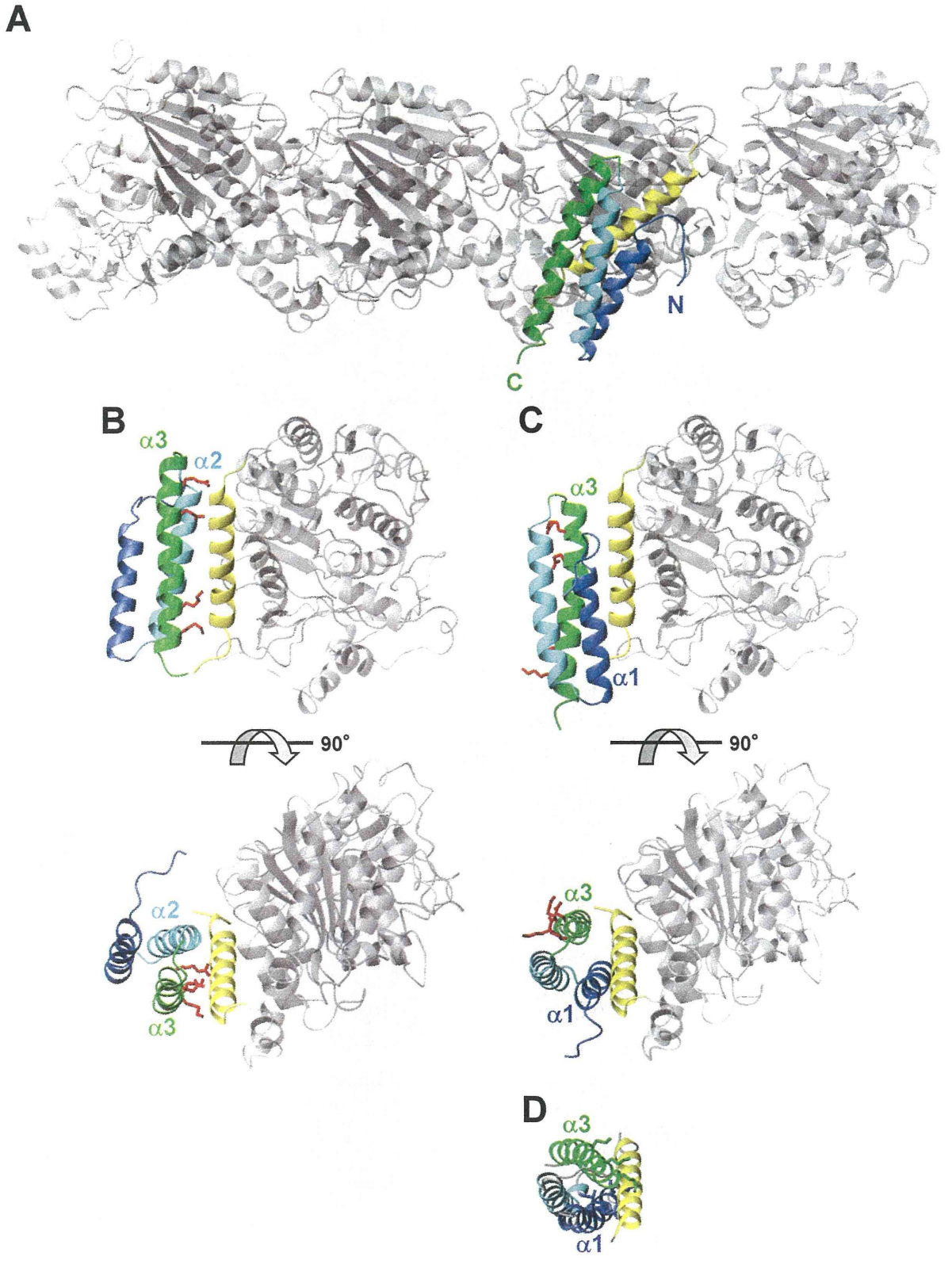
Supplementary Figure 4



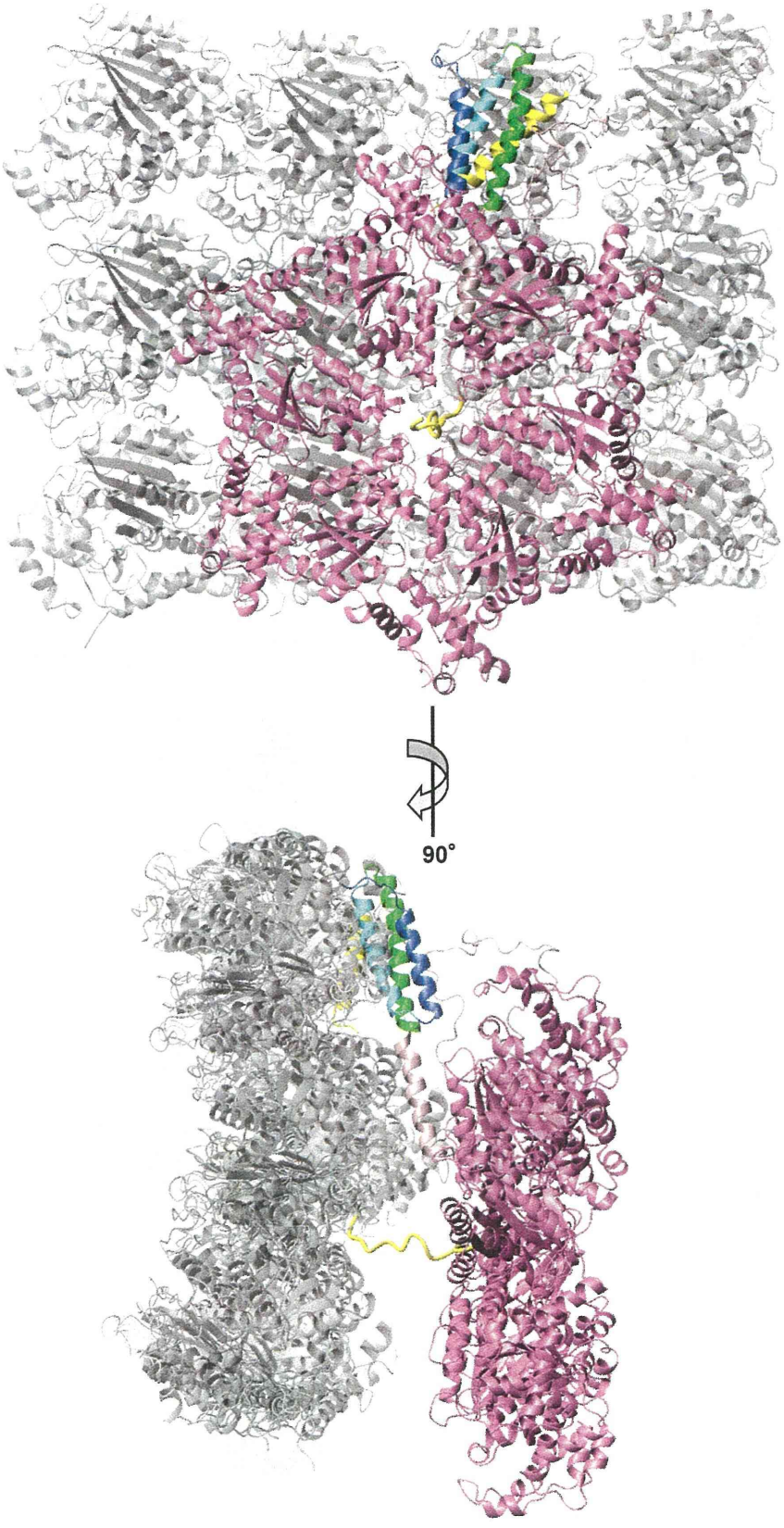
Supplementary Figure 5



Supplementary Figure 6



Supplementary Figure 7



Supplementary Figure 8

Published online 3 November 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D487–D493***
doi:10.1093/nar/gkq1057

SAHG, a comprehensive database of predicted structures of all human proteins

Chie Motono[1,2,*], Junichi Nakata[1,2], Ryotaro Koike[2,3], Kana Shimizu[1,2], Matsuyuki Shirota[2,4], Takayuki Amemiya[2,5], Kentaro Tomii[1,2], Nozomi Nagano[1,2], Naofumi Sakaya[1,2,6], Kiyotaka Misoo[1,2,6], Miwa Sato[1,2,5,7], Akinori Kidera[2,5,8], Hidekazu Hiroaki[2,9], Tsuyoshi Shirai[2,10], Kengo Kinoshita[2,4], Tamotsu Noguchi[1,2] and Motonori Ota[2,3,*]

[1]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, [2]Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST), Tokyo 102-0081, [3]Graduate School of Information Science, Nagoya University, Nagoya 464-8601, [4]Graduate School of Information Science, Tohoku University, Sendai 980-8579, [5]Department of Supramolecular Biology, Yokohama City University, Yokohama 230-0045, [6]Information and Mathematical Science Laboratory Inc., Tokyo 112-0012, [7]Mitsui Knowledge Industry Co., Ltd, Tokyo 105-6215, [8]Department of Computational Science Research Program, RIKEN, Wako 351-0198, [9]Graduate School of Medicine, Kobe University, Kobe 650-0017 and [10]Department of Bioscience, Nagahama Institute of Bioscience and Technology, Nagahama 526-0829, Japan

Received August 15, 2010; Revised October 2, 2010; Accepted October 13, 2010

ABSTRACT

Most proteins from higher organisms are known to be multi-domain proteins and contain substantial numbers of intrinsically disordered (ID) regions. To analyse such protein sequences, those from human for instance, we developed a special protein-structure-prediction pipeline and accumulated the products in the Structure Atlas of Human Genome (SAHG) database at http://bird.cbrc.jp/sahg. With the pipeline, human proteins were examined by local alignment methods (BLAST, PSI-BLAST and Smith–Waterman profile–profile alignment), global–local alignment methods (FORTE) and prediction tools for ID regions (POODLE-S) and homology modeling (MODELLER). Conformational changes of protein models upon ligand-binding were predicted by simultaneous modeling using templates of apo and holo forms. When there were no suitable templates for holo forms and the apo models were accurate, we prepared holo models using prediction methods for ligand-binding (eF-seek) and conformational change (the elastic network model and the linear response theory). Models are displayed as animated images. As of July 2010, SAHG contains 42 581 protein-domain models in approximately 24 900 unique human protein sequences from the RefSeq database. Annotation of models with functional information and links to other databases such as EzCatDB, InterPro or HPRD are also provided to facilitate understanding the protein structure-function relationships.

INTRODUCTION

Nowadays, genome sequencing projects are producing complete genome sequences at an extremely high rate (1,2). With the rise of next-gen sequencers (3–5), this is the continuous trend for the future without a doubt. Consequently, the number of known protein sequences (6) grows more rapidly than the number of known protein structures experimentally determined (7). However, to make full use of genome sequences, proteins encoded in genomes should be analysed and for this purpose, protein three-dimensional (3D) structures provide much information (8,9). Computational methods for protein 3D structure prediction are anticipated to

*To whom correspondence should be addressed. Tel: +81 3 3599 8067; Fax: +81 3 3599 8081; Email: c-motono@aist.go.jp
Correspondence may also be addressed to Motonori Ota. Tel: +81 52 789 4782; Fax: +81 52 789 4782; Email: mota@is.nagoya-u.ac.jp

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

bridge the gap between the number of known protein sequences and the number of known protein structures. According to assessments of the accuracy of those methods, e.g. recent Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiments (10,11), template-based protein structure prediction often produced 3D models accurate enough for functional annotations, modification of protein functions or even for structure-based drug design (12,13). In addition, in the CASP7 and 8 experiments, fully automated structure prediction methods had reached a comparable level to the best prediction performance by methods with human intervention (14).

In the CASP experiments, target protein sequences are ones whose 3D structures will be determined. It means that such protein structures are expected to be single domains or a couple of domains and suitable for the experimental structure determination. Therefore, sometimes protein sequences are truncated from their full-length forms. On the other hand, most protein sequences coded in genomes from higher organisms are known to be long and should be multi-domain proteins (15), and contain a significant portion of intrinsically disordered (ID) regions (16–19). Clearly, these proteins are unsuitable for experimental structure determination in the full-length form and distinct from the target protein sequences of CASPs. To analyse such proteins, we have developed a special protein-structure-prediction pipeline, by integrating and arranging various computational tools, either developed by us or widely used as global standards. This pipeline was applied to all proteins coded in the human genome. The resulting 3D models as well as other annotations for protein functions were accumulated in the Structural Atlas of Human Genome (SAHG) database and presented through the web interface at http://bird.cbrc.jp/sahg.

There are other databases of protein structure models, e.g. SWISS-MODEL Repository (20) or ModBase (21). Both databases contain annotated protein structure models generated by original automated modeling pipelines. They also allow the users to build models on demand. Compared with them, the SAHG database is distinct mainly in the following points: (i) The 3D models in SAHG were generated by an original pipeline, specific for multi-domain proteins with substantial ID regions; (ii) Conformational changes of proteins upon ligand-binding are predicted by simultaneous modeling using templates of the ligand-bound state (holo form) and the unbound state (apo form) and displayed as animated images; and (iii) Functional annotations for protein interactions, e.g. ligand-binding and protein–protein interactions, are available. All these features are suitable for analysing eukaryotic proteins toward a deep understanding of their functions and interactions.

PREDICTION SCHEME AND CONTENTS

Overview

Schematically, two types of prediction systems were used to analyse protein sequences [RefSeq sequence (22)] automatically. One is the 'Structure prediction pipeline' (right pink regions in Figure 1) in which several homology search and protein structure prediction tools, conducting sequence–sequence, sequence–profile and profile–profile alignments, are combined sequentially, and it processes protein sequences, assigns them with 3D templates and finally produces 3D models. If available, 3D models of apo and holo forms were generated. The other components are 'Other structure and function predictors' (bottom light blue regions in Figure 1). They are an ensemble of independent prediction tools, which analyse protein sequences. All the results from these systems were accumulated in SAHG in XML formats.

Structure prediction pipeline

Construction of 3D models. Protein structure prediction consists of the following procedures: template searches and selection, alignment of target sequence and template, building 3D models and evaluation of model quality.

The template searches and their assignments to a target protein are the 'step-wise-multi-methods' approach. In the first step, a BLAST (23) search against all the latest Protein Data Bank (PDB) (7) and Structure Classification of Proteins (SCOP) (24,25) sequences is performed with 10^{-5} *E*-value cut-off. We selected templates, at least 90% of whose sequence could be aligned with the target, to ensure that the 3D models corresponded to stable domains or proteins. The resulting target sequence-template alignments were ranked based on their *E*-values. The best combination of templates for each domain was determined using an original algorithm to maximize the coverage of the target sequence (label I in Figure 1). In the second step, a PSI-BLAST (23) search with the same parameters was conducted for the remaining regions of the target sequence, where no models had been assigned and the best templates were assigned onto the target sequence (II in Figure 1). Protein sequence profiles were prepared using the latest NCBI-nr database. In the third step, a Smith–Waterman profile–profile alignment method (SWPPA) (26) was applied to the remaining regions against restricted templates (SCOP and PDB subsets with less than 40% sequence identity) with a cut-off of Z-score > 10, the comparable threshold to E-value $< 10^{-5}$ in PSI-BLAST (III in Figure 1). Finally, the FORTE (27) search, a profile–profile comparison method, was performed for the remaining regions, with a strict cut-off of Z-score > 20, to detect distantly related templates (V in Figure 1). FORTE is based on the global–local alignment method and was adjusted to perform best (28) when the target proteins were almost the same length as the PDB entries (around 400 aa) (29). However, more than half of human proteins (53%) are larger than 400 amino acids and even the remaining regions are sometimes over 2000 amino acids. Thus, prior to the FORTE search, potential domains were carved out from the remaining regions using an algorithm based on the prediction of ID regions (IV in Figure 1) and fed into FORTE (see 'Prediction of potential domains' section for details).

Once the target sequence-template alignments were obtained, all templates were checked against our 'apo

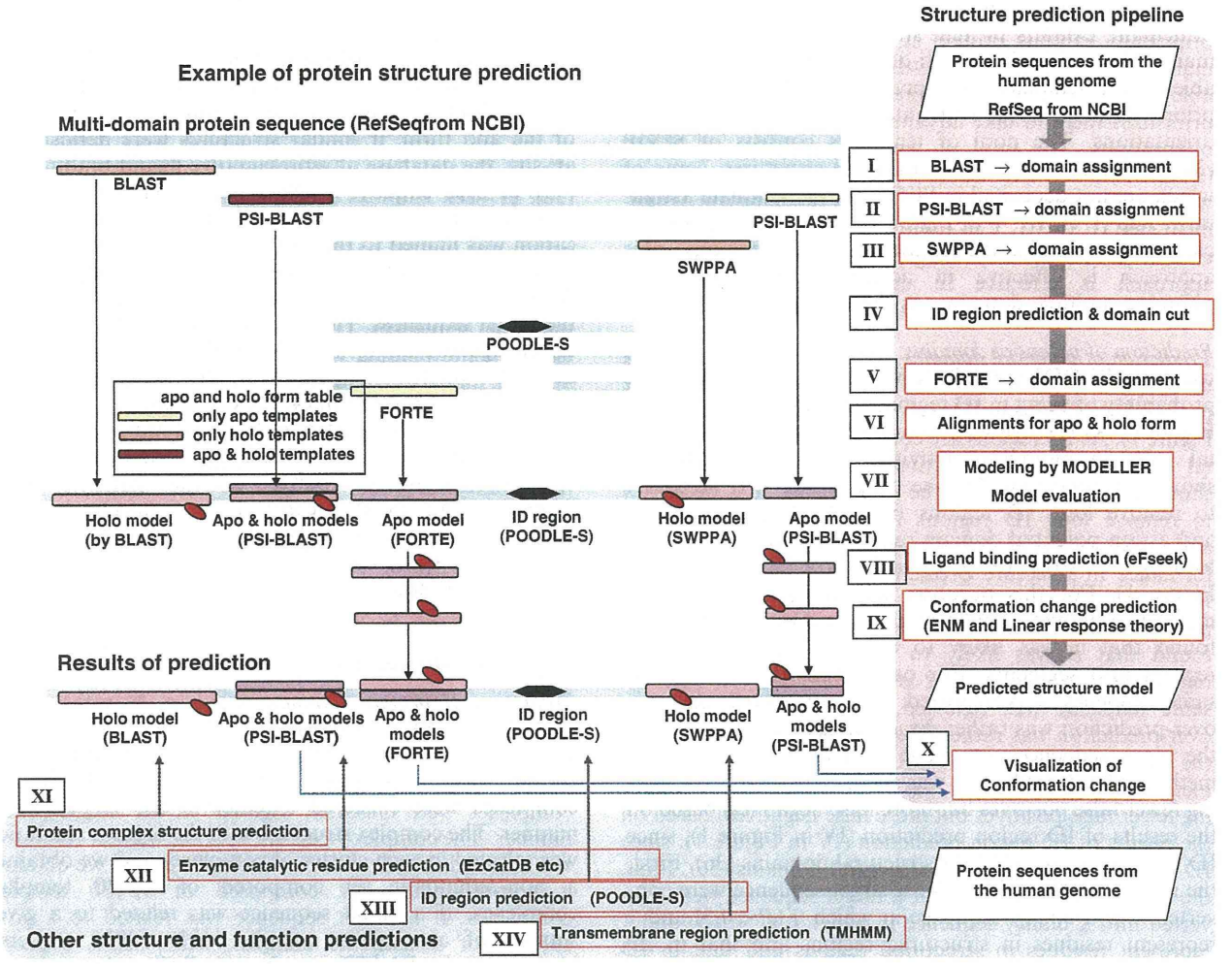


Figure 1. SAHG prediction systems. 'Structure prediction pipeline' and 'Other structure and function predictions' are shown in the right pink regions and bottom light-blue regions, respectively. The center panel illustrates each procedure in the flow of the structure prediction pipeline, showing how the results of systems are integrated. SWPPA: Smith–Waterman profile–profile alignment method; ID: intrinsically disordered; ENM: elastic network model.

and holo form table' originally prepared by us (see 'Apo and holo form table' section in Supplementary data). For the template in apo form, the corresponding template (>90% sequence identity) in holo form was selected from the table and *vice versa*. For both the templates, alignments to target sequences were prepared (VI in Figure 1). In the model building and quality assessment step, 10 models were constructed using the MODELLER (30) software. The quality of the models was evaluated using Stability score (31) and the best 3D model for each alignment was chosen (VII in Figure 1).

As of July 2010, 24 878 RefSeq sequences [(22), 14 012 591 residues] encoded in the human genome were processed by the pipeline. In total, 42 581 structure models were constructed, of which 18 228, 14 577, 9163 and 613 templates were detected by BLAST, PSI-BLAST, SWPPA and FORTE, respectively. For 4083 models (9% of all models), both the apo and holo forms were assigned. In total, 35 275 residues were predicted to form long ID regions and removed from target sequences, in advance of the FORTE search. In total, 295 309 residues were eliminated because they were fragmented into small pieces (<26 residues). Multiple models were generated for 9057 RefSeq sequences, while only one model was generated for 12 310 RefSeq sequences. In total, 3511 RefSeq sequences remain without any predicted model. Note that one model does not necessarily correspond to one domain (sometimes it corresponds to a protein chain), but at least more than one-third of human proteins were estimated to be multi-domain proteins. In some cases, we assessed predictions by comparing models with the protein structures recently revealed. Even the sequence identities of the alignments are quite low (<20%), more than half predictions detect correct folds (Supplementary Table S1), indicating that our prediction pipeline worked well.

Treatments of multi domain proteins. Many human proteins are composed of multiple domains and contain a significant fraction of ID regions, as was described above. These factors often prevent predicting protein

structures in their full-length forms. As a result, SAHG principally exhibits protein structure as an array of domains. However, when multi-domain structures are available in the templates, the prediction pipeline implicitly prioritizes them to take advantage of the relative domain orientations. The pool of templates consists of SCOP (24,25) domains and whole PDB (7) structures, some of which are not deposited in SCOP. At the template assignment step (I, II, III, V in Figure 1), a set of templates was chosen to maximize the length of modeled regions. This approach is effective in accepting PDB structures spanning multiple domains, as the templates.

Prediction of potential domains. ID regions were predicted using the POODLE-S (18) software, which calculates the probability of being in ID regions for each residue (XIII in Figure 1). As ID regions are considered to play fundamental roles in biological activities (17), their detections should be important. On the other hand, it is necessary to remove long ID regions from the target sequences and assign potential domain regions to assure better performance in structure prediction (FORTE search, V in Figure 1). For this purpose, we evaluated an existing method to predict domain boundaries [Domcut (32)] and found that it was likely to overcut potential domain regions into segments. For other methods (33–35), the same tendency was reported. We considered that the over-prediction was rather disadvantageous for arranging the input sequences for FORTE and developed a new method whose prediction was more 'moderate' (containing fewer false positives but more false negatives) based on the results of ID region prediction (IV in Figure 1), since ID regions act as linkers of structural domains (36). First, the results of POODLE-S for a target sequence were converted into a binary sequence in which 0 ($P < 0.5$) and 1 represent residues in structured regions and that in ID regions, respectively. Next, to detect regions where 0 were continuously abundant, we employed a simple two-state Hidden Markov Model. In this model, one state, 'a mostly structured region' (STR), emits 0 more frequently than 1 and the other state, 'a mostly ID region' (IDR), emits 1 more frequently than 0. The transition probability between STR and IDR and all the emission probabilities were empirically adjusted to eliminate over-prediction by referring to known domain data in PDB. Finally, the STR regions were estimated from the input binary sequence by calculating a Viterbi path.

Prediction of conformational change upon ligand binding. When templates for both the ligand-bound state (holo form) and unbound state (apo form) were detected using the 'apo and holo form table', two types of models were constructed and their structural changes upon ligand-binding are visualized by means of a morphing technique (the MORPH2 program in Martz-Authored PDB Tools see http://www.umass.edu/microbio/rasmol/pdbtools.htm) (X in Figure 1). The animation of conformational change provides significant information for protein function when it is shown with functional residues and ligands.

When there was only the template for apo form available and accordingly, only the model for apo form was constructed, its putative ligand and the binding sites were predicted by the eF-seek software (37) (VIII in Figure 1). eF-seek finds potential ligand-binding sites in the model of the apo form, if similar structures were deposited in eF-site, the database of representative ligand-binding sites (38). eF-seek employs a clique search algorithm. As this method is sensitive to the input 3D coordinates, the application was limited to the case of highly accurate structure models being available, i.e. the templates were detected by BLAST search with more than 90% sequence identity to the target sequences. The structural changes upon the predicted ligand-binding were then deduced using the elastic network model (39) and linear response theory to construct a model of the holo form (40) (IX in Figure 1).

Note that this approach and presentation is one of the key features of the SAHG database. Animated views of the conformational change of the domains upon ligand-binding could present a deep insight into the protein structure and function relationship (X in Figure 1). As of July 2010, conformational changes upon ligand-binding were predicted for 4083 modeled domains among 42 581 3D models.

Other structure and function predictors

Prediction of protein complex structure. In total, 33 687 protein complex structures were gathered from the PQS database (41). If all the subunits from two complexes were paired with more than 95% sequence identity, the complexes were clustered together in the single-linkage manner. The complex structure with the highest resolution was selected in each cluster of complexes and we obtained a non-redundant set composed of 12 730 template complexes. If a target sequence was related to a given subunit of a template complex with >80% sequence identity by the BLAST search and all the other subunits were related to any target sequences, the complex model was constructed by MODELLER. In total, 8667 complex models were prepared for 3650 target sequences (XI in Figure 1).

Ligand binding information. The ligands and their binding sites were retrieved from constructed models. The ligands were mainly small molecules, such as peptides, nucleotides, metal ions, etc. and some trivial chemicals from buffers or precipitants were excluded. Binding sites were residues whose distances from any ligand atoms were within 5 Å.

Prediction of catalytic residues. For the target sequences of enzymes, catalytic residues were predicted using the EzCatDB database (42) (XII in Figure 1). The EzCatDB database provides annotations on catalytic residues with PDB structure data. The catalytic residues and their positions were already denoted for sequences in the UniProt database (6), as mapped from the catalytic residues on the PDB sequence data, by BLAST search with 10^{-10} *E*-value cut-off and POA ver. 2.0 (43). From the human proteins in the UniProt database, target sequences were detected and catalytic residues were assigned in the same manner. Only

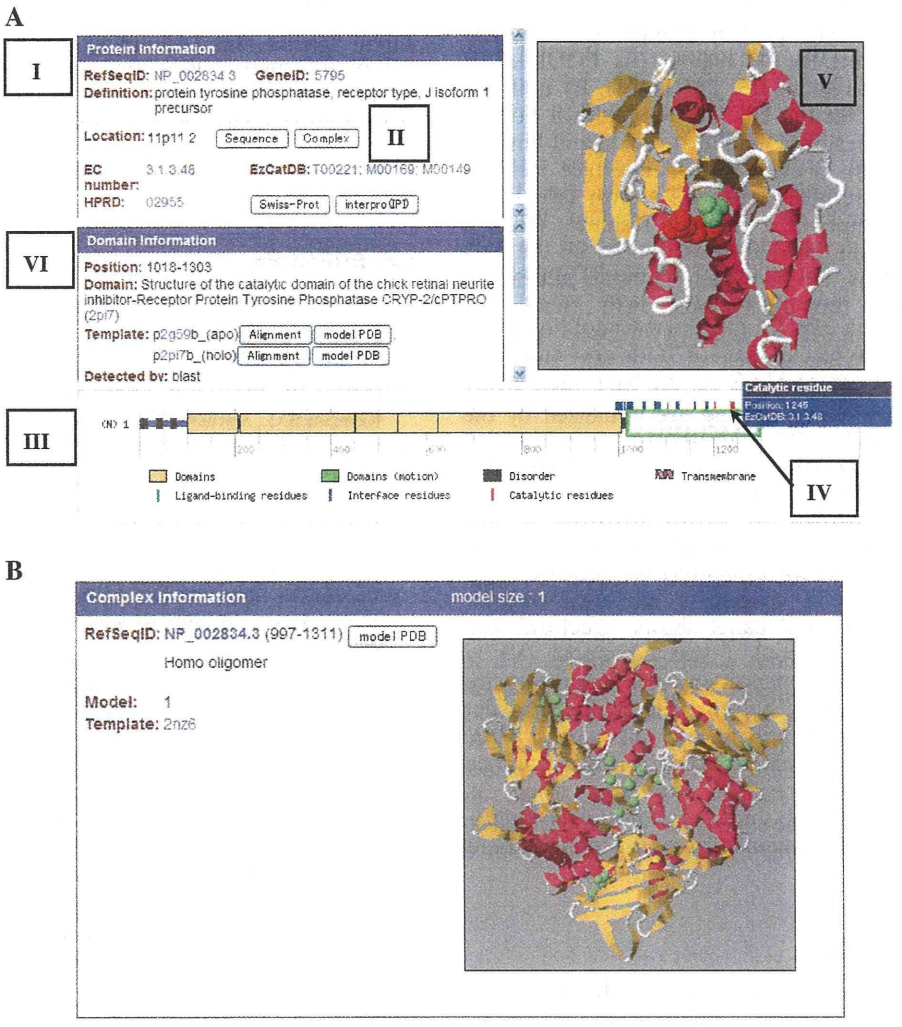


Figure 2. (**A**) Example view of SAHGs detailed information page [RefSeqID: NP_002834.3, protein tyrosine phosphatase, receptor type, J isoform 1 precursor (48)]. Labels I, II, III, IV, V and VI indicate the 'Protein information' panel, the 'Complex' button, the 'bar indicator', the 'Domain information' panel, the 'Jmol Window' and the 'Catalytic residue' pin on the bar indicator, respectively. (**B**) Example view of a 'Complex information' page (NP_002834.3). For this protein, only one complex structure in a homo-trimeric form was predicted.

chemically consistent residues were regarded as catalytic residues. The annotated 'ACT_SITE' residues for the human proteins in the UniProt database were also mapped on the target sequences using BLAST search.

Prediction of ID and transmembrane regions. ID regions were predicted by the POODLE-S software (XIII in Figure 1). Transmembrane regions were assigned by the TMHMM software (44) (XIV in Figure 1). If these predicted regions were overlapped with 3D models, the latter take priority over the former.

ACCESS AND INTERFACE

SAHG provides its graphical web interface at http://bird.cbrc.jp/sahg. By clicking a chromosome's image, all proteins coded in the chromosome are listed with the predicted models. By choosing an image of a domain, detailed information of the target protein is shown. More practically, detailed information of specific proteins can be accessed by querying with Gene ID, RefSeq ID, annotation keywords or their combinations or by sequence homology search (BLAST), from an 'Advanced search page'. In the detailed information page (Figure 2A), all contents for a given protein are shown. The 'Protein information' panel provides the information of the protein's RefSeq ID (I in Figure 2A). The sequence in FASTA format is displayed by clicking a 'Sequence' button. Predicted protein complexes are shown via a 'Complex' button if available (II in Figure 2A). An example of a 'complex information' page is shown in Figure 2B. Links to EC number, EzCatDB (42), HPRD (45), Swiss-Prot(6) and InterPro (46) are provided if available. A bar indicator is convenient for seeing the position of the predicted models in the full-length protein (III in Figure 2A). It also shows the annotation of ligand-binding

residues (retrieved from the holo models), protein–protein interface residues (from protein complexes), catalytic residues (from EzCatDB), ID regions (by POODLE-S) and transmembrane regions (by TMHMM). By pointing at the colored pins on the bar indicator with a mouse, precise locations (residue numbers) of ligand-binding residues (green pins), protein–protein interface residues (blue) or catalytic residues (red) are shown (see IV in Figure 2A, an example of a catalytic residue). When a modeled region in the bar indicator (blocks on the bar) is selected by clicking, the predicted 3D model appears in the Jmol window (an open-source Java viewer for chemical structures in 3D; see http://www.jmol.org/Jmol) (V in Figure 2A). When models of both apo and holo forms are available (green block on the bar), their structural changes upon ligand-binding are visualized by the morphing technique (the MORPH2 program in Martz-Authored PDB Tools; see http://www.umass.edu/microbio/rasmol/pdbtools.htm) and displayed as an animated image including the ligand molecules in this window. By clicking the bar indicator of ligand-binding or catalytic residues, the corresponding residues are highlighted in 'CPK spacefill' scheme in the Jmol window. The 'Domain Information' panel shows structural and functional information about a selected model (VI in Figure 2A). The target sequence-template alignments are displayed by an 'Alignment button'. The predicted model can be downloaded in a pdb format via 'model PDB' button. Ligand-binding residues, protein–protein interface residues and catalytic residues are also listed as 'Functional Residues' in the same color of the bar indicator. (In Figure 2A, the 'Domain information' panel should be scrolled up).

FUTURE DIRECTIONS

To improve the accuracy of structure prediction we are implementing a probabilistic profile–profile alignment method in our prediction pipeline. The method is an enhanced version of the probabilistic sequence–sequence alignment method (47), which has been proven to perform better than PSI-BLAST, in particular for orphan proteins. New versions of structure models provided by the new pipeline will appear in fall of 2010. The results of predictions are being examined to clarify the function and the interaction of human proteins. For some proteins, predicted ligands are being verified experimentally. The structure model set in SAHG will be downloadable in bulk in future.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to Takatsugu Hirokawa and Kiyoshi Asai for their support of the project, to Martin Frith for his critical reading of the article and to Mari Saito for her contribution to website design.

FUNDING

Japan Science and Technology Agency (JST) – Institute for Bioinformatics Research and Development (BIRD). Funding for open access charge: National Institute of Advanced Industrial Science and Technology (AIST).

Conflict of interest statement. None declared.

REFERENCES

1. Nelson,K.E., Weinstock,G.M., Highlander,S.K., Worley,K.C., Creasy,H.H., Wortman,J.R., Rusch,D.B., Mitreva,M., Sodergren,E., Chinwalla,A.T. *et al.* (2010) A catalog of reference genomes from the human microbiome. *Science*, **328**, 994–999.
2. Drmanac,R., Sparks,A.B., Callow,M.J., Halpern,A.L., Burns,N.L., Kermani,B.G., Carnevali,P., Nazarenko,I., Nilsen,G.B., Yeung,G. *et al.* (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, **327**, 78–81.
3. Zhang,W. and Dolan,M.E. (2010) Impact of the 1000 genomes project on the next wave of pharmacogenomic discovery. *Pharmacogenomics*, **11**, 249–256.
4. Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**, 31–46.
5. MacLean,D., Jones,J.D. and Studholme,D.J. (2009) Application of 'next-generation' sequencing technologies to microbial genetics. *Nat. Rev. Microbiol.*, **7**, 287–296.
6. Consortium,U. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
7. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33**, D233–D237.
8. Xie,L. and Bourne,P.E. (2005) Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. *PLoS Comput. Biol.*, **1**, e31.
9. Thornton,J.M., Todd,A.E., Milburn,D., Borkakoti,N. and Orengo,C.A. (2000) From structure to function: approaches and limitations. *Nat. Struct. Biol.*, **7(Suppl.)**, 991–994.
10. Cozzetto,D., Kryshtafovych,A., Fidelis,K., Moult,J., Rost,B. and Tramontano,A. (2009) Evaluation of template-based models in CASP8 with standard measures. *Proteins*, **77(Suppl. 9)**, 18–28.
11. Kopp,J., Bordoli,L., Battey,J.N., Kiefer,F. and Schwede,T. (2007) Assessment of CASP7 predictions for template-based modeling targets. *Proteins*, **69(Suppl. 8)**, 38–56.
12. Grant,M.A. (2009) Protein structure prediction in structure-based ligand design and virtual screening. *Comb. Chem. High Throughput Screen.*, **12**, 940–960.
13. Katritch,V., Rueda,M., Lam,P.C., Yeager,M. and Abagyan,R. (2010) GPCR 3D homology models for ligand screening: lessons learned from blind predictions of adenosine A2a receptor complex. *Proteins*, **78**, 197–211.
14. Zhang,Y. (2009) I-TASSER: fully automated protein structure prediction in CASP8. *Proteins*, **77(Suppl. 9)**, 100–113.
15. Apic,G., Gough,J. and Teichmann,S.A. (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.*, **310**, 311–325.
16. Dunker,A.K., Obradovic,Z., Romero,P., Garner,E.C. and Brown,C.J. (2000) Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.*, **11**, 161–171.
17. Dunker,A.K., Silman,I., Uversky,V.N. and Sussman,J.L. (2008) Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.*, **18**, 756–764.
18. Shimizu,K., Muraoka,Y., Hirose,S., Tomii,K. and Noguchi,T. (2007) Predicting mostly disordered proteins by using structure-unknown protein data. *BMC Bioinformatics*, **8**, 78.
19. Ward,J.J., Sodhi,J.S., McGuffin,L.J., Buxton,B.F. and Jones,D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.

20. Kiefer,F., Arnold,K., Kunzli,M., Bordoli,L. and Schwede,T. (2009) The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res.*, **37**, D387–D392.
21. Pieper,U., Eswar,N., Webb,B.M., Eramian,D., Kelly,L., Barkan,D.T., Carter,H., Mankoo,P., Karchin,R., Marti-Renom,M.A. *et al.* (2009) MODBASE, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.*, **37**, D347–D354.
22. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**, D32–D36.
23. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
24. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
25. Chandonia,J.M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.
26. Wang,G. and Dunbrack,R.L. Jr (2004) Scoring profile-to-profile sequence alignments. *Protein Sci.*, **13**, 1612–1626.
27. Tomii,K. and Akiyama,Y. (2004) FORTE: a profile-profile comparison tool for protein fold recognition. *Bioinformatics*, **20**, 594–595.
28. Tomii,K., Hirokawa,T. and Motono,C. (2005) Protein structure prediction using a variety of profile libraries and 3D verification. *Proteins*, **61(Suppl. 7)**, 114–121.
29. Thornton,J.M., Orengo,C.A., Todd,A.E. and Pearl,F.M. (1999) Protein folds, functions and evolution. *J. Mol. Biol.*, **293**, 333–342.
30. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
31. Ota,M., Isogai,Y. and Nishikawa,K. (2001) Knowledge-based potential defined for a rotamer library to design protein sequences. *Protein Eng.*, **14**, 557–564.
32. Suyama,M. and Ohara,O. (2003) DomCut: prediction of inter-domain linker regions in amino acid sequences. *Bioinformatics*, **19**, 673–674.
33. Cheng,J. (2007) DOMAC: an accurate, hybrid protein domain prediction server. *Nucleic Acids Res.*, **35**, W354–W356.
34. Ebina,T., Toh,H. and Kuroda,Y. (2009) Loop-length-dependent SVM prediction of domain linkers for high-throughput structural proteomics. *Biopolymers*, **92**, 1–8.
35. Kim,D.E., Chivian,D., Malmstrom,L. and Baker,D. (2005) Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. *Proteins*, **61(Suppl. 7)**, 193–200.
36. Dyson,H.J. and Wright,P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, **6**, 197–208.
37. Kinoshita,K., Murakami,Y. and Nakamura,H. (2007) eF-seek: prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape. *Nucleic Acids Res.*, **35**, W398–W402.
38. Kinoshita,K. and Nakamura,H. (2004) eF-site and PDBjViewer: database and viewer for protein functional sites. *Bioinformatics*, **20**, 1329–1330.
39. Tirion,M.M. (1996) Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.*, **77**, 1905–1908.
40. Ikeguchi,M., Ueno,J., Sato,M. and Kidera,A. (2005) Protein structural change upon ligand binding: linear response theory. *Phys. Rev. Lett.*, **94**, 078102.
41. Henrick,K. and Thornton,J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358–361.
42. Nagano,N. (2005) EzCatDB: the enzyme catalytic-mechanism database. *Nucleic Acids Res.*, **33**, D407–D412.
43. Grasso,C. and Lee,C. (2004) Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. *Bioinformatics*, **20**, 1546–1556.
44. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
45. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human protein reference database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
46. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
47. Koike,R., Kinoshita,K. and Kidera,A. (2007) Probabilistic alignment detects remote homology in a pair of protein sequences without homologous sequence information. *Proteins*, **66**, 655–663.
48. Ostman,A., Yang,Q. and Tonks,N.K. (1994) Expression of DEP-1, a receptor-like protein-tyrosine-phosphatase, is enhanced with increasing cell density. *Proc. Natl Acad. Sci. USA*, **91**, 9680–9684.